Comparison of Data-Driven Bandwidth Selectors

by

Byeong Park and J. S. Marron

University of North Carolina, Chapel Hill

May 3, 1988

This paper provides a comparison, on three levels, of several promising data-driven methods for selecting the bandwidth of a kernel density estimator. The methods compared are: least squares cross-validation, biased cross-validation, partitioned cross-validation, and a plug-in rule. The levels of comparison are: asymptotic rate of convergence to the optimum, explicit calculation in the case of one target density, and a simulation study. It is seen that the plug-in bandwidth is usually most efficient when the underlying density is sufficiently smooth, but is less robust when there is not enough smoothness present.

KEY WORDS: cross-validation, data driven bandwidth selection, density estimation, kernel estimators, plug-in method
1. Introduction

Kernel density estimation is a very useful tool for exploring the distribution structure of unknown populations. See Silverman (1986) for a variety of real data examples which illustrate the power of this technique.

Practical application of this method is crucially dependent on choice of the smoothing parameter or bandwidth. While effective data analysis has often be done by a subjective, trial and error approach to this choice, the usefulness of density estimation would be greatly enhanced if an efficient and objective method of using the data to determine the amount of smoothing could be agreed upon. Hence a number of data-driven methods for choosing the bandwidth have been proposed and studied. See the survey paper Marron (1988a) for a listing of proposed methods and discussion. The present paper provides several different means for a comparison of the best known and most promising of these. Section 2 contains precise definitions of the bandwidth selectors discussed in this paper.

The most widely studied bandwidth selector is least squares cross-validation, proposed by Rudemo (1982) and Bowman (1984). The basic idea is to choose the bandwidth to minimize a score function which provides an estimate of the mean integrated squared error. This method has become the benchmark against which other selection methods are usually compared. It has been shown to have the attractive asymptotic property of giving an answer which converges to the optimum under very
weak conditions (see Stone (1984) for the best known such result). However, in a number of simulation studies and real data examples, the performance of this method has been often disappointing. The reason for this is that the least squares cross-validated bandwidth suffers from a great deal of sample variability. This large variability has been asymptotically quantified by Hall and Marron (1987a), who obtain a very slow rate of convergence of the cross-validated bandwidth to the optimum.

Because of these limitations of least squares cross-validation, there has been serious investigation made into other methods of bandwidth selection. The most appealing of these are plug-in rules, biased cross-validation and partitioned cross-validation.

A version of the plug-in selector is the first proposed method for using the data to choose the bandwidth of a kernel density estimator, see Woodroofe (1970). The basic idea is to substitute estimates into an asymptotic representation of the optimal bandwidth. Such methods have been slow to gain acceptance, because care must be taken concerning which estimates are plugged in. An effective method of overcoming the early difficulties was discovered independently by Hall (1980) and by Sheather (1983,1986).

Biased cross-validation was proposed by Scott and Terrell (1987). This method is actually a hybrid of cross-validation and plug-in methods, in that a score function is minimized as for least squares cross-validation, but the score function makes use of some plug-in ideas. The effect of this is to provide a data-driven bandwidth with
substantially less sample variability than ordinary cross-validation.

Partitioned cross-validation was proposed by Marron (1988b). The idea here is to first partition the data into subsamples, compute the least squares cross-validation score for each subsample, and use the minimum of the combined score functions. This has the effect of also cutting down on the large variability of ordinary cross-validation. A substantial drawback to this method, not shared by the others, is that one must select the number of subsamples. It was shown in Marron (1988b) that this choice can be thought of as being of the same level of difficulty as the choice of a smoothing parameter.

The main point of this paper is a comparison of these methods of smoothing parameter selection. The comparison is done on three levels. Section 3 treats asymptotic rate of convergence. Section 4 gives more explicit information in the special case of a Gaussian kernel and density. Section 5 presents the results of a simulation study.

The exponents of the asymptotic rate of convergence of the various data-driven bandwidths to the optimum provide an effective means of understanding their asymptotic performance. As for the rate of convergence of the density estimator to the density, this rate depends on the amount of smoothness, typically quantified in terms of number of bounded derivatives, of the underlying density. In Section 3, precise results, which quantify these rates as a function of smoothness, are given. The main result is that, for small amounts of smoothness (i.e. weak assumptions on the underlying density), cross-validation is the most effective. However for more smoothness (i.e. stronger
assumptions), biased cross-validation is superior to cross-validation, but the plug-in method is best of all. The reason for this is the extra estimation steps done by biased cross-validation and the plug-in rule. This extra estimation requires stronger assumptions to work effectively. When the additional smoothness is present, there is a payoff in terms of reduced variability for biased cross-validation and the plug-in rules. But when there is not enough smoothness, the additional estimation is much less effective. This trade-off is analogous to that in robustness theory. In particular, this provides a sense in which cross-validation is "more robust" at some cost in "efficiency", while the plug-in rule is "more efficient" when stronger assumptions hold.

In section 4, an idea concerning the behavior of the constant coefficients in the limiting distribution of the automatically selected bandwidths is obtained through evaluation in a special case. It is seen that some of the effects described by studying only the exponents in the rate of convergence can require very large sample sizes before they are valid.

Section 5 contains the results of a simulation study. It is seen there that once again the plug-in bandwidth usually gives quite good results. Also insight is given into what drives the various results.
2. Selection methods

The goal of kernel density estimation is to estimate a probability density \( f(x) \) using a random sample \( X_1, \ldots, X_n \) from \( f \). The kernel density estimator is given by

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

where \( K_h(\cdot) = K(\cdot/h)/h \). The function \( K \) is called the kernel. The scale parameter \( h \) is called the bandwidth, because it is the effective window width of the local averaging being done by \( \hat{f}_h \). Another name for \( h \) is smoothing parameter, which comes from the fact that \( h \) controls the roughness of the underlying estimate. It is mathematically not necessary to use \( h \) as a separate parameter, but this is done because choice of the scale \( h \) is far more important, than choice of the shape of the window \( K \), to the performance of the estimator, see section 3.3.2 of Silverman (1986) for example.

A very readable and informative introduction to many of the properties of the kernel density estimator may be found in Silverman (1986). A number of these properties will be used here without giving further reference.

In this paper it is assumed that,

(2.1) \( K \) is a symmetric probability density

(2.2) \( K \) has four Hölder continuous derivatives and satisfies the moment conditions (6.1) stated explicitly in section 6.

(2.3) \( h \in [Bn^{-1/5}, Bn^{-1/5}] \), for some constants \( 0 < B < B \) (thought of as quite small and large respectively).
Assumption (2.1) is quite standard and, except for the implied nonnegativity (see Section 3.6 of Silverman (1986) for further discussion), simply guarantees the effective performance of the estimator \( \hat{f}_h \). Assumption (2.2) is for convenience, both in stating results and in proving them. Assumption (2.3) is also for convenience, but is not restrictive because this is well known to be the range of reasonable bandwidths (under standard conditions, such as those of this paper) in the limit as \( n \to \infty \). Both (2.2) and (2.3) can be weakened substantially, however we do not do so here because the added notational burden and length of proof would tend to obscure our main points.

A common means of assessing the performance of a density estimator is the Mean Integrated Squared Error,

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

although see Devroye and Gyorfi (1984) for another viewpoint on this. Most data driven bandwidths can be viewed as an attempt to estimate \( h_{\text{MISE}} \), the minimizer of MISE(h).

The least squares cross-validated bandwidth, \( \hat{h}_{\text{CV}} \), may be motivated by finding a function of \( h \) which is an unbiased estimate of a vertical shift of MISE(h). In particular, take \( \hat{h}_{\text{CV}} \) to be the minimizer, over the range (2.3), of the cross-validation score function,

\[
\text{CV}(h) = R(\hat{f}_h) - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{j,h}(X_j),
\]

where here and at all points below, for any function \( g(x) \),

\[
R(g) = \int g(x)^2 dx,
\]

and where \( \hat{f}_{j,h} \) denotes the leave-one-out kernel estimator constructed from the data with \( X_j \) deleted. For reasonable choice of \( B \) and \( \overline{B} \)
there will usually be at least one minimizer of $CV(h)$ although local minima are known to occur reasonably often. For definiteness, we take here (there is no common agreement on this) $\hat{h}_{CV}$ to be the largest local minimizer, over the range (2.3).

The biased cross-validated bandwidth, introduced by Scott and Terrell (1987), is based on the following asymptotic representation for $\text{MISE}(h)$. If it is assumed that

(2.5) $f$ has a Hölder continuous, square integrable second derivative, then as $n \to \infty$,

$$\frac{\text{MISE}(h)}{\text{AMISE}(h)} \to 1,$$

where

(2.6) $\text{AMISE}(h) = n^{-1}h^{-1}\int k^2 + 4h^4 \sigma^4_K R(f'''),$

where here and below, for any mean zero probability density $g(x)$, the variance of the $g$ distribution is denoted by

$$\sigma^2_g = \int x^2 g(x) dx$$

Note that the only part of $\text{AMISE}(h)$ which is unavailable to the experimenter is $R(f''')$, so it makes sense to consider using estimates of this unknown quantity. Scott and Terrell propose doing this with $R(\hat{f}_h''')$ where $\hat{f}_h$ is the kernel estimator whose bandwidth $h$ is the same as that under consideration. The minimizer, over the range (2.3), of the resulting estimate of $\text{MISE}(h)$ is called $\hat{h}_{BCV}$. Section 5 discusses problems that can occur for small samples, where the minimizer is at the left endpoint of (2.3).
The plug-in idea is based on closely related considerations. In particular, under the same assumption (2.5) it can be shown that as \( n \to \infty \)

\[
h_{MISE} / h_{AMISE} \to 1,
\]

where \( h_{AMISE} \) is the minimizer of \( AMISE(h) \), and can be written as

\[
(2.7) \quad h_{AMISE} = C_0 n^{-1/5},
\]

\[
C_0 = \left( \frac{R(K)}{\sigma_K^4 R(f''')} \right)^{1/5}.
\]

At this point it is apparent that the range (2.3) should chosen so that \( B << C_0 \) and \( B >> C_0 \), so this assumption is made for the rest of the paper. As above the only unknown part of \( h_{AMISE} \) is \( R(f''') \), so it makes sense to consider estimates. A candidate for such an estimate is \( \hat{R}(\hat{f}'''') \), where \( \hat{f}_a \) is a kernel density estimate, with bandwidth now represented by \( a \) (allowed to be different from \( h \) because estimation of this integral is a different smoothing problem from estimation of \( f(x) \)). The fact that a different bandwidth is used here appears to be the crucial difference between this approach and biased cross-validation. As for the curve estimation problem, the choice of the bandwidth is crucial to the performance of this estimator of \( R(f''') \). Note in particular that for any fixed set of data, \( \hat{R}(\hat{f}'''') \) takes on all values between \( 0 \) and \( \infty \) as \( a \) ranges from \( 0 \) to \( \infty \). The approach to this problem developed by Hall (1980) and by Sheather (1983, 1986) is to find a reasonable representation of \( a \) in terms of \( h \) and then solve the resulting version of the equation (2.7) for \( h \). Such a representation comes from the fact that, if (2.2) holds and

\[
(2.8) \quad f \text{ has a Hoelder continuous, square integrable fourth derivative,}
\]
then Hall and Marron (1987b) have shown that \( a_{\text{MSE}} \), the minimizer of the Mean Squared Error,

\[
\text{MSE}(a) = E \left[ (\hat{f}_a - R(f)) \right]^2
\]

has the asymptotic, as \( n \to \infty \), representation

\[
a_{\text{MSE}}/a_{\text{AMSE}} \to 1,
\]

where \( a_{\text{MSE}} \) is the minimizer of an asymptotic representation of \( \text{MSE}(a) \) and can be expressed as

\[
a_{\text{AMSE}} = C_1(K)C_2(f)n^{-2/13},
\]

\[
C_1(K) = \left\{ 18R(K^{(4)})/\sigma_K^4 \right\}^{1/13},
\]

\[
C_2(f) = \left\{ (R(f)/R(f^{(3)})^2 \right\}^{1/13},
\]

where \( K \ast K \) denotes the convolution, \( K \ast K(x) = \int K(x-t)K(t)dt \). Note that \( (2.7) \) can be combined with \( (2.9) \) to give

\[
a_{\text{AMSE}} = C_3(K)C_4(f)h_{\text{AMISE}}^{10/13}
\]

where

\[
C_3(K) = \left\{ 18R(K^{(4)})/\sigma_K^8 \right\}^{1/13},
\]

\[
C_4(f) = \left\{ (R(f)/R(f^{(3)})^2 \right\}^{1/13}.
\]

Now that we understand how a should relate to \( h \), we can consider attempting to solve, for \( h \), a version of the equation \( (2.7) \) (after making an appropriate substitution for \( a \) in terms of \( h \)), except for the fact that \( C_4(f) \) is still unknown. Since the dependence of \( C_4(f) \) on \( f \) at this point appears to be less crucial than dependence on \( f \) at other stages, it seems to be enough to simply use a scale parameter model for \( f \) (i.e. the scale of the distribution is considered to be the only crucial characteristic at this stage). It is demonstrated in Theorem 3.3 below that in fact this approach is appropriate. Let \( g_1(x) \)
be any fixed probability density, which has been normalized so that some
measure of scale such as the interquartile range or the standard deviation
is equal to 1. Then replace \( f \) in \( C_4(f) \) by \( g_\lambda \), where
\( g_\lambda(x) = g_1(x/\lambda)/\lambda \). Since \( C_4(g_\lambda) = \lambda^{3/13}C_4(g_1) \),
the relationship (2.10) motivates the definition
\[
a_\lambda(h) = C_3(K)C_4(g_1)\lambda^{3/13}h^{10/13}.
\]
The plug-in bandwidth \( \hat{h}_{PI} \) is taken to be the root (when it exists, the
largest if there are more than one), over the range of (2.3) of the equation
\[
h = \left( \frac{R(K)}{\sigma_\lambda^4}R(\hat{f}_a^\lambda(h)) \right)^{1/5}n^{-1/5},
\]
where \( \hat{\lambda} \) denotes a good (i.e. \( n^{-1/2} \) consistent) estimate of \( \lambda \), the scale
(e.g. standard deviation or interquartile range) of \( f \). In cases where the
equation has no root over the range of (2.3), let \( \hat{h}_{PI} = \bar{h}n^{-1/5} \), for
definiteness, although there seem to be no reported cases of either
nonexistence of a root, or multiple roots. A minor improvement that can be
made, is based on the fact that \( R(\hat{f}_a^{\lambda''}) \) is a double sum where the diagonal
terms are constant. It is seem in the proofs of section 6 that this
diagonal creates an asymptotically negligible contribution, however it seems
best from a small sample point of view to delete this diagonal, and replace
the normalization \( n^{-2} \) by \( n^{-1}(n-1^{-1}) \), see Hall and Marron (1987b) for more
discussion. This slight modification has been used in the simulations
presented in section 5.

Motivation for the partitioned cross-validated bandwidth may be found
in Marron (1988b). It may be thought of as another approach to
reducing the large random variability of the cross-validated bandwidth \( \hat{h}_{CV} \). Let \( m \) be an integer which divides the sample size \( n \) (this is strictly for notational convenience, the generalization to arbitrary \( m \) is straightforward). Randomly (independent of \( X_1, \ldots, X_n \)) partition the integers \( 1, \ldots, n \) into subsets \( S_1, \ldots, S_m \) each with cardinality \( n/m \). For \( j = 1, \ldots, m \), let \( CV_j(h) \) denote the least squares cross-validation score function (2.4) calculated over the subsample \( S_j \). Define the average score function,

\[
\overline{CV}(h) = m^{-1} \sum_{j=1}^{m} CV_j(h),
\]

and let \( \hat{h}_{CV} \) denote the minimizer of \( \overline{CV}(h) \) (with existence and unique imposed through the conventions used above for \( \hat{h}_{CV} \)). The effect of averaging the score functions is to cut down dramatically on the variability (see the theoretical results in the next section), but the bandwidth \( \hat{h}_{CV} \) is appropriate for a sample of size \( n/m \), not of size \( n \). Taking this into account, through (2.7) yields the partitioned cross-validated bandwidth,

\[
\hat{h}_{PCV} = m^{-1/5} \hat{h}_{CV}.
\]
3. Rates of Convergence

A useful tool, for understanding the large sample characteristics of the automatically selected bandwidths described in section 2, is the calculation of limiting distributions. While the performance of the estimator \( \hat{f}_h \), using the automatically selected bandwidths, is the chief concern, it will be shown (see Remark 3.7 below) that this performance is directly dependent on the sample variability of the distribution of the bandwidths themselves. For this reason, the main results of this section are stated in terms of the bandwidth distributions.

The aspect of the limiting distributions which is most important in the limit is the exponent in the rate of convergence. As noted above, this rate depends on the amount of smoothness of the underlying density function \( f(x) \). In the present context, a useful means of parametrizing the amount of smoothness is the following.

Let \( \nu = \ell + \zeta \), where \( \ell \) is an integer and \( \zeta \in (0,1] \). The density function \( f \) is said to have "smoothness of order \( \nu \)" when (2.5) is satisfied and there is a constant \( M > 0 \), so that

\[
|f^{(2+\ell)}(x) - f^{(2+\ell)}(y)| \leq M |x - y|^{\zeta}, \text{ for all } x \text{ and } y.
\]

Conditions under which all of the results in this section are valid are contained in

Condition C: the underlying density has smoothness of order \( \nu > 0 \), the bandwidths under consideration fall in the range (2.3), and the kernel function \( K \) satisfies (2.1), (2.2) and (6.1).
Hall and Marron (1987a), although see Scott and Terrell (1987) for a more convenient representation, have quantified the amount of sample variability of the cross-validated bandwidth $\hat{h}_{CV}$ by the following:

**Theorem 3.1** Under condition C,

$$n^{1/10}(h_{CV}/h_{MISE} - 1) \xrightarrow{d} N(0, \sigma_{CV}^2)$$

where,

$$\sigma_{CV}^2 = 2R(\rho)R(f)/(25\sigma_K^{36/5}R(f')^{9/5}R(K)^{1/5}).$$

$$\rho(x) = x \int K(t)K'(x)dt - 2xK'(x).$$

**Remark 3.1**: The rate of convergence $n^{-1/10}$ is very slow. This together with $\sigma_{CV}^2$ quantifies the large amount of sample variability, discussed in the earlier sections, for the least squares cross-validated bandwidth.

Scott and Terrell (1987) have established a related result, which asymptotically quantifies the amount of sample variability of the biased cross-validated bandwidth. An extension of their result is

**Theorem 3.2** Under condition C,

(a) when $0 < \nu \leq 1/2$,

$$n^{1/10}(h_{BCV}/h_{MISE} - 1) = o_p(n^{-\nu/5})$$

(b) when $\nu > 1/2$

$$n^{1/10}(h_{BCV}/h_{MISE} - 1) \xrightarrow{d} N(0, \sigma_{BCV}^2)$$

where,
\[ \sigma_{\text{BCV}}^2 = \sigma_K^{36/5} R(\phi)R(f)\left[200R^{1/5}(f')R^{9/5}(K)\right], \]
\[ \psi(x) = x \int K''(t)K'''(x+t)dt. \]

The proof of Theorem 3.2 is given in section 6.

Remark 3.2: Here again there is a very slow rate of convergence of \( n^{-1/10} \), however Scott and Terrell (1987) have shown that often \( \sigma_{\text{BCV}}^2 \) may be expected to be much smaller than \( \sigma_{\text{CV}}^2 \) (Section 4 of this paper provides further evidence along this line). This is crucial, because when the power of \( n \) is so close to 0, it is the constant coefficients which really determine the practical behavior.

Remark 3.3: Observe that general \( \nu > 0 \) is considered here, instead of \( \nu \geq 2 \), which is essentially what was assumed by Scott and Terrell (1987). Their assumption was an artifact of the method of proof used, and is not intrinsic to the method of biased cross-validation.

The amount of sample variability in the plug-in bandwidth is asymptotically quantified by

Theorem 3.3: Under condition C,
(a) when \( 0 < \nu \leq 2 \),
\[ \hat{h}_{\text{PI}}/h_{\text{MISE}} - 1 = O_p(n^{-2\nu/13}) \]
(b) when \( \nu > 2 \),
\[ n^{4/13}(\hat{h}_{\text{PI}}/h_{\text{MISE}} - 1) \rightarrow N(0, \sigma_{\text{PI}}^2). \]

where
\[ \sigma_{\text{PI}}^2 = (2/25)\sigma_K^{72/13} R(\phi)R(f)R^{-18/13}(K)R^{-8/13}(f')/\left[C_3(K)C_4(g)\right]. \]
\[ \phi(x) = K''K'''(x) = \int K''(t)K'''(x+t)dt. \]
The proof of Theorem 3.3 is also given in Section 6.

Remark 3.4: Theorem 3.3 demonstrates that the idea of Hall (1980) and Sheather (1983, 1986), discussed in the previous section, of replacing $C_4(f)$ by $C_4(g_\lambda)$ is very reasonable. In particular, observe that the rate of convergence obtained in Theorem 3 is the same as the rate using the theoretically best (but unavailable in practice) plug-in kernel estimator as given in Remark 4.6 of Hall and Marron (1987b).

Remark 3.5: Note that the sample variability of $\hat{h}_{PI}$ decreases much faster than that of $\hat{h}_{CV}$ or $\hat{h}_{BCV}$. Hence this bandwidth will always be superior for $n$ sufficiently large. The issue of how large $n$ needs to be for this to happen is addressed in Sections 4 and 5.

Quantification of the sample variability of the partitioned cross-validated bandwidth is trickier, because of the need to choose the number of subsamples $m$. Sample variability issues suggest choosing $m$ large. However this has the effect of causing large errors in the rescaling done at (2.11). The trade-off involved in choice of $m$ is actually analogous to the trade-off made in choice of $h$ (the bandwidth of the underlying estimator $\hat{f}_h$), for $m$ small "variance" is too large, while $m$ large creates too much "bias". See Marron (1988b) for a precise formulation, and theoretical solution, in the same spirit as $h_{AMISE}$ at (2.7), to this problem. It appears that this theoretical solution to the problem of practical choice of $m$ could be made practical through some ideas related to the plug-in bandwidth selector. However as the details of this have not yet been worked out, we work
here only with the theoretical solution, as interesting comparison can still be made. For optimal choice of \( m \), the variance-squared bias tradeoff entails that the \( \hat{h}_{\text{PCV}} \) limiting distribution will have nonzero mean. In particular, it is straightforward to adapt the results of Marron (1988b) to obtain

**Theorem 3.4** Under condition \( C \),

(a) when \( 0 < v \leq 1/2 \),

\[
(\hat{h}_{\text{PCV}}/h_{\text{MISE}} - 1) = O_p(n^{-1/10}),
\]

by choosing

\[
m = 1.
\]

(b) when \( 1/2 < v \leq 2 \),

\[
(\hat{h}_{\text{PCV}}/h_{\text{MISE}} - 1) = O_p(n^{-v/(2v+4)}),
\]

by choosing

\[
m \approx n^{(2v-1)/(2v+4)}.
\]

(c) when \( v > 2 \),

\[
n^{1/4}(\hat{h}_{\text{PCV}}/h_{\text{MISE}} - 1) \overset{d}{\to} N(\mu_{\text{PCV}}, \sigma_{\text{PCV}}^2),
\]

by choosing

\[
m = (\sigma_{\text{CV}}/C_1)^{5/4} n^{3/8}.
\]

where,

\[
\mu_{\text{CV}} = (C_1 \sigma_{\text{CV}})^{1/2}/C_0,
\]

\[
\sigma_{\text{PCV}}^2 = C_1 \sigma_{\text{CV}}^2/C_0^2.
\]

\[
C_1 = \{R(K)^{3/5}(f^4 K R(f^\cdots))/\{20\sigma_k^{22/5} R(f^\cdots)^{8/5}\}\}
\]

The proof of Theorem 3.4 is omitted because it is essentially the same
combination of Theorems 1 and 2 of Marron (1988b) as was used at (3.2) of that paper.

Remark 3.6: Note that the sample variability of the theoretically best version of \( \hat{h}_{PCV} \) will be much smaller in the limit than that of \( \hat{h}_{CV} \) or \( \hat{h}_{BCV} \). However, for reasonably large \( v \), it will be larger than that of \( \hat{h}_{PI} \). See Marron (1988b) for discussion of possible improvements of \( \hat{h}_{PCV} \).

Remark 3.7: Note that the above results all concern the asymptotic distribution of the various automatically selected bandwidths. This may be surprising, because the performance of the estimator \( \hat{f}_h \), using the chosen bandwidth should clearly be the primary concern here. The reason this is done is that all of these limiting distributions, for the \( \hat{h} \)'s, can be directly translated into analogous limiting distributions for \( MISE(\hat{h}) \). In particular, using a simple Taylor expansion argument, such as that leading to Theorem 2.2 of Hall and Marron (1987a), for \( \hat{h} \) any of \( \hat{h}_{CV} \), \( \hat{h}_{BCV} \), or \( \hat{h}_{PI} \), as \( n \to \infty \),

\[
n^{1/5}(MISE(\hat{h})/MISE(\hat{h}_{MISE}) - 1) \overset{d}{\to} 2\sigma^2 \chi_1^2
\]

for \( \sigma^2 = \sigma_{CV}^2, \sigma_{BCV}^2, \sigma_{PI}^2 \), respectively. An analogous noncentral chi-square limiting distribution can be derived for \( MISE(\hat{h}_{PCV}) \). Additional, and more visual, insight into why it is enough to consider only the \( \hat{h} \) distributions can be gained from Figure 3 and the discussion in Section 5.

Remark 3.8: The limiting distributions of Theorems 3.1 - 3.4 can be formulated in other ways as well. In particular, through simple Taylor expansions, these give limiting normal distributions for \( (\hat{h} - h_{MISE}) \).
and \((\log(h) - \log(h_{MISE}))\). The relative error representation here is simply a matter of personal preference, motivated by the fact that the bandwidth is a scale parameter. One nice feature of the chosen version is that Theorem 3.2 is then scale invariant. In particular, note that if \(f\) is rescaled, \(\sigma_{BCV}^2\) still keeps the same value.

Remark 3.9: Section 3.6 of Silverman (1986) discusses how faster rates of convergence of \(\hat{f}_h\) to \(f\) can be obtained through use of kernels which take on carefully chosen negative values. It is straightforward to adapt the results of this paper to that case. The main change in Theorems 3.1 - 3.4 will be that the rates of convergence typically become slower. This extension is not done explicitly here, because it does not seem to add enough insight to justify the additional space, and because this type of estimator is rarely used in practice.

A less appealing feature of Theorems 3.1 - 3.4 is that it is not easy to compare the rates of convergence for small values of \(\nu\). To help with this comparison, see Figure 1.

[Put Figure 1 about here]

Figure 1 shows essentially how the power of \(n\) in each of these limiting distribution depends on \(\nu\). In interpreting this picture, recall that the power for CV is identically 1/10, while BCV and PCV take on this value on complementary sets. Figure 1 makes it is easy to see which bandwidth selector has the best asymptotic performance for the various smoothness assumptions. This quantifies the "robustness" issues discussed in Section 1. In particular observe that \(\hat{h}\) may be though of as the most robust, because it handles the difficult small \(\nu\) situation
without breaking down, but pays a price by not being so efficient when there is more smoothness available. On the other hand \( \hat{h}_{PI} \) is more efficient for large \( v \), but breaks down when the underlying density is not very smooth. While \( \hat{h}_{BCV} \) appears to be dominated here by \( \hat{h}_{CV} \), it is important to keep in mind that this picture considers only the exponent of the rate of convergence, while, for reasonable sample sizes, it is in fact the constant coefficient which typically has the most effect on the errors involved (See Section 4 for another point of view on these issues). The bandwidth \( \hat{h}_{PCV} \) may be thought of as representing some sort of middle ground between the robust \( \hat{h}_{CV} \) and the efficient \( \hat{h}_{PI} \), but it must be kept in mind that this good behavior is not yet practically available, because it depends on reasonable choice of \( m \), which appears to be a problem that is at least as difficult as bandwidth selection.

A crucial part of these robustness issues is the extent to which each bandwidth selector relies on estimation of \( R(f'') \). The selector which relies most on this is \( \hat{h}_{PI} \), while \( \hat{h}_{CV} \) makes no use of it at all. The fact that when \( v \) is small, \( R(f'') \) should be very difficult to estimate, is what causes the poor robustness properties of \( \hat{h}_{PI} \) and \( \hat{h}_{BCV} \).
4. Normal Example

The results of the previous section provide a good deal of insight into the relative performances of the various bandwidth selectors, but an important point is that comparison was only done on the basis of the exponent of the rate of convergence. Certainly attention must be paid also to the sizes of the constant coefficients, especially as very large sample sizes will be required before the dominating effects of one fractional power of \( n \) over another will be realized. Again recall, from Remark 3.7, that the performance of \( \hat{f}_h \), using these automatic bandwidths is driven by the sample variability of the bandwidths, so the focus is again on this variability.

The difficulty with looking at the constant coefficients is that they are quite complicated, and depend on the unknown density \( f \) in rather intricate ways. In this section we approach this problem by comparing the asymptotic sample variabilities of the data driven bandwidths under consideration, for one setting. In particular, we take all of \( f, K, \) and \( g_1 \) to be standard normal densities.

In this case, it follows from Theorem 3.1 that

\[
\text{asd}(\hat{h}_{CV}/h_{MISE}) = \sigma_{CV}^2
\]
\[
\approx 0.33859 \, n^{-1/10}.
\]

from Theorem 3.2 that

\[
\text{asd}(\hat{h}_{BCV}/h_{MISE}) = \sigma_{BCV}^2
\]
\[
\approx 0.08541 \, n^{-1/10},
\]

from Theorem 3.3 that
\[(4.3) \quad \text{asad}(\hat{h}_{\text{PI}}/h_{\text{MISE}}) = \sigma_{\text{PI}}^2 \]
\[\approx .09722 \, n^{-4/13}.\]

Because the limiting distribution of \(\hat{h}_{\text{PCV}}\), given in Theorem 3.4, does not have zero mean, it makes sense to compare the above standard deviations, with the asymptotic mean square error which results from combining the asymptotic variance and squared bias,
\[(4.4) \quad \text{mse}(\hat{h}_{\text{PCV}}/h_{\text{MISE}})^{1/2} = (\sigma_{\text{PCV}}^2 + \mu_{\text{PCV}}^2)^{1/2}n^{-1/10} \]
\[\approx .53379 \, n^{-1/4}.\]

For insight into how these compare in terms of various values of the sample size \(n\), consider Figure 2. This is a plot of \(\log_{10}(\text{asad}(h/h_{\text{MISE}}))\) as a function of the \(\log_{10}(n)\), for each of the bandwidth selectors. The log scale of course makes the functions linear for easier comparison. Also indicated are the values of \(n\) for which the functions intersect. Of course, it must be remembered that these are only asymptotic representations of the standard deviations, so care must be taken in interpretation, especially for small values of \(n\).

[Put Figure 2 about here]

The fact that constant coefficients are very important here is demonstrated very clearly in Figure 2. Although the variability of \(\hat{h}_{\text{PCV}}\) has a faster rate of convergence than \(\hat{h}_{\text{BCV}}\) (and hence must eventually be far smaller), note that the enormous sample size of 202,117 is the break even point, simply because of the substantially smaller coefficient for \(\hat{h}_{\text{BCV}}\). Note that the impression one receives here, of the performance of \(\hat{h}_{\text{BCV}}\), is much different that that received in Figure 1, again because of the effect of the constant coefficients.
Another point which is demonstrated by Figure 2 is the superiority of the plug-in bandwidth in this sense. Of course this picture will doubtless change quite substantially for other choices of the underlying density $f$. 
5. Simulations

While the methods of the preceding sections provide an informative basis for comparison of bandwidth selectors, it is important to keep in mind that they are only asymptotic in character. As with all types of asymptotics, it is important to see if the effects described indicate what is happening for reasonable sample sizes. In this section, simulation results are presented for this purpose.

The underlying density functions chosen here were:

1. a standard normal density,
   \[ N(0,1), \]
2. a mixture of normals with different means,
   \[ .5N(-1.5,1) + .5N(1.5,1), \]
3. 2 mixtures of normals with different variances,
   \[ .5N(0,1) + .5N(0,.1), \]
   \[ .5N(0,1) + .5N(0,.01). \]

The reason that normal mixtures were used is that they greatly facilitate the exact computation of MISE(h), using an obvious extension of the ideas of Fryer (1976). For this same reason (and also because this was used in section 4 as well), only the Gaussian kernel was considered. An exact version of MISE is preferred because asymptotic representations, such as (2.6), can sometimes provide rather different conclusions (see Scott (1986) and Dodge (1986) for some interesting examples connected to this issue), which we did indeed observe in the present context.
The sample sizes considered here were \( n = 25, 50 \) and 100 for the first two densities, and \( n = 25 \) and 100 for the last two. Exact computation for the larger sample sizes was intractable using the machinery available. Approximate calculations were attempted, using the Fourier Transform ideas discussed in section 3.5 of Silverman (1986), but these were not deemed sufficiently accurate. (Although they were really unacceptable only for the computationally more complicated \( \hat{h}_{BCV} \) and \( \hat{h}_{PI} \).)

Because bandwidth selection seems to generate "heavy tailed" distributions, 500 Monte Carlo replications were required for each setting to get reasonable resolution between the various bandwidth selectors. See Marron (1988c) for further discussion of this issue, and for a related simulation study.

For comparison of the bandwidth selectors, we considered the MISE at each bandwidth relative to the best possible MISE. In particular our goal was to compare

\[
E(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})),
\]

for \( \hat{h} \) being each of \( \hat{h}_{CV} \), \( \hat{h}_{BCV} \) and \( \hat{h}_{PI} \). We did not include \( \hat{h}_{PCV} \) in the simulations because it requires specification of the number of subsamples \( m \). We could have tried a range of values of \( m \), but this seems to move too far away from the main point of the present paper.

Since it was simple to do, we also included a bandwidth which uses the "oversmoothing" idea of Terrell and Scott (1985). The particular version used here is

\[
\hat{h}_{OS} = 7^{1/2}(2R(K)/(45\sigma_K^2))^{1/5} \sigma_n^{-1/5},
\]
where \( \hat{\sigma} \) denotes the sample standard deviation. This bandwidth is based on the clever observation that there is an upper bound on the bandwidth \( h_{\text{AMISE}} \) and provides a simple estimate of this upper bound. Intuitively it is clear that this bandwidth should perform very well when the target density \( f \) has a very small amount of structure, as with the standard normal, and arbitrarily badly when there is more structure present (because it will tend to smooth too much).

In order to take the Monte Carlo variability properly into account, it is convenient to replace (5.1) by

\[
C^\dagger = E(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})) - 1,
\]

for \( \hat{h} = \hat{h}_{\text{CV}}, \hat{h}_{\text{BCV}}, \hat{h}_{\text{PI}} \) and \( \hat{h}_{\text{OS}} \). The expected value was estimated by \( \hat{C}^\dagger \), the Monte Carlo average. To use \( \hat{C}^\dagger \) to obtain confidence intervals for \( C^\dagger \), we use the limiting chi-square one distribution of Remark 3.7. This ensures that

\[
\text{NSIM}^{1/2}(\hat{C}^\dagger - C^\dagger) \overset{d}{\rightarrow} N(0, 2C^2),
\]

where \( \text{NSIM} \) is the number of Monte Carlo replications. This limiting distribution motivates comparing the different bandwidths, with proper attention paid to the Monte Carlo variability, by using the pivoted 95\% Confidence Intervals

(5.2)

\[
(\hat{C}^\dagger/(1+T), \hat{C}^\dagger/(1-T)).
\]

where

\[
T = 1.96(2/\text{NSIM})^{1/2}.
\]

The limiting chi-square distribution does not hold up for \( \hat{h}_{\text{OS}} \), because it is not centered at \( h_{\text{AMISE}} \), however there is still considerable insight to be gained concerning the effectiveness of this bandwidth, by
considering intervals of the type (5.2).

Table 1 contains the main results of the simulation study. For each sample size and each distribution, it shows the 95% confidence intervals for the comparison number $\hat{C}^\dagger$, given in (5.2). Of course for simultaneous inference, a simultaneous interval scheme should really be used, but this is not done because it tends to obscure the points being made. The fact that these intervals do indeed give a reasonable idea of the amount of Monte Carlo variability involved was verified by experimenting with different seeds of the random number generator.

[Put Table 1 about here]

As expected, since all of these densities are quite smooth, $\hat{h}_{PI}$ has generally the best performance (even better than the asymptotically optimal value in one case, although not significantly so). Note that Figure 2, which treated the standard normal case, gives only a rough indication of how things behave here (this is because Figure 2 presents only asymptotic values).

Observe that $\hat{h}_{BCV}$ exhibits very poor behavior for $n = 25$ in all cases, and for larger sample sizes for all but the standard normal distribution. What seems to have happened in those situations was that, for a substantial number of the Monte Carlo data sets, the score function $BCV(h)$ had its only minimum at $h = \infty$, so $\hat{h}_{BCV}$ was taken to be the endpoint of the range of bandwidths under consideration. This is of course an unattractive feature of $\hat{h}_{BCV}$, but at least one knows when this trouble occurs.
The oversmoothed bandwidth \( \hat{h}_{OS} \) works very well, as expected, for the standard normal, but, again as expected, not so well for the other cases. It is especially bad for the more extreme scale mixture, which is not surprising because this density has much more "structure" (this can be quantified in terms of \( R(f') \) when scale is accounted for) than the standard normal, which is ignored by \( \hat{h}_{OS} \). Actually this comparison is not quite fair, since for the extreme cases \( \hat{h}_{OS} \) was often outside the interval of bandwidths under consideration for the other selectors.

Visual insight into what drives the results of Table 1 can be obtained from Figure 3. This shows an overlay of MISE(h) (MISE(h) is much higher for the variance mixture example because this is a much "harder" density estimation setting), together with kernel estimates of the density of the distribution of the various automatically selected bandwidths, for \( n = 100 \) observations from (a) the standard normal and (b) the first variance mixture. The bandwidth used for these kernel estimates was the oversmooother, which seems reasonable here in view of the limiting normal distributions available.

[Put Figures 3a and 3b about here]

Figure 3 demonstrates how the ordering of the selectors in Table 1 is determined by their sample variability. In particular the more variable bandwidths take on occasional values that are far from the optimum, with a resulting increase in MISE. This provides another view of the fact, derived theoretically in Remark 3.7, that to study the performance of \( \hat{f}_h \) at a data-driven bandwidth, it is enough to consider the noise in the bandwidth.
The superior performance of $\hat{h}_{OS}$ in the standard normal case is clearly seen to come from the fact that its distribution is tightly clustered near $h_{MISE}$. It performs poorly in the other cases because, while it still has a very tight distribution, the center point is too large.

Note that in both cases, the distribution of $\hat{h}_{CV}$ has substantial skewness towards a heavy tail, and $\hat{h}_{BCV}$ has a bias towards oversmoothing. Both of these points were also noted by Scott and Terrell (1987). It is interesting that $\hat{h}_{PI}$ seems to follow the limiting mean zero normal distribution, predicted by the theory of section 3, more closely than the others.
6. Proofs

The technical moment assumptions of $K$ made in these proofs (mentioned at (2.2)) are:

\[(6.1) \quad \int |u|^{2+p}K(u) \, du < \infty,\]
\[\int |u|^{1+p} |K'(u)| \, du < \infty,\]
\[\int |u|^{1+p} |K''(u)| \, du < \infty,\]
\[\int |u|^{2+p} \phi(u)^2 \, du < \infty,\]
\[\int |u|^{2+p} \psi(u)^2 \, du < \infty.\]

Useful notation is, for $i,j = 1,\ldots,n$,

\[c_{ij} = (X_i - X_j) / h.\]

For sequences $a_n$ and $b_n$, "$a_n \sim b_n$" is understood to mean that

\[\lim_{n \to \infty} a_n / b_n = c, \text{ where } c \text{ is a nonzero constant.}\]

For a sequence of random variables $Y_n$, "$Y_n = \text{AN}(a_n, b_n)$" means

\[b_n^{1/2} (Y_n - a_n) \overset{d}{\to} N(0,1).\]

Proof of Theorem 3.2: This proof uses the approach of Scott and Terrell (1987), who point out that minimizing $BCV(h)$ is the same as finding the $h$ that is the root of (recall $c_{ij}$ depends on $h$)

\[(6.2) \quad \sum_{i < j} \{\phi(c_{ij}) + \psi(c_{ij})\} = -2nR(K)/\sigma_K^4.\]

Key steps in the proof are

Lemma 6.1: Under condition C, if $h \sim n^{-1/5}$, then

\[\sum_{i < j} [\phi(c_{ij}) + \psi(c_{ij})] = \text{AN}(\mu_1, \sigma_1^2),\]
where

\[
\begin{align*}
\mu_1 &= -2n^2h^5R(f'') + O(n^2h^{5+v}) \\
\sigma_1^2 &= (1/2)n^2hR(\psi)R(f)
\end{align*}
\]

Lemma 6.2: Under condition C,

\[\hat{h}_{BCV} \leq h_{AMISE}.\]

Lemma 6.3. Under condition C,

\[
\frac{h_{AMISE}}{h_{MISE}} - 1 = O(n^{-u/5}) \quad \text{if} \quad 0 < u < 3
\]

\[
= O(n^{-3/5}) \quad \text{if} \quad v \geq 3
\]

The proofs of Lemmas 6.1 and 6.2 follow the proof of Theorem 3.2. The proof of Lemma 6.3 is omitted because it is essentially the same as the proof of (3.2) in Marron (1988b).

From Lemma 6.1, if \( h \sim n^{-1/5} \),

\[
(6.3) \quad n^{-1}h^{-1/2} \left[ \sum_{i<j} \Phi(c_{ij}) + \psi(c_{ij}) \right] + 2n^2h^5R(f'') + O(n^2h^{5+v})
\]

\[
= AN(0, R(\psi)R(f)/2)
\]

Since the right hand side of (6.3) does not depend on \( h \), we can replace \( h \) by \( \hat{h}_{BCV} \) in the left hand side of (6.3). Hence, since \( \hat{h}_{BCV} \) is the root of (6.2), we get

\[
(6.4) \quad n^{-1}\hat{h}_{BCV}^{-1/2} \left[ -2nR(K)/\sigma_K^4 + 2n^2\hat{h}_{BCV}^5R(f'') + O(n^2\hat{h}_{BCV}^{5+v}) \right]
\]

\[
= AN(0, R(\psi)R(f)/2).
\]

Dividing both sides of (6.4) by \( 2R(f'') \), we have

\[
(6.5) \quad n^{-1}\hat{h}_{BCV}^{-1/2} \left[ -nR(K)/\sigma_K^4R(f'') + n^2\hat{h}_{BCV}^5 + O(n^2\hat{h}_{BCV}^{5+v}) \right]
\]

\[
= AN(0, R(\psi)R(f)/8R(f'')^2).
\]
From (2.7), Lemma 6.2, and (6.5) it follows that

\[(6.6) \quad n^{1/2} \left[ \hat{h}_\text{BCV} - n^{4.5+\nu} \right] + o_p(n^{4.5+\nu}) = \mathcal{A}(0, R(\psi) R(f)/8R(f''))^2. \]

If \(0 < \nu \leq 1/2\), then from (6.6)

\[\hat{h}_\text{BCV} - n^{5+\nu} = o_p(n^{5+\nu}) = o_p(n^{-(5+\nu)/5}).\]

Taylor expansion shows that

\[\hat{h}_\text{BCV} - h_\text{AMISE} = o_p(n^{-(1+\nu)/5}).\]

Now Lemma 3 can be applied to conclude the first part of the theorem. When \(\nu > 1/2\), the second term in the left hand side of (6.6) is negligible. So in this case, it is straightforward to get the second part of the theorem using Lemma 3.

Proof of Lemma 6.1: For the expected value component, note that

\[(6.7) \quad E \Phi(c_{ij}) = \iint K''(\omega) K''(\omega + (x-y)/h) f(x)f(y) d\omega \ dx \ dy \]
\[= h \iint K''(\omega) K''(\omega + u) f(x) f(x+h\omega) d\omega \ dx \ du \]
\[= h^5 \iint K(\omega) K(\omega + u) f''(x) f''(x+h\omega) d\omega \ dx \ du \]
\[= h^5 R(f'') + O(h^{\zeta+5}),\]

where \(\zeta = \min(\nu, 2)\), and where the last equality follows from condition C.

Similarly

\[(6.8) \quad E \psi(c_{ij}) = -5h^5 R(f'') + O(h^{\zeta+5}).\]

Hence

\[(6.9) \quad E(\Phi(c_{ij}) + \psi(c_{ij})) = -4h^5 R(f'') + O(h^{\zeta+5}).\]

For the variance part, note that

\[(6.10) \quad \text{var}(\Phi(c_{ij})) = E \Phi^2(c_{ij}) - (E \Phi(c_{ij}))^2 \]
\[= \iint K''(\omega) K''(\omega + (x-y)/h) d\omega \ f(x)f(y) dx \ dy - (E \Phi(c_{ij}))^2.\]
\[ = h \int \int (K'' \ast K''(u))^2 f(x) f(x-hu) dx \, du - \{ E \Phi(c_{ij}) \}^2 \]
\[ = h R(\Phi) R(f) + O(h^3), \]

where the last equality makes use of the symmetry of \((K'' \ast K''(u))^2\) as a function of \(u\), and of (6.7). Similarly

\[ \text{var}[\psi(c_{1j})] = h R(\psi) R(f) + O(h^3). \]  

Observe that for \(j \neq k, \)

\[ E \Phi(c_{1j}) \Phi(c_{1k}) \]
\[ = h^2 \int \int \int K''(u_1) K''(u_2) f(x) f(x-hu_1) f(x-hu_2) du_1 \, du_2 \, dx \]
\[ = h^2 \int \{ \int \int K''(w) K''(w+u_1) f(x-hu_1) d\, du_1 \} \]
\[ \{ \int \int K''(t) K''(t+u_2) f(x-hu_2) dt \, du_2 \} f(x) dx \]
\[ = h^6 \int \{ \int \int K''(w) K''(w+u) f''(x-hu) d\, du \}^2 f(x) dx \]
\[ = O(h^{6+2\xi}). \]

again using condition C and the fact that \( \int \int K''(\omega) K(\omega+u) d\omega \, du = 0. \)

Similarly we get, for \(j \neq k, \)

\[ E[\psi(c_{1j}) \psi(c_{1k})] = 0(h^{6+2\xi}) \]

and

\[ E[\Phi(c_{1j}) \psi(c_{1k})] = 0(h^{6+2\xi}). \]

By the same techniques, it is straightforward to show

\[ E[\Phi(c_{1j}) \psi(c_{1j})] = h R((\Phi \psi)^{1/2}) R(f) + O(h^3). \]

Using the notation, \( \Phi_{ij} = \Phi(c_{ij}) \) and \( \psi_{ij} = \psi(c_{ij}), \) it follows from (6.7), (6.8) and (6.12)-(6.15) that, for \(i, j\) and \(k\) all different,

\[ \text{Cov}(\Phi_{ij}, \Phi_{ik}) = 0(h^{6+2\xi}) \]
\[ \text{Cov}(\Phi_{ij}, \psi_{ik}) = 0(h^{6+2\xi}) \]
\[ \text{Cov}(\psi_{ij}, \Phi_{ik}) = h R((\Phi \psi)^{1/2}) R(f) + O(h^3) \]
\[ \text{Cov}(\psi_{ij}, \psi_{ik}) = O(h^{6+2\xi}) \quad j \neq k. \]
Thus from (6.10), (6.11) and (6.16), we have

\begin{equation}
\text{Var}(\sum_{i<j} \phi(c_{ij}) + \psi(c_{ij}))
= \frac{n(n-1)}{2} h R(\phi + \psi) R(f) + O(n^{-1} h^3 + n^{-1} h^{6+2\delta})
\end{equation}

= \frac{n}{2} h R(\phi) R(f) + O(n^{-1} h^3 + n^{-1} h^{6+2\delta})

since \( R(\phi + \psi) = R(\psi) \). Similar arguments as in the proof of Theorem 3.2 of Scott and Terrell (1987) together with calculations similar to the foregoing entail the asymptotic normality.

Proof of Lemma 6.2: Observe that from (2.6), (6.2) and (6.7)

\[ E(BCV(h)) = AMISE(h) + O(h^{4+\delta}) \]

and from (6.10) and (6.12)

\[ \text{var}(BCV(h)) = \sigma_K^2 R(\phi) R(f) / 8n^2 h + O(n^{-2} h + n^{-1} h^{4+2\delta}) . \]

Now applying the same arguments as in the proof of Corollary 3.2 of Scott and Terrell (1987) gives the lemma.

Proof of Theorem 3.2: Recall that \( \hat{h}_{PI} \) is the value of \( h \) which makes \( L_\lambda(h) = 0 \), where

\[ L_\lambda(h) = h[\sigma_K^4 R(\hat{f}_{a_\lambda(h)}^{\prime\prime\prime})]^{1/5} - n^{-1/5}[R(K)]^{1/5} . \]

The fact that using an estimate of \( \lambda \) causes no problems, is shown by

Lemma 6.4. Under condition C, if \( h \sim n^{-1/5} \), then

\[ R(\hat{f}_{a_\lambda(h)}^{\prime\prime\prime}) - R(\hat{f}_{a_\lambda(h)}^{\prime\prime\prime}) = o_p(n^{-1/2}) \]
As some of the steps in the proof of Lemma 6.4 are useful at other points in the proof of Theorem 3.2, a proof is given now.

Proof of Lemma 6.4: Let \( c_{ij}(a) = (X_i - X_j)/a \). Observe that

\[
R(f''_a) = R(K'')/n^2a^5 + 2n^{-2}a^{-5}\sum_{i<j} \Phi(c_{ij}(a)).
\]

(6.18)

Studying means as in (6.7) and (6.8) and variances as in (6.10), (6.11) and (6.16), we obtain the representations

\[
\sum_{i<j} \Phi(c_{ij}(a)) = (1/2)n^2a^5R(f'') + o_p(n^2a^5)
\]

and

\[
\sum_{i<j} \psi(c_{ij}(a)) = -(5/2)n^2a^5R(f'') + o_p(n^2a^5).
\]

(6.19)

(6.20)

Since the first term in the right hand side of (6.18) is negligible by (6.19),

\[
R(\hat{f}''_a) - R(\hat{f}''_{a'}) \sim -10n^{-2}a_{\star}^{-6}\sum_{i<j} \Phi(c_{ij}(a_{\star})) - 2n^{-2}a_{\star}^{-6}\sum_{i<j} \psi(c_{ij}(a_{\star}))(a' - a)
\]

(6.21)

where \( a_{\star} \) lies in between \( a \) and \( a' \). Plugging \( a_{\lambda}(h) \) and \( a_{\lambda}(h) \) into \( a' \) and \( a \) in (6.21) and using (6.19) and (6.20), we have

\[
R(\hat{f}''_{a_{\lambda}(h)}) - R(\hat{f}''_{a_{\lambda}(h)}) = o_p(n^{-1/2})
\]

since \( \lambda - \lambda = 0_p(n^{-1/2}) \).

Now we finish the proof of Theorem 3.2. A consequence of Lemma 6.4 is

\[
0 = L_{\lambda}(\hat{h}_{PI}) = L_{\lambda}(\hat{h}_{PI}) + o_p(n^{-7/10}).
\]

(6.22)
Now
\begin{equation}
L_\lambda(\hat{h}_{\text{PI}}) = L_\lambda(h_{\text{AMISE}}) + L_\lambda(h^*)(\hat{h}_{\text{PI}}-h_{\text{AMISE}})
\end{equation}
where \( h^* \) lies in between \( \hat{h}_{\text{PI}} \) and \( h_{\text{AMISE}} \). Using arguments similar to those in the proof of Lemma 1,
\[ \Sigma_{i<j} \Phi(c_{ij}(a_\lambda(h))) = AN(\mu_\lambda^{**}, \sigma_\lambda^{**2}), \]
where
\[ \mu_\lambda^{**} = (1/2)n^2 a_\lambda^5(h)R(f'') + O(n^2 a_\lambda(h)^{5+\xi}), \]
\[ \sigma_\lambda^{**2} = (1/2)n^2 a_\lambda(h)R(\phi)R(f) + O(n^3 a_\lambda^{6+2\xi}(h)) + O(n^2 a_\lambda^3(h)). \]
Thus, it follows from (6.18) that
\begin{equation}
L_\lambda(h^*(\hat{f}_{\lambda}(h)))^{1/5} = AN(\mu_\lambda^{**}, \sigma_\lambda^{**2}),
\end{equation}
where
\[ \mu_\lambda^{**} = h(\sigma_K^4 R(f''))^{1/5} + O(h a_\lambda(h)^\xi), \]
\[ \sigma_\lambda^{**2} = (2/25)n^{-2} a_\lambda^{-9}(h)R(f'')^{-8/5} h^2 a_\lambda^{8/5} R(\phi)R(f) +
0(h^2 n^{-1} a_\lambda(h)^{-4+2\xi}) + O(h^2 n^{-2} a_\lambda(h)^{-7}). \]
Plugging in \( h = h_{\text{AMISE}} \) and using (2.7), we have
\begin{equation}
L_\lambda(h_{\text{AMISE}}) = AN(\mu_\lambda^{\uparrow}, \sigma_\lambda^{\uparrow2}),
\end{equation}
where
\[ \mu_\lambda^{\uparrow} = 0(h a_\lambda(h)^\xi), \]
\[ \sigma_\lambda^{\uparrow2} = (2/25)R(\phi)R(f)[R(\hat{f})]^{-2/5} R^{-2}(f'') n^{-12/5} a_\lambda(h_{\text{AMISE}})^{-9} +
0(n^{-(51+20\xi)/65}) + O(n^{-(86/65}). \]
Hence when \( 0 < \nu \leq 2 \),
\begin{equation}
L_\lambda(h_{\text{AMISE}}) = 0(n^{-(13+10\nu)/65}),
\end{equation}
and when \( \nu > 2 \),
\begin{equation}
n^{33/65} L_\lambda(h_{\text{AMISE}}) = AN(0, \sigma_1^2),
\end{equation}
where
\[ \sigma_1^2 = (2/25)\sigma_K^{72/13}R(\phi)R(f)R(K)^{-64/65}R(f'')^{-8/13}\{C_3(K)C_4(\alpha)\}^{-2}. \]

Using (6.18), (6.20) and (6.21), it is easy to see that

\[ L(\lambda^*) \sim \{\sigma_K^4R(f'')\}^{1/5}. \tag{6.28} \]

Putting (6.22), (6.23), (6.26), (6.27), (6.28) and Lemma 6.3 together, the theorem follows.
REFERENCES


CAPTIONS

Figure 1: Powers of the sample size in the limiting distributions of Theorems 3.1 - 3.4, as a function of $\nu$. This demonstrates relative performance (in terms of rate of convergence) of the bandwidth selectors in terms of the smoothness of $f$.

Figure 2: $\log_{10}$ of the asymptotic standard deviations (mean squared error for PCV) of the limiting distributions of the selected bandwidths, as a function of $\log_{10}$ of the sample size. This is the special case where $f$, $K$, and $g_1$ are all $N(0,1)$.

Figure 3: MISE$h$ and kernel density estimates of the distributions of the automatically distributed bandwidths (on $\log_{10}$ scale). From 500 Monte Carlo replications of samples of size 100, from (a) $N(0,1)$, (b) $.5N(0,1) + .5N(0,.1)$. 
Table 1: Approximate Monte Carlo 95% confidence intervals for the "comparison numbers", which allow comparison of the MISE performance of the automatically selected bandwidths. Based on 500 replications in each case.
Table 1

<table>
<thead>
<tr>
<th>n</th>
<th>Standard Normal</th>
<th>Mean Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>AMISE   (0.01757, 0.02254)</td>
<td>OS   (0.03089, 0.03964)</td>
</tr>
<tr>
<td></td>
<td>OS      (0.05662, 0.07264)</td>
<td>PI      (0.08031, 0.10304)</td>
</tr>
<tr>
<td></td>
<td>PI      (0.28839, 0.37001)</td>
<td>AMISE   (0.11363, 0.14579)</td>
</tr>
<tr>
<td></td>
<td>CV      (0.29432, 0.37761)</td>
<td>CV      (0.25094, 0.32196)</td>
</tr>
<tr>
<td></td>
<td>BCV     (0.34173, 0.43844)</td>
<td>BCV     (0.26143, 0.33541)</td>
</tr>
<tr>
<td>50</td>
<td>AMISE   (0.01066, 0.01368)</td>
<td>PI      (0.05860, 0.07518)</td>
</tr>
<tr>
<td></td>
<td>OS      (0.02499, 0.03207)</td>
<td>AMISE   (0.05970, 0.07659)</td>
</tr>
<tr>
<td></td>
<td>PI      (0.12683, 0.16272)</td>
<td>OS      (0.08825, 0.11323)</td>
</tr>
<tr>
<td></td>
<td>BCV     (0.13326, 0.17097)</td>
<td>CV      (0.23758, 0.30482)</td>
</tr>
<tr>
<td></td>
<td>CV      (0.24561, 0.31511)</td>
<td>BCV     (0.55938, 0.71769)</td>
</tr>
<tr>
<td>100</td>
<td>MISE    (0.00641, 0.00822)</td>
<td>AMISE   (0.03248, 0.04167)</td>
</tr>
<tr>
<td></td>
<td>OS      (0.01183, 0.01517)</td>
<td>PI      (0.05856, 0.07513)</td>
</tr>
<tr>
<td></td>
<td>BCV     (0.06713, 0.08613)</td>
<td>OS      (0.16502, 0.21173)</td>
</tr>
<tr>
<td></td>
<td>PI      (0.09362, 0.12012)</td>
<td>CV      (0.21859, 0.28045)</td>
</tr>
<tr>
<td></td>
<td>CV      (0.21243, 0.27255)</td>
<td>BCV     (0.85551, 1.09763)</td>
</tr>
</tbody>
</table>

Variance Mixture 1

<table>
<thead>
<tr>
<th>n</th>
<th>AMISE   (0.05855, 0.07512)</th>
<th>AMISE   (0.06000, 0.07699)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PI      (0.26445, 0.33929)</td>
<td>PI      (0.20090, 0.25775)</td>
</tr>
<tr>
<td></td>
<td>OS      (0.31323, 0.40187)</td>
<td>CV      (0.25418, 0.32611)</td>
</tr>
<tr>
<td></td>
<td>CV      (0.32394, 0.41562)</td>
<td>BCV     (1.09220, 1.40130)</td>
</tr>
<tr>
<td></td>
<td>BCV     (0.89352, 1.14639)</td>
<td>OS      (1.87249, 2.40241)</td>
</tr>
<tr>
<td>100</td>
<td>AMISE   (0.02042, 0.02620)</td>
<td>AMISE   (0.02093, 0.02686)</td>
</tr>
<tr>
<td></td>
<td>PI      (0.07284, 0.09346)</td>
<td>PI      (0.04483, 0.05752)</td>
</tr>
<tr>
<td></td>
<td>CV      (0.16944, 0.21739)</td>
<td>CV      (0.10303, 0.13219)</td>
</tr>
<tr>
<td></td>
<td>BCV     (0.38594, 0.49516)</td>
<td>BCV     (0.73956, 0.94885)</td>
</tr>
<tr>
<td></td>
<td>OS      (0.61895, 0.79412)</td>
<td>OS      (5.18076, 6.64694)</td>
</tr>
</tbody>
</table>