Comparison of Data-Driven Bandwidth Selectors

BYEONG U. PARK and J. S. MARRON*

This article compares 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, and a plug-in rule. The comparison is done by asymptotic rate of convergence to the optimum 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. We believe the plug-in rule is the best of those currently available, but there is still room for improvement.

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. Figure 1 provides an example of how this method can show structure that can be very difficult to see by classical methods. The figure shows overlays of estimates of net income densities for the years 1968–1983, with each year containing roughly 7,000 data points (to study only relative effects, the observations for each year have been divided by the mean for that year). The data utilized here were made available by the ESCR Data Archive at the University of Essex: Family Expenditure Survey, Annual Tapes, 1968–1983, Department of Employment, Statistics Division, Her Majesty's Stationery Office, London.

Figure 1a shows the 16 densities resulting from lognormal fits, and Figure 1b shows kernel density estimates for the same set of data. The impression given from Figure 1a is that all of the populations are unimodal and there is essentially no change across years. Nevertheless, Figure 1b indicates that these populations each have at least two modes, and that there is enormous change in the structure over time.

An even more striking impression, and the additional interesting fact that there is a very systematic shift over time, comes from Figure 2, which shows the same curve estimates as Figure 1b, placed one behind the other in chronological order. Note that the left mode increases while the right mode decreases over time, which indicates a dramatic shift in the income distribution. See Marron and Schmitz (1988) for a deeper analysis of this data.

Practical application of kernel density estimation is crucially dependent on the choice of the smoothing parameter or bandwidth. Although effective data analysis has often been 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 various data-driven methods for choosing the bandwidth have been proposed and studied. See Marron (1988a) for a listing of proposed methods and discussion. This article 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 article. The method used for bandwidth selection in the example illustrated in Figures 1 and 2 is discussed in Section 5.

The most widely studied bandwidth selector is least squares cross-validation, proposed by Rudemo (1982) and Bowman (1984). This method has been shown to have the attractive asymptotic property of giving an answer that converges to the optimum under very weak conditions (Stone 1984). In many simulation studies and real data examples, however, the performance of this method has been often disappointing, since least squares cross-validation suffers from a great deal of sample variability (Hall and Marron 1987a).

Because of the 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 and biased 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.

Another method of improving on least squares cross-validation is partitioned cross-validation, as proposed by

* Byeong U. Park is Assistant Professor, Department of Computer Science and Statistics, College of Natural Sciences, Seoul National University, SAN 56-1, Shinrim-Dong, Kwanak-ku, Seoul, Korea. J. S. Marron is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. Marron's research was partially supported by National Science Foundation Grant DMS-8701201 and the Deutsche Forschungsgemeinschaft, and partially performed while he was visiting the Institut für Wirtschaftstheorie II, Universität Bonn.

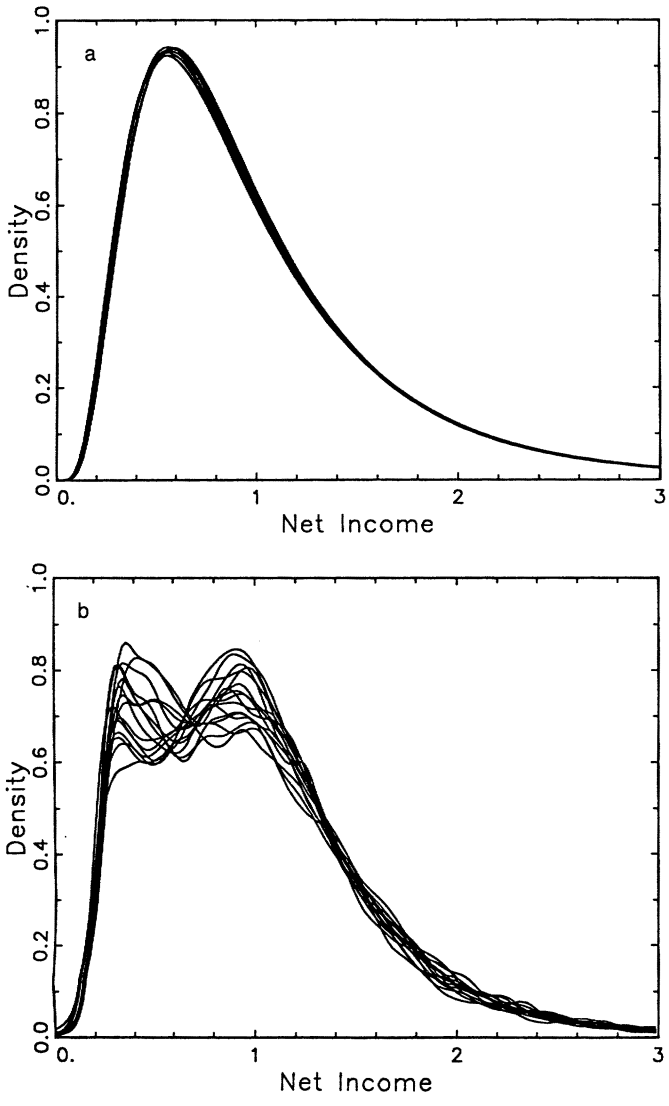

Figure 1. Net Income Density Estimates From the Family Expenditure Survey, 1968–1983: (a) Parametric Lognormal Fit; (b) Kernel Density Estimates.

Marron (1988b). This method will not be discussed in detail here, because it is not quite fully objective.

The main point of this article is a comparison of these methods of smoothing parameter selection. We believe that our results show that the plug-in method is the most practical method currently available, but we feel there is substantial room for improvement. These points are demonstrated through an asymptotic analysis in Section 3, a simulation study in Section 4, and the example of Figures 1 and 2 in Section 5.

The asymptotic rates 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 the number of bounded derivatives, of the underlying density. In Section 3, precise results that quantify these rates as a function of smoothness are given. The main result is that, under strong enough smoothness assumptions on the underlying density, the plug-in bandwidth will dominate in the limit. Nevertheless, there is some trade-off for this, which is caused by the fact that for small amounts of smoothness least squares cross-validation is the most effective. 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 an asymptotic 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, whereas the plug-in rule is more efficient when stronger assumptions hold.

Section 4 contains the results of a simulation study. Once again, we see that the plug-in bandwidth usually gives good results. In addition, 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 $K_h(\cdot) = K(\cdot/h)/h$ and K is the kernel. The scale parameter h is called the bandwidth or smoothing parameter and is crucial to the performance of the estimator.

For the theoretical results in this article, it is assumed that

K is a symmetric probability density with finite support, (2.1)

K has four Hölder (i.e., Lipschitz) continuous derivatives, (2.2)

and

$h \in [\underline{B}n^{-1/5}, \overline{B}n^{-1/5}]$, for some constants $0 < \underline{B} < \overline{B}$. (2.3)

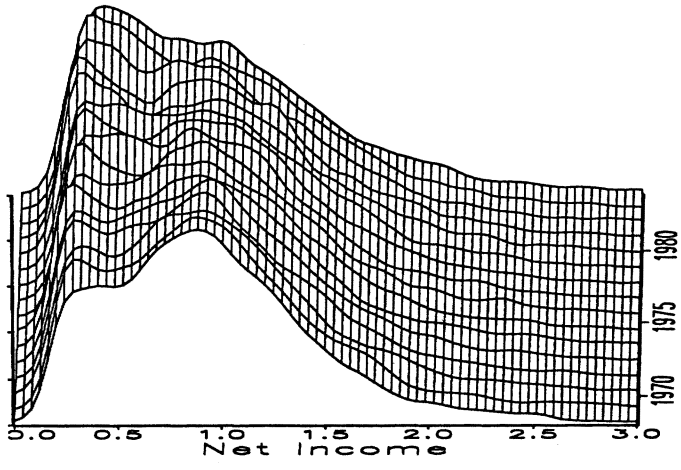

Figure 2. Expanded Representation of the Density Estimates in Figure 1b.

The assumptions (2.1) and (2.2) can be weakened considerably. Assumption (2.3) is made mostly for convenience, but it is not restrictive because this is well known to be the range of reasonable bandwidths.

A common means of assessing the performance of a density estimator is the mean integrated squared error $\text{MISE}(h) = E\int(\hat{f}_h - f)^2$. Most data-driven bandwidths can be viewed as an attempt to estimate h_{MISE}, the minimizer of $\text{MISE}(h)$.

The least squares cross-validated bandwidth $\hat{h}_{\text{CV}}$ is the minimizer [over the range (2.3)] of the cross-validation function

$$\text{CV}(h) = R(\hat{f}_h) - 2n^{-1}\sum_{j=1}^{n}\hat{f}_{j,h}(X_j), \tag{2.4}$$

where here and throughout 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 a reasonable choice of $\underline{B}$ and $\overline{B}$ there is usually at least one minimizer of $\text{CV}(h)$, although local minima are known to occur reasonably often (Hall and Marron 1988). For definiteness, here we take $\hat{h}_{\text{CV}}$ to be the largest local minimizer, over the range (2.3).

The biased cross-validated bandwidth uses the fact that when

f has a Hölder-continuous, square-integrable second derivative, (2.5)

$\text{MISE}(h)$ admits the asymptotic (as $n \to \infty$) representation

$$\text{AMISE}(h) = n^{-1}h^{-1}\int K^2 + h^4\sigma_K^4 R(f'')/4, \tag{2.6}$$

where here and throughout for any mean-zero probability density $g(x)$, the variance of the g distribution is denoted by $\sigma_g^2 = \int x^2 g(x)\,dx$. The biased cross-validated bandwidth $\hat{h}_{\text{BCV}}$ is the minimizer over (2.3) of the estimate of $\text{AMISE}(h)$, obtained by replacing $R(f'')$ by $R(\hat{f}_h'') - n^{-1}h^{-5}R(K'')$.

The plug-in idea is based on closely related considerations. In particular, h_{AMISE}, the minimizer of $\text{AMISE}(h)$, can be written as

$$h_{\text{AMISE}} = C_0 n^{-1/5}$$

$$C_0 = \{R(K)/\sigma_K^4 R(f'')\}^{1/5}. \tag{2.7}$$

At this point it is apparent that the range (2.3) should be chosen so that $\underline{B} \ll C_0$ and $\overline{B} \gg C_0$, so this assumption is made for the rest of the article. As before, the only unknown part of h_{AMISE} is $R(f'')$, so it makes sense to consider estimates. One candidate for such an estimate is $R(\hat{f}_a'')$, 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 $f(x)$]. A somewhat improved estimator (Hall and Marron 1987b) is $\hat{R}_{f''}(a) = R(\hat{f}_a'') - n^{-1}a^{-5}R(K'')$. The fact that a is used instead of h is a 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 $R(f'')$. In particular, note that for any fixed set of data, $R(\hat{f}_a'')$ takes on all values between 0 and ∞ as a ranges from 0 to ∞. The approach to this problem, developed by Hall (1980) and 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

f has a Hölder-continuous, square-integrable fourth derivative, (2.8)

then Hall and Marron (1987b) have shown that a_{MSE}, the minimizer of the mean squared error for using $\hat{R}_{f''}(a)$ to estimate $R(f'')$, has the asymptotic (as $n \to \infty$) representation

$$a_{\text{AMSE}} = C_1(K)C_2(f)n^{-2/13}, \tag{2.9}$$

where $C_1(K) = \{18R(K^{(4)}*K)/\sigma_{K*K}^4\}^{1/13}$ and $C_2(f) = \{R(f)/R(f''')^2\}^{1/13}$, and where $K*K$ denotes the convolution $K*K(x) = \int K(x-t)K(t)\,dt$. Note that (2.7) can be combined with (2.9) to give

$$a_{\text{AMSE}} = C_3(K)C_4(f)h_{\text{AMISE}}^{10/13}, \tag{2.10}$$

where

$$C_3(K) = \{18R(K^{(4)})\sigma_K^8/\sigma_{K*K}^4 R(K)^2\}^{1/13}$$

and

$$C_4(f) = \{R(f)R(f'')^2/R(f^{(3)})^2\}^{1/13}.$$

Now that we understand how a should relate to h, we can consider attempting to solve a version of Equation (2.7) for 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 use a scale parameter model for f. Let $g_1(x)$ be any fixed probability density that 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_λ, 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 $\hat{h}_{\text{PI}}$ is taken to be the root (when it exists, the largest if there are more than one) over the range (2.3) of the equation

$$h = \{R(K)/\sigma_K^4\hat{R}_{f''}(a_{\hat{\lambda}}(h))\}^{1/5}n^{-1/5};$$

here $\hat{\lambda}$ denotes a good (i.e., $n^{1/2}$ consistent) estimate of λ, the scale of f. In cases where the equation has no root over the range (2.3), let $\hat{h}_{\text{PI}} = \overline{B}n^{-1/5}$ for definiteness, although there seem to be no reported cases of either nonexistence of a root or multiple roots. We take g_1 to be the normal density. The fact that this choice does not critically affect the performance of $\hat{h}_{\text{PI}}$ is demonstrated in Theorem 3.3.

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. Although the performance of the estimator $\hat{f}_h$ using the automatically selected bandwidths is the chief concern, it will be shown 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 asymptotically dominant aspect of the limiting distributions is the exponent in the rate of convergence. As noted previously, this rate can depend on the amount of smoothness of the underlying density function $f(x)$. In the present context, a useful means of parameterizing the amount of smoothness is the following.

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

$$|f^{(2+l)}(x) - f^{(2+l)}(y)| \leq M|x - y|^{\eta} \quad \text{for all } x \text{ and } y. \tag{3.1}$$

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) and (2.2).

The amount of sample variability of the cross-validated bandwidth $\hat{h}_{CV}$ is conveniently demonstrated by the following theorem, which is a straightforward consequence of remark 2.3 of Hall and Marron (1987a).

Theorem 3.1. Under condition C,

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

where $\sigma_{CV}^2 = 2R(\rho)R(f)/\{25\,\sigma_K^{4/5}R(f'')^{1/5}R(K)^{9/5}\}$ and $\rho(x) = x\int K(t)K'(t + x)\,dt - 2xK'(x)$.

Remark 3.1. The rate of convergence $n^{-1/10}$ is very slow. This, together with σ_{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) established a related result that asymptotically quantifies the amount of sample variability of the biased cross-validated bandwidth. An extension of their result is the following theorem.

Theorem 3.2. Under condition C, (a) when $0 < \nu \leq \frac{1}{2}$,

$$(\hat{h}_{BCV}/h_{MISE} - 1) = O_p(n^{-\nu/5}),$$

and (b) when $\nu > \frac{1}{2}$,

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

where

$$\sigma_{BCV}^2 = \sigma_K^{36/5}R(\psi)R(f)/[200R^{1/5}(f'')R^{9/5}(K)]$$

and

$$\psi(x) = x\int K''(t)K'''(t + x)\,dt.$$

Remark 3.2. Here again there is a very slow rate of convergence of $n^{-1/10}$. Nevertheless, Scott and Terrell (1987) showed that often σ_{BCV}^2 may be expected to be much smaller than σ_{CV}^2. This is an important improvement, because when the power of n is so close to 0, it is the leading constant coefficients that 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 the following theorem.

Theorem 3.3. Under condition C, (a) when $0 < \nu \leq 2$,

$$\hat{h}_{PI}/h_{MISE} - 1 = O_p(n^{-2\nu/13})$$

and (b) when $\nu > 2$,

$$n^{4/13}(\hat{h}_{PI}/h_{MISE} - 1) \to N(\mu_{PI}, \sigma_{PI}^2),$$

where

$$\mu_{PI} = [C_3(K)C_4(g_\lambda)]^2R(f''')R(K)^{4/13}\sigma_K^{10/13}R(f'')^{-17/13}/5,$$

$$\sigma_{PI}^2 = (2/25)\sigma_K^{72/13}R(\phi)R(f)R^{-18/13}(K)R^{-8/13}(f'') \div [C_3(K)C_4(g_\lambda)]^9,$$

and

$$\phi(x) = K''*K''(x) = \int K''(t)K''(x + t)\,dt.$$

An editorial decision was made to delete the proofs of Theorems 3.2 and 3.3. Details are available from us.

Remark 3.4. Theorem 3.3 demonstrates that the idea of Hall (1980) and Sheather (1983, 1986), replacing $C_4(f)$ by $C_4(g_\lambda)$, is very reasonable. In particular, observe that the rate of convergence obtained in Theorem 3.3 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 is always superior for n sufficiently large. The issue of how large n needs to be for this to happen is discussed in Section 4.

Remark 3.6. The trade-off for the greater efficiency of $\hat{h}_{PI}$ is that it is dependent on strong assumptions about the underlying smoothness. In particular, note that the asymptotic rate of convergence can be arbitrarily small for ν close to 0. On the other hand, the limiting distribution of $\hat{h}_{CV}$ is independent of ν. This quantifies the statement that $\hat{h}_{PI}$ is

more efficient, whereas $\hat{h}_{CV}$ is more robust. We feel that the first consideration is more important, because it seems that very large sample sizes are required before there is any practical difference between f being smooth or not smooth.

Remark 3.7. Note that the aforementioned results all concern the asymptotic distribution of the various automatically selected bandwidths rather than MISE(h). 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 $\hat{h}$ of either $\hat{h}_{CV}$ or $\hat{h}_{BCV}$ (as $n \to \infty$),

$$n^{1/5}(\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}}) - 1) \xrightarrow{d} 2\sigma^2 \cdot \chi_1^2,$$

for $\sigma^2 = \sigma_{CV}^2$ or σ_{BCV}^2, respectively. An analogous noncentral (since the limiting mean is nonzero) chi-squared limiting distribution can be derived for MISE($\hat{h}_{PI}$).

Remark 3.8. Silverman (1986, sec. 3.6) discussed how faster rates of convergence of $\hat{f}_h$ to f can be obtained through the use of kernels that take on carefully chosen negative values. It is straightforward to adapt the results of this article to that case. The main change in Theorems 3.1–3.3 is that the rates of convergence typically become slower.

Remark 3.9. It is straightforward to use a truncation argument to extend our results to noncompactly supported kernels.

4. SIMULATIONS

Although 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. Only a brief outline of the various choices made for this study is given, because the present setting is closely related to that of Marron (1988c).

The underlying density functions chosen here were the standard normal density, $N(0, 1)$, a mixture of normals with different means, $.5N(-1, 4/9) + .5N(1, 4/9)$, an outlier mixture of normals, $.75N(0, 1) + .25N(0, .04)$, and a variance mixture, $.5N(0, 1) + .5N(0, .09)$.

The sample sizes considered here were $n = 100$ and 400. For each setting 500 repetitions were used. To speed computation we used a binned implementation (Scott and Terrell 1987) with 400 equally spaced bins from -3 to 3.

For comparison of the bandwidth selectors, we chose to compare

$$E\{\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})\}, \tag{4.1}$$

for each $\hat{h}$ of $\hat{h}_{CV}$, $\hat{h}_{BCV}$, and $\hat{h}_{PI}$. These were each calculated from the range of bandwidths $[h_{\text{MISE}}/3, 3h_{\text{MISE}}]$. The scale λ used in $\hat{h}_{PI}$ was the interquartile range.

A bandwidth that uses the "oversmoothing" idea of Terrell and Scott (1985) was included as well. The particular version used here is

$$\hat{h}_{OS} = 7^{1/2}\{2R(K)/(45\sigma_K^2)\}^{1/5}\hat{\sigma}_n^{-1/5},$$

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_{AMISE}, and it 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.

To take the Monte Carlo variability properly into account, we used the pivoted 95% confidence intervals

$$(\hat{C}/(1 + T), \hat{C}/(1 - T)), \tag{4.2}$$

where $\hat{C}$ is an estimate of $C = E\{\text{MISE}(\hat{h})/\text{MISE}(h_{\text{MISE}})\} - 1$ and $T = 1.96(2/\text{NSIM})^{1/2}$. See Marron (1988c) for derivation and discussion of these intervals.

Table 1 contains the main results of the simulation study. For each sample size and distribution, it shows the 95% confidence intervals for the comparison number C given in (4.2). A simultaneous interval scheme should really be used, but this is not done because it tends to obscure the points being made. To provide insight into what affects the relative performances, for each selector the sample means and standard deviations of the bandwidths distributions are given. To attach additional insight to the means of these distributions, the minimizer of MISE(h) is given at the head of each mean column. Other reported statistics include the number of times an endpoint was hit in the minimization/root-finding process and the number of times there was a local minimum/multiple root.

As expected in view of the zero mean in Theorem 3.1, $\hat{h}_{CV}$ was on the average closest to h_{MISE}. Nevertheless, $\hat{h}_{CV}$ rarely performed best, because of its large amount of variability (indicated by the $n^{1/10}$ in Theorem 3.1). Except for the outlier mixture case, the much smaller sample variability (predicted by $n^{4/13}$ in Theorem 3.3) of $\hat{h}_{PI}$ gave it much better performance. In the outlier mixture case, the bias inherent to $\hat{h}_{PI}$ (observe the nonzero mean in Theorem 3.3) turned out to be bad enough to drown out this effect.

Visual insight into these effects can be obtained from Figure 3. These show an overlay of MISE(h), together with kernel estimates of the density of the distribution of the various automatically selected bandwidths, for $n = 400$ observations from (a) the standard normal and (b) the mean mixture. The bandwidth used for these kernel estimates was $\hat{h}_{OS}$, which seems reasonable here in view of the limiting normal distributions available.

The connection between confidence interval performance and means and standard deviations is easily understood by thinking of plugging the bandwidth distributions into the MISE function. For example, in Figure 3a observe that the slight Table 1 superiority ($N(0, 1)$, $n = 400$) of $\hat{h}_{PI}$ over $\hat{h}_{BCV}$ is clearly caused by the fact that $\hat{h}_{PI}$ has slightly

Table 1. Approximate Monte Carlo 95% Confidence Intervals for the Comparison Numbers, Which Allow Comparison of the MISE Performance of the Automatically Selected Bandwidths

Density function	*Confidence intervals for C*	μ_h	σ_h	*Endpoint*	*Multiple root*
Standard normal					
$n = 100$		.445			
OS	(.011, .014)	.441	.031		
PI	(.061, .078)	.479	.066	0	0
BCV	(.073, .094)	.508	.056	0	0
CV	(.195, .251)	.439	.118	9	31
$n = 400$		.330			
OS	(.003, .004)	.334	.012		
PI	(.017, .022)	.344	.026	0	0
BCV	(.019, .025)	.348	.026	0	0
CV	(.118, .156)	.322	.069	7	29
Mean mixture					
$n = 100$		.385			
PI	(.162, .208)	.508	.083	0	0
OS	(.163, .209)	.526	.027		
CV	(.182, .234)	.410	.130	7	43
BCV	(.865, 1.110)	.787	.116	3	19
$n = 400$		.272			
PI	(.066, .085)	.317	.031	0	0
CV	(.093, .119)	.271	.058	5	8
OS	(.331, .424)	.399	.010		
BCV	(.435, .559)	.375	.140	26	55
Outlier mixture					
$n = 100$		.185			
CV	(.182, .234)	.204	.077	4	32
PI	(.223, .286)	.263	.069	0	0
OS	(.697, .894)	.382	.033		
BCV	(1.006, 1.290)	.441	.093	134	41
$n = 400$		.126			
CV	(.083, .106)	.129	.029	0	7
PI	(.108, .139)	.156	.021	0	0
BCV	(.515, .661)	.195	.066	31	45
OS	(1.546, 1.983)	.290	.012		
Variance mixture					
$n = 100$		.186			
PI	(.090, .116)	.212	.038	0	0
CV	(.152, .195)	.191	.052	2	33
BCV	(.635, .815)	.290	.109	53	57
OS	(.726, .931)	.324	.033		
$n = 400$		.134			
PI	(.029, .037)	.145	.013	0	0
BCV	(.038, .049)	.148	.014	0	0
CV	(.096, .123)	.133	.029	3	24
OS	(1.108, 1.421)	.245	.013		

NOTE: Results are based on 500 replications in each case. Included are the means and standard deviations for the populations of selected bandwidths, the number of times an endpoint was hit, and the number of times there were multiple local minima (or roots). The top of each mean column is the minimizer of MISE(h).

less bias and variability. In addition, very poor performance of $\hat{h}_{CV}$ is caused by its large spread.

As expected, $\hat{h}_{OS}$ is clearly superior in the standard normal case because its distribution is tightly clustered near h_{MISE}. It performs poorly in the other cases because, though it still has a very tight distribution, the center point is too large.

Note that in Figure 3, the distribution of $\hat{h}_{CV}$ has substantial skewness toward a heavy tail, and $\hat{h}_{BCV}$ has a bias toward oversmoothing. Both of these points were noted by Scott and Terrell (1987).

Unlike $\hat{h}_{PI}$, which never exhibited a problem of either type, both $\hat{h}_{BCV}$ and $\hat{h}_{CV}$ occasionally had problems with the minimizer appearing outside the range $[h_{MISE}/3, 3h_{MISE}]$, and with the appearance of multiple local minima. The rule used in such cases was to take the largest local minimizer, because this seems to give $\hat{h}_{CV}$ its best chance. This is rather unfair to $\hat{h}_{BCV}$, because in situations where it has its maximum (Scott and Terrell 1987) inside the interval, it was taken as the right endpoint, even when there may have been a reasonable local minimum available. Had our minimization algorithm instead taken the smallest local minimizer, then we believe $\hat{h}_{BCV}$ would give substantially better performance for those data sets where it exhibited many multiple roots (and $\hat{h}_{CV}$ would have been substantially worse using this rule). Nevertheless, we do not feel motivated to investigate this further because of the theoretical superiority of $\hat{h}_{PI}$ and because $\hat{h}_{PI}$ seems far less prone to such problems, and it seems to work better even in those situations where these problems do not come up as often.

We ran some simulations with $n = 25$ and $n = 50$. An

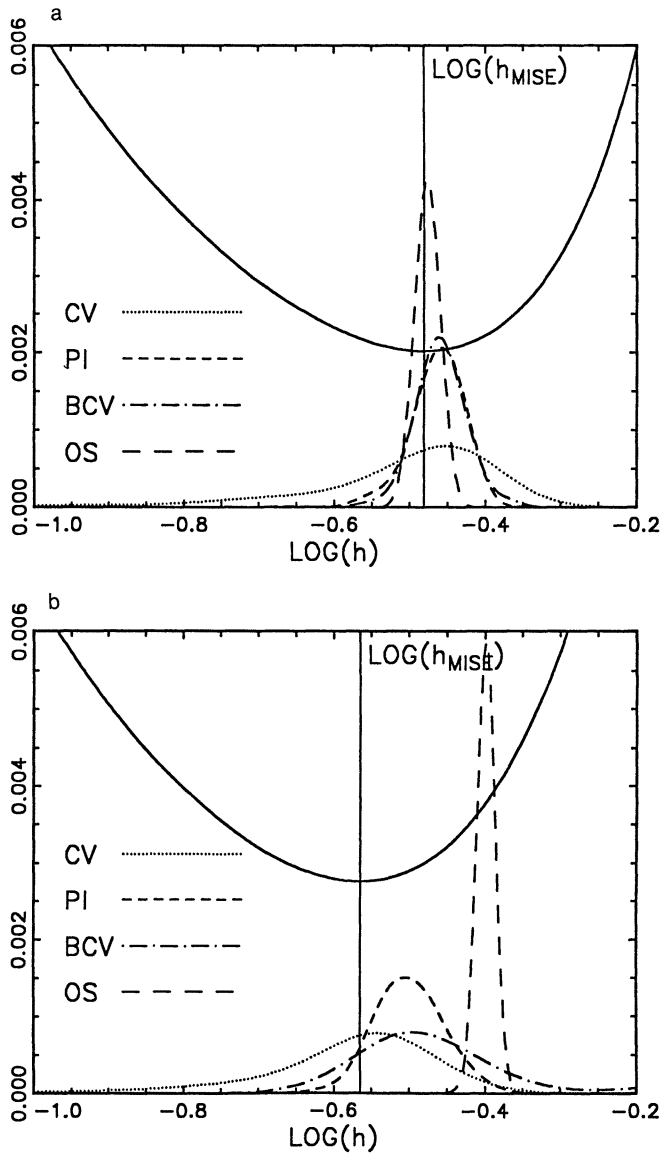

Figure 3. MISE(h) and Kernel Density Estimates of the Distributions of the Automatically Selected Bandwidths (on $\log_{10}$ scale), Taken From 500 Monte Carlo Replications of Samples of Size 400, From (a) N(0, 1) and (b) .5N(−1, 4/9) + .5N(1, 4/9).

editorial decision was made to delete these, but details are available from us. The lessons are roughly the same as indicated in Table 1.

Another issue we investigated was the effect of measuring λ, the scale of f with respect to the reference distribution g_1, in terms of the sample standard deviation instead of the interquartile range. In the standard normal case there was little difference, but the bandwidth selector $\hat{h}_{PI}$ based on the standard deviation performed slightly better. In the mean mixture case, the standard deviation performed substantially better because of less bias (although the variance was slightly bigger). The exact opposite occurred for the outlier and variance mixtures. A way of explaining this comes from comparing $\sigma_{g_1}/\mathrm{IQR}_{g_1}$ to σ_f/IQR_f. The former is smaller in the mean mixture case and bigger in the other cases. As it makes a significant difference in the performance of $\hat{h}_{PI}$, choice of scale measure is an important issue for future research. In addition, it would be interesting to investigate other adjustments, such as using a pilot density estimator instead of the reference distribution g_1.

5. THE EXAMPLE

The fact that the superior performance of $\hat{h}_{PI}$ (demonstrated theoretically in Sec. 3) can yield big dividends in real data situations is demonstrated by Figures 1b and 2. There the bandwidth is chosen by the plug-in rule, and the result is essentially as good as figure 4 of Marron and Schmitz (1988). In that paper a much more complicated pooling procedure was proposed for bandwidth selection.

For these data $\hat{h}_{CV}$ gives completely unreasonable answers (all drastically undersmoothed, apparently because of small-scale clustering in the data). Though $\hat{h}_{BCV}$ performed much better, giving quite reasonable bandwidths for some of the data sets, it was still far too variable to allow anything so demanding as construction of an effective Figure 2. The performance of $\hat{h}_{PI}$ was even better than we had expected, in view of the results of Section 3 and 4, for this challenging data set.

[*Received May 1988. Revised July 1989.*]

REFERENCES

Bowman, A. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika,* 71, 353–360.
Hall, P. (1980), "Objective Methods for the Estimation of Window Size in the Nonparametric Estimation of a Density," unpublished manuscript.
Hall, P., and Marron, J. S. (1987a), "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Probability Theory and Related Fields,* 74, 567–581.
——— (1987b), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters,* 6, 109–115.
——— (1988), "Local Minima in Cross-Validation Functions," unpublished manuscript.
Marron, J. S. (1988a), "Automatic Smoothing Parameter Selection: A Survey," *Empirical Economics,* 13, 187–208.
——— (1988b), "Partitioned Cross-Validation," *Econometric Reviews,* 6, 271–283.
——— (1988c), "Comments on a Data Based Bandwidth Selector," *Computational Statistics and Data Analysis.*
Marron, J. S., and Schmitz, H. P. (1988), "Simultaneous Estimation of Several Size Distributions of Income," Discussion Paper A-186, Sonderforschungsbereich 303, Universität Bonn.
Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics,* 9, 65–78.
Scott, D. W., and Terrell, G. R. (1987), "Biased and Unbiased Cross-Validation in Density Estimation," *Journal of the American Statistical Association,* 82, 1131–1146.
Sheather, S. J. (1983), "A Data-Based Algorithm for Choosing the Window Width When Estimating the Density at a Point," *Computational Statistics and Data Analysis,* 1, 229–238.
——— (1986), "An Improved Data-Based Algorithm for Choosing the Window Width When Estimating the Density at a Point," *Computational Statistics and Data Analysis,* 4, 61–65.
Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* New York: Chapman & Hall.
Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics,* 12, 1285–1297.
Terrell, G. R., and Scott, D. W. (1985), "Oversmoothed Density Estimates," *Journal of the American Statistical Association,* 80, 209–214.
Woodroofe, M. (1970), "On Choosing a Delta Sequence," *The Annals of Mathematical Statistics,* 41, 1665–1671.