On Locally Adaptive Density Estimation

Stephan R. Sain and David W. Scott[1]

January 8, 1996

ABSTRACT:

In this paper, theoretical and practical aspects of the sample-point adaptive positive kernel density estimator are examined. A closed-form expression for the mean integrated squared error is obtained through the device of preprocessing the data by binning. With this expression, the exact behavior of the optimally adaptive smoothing parameter function is studied for the first time. The approach differs from most earlier techniques in that bias of the adaptive estimator remains $O(h^2)$ and is not "improved" to the rate $O(h^4)$. A practical algorithm is constructed using a modification of least-squares cross-validation. Simulated and real examples are presented, including comparisons with a fixed bandwidth estimator and a fully automatic version of Abramson's adaptive estimator. The results are very promising.

KEY WORDS: Kernel Function, Variable Bandwidth, Binning, Cross-Validation.

[1]Stephan R. Sain is Research Associate, Department of Statistical Science, Southern Methodist University, POB 750332, Dallas, TX 75275. David W. Scott is Professor, Department of Statistics, Rice University, POB 1892, Houston, TX 77251. This research was supported in part by the National Science Foundation under grant DMS-9306658 and the National Security Agency under grant MOD 9086-93. The authors would like to thank the readers for many helpful suggestions.

1. Introduction

Precise theoretical understanding of adaptive density estimators as well as the availability of sound practical algorithms has proven surprisingly difficult. Note that the term *adaptive* in this setting does not refer to automatic or data-based bandwidth selection, but rather to local smoothing of the estimated density in order to obtain an improved global estimate. This local smoothing can be achieved by varying the functional form of the kernel or the bandwidth or both. In this work, we will only consider taking the variable bandwidth as a function of the data points.

Let $K_h(t) = h^{-1}K(h^{-1}t)$. For data $\{x_1, \ldots, x_n\}$, the *fixed* kernel estimator is given by

$$\hat{f}_0(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i), \tag{1}$$

where the kernel, K, with finite variance, σ_K^2, generally satisfies $K \geq 0$; $K(-x) = K(x)$; and $\int K = 1$. The smoothing parameter, h, is held constant for all $x \in \Re^1$ and for all the data points. Much of the research in the literature has focused on choosing the proper value of h. For more information on details concerning the fixed bandwidth kernel estimators as well as other density estimators see Silverman (1986), Scott (1992), and Wand and Jones (1995).

In general, two rather intuitive formulations of adaptive or variable bandwidth estimators have been considered (Jones, 1990). The first varies the fixed bandwidth with the estimation point and is often referred to as a *balloon* estimator. Its form is given by

$$\hat{f}_1(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - x_i), \tag{2}$$

where $h_x = h(x)$. The balloon estimator was first introduced by Loftsgaarden and Quesenberry (1965) in the particular form of the kth nearest neighbor estimator. Define $R(\phi) = \int \phi(t)^2 dt$. Then standard approximations to the mean squared error lead to the optimal error rate of $O(n^{-4/5})$ with the optimal choice for the bandwidth of the form

$$h_x^* = \left[\frac{R(K) f(x)}{n \sigma_K^4 f''(x)^2} \right]^{1/5}, \tag{3}$$

which is well-defined, except at points of inflection; see Rosenblatt (1956) and Terrell and Scott (1992). While Terrell and Scott (1992) demonstrated that the kth nearest neighbor and other more general balloon estimators show some promise in higher dimensions, such formulations can suffer severe drawbacks in the univariate and bivariate settings.

The second variable bandwidth procedure is referred to as the *sample-point* estimator, in which the bandwidth is varied with each data point and not with the estimation point. The functional form of the sample-point estimator is given by

$$\hat{f}_2(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i} \right) = \frac{1}{n} \sum_{i=1}^{n} K_{h_i}(x - x_i), \tag{4}$$

where $h_i = h_{x_i} = h(x_i)$. This type of estimator was introduced by Breiman, Meisel, and Purcell (1977), who suggested using $h_i \propto f(x_i)^{-1/d}$, where d is dimension of the data. One clear advantage of this procedure over the balloon estimator is that the sample-point estimator will always integrate to one.

Abramson (1982) suggested using $h_i \propto f(x_i)^{-1/2}$ regardless of dimension and showed that this adaptive choice can outperform the fixed bandwidth estimator for pointwise estimation. In fact, Abramson and others (including Silverman (1986) who termed such estimators "adaptive") have shown that this "square-root" law yields $O(h^4)$ local bias, clearly superior to the $O(h^2)$ bias exhibited by the usual fixed bandwidth approach with positive kernels. The rate is consistent with the "higher-order" kernels (Bartlett, 1963) that remove the restriction of non-negativity. However, recent research by Terrell and Scott (1992) and Hall, Hu, and Marron (1994) shows that in certain cases, the square-root law can in fact perform significantly worse than the fixed bandwidth approach. While the improved bias properties of Abramson's square-root law are promising, finding a reasonable global implementation of Abramson's approach has proven rather elusive. Other adaptive procedures, some more practical, have been studied by Schucany and Sommers (1977), Terrell and Scott (1980), and Jones et al. (1994). Each modifies a $O(h^2)$ bias kernel estimator to obtain a global $O(h^4)$ bias estimator.

While much research has centered on characterizing the improved bias properties of the square-root law, studying global properties of the general sample-point estimator is difficult without knowledge of the form of the optimal bandwidth function. In this work, a binning preprocessor is utilized in order to characterize and estimate the mean integrated squared error (MISE) of the sample-point estimator, to determine the behavior of the optimal adaptive bandwidth function, and to study the resulting adaptive estimates. In addition, a cross-validation procedure is proposed to estimate the unknown smoothing parameters.

Our philosophy differs substantially from previous algorithms. We do not seek to use our degrees of freedom for adaptivity in order to eliminate the $O(h^2)$ bias term. Rather, we wish to optimize within the family of $O(h^2)$ bias, since we believe a well-constructed adaptive $O(h^2)$ procedure will have more practical and graphical appeal than many $O(h^4)$ algorithms. We note our methodology can be extended to construct adaptive $O(h^4)$ algorithms.

2. Binning in Kernel Density Estimation

Scott (1981), Silverman (1982), Scott and Sheather (1985), and Jones (1989) describe the notion of a binned kernel density estimator as a practical approach to the fixed bandwidth kernel density estimator, $\hat{f}_0(x)$, in (1). Bin counts $\{n_j\}$ for bins $\{B_j\}$ are computed for an equally spaced mesh with bin centers $\{t_j\}$. If δ is the bin width, then $t_{j+1} - t_j = \delta$ and $\sum n_j = n$. In practice, $\delta < h$. The binned kernel density estimator is then given by

$$\hat{f}_b(x) = \frac{1}{n} \sum_{j=-\infty}^{\infty} n_j K_h\left(x - t_j\right) = \frac{1}{n} \sum_{j=1}^{m} n_j K_h\left(x - t_j\right), \tag{5}$$

since, in reality, only a finite number of bins contain data; the summation is taken over the

m nonempty bins.

Hall (1982) derives the pointwise MSE of the kernel estimator using both rounded and truncated data. Scott and Sheather (1985) expand on Hall's result for rounded data to give an asymptotic form of the MISE (AMISE) for the binned estimator (5),

$$\text{AMISE}_b = \frac{R(K)}{nh} + \frac{1}{4}h^4\sigma_K^4\left(1 + \frac{\delta^2}{12h^2\sigma_K^2}\right)^2 R(f'').$$

Note that binning inflates only the bias term. However, Scott and Sheather (1985) note that the AMISE is relatively insensitive to reasonable amounts of binning. Hall and Wand (1994) have shown that fractional binning (in which each datum is partially split with an adjacent bin) reduces the bias inflation from $O(\delta^2)$ to $O(\delta^4)$.

The binned version of the sample-point estimator (4) is given by

$$\tilde{f}(x) = \frac{1}{n}\sum_{j=1}^{m} n_j K_{h_j}(x - t_j). \tag{6}$$

The introduction of binning in the fixed kernel estimator was an effort to ease the computational burden. In this work, the binned version (6) is used solely to obtain a set of bandwidths $\{h_j\}$ to use in (4). Form (6) facilitates the derivation of a useful approximate expression for the adaptive MISE. Efforts to use the usual asymptotic error approximations directly on the unbinned adaptive estimator (4) have proven futile. While Terrell and Scott (1992) give an asymptotic expression for the MSE of the general sample-point estimator, calculating exact expressions or even asymptotic approximations for the MISE of the sample-point estimator to study global properties are impossible since the form of the bandwidth function, $h_i = h(x_i)$, is unknown. However, it is a straightforward exercise to derive the exact MISE for (6). An alternative to (6) is described in Priebe (1994) and involves modeling the data with a finite normal mixture density where the mixing parameters are chosen by maximum likelihood.

3. MISE Calculations

Using the normal kernel in (6), the MISE of the binned sample-point estimator can be derived in closed form:

$$\begin{aligned}
\text{MISE} &= E\int[\tilde{f}(x)-f(x)]^2dx \\
&= E\int \tilde{f}^2(x)\,dx - 2E\int \tilde{f}(x)f(x)\,dx + \int f^2(x)\,dx \\
&= \frac{1}{n^2}\sum_j E\left[n_j^2\int \phi_{h_j}^2(x-t_j)\,dx\right] + \frac{1}{n^2}\sum_{i\neq j} E\left[n_i n_j \int \phi_{h_i}(x-t_i)\phi_{h_j}(x-t_j)\,dx\right] \\
&\quad -\frac{2}{n}\sum_j E\left[n_j\int \phi_{h_j}(x-t_j)f(x)\,dx\right] + R(f).
\end{aligned}$$

Since the only random quantities are the bin counts, n_j, which are multinomial with parameters given by the bin probabilities, p_j, where $p_j = \int_{B_j} f(x)\,dx$, evaluating the expectations becomes a trivial exercise. Taking expectations and integrating (see Wand and Jones (1995), Appendix C), the final form of the exact MISE is

$$\begin{aligned}
\text{MISE} &= \frac{1}{2n\sqrt{\pi}}\sum_j \frac{p_j(1-p_j)+np_j^2}{h_j} + \frac{n-1}{n}\sum_{i\neq j} p_i p_j \phi_{\sqrt{h_i^2+h_j^2}}(t_i-t_j) \\
&\quad -\frac{2}{n}\sum_j p_j \int \phi_{h_j}(x-t_j)f(x)\,dx + R(f).
\end{aligned} \tag{7}$$

The MISE expression given in (7) can now be used to study the characteristics of the sample-point estimator. Partitioning the support of a known density f into m distinct bins, the bin probabilities can be found, the integral in the cross-product term evaluated, and the criterion optimized over the space of possible smoothing parameters $\{h_j,\ j=1,\ldots,m\}$.

Figure 1 shows the behavior of the optimal binned sample-point estimator for $f = N(0,1)$ as both the number of bins and the sample size increase. The dotted line in the figure shows the MISE for the (optimal) fixed bandwidth estimator. Note that with too few bins, the binned estimator performs poorly with respect to the (unbinned) fixed bandwidth estimator, both in terms of the absolute improvement for a fixed n as well as the rate at which the MISE goes to zero. As the number of bins increases, the binned estimator (6) more closely approximates the behavior of the (unbinned) sample-point estimator (4), and

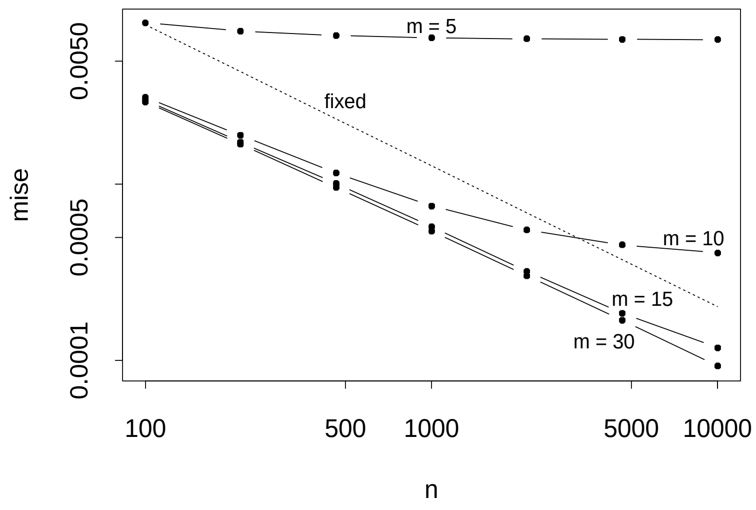

Figure 1: MISE($\tilde{f}$) versus sample size for several levels of binning for a standard normal density. The dotted line represents MISE of the optimal fixed bandwidth estimator.

yields considerable improvement over the fixed bandwidth estimator (1) for most reasonable sample sizes. For example, when $m = 30$, the binned estimator will outperform the fixed bandwidth estimator until $n = 2.65e10$. Furthermore, since the lines are essentially parallel for large m, the MISE for the binned estimator shares the $n^{-4/5}$ rate for the fixed bandwidth estimator. However, the MISE is reduced by slightly more than 50% by the adaptive scheme. This is particularly noteworthy and significant since the adaptive balloon estimator (2) was shown by Terrell and Scott (1992) to afford a theoretical reduction in MISE of only 8.5% in this setting, regardless of sample size. In the following, the reasons for this unexpected improvement are explored.

Figure 2(a) displays the set of optimal smoothing parameters with $m = 20$ equally-spaced bins for sample sizes $n = 10^2$, 10^3, 10^4 for the standard normal density. There are several interesting aspects of the pattern of optimal smoothing parameters that deserve attention. First, the optimal bandwidths follow the general rule of having small values of h_i for data in the neighborhood of the modes and larger values of h_i in the tails of the distribution. As n gets larger, the behavior of the smoothing parameters for the different

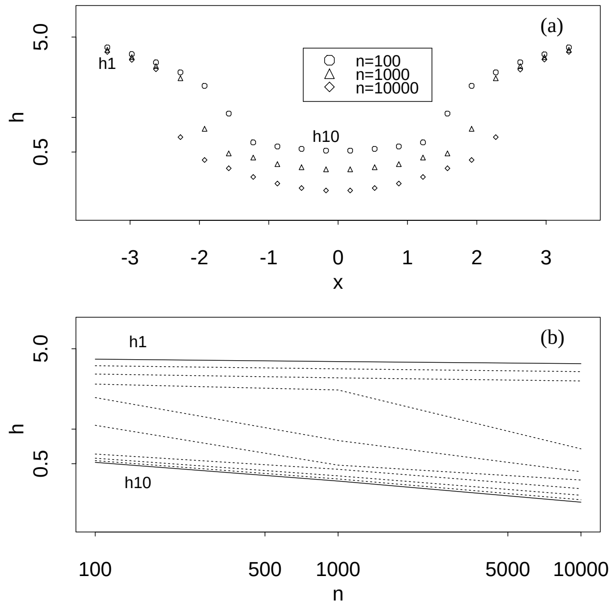

Figure 2: (a) Optimal adaptive binned smoothing parameters for $m = 20$ and $f(x) = N(0,1)$. (b) A plot of h_i versus n for h_1 through h_{10} (top-to-bottom)— by symmetry, $h_{11} = h_{10}, \ldots, h_{20} = h_1$.

bins is quite different as shown in the Figure 2(b). The smoothing parameters near the mode of the distribution appear to be going to zero at an approximate rate of $n^{-1/5}$, similar to that of the optimal fixed bandwidth choice of h. However, the rate at which the smoothing parameters in the tails is going to zero is almost ten times as slow. In fact, it is not clear from the figures that these values will go to zero at all. Note that this behavior is reminiscent of the zero-bias bandwidths associated with balloon estimators discussed in Sain and Scott (1995). In that work, the authors show that the theoretically optimal pointwise bandwidths in many regions of the underlying density (convex regions, for example, in the tails) converge, not to zero, but rather to constant (positive) bandwidths yielding zero bias. Another interesting feature in Figure 2(b) is apparent for the "transition" bins over the intervals $1.5 < |x| < 2.5$, where the optimal bandwidths switch from larger values in the tails to smaller values near the mode.

Five examples of "optimal" estimates with samples of standard normal data of size

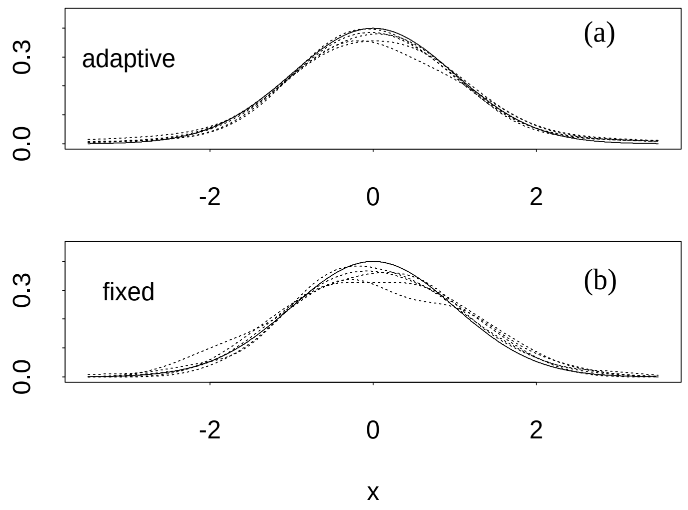

Figure 3: Five estimates of standard normal data with $n = 100$ for (a) the binned sample-point estimator and (b) the fixed kernel estimator.

$n = 100$ are displayed in Figure 3. Note again that the binned sample-point estimator (6) is not actually used to estimate the densities. After the optimal smoothing parameters are found by the numerical optimization procedure in Splus, the smoothing parameter associated with that bin is assigned to each data point in that particular bin. The unbinned sample-point estimator in Equation (4) is then used to calculate the estimates. (A more sophisticated implementation would have two levels of binning, one for estimating the $\{h_j\}$ and the other for estimating the density.) The adaptive estimates using this procedure are shown in the top frame of Figure 3 with the corresponding fixed bandwidth estimates shown below, using $h = (4/3n)^{1/5}$, which minimizes the AMISE for standard normal data. The improvement predicted in Figure 1 is reflected across the entire range of the estimates displayed in Figure 3(a).

Figure 4 compares the theoretical behavior of the bandwidths for the binned approach and the Abramson square-root proposal. The top plot uses $m = 15$ equally spaced bins for the standard normal density, while the bottom plot uses $m = 25$ bins for the bimodal mixture density, $f(x) = \frac{3}{4}\,\phi_1(x+\frac{3}{2}) + \frac{1}{4}\,\phi_{1/3}(x-\frac{3}{2})$. The solid line represents the bandwidth

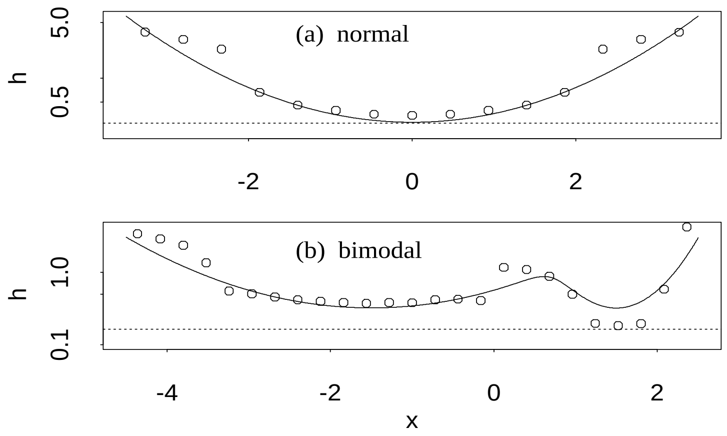

Figure 4: Plots of optimal binned smoothing parameters, h_j (circles), for the (a) $N(0,1)$ and (b) bimodal mixture densities when $n = 1000$. The solid line indicates the optimal Abramson bandwidth function; the dotted line represents the optimal fixed bandwidth.

function defined by Abramson's proposal, $h/f^{1/2}(x)$, using the unknown but true density to find the constant h minimizing the numerically integrated form of MISE. The horizontal dotted lines in the figure locate the optimal fixed bandwidths (see the exact MISE formulas of Marron and Wand (1992)). It may seem like an obvious error that all of the adaptive h_j are above the dotted line. (For example, does this imply that $\tilde{f}(0)$ is severely biased downward?) However, the adaptive estimate is increased at the mode by contributions from the unusually large bandwidths in the tails.

Compare the optimal bandwidths found by the binned and Abramson's procedures. For the standard normal case, the optimal binned bandwidths near the mode are slightly larger than that of Abramson's method, while in the tails the bandwidths are flatter, as opposed to the continued rapid increase of Abramson's proposal. This does suggest that the clipping procedures originally suggested, but largely ignored, by Abramson (1982) would be beneficial if the proper amount and form of clipping could be determined; see also McKay (1993). However, as noted in Sain (1994), no significant improvement in the MISE for the square-root law was found by implementing the clipping procedures proposed by McKay.

In the bimodal case, the two variable bandwidth functions are increasingly dissimilar. In the neighborhood of the left-hand mode at $x = -1.5$, the two bandwidth functions are fairly consistent with those of the standard normal discussed above. However, near the right mode at $x = 1.5$, which has a smaller scale parameter than that of the left mode, the optimal binned bandwidths in the neighborhood of the mode are significantly smaller than Abramson's. The true mixture density was selected so that the levels at the two modes were equal, but the curvature was greater for the right mode. The magnitude of the bandwidth chosen by the square-root law, $h/f(x_i)^{1/2}$, is determined solely by the height of the density near the mode and does not account for differences in curvature (scale). Formula (3) for h_x^* suggests that both f and f'' are relevant. In this case, Abramson's method lacks a certain flexibility, in part accounting for its poor large-sample performance (Terrell and Scott, 1992).

These differences in bandwidth functions result in quite different estimates, as shown in Figure 5. Five estimates of the bimodal mixture density for $n = 1000$ with the binned sample-point estimator, Abramson's square-root law, and the fixed kernel estimator are shown from top to bottom. (Again, the optimal smoothing parameters were found for all three.) Note that the Abramson estimator does a fairly good job of smoothing the left mode, but fails to adequately emphasize the right mode as only the level and not curvature is accounted for. Even at this sample size, the fixed bandwidth procedure has obvious trouble, exhibiting a large amount of variability near the left mode as the optimal fixed bandwidth reflects a compromise between the two optimal fixed bandwidths for each mode. In general, the fixed bandwidth procedure simultaneously undersmooths and oversmooths in different regions.

To study the differences between the binning procedure and Abramson's approach, Figures 6(a) and 6(b) display the MISE as a function of sample size for the standard

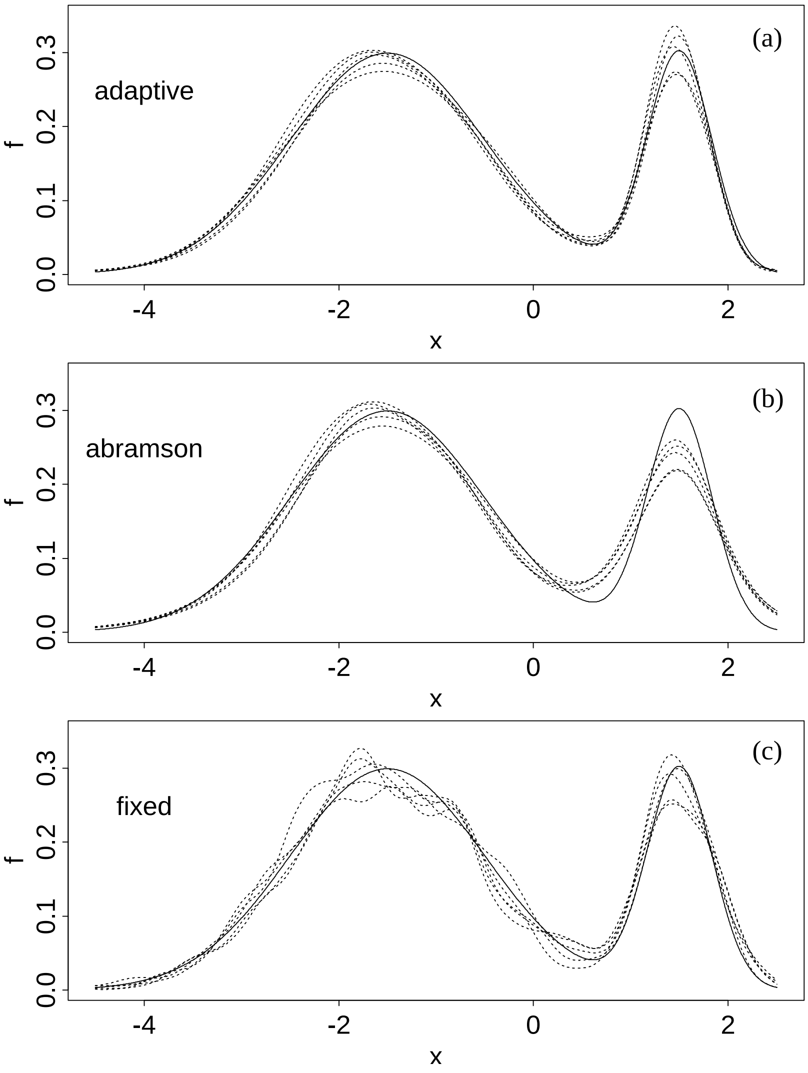

Figure 5: Five estimates of a bimodal mixture density with $n = 1000$ for (a) the binned sample-point estimator; (b) Abramson's square-root law; and (c) the fixed bandwidth estimator.

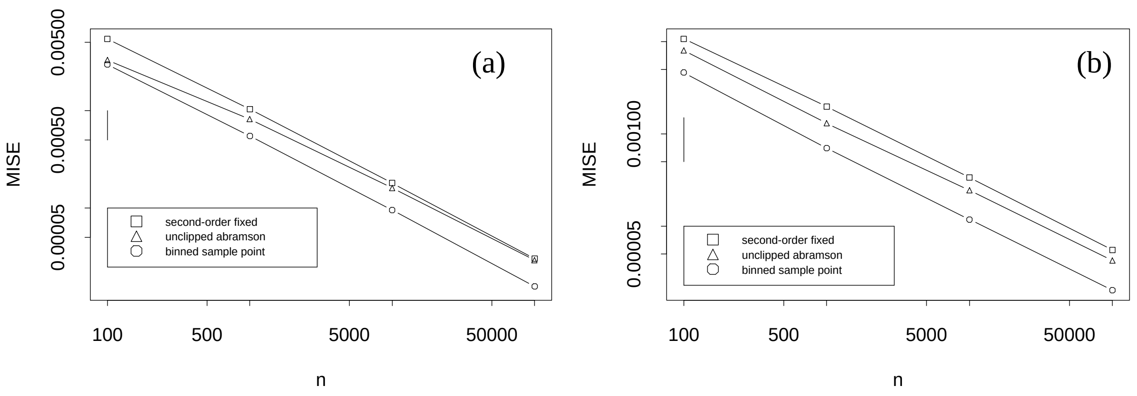

Figure 6: Sample size versus MISE for various methods with (a) standard normal data and (b) bimodal data. The height of the vertical line in (a) on the left side of the plot represents a factor of two in the magnitude of the MISE, and a factor of 3 in (b).

normal and bimodal mixture to study the convergence of the MISE. In both cases, for small samples, the binned sample-point estimator and Abramson's approach both outperform the fixed bandwidth. However, the gains of Abramson's square-root law over the fixed kernel procedure diminish as the sample size increases, while the proportional gain in the MISE continues for the binned sample-point estimator. In fact, the binned estimator has a MISE that is about half that of the fixed bandwidth estimator in the normal case and about a third in the bimodal case. While an unclipped version of Abramson's estimator is shown in the plots, similar results where obtained by implementing the clipping proposals of Abramson (1982) and McKay (1993).

4. A Practical Cross-Validation Approach

Any practical algorithm utilizing the binning approach must address two crucial issues. The first is selecting the proper number of bins and the second is accurate estimation of the smoothing parameters associated with each of the bins. In this section, the least-squares or unbiased cross-validation (UCV) criterion (Rudemo, 1982; Bowman, 1984) is shown to accomplish both tasks simultaneously. In fact, Rudemo discussed estimating adaptive histogram bin widths in his paper.

UCV seeks to minimize an estimate of the integrated squared error (ISE), given by

$$\text{ISE} \;=\; \int [\tilde{f}(x) - f(x)]^2 dx = R(\tilde{f}) - 2\int \tilde{f}(x) f(x)\, dx + R(f).$$

Note that the third term, $R(f)$, is constant with respect to the unknown values of the smoothing parameters and can be ignored in the minimization, while $R(\tilde{f})$ is solely a function of the binned sample-point estimator and, using a normal kernel, is given by,

$$\begin{aligned} R(\tilde{f}) &= \frac{1}{n^2}\sum_{j=1}^{m} n_j^2 \int \phi_{h_j}^2(x - t_j)\, dx + \frac{1}{n^2}\sum_{i \neq j} n_i n_j \int \phi_{h_i}(x - t_i)\phi_{h_j}(x - t_j)\, dx \\ &= \frac{1}{2\sqrt{\pi} n^2}\sum_{j=1}^{m} \frac{n_j^2}{h_j} + \frac{1}{n^2}\sum_{i \neq j} n_i n_j \phi_{\sqrt{h_i^2 + h_j^2}}(t_i - t_j). \end{aligned} \tag{8}$$

The ISE cross-product term, $\int \tilde{f}(x) f(x)\, dx = E[\tilde{f}(X)]$, can be estimated in an unbiased manner by using the leave-one-out estimator,

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{-i}(x_i), \tag{9}$$

where

$$\tilde{f}_{-i}(x_i) = \frac{1}{n-1}\sum_{j=1}^{m} n_{ij}^* K\left(\frac{x_i - t_j}{h_j}\right)$$

and

$$n_{ij}^* = \begin{cases} n_j - 1 & x_i \in B_j \\ n_j & \text{otherwise.} \end{cases}$$

Combining (8) and (9), the adaptive cross-validation criterion function is given by

$$\text{UCV}(h_1, \ldots, h_m) = R(\tilde{f}) - \frac{2}{n}\sum_{i=1}^{n} \tilde{f}_{-i}(x_i). \tag{10}$$

Estimates of the optimal smoothing parameters are found by numerically minimizing the UCV criterion over $\{h_j,\, j = 1, \ldots, m\}$ and the integer m.

Rudemo (1982) and Bowman (1984) showed that UCV is unbiased for any (nonrandom) choice of smoothing parameter in the sense that $E[\text{UCV}] + R(f) = \text{MISE}$. Hence the alternate terminology, "unbiased" cross-validation. It is straightforward to show that such is true for the case of the binned sample-point estimator.

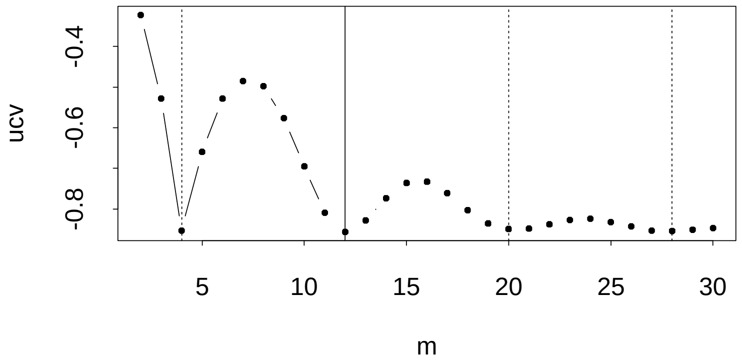

Figure 7: Cross-validation criterion for bimodal data example. Local minima are indicated by dotted lines while the global minima at $m = 12$ is shown by a solid line.

Theorem (*Unbiasedness of UCV for $\tilde{f}$*)**.** Let $\tilde{f}$ be the binned sample-point estimator defined in (6). Also, let UCV be the least-squares cross-validation criterion defined for $\tilde{f}$ in (10). Then UCV is unbiased in the sense that

$$E[\mathrm{UCV}] + R(f) = \mathrm{MISE}(\tilde{f}).$$

Proof: See Appendix.

A practical procedure utilizing cross-validation must use enough bins to achieve a sufficient approximation of the adaptive sample-point procedure to gain an edge over a fixed bandwidth approach, while limiting the number of bins to achieve reasonable approximations of optimal smoothing parameters (depending on sample size) and to avoid numerical optimization instabilities when n_j is small. To achieve this tradeoff, data are first binned according to a fixed size mesh, and the optimal smoothing parameters are found by minimizing the cross-validation criterion given above. Then the number of bins is increased, and the optimization repeated. Selection of the proper number of bins is accomplished by choosing the number of bins yielding the smallest cross-validation criterion evaluated at the optimal set of smoothing parameters.

To study the behavior of this procedure, a sample of size $n = 500$ was drawn from a bimodal normal mixture, $f(x) = 0.5 * [\phi_{0.25}(x) + \phi_{0.01}(x - 1.5)]$, and was binned over the

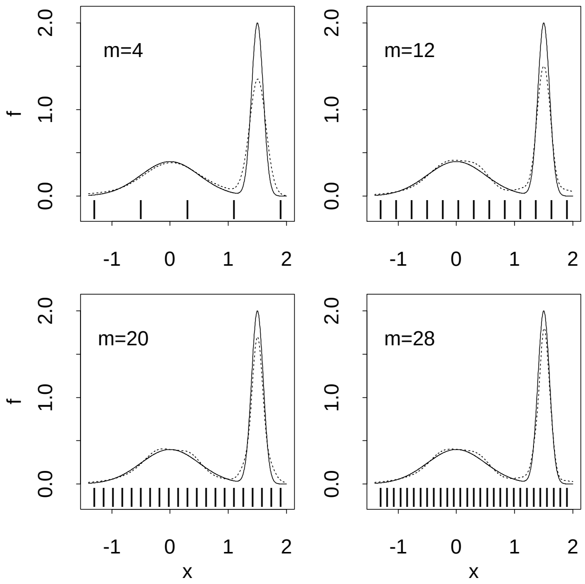

Figure 8: Estimates ($n = 500$) from using smoothing parameters from local minima of the UCV. Bin locations are indicated under each plot.

interval $(-1.3, 1.9)$. The maximum number of bins is limited by the largest number of non-empty bins. In Figure 7, the optimal values of the cross-validation criterion across a number of bins are displayed, showing four local minima with a global minima at $m = 12$. The fluctuations in the UCV criterion are due in part to an interaction between locations of the bin edges and the modes; shifting meshes could be considered as well. The UCV criterion can be used to pick both the number of bins and the shift, although our experience indicates that the general appearance of the final density estimate is not affected much by the choice of bin edge (see discussion of the example in Figure 11).

Estimates using the optimal smoothing parameters for each local minima of the UCV criterion (m = 4, 12, 20, and 28) are shown in Figure 8. There is little difference in the optimal estimates, except near the right mode, where the increased flexibility of the larger number of bins gives a better estimate near the right mode but at the expense of slightly more variability near the left mode.

The increased flexibility of larger number of bins also incurs a higher cost in terms of

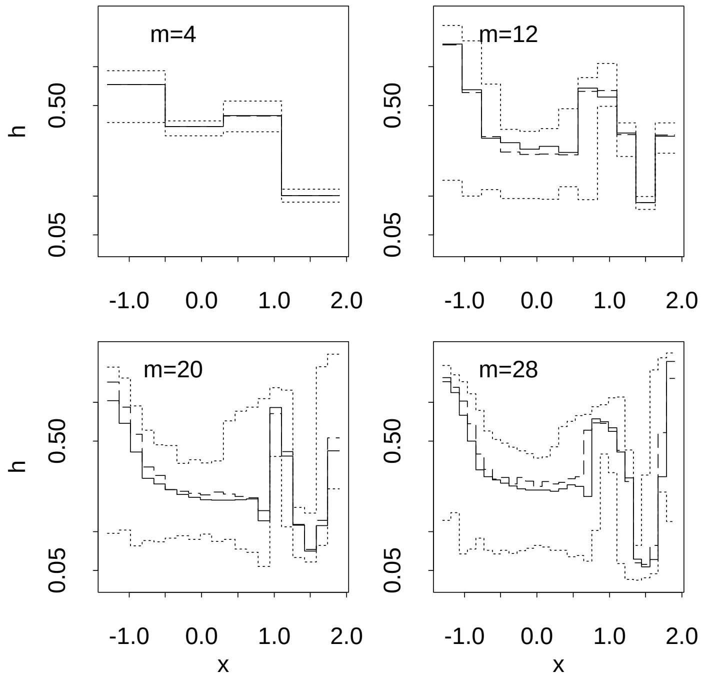

Figure 9: Results from a simulation study using the bimodal density: solid lines indicate the optimal MISE bandwidth function; dashed lines indicate the mean bandwidth function calculated from 100 replications; and dotted lines represent empirical 95% confidence bands.

variability of the estimated smoothing parameters. To illustrate, one hundred samples of size $n = 500$ were drawn from this bimodal density, and estimated smoothing parameters were calculated for each number of bins $m = 4$, 12, 20, and 28. The results are shown in Figure 9. As the number of bins increase, so does the variability associated with the estimation of the bandwidths. Thus, the number of bins can be considered a "smoothing parameter" that attempts to balance flexible estimation of the unknown bandwidth function with accurate estimation of the unknown smoothing parameters. However, direct examination of estimates resulting from each local minima would be necessary to gain some insight to the structure of the data.

5. Examples

In this section, four examples utilizing actual data are shown to further illustrate the

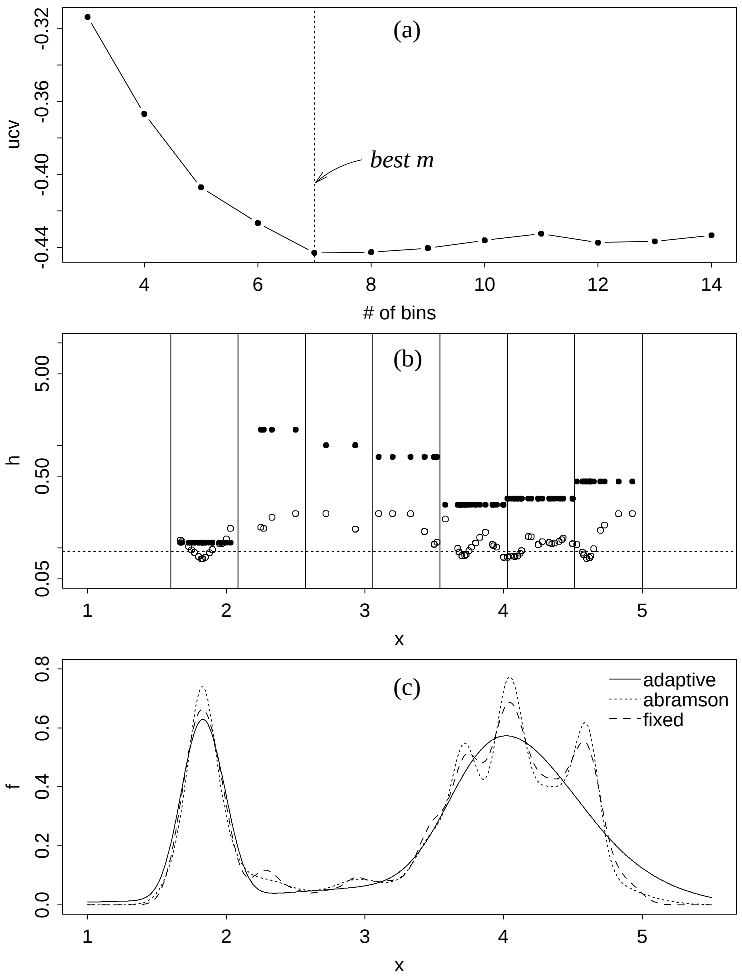

Figure 10: Old Faithful example: (a) UCV criterion; (b) bandwidth functions — UCV (filled circles), Abramson (open circles), fixed (dashed line); (c) density estimates.

use of the methodology outlined in the previous sections. The first example is shown in Figure 10 and uses 107 measurements of the length in minutes of eruptions of the Old Faithful geyser taken from Weisberg (1985). In the frame (c), the data are clearly bimodal, with well-separated modes. The two modes, while approximately the same height, differ in scale. The (a) frame in the figure shows the number of bins versus the value of the adaptive UCV cross-validation criterion. The interval of interest is taken to be fixed, and the maximum number of bins restricted so that empty bins are not allowed. A global minima is achieved at $m = 7$. The middle frame shows the estimated bandwidths for each bin (solid circles), as well as the final bin boundaries. Note that the bandwidth for each data point is plotted, as the binned estimator is not actually used for estimation. Observe the difference in bandwidth near the two modes. Data points near the left mode with smaller scale (larger curvature) are given a significantly smaller bandwidth than those near the right mode with larger scale (smaller curvature).

Estimates of the density are shown in the bottom frame. A fixed bandwidth estimator with bandwidth chosen by ordinary UCV ($h = 0.0924$) as well as an Abramson-style estimator are shown for comparison. The Abramson estimator utilizes a fixed kernel pilot estimator in calculating the bandwidths $h_i = h/\hat{f}_\lambda(x_i)^{1/2}$. The smoothing parameter for the pilot estimator, λ, and h were jointly chosen by minimizing the least-squares cross-validation criterion ($\hat{\lambda} = 0.0309$ and $\hat{h} = 0.0753$).

Note that the fixed and Abramson estimates are remarkably similar, as the estimated bandwidths for both methods are roughly equal. Both procedures do an adequate job of estimating the left mode while leaving the right mode undersmoothed. The adaptive binning procedure appears to do a better job at estimating both modes.

A second example is shown in Figure 11, with the Buffalo snowfall data ($n = 63$). Here, choosing the appropriate number of bins is a bit more difficult, as shown in the top plot of Figure 11. Mesh shifting has some affect here due to ties and the small sample size. This data set does exhibit evidence of trimodality, and the rough nature of the UCV criterion as a function of the number of bins results from the placement of bin edges near these modes. The unshifted choice of $m = 10$ is reasonable, but the best UCV occurs with a 40% shift when $m = 7$. (This solution actually has 8 bins, since the shifting introduces an extra nonempty bin on the left.) Bandwidth functions are shown in the middle plot. The binned procedure shows clear evidence of the trimodality, with smaller bandwidths near $x = 50$, 80, and 110, while finding considerable larger bandwidths between these modal bins.

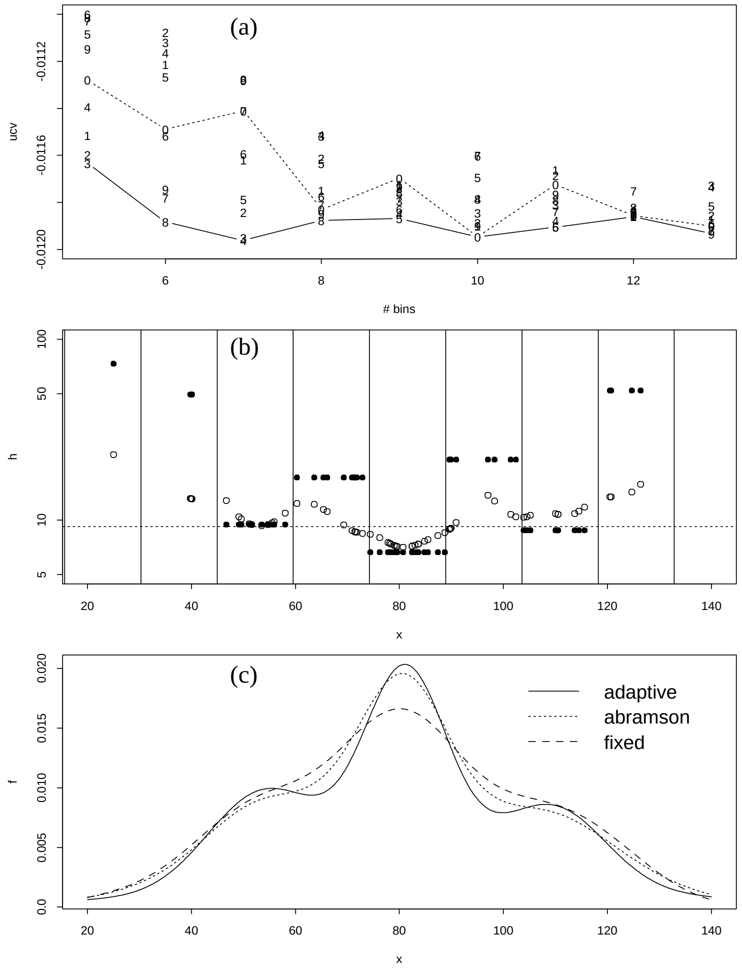

Figure 11: Buffalo snowfall data example (see legend for previous figure). In frame (a), nine equally shifted meshes were evaluated in addition to the original unshifted mesh (0 – 9). The best and original UCV values are connected by solid and dotted lines, respectively.

Abramson's procedure ($\hat{\lambda} = 2.93$ and $\hat{h} = 1.07$) and the fixed bandwidth procedure ($\hat{h} = 9.18$) are also shown. The Abramson method yields a bandwidth function that does show some evidence of three bumps, but it is not as pronounced as the binned procedure.

The estimates are shown in the bottom frame. The most conservative of the three, the fixed bandwidth procedure, barely hints that the data may be trimodal. Abramson's method emphasizes the primary mode, while exhibiting two shoulders or bumps rather than two additional modes. The adaptive procedure, however, gives clearer evidence that the data are trimodal. Parzen (1979) speculated that trimodality might reflect overfitting.

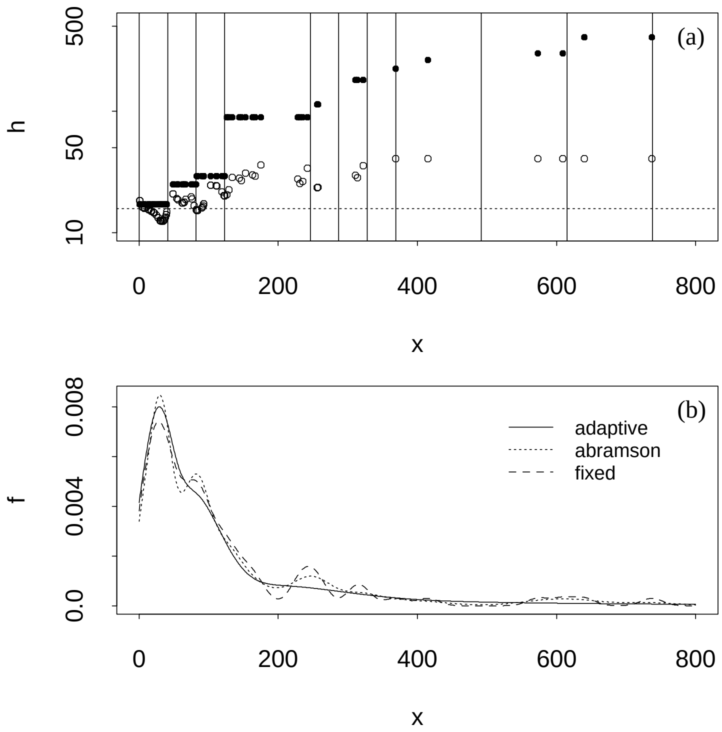

Figure 12: Suicide data example (see previous legend).

The next data set is the suicide example ($n = 86$); see Silverman (1986). The density is strongly skewed, exhibiting a long right tail. In this case, $m > 6$ leads to empty bins. Unfortunately, six bins does not yield improvement over the fixed bandwidth estimator.

In this case, a procedure based on the "partitions of locally equisized cells" (POLEC) studied by Kogure (1987) and Simonoff and Hurvich (1993) is employed. First, the data are partitioned into six equally-spaced bins. Then, each of those bins is further split into equally-spaced bins until no further splitting reduces the UCV criterion. The result of this splitting and the corresponding bandwidths is shown in Figure 12(a). Note that not all bins required additional splitting.

Also shown in the top frame are the estimated bandwidth functions for the Abramson procedure ($\hat{\lambda} = 4.71$ and $\hat{h} = 1.27$) and the fixed bandwidth procedure ($\hat{h} = 15.7$). Abramson's bandwidth function again represents a compromise between the other two procedures.

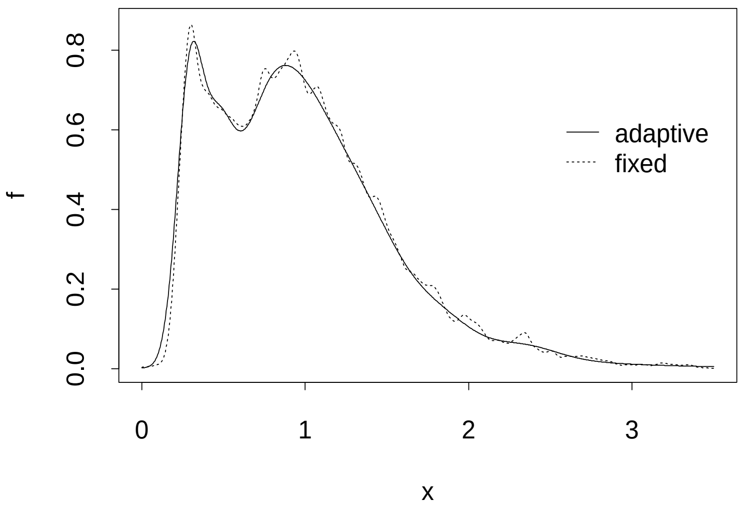

Figure 13: Economic data example.

Estimates are shown in Figure 12(b). The adaptive procedure smoothes out most of the roughness in the tail, including the spurious modes near $x = 250$, while still adequately smoothing the mode. The adaptive procedure indicates a shoulder near $x = 100$ rather than the clear mode shown by the other two procedures. Both Abramson and fixed methods are dominated by the large curvature near the mode, giving smaller smoothing parameters throughout the range of the estimated density and larger noise levels in the tails.

Our final example is a much larger sample ($n = 7201$) of economic data from an income survey taken in 1975 in the UK. The data have been normalized so the average is one. These data have been analyzed by several authors, for example, Park and Marron (1990), Wand et al. (1991), and Kooperberg and Stone (1991). In Figure 13, a fixed kernel estimate ($\hat{h} = 0.0287$) and an adaptive estimate ($m = 20$) are displayed. The adaptive estimate is able to capture the bimodal structure as well as adequately smooth the tail as opposed to the fixed bandwidth estimate which leaves much of the distribution undersmoothed.

6. Discussion

In this paper, we have advocated the use of binning to approximate an adaptive sample-point density estimator. This simple device allows theoretical study of a powerful alternative to the fixed bandwidth estimator. This long-overdue study has been elusive in the literature. At its heart, the binning procedure discussed here as an adaptive or variable bandwidth procedure does not seek to eliminate the $O(h^2)$ bias term as in other proposed procedures, but rather, through deliberate construction, to stay within the $O(h^2)$ family while still improving upon the fixed bandwidth approach in important cases.

On the practical side, the cross-validation algorithm is powerful in its simplicity and performs as the theory predicts. The dimension of the optimization problem is reduced from n to m as only the m unknown bandwidths must be estimated. Furthermore, the UCV criterion is relatively easy to calculate, being primarily a function of the bin counts. It was also shown that the UCV procedure does an excellent job of estimating the unknown parameters, and the examples demonstrate the algorithms superior performance with real data. We believe it may be possible to prove that the notorious variability of UCV for fixed kernels is reduced with binned variable estimators.

This procedure shows real promise in the low-dimensional multivariate setting, as binning could be a practical solution to the problems of the scarcity of data in higher dimensions. However, continued work on properly choosing the form of the mesh, including variable sized meshes, will be more important for the multivariate case.

A related practical idea is to define the adaptive smoothing parameter function $h(x)$ to be a cubic spline and to use UCV on that; in one sense, our procedure is of this variety but using a zero-order spline. There are no simple closed-formed theoretical MISE expressions available for this unbinned adaptive UCV algorithm. This formulation was proposed independently by Fan, et al. (1993) and Sain (1994). The relative performance of these ideas has not been thoroughly tested to date. Another alternative is to smooth the estimated bandwidth function by fitting a spline to the binned estimates in a manner similar to that suggested for regression by Härdle and Marron (1991)

Our findings have implications for other adaptive approaches, such as wavelet density smoothing. Appropriate smoothing of such estimates is determined by applying a thresholding function to the wavelet coefficients (Donoho, et al., 1993). However, since such wavelet smoothing is local in nature, its quality should be more like the balloon algorithm (2) than algorithm (4), which optimally requires non-local smoothing in the tails. Extending our non-local algorithm to wavelet thresholding would be an interesting project.

Appendix: (Proof of Theorem)

The proof proceeds by examining the terms in MISE $= E$[ISE] and E[UCV] and verifying that they are equal term by term. Note that the first and last terms of the two criterion, $R(\tilde{f})$ and $R(f)$, are identical. The second term, however, must be examined in detail. Recall that the bin counts $\{n_j\}$ are multinomial with parameters $\{p_j\}$; then

$$E[\int \tilde{f}(x)f(x)\,dx] = \sum_{j=1}^{m} p_j \int \phi_{h_j}(x-t_j)f(x)\,dx. \tag{11}$$

The expectation of the cross-product term of UCV in (10) is given by

$$E[\frac{1}{n}\sum_{i=1}^{n}\tilde{f}_{-i}(x_i)] = E[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}^{*}\phi_{h_j}(x_i-t_j)]. \tag{12}$$

The expectation (12) can be found by noting (i) the n_{ij}^* are multinomial with parameters p_j, (ii) $E[n_{ij}^*] = (n-1)p_j$, and (iii) the n_{ij}^* are independent of x_i because the bin counts, n_{ij}^*, for a given i do not include the ith observation. The expectation (12) then reduces to

$$E[\frac{1}{n}\sum_{i=1}^{n}\tilde{f}_{-i}(x_i)] = \sum_{j=1}^{m} p_j \int \phi_{h_j}(x-t_j)f(x)\,dx, \tag{13}$$

which equals (11). Thus the cross-product terms also agree, which completes the proof. □

References

Abramson, I. (1982), "On Bandwidth Variation in Kernel Estimates -A Square Root Law," *The Annals of Statistics*, **10**, 1217-1223.

Bartlett, M.S. (1963), "Statistical Estimation of Density Functions," *Sankhya Series A*, *25*, 245-254.

Bowman, A.W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, **71**, 353-360.

Breiman, L., Meisel, W., and Purcell, E. (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, **19**, 135-144.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1993), "Density Estimation by Wavelet Thresholding," Technical Report, Department of Statistics, Stanford University.

Fan, J., Hall, P., Martin, M., and Patil, P. (1993), "On Local Smoothing of Nonparametric Curve Estimators," Technical Report, Department of Statistics, Stanford University.

Hall, P. (1982), "The Influence of Rounding Errors on Some Nonparametric Estimators of a Density and its Derivatives," *SIAM Journal of Applied Mathematics*, **42**, 390-399.

Hall, P., Hu, T.C., and Marron, J.S. (1994), "Improved Variable Window Kernel Estimates of Probability Densities," *The Annals of Statistics*, in press.

Hall, P. and Wand, M.P. (1994), "On the Accuracy of Binned Kernel Density Estimators," unpublished manuscript.

Härdle, W. and Marron, J.S. (1991), "Fast and Simple Scatterplot smoothing," *Discussion Paper 9143*, C.O.R.E., Voic du Roman Pays 34, B-1348 Louvain-la-Neuve.

Jones, M.C., McKay, I.J., and Hu, T.C. (1994), "Variable Location and Scale Density Estimation," *Annals of the Institute of Statistical Mathematics*, **46**, in press.

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

Jones, M.C. (1990), "Variable Kernel Density Estimates," *Australian Journal of Statistics*, **32**, 361-371.

Kogure, A. (1987), "Asymptotically Optimal Cells for a Histogram," *Annals of Statistics*, **15**, 1023-1030.

Kooperberg, C. and Stone, C.J. (1991), "A Study of Logspline Density Estimation," *Computational Statistics and Data Analysis*, **12**, 327-347.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *The Annals of Mathematical Statistics*, **36**, 1049-1051.

Marron, J.S. and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, **20**, 712-736.

McKay, I.J. (1993), "A Note on the Bias Reduction in Variable Kernel Density Estimates," *Canadian Journal of Statistics*, **21**, 367-375.

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

Parzen, E. (1979), "Nonparametric Statistical Data Modeling (with discussion)," *Journal of the American Statistical Association*, **74**, 105-131.

Priebe, C.E. (1994), "Adaptive Mixtures," *Journal of the American Statistical Association*, **89**, 796-806.

Rosenblatt, M. (1956), "Remarks On Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, **27**, 832-837.

Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimators," *Scandinavian Journal of Statistics*, **9**, 65-78.

Sain, S.R. (1994), "Adaptive Kernel Density Estimation," Unpublished dissertation, Department of Statistics, Rice University.

Sain, S.R. and Scott, D.W. (1995), "Zero-Bias Bandwidths for Locally Adaptive Kernel Density Estimators," unpublished manuscript.

Schucany, W.R. and Sommers, J.P. (1977), "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, **72**, 420-423.

Scott, D.W. (1981), "Using Computer-binned Data for Density Estimation," *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, W.F. Eddy, Ed., Springer-Verlag, New York, pp. 292-294.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

Scott, D.W. and Sheather, S.J. (1985), "Kernel Density Estimation with Binned Data", *Communications in Statistics - Theory and Methods*, **14**, 1353-1359.

Silverman, B.W. (1982), "Kernel Density Estimation using the Fast Fourier Transform," *Applied Statistics*, **31**, 93-97.

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

Simonoff, J.S. and Hurvich, C.M. (1993), "A Study of the Effectiveness of Simple Density Estimation Methods," *Computational Statistics*, **8**, 259-278.

Terrell, G.R. and Scott, D.W. (1980), "On Improving Convergence Rates for Nonnegative Kernel Density Estimators," *Annals of Statistics*, **8**, 1160-1163.

Terrell, G.R. and Scott, D.W. (1992), "Variable Kernel Density Estimation," *The Annals of Statistics*, **20**, 1236-1265.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Weisberg, S. (1985), *Applied Linear Regression*, New York: John Wiley.