SiZer for Exploration of Structures in Curves

Probal CHAUDHURI and J. S. MARRON

In the use of smoothing methods in data analysis, an important question is which observed features are "really there," as opposed to being spurious sampling artifacts. An approach is described based on scale-space ideas originally developed in the computer vision literature. Assessment of SIgnificant ZERo crossings of derivatives results in the SiZer map, a graphical device for display of significance of features with respect to both location and scale. Here "scale" means "level of resolution"; that is, "bandwidth."

KEY WORDS: Confidence bands; Curve estimation; Kernel estimates; Local polynomials; Nonparametric smoothing; Scale space; Significant features; SiZer map.

1. INTRODUCTION

Smoothing for curve estimation in statistics is a useful tool for discovering features in data. Some examples of this are shown in Figure 1. For many more such examples, see, for example, the monographs of Bowman and Azzalini (1997), Eubank (1988), Fan and Gijbels (1996), Green and Silverman (1994), Härdle (1990), Müller (1988), Scott (1992), Silverman (1986), Simonoff (1996), Wahba (1991), and Wand and Jones (1995).

Figure 1(a) is an example of density estimation, where the typical goal is to present a density f that reveals structure in univariate data $X_1, \ldots, X_n$. The kernel approach involves centering small pieces of probability mass (having a Gaussian shape here) at each data point, using the formula given in (1). As seen, the window width h controls the amount of smoothing. The data here are $n = 7{,}211$ family incomes (rescaled so that the mean is 1) for the year 1975, from the Family Expenditure Survey in the United Kingdom. (See Schmitz and Marron 1992 for a detailed discussion and analysis of these data.) Note that the midrange bandwidth, $h = .05$, shows two prominent modes—perhaps an indication of an economic class structure? However, these modes can be made to disappear simply by using the larger bandwidth $h = .2$. Also, many more modes, which are likely to be spurious sampling artifacts, can be made to appear by using the smaller bandwidth $h = .0125$. Which modes are "really there"? The detailed analysis of Schmitz and Marron (1992) reveals that the two important modes are (perhaps surprisingly) important features of this dataset. That analysis also reveals an interesting shift in the size of these modes over time.

Figure 1(b) is an example of scatterplot smoothing, also called nonparametric regression estimation, where bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are smoothed (e.g., by a moving average) to give a curve that can be viewed as an estimated conditional mean, $f(x) = E\,(Y|X = x)$. The smooths actually used here are local linear smooths, with Gaussian weights, explicitly defined in (2). These have some preferable properties, as summarized in, for example, the monographs of Fan and Gijbels (1996) and Wand and Jones (1995). Again the window width is crucial to the smooth, with $h = .3$ and 4.8 representing substantial undersmoothing and oversmoothing. The data, provided by T. Bralower of the University of North Carolina, reflect global climate millions of years ago, through ratios of strontium isotopes found in fossil shells. The ratios had .70 subtracted, and then were multiplied by 100, because all are very close to .70. The shells are dated by biostratigraphic methods (see Bralower, Fullager, Paull, Dwyer, and Leckie 1997), so the strontium ratio can be studied as a function of time. Both the scatterplots and the smooths have a relatively high ratio for fossils less than 105 million years old, have a substantial dip with a minimum for those near 115 million years old, and then perhaps an increase for those around 120 million years old. These features are shown nicely by the larger bandwidth $h = 4.8$. However, at the dip this bandwidth seems to be substantially oversmoothing, so there is a good chance that it could be smoothing away some features that are really there. The bandwidth $h = 1.2$ seems closer to a reasonable amount of smoothing; note that this suggests additional possible features, such as an increase from 92 to 95 million years ago, and perhaps a dip around 98 million years ago. But the significance of at least this last dip is quite suspect, because a look at the data shows that it appears to be based on only two isolated observations.

Both examples in Figure 1 illustrate a major hurdle in the practical use of smoothing methods: Which features observed in a smooth are really there? Data analysts familiar with smoothing methods are usually very good at answering this question (although even for them gray areas exist where quantification would be helpful), when they have the time for a careful trial and error approach. However, such analysts do not always have lots of time, and even worse, such skilled people are all too often just not available. In this article we propose a graphical device, the SiZer map, that has two important benefits. First, it speeds up the process of deciding "which features are really there" for the experienced analyst, while at the same time quantitatively resolving gray area problems. Second, it allows even inexperienced analysts to make inferences about which features are really there.

Probal Chaudhuri is Associate Professor, Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Calcutta 700035, India (E-mail: *probal@isical.ac.in*). J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *marron@stat.unc.edu*). This research has been partially supported by grants from the National Science Foundation and the Indian Statistical Institute. The authors are grateful to the reviewers for many helpful comments.

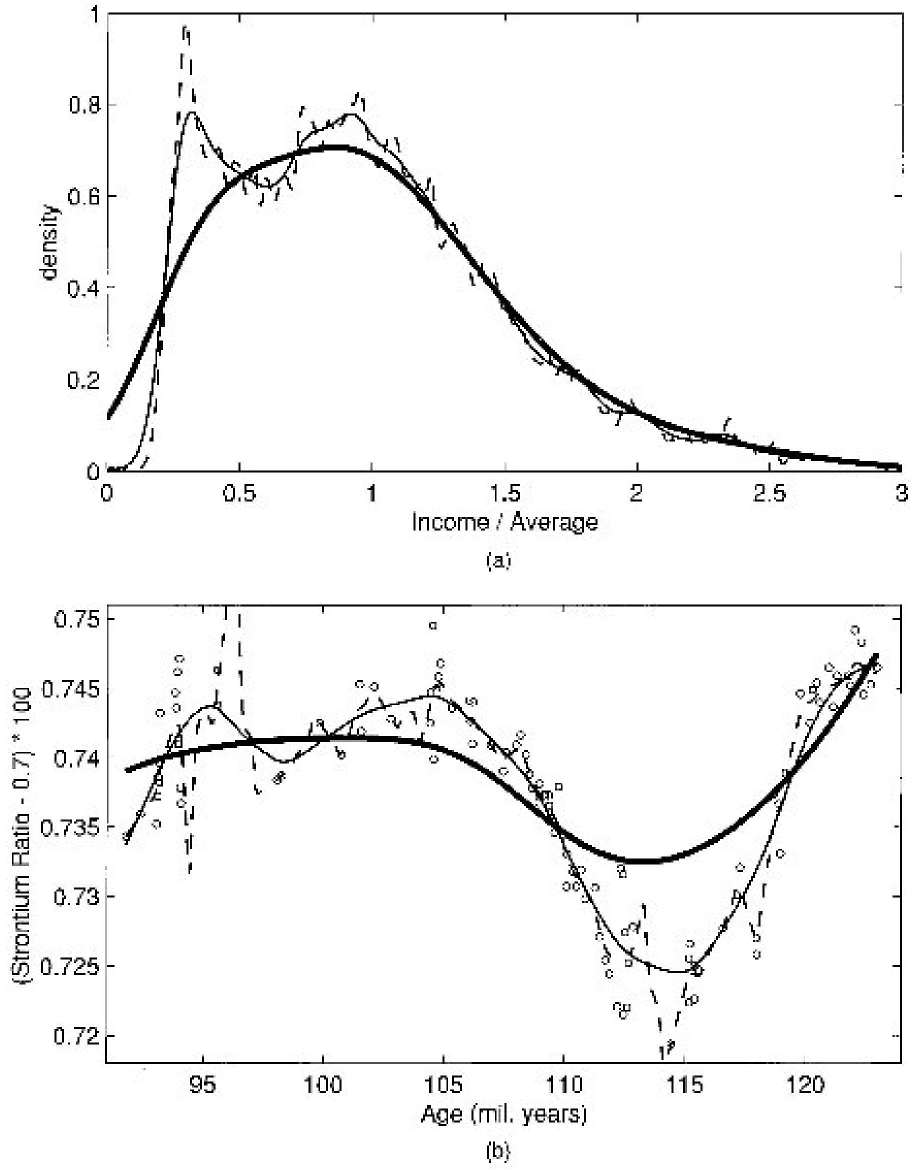


Figure 1. Examples of Features Revealed by Smoothing. (a) Kernel density estimates, with three different bandwidths h = .0125, .05, and .2, for the 1975 Income data. (b) A scatterplot and local linear regression estimates, with three different bandwidths h = .3, 1.2, and 4.8 for the Fossil data, with the raw data shown as small circles.

Our approach involves a view of smoothing and of the statistical inference problem at hand that is radically different from most of the literature. The traditional approach is to focus on a true underlying curve and do inference about that. In particular, much work has been done on choosing the bandwidth from the data, and many proposals have been made for inference based on confidence intervals/bands. For reasons discussed in detail in Section 6.2, such inference has not been very useful, especially for the problem of finding important features. The main problem is that curve estimators suffer inherently from a bias that is hard to deal with. This bias is not present in classical parametric statistics, where one operates under the assumption that a parametric model is "truth." We believe this is why attempts to extend the classical notion of parametric confidence intervals to smoothing seem to have not yielded the same useful results.

Our methodology is motivated by "scale-space" ideas from computer vision. (See Lindeberg 1994 for an introduction and detailed discussion). Our approach departs from the classical in two ways. First, we simultaneously study a very wide range of bandwidths, avoiding the classical need to choose a bandwidth. This idea is not foreign to good data analysts, who know well that different useful information can be available at different levels of smoothing. The *family approach* of Marron and Chung (1997) is one way of tapping into this information, but does not address the key question of which features are really present.

Our second departure from the classical view, again following scale-space ideas from computer vision, is that we avoid the bias problem in doing inference by shifting the focus from the true underlying curve to the true curve, viewed at varying levels of resolution. In particular, our inference focuses on smoothed versions of the underlying curve, with the idea that this contains all the information available in the data when working with that bandwidth. Detailed discussion of this view of smoothing is given in Section 2.

In Section 3 our main inferential tool, the SiZer map, is developed. This studies features simultaneously over both location and scale (i.e., bandwidth) by using a color map, as shown in Figure 2. The idea is to highlight significant features, such as bumps, by displaying where (with respect to both location and scale) the curve significantly increases and decreases. Note that significant bumps will be at *zero crossings of the derivative* between regions of significant increase and decrease. The name "SiZer" is a shortening of "SIgnificant ZERo crossings of derivatives." The color scheme is blue (red) in locations where the curve is significantly increasing (decreasing), and the intermediate color of purple is used where the curve cannot be concluded to be either decreasing or increasing. Here the term "location" is used in the scale-space sense of both "x-location" and "bandwidth location." Gray is used to indicate regions where the data are too sparse to make statements about significance, because there are not enough points in each window, as defined precisely in Section 3.

Note that for both sets of data, the family approach reveals potential interesting structure, in addition to lots of likely spurious structure. Perhaps the worst spurious structure is in the fossil data, where the smallest bandwidth smooth actually leaves the range of the data around 95 and 97 million years ago. This is caused by data sparsity in that region and is an unappealing feature of the local linear smoother. (See Hall and Marron 1997 for detailed discussion, and access to the literature on various fixes that have been proposed.) The SiZer maps for each make it clear which structure seen in the family plots is "statistically significant" and which one cannot be separated from the natural variability.

For the income data [Fig. 2(c)], at very coarse levels of resolution (i.e., large bandwidths), the smooths are significantly increasing (shaded blue) and then significantly decreasing (shaded red), meaning that these features are "really there" *at this level of resolution.* For bandwidths near $\log_{10}(h) = -1.3$, the two modes become apparent and are both statistically significant, because the shading changes from blue (↑) to red (↓) to blue (↑) to red (↓). Hence SiZer gives an answer consistent with the results obtained from other independent analysis discussed earlier. SiZer further suggests that the other features that can be seen in the family of smooths [including the three small bumps near the broader peak in the smooth with the thickest width in Fig. 2(a)] are just sampling artifacts, because the color is purple in these regions. The gray areas in each lower corner are where the data are too sparse for SiZer to be effective, as described in detail in Section 3.

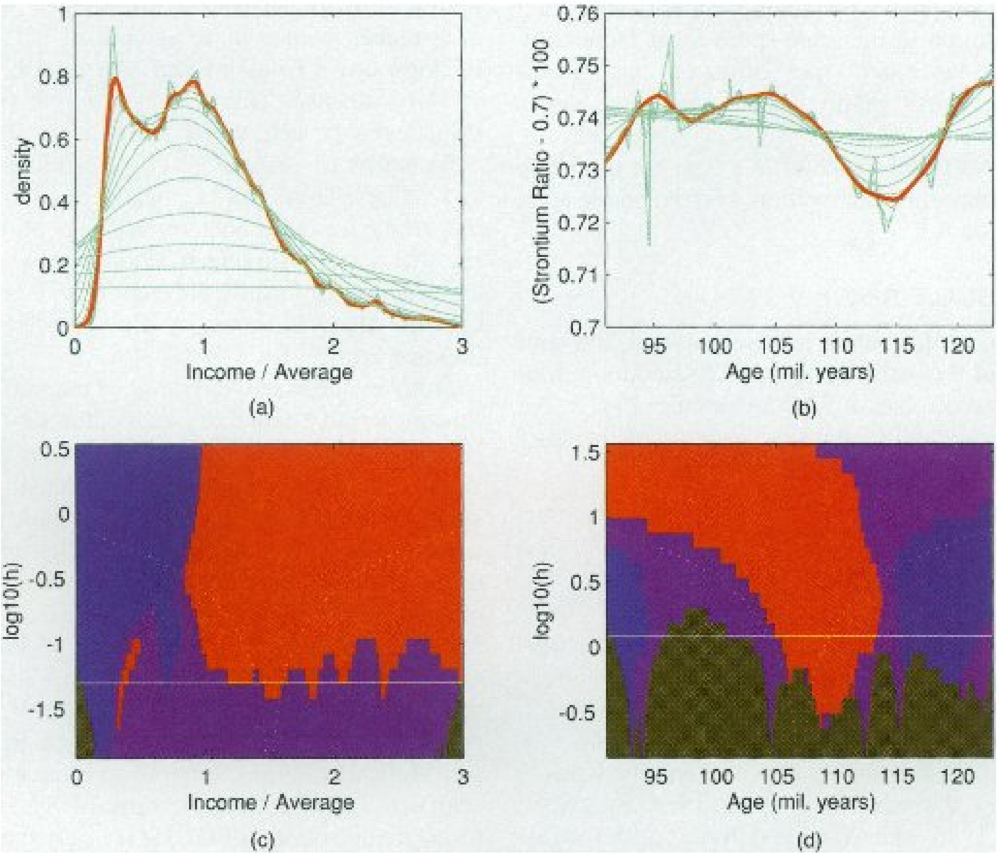


Figure 2. Combination of Family Plots [(a) and (b)] and SiZer Maps [(c) and (d)] for the Datasets in Figure 1, Using Level of Significance α = .05. (a) and (c) The income data; the important bandwidth h = .05 is highlighted in both plots. (b) and (d) The fossil data; again the important bandwidth h = 1.2 is highlighted in both plots. The dotted curves in the SiZer maps show effective window widths for each bandwidth, as intervals representing ±2h (i.e., ±2 standard deviations of the Gaussian kernel).

For the fossil data [Fig. 2(d)], at the coarsest levels of resolution (largest bandwidths) the smooth is not far from a simple least squares fit line (because the window is extremely large), although not the same, because SiZer shows significant decrease up to around 105 million years ago and then no significant change. For bandwidths that are less grossly oversmoothed, such as the bandwidth $h = 4.8$ [note that $\log_{10}(4.8) = .68$)] shown in Figure 1(b), the estimate has no significant slope on the left, significantly decreases in the center, and significantly increases on the right. When one looks at finer levels of resolution (smaller bandwidths), the curve is seen to be significantly increasing at around 93 million years ago. However, the dip in the thick curve of Figure 2(b), at about 97 million years ago, is shown to be spurious, because this feature is in the gray area, where there is not enough data to conclude that this dip is really there.

These examples demonstrate the great potential of SiZer as a tool for data analysis. More examples to this effect that also illustrate potential pitfalls are given in Section 4. David Scott has pointed out that the SiZer map can be viewed as an "enhancement" of the mode tree of Minnotte and Scott (1993); Bowman and Azzalini (1997) have given some related ideas.

Our main ideas can easily be adapted to many different types of smoothing methods, such as smoothing splines, regression splines, or wavelets. But in this article we concentrate on kernel–local polynomial smoothers, because of their simplicity and interpretability and their very direct connection to the scale-space ideas from computer vision. Various other types of extensions of this methodology are worth pointing out, which we do in Section 5.

Other approaches to inference of this type are discussed in Section 6. An important competitor is formal mode tests, reviewed in Section 6.1.

2. SCALE-SPACE VIEWPOINT

In this section we introduce precise notation and give some discussion of the scale-space view of smoothing. Kernel density estimation uses a random sample $X_1, \ldots, X_n$ from a smooth probability density $f(x)$, to estimate f through

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \qquad (1)$$

where h is the bandwidth (i.e., smoothing parameter) and K_h is the "h-rescaling" of the kernel function K, $K_h(\cdot) = 1/hK(\frac{\cdot}{h})$. The main idea is to put probability mass $\approx 1/n$ near each X_i. As shown in Figure 1(a), the bandwidth controls the amount of smoothing; $\hat{f}_h(x)$ is wiggly when h is small, and very flat when h is large. (See, e.g., Scott 1992, Silverman 1986, and Wand and Jones 1995 for discussion of many important properties and aspects of this estimator.)

The local linear regression estimate uses a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ to estimate the conditional expected value; that is, the regression function,

$$f(x) = E(Y_i | X_i = x),$$

through

$$\hat{f}_h(x) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} [Y_i - (a + b(X_i - x))]^2 \times K_h(x - X_i), \qquad (2)$$

where argmin_a is interpreted to mean minimize jointly over a and b, but use the a value. The main idea is that for each x, a line is fitted to the data, using K_h-weighted least squares. Again the bandwidth controls the amount of smoothness of $\hat{f}_h(x)$, as shown in Figure 1(b). (See, e.g., Fan and Gijbels 1995 and Wand and Jones 1996 for discussion of many properties and important aspects of this estimator.)

Scale-space ideas from computer vision provide a viewpoint on kernel smoothing that is new to statisticians. The "scale space surface," the family of all kernel smooths indexed by the bandwidth h, is a model used in computer vision. The essential idea is that large h models macroscopic (distant) vision where only large-scale features can be resolved, and small h models microscopic (zoomed in) resolution of small-scale features. In particular, for a given function f (i.e., underlying signal), various amounts of signal blurring (at least some is present in any real visual system) are represented by the convolution $f * K_h$ for different values of h. In fact, this family of convolutions becomes the focus of the analysis, with the idea that this is all that is available from a finite amount of data in the presence of noise (see Chaudhuri and Marron 1997 and Lindeberg 1994 for details). This is very different from the classical statistical approach, where the focus is f.

Examples of features in curve estimation include peaks and valleys. These can be characterized in several ways. In this article we focus on zero crossings of the derivative. We say that a zero crossing is significant when the derivative estimate is significantly different from 0 on both sides, with opposite signs as shown by blue and red areas in Figures 2(c) and 2(d).

Studying these zero crossings of the smooth derivative estimates across a range of bandwidths shows that the Gaussian kernel $K(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ has an important advantage over other kernels. In particular, for convolution smoothers, the number of zero crossings of the derivative smooth is always a decreasing function of h (which is not true for any other kernel used for kernel smoothing). In other words, only Gaussian blurring has monotonicity of features with respect to the amount of smoothing. Several ways to see this have been given by Chaudhuri and Marron (1997, sec. 2) and Lindeberg (1994), based on "total positivity" (see also Karlin 1968 and Brown, Johnstone, and McGibbon 1981). Interesting related references in the statistical literature include work of Silverman (1981) and Minnotte and Scott (1993). Hence only the Gaussian kernel is used in this article.

The main point of this article is the development of color maps as shown in Figures 2(c) and 2(d), called SiZer maps. These maps, which can be used for exploratory data analysis, show regions in scale space (i.e., with respect to both x and h) where the derivative is significantly increasing and decreasing. As discussed in Section 6.2, classical approaches to significance of features based on confidence bands are either much too conservative for useful inference, or are grossly invalid because of bias problems. In this article we take a novel approach to this old bias problem by adopting the scale-space point of view. In particular, instead of seeking confidence intervals for $f'(x)$, we seek confidence intervals for the scale-space version $f'_h(x) \triangleq E\hat{f}'_h(x)$. (For regression, we take this E to be conditional on $X_1, \ldots, X_n$.) The center point of such intervals is automatically correct, and the variance is estimated simply and effectively, as detailed later. From this point of view, significance of any feature depends on the scale of resolution (i.e., on h) and must be interpreted in that way. Figure 2(c) shows that the bimodal structure is present at some levels of resolution but disappears at coarser levels (i.e., there is only one mode at large bandwidths).

Note that this approach is rather different from traditional mode testing. In particular, the SiZer map not only counts the number of significant modes at different levels of resolution, but also gives information about mode locations. There is a trade-off, however, in that the SiZer map tends to be more conservative than mode tests that specifically target the number of modes; see Section 4.

3. DEVELOPMENT OF SIZER

Our approach to the visual assessment of the significance of features such as peaks and valleys in a family of smooths $\{\hat{f}_h(x) \colon h \in [h_{\min}, h_{\max}]\}$ is based on confidence limits for the derivative in scale space, $f'_h(x)$. (The choice of $h_{\min}$ and $h_{\max}$ is discussed in Sec. 3.1.) Behavior at x and h locations is presented via the SiZer color map, where blue (black in versions where only black and white are available) indicates locations where $\hat{f}'_h(x)$ is significantly positive, red (white in black and white versions) shows where $\hat{f}'_h(x)$ is significantly negative, and purple (gray in black and white versions) indicates where $\hat{f}'_h(x)$ is not significantly different from 0.

Because repeated calculation of smoothers is required for such color maps, fast computational methods are very important. Binned (also called "WARPed") methods are natural for this, because the data need be binned only once. (See Fan and Marron 1994 for detailed discussion of this and other fast computation methods.) The main idea is that calculation of $\hat{f}'_h(x)$ becomes a rapidly computed discrete convolution when the data are approximated by bin counts on an equally spaced grid, which can result in speed savings of factors of 100 (for larger sample sizes). For the reasons discussed by Fan and Marron (1994), we use $g = 401$ grid points for most examples in this article, although in some situations other values can be desirable, as discussed later.

Confidence limits for $f'_h(x)$ are of the form

$$\hat{f}'_h(x) \pm q \cdot \widehat{\mathrm{SD}}(\hat{f}'_h(x)), \tag{3}$$

where q is an appropriate quantile, and the standard deviation is estimated as discussed in Section 3.1. An (x, h) location (in scale space) is called significantly increasing, decreasing, or not significant when 0 is below, above, or within these confidence limits.

Candidates for calculation of the quantile q include:

- pointwise Gaussian quantiles: $q_1(h) = q_1 = \Phi^{-1}[1 - (\alpha/2)]$
- approximate simultaneous over x Gaussian quantiles: based on "number of independent blocks," defined as q_2 later
- bootstrap simultaneous over x, defined as q_3 later
- bootstrap simultaneous over x and h, defined as q_4 later.

Although Gaussian approximations work quite well (because smoothers are local averages), the pointwise quantiles q_1 are not recommended. This is because this version of the SiZer map suggests that too many features are "significant," as shown in Figure 3.

Each panel in Figure 3 is for the same simulated dataset of size $n = 100$ from the density #3 of Marron and Wand (1992). This density, shown as the heavy yellow curve in Figure 3(a), is a mixture of eight normals, intended to reflect much of the structure present in the lognormal distribution: a single large peak, with a very long right tail. As shown in the family of smooths, based on the single dataset in Figure 3(a), this density is challenging to estimate. In particular, small window widths are most appropriate near the peak to avoid smoothing that down to too low a level, but large bandwidths are more sensible in the tail to smooth out the spurious clusters that arise just by chance. The pointwise SiZer map, shown in Figure 3(b), incorrectly indicates that some of these spurious clusters are "significant"; for example, the peaks near $x = -1.7, -1.4$, and $.6$. The problem is understood via the classical frequentist interpretation of confidence intervals; looking at many replications should result in roughly proportion α intervals that do not cover the true value. A natural solution to the problem is to adjust the length of the intervals to do simultaneous inference, which is the goal of the other approaches to q mentioned earlier, which are discussed in detail in the next section. The approximate simultaneous approach is shown in Figure 3(c), where these spurious modes are now shaded correctly as purple. However, this version has a curious red stripe in the lower right corner that we do not fully understand. We have not carefully analyzed this, because it is in a region in scale space where the data are very sparse. Both because of effects like this and because we do not trust confidence intervals based on too few points, regions in scale space where the data are too sparse for meaningful inference are grayed out. Based on the classical rule of thumb, a location gray is shaded gray when the effective sample size in the window (defined later) is less than 5. This gives the map shown in Figure 3(d).

Our first suggestion, q_2, for approximate simultaneous confidence limits is based on the fact that when x and x' are sufficiently far apart, so that the kernel windows centered at x and x' are essentially disjoint, the estimates $\hat{f}'_h(x)$ and problems with small **ESS** as

$$m(h) = \frac{n}{\operatorname{avg}_{x \in D_h} \mathrm{ESS}(x, h)},$$

where D_h is the set of x locations where the data are dense,

$$D_h = \{x : \mathbf{ESS}(x, h) \geq n_0\}.$$

These approximate simultaneous confidence limits are somewhat crude and also are only simultaneous over x, not h. To improve them, we explored several classical multivariate normal simultaneous confidence sets (both elliptical and rectangular). These are based on the standard principal component analysis. Unfortunately, they tended to be far too conservative, because the orientation of the usual confidence sets along the eigenvector directions gave a region that did not efficiently project back to confidence intervals for $\hat{f}'_h(x)$ for each x. The projections (i.e., the resulting confidence intervals) tended to be far too long to allow us to find important features.

Simultaneous confidence sets that are hypercubes, whose edges are parallel to the axes with lengths of the form (3), are much better oriented to reflect the significance of our derivative estimates. Direct calculation of the probabilities of such rectangular sets in high dimensions is very difficult for these highly correlated normal distributions. Because simulation is the only tractable approach, it is natural to use the more direct method of the bootstrap (i.e., simulate from the empirical distribution of the data, instead of from the approximating Gaussian). For each bootstrap sample (i.e., random sample drawn with replacement from the data; see Efron and Tibshirani 1993 for an introduction to bootstrap ideas), we compute $\hat{f}'_h(x)^*$ (again a fast implementation is crucial) and the standardized version

$$Z^*(x, h) = \frac{\hat{f}'_h(x)^* - \hat{f}'_h(x)}{\widehat{\mathrm{SD}}(\hat{f}'_h(x))}.$$

For each h, the bootstrap quantile $q_3 = q_3(h)$ that is simultaneous over x (where the data are reasonably dense) is the empirical quantile of $\max_{x \in D_h} |Z^*(x, h)|$ calculated over the bootstrap replications. Similarly, the bootstrap quantile q_4 that is simultaneous over both x and h is the empirical quantile of $\max_h \max_{x \in D_h} |Z^*(x, h)|$ taken over the bootstrap replications.

Study of many SiZer maps based on q_2, q_3, and q_4 showed that in many cases there was not much difference between the quick and approximate quantile q_2 and the bootstrap quantile q_3 that is simultaneous over x. As expected, somewhat fewer features generally appeared as significant for q_4, the bootstrap value that is simultaneous over both x and h, although surprisingly often q_4 was quite similar to q_2 and q_3. The maps based on different choices for q were most similar for examples that were homogeneous in x, meaning either equally spaced regression or density estimation examples where the local average height of the density is roughly constant. This is because there is an implicit homogeneity assumption made by q_2 that is a reasonable approximation in this case.

An example where this homogeneity is lacking (thus giving interesting differences) is the income dataset, from Figures 2(a) and 2(c). SiZer maps based on the bootstrap quantiles q_3 and q_4 are shown in Figure 4.

The SiZer map for q_3 with $\alpha = .05$, shown in Figure 4(a), is fairly similar to that for q_2 shown in Figure 2(c), except the lower right red region above $x = .4$ is quite a bit thinner. That red region actually disappears for the fully simultaneous SiZer map based on q_4 with $\alpha = .05$ shown in Figure 4(c). This shows that the q_4 SiZer map can be rather conservative, because it does not show that there are two significant modes here (at the level $\alpha = .05$), although these have been verified by other means. But when the level of significance is raised to $\alpha = .10$, as shown in Figure 4(d), the red region reappears, so both modes are now statistically significant in this sense. Note that for both q_3 and q_4, as α increases, the red and blue regions grow, as expected.

Because the bootstrap versions of SiZer are much slower to compute, we suggest using q_2 for a first look at the data. This version of SiZer is called SiZer1 in our software, available at the URL *http://www.stat.unc.edu/faculty/marron/marron_software.html*. But when there are any doubts (there should be more doubts in settings that are not homogeneous in x), we recommend using q_3 and q_4 (implemented as SiZer5 in our software) for verification. Although q_4 is our only procedure that gives a rigorous test of significance of features, it is also generally somewhat conservative, so we recommend that features found in q_2 or q_3 SiZer maps that do not appear in the q_4 version be independently investigated by a mode testing method, as explained in section 6.1. For example, the mode test of Fisher and Marron (1998) shows that the existence of two modes in the income data can be established with $\alpha < .01$ by a test that focuses explicitly on number of modes.

Bootstrap theory suggests improvement by a studentized modification (see, e.g., Hall 1992) or other methods (see, e.g., Efron and Tibshirani 1993). Such methods have not been implemented here, because they involve recalculation of the variance estimate for each bootstrap sample, which would entail substantial computational cost.

3.1 Numerical Implementation

The bandwidth range $[h_{\min}, h_{\max}]$ can be chosen in several ways. One approach is a broad range of smooths which should catch most interesting features, as developed in the family approach to smoothing of Marron and Chung (1997). Another approach is "a very wide range of smooths," which is determined more by the curve estimation setting than by the data. In the examples of this article, we have used the latter, and we took $h_{\min}$ to be the smallest bandwidth for which there is no substantial distortion in construction of the binned implementation of the smoother, $h_{\min} = 2 * (\text{binwidth})$, and took $h_{\max}$ to be the range of the data.

3.1.1 Density Estimation Specifics. The main idea behind the calculation of $\widehat{\mathrm{SD}}$ in this context is that the derivative estimator $\hat{f}'_h(x)$ is an average (of the derivative kernel functions), so we use the corresponding sample standard

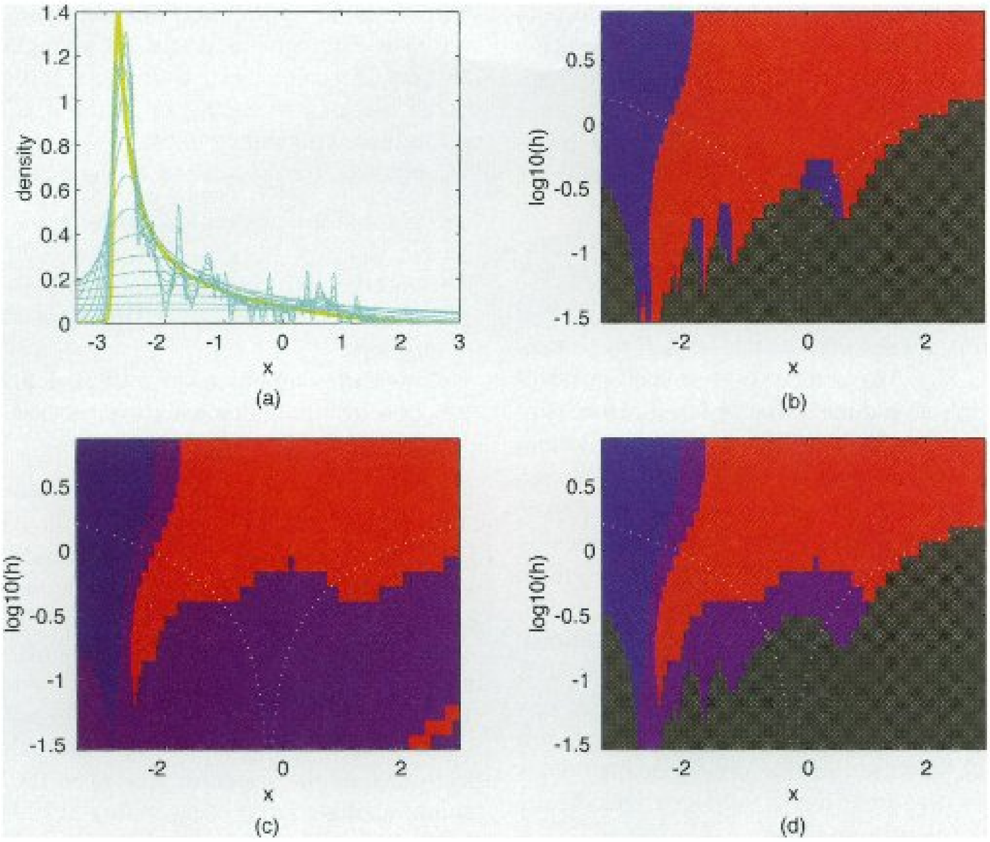


Figure 3. A Simulated Dataset, of Size n = 100, From the Marron–Wand Density #3. (a) The family approach, overlaid with the true underlying density (thick green curve); (b) pointwise SiZer; (c) simultaneous SiZer with no gray shading for sparse data regions; (d) approximate simultaneous SiZer with sparse data regions shown in gray.

$\hat{f}'_h(x')$ are essentially independent, but when x and x' are close together, the estimates are highly correlated. The simultaneous confidence limit problem is then approximated by m independent confidence interval problems, where m reflects the number of independent blocks. We estimate m through an estimated effective sample size (ESS), defined for each (x, h) as

$$\text{ESS}(x,h) = \frac{\sum_{i=1}^{n} K_h(x - X_i)}{K_h(0)}.$$

Note that when K is a uniform (i.e., boxcar) kernel, $\text{ESS}(x,h)$ is the number of data points in the kernel window centered at x. For other kernel shapes, points are downweighted according to the height of the kernel function, just as they are in the averages represented by the kernel estimators. Next, we choose m to be essentially the number of independent blocks of average size available from our dataset of size n,

$$m(h) = \frac{n}{\text{avg}_x \text{ESS}(x,h)}.$$

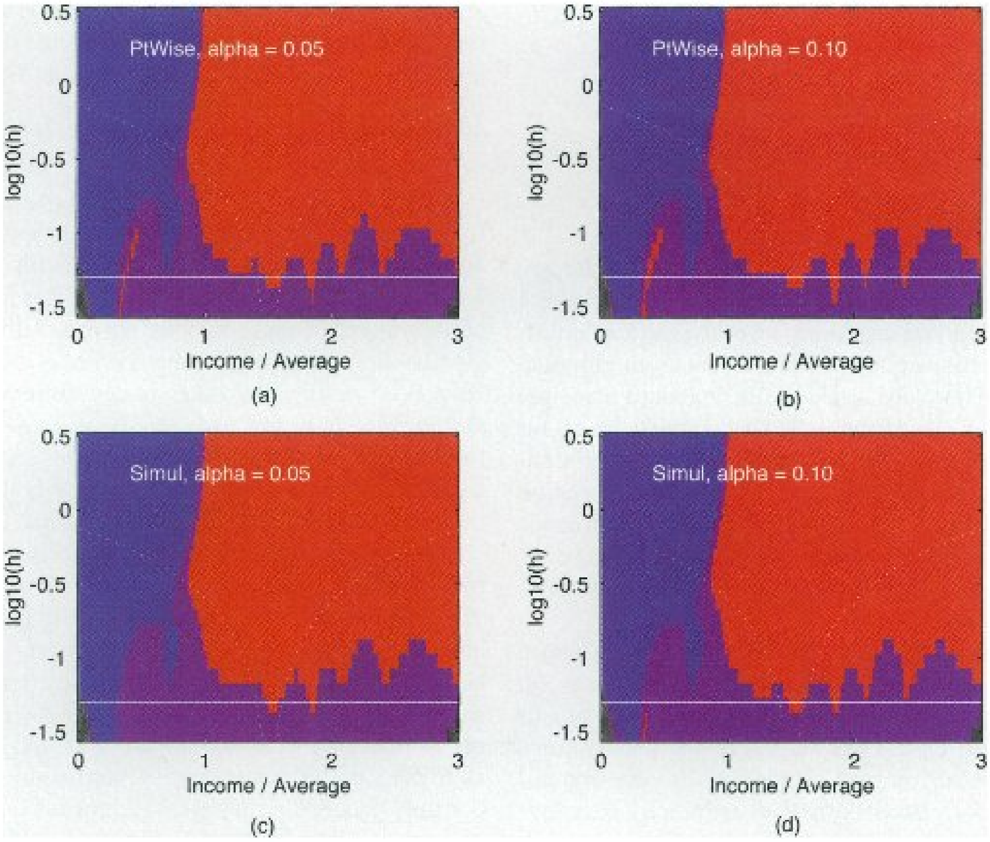
A reviewer pointed out an interesting connection between $m(h)$ and the concept of effective degrees of freedom of Hastie and Tibshirani (1990), which is the trace of the smoother matrix. Ignoring edge effects, this trace is

$$n \cdot \left(\frac{1}{n} K_h(0)\right) = \frac{n}{\left(\frac{n}{K_h(0)}\right)} \approx \frac{n}{\left(\frac{\text{avg}_x \sum_{i=1}^{n} K_h(x-X_i)}{K_h(0)}\right)} = m(h),$$

because $\text{avg}_x (1/n) \sum_{i=1}^{n} K_h(x - X_i) \approx 1$ (the area under a kernel density estimate). Now, assuming independence of these $m(h)$ blocks of data, the approximate simultaneous quantile is

$$q_2 = q_2(h) = \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right).$$

The quantity ESS is also useful to highlight regions where the normal approximation implicit in (3) could be inadequate. This plays a role similar to np in the Gaussian approximation to the binomial. So regions where $\text{ESS}(x,h) < n_0$ (we have followed the standard practice of $n_0 = 5$ at all points here) are shaded gray, to rule out spurious features and also to indicate regions where the smooth is essentially based on sparse data, as shown in Figure 3(d). The foregoing calculation of the block size $m(h)$ is modified to avoid



(a)

(b)

(c)

(d)

Figure 4. SiZer Maps for the Income Data, Based on 1,000 Bootstrap Replications. Quantiles are q_3 pointwise over h in the (a) and (b), and q_4 simultaneous over h in (c) and (d). Significance levels are α = .05 in (a) and (c), α = .10 in (b) and (d).

deviation,

$$\widehat{\mathrm{var}}(\hat{f}'_h(x)) = \widehat{\mathrm{var}}\left(n^{-1}\sum_{i=1}^{n} K'_h(x - X_i)\right)$$
$$= n^{-1}s^2(K'_h(x - X_1), \ldots, K'_h(x - X_n)),$$

where s^2 is the usual sample variance of n numbers.

Details of the binned implementation of $\hat{f}'_h(x)$ are similar to those given by Fan and Marron (1994), except that the kernel is now replaced by the derivative of the kernel. In particular, for the equally spaced grid of points $\{x_j\colon j = 1, \ldots, g\}$, let the corresponding bincounts (computed by, for example, the linear binning described in Fan and Marron 1994) be $\{c_j\colon j = 1, \ldots, g\}$. Then

$$\hat{f}'_h(x_j) \approx n^{-1}\tilde{S}'_0(x_j),$$

where

$$\tilde{S}'_0(x_j) = \sum_{j'=1}^{g} \kappa'_{j-j'} c_{j'} \tag{4}$$

and

$$\kappa'_{j-j'} = K'_h(x_j - x_{j'}). \tag{5}$$

To similarly approximate $\widehat{\mathrm{SD}}$, use

$$\widehat{\mathrm{SD}}(x_j) = n^{-1/2}\sqrt{n^{-1}\sum_{j'=1}^{g} (\kappa'_{j-j'})^2 c_{j'} - (\hat{f}'_h(x_j))^2}.$$

3.1.2 Regression Estimation Specifics. We prefer the local linear smoother to a number of other sensible smoothers, because the derivative estimate is the simple and appealing slope of the local line,

$$\hat{f}'_h(x) = \underset{b}{\mathrm{argmin}} \sum_{i=1}^{n} [Y_i - (a - b(X_i - x))]^2$$
$$\times K_h(x - X_i). \tag{6}$$

This is very similar to (2), except that the slope is kept instead of the intercept. This slope estimate is preferable to the quotient rule form of the derivative estimate, which has an unpleasant form. (See, e.g., Fan and Gijbels 1996 and Wand and Jones 1995 for further discussion, and Fan and Marron 1994 for a fast binned implementation of the local linear smoother.)

Our proposed $\widehat{\mathrm{SD}}$ is motivated by the fact that the derivative estimator is a weighted sum of the observed responses, and we essentially use the conditional (given $X_1, \ldots, X_n$) weighted sample variances,

$$\mathrm{var}(\hat{f}'_h(x)|X_1, \ldots, X_n)$$
$$= \mathrm{var}\left(n^{-1}\sum_{i=1}^{n} W_h(x, X_i)Y_i|X_1, \ldots, X_n\right)$$
$$= \sum_{i=1}^{n} \sigma^2(Y_i|X_i)(W_h(x, X_i))^2.$$

To estimate $\sigma^2(Y_i|X_i)$, we use a simple smooth of the residuals; for example,

$$\hat{\sigma}^2\ (Y|X=x) = \frac{\sum_{i=1}^n \hat{e}_i^2 K_h(x-X_i)}{\sum_{i=1}^n K_h(x-X_i)},$$

where $\hat{e}_i = Y_i - \hat{f}_h(X_i)$.

An efficient binned approximation of the local linear derivative estimate $\hat{f}_h'(x)$ (6) is

$$\hat{f}_h'(x_j) \approx \frac{\bar{T}_1(x_j) - T_0(x_j)\bar{X}(x_j)}{\bar{S}_2(x_j) - 2\bar{S}_1(x_j)\bar{X}(x_j) + \bar{S}_0(x_j)\bar{X}(x_j)^2}$$

$$= \frac{T_1(x_j) - \bar{T}_0(x_j)\bar{X}(x_j)}{\bar{S}_2(x_j) - \bar{S}_1(x_j)^2/\bar{S}_0(x_j)},$$

where the notations

$$\bar{S}_l(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} c_{j'} x_{j'}^l,$$

$$\bar{T}_l(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} Y_{j'}^{\Sigma} x_{j'}^l,$$

and

$$\bar{X}(x_j) = \bar{S}_1(x_j)/\bar{S}_0(x_j), \tag{7}$$

are used together with

$$\kappa_{j-j'} = K_h(x_j - x_{j'}) \tag{8}$$

and $Y_{j'}^{\Sigma}$ for the bin sums of the Y_i. Note that using the Gaussian kernel ensures that $\bar{S}_0(x_j)$ is theoretically nonzero. Rounding errors can create zero values of $\bar{S}_0(x_j)$, but SiZer is unaffected because this happens only in the gray regions where not enough data are present.

A binned approximation to $\hat{\sigma}^2\ (Y|X=x_j)$, based on calculations familiar from simple linear regression, is

$$\hat{\sigma}^2\ (Y|X=x_j) \approx (1-\hat{\rho}(x_j)^2)\hat{\sigma}(x_j)^2,$$

where

$$\hat{\sigma}(x_j)^2 = \frac{\bar{U}_0(x_j)}{\bar{S}_0(x_j)} - \left(\frac{\bar{T}_0(x_j)}{\bar{S}_0(x_j)}\right)^2$$

and

$$\hat{\rho}(x_j)^2 = (\hat{f}_h'(x_j))^2 \left(\frac{\bar{S}_0(x_j)\bar{S}_2(x_j) - \bar{S}_1(x_j)^2}{\hat{\sigma}(x_j)^2\bar{S}_0(x_j)^2}\right) \tag{9}$$

using the notation (7) and

$$\bar{U}_0(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} Y_{j'}^{2\Sigma}$$

for $Y_{j'}^{2\Sigma}$ denoting the bin sums of the Y_i^2. Our binned approximation to the conditional variance is now

$$\operatorname{var}(\hat{f}_h'(x_j)|X_1,\ldots,X_n) \approx \frac{\bar{V}_2(x_j) - 2\bar{V}_1(x_j)\bar{X}(x_j) + \bar{V}_0(x_j)\bar{X}(x_j)^2}{(\bar{S}_2(x_j) - \bar{S}_1(x_j)^2/\bar{S}_0(x_j))^2},$$

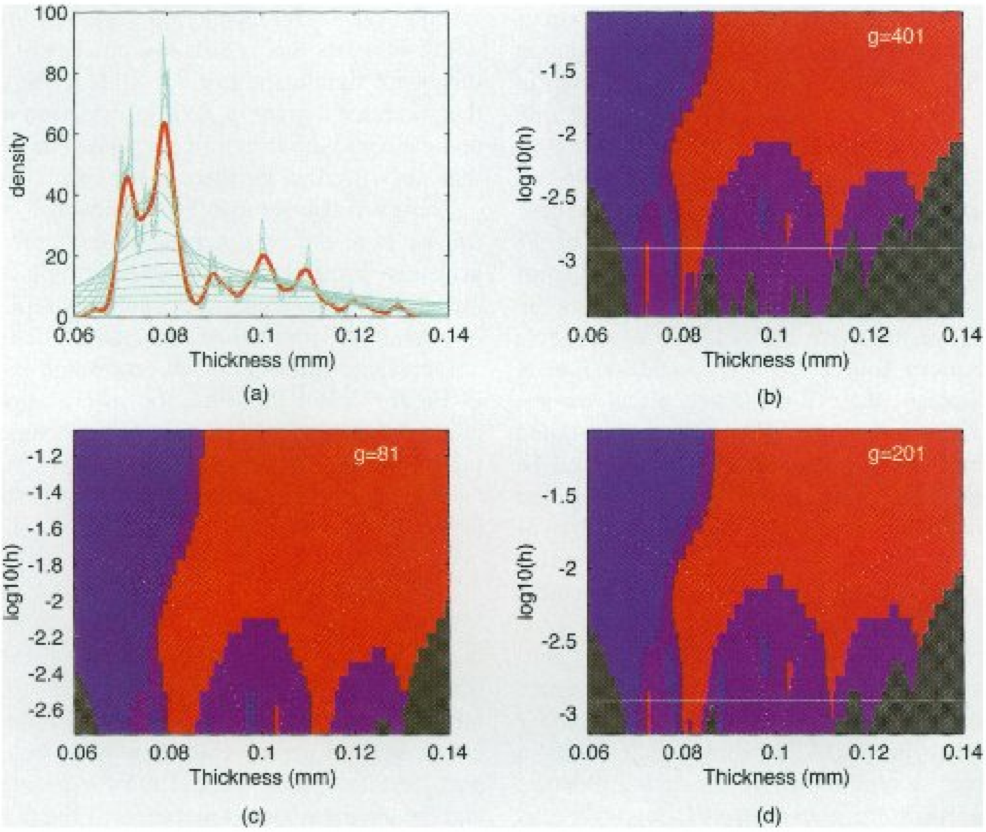


Figure 5. Family Plot (a) and SiZer Maps, Based on 401 Grid Points (b), 81 Grid Points (c), and 201 Grid Points (d), for the Hidalgo Stamp Data. The Sheather–Jones plug-in bandwidth is the thick curve in the family plot and corresponds to the highlighted horizontal bar. The SJPI bandwidth suggests seven modes, but not all are "significant" from the SiZer point of view.

where

$$V_l(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} c_{j'} x_{j'}^{l} \hat{\sigma}^2 \, (Y|X = x_{j'}),$$

and the notations (7), (8), and (9) have been used. This results in

$$\widehat{SD}(x_j) = \sqrt{\mathrm{var}(\hat{f}_h'(x_j)|X_1, \ldots, X_n)}.$$

4. MORE APPLICATIONS AND EXAMPLES

In this section, we present additional examples illustrating both the usefulness of SiZer and also some potential pitfalls. Refer to Marron and Chaudhuri 1998a, 1998b for several other interesting applications of SiZer as a powerful data analytic tool.

The Hidalgo stamp data set was brought to the mode-testing literature by Izenman and Sommer (1988). This is a univariate dataset comprised of the thicknesses of stamps issued in Mexico during the last century. These thicknesses have a remarkable amount of variability and clustering, which suggests a number of sources for the paper. An interesting philatelic question is to determine the number of paper sources, which was addressed by Izenman and Sommers (1988) via nonparametric density estimation. Figure 5 analyzes these data with the family approach and SiZer.

The thick curve in Figure 5(a) is a kernel density estimate using the Sheather–Jones plug-in bandwidth, as recommended by Jones, Marron, and Sheather (1996a,b). This suggests seven modes in the data (i.e., at least seven sources for the paper), which agrees with the findings of Izenman and Sommers (1988) and some others. The SiZer map in Figure 5(b) shows that the two largest modes, at .072 mm and .079 mm, are indeed significant, as is the mode at .1 mm. The mode at .09 mm is less certain, as SiZer finds a significant increase on the left but no significant decrease on the right. Similar results were found for the mode at .11 mm, where there is only a significant decrease. SiZer completely misses the modes at .12 mm and .13 mm, but the existence of these is perhaps debatable. If one has a priori knowledge that no paper source has a very wide variance, then one may be able to believe these are actual modes. However, if one accepts the possibility of a heavy-tailed distribution, then the family plot suggests these could be just random clustering in such a heavy tail. Also note the thick density estimate is heavily into the gray region of the SiZer map, which says the data are very sparse in this region, which also casts doubt on these modes, from this point of view.

Note that at the finest level of resolution (smallest bandwidth), the SiZer map in Figure 5(b) suggests the existence of more "modes" between .068 and .083. This is caused by the data being heavily rounded, to .001 mm, which results in many replicate values in regions where the data are dense. When such rounded data are binned to 401 bins over this range (i.e., a binwidth of .0002), there are a number of bins that receive no observations. When these bincounts, which alternate between 0 and very large numbers (because of the rounding) are smoothed with a very small bandwidth, one gets a kernel estimate that significantly increases and decreases, as shown. In this sense these feature are "really there," although the only conclusion is that the data have been rounded. We have seen this same phenomenon in other datasets. A natural solution is that in Figure 5(c), where the number of grid points is reduced to $g = 81$, which makes each rounded data value a bin center. Unfortunately, the heavy rounding in the data entails a SiZer map that misses some of the most interesting levels of smoothing, such as the Sheather–Jones plug-in bandwidth. One way to fix this would be to expand the range of bandwidths, but, as noted in Section 3.1, this entails using small bandwidths, which results in distorted density estimates. A better fix is shown in Figure 5(d), where $g = 201$ is used. Note that this SiZer map has all the same important features as in Figure 5(b).

Next we study the performance of SiZer in some simulation settings, which highlight how SiZer displays the information available in the data. The first of these is shown in Figure 6, where we study the effect of increasing sample size n (i.e., increasing information in the data) in density estimation.

The family plot, in combination with the Sheather–Jones plug-in bandwidth, for $n = 100$ suggests no significant modal structure in the data. This is also reflected in the SiZer map. There are just not enough data to resolve even the two larger modes present in the underlying density. For $n = 1{,}000$, the situation is different, and now the two large modes are clearly present in the data. More interesting is the third central mode. It is not clear from the family plot whether this is significant, the Sheather–Jones plug-in bandwidth suggests this is dubious, and the SiZer map confirms this is not significant. For $n = 10{,}000$, the family plot shows that we have a great deal of information about this density and can estimate it extremely well. The SiZer plot verifies this, showing that all three modes are clearly present.

There are also some interesting overall trends present that can be expected in general. For example, as n grows, the gray area diminishes and tends to be replaced by purple. The purple areas also tend to be eventually replaced by either red or blue, both from below and also in the boundary regions.

Increasing information in regression is also investigated in Figure 7, but this time the information in the data increases through decreasing the error variance, rather than increasing the sample size.

The underlying regression curve in each part of Figure 7 is

$$f(x) = x + 1.5 \times \varphi\left(\frac{(x - .35)}{.15}\right) - \varphi\left(\frac{(x - .8)}{.04}\right),$$

where φ is the standard Gaussian probability density function. For the very low noise case, $\sigma = .02$, the data contain a lot of information about the underlying regression curve, so the Ruppert–Sheather–Wand bandwidth (see Ruppert, Sheather, and Wand 1995 for a detailed description) and the undersmoothed members of the family are all essentially the same as the target curve. The SiZer map shows that all features of the target curve are significant, for a wide range of different resolutions (i.e., bandwidths). Even

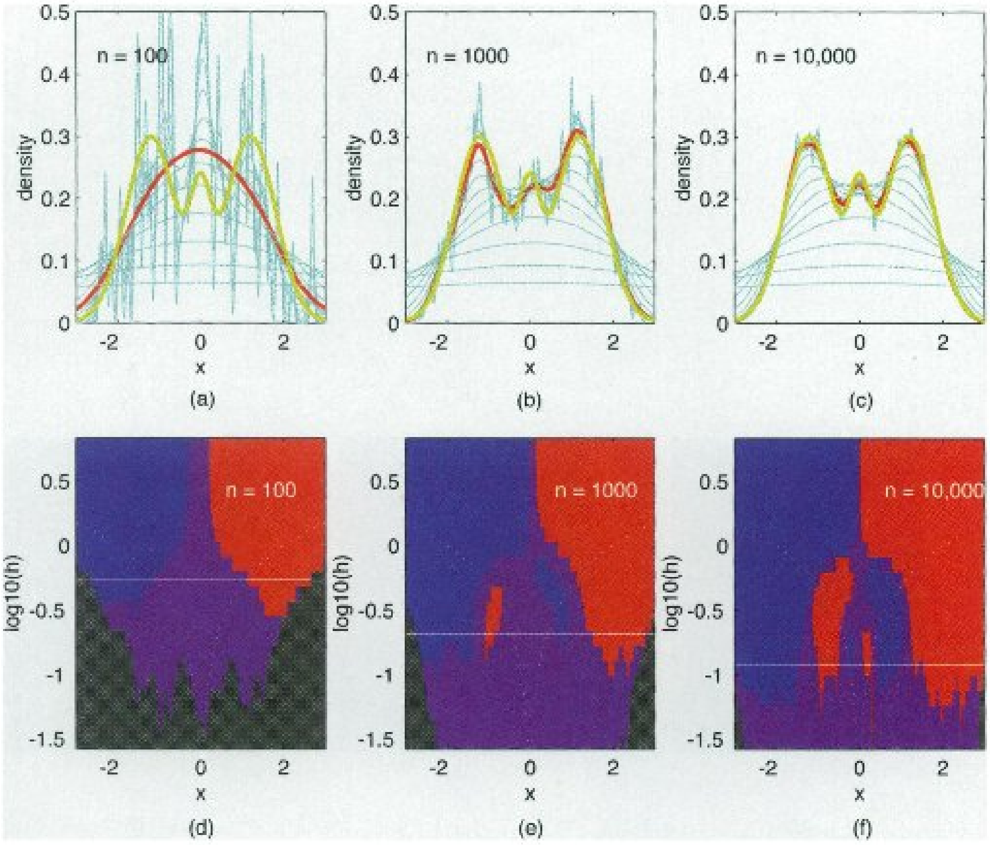


Figure 6. Family Plots [(a), (b), and (c)] and Corresponding SiZer Maps [(d), (e), and (f)] for Kernel Density Estimates, Based on Simulated Data, From the Marron and Wand Density #9, Trimodal, Shown as the Thick Green Curve in the Family Plots. Sample sizes are n = 100 in (a) and (d), n = 1,000 in (b) and (e), and n = 10,000 in (c) and (f). The thick red curve in the family plots is the Sheather–Jones plug-in bandwidth, which is the highlighted horizontal bar in the SiZer maps.

the "flat spot" near $x = .6$, which is not easy to find in the smooths, shows up as purple. When the noise level is increased substantially to $\sigma = .18$, the family plot shows that the estimation problem is now harder, and the SiZer map shows fewer significant features. However, the regions of increase are still significant, and two regions of decrease still appear, although at different levels of resolution. Increasing the noise still further, to $\sigma = .66$, results in a very challenging estimation problem, and now SiZer does not indicate any of the decreases and indicates only one of the two increases as being significant. The family plot shows that this is reasonable, because the noise level is so high. The Ruppert–Sheather–Wand bandwidth suggests a decrease (although it completely misses the small valley at $x = .8$), but it is not clear with this noise level that it is significant, and SiZer shows that it is not.

SiZer is also useful even in settings where the underlying target curve is not smooth, and in fact is quite useful at highlighting "jumps." This is shown in Figure 8, where Donoho and Johnstone's blocks function (famous from many papers on wavelets) is used as a regression target, but the added Gaussian noise is larger than is typical in wavelet examples. Note that the location of each jump is highlighted by a colored streak (blue for up and red for down) that reaches all the way to the bottom of the SiZer map. The streaks are caused by the fact that even at very small bandwidths, the estimates are changing significantly at these points. This phenomenon appeared in a number of other examples we have studied where the target curve has jump discontinuities. These indicated jumps could be used to construct a step function estimator with much better properties than the usual wavelet estimators for this example.

5. FUTURE RESEARCH DIRECTIONS

In this section we discuss a number of future research directions that are motivated by SiZer.

5.1 Local Likelihood

George Terrell has pointed out that for density estimation, and for special types of regression such as logistic regression, symmetric confidence intervals such as those proposed here can be improved upon using context-specific information. We suggest a local likelihood approach to this. Local likelihood is a smoothing method that is more efficient than simple kernel methods in some cases; for example, discrete response variables. (See Chaudhuri and Dewanji 1995, Fan, Heckman, and Wand 1995, Simonoff 1997, Staniswallis 1989, and Tibshirani and Hastie 1987 for detailed discussion and more references.) We anticipate that SiZer may be extended in a fairly straightforward way to this important smoothing context.

5.2 Handling Dependency

In nonparametric regression, our current SiZer develop-

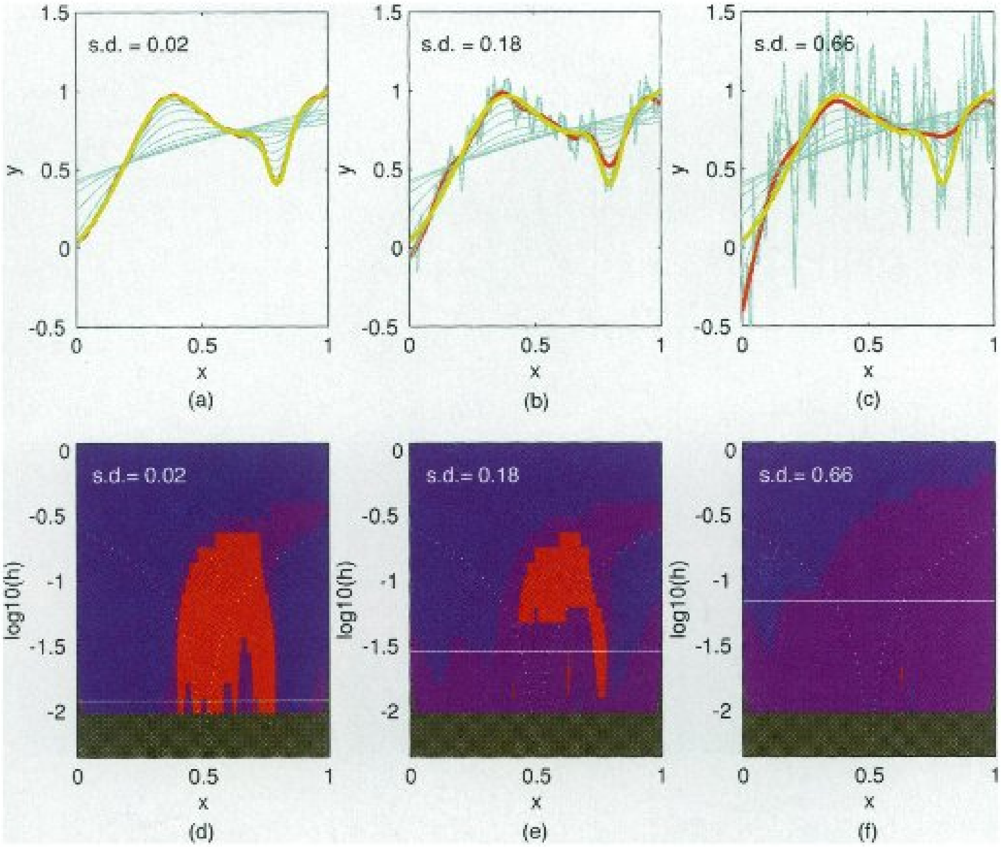


Figure 7. Family Plots [(a), (b), and (c)] and Corresponding SiZer Maps [(d), (e), and (f)] for Local Linear Regression Estimates, Based on n = 200 Simulated Data, Shown as Green Dots, From an Equally Spaced Design, and the Regression Curve Shown as the Thick Green Curve. Simulated errors are independent Gaussian, with standard deviations σ = .02 in (a) and (d), σ = .18 in (b) and (e), and σ = .66 in (c) and (f). The thick red curve is the Ruppert–Sheather–Wand direct plug-in bandwidth, which is highlighted in the corresponding SiZer maps.

ment assumes independent errors, which is not always realistic (e.g., in time series contexts). But SiZer has the potential to become an important tool in such contexts where "significance of trends" is often an important issue. We believe that such applications will require appropriate modeling of the error structure (e.g., by some autoregressive moving average or even long-range dependent models) before useful inference can be done.

5.3 Testing Other Types of Hypotheses

SiZer focuses on regions where the derivatives are significantly increasing and decreasing, but for some situations other aspects of the underlying curve, such as the second derivative, or even the curve itself could be more appropriate to study in this way. Variations of SiZer could also be used to address other problems, such as whether or not two curves are significantly different.

5.4 Other Estimation Settings

Smoothing is useful in other settings besides just density and regression estimation. For example, SiZer can be extended to estimation of the hazard function and other functions appearing in survival analysis. Another interesting extension would be to various censored-data contexts.

5.5 Local Bandwidth Selection

A separate potential application of SiZer is to the old field of location varying bandwidth selection. The need for this is demonstrated in Figure 9, where the family approach shows that one would prefer a smaller bandwidth on the right, where the underlying density has finer features, and a larger bandwidth on the left, where the density has less curvature. The Sheather–Jones plug-in bandwidth does a reasonable job with the fatter peaks, but could be much improved on the smaller peaks. In particular, SiZer shows that the smaller peaks really are significant, but only at a finer level of resolution (smaller bandwidth). However, although the need for it has been clearly understood, data-based local bandwidth selection has proven to be a very challenging problem. In particular, the simulation study of Farmen (1996) and Farmen and Marron (1997) shows that most of the available methods do not fare much better overall than the simple global bandwidth chosen by the Sheather–Jones plug-in method. A likely intuitive explanation for this is that local bandwidth selectors essentially require knowledge of the local curvature, which is very hard to estimate.

Note that the SiZer map gives some interesting visual cues as to how one might choose a local bandwidth function, which is described as a curve running across the map. For example, the Sheather–Jones plug-in bandwidth could be used for $x \in (-3, 2)$, then the bandwidth curve could move down to around $\log_{10}(h) = -1.4$ for $x \in (2.3, 3)$. An interactive approach to local bandwidth selection could be based on tracing a "bandwidth curve" with

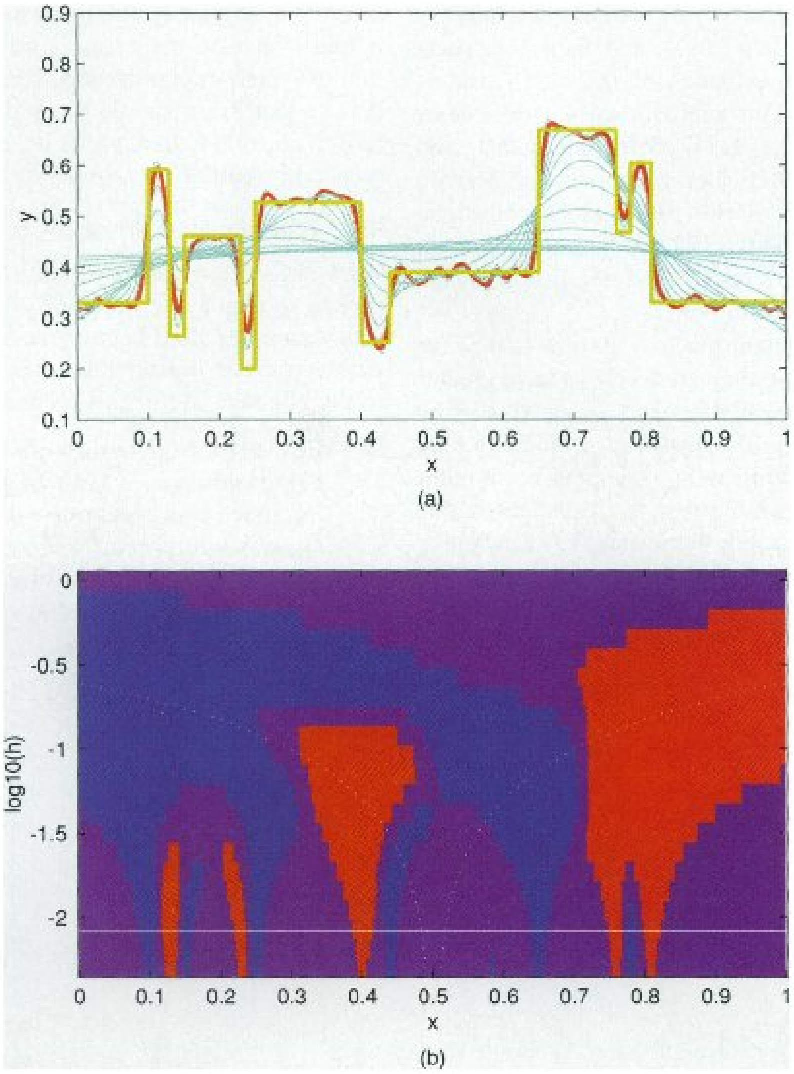


Figure 8. Family Plot and SiZer Map, for Local Linear Regression, Based on $n = 1,024$ Simulated Data, Shown as Green Dots, From an Equally Spaced Design, and the Donoho–Johnstone Blocks Regression Curve Shown as the Thick Green Curve in (a). Simulated errors are independent Gaussian, with standard deviations $\sigma = .05$. The thick red curve in (a) is the Ruppert–Sheather–Wand direct plug-in bandwidth, which is highlighted in the SiZer map in (b).

a mouse on the SiZer map. The resulting local bandwidth smooth could be shown in another window. If the family has already been computed, then computation of the local bandwidth smooth would be very fast, because it only needs interpolation among the family members. Marron and Udina (1997) discussed a different approach to interactive local bandwidth selection.

A natural question is, with SiZer, why do we need local bandwidth smoothing? The answer is that for presenting conclusions to nonexperts (who are not interested in details behind the conclusions), a single location-varying smooth will be very simple and attractive.

5.6 Higher Dimensions

The problem of which features are really present? is also very important in smoothing settings of more than one dimension. In particular, the two-dimensional smoothing problems arise in "image analysis," which has a very large literature. An important problem with extending SiZer to higher dimensions is how to present the "map." The very simplest two-dimensional version that one might try is to study the magnitude of the gradient, and highlight scale-space regions where this is significantly above 0. But now the map would be shaded regions in three dimensions, which is fairly challenging to visualize. Minnotte and Scott (1993) faced analogous challenges in developing a two-dimensional version of their mode tree.

Other applications would likely result in the need to visualize even higher-dimensional maps. For example, one could replace the magnitude of the gradient by directional derivatives. As another example, in some cases it could be desirable to use different bandwidths in different directions. Even with a two-dimensional image, implementation of both ideas would result in a six- or seven-dimensional map.

6. OTHER APPROACHES

6.1 Mode Testing

An older approach to the analysis of significant features in a smooth is mode testing. Here one formulates a null hypothesis of "few modes" (e.g., one), and then constructs a test which seeks strong evidence of the alternative of "more modes" (e.g., two). This approach goes back at least to Good and Gaskins (1980); later work includes Cheng and Hall (1997), Donoho (1988), Fisher, Mammen, and Marron (1994), Fisher and Marron (1998), Hartigan and Hartigan (1985), Hartigan and Mohanty (1992), Mammen, Marron, and Fisher (1992), Minnotte and Scott (1993), Müller and Sawitzki (1991), and Silverman (1981).

Such tests have an important place, even now that SiZer has been developed, because they are likely to have greater power than the inferences available from SiZer. This is because they focus directly on the question of modality, and also because they are not hampered by trying to be simultaneous over all of scale space. However, most available mode tests have the weakness that they determine only how many modes are present, and do not say where the modes are, or even which features in the smooth are the modes. (See Mammen, Marron, and Udina 1997 and Minnotte 1997 for some interesting exceptions to this.) The strength of SiZer is that it gives a much faster way of addressing the question of which modes are significant and which are not. We believe that SiZer should be used mostly as an exploratory tool, with follow-up analysis by explicit mode tests recommended in borderline cases.

6.2 Why Not Conventional Confidence Bands?

In classical parametric statistics, a time-honored approach to displaying variability is the confidence interval. Many attempts have been made to extend this idea to nonparametric curve estimation. There are two major hurdles to the effective use of this technique:

- Instead of a single real-valued parameter, the quantity being estimated is now an entire curve. Furthermore, inference about features will involve aspects of simultaneous inference.
- Unlike conventional parameter estimation, curve estimation necessarily involves an important bias component.

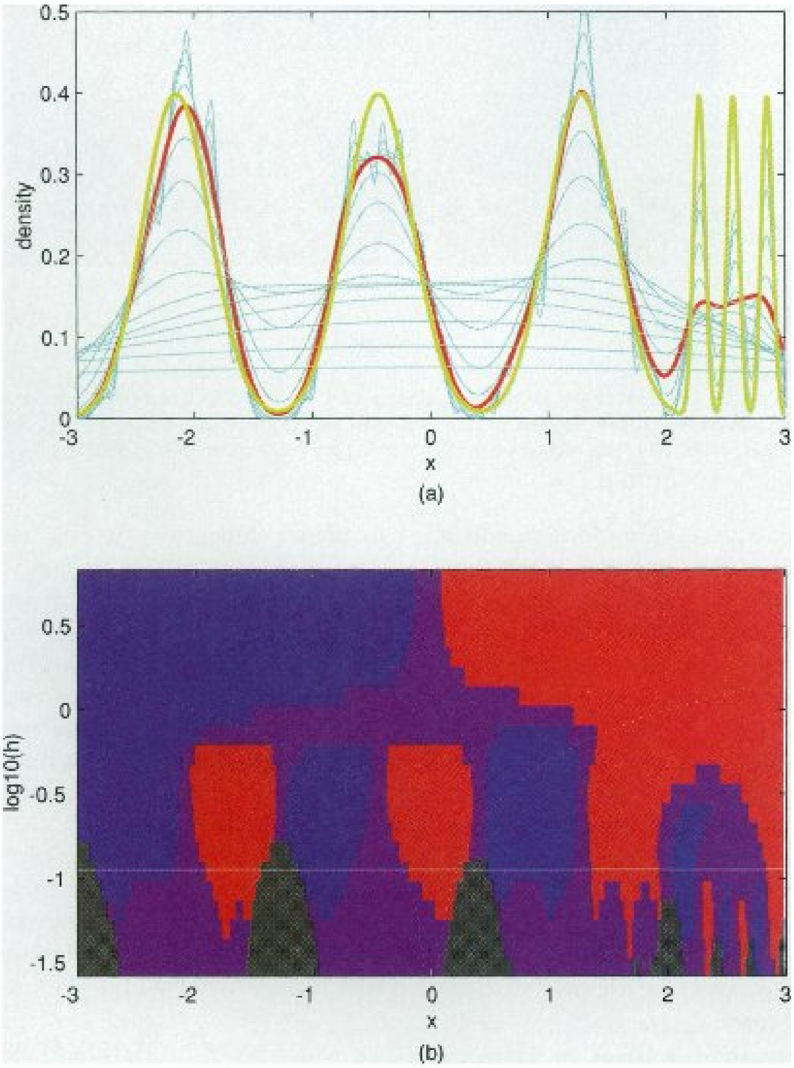


Figure 9. Family Plot (a) and SiZer Map (b) for Kernel Density Estimates, Based on n = 1,000 Simulated Data, From the Marron and Wand Density #15, "Discrete Comb," Shown as the Thick Yellow Green in the Family Plots. The thick red curve in the family plots is the Sheather–Jones plug-in bandwidth, which is the highlighted horizontal bar in the SiZer maps.

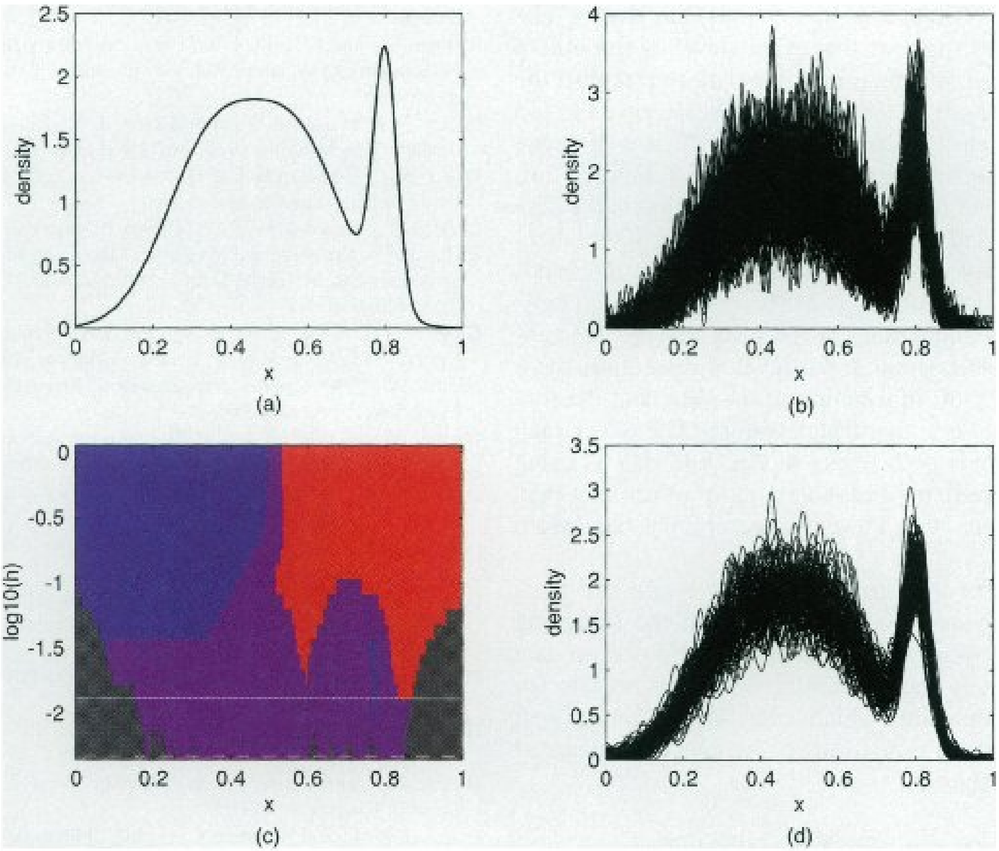


Figure 10. Underlying Normal Mixture Density f (a); SiZer Map for a Simulated Dataset of Size n = 500 (c); 100 Replicates of Kernel Density Estimates, Using the Coverage-Optimal Bandwidth (b); and the Bandwidth That is MSE Optimal at x = .8 (d). These two bandwidths are highlighted in the SiZer map as a solid line for the MSE optimal and a dashed line for the coverage optimal.

There is a large literature on attempts to address these problems in the context of smoothing. Good access is provided through the monograph of Hall (1992).

A quick and simple approach has been suggested by, for example, Hastie and Tibshirani (1990), where one ignores both of these hurdles and simply writes down standard confidence intervals that capture only the variability part of the error. If the goal is to make confidence statements about the true underlying curve $f(x)$, then this approach is inappropriate, because the bias is ignored, and the pointwise nature of the intervals makes them too short for valid inference.

A time-honored approach to handling bias is to make it negligible by undersmoothing; that is, using a very small bandwidth. Many do this simply by assuming that asymptotically as the sample size grows, the bandwidth tends to 0 faster than the optimal, which causes the bias to tend to 0 at a faster rate. This still leaves open the problem of how the bandwidth should be chosen, and the fact that for any fixed set of data, any bandwidth will have at least a little bias. But even ignoring these problems, confidence intervals based on such bandwidths are not intuitively appealing, because they may be expected to be unnecessarily long; that is, significant features can be missed.

Another approach is to try to estimate the bias and adjust accordingly. An attempt at this presented by Härdle and Marron (1991) was asymptotically successful, but gave incorrect coverage in simulations, as discussed in their section 3. They also showed that the reason for the error was because the bias estimate was inefficient. D. Nychka (personal communication, 1988) provided an intuitive explanation of this with the statement "if you could estimate bias effectively, then you could get an improved estimate."

Nice insight into the failure of bias-correction methods was developed in several papers by Hall (and is well summarized in Hall 1991, sec. 4.4). The approach taken there is to choose the bandwidth to make coverage probabilities as close as possible to the desired values. Asymptotic theory is developed for optimal bandwidths according to this criterion, and it is shown that when optimal bandwidths are used, simple undersmoothed bandwidths give shorter confidence intervals than if one attempts any type of bias correction.

This motivates a more careful look at undersmoothed bandwidths, and a natural question is: How long are the coverage optimal confidence intervals? Figure 10 shows an example addressing this point, using the explicit representation given just after (3.5) of Hall (1991).

The true underlying density shown in Figure 10(a) is the Gaussian mixture density

$$.425 \cdot \mathrm{N}(.35, .0144) + .425 \cdot \mathrm{N}(.575, .0144) + .15 \cdot \mathrm{N}(.8, .0009).$$

Here we study its estimation when $n = 500$ data points are used, and focus on the thinner peak; that is, on estimation at $x = .8$. The practical effect of the coverage-optimal bandwidth is shown in Figure 10(b), where overlays of kernel density estimates for 100 independent replicates (i.e., regeneration of the $n = 500$ pseudo–data points) are shown. The envelope of curves suggests that at this level of smoothing there is not enough information in the data to establish the statistical significance of the thinner peak, because the top of the envelope near the valley point $x = .72$ is well above the bottom of the envelope at the peak, $x = .8$. Figure 10(d) investigates whether there is enough information in the data to resolve the second peak, by again overlaying 100 realizations of the density estimate, but this time with the bandwidth chosen to minimize the MSE $= E[\hat{f}_h(x) - f(x)]^2$ (approximated by simulation) at the peak $x = .8$. This envelope of curves shows that at this level of resolution, there seems to be plenty of information in the data, and the second mode should be a significant feature. The SiZer map in Figure 10(c) finds both of the modes, and thus is using the information available in the data more effectively than confidence intervals with the coverage optimal bandwidth can do.

Note that even if it were possible to get effective classical confidence bands (doubtful in view of the foregoing discussion), then SiZer would still be a more powerful data analytic tool. This is because confidence bands need to focus on a single bandwidth, which (even when it can be well chosen from the data) can still miss features that appear at other levels of resolution.

[Received October 1997. Revised March 1999.]

REFERENCES

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques: the Kernel Approach With S-Plus*, Oxford, U.K.: Oxford University Press.

Bralower, T. J., Fullagar, P. D., Paull, C. K., Dwyer, G. S., and Leckie, R. M. (1997), "Mid-Cretaceous Strontium-Isotope Stratigraphy of Deep-Sea Sections," *Geological Society of America Bulletin*, 109, 1421–1442.

Brown, L. D., Johnstone, I. M., and McGibbon, K. B. (1981), "Variation Diminishing Transformations: A Direct Approach to Total Positivity and Its Statistical Applications," *Journal of the American Statistical Association*, 76, 824–832.

Chaudhuri, P., and Dewanji, A. (1995), "On a Likelihood-Based Approach in Nonparametric Smoothing and Cross-Validation," *Statistics and Probability Letters*, 22, 7–15.

Chaudhuri, P., and Marron, J. S. (1997), "Scale-Space View of Curve Estimation," Mimeo Series #2357, North Carolina Institute of Statistics.

Cheng, M. Y., and Hall, P. (1997), "Calibrating the Excess Mass and Dip Tests of Modality," *Journal of the Royal Statistical Society*, Ser. B, 60, 579–589.

Donoho, D. (1988), "One Sided Inference About Functionals of a Density," *The Annals of Statistics*, 16, 1390–1420.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman and Hall.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Farmen, M. (1996), "The Smoothed Bootstrap for Variable Bandwidth Selection and Some Results in Nonparametric Logistic Regression," PhD dissertation, Technical Report Series 2342, North Carolina Institute of Statistics.

Farmen, M., and Marron, J. S. (1997), "An Assessment of Finite Sample Performance of Adaptive Methods in Density Estimation," unpublished manuscript.

Fisher, N. I., Mammen, E., and Marron, J. S. (1994), "Testing for Multimodality," *Computational Statistics and Data Analysis*, 18, 499–512.

Fisher, N. I., and Marron, J. S. (1998), "Mode Testing via the Excess Mass Estimate," unpublished manuscript.

Good, I. J., and Gaskins, R. A. (1980), "Density Estimation and Bump-Hunting by the Penalized Maximum Likelihood Method Exemplified by Scattering and Meteorite Data" (with discussion), *Journal of the American Statistical Association*, 75, 42–73.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Härdle, W., and Marron, J. S. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *The Annals of Statistics*, 19, 778–796.

Hall, P. (1991), "Edgeworth Expansions for Nonparametric Density Estimators," *Statistics*, 2, 215–232.

——— (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Hall, P., and Marron, J. S. (1997), "On the Role of the Ridge Parameter in Local Linear Smoothing," *Probability Theory and Related Fields*, 108, 495–516.

Hartigan, J. A., and Hartigan, P. M. (1985), "The DIP Test of Multimodality," *The Annals of Statistics*, 13, 70–84.

Hartigan, J. A., and Mohanty, S. (1992), "The RUNT Test From Multimodality," *Journal of Classification*, 9, 63–70.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Izenman, A. J., and Sommer, C. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, 83, 941–953.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996a), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

——— (1996b), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," *Computational Statistics*, 11, 337–381.

Karlin, S. (1968), *Total Positivity*, Stanford, CA: Stanford University Press.

Lindeberg, T. (1994), *Scale-Space Theory in Computer Vision*, Boston: Kluwer.

Mammen, E., Marron, J. S., and Fisher, N. I. (1992), "Some Asymptotics for Multimodality Tests Based on Kernel Density Estimates," *Probability Theory and Related Fields*, 91, 115–132.

Mammen, E., Marron, J. S., and Udina, F. (1997), "Interactive Mode Testing," unpublished manuscript.

Marron, J. S., and Chaudhuri, P. (1998a), "Significance of Features via SiZer, Statistical Modeling," *Proceedings of the 13th International Workshop on Statistical Modeling*, New Orleans, eds. B. Marx and H. Friedl.

——— (1998b), "When is a Feature Really There? The SiZer Approach," in *Automatic Target Recognition VII*, ed. Firooz A. Sadjadi, Proceedings of the Society of Photooptic and Industrial Engineering, Vol. 3371, 306–312.

Marron, J. S., and Chung, S. S. (1997), "Presentation of Smoothers: The Family Approach," unpublished manuscript.

Marron, J. S., and Udina, F. (1997), "Interactive Local Bandwidth Choice," unpublished manuscript.

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

Minnotte, M. C. (1997), "Nonparametric Testing of the Existence of Modes," *The Annals of Statistics*, 25, 1646–1660.

Minnotte, M. C., and Scott, D. W. (1993), "The Mode Tree: A Tool for Visualization of Nonparametric Density Features," *Journal of Computational and Graphical Statistics*, 2, 51–68.

Müller, D. W., and Sawitzki, G. (1991), "Excess Mass Estimates and Tests for Multimodality," *Journal of the American Statistical Association*, 86, 738–746.

Müller, H. G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Heidelberg: Springer-Verlag.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Schmitz, H. P., and Marron, J. S. (1992), "Simultaneous Estimation of Several Size Distributions of Income," *Econometric Theory*, 8, 476–488.

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

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Ser. B, 43, 97–99.

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

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

——— (1997), "Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation," *International Statistical Review*, 66, 137–156.

Staniswalis, J. G. (1989), "The Kernel Estimate of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association*, 84, 276–283.

Tibshirani, R. J., and Hastie, T. J. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–568.

Wahba, G. (1991), *Spline Models for Observational Data*, Philadelphia: SIAM.

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