Curvature vs. Slope Inference for Features in Nonparametric Curve Estimates

March 11, 2002

Abstract

First derivative based tools have been very popular for detecting features in nonparametric curve estimators. However, in many applications second derivative information is quite important for identifying statistically signi…cant features. This paper illustrates several di¤erent ways in which second derivative based inference signi…cantly improves upon methods based on …rst derivatives. The scale space viewpoint provides the foundation for e¤ective use of second derivative information in our inference.

1 Introduction: why study the second derivative?

A traditional limitation of the nonparametric curve estimation tools, when applied to real data, is the challenge of assessing the statistical signi…cance of observed features in the smoothed data. For example, when a bump appears in a curve estimate (in the course of an exploratory data analysis), this could be a discovery of scienti…c importance, or it could be simply due to sampling variation.

The SiZer method, developed by Chaudhuri and Marron (1999), has overcome this limitation, by an e¤ective combination of statistical inference in scale space and visualization. See Lindeberg (1994) and ter Haar Romeny (2001) for introduction to scale space ideas. Other approaches to this problem are bump hunting and mode testing, see Good and Gaskins (1980) Silverman (1981), Hartigan and Hartigan (1985), Izenman, A. J. and Sommer, C. (1988), Müller and Sawitzki (1991), Hartigan and Mohanty (1992), Minnotte and Scott (1993), Fisher, Mammen and Marron (1994), Donoho (1998), Cheng and Hall (1997), Minnotte, M. C. (1997) and Fisher and Marron (2001). A less attractive approach to inference for curve estimation is classical con…dence bands, see Section 6.2 of Chaudhuri and Marron (1999) for discussion. An important advantage of SiZer over these other approaches is that not only is the number of bumps investigated, but also their location, as well as other types of features. Another difference is that the inference of SiZer focuses on the underlying curve, at a given scale of resolution (i.e. for a given level of the smoothing parameter).

Many of the above methods are based on the first derivative of the curve estimate, and none explicitly uses the second derivative. But some features are better detected using information about the second derivative. The main contribution of this paper, is the study of the importance of second derivative information for exploratory data analysis (both density estimation and regression problems). We observe that this information is especially powerful when used in conjunction with first derivative information.

A first example showing the usefulness of second derivative information is shown in Figure 1, where the data are half marathon times from a full marathon foot race in Raleigh, North Carolina in December 2000. It was suspected that early in this race, a leading group of runners was mistakenly sent on a shorter route. When the mistake was discovered, the remaining runners were sent on a longer route, thus opening up a large gap between them and the first group. The first official times of the runners were measured half way through the race, by which time some mixing of the groups had taken place, leading to a mixture of these two distributions in the data. The top panel of Figure 1 shows the $n = 1056$ half marathon times (in minutes), as green dots (with a random vertical "jitter", see pages 121-122 of Cleveland), together with a family of kernel density estimates (defined in Section 3.1), indexed by the smoothing parameter, shown as blue curves. See e.g. Silverman (1986), Scott (1992) and Wand and Jones (1995) for additional discussion of kernel density estimation.

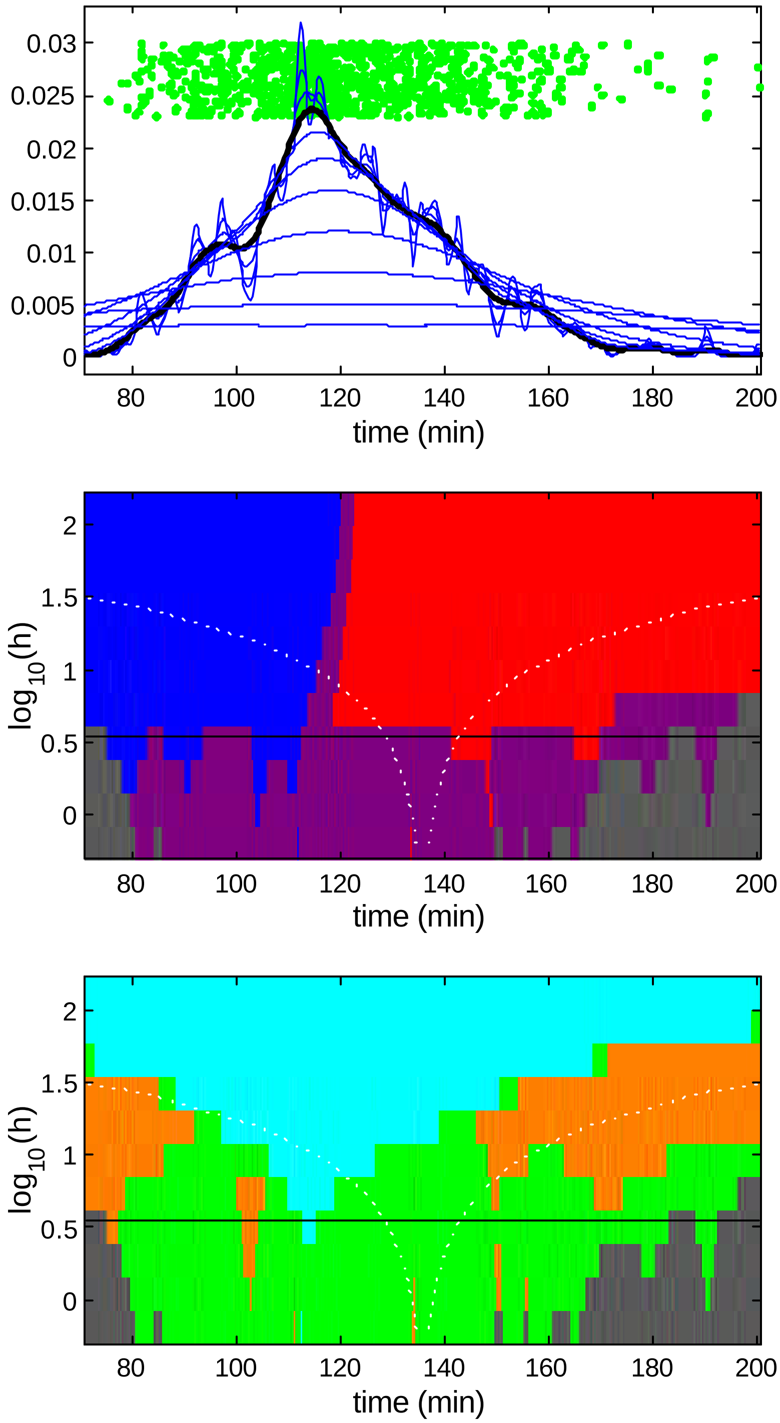

Figure 1: Half Marathon times, for the Raleigh Marathon. The black density estimate curve in the top panel suggests a mixture of two distributions. The first derivative analysis in the second panel does not confirm bimodality. The second derivative analysis in the bottom panel indicates that the shoulder on the left is statistically significant.

The thick black curve is the Sheather-Jones Plug In bandwidth (see Jones, Marron and Sheather (1996a,b) for a discussion of data driven bandwidth selection), and the bimodal structure suggests that the data may have come from a mixture of two subpopulations. However, the other bandwidths cast some doubt on the strength of the evidence in favor of bimodality. In particular, by oversmoothing with a larger bandwidth, the two modes converge into a single unimodal distribution. By undersmoothing with a smaller bandwidth, the small mode on the left and the valley just to the right of it can be sharpened, but many spurious modes (some of even larger magnitudes) also appear. This is an example of a problem routinely encountered in exploratory data analysis. There is some suggestion of an important feature, but the question of statistical significance of the feature is critical and it is not easy to resolve.

The middle panel of Figure 1 shows a first derivative based analysis of the Raleigh Marathon data. This map is a visual representation of statistical significance of the slopes of the family of kernel density estimates with varying choices of the bandwidth (i.e. over the scale space). The horizontal axis is the same as in the top panel, and the vertical axis shows the bandwidth (i.e. level of resolution of the data) in a logarithmic scale. The funnel shaped dotted white curves indicate the amount of smoothing being done at each level of resolution, i.e. the width of the Gaussian kernel window as ± 2 standard deviations. Blue regions show significant increase of the curves (at the level $\alpha = 0.05$), red shows significant decrease, and the intermediate color of purple shows lack of significance (i.e. there is no strong evidence for the slope being either positive or negative). One more color shown in this SiZer map is gray, used in regions where the data are too sparse for drawing inference. The first derivative map is blue on the left, and red on the right, thus not supporting the existence of two modes in the data (which would follow from an additional red patch near the shoulder). In other words, for the present sample size, the small mode on the left does not appear to be statistically significant at any level of resolution, using first derivative based inference.

The reason that our first derivative based method fails in this example is that the decrease to the left of the first mode is very small. In fact this feature is more like a "shoulder" than a mode. Instead of being well highlighted by slope, this shoulder is better quantified in terms of curvature, and an inflection point. The color map shown in the bottom panel does second derivative inference, using the color cyan (light blue) to indicate statistically significant concavity (curving downwards, i.e. negative second derivative), and orange to flag statistically significant convexity (curving upwards, i.e. positive second derivative), and green where the curve is very flat or linear with insignificant curvature (i.e. "zero second derivative"). Regions of data sparsity are again indicated using the color gray. Mathematical details of this new second derivative based inference are developed in Section 3.

The map summarizing the second derivative inference, in the bottom panel of Figure 1, shows a statistically significant region of convexity at times around 103 minutes, indicating that the "shoulder" to the left of the central mode is statistically significant. Thus we conclude that the data are indeed a mixture of two populations, and thus that a group of runners received an unfair advantage.

In addition to …nding features not easily visible using …rst derivative tools, second derivative analysis is also of fundamental interest in change point problems. Change points can be studied in terms of the …rst derivative of the smooth, see e.g. Carlstein, Müller and Siegmund (1994). Figure 2 studies a change point example. This time the data are generated as a step function with four steps of integer heights, and additive Gaussian noise with standard deviation = 0.5. The top panel shows the $n = 1024$ data points as green dots. The blue curves are a family of local linear smooths of the data (de…ned in Section 3.2), for di¤erent choices of the bandwidth. See e. g. Wand and Jones (1995) and Fan and Gijbels (1996) for additional discussion of local polynomial regression. Each jump (either up or down) corresponds to a change point in the underlying signal.

The second panel in Figure 2 shows how a …rst derivative analysis can be used to …nd change points, using the distinctive funnel shape of the red and blue regions for small scales. This shape has been mathematically explained by Kim and Marron (2001), who used this to develop a separate visualization tool for …nding jumps.

Change points are even more strongly indicated by using second derivative information because a change point is a local maximum of the …rst derivative, thus a local zero crossing of the second derivative. Because of this zero crossing property of the second derivative at jumps, the curvature based analysis in the bottom panel clearly indicates change points with abrupt changes in color. The information conveyed in the second panel is less e¤ective at highlighting change points because it only shows statistical signi…cance of the …rst derivative, but does not clearly ‡ag local maxima of the …rst derivative.

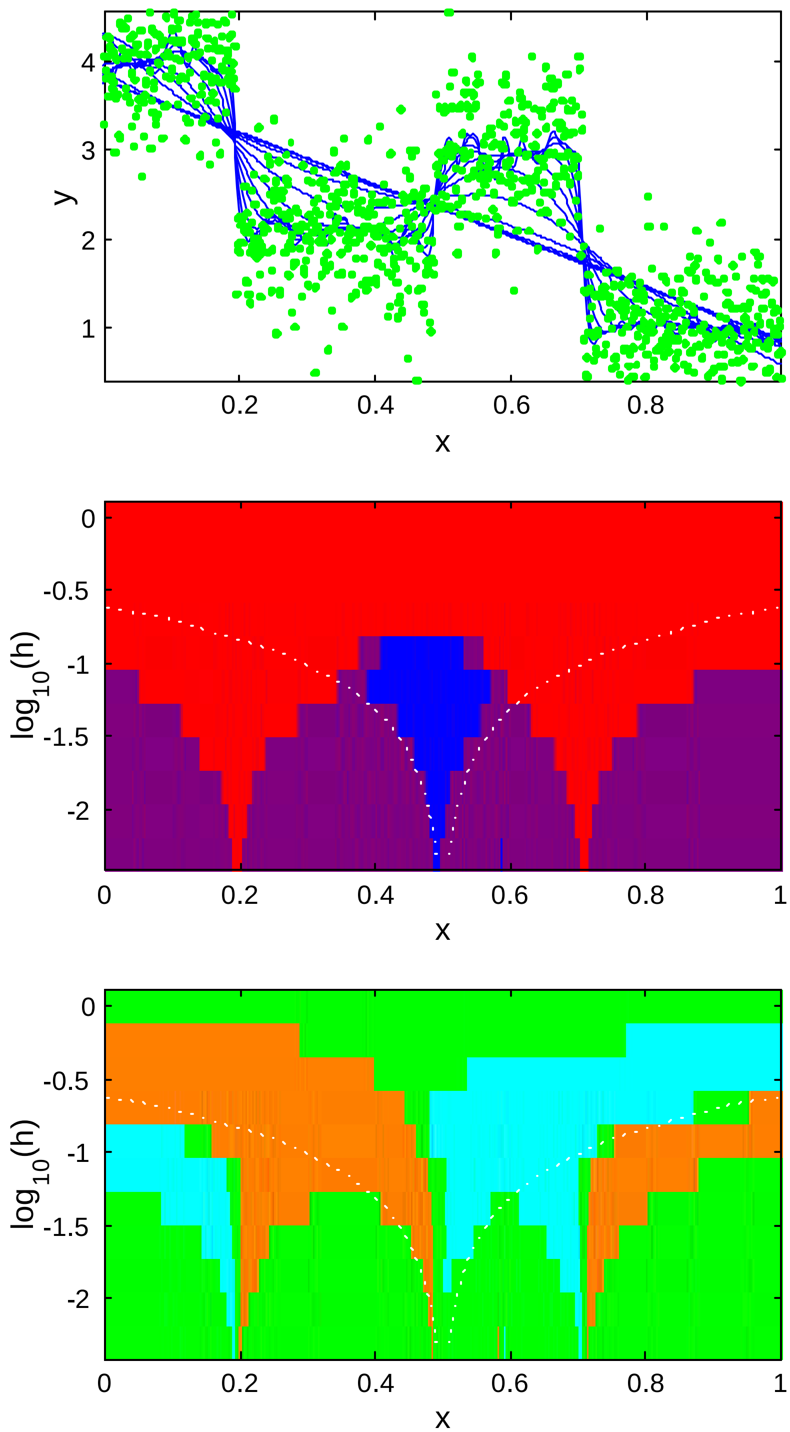

Figure 2: Simulated step function example, about change points. Middle panel shows …rst derivative analysis. Bottom panel shows that signi…cant zero crossings of the second derivative can better highlight jumps.

Additional illustrations of the usefulness of second derivative information are given in Section 2. The details of the statistical inference that underlies SiZer are given in Chaudhuri and Marron (1999). Mathematical and computational details, for statistical inference using second derivative information, are outlined in Section 3.

2 More examples with simulated and real data

In this section additional examples are analyzed, which again show that second derivative information can be very important to statistical inference for features in smooth curves.

Figure 3 shows a simulated example in the context of nonparametric regression. Simulated data points $(X_i; Y_i)$ for $i = 1; :::; 200$, were generated as a tilted sine wave signal with additive noise. Speci…cally, the X_i's are equally spaced on $[0; 1]$ and $Y_i = \sin(8\pi X_i) + 2 + 20X_i + \varepsilon_i$, where the $\varepsilon_1; :::; \varepsilon_n$ are i. i. d. $N(0; 2^2)$. The data points are shown as green dots in the top panel of Figure 3. Because of the tilt in the sin wave, there are regions of strong increase, that alternate with regions of ‡atness. The blue curves are local linear scatterplot smooths.

The …rst derivative analysis is shown in the middle panel of Figure 3. The only colors present are blue and purple, indicating regions of increase and of uncertainty. Thus this analysis provides no conclusive evidence for any interesting features, such as the wiggles of the sine wave, though they are present in the underlying signal. The reason is that the general upward trend provides a "masking e¤ect" that downplays the waves around the line.

This example is deliberately constructed to show that second derivative analysis can be very useful in situations where interesting features are masked by other strong behavior of the …rst derivative In particular, the second derivative analysis, shown in the bottom panel, ‡ags all of the arches of the sine wave as statistically signi…cant structures, using the colors cyan (orange) for signi…cant downward (upward) curvature.

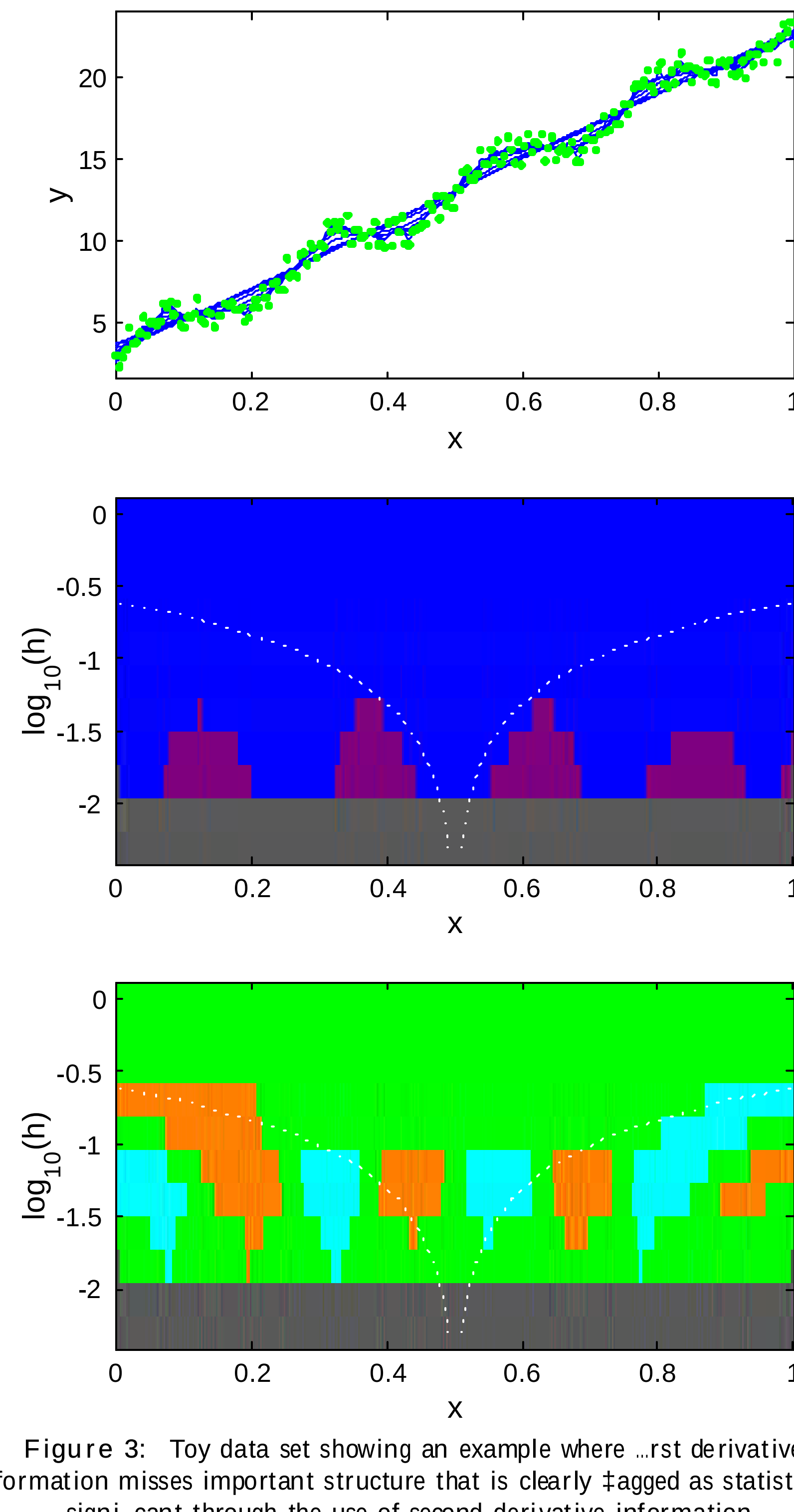

Figure 3: Toy data set showing an example where first derivative information misses important structure that is clearly flagged as statistically significant through the use of second derivative information.

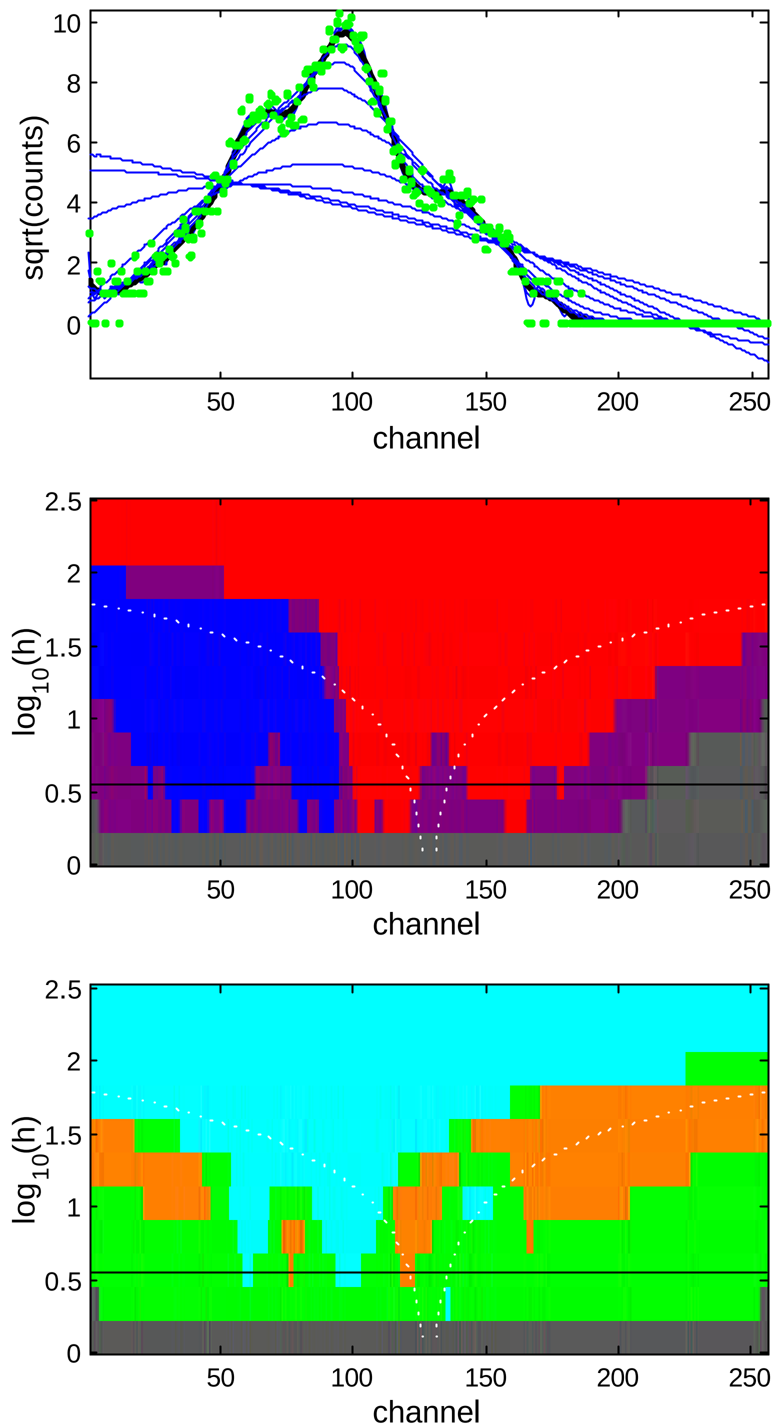

Figure 4: Analysis of Flow Cytometry data. This shows two shoulders in the curve, that are found by the second derivative analysis, but are not statistically significant in the first derivative analysis.

Figure 4 shows another real data example demonstrating the importance of using second derivative information. These data are from the …eld of ‡ow cytometry, where the presence and percentage of ‡orescence marked antibodies on cells are measured. The medical goal is the determination of quantities such as the percentage of lymphocytes among cells. The data come from the laboratory of Drs. S. Mentzer and J. Rawn, Brigham and Women's Hospital, Boston, Massachusetts, and we are grateful to M. P. Wand for putting us in contact with them. In a single experiment, many cells are run through a laser, and the intensity of ‡orescence of each cell is measured, and the data are stored as 256 bin counts, where bins are called "channels". These bin counts are traditionally viewed on the square root scale. The green dots in the top panel are square root bin counts for one such experiment, based on 5000 total cells.

For some ‡ow cytometry data sets, the cells are of the same type, and the marked antibodies have a nearly uniform distribution on the cell, resulting in an approximately Gaussian population in the presence of measurement error. In other data sets, there are two di¤erent subpopulations of cells, with markedly di¤ering degrees of attraction for the marked antibodies, resulting in a clear bimodal population. There are also "in between cases", where there is a suggestion of bimodality, but it is not clear cut, an example is shown in Figure 4. Examples of all of these three cases are not shown in this paper, to save space, but can be viewed on the web page

http://www.stat.unc.edu/faculty/marron/DataAnalyses/
SiZer/SiZer_Examples.html#Eg2:FlowCytometry

The …rst derivative analysis in the middle panel of Figure 4 shows blue on the left and red on the right at a wide range of di¤erent scales, indicating a signi…cant increase then decrease, i.e. unimodality. However, the second derivative analysis shown in the bottom panel, indicates much more structure. In particular, the small orange region near Channel 75, and the small cyan region near Channel 150, ‡ag the two shoulders in the curve that are visible in the top panel as being statistically signi…cant. These shoulders suggest that there are three mixture components in this distribution. Again …rst derivative inference failed to …nd these components, because of the" masking e¤ect" of the overall strong increase and decrease of the curve in those regions.

The data set shown in Figure 4 was chosen from a set of 42 similar analyses. This data set is special because there are actually two di¤erent features that are found by the second derivative analysis, but not from the …rst derivative. However, features of this type were rather frequent, in fact occurring in13 of the 42 data sets considered.

Next is an example which demonstrates that statistical signi…cance of a feature can be observed simultaneously using …rst and second derivatives, however the signi…cance may show up at di¤erent levels of smoothing for the two derivatives. Figure 5 shows the 1975 British Family Incomes data, that were analyzed in Figures 1 and 2 of Chaudhuri and Marron (1999). Again the green dots show the raw data, with a random vertical jitter for good separation. The blue family of kernel density estimates reveals both expected features of income distributions, such as a long right tail, and a large number of lower to middle income families on the left, and some unexpected features, such as two modes. At one point an important question was whether these modes were signi…cant structures, and an a¢rmative answer was provided by Schmitz and Marron (1992), and corroborated using a scale space analysis by Chaudhuri and Marron (1999).

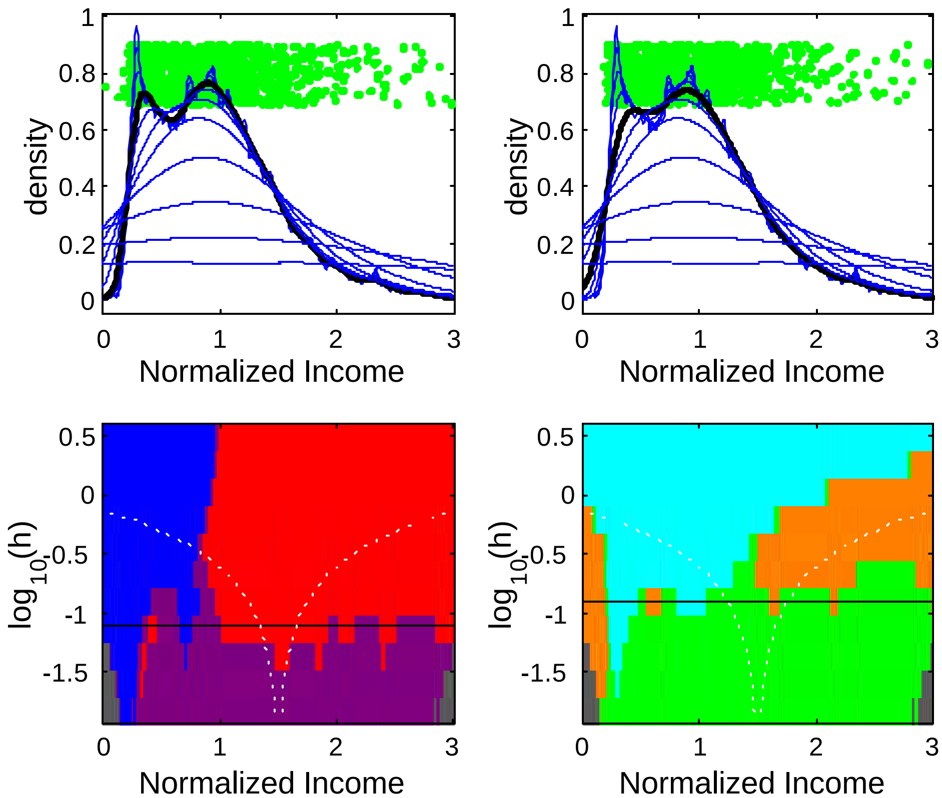

Figure 5: Analysis of the British Incomes Data. Shows that signi…cant structure appears at di¤erent scales for …rst and second derivative based inference.

Here the …rst and the second derivative analyses are done side by side to illustrate several important di¤erences. The …rst and second derivative color maps shown in the bottom panels both indicate the signi…cance of the bimodal structure. In particular, at the scale indicated by the black horizontal bar in the bottom left map (this same scale is highlighted in the family of smooths directly above), the …rst derivative color changes, of blue-red-blue-red, ‡ag both modes as statistically signi…cant. Similarly at the scale indicated by the horizontal black bar in the lower right map, the signi…cance of the two modes is ‡agged by the orange-cyan-orange-cyan-orange color changes. A very important point is that the statistical signi…cance of the modes shows up at two quite di¤erent scales in the two maps. This highlights a key difference between inferences that can be drawn using first and second derivative information in the presence of noise. In particular, the second derivative estimate is more strongly affected by sample variations at small scales than the first derivative estimate. This appears clearly in the bottom panels, because the red-blue regions appear at smaller scales in the map on the left highlighting significance of the first derivative, than the cyan - orange regions in the map on the right highlighting significance of the second derivative. There are some well known mathematics behind this phenomenon, discussed in Section 3.3 below.

3 Mathematical Details

Let $\hat{f}_h(x)$ denote a nonparametric curve estimate. Our approach to statistical inference is based on confidence limits for the first and second derivatives. Behavior at x and h locations is presented via color maps where different colors indicate regions where the derivatives are significantly positive, significantly negative or insignificant. This inference is based on confidence limits of the form

$$\hat{f}'_h(x) \pm q \cdot \widehat{sd}\left(\hat{f}'_h(x)\right)$$

or

$$\hat{f}''_h(x) \pm q \cdot \widehat{sd}\left(\hat{f}''_h(x)\right) ;$$

depending on the derivative of interest, where q is an appropriate quantile (see Section 3 of Chaudhuri and Marron 1999), and the standard deviation is estimated as discussed below. The derivative is significantly positive (negative) when both confidence limits are above (below) 0, and insignificant when the confidence limits bracket 0.

It can be shown that when the first derivative $E\hat{f}'_h(x)$ and the second derivative $E\hat{f}''_h(x)$ curves, viewed at a specific level of smoothing h have a finite number of zero crossings over a compact interval, all those zero crossings will be detected with probability tending to one as the sample size grows by the our procedures for assessing statistical significance based on confidence limits, and they will be marked by color changes in the respective color maps summarizing the first and second derivative inferences. For this and other related asymptotic consistency results and their implications see Chaudhuri and Marron (2000).

Because repeated calculation of smoothers is required for these color maps, fast computational methods are very important. Binned (also called "WARPed") methods are natural for this, because the data need only be binned once. See Fan and Marron (1994) for detailed discussion of this, and other fast computation methods.

Further details are substantially different for density estimation, as illustrated in Figures 1 and 5, and regression, as illustrated in Figures 2, 3 and 4. Density estimation is treated in Section 3.1 and regression in Section 13.

3.1 Density Estimation

Given a set of data $X_1, \dots, X_n$ from a smooth probability density $f(x)$, the kernel estimate of f is

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

where h is the "bandwidth" and K_h is the "h-rescaling" of the kernel function K, $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. The main idea is to "put probability mass $\approx \frac{1}{n}$ near each X_i". See for example Silverman (1986), Scott (1992) and Wand and Jones (1995). Density derivative estimates are obtained by differentiating $\hat{f}_h(x)$,

$$\begin{aligned}
\hat{f}'_h(x) &= \frac{1}{n}\sum_{i=1}^{n} K'_h(x - X_i), \\
\hat{f}''_h(x) &= \frac{1}{n}\sum_{i=1}^{n} K''_h(x - X_i),
\end{aligned}$$

where $K'_h(\cdot) = \frac{1}{h^2}K'\left(\frac{\cdot}{h}\right)$, $K''_h(\cdot) = \frac{1}{h^3}K''\left(\frac{\cdot}{h}\right)$. Using the same scale space viewpoint as in Chaudhuri and Marron (1999, 2000), $\hat{f}'_h(x)$ and $\hat{f}''_h(x)$ are considered to be estimates of $E\hat{f}'_h(x)$ and $E\hat{f}''_h(x)$, respectively, which represent the derivatives of f at the level of resolution h. Since both of these estimates are simple averages of i. i. d. random variables, their variances are simply estimated as the corresponding sample standard deviations,

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

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

3.2 Regression

Given a sample of paired data $(X_1, Y_1), \dots, (X_n, Y_n)$, the local linear and local quadratic regression estimates of the conditional expected value, i.e. the regression function,

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

are obtained as the solutions a of the locally weighted least squares problems

$$\min_{a,b} \sum_{i=1}^{n} \left[Y_i - (a + b(X_i - x))\right]^2 K_h(x - X_i), \tag{2}$$

$$\min_{a,b,c} \sum_{i=1}^{n} \left[Y_i - \left(a + b(X_i - x) + \frac{c}{2}(X_i - x)^2\right)\right]^2 K_h(x - X_i). \tag{3}$$

See e.g. the monographs of Wand and Jones (1996) and Fan and Gijbels (1996). The local linear estimate of the slope is given by $\hat{f}'_h(x) = b$ from (2), while the local quadratic estimate of the second derivative is $\hat{f}''_h(x) = c$ from (2). Here again, following the scale space idea, $\hat{f}'_h(x)$ and $\hat{f}''_h(x)$ are considered to be estimates of their expected values, which again represent derivatives of the regression function f at the level of resolution h. In this paper, local linear fits are used for curve estimation and estimation of the first derivative, while local quadratic fits are used for second derivative estimation. This choice was made for reasons of simplicity, see e.g. Fan and Gijbels (1996) for detailed discussion of other choices of local polynomial order and the relation to derivative estimation.

Since these estimates are solutions of weighted least squares problems, their variances can be obtained from standard formulas, using the estimate of residual variance, $\sigma^2(x) = \mathrm{var}(Y|X = x)$, based on the minimum value of (2) or (3) as appropriate. For example, in the local linear case, the variance of the slope estimates is the lower right entry of the 2×2 matrix

$$\sigma^2(x)\begin{pmatrix} 1 & \frac{\sum_i (X_i - x)K_h(x - X_i)}{\sum_i K_h(x - X_i)} \\ \frac{\sum_i (X_i - x)K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^2 K_h(x - X_i)}{\sum_i K_h(x - X_i)} \end{pmatrix}^{-1}.$$

Similarly, in the local quadratic case, the variance of the second derivative estimate is the lower right entry of the 3×3 matrix

$$4\sigma^2(x)\begin{pmatrix} 1 & \frac{\sum_i (X_i - x)K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^2 K_h(x - X_i)}{\sum_i K_h(x - X_i)} \\ \frac{\sum_i (X_i - x)K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^2 K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^3 K_h(x - X_i)}{\sum_i K_h(x - X_i)} \\ \frac{\sum_i (X_i - x)^2 K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^3 K_h(x - X_i)}{\sum_i K_h(x - X_i)} & \frac{\sum_i (X_i - x)^4 K_h(x - X_i)}{\sum_i K_h(x - X_i)} \end{pmatrix}^{-1}.$$

3.3 Statistical Variation in Derivative Estimation

In Figure 5 above, it was noted that first derivative inference can be done at smaller scales than second derivative inference. This can be easily understood by studying the variances of $\hat{f}'_h(x)$ and $\hat{f}''_h(x)$. For either density estimation or regression, these have asymptotic (as $n \to \infty$) order

$$\mathrm{var}\left(\hat{f}'_h(x)\right) \approx \frac{C'}{nh^3};$$
$$\mathrm{var}\left(\hat{f}''_h(x)\right) \approx \frac{C''}{nh^5};$$

for some constants C' and C''. For detailed calculation of this, and other scale space asymptotic results, see Chaudhuri and Marron (2000). Thus for small bandwidths h (important for good performance of smoothing methods), the second derivative will have larger variance. More specifically, it is clear that our inference in scale space will never flag significance at "small scales", in particular of the order $h = o\left(n^{-1/3}\right)$ for the first derivative, and of the order $h = o\left(n^{-1/5}\right)$ for the second derivative, because in those cases the variance will tend to infinity. The actual features that are found in a specific case will be determined by a trade off of this variance, with the sample size and the strength of the underlying features, as reflected in the magnitudes of the derivatives.

References

[1] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, Journal of the American Statistical Association, 94, 807-823.

[2] Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation, Annals of Statistics, 28, 408-428.

[3] Cheng M. Y. and Hall, P. (1997) Calibrating the excess mass and dip tests of modality, Journal of the Royal Statistical, Series B, 60, 579–589.

[4] Cleveland, W. S. (1993) Visualizing Data, Hobart Press, Summit, New Jersey, U.S.A

[5] Donoho, D. (1988) One sided inference about functionals of a density, Annals of Statistics, 16, 1390-1420.

[6] Fan, J. and Gijbels, I. (1996) Local polynomial modeling and its applications, Chapman and Hall, London.

[7] Fan, J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, Journal of Computational and Graphical Statistics, 3, 35-56.

[8] Fisher, N. I., Mammen, E. and Marron, J. S. (1994) Testing for multimodality, Computational Statistics and Data Analysis, 18, 499-512.

[9] Fisher, N. I. and Marron J. S. (2001) Mode Testing Via the Excess Mass Estimate, Biometrika, 88, 499-517.

[10] 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.

[11] Hartigan, J. A. and Hartigan, P. M. (1985) The DIP test of multimodality, Annals of Statistics, 13, 70-84.

[12] Hartigan, J. A. and Mohanty, S. (1992) The RUNT test from multimodality, J. Classification, 9, 63-70.

[13] Izenman, A. J. and Sommer, C. (1988) Philatelic mixtures and multimodal densities, Journal of the American Statistical Association, 83, 941-953.

[14] 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.

[15] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b) Progress in data-based bandwidth selection for kernel density estimation, Computational Statistics, 11, 337-381.

[16] Kim, C. S. and Marron, J. S. (2001) SiZer for jump detection, to appear in Journal of Nonparametric Statistics.

[17] Lindeberg, T. (1994) Scale-Space Theory in Computer Vision, Kluwer, Dordrecht.

[18] Minnotte, M. C. (1997) Nonparametric testing of the existence of modes, Annals of Statistics, 25, 1646-1660.

[19] 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.

[20] Müller, D. W. and Sawitzki, G. (1991) Excess mass estimates and tests for multimodality, Journal of the American Statistical Association , 86, 738-746.

[21] Carlstein, E., Müller, H.G. and Siegmund, D. (1994) Change-point problems, IMS Lecture Notes Monograph Series, 23, Institute of Mathematical Statististics, Hayward, CA.

[22] Scott, D. W. (1992) Multivariate density estimation, theory, practice and visualization, John Wiley: New York.

[23] Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, Econometric Theory, 8, 476-488.

[24] Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality, Journal of the Royal Statistical Society, Series B, 43, 97-99.

[25] Silverman, B. W. (1986) Density estimation for statistics and data analysis, Chapman and Hall: London.

[26] ter Haar Romeny, B. M. (2001) Front-End Vision and Multiscale Image Analysis, Kluwer Academic Publishers, Dordrecht, the Netherlands.

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