When is a feature really there? The SiZer approach

J. S. Marron[a] and P. Chaudhuri[b]

[a]Department of Statistics, University of North Carolina, Chapel Hill, NC, USA

[b]Indian Statistical Institute, Calcutta, India

ABSTRACT

Statistical smoothing methods are useful for finding important and nonobvious structure in data. However, some of the features discovered in this way can be spurious sampling artifacts. The SiZer approach (based on studying statistical SIgnificance of ZERo crossings of smoothed estimates) to analyzing which visible features represent important underlying structures, is discussed.

1. BACKGROUND

Smoothing methods in statistics, also called nonparametric curve estimates, are a powerful method for finding often unexpected structure in data. Two examples of this are shown in Figure 1. Many more examples are available in the many monographs on smoothing in the statistical literature. In the last 5 years these include Green and Silverman (1994), Wand and Jones (1995), Simonoff (1996), Fan and Gijbels (1996), Hart (1997) and Bowman and Azzelini (1997).

Scatterplot smoothing is illustrated in Figure 1. Precise formulation is given in Section 2. A smooth is essentially a local average, indexed by a "smoothing parameter", or "window width" h, which controls the amount of smoothing done. When h is small, i.e. undersmoothed, the resulting curve estimates are quite wiggly, because too much of the sampling variability is present. When h is large, i.e. oversmoothed, the resulting curve estimates are very smooth, and important underlying structure can be smoothed away.

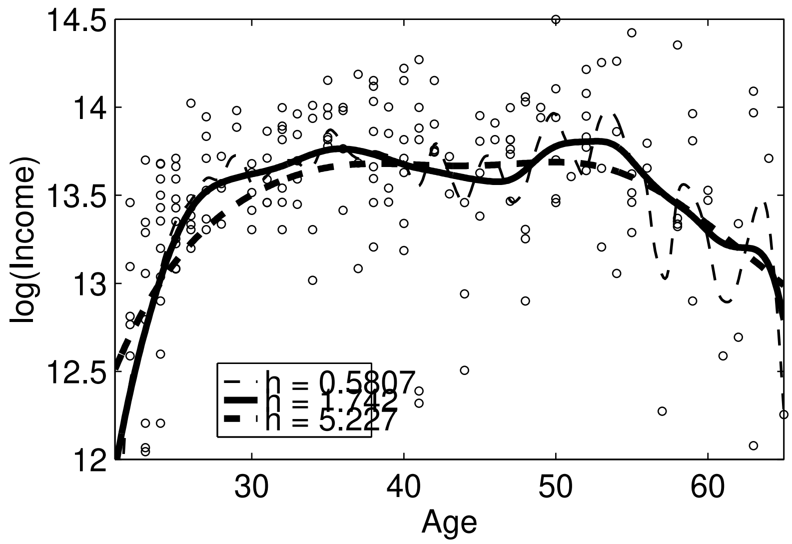

Figure 1: Scatterplot smooths, with different window widths h, for the Canadian Earning Power data (shown as small circles).These show that observed features depend on the level of smoothing used, and motivate the question of which observed features are "really there?".

Other author information: (Send correspondence to J.S.M.)
J.S.M.: E-mail: marron@stat.unc.edu
P.C.: E-mail: probal@isical.ernet.in

The Canadian Earning Power data shown in Figure 1 comes from Ullah (1985). This was from a survey done in Canada, and shows income as a function of age at a particular time. The large window width shows the expected features: young people earn significantly less than others but can expect substantial increases, and as older people leave the work force their earning power declines. The small window width smooth is too strongly affected by the sampling noise, and has many features that are clearly not important underlying structure. But the medium smooth shows features that are uncertain, in particular a marked dip in the middle. If this is an important underlying feature, then a surprising discovery has been made, that is well worth deeper investigation. However, if the dip is just a spurious sampling artifact, then such deeper investigation would be wasted effort.

This example illustrates that for practical use of smoothing methods, a critical question is "which observed features are important underlying structure, and which are spurious sampling artifacts?". SiZer provides an answer to this question, based on studying SIgnificance of ZERo crossings of derivatives of a family of smooths. This method was developed in Chaudhuri and Marron (1997), and is discussed here in Section 2. Illustrations and examples are given in Section 3. Conclusions, and suggestions for future work are given in Section 4

2. THE SIZER APPROACH

SiZer was developed in Chaudhuri and Marron (1997) as a first attempt at a statistical application of "scale space" ideas from computer vision. An introduction to scale space ideas, from a statistician's viewpoint, is given in Chaudhuri and Marron (1998). That paper discusses a number of interesting ideas from the monograph Lindeberg (1994), where a mathematical mdoel for "vision" is developed, which is simply a family of Gaussian kernel smooths indexed by the bandwidth. The idea is that large bandwidths represent distant, i.e. "blurred" views, and small bandwidths represent close in, i.e. "zoomed" views.

One very important scale space idea is that one should consider not a single smooth, but in fact a wide range of different levels of smoothing. This idea is not new to many statisticians, who have been aware that different information can be available from different levels of smoothing. See Marron and Chung (1997) for some specific proposals concerning use of a "family of smooths" for data analysis.

A second critical concept of scale space, is the idea that one should no longer think of "estimating a true underlying curve", and should instead focus on "aspects of the true curve which are recoverable from the data". In particular, the target of the estimation process becomes a family of convolutions of the true curve with the window function. These convolutions are interpreted as reflecting "how much of the true curve is available at that level of resolution".

SiZer has been developed for both density estimation, i.e. smoothed histograms, and for nonparametric regression, i.e. scatterplot smoothing. Kernel density estimation reveals structure in a univariate set of data, $X_1, ..., X_n$, by putting some probability mass "near" each observation:

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

where K_h is the "h-rescaling" of the kernel function K, $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. In the examples in this paper, K is the standard Gaussian density $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, so K_h is the Gaussian density with standard deviation h. See Chaudhuri and Marron (1998) for detailed reasons behind this choice. Kernel density estimation is illustrated in figure 2. Nonparametric regression is for the study of dependence of variables, and in particular focuses on the conditional mean $E\left(Y|X = x\right)$ based on pairs of variables $\left(X_1, Y_1\right), ... \left(X_n, Y_n\right)$. There are a number of popular methods for smoothing such data, but here we study the local linear method, see e.g. Fan and Gijbels (1996) for reasons for this choice,

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

where $\arg\min_a$ 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.

The first component of the SiZer methodology used in all the examples below is a family of smooths, for a very broad range of bandwidths, h. Suggested defaults are: evaluate the smooths at a grid of 401 equally spaced points (this gives smooth curves in most instances), with the smallest h equal to twice the grid spacing, and the largest h

equal to the range. To give some frame of reference, it is often useful to highlight a good data based choice of h. Here this is done using the Sheather Jones Plug In bandwidth, see e.g. Jones, Marron and Sheather (1996a,b), in density estimation, and the Ruppert Sheather and Wand (1995) method in regression.

SiZer investigates which of the features seen in the smooths are statistically significant by studying derivatives (in the x direction) of the smooths. When a confidence interval (that is simultaneous in a sense described in Chaudhuri and Marron (1997)) for the derivative has both endpoints above 0, the estimate has significant positive slope, and vice versa for significantly negative slope. This information is displayed in the second component of SiZer, a color map of scale space, where each pixel represents a "location" with respect to *both* position x and bandwidth h. Each pixel is shaded dark where there is a significant increase, light where there is a significant decrease, and an intermediate gray where there is no significant change. In some scale space locations, there is not enough data present (taken here to mean the effective number of points in each window is less than 5) so a slightly darker shade of gray is used.

Publication constraints make the use of color too expensive in the present paper, but the full color version of SiZer is recommended. This replaces the dark regions by blue, the light regions by red, and the intermediate gray with purple. S. G. Eick summed up the situation well by pointing out that "the color version of SiZer is much more engaging."

Because the family approach and SiZer are based on replicated smoothing, a fast numerical implementation can be quite important. Because of the replication, the "binning" method discussed in Fan and Marron (1994) is very appropriate, and as noted there this can make the difference between obtaining a SiZer map in a few seconds, as opposed to several minutes for a naive direct implementation. Matlab versions of SiZer, and electronic copies of Chaudhuri and Marron (1997,1998), are available on the World Wide Web at the URL:

`http://hotelling.stat.unc.edu/faculty/marron/marron_software.html`

For a quick first try of SiZer, use `gpanal.m`.

3. EXAMPLES

3.1. Simulations

Figure 2 shows simulated density estimation, and illustrates how the information that is available in the data increases with the sample size. The underlying density in each case is the #10 Claw normal mixture density from Marron and Wand (1972). This density has 5 thin peaks, which should be quite difficult to find with small sample sizes.

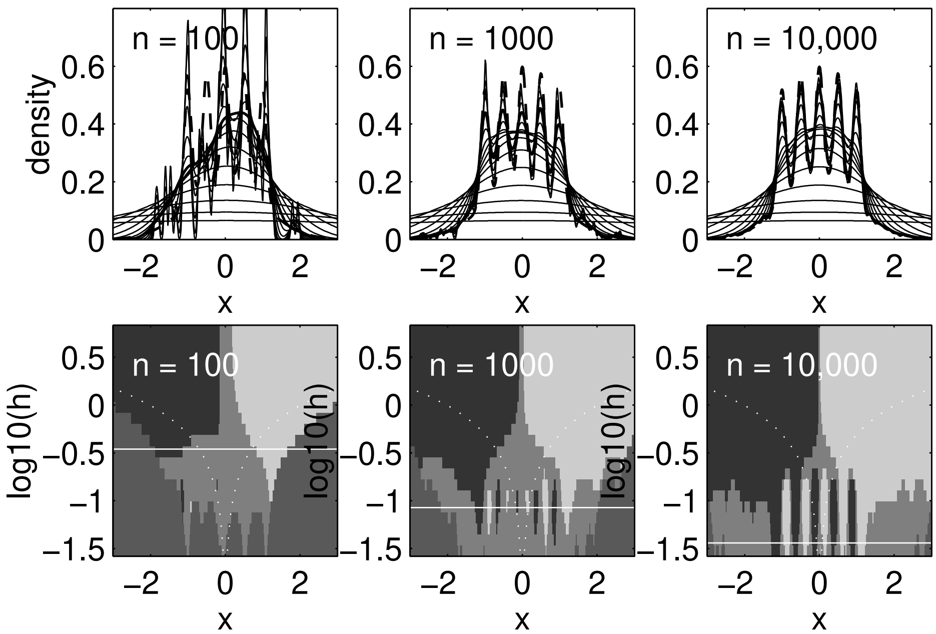

Figure 2: Simulated kernel density estimation, from the underlying density shown as the thick dashed curve, for sample sizes $n = 100$, 1000, $10,000$. Families of smooths are shown in the top panels, and corresponding SiZer maps are shown in the bottom panels. The horizontal white bar in the SiZer map shows the Sheather Jones Plug In data driven bandwidth, which is shown as the thicker solid line type in the family plots. The white dots in the SiZer maps show $\pm 2h$ to give a feeling for the effective window width of K_h. This figure shows how the number of features of the underlying density, which can be resolved as "significant", increases with the sample size.

The family in the top left part of Figure 2, for $n = 100$, shows a wide range of smoothing, from grossly oversmoothed (with most of the probability mass spread outside the range under consideration), to very undersmoothed (with peaks that reach far higher than the underlying density). The corresponding SiZer map (bottom left) shows that for large bandwidths, the estimates slope significantly upward on the left, and downward on the right. For smaller bandwidths, there is a central gray area, which shows that the wiggles in the estimate at that level of resolution can not be separated from spurious sampling noise. Even though many of the wiggles are actually caused by the shape of the underlying density, there is not enough information available to determine this from the background noise. The darker shade of gray in the SiZer map shows regions where there is not enough data to construct meaningful confidence intervals of the type underlying SiZer, at that level of resolution.

The family plot in the center top of Figure 2, for $n = 1000$, is similar to the $n = 100$ case for large bandwidths, but less wiggly for smaller bandwidths, since more data tends to stabilize the wiggliness. More members of the family are now doing a better job of resolving the 5 think peaks, and the SiZer map shows that now these are significant features, since 5 dark regions, followed by 5 light regions are visible for h near the data driven bandwidth shown at the horizontal white bar. Once again for large bandwidths, there is only a single increase, followed by a single decrease, which reflects what can be seen in the underlying density *at this level of resolution.*

For $n = 10,000$, there is very abundant information available in the data, so now the 5 peaks stand out quite clearly in both the family plot, and also the SiZer map. Note that the data driven bandwidth (thicker curve in the family plot, and white horizontal bar in the SiZer map) is now quite small, since the increased sample size means the optimal trade-off between variance and bias happens for a small window width.

Figure 3 shows a simulated regression example. The underlying target curve is the sum of a line and a Gaussian density, together with a jump discontinuity. Gaussian noise has been added at an equally spaced rid of $n = 1024$ design points.

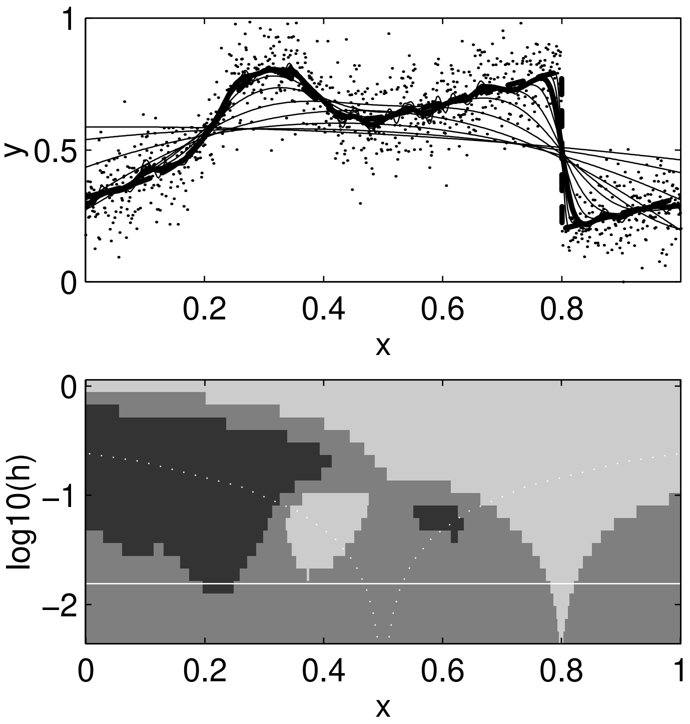

Figure 3: Family and SiZer analysis for a simulated regression example. Target curve is shown as the thick dashed line. Data points are shown as small dots. This shows how the jump appears as a funnel shaped shaded region in the SiZer map.

Although the noise level is relatively high, the large number of data points means that there is enough information to discern that both the smooth peak around $x = 0.3$ (the gray between the large dark region and the small light region) and the smooth valley around $x = 0.5$ (the gray between the small light region and the even smaller dark region) are significant features. Especially interesting is the shape of the light shaded region near the jump at $x = 0.8$. The funnel shape of the region is caused by this jump being a region of significant decrease, even at the finest levels of resolution of the data (i.e. smallest bandwidths). This shape has appeared in a number of other examples with jumps, and seems to be characteristic of jumps in regression curves that are large with respect to the noise level. This shows that even though SiZer is based on smoothing ideas, it can be quite effective at finding points where the underlying curve is not smooth.

3.2. The initial example

Now we return to the Canadian Earning Power Data of Figure 1.

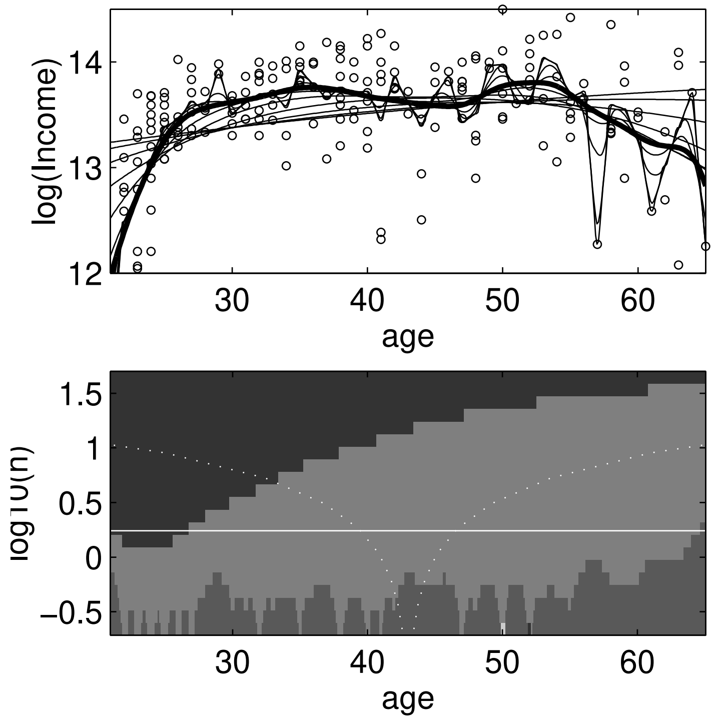

Figure 4: Family and SiZer analysis of the Earning Power Data from Figure 1. This shows that the only significant feature is the large scale increase. The middle age dip apparent in Figure 1 can not be distinguished from the sampling variability.

Figure 4 suggests that further investigation of the middle age dip discussed in section 1 may not be worthwhile, since there is no strong evidence that this is anything besides the level of variability expected from the sampling process. Of course this is not conclusive evidence against such a dip, but only a statement that more evidence is needed before it can be concluded to be an important part of the underlying structure. Perhaps more surprising is that the decrease that seems apparent for older ages is not recognized as significant by SiZer (since there is no light region on the right hand side). This can be explained by the relatively small sample size of $n = 205$, and by the increasing heteroscedasticity, making it more difficult to make sharp inference on the right side of the plot.

4. SECOND GENERATION THEORY FOR SMOOTHING

Note that SiZer represents a radically different approach to smoothing. Most of the statistical theory of smoothing concerns the question:

- What are good curve estimates? How good are they?

There is a very large literature addressing this issue from many points of view, such as asymptotic rates of convergence (and deeper asymptotics involving constant coefficients and minimax lower bounds), simulation studies, and exact risk analysis. However, "good" estimates of this type are typically those that optimize some type of error criterion. Unfortunately such optimal smooths will typically be quite different from smooths with "the right features", e.g. optimal smooths will often have spurious "bumps" that are accepted as simply being part of the optimal trade-off present in bandwidth choice.

For this reason, we believe that it is time to shift the bulk of research efforts away from the above *first generation* research question, which has now been answered quite successfully, towards an emphasis on the *second generation* question:

- Which features in a smooth are "really there"?

This question is not completely unexplored, e.g. there is some literature on "bump hunting", see for example Good and Gaskins (1980), Hartigan and Hartigan (1985), Müller and Sawitzki (1991), Hartigan and Mohanty (1992), Mammen, Marron and Fisher (1992), Minnotte and Scott (1993), Fisher, Mammen, and Marron (1994), Minnotte (1997) Cheng and Hall (1997) and Mammen, Marron and Udina (1997).

We hope that SiZer represents a reasonable first attempt at finding other ways to tackle this second generation research question, and that many others will be developed in the future.

REFERENCES

1. Bowman, A. W. and Azzelini, A (1997) *Applied Smoothing Techniques for Data Analysis, the Kernel Approach with S-plus Illustrations*, Oxford Statistical Science Series 18, Oxford Science Publications, Oxford.
2. Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, second edition, Springer Verlag, New York.
3. Chaudhuri, P. and Marron, J. S. (1998) SiZer for exploration of structures in curves, North Carolina Institute of Statistics Mimeo Series # ????.
4. Chaudhuri, P. and Marron, J. S. (1997) Scale space view of curve estimation, North Carolina Institute of Statistics Mimeo Series # ????.
5. Cheng M. Y. and Hall, P. (1997) Calibration methods for excess mass and DIP tests of homogeneity, unpublished manuscript.
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. 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.
10. Green and Silverman (1994) *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.
11. Hart, J. D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.
12. Hartigan, J. A. and Hartigan, P. M. (1985) The DIP test of multimodality, *Annals of Statistics*, 13, 70-84.
13. Hartigan, J. A. and Mohanty, S. (1992) The RUNT test from multimodality, *J. Classification*, 9, 63-70.
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. Lindeberg, T. (1994) *Scale space theory in computer vision.* Kluwer: Boston.
17. 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.
18. Mammen, E., Marron, J. S. and Udina, F. (1997) Interactive mode testing, unpublished manuscript.
19. Marron, J. S. and Chung, S. S. (1997) Presentation of smoothers: the family approach, unpublished manuscript.
20. Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Annals of Statistics*, 20, 712-736.
21. Minnotte, M. C. (1997) Nonparametric testing of the existence of modes, *Annals of Statistics*, 25, 1646-1660.
22. 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.
23. Müller, D. W. and Sawitzki, G. (1991) Excess mass estimates and tests for multimodality, *Journal of the American Statistical Association* , 86, 738-746.
24. 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.
25. Simonoff, J. S. (1996) *Smoothing methods in statistics*, Springer, New York.
26. Ullah, A. (1985) Specification analysis of econometric models. *Journal of Quantitative Economics*, 2, 348-354.
27. Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall: London.