# Significance of Features via SiZer

J. S. Marron and P. Chaudhuri

April 28, 1998

**Abstract**

SiZer is an exploratory data analysis tool, that works in conjuction with smoothing methods. It addresses the issue that is often central to the use of smoothing in data analysis: which observed features are "really there", and which cannot be distinguished from the background sampling variability? The SiZer approach makes heavy use of scale space ideas.

Research partially supported by the National Science Foundation, and by the Indian Statistical Institute.

## 1. Introduction

Smoothing methods have shown themselves to be quite useful tools for exploratory data analysis. Monographs on smoothing in the last five years include Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Hart (1997) and Bowman and Azzalini (1997). These contain useful introductory material, and also access to the large early literature on smoothing.

When analyzing data by these methods one often finds "features", e.g. bumps or valleys, which may be important underlying structure, or may be spurious artifacts of the sampling process. Determining which is quite challenging, since both types can be made to disappear by increasing the amount of smoothing done. Conversely, sufficiently decreasing the amount of smoothing can make many spurious features appear.

An interesting approach to this problem is "mode testing", i.e. "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), Fisher, Mammen and Marron (1994), Minnotte (1997) and Cheng and Hall (1997) for a number of interesting ideas in this direction, and access to other work. This type of formal mode testing is quite useful when one is investigating the validity of a prior idea about the number of modes, but is less well adapted to exploratory work (because of multiple comparison problems in interpreting the result). Also most such mode tests only study the number of modes, without regard to their location.

For exploratory data analysis desirable properties of a bump hunting method include no prior assumptions on the number of modes, and an indication of the location of "features that are really present". The mode tree of Minnotte and Scott (1993) is a first such exploratory tool.

In this paper, we discuss a tool which can be viewed as an extension of the mode tree. There are two graphical components, as indicated in Figure 1. The first is an overlaid family of smooths, indexed by the smoothing parameter, and the second is called a SiZer map. Both of these visualizations are based on the "scale space" view of smoothing, see Lindeberg (1994) for general details, and Chaudhuri and Marron (1998) for discussion of several statistical aspects.

The idea of the family plot is that no single smooth can contain all of the information available in the data, and indeed several smooths are needed, since different types of features will be available at different "smoothing levels". The overlay of the family attempts to display as many of these features as possible. Specific guidelines for choice of family (e.g. the number of family members, and the endpoints of the smoothing range) have been suggested in Marron and Chung (1997).

Figure 1 is an analysis of the Old Faithful Geyser data that have been widely studied. See Section 1.4.1 of Scott (1992) for detailed description, and useful insights. Here the smooths are kernel density estimates, using a Gaussian kernel (see e.g. Section 2.1 of Wand and Jones (1995) for an introduction). Note that in the top of Figure 1, two strong modes seem to be a clear feature of the data, and the other much smaller bumps are likely to be just sampling noise. The thick curve in the family is the Sheather Jones Plug In bandwidth choice, see e.g. Jones, Marron and Sheather (1995) for further discussion.

The lower visualization is a SiZer map, as developed by Chaudhuri and Marron (1997), which shows which features visible in the family of smooths in the top panel are "statistically significant", by studying SIgnificance of ZERo crossings. Each pixel in the gray level map represents a point in scale space, indexed by location (on the horizontal axis) and smoothing parameter (i.e. the bandwidth $h$ for kernel density estimation, on the vertical axis). For each location and each bandwidth, a simple confidence interval for the derivative is constructed (thus studying the local slope of the smooth). When this confidence interval is completely above zero, the smooth is "significantly increasing", and the pixel is shaded dark (blue is used in full color versions). When this confidence interval is completely below zero, the smooth is "significantly decreasing", and the pixel is shaded light (red in full color versions). When the confidence interval contains zero, there is no statistically significant slope, so the pixel is shaded an intermediate gray (purple in full color versions). Thus modes that are "really there" have a dark region on the left, and a light region on the right, as seen around $x = 2$ and $x = 5$ in Figure 1. Similarly "significant valleys" have the opposite coloring, as seen around $x = 3$ in Figure 1. The other peaks and valleys apparent in the family of smooths all appear in the gray region of the SiZer map, which indicates these modes are not strong enough to be distinguished form the background noise. This exploratory conclusion is in keeping with that from the deeper analysis of Sections 1.4.1 and 8.3.2 of Scott (1992).

Note that these two modes are not seen at all "levels of resolution", i.e. at all bandwidths. When the smoothing window is very large, as at the top quarter of the SiZer map, the two modes have been smoothed down to one (the shading goes from light to medium to dark), because a single bump is the only feature in the data *at this level of resolution*, as shown in the above family plot. For more moderate bandwidths (middle third of the SiZer map), the two bumps are seen to be important features of this data set *at this level of resolution* as noted above. At finer resolution scales, i.e. smaller bandwidths, (bottom part of the SiZer map), the dominant color is gray, which means that the small spikes visible only at these levels of resolution in the family plot are not statistically significant.

The multiple comparison problem that comes from studying a large number of confidence intervals is addressed using methods described in Chaudhuri and Marron (1997). Since these confidence intervals are based on asymptotics, they become unreliable when here are not enough points in each window, so a darker shade of gray is used to indicate such locations (visible in the lower center, and lower left and right of the SiZer map in Figure 1). The dashed curves overlayed on the SiZer map show the effective amount of smoothing in each row, by showing plus and minus two standard deviations of the Gaussian window function. The horizontal bar shows where the thick curve in the family appears on the SiZer map.

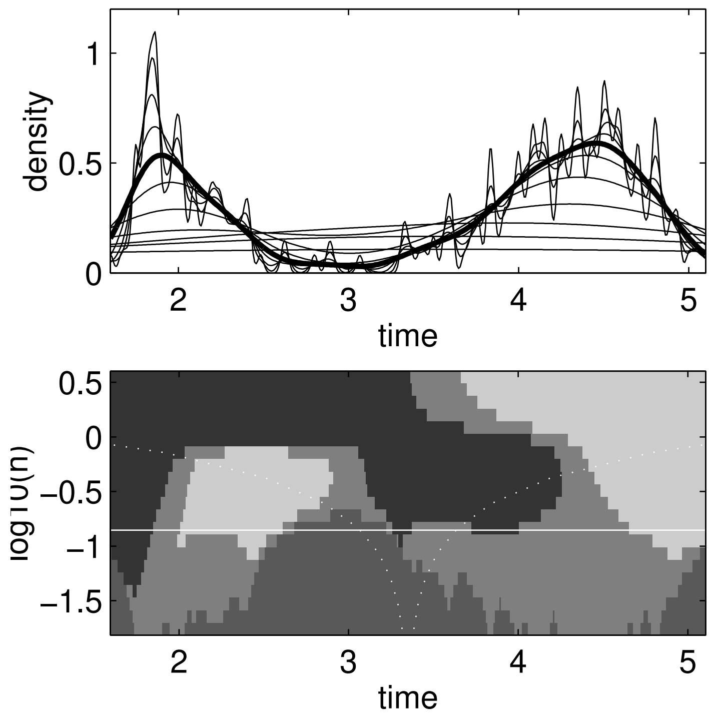


FIGURE 1: *Family approach (top panel), and SiZer map for the Yellowstone Old Faithful Geyser data. This indicates that the two strong modes at $x = 2$ and $x = 4.5$ are important underlying features, but the other modes can not be distinguished from spurious sampling fluctuations.*

Since the family approach and SiZer are based on repeated iterations of smoothing, fast computational methods are important for implementation. See Fan and Marron (1994) for discussion of some non-obvious methods for fast smoothing.

The remaining sections show examples of various types. Section 2 studies a simulated example to understand the effects of increasing sample size, in particular demonstrating how increased information in the data results in being able to distinguish more features of the true underlying curve from the background noise. Section 3 is a particularly challenging real data set, which shows some unusual features. Section 4 shows a simple use of SiZer in a time series context, with the additional lesson that use of SiZer can sometimes be enhanced by pre-processing.

In particular, because SiZer is based on sign changes in derivatives, it can be useful to subtract a linear fit from the data before applying SiZer.

## 2. A simulation example

This section demonstrates how SiZer quantifies the idea of "more information in the data mean more features are statistically significant". The setting is kernel density estimation, and the underlying density is the #13 Asymmetric Double Claw Density from Marron and Wand (1982). The true density has two broad modes at $x = \pm 1$, three quite thin modes at $x = 0.5$, 1.0 and 1.5, and three extremely thin modes at $x = -1.5, -1.0, -0.5$. Figure 2 shows the family approach and SiZer for simulated data sets with sample sizes $n = 100$, 1000 and $10,000$.

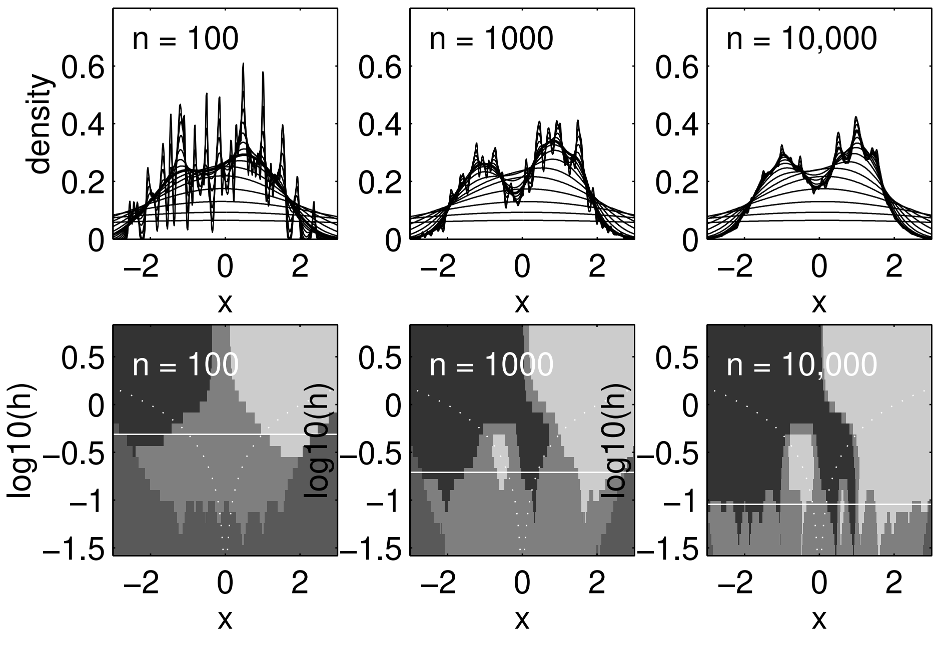


FIGURE 2: *Family approach (top panels), and SiZer maps for simulated data sets from Marron and Wand density # 13, for sample sizes $n = 100$ (left) $n = 1000$ (center), $n = 10,000$ (right). This shows how more information in the data results in more significant features.*

Note that when there only $n = 100$ observations available, the SiZer map show only one significant mode, i.e. there is not enough information in the data to even find the two large modes. The family plot suggests there may be two modes, but note that the background noise visually apparent in the family plot is quite high, so it is not surprising that the two modes are not shown to be significant.

When the sample size is increased substantially, to $n = 1000$, there is now enough information in the data to clearly resolve the two main modes. The three thinner modes on the right are also beginning to be visible in the family plot, although the SiZer map shows these are not clearly distinguishable from the sampling variability.

Another order of magnitude increase in the sample size, to $n = 10,000$, indicates that not only are the two main modes significant, but so are the three quite thin modes at $x = 0.5$, $1.0$ and $1.5$. The first two are highlighted by both dark and light regions, and the third has a light region to the right but no corresponding dark region on the left, which means that the upslope on the left is not statistically significant, but the down slope on the right is. There are some hints at the very thin modes on the left in the family plot, but the SiZer map shows there are not strong enough to be called "really there".

## 3. A challenging real data example

Figure 3 shows a family plus SiZer analysis of a challenging real data set, whose source is confidential. The goal was to find the number of significant clusters in the data. The family plot only shows a very heavily skewed distribution, with an obvious very large mode at the extreme left. But the SiZer map reveals additional significant structure. The mode at the left is indeed very significant, but there is also a region of significant increase, shown as a dark blob, around $x = 8$, and a "sudden drop off", near $x = 33$.

A surprising feature of the SiZer map is the "fringe of dark and light shadings" at the smallest bandwidth, i.e. the finest level of resolution. A careful look at a listing of the data showed that there were many replications among the data (i.e. most numbers that were in the set appeared a number of times), although it was not simple data rounding. This fringe is present because these replications are sufficiently numerous that *at this scale of resolution* then when the data are viewed as coming from a continuous probability density (a basic assumption of SiZer), this density has these many small but significant bumps. While this particular structure was not useful (it turned out the data had first been rather coarsely rounded, and then transformed) because the interesting analysis was at a much different level of resolution, this still shows how SiZer can find structure at wide range of granularities.

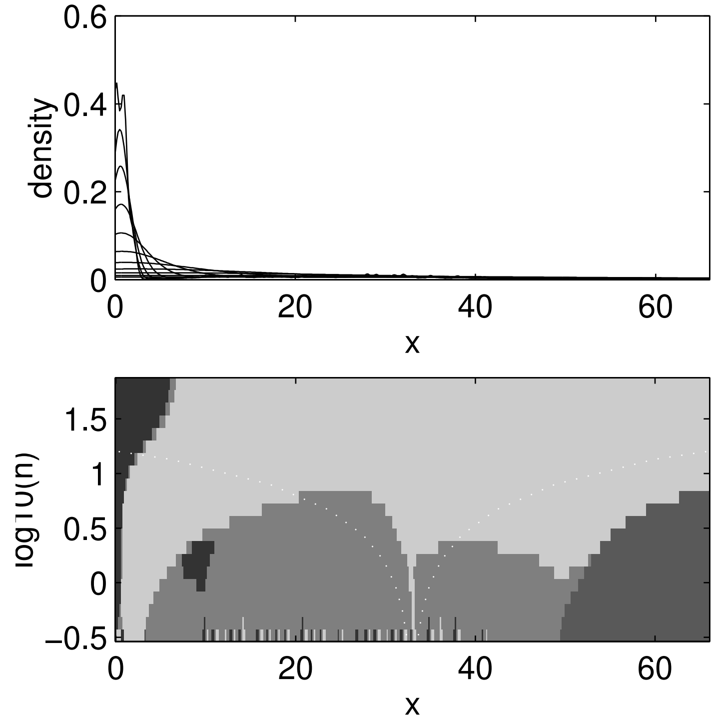


FIGURE 3: *Family approach (top), and SiZer map (bottom) for a real data set. This shows how SiZer can reveal the presence of rounding/prebinning in data.*

In the absence of other information, it would be natural to analyze these data on the log scale, however, it needs to be kept in mind that such transformations have a substantial effect, both on SiZer and even on the modality of the true underlying density.

## 4. A time series example

In this section an example from time series is considered. Here the smoothing is done on a scatterplot of values, by the local linear smoother (see e.g. Section 2.3.1 of Fan and Gijbels (1996) for good discussion), using a Gaussian kernel.

Figure 4 shows the use of the family approach and SiZer for the "monthly Australian Chocolate Consumption (in kilotonnes, for the late 1950's to early 1990's)" data set that comes with the software companion to Brockwell and Davis (1991). Smoothing can be useful for understanding trends in time series, and SiZer is suggested here as a method for understanding which trends are "really there", and which are just artifacts of the sampling variability. The raw data have an obvious annual (i.e. lag 12) component, which has already been removed in the examples considered here.

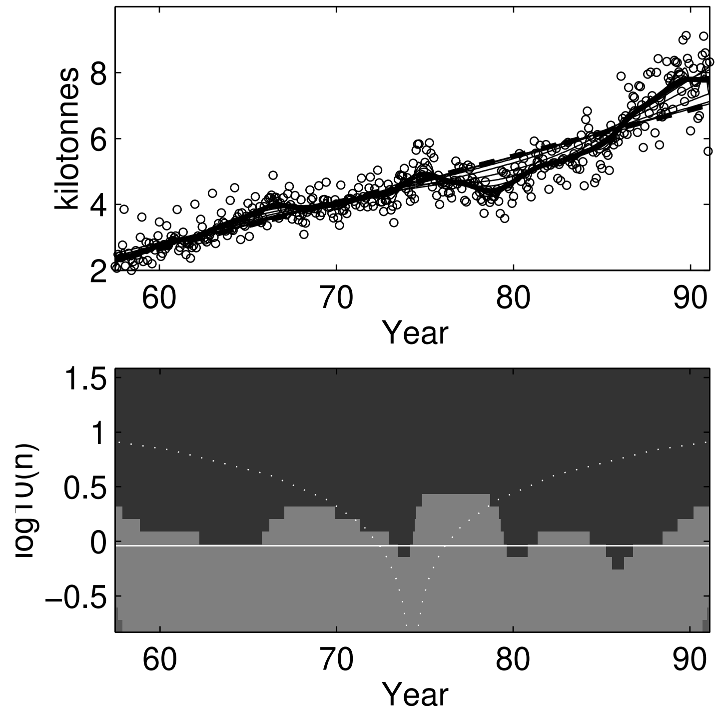


FIGURE 4: *Family and SiZer analysis of trends in the deseasonalized Chocolate data. Only the obvious upward linear trend is flagged as significant by SiZer. Simple linear regression line is shown as the heavy dashed line.*

The family plot shows a strong upward trend, driven by increasing consumption due to increasing population size. The family of smooths also suggests some potential nonlinear trends, however, these do not show up as significant in SiZer.

A potential reason for this is that SiZer works in terms of slopes, and the large upward sloping linear component can drown out other relatively smaller changes in slope.

To investigate this possible "masking effect", Figure 5 shows the result of applying the family approach and SiZer to the data in Figure 4, after the least squares linear fit has been subtracted. Now the family of smooths suggests some rather strong non-stationary trends in the mean of this time series. In particular, the dip around the year 78 is shown to be significant by the light region to the left, and the dark region to the right.

The light to gray to dark regions near the top of the SiZer map suggest some "coarse resolution shift in the trend", which could be a result of the gross scale trend being not linear but exponential. This would be intuitively sensible if this trend is driven by population growth. It may be possible to find additional significant small scale trends, by replacing the initial linear fit by an exponential, or by working with the logarithm of the data.

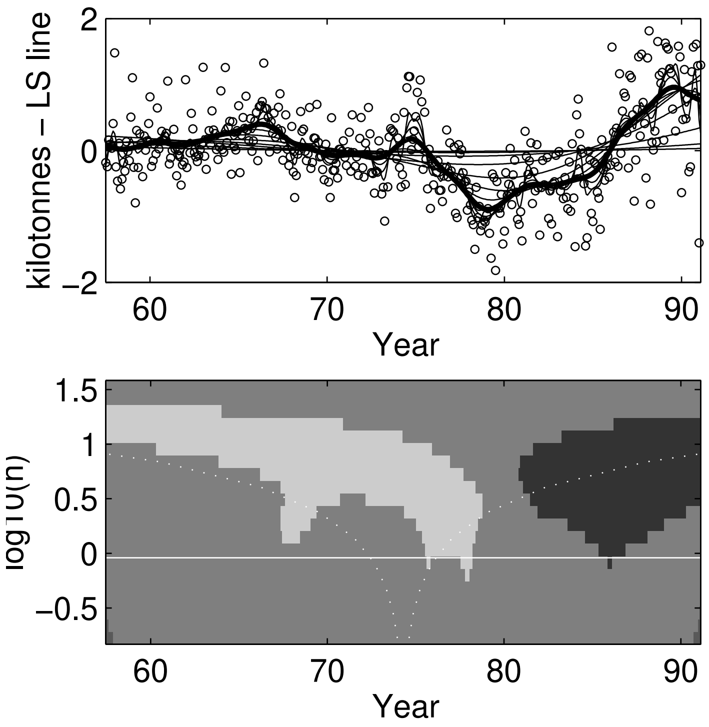


FIGURE 5: *Family and SiZer analysis of trends in the deseasonal-*

*ized, and linearly detrended Chocolate data. This shows that additional trends are significant, that were masked by the strong upward component in Figure 4.*

In time series contexts, the lessons of SiZer need to be taken with a grain of salt, because there is an underlying assumption of a smooth mean curve, with independent and identically distributed error, which is often not a good model in time series contexts. An interesting open problem is the development of a time series version of SiZer which correctly uses correlation information in its inference (i.e. the underlying confidence intervals are adjusted according to the correlation present).

Another interesting potential time series application of SiZer is to spectral density estimation. This is usually accomplished by smoothing the periodogram, and the same questions of which features observed in the periodogram also arise. In this context applications of SiZer will be straightforward, since for stationary time series, the periodogram values are uncorrelated.

## 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. (1997) SiZer for exploration of structures in curves, North Carolina Institute of Statistics Mimeo Series # 2355.

[4] Chaudhuri, P. and Marron, J. S. (1998) Scale space view of curve estimation, North Carolina Institute of Statistics Mimeo Series # 2356.

[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] Lindeberg, T. (1994) *Scale space theory in computer vision.* Kluwer: Boston.

[16] 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.

[17] Mammen, E., Marron, J. S. and Udina, F. (1997) Interactive mode testing, unpublished manuscript.

[18] Marron, J. S. and Chung, S. S. (1997) Presentation of smoothers: the family approach, unpublished manuscript.

[19] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Annals of Statistics*, 20, 712-736.

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

[21] 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.

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

[23] Simonoff, J. S. (1996) *Smoothing methods in statistics*, Springer, New York.

[24] Scott, D. W. (1992) *Multivariate density estimation: Theory Practice and Visualization*, Wiley, New York.

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