SiZer for jump detection

C. S. Kim
Department of Computer Science and Statistics
Cheju National University
Ara-Dong , Cheju City, 690-756
Korea

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
USA

January 15, 1999

Abstract

SiZer is an exploratory graphical method for finding structure in data. When the structure is a jump in the underlying curve, a "jump funnel" is created in the SiZer map. The shape of this funnel is analyzed. The result is the basis of a proposed variation of SiZer that is specially tuned to finding jumps.

Acknowledgement 1 *Research of C. S. Kim was supported by a Cheju National University Research Grant, 1997.*

1 Introduction

SiZer was introduced by Chaudhuri and Marron (1999), for finding structure in smooths of data. This method is based on "scale space" ideas, as discussed in Chaudhuri and Marron (1998). Scale space is a family of Gaussian smooths indexed by the bandwidth, shown as blue curves in the top panel of Figure 1, based on data shown as green dots. SiZer studies SIgnificance of ZERo crossings of the derivative of the smooths in scale space, as shown in the lower panel of Figure 1. It represents regions, with respect to both location and scale (i.e. bandwidth) with colors. Red is used where the smooth is significantly decreasing, blue is used for significantly increasing, and the intermediate color of purple is used for no significant slope (i.e. a confidence interval for the slope contains the origin). The SiZer map shows that the increases on the left side, at large bandwidths, are statistically significant, as are the decreases on the right

side, although only for relatively smaller bandwidths. The solid white horizontal bar in the SiZer map shows a good data driven bandwidth, as suggested by Ruppert, Sheather and Wand (1995). The smooth with this bandwidth is represented as the thick red curve in the top panel. The dotted white curves give an idea about "effective window widths", by showing $\pm 2h$, where h is the bandwidth. The dashed horizontal bar in the SiZer map is discussed in Section 3. See Marron and Chaudhuri (1998a,b) for additional examples and insights about SiZer.

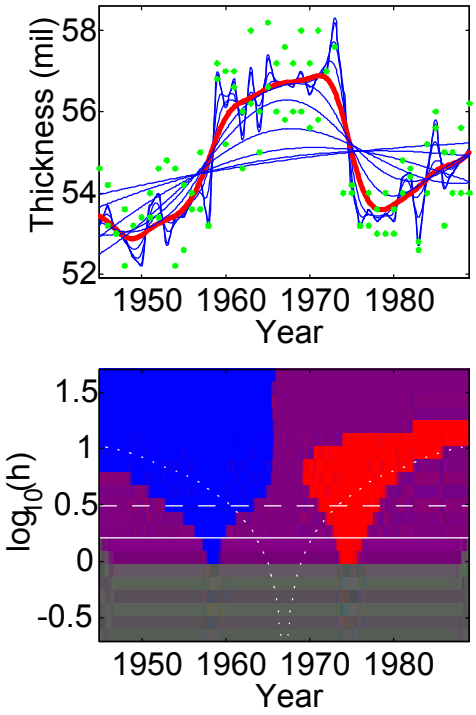

FIGURE 1: *Family of smooths (blue curves, top panel) and SiZer map (bottom panel) for the Penny Thickness data (green dots, top panel). Funnels highlight two jumps.*

The data in Figure 1 come from Table B.4 of Scott (1992), and are nicely analyzed in Section 8.3.2 of that book. The data are thicknesses (in mil) of pennies, two measurements made for each year from 1945 to 1989. Scott comments that there are known to be two jumps in the mean of this regression, when the manufacturing process was changed. First in 1958, where the penny was returned to its pre-war thickness after having been thinned for World War II. Second in 1974, when it was thinned again. Scott uses the nice idea of "modal regression" to show these features of the data.

The SiZer map in Figure 1 clearly shows these same two jumps. They appear as "funnel shapes", caused by the slopes being significant, even for very small bandwidths. This was discussed at the end of Section 4 of Chaudhuri

and Marron (1999) and in Section 3.1 of Marron and Chaudhuri (1998b). It was remarked that this shape appears quite often when there are jumps in the true underlying regression. In this paper, we analyze that characteristic funnel shape, called the "jump funnel", and propose a new visual device for using this shape to learn about potential jumps in the regression curve.

In Section 2, we present some mathematical analysis, which reveals the shape of the region caused by the underlying jumps. In particular, the boundary of the jump funnels is seen to grow linearly with the bandwidth. In Section 3 this mathematical insight is used to motivate a modification of the SiZer map, which is tuned to visually highlight jumps, by making the funnel boundaries appear as straight lines. This is illustrated with a modified SiZer map for the penny data of Figure 1. Some additional simulated examples are discussed in Section 4.

2 Mathematical Analysis

Jump funnels, are most easily studied in the case of regression with an equally spaced design, i.e. the data are (perhaps rescaled to be) of the form $\left\{\left(\frac{i}{n}, Y_i\right) : i = 1, ..., n\right\}$. SiZer for nonparametric regression, is based on the local linear smoother, where one fits a line, say having equation $y = mx + b$, to the scatterplot of the data, using weighted least squares, in a moving window. In particular, at each location $x \in [0, 1]$, a slope estimate $\widehat{m}_h(x)$ and an intercept estimate $\widehat{b}_h(x)$ are found by minimizing, over m and b, the kernel weighted least squares criterion

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

where $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, for a kernel function K controlling the shape of the local window around x, and a bandwidth h controlling the width of the window. SiZer uses the standard Gaussian density for K, for scale space reasons as discussed in Chaudhuri and Marron (1998). However, for ease of illustration, we consider here the uniform kernel. The "moving intercept" curve, $\widehat{b}_h(x)$, provides a reasonable estimate of the underlying regression curve, see for example Wand and Jones (1995) and Fan and Gijbels (1996) for good discussion of the properties of this smoothing method. SiZer is based on the slope estimate $\widehat{m}_h(x)$, and whether or not it is significantly positive or negative. This significance is determined by its normalized version

$$\frac{\widehat{m}_h(x)}{\widehat{\sigma}_m(x)},$$

where $\widehat{\sigma}_m$ is the usual (locally weighted) estimate of the standard deviation of $\widehat{m}$. SiZer uses the color blue (red) when $\frac{\widehat{m}_h(x)}{\widehat{\sigma}_m(x)}$ is larger than (smaller than, resp.) q, an appropriate quantile to the Gaussian distribution (where q is chosen

to give simultaneous inference). Thus the boundary of the jump funnels, is

$$B = \left\{ (x,h) : \frac{\widehat{m}_h(x)}{\widehat{\sigma}_m(x)} = q \right\}.$$

Simplest analysis of the jump funnels comes from working with data which are constant except for a single jump. In particular, consider

$$Y_i = \begin{cases} 0 & i = 1, ..., j \\ 1 & i = j+1, ..., n \end{cases},$$

where it is enough to keep the constant height at 1, since $\frac{\widehat{m}}{\widehat{\sigma}_m}$ is vertically scale free. For the moment, ignore the kernel weights, i.e. take $K_h \equiv 1$. Straightforward calculation, using the standard formulas for simple linear regression gives

$$\frac{\widehat{m}_h}{\widehat{\sigma}_m} = \left(\frac{3j(n-j)(n-2)}{n^2 - 1 - 3jn + 3j^2} \right)^{1/2} \sim (3j)^{1/2}, \tag{1}$$

in the limit as $n \to \infty$, with $\frac{j}{n} \to 0$.

Now to use (1) for analysis of SiZer, consider a uniform window local linear estimate, with bandwidth h, centered at the point x_w and suppose the jump occurs at a point x_j. When the centerpoint x_w is far enough from the boundary (in particular, $x_w > h$ and $x_w < 1 - h$) the number of points in each window is $n_w \cong 2nh$ (where $\cong$ is used since appropriate rounding may be needed to make $2nh$ and integer). When the jump point x_j is inside the window, and to the left of x_w (i.e. $x_j \in (x_w - h, x_w)$), then the above calculations apply with the changes of variable:

$$\begin{aligned} n &\leftrightarrow n_w, \\ j &\leftrightarrow n\left(\frac{1}{2} + \frac{x_j - x_w}{2h} \right). \end{aligned}$$

This gives the approximation:

$$\begin{aligned} B_R &\approx \left\{ (x_w, h) : n\left(\frac{1}{2} + \frac{x_j - x_w}{2h} \right) = \frac{q^2}{3} \right\} \\ &\approx \left\{ (x_w, h) : x_w = x_j - 2h\left(\frac{q^2}{3n} - \frac{1}{2} \right) \right\}, \end{aligned}$$

for the boundary of the funnel on the right side of the jump (i.e. $x_j < x_w$). For the left side of the jump, i.e. $x_w < x_j$, similar calculations show that the right boundary of the funnel is approximated as

$$B_L \approx \left\{ (x_w, h) : x_w = x_j + 2h\left(\frac{q^2}{3n} - \frac{1}{2} \right) \right\}.$$

This shows that the jump boundary is approximately linear as a function of the bandwidth h. The curved funnel shapes shown in the bottom panel of Figure 1, come from the fact that the $\log_{10}$ scale is used for h in that display. The $\log_{10}$ bandwidth scale was chosen for SiZer, since that scale gives smooths that are more "equally spaced", as shown in Marron and Chung (1997). In Section 3, changing this choice in the context of jump detection is proposed.

Another consequence of (1) is that SiZer will not show a significant increase, if only a single point is much larger than all of the others, regardless of how large that is. In particular, if all the data in a window have a constant value, except at a single point, SiZer has the best chance of indicating a significant slope, when the point is at the edge of the window. From (1), this is the case $j = n - 1$, which gives (after some simplification)

$$\frac{\widehat{m}_h}{\widehat{\sigma}_m} = 3^{1/2},$$

(independent of n and of h) which is not large enough to be significant for reasonable values of the critical level α. This shows that SiZer cannot find "single point features", such as might arise in spectral analysis, where a single frequency is dominant.

3 Jump SiZer

An important lesson from Section 2 is that the boundaries of the jump funnels grow linearly with the bandwidth, h. Since the human eye is better at noticing lines than particular curves, it is sensible to modify the SiZer map so that the jump funnels have linear boundaries. This is done by replacing the usual $\log_{10}(h)$ scale in the SiZer map by the linear h scale. The result of this for the data of Figure 1 is shown in Figure 2. As expected from the theory developed in Section 2, the boundaries of the jump funnels are now linear. Hence, for situations where exploratory jump detection is of interest, we recommend appending a linear scale SiZer map (as in Figure 2) to the usual SiZer analysis (shown in Figure 1).

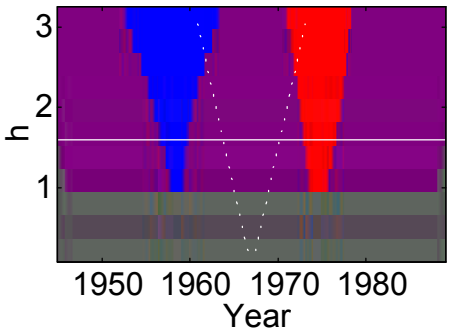

FIGURE 2: *Jump Sizer, for Penny Data shown in Figure 1. Linear bandwidth scale highlights jump funnels.*

Note that the curved lines, showing $\pm 2h$, are also now linear. A price to be paid in looking at this modification of the SiZer map is that a smaller range of bandwidths is needed to see the important jump structure of the data. The reason is that if the full range of bandwidths is used, then the linear scale concentrates too much on the larger bandwidths, which is not where the jump funnel occurs. We recommend addressing this problem, by using only half the range (on the log scale). In particular, if the range $[\log_{10}(h_{\min}), \log_{10}(h_{\max})]$ is used for SiZer, the Jump SiZer should use the range

$$\left[h_{\min}, e^{(\log 10(h_{\min})+\log_{10}(h_{\max}))/2}\right].$$

The upper endpoint of this range is shown on the original SiZer map (bottom of Figure 1) as the dashed horizontal bar.

4 Examples

In this section, we show some examples illustrating the use of this version of SiZer. Figure 3 shows a simulated example, based on the same data as considered in Figure 8 of Chaudhuri and Marron (1999), where the funnel shapes were first noticed. It was also pointed out there that each jump in the regression function corresponds to a funnel. The underlying regression function, called "Blocks" comes from Donoho and Johnston (1994), and the additive i.i.d. Gaussian noise is the same "high noise" used by Marron, Adak, Johnstone, Neumann and Patil (1998).

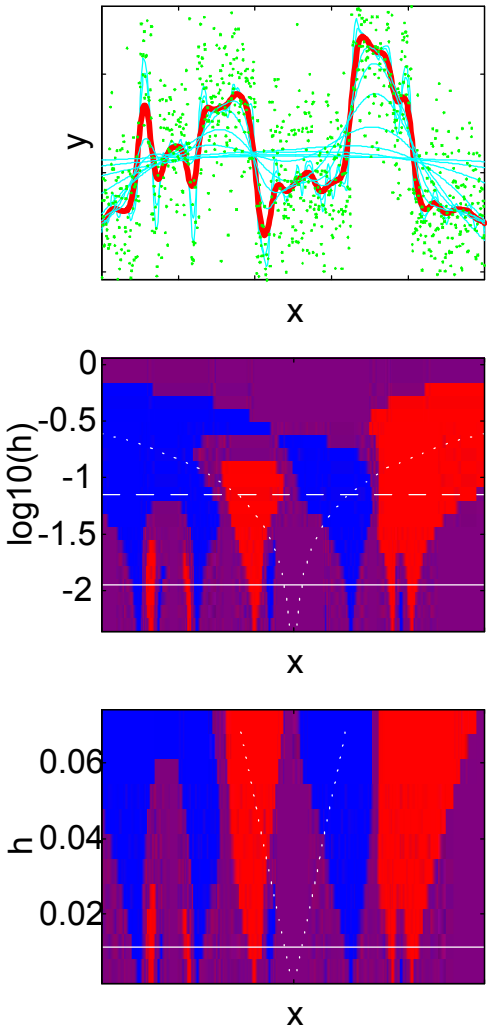

FIGURE 3: *Jump SiZer for the Blocks target curve, and high level Gaussian noise. Every jump point is found.*

Note that the new bottom panel in Figure 3 shows that many of the funnels in the middle panel are clearly of the type associated with jumps. However, some of the funnels are too small to clearly see the linear structure, because they are too close together.

This type of jump detection is intended as exploratory. It is recommended that after potential jumps are found in this way, their significance should be investigate by any of many conventional change point testing methods. See Carlstein, Müller and Siegmund (1994) for access to that literature.

Another example, showing some additional insights is shown in Figure 4. Here the setting is the low noise level, "Angles" target, from Marron, Adak,

Johnstone, Neumann and Patil (1998).

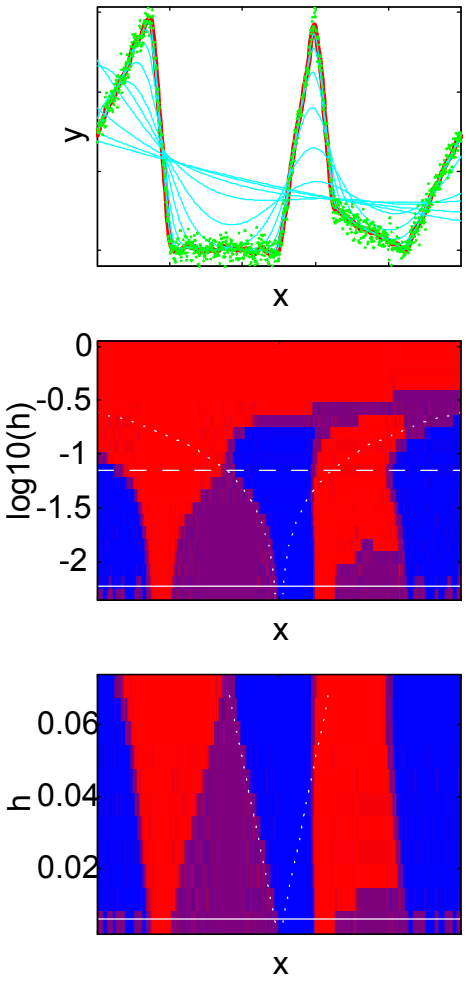

FIGURE 4: *Jump SiZer for the Angles target curve, and low level Gaussian noise. Shows effects of sloped regions in the regression.*

Figure 4 explores the impact on the jump SiZer map, when the underlying regression is piecewise linear, without jumps. Note that in regions where the slope is very steep, even the smallest bandwidth parts of the SiZer maps are colored (as happens in jump regions), because the noise level is low relative to the slope. The blue-red boundary on the left side is nearly vertical, but slopes a little to the left, since the slopes are different on each side of the first angle. This same effect is visible at the last red blue boundary. The boundaries of the large purple triangular region have a slope that is more similar to those observed for jumps in Figures 2 and 3. A very careful look shows that e.g. the purple blue boundary is slightly less steep than the nearby dotted line, while in

Figures 2 and 3, the full jump funnels are slightly more steep. This is caused by the sharp increase that generates the blue region being not quite vertical. This shows that jump SiZer gives meaningful impressions of the data, even when there is only a rapid increase in the regression, instead of an actual jump.

References

[1] Carlstein, E., Müller, H. - G. and Siegmund, D. (1994) *Change-point Problems*, Institute of Mathematical Statistics, Hayward, California, Lecture Notes, Volume 23.

[2] Chaudhuri, P. and Marron, J. S. (1998) Scale space view of curve estimation, North Carolina Institute of Statistics Mimeo Series # 2357.

[3] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structures in curves, *Journal of the American Statistical Association*, to appear.

[4] Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425-455.

[5] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall, New York.

[6] Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1998) Exact risk analysis of wavelet regression, *Journal of Computational and Graphical Statistics*, 7, 278-309.

[7] Marron, J. S. and Chaudhuri, P. (1998a) Significance of Features via SiZer", in *Statistical Modelling, Proceedings of 13th International Workshop on Statistical Modelling*, Brian Marx and Herwig Friedl, Eds., 65-75.

[8] Marron, J. S. and Chaudhuri, P. (1998b) When is a feature really there? The SiZer approach, in *Automatic Target Recognition VII*, Firooz A. Sadjadi, Editor, Proc. of SPIE vol. 3371, 306-312.

[9] Marron, J. S. and Chung, S. S. (1997) Presentation of smoothers: the family approach, Institute of Statistics, Mimeo Series # 2347.

[10] Ruppert, D., Sheather, M. J. and Wand M. P. (1995) An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, 90, 1257-1270.

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

[12] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, New York.