Significance in Scale-Space for Clustering

F. Godtliebsen
Department of Mathematics and Statistics
University of Tromsø
N-9037 Tromsø, Norway

J. S. Marron
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853
and
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
USA

S. M. Pizer
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599-3175

1 Introduction

An intuitive, visual approach to finding clusters in low dimensions is through the study of smoothed histograms, e.g. kernel density estimates. Scale-space provides a useful framework for understanding data smoothing. See Lindeberg (1994) and ter Haar Romeny (2001) for excellent overview of the very large scale-space literature.

The scale-space approach has allowed practical resolution of several long-standing problems in the statistical smoothing literature. See Chaudhuri and Marron (1999, 2000) for detailed discussion. For example, the classical problem of choice of the level of smoothing (bandwidth) can be viewed in an entirely new way using scale-space ideas. In particular, instead of choosing one level of smoothing, one should consider the full range of smooths (the whole scale-space). This corresponds to viewing the data at a number of different levels of resolution, each of which may contain useful information.

For clustering purposes, this simultaneous viewing of several different levels of smoothing incurs an added cost of interpretation. In particular, it becomes more challenging to decide which of the many clusters that are found at different levels represent important underlying structure, and which are insignificant sampling artifacts. An overview of some solutions to this problem is given in Section 2. These solutions involve scale-space views of the data (i.e. a family of smooths), which are enhanced by visual devices that reflect the statistical significance of the clusters that are present.

In keeping with the visual nature of these new methods, only one and two dimensional cases are presented. Certainly higher dimensional clustering is of keen interest, but visual implementation in higher dimensions represents a very significant hurdle. For now, dimension reduction methods need to be applied first, before these approaches can be used in higher dimensions.

In Section 3 we propose a new enhancement of the two dimensional version, based on the natural idea of contour lines.

Finally there is some discussion of interesting future research directions in Section 4.

2 Overview

There are a number of different approaches to assessing the statistical significance of clusters in one and two dimensions. This problem was called "bump hunting" by Good and Gaskins (1980). A wide range of approaches to this topic may be found in the papers Silverman (1981), Hartigan and Hartigan (1985), Donoho (1988), Izenman and Sommer (1988), Müller and Sawitzki (1991), Hartigan and Mohanty (1992), Minnotte and Scott (1993), Fisher et al (1994), Cheng and Hall (1997), Minnotte (1997) and Fisher and Marron (2001). Many of these approaches are only concerned with the number of significant clusters.

In this paper we discuss more visual approaches to significant clusters, that make more explicit use of scale-space ideas. An advantage of the visual approach is that one also learns where clusters are located. Viewing through scale-space reveals the levels of resolution at which each cluster appears.

Statistical inference for clustering in one dimension was developed by Chaudhuri and Marron (1999). Their method is called SiZer, for "SIgnificance of ZERo crossings". SiZer finds clusters through the study of the slope of the smooth histogram. A cluster is significant when the slope of the curve is significantly positive on the left, and significantly negative on the right. In particular, when there is a statistically significant zero crossing of the derivative. An example is given in Figure 1.

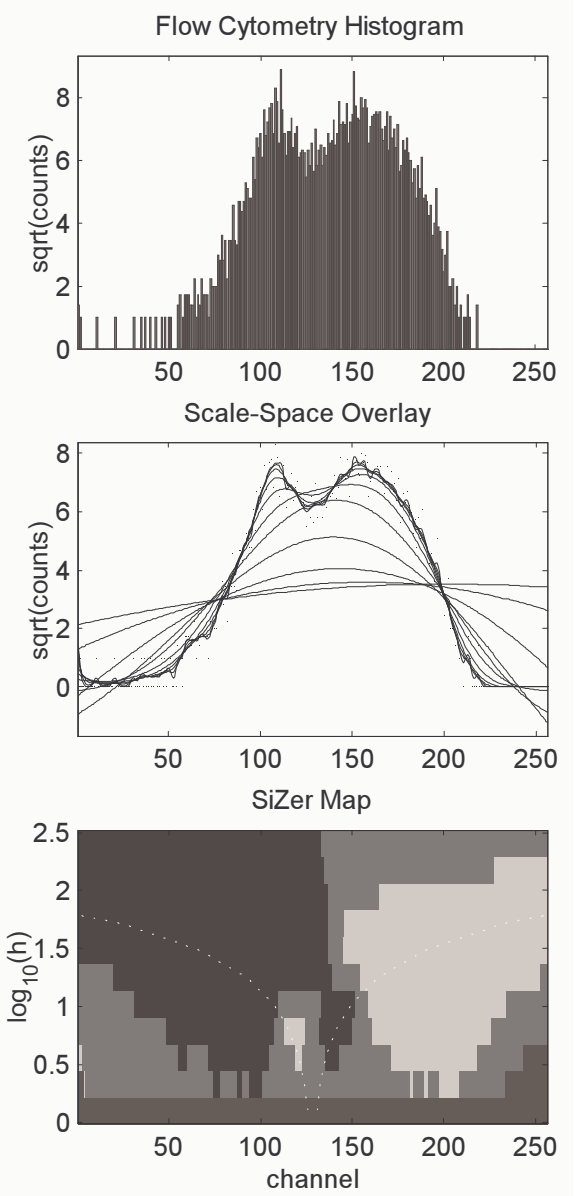

FIGURE 1: *SiZer analysis of the flow cytometry data. Original square root bin counts in the top panel. Family of smooths in the middle panel. SiZer analysis, showing two significant clusters in the bottom panel.*

Flow cytometry studies the presence and percentage of fluorescence marked antibodies on cells. The fluorescence of individual cells is measured, and the results are binned, into bins called "channels". The top of Figure 1 shows

a bar graph of square root bincounts, from a single experiment. The bar graph suggests that there are two clusters in the data. Are the clusters statistically significant? Or could they be mere artifacts of the sampling process? The issue is not 100% clear, because the peaks contain some bars that dip below the taller bars located in the valley between. SiZer aims to address this issue.

The heights of the bars in the top panel of Figure 1 are shown as dots in the middle panel, which also shows the scale-space as the family of curves. If the raw fluorescence levels were available, the scale-space curves would be kernel density estimates. However SiZer also works in terms of counts, using local linear smoothing to obtain the scale-space, which is done here. The inference done by SiZer is shown in the map in the bottom panel of Figure 1. The horizontal (x) axis of this map is the same as the x-axis of both of the panels above, i.e. "location". The vertical (y) axis of the map is "scale", i.e. bandwidth of the smooth, on the log scale. Thus each row of the SiZer map corresponds to one of the curves in the middle panel. The SiZer statistical inference is based on a confidence interval, for slope (derivative) of the smooth at each location, and at each scale. When the confidence interval is completely above 0, the smooth is significantly increasing, and the color blue (shown here as a dark shade of gray to minimize the need for color printing) is used. When the confidence interval is entirely below 0, the smooth significantly decreases, and this location in the map is colored red (shown here as a light shade of gray). In the indeterminate case, when the confidence interval contains 0, the intermediate color of purple (shown here as a the lighter intermediate shade of gray) is used. The fourth SiZer color is gray (the darker intermediate shade here), which is used at locations and scales where there is not enough data in each kernel window for reliable statistical inference. This SiZer map shows that both clusters are statistically significant. In particular the left hand cluster around channel 110 is seen to be "really there" because of the large blue (dark) patch to the left, and the smaller red (light) patch on the right (flagging significant increase, followed by decrease). The same holds for the larger cluster around channel 160.

The SiZer visualization is very useful in one dimension, but does not extend easily to the two dimensional case. One reason is that an overlay view of the family of curves (the scale-space) is no longer possible. A more serious reason is that the SiZer foundation of "significantly sloping up or down" no longer makes sense in two dimensions. Proposals for some completely new visualizations of statistically significant features, called S^3 for "signifi-

cance in scale-space", were made in the case of two dimensional images by Godtliebsen et al (1999). Some closely related proposals for two dimensional smooth histograms, and thus for finding significant clusters, were made by Godtliebsen et al (2001). The problem of lack of availability of overlays in two dimensions is addressed by the construction of movies where time is the scale (i.e. level of smoothing). The problem of statistical inference is addressed by adding visual enhancements to the movie. Some of these are illustrated in Figure 2.

Figure 2 provides a visual clustering of the Earthquake data from Section 4.2 in Wand and Jones (1995), shown as a scatterplot in the top panel. These data record the locations of epicenters, in longitude (the x coordinate, with 122 subtracted for numerical convenience) and latitude (the y coordinate, with 46 subtracted) of earthquakes in the Mount St. Helens area of the United States.

The lower left hand panel of Figure 2 demonstrates the streamline version of S^3. The green "streamlines" are the visual cues indicating statistically significant structure. These are based on the gradient of the gray level surface, which is the direction of maximal change. These green curves essentially show the direction that a drop of water would follow as it moves down the surface. However, lines are only drawn in regions where the gradient is statistically significantly different from 0, i.e. where there is a significant slope.

The gray level plot, together with the streamlines suggest three clusters, as does the scatterplot on the top. The streamlines show that there is strong evidence only for the right cluster being statistically significant, as indicated by the ring of streamlines pointing towards the peak, that are completely around this cluster. The middle cluster has streamlines pointing towards the peak most of the way around, which are a suggestion of a cluster, but not conclusive statistical evidence. The left cluster is much less convincing (in the sense of statistical significance), because there are few streamlines pointing towards its peak. Of course it must be kept in mind that statistical significance is necessarily one sided. Streamlines give strong evidence of presence of a feature, but lack of streamlines only indicates the evidence is not strong enough to be sure, and does not prove absence of a cluster.

To save space, only one scale, i.e. bandwidth is shown in Figure 2. This is $h = 4$, which was chosen for presentation purposes, after viewing the full scale-space. For data analytic purposes, this viewing of the full scale-space is essential, and we suggest doing this as a movie. We recommend viewing the

movie version of the left side of Figure 2, in the file `SSScntr1Fig2a.avi` in the web directory http://www.unc.edu/depts/statistics/postscript/papers/marron/SSS_cntr/. The movie format is AVI, which is easily viewable on most computers without the need of downloading an extra viewer.

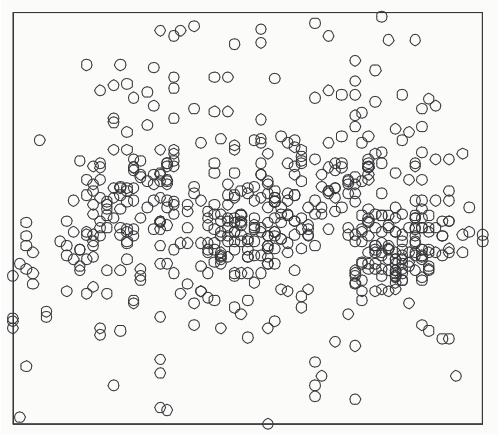

FIGURE 2A: *The raw Earthquake data (shown as a scatterplot). An S^3 analysis appears in Figure 2b (which has been separated, since color pictures have been combined).*

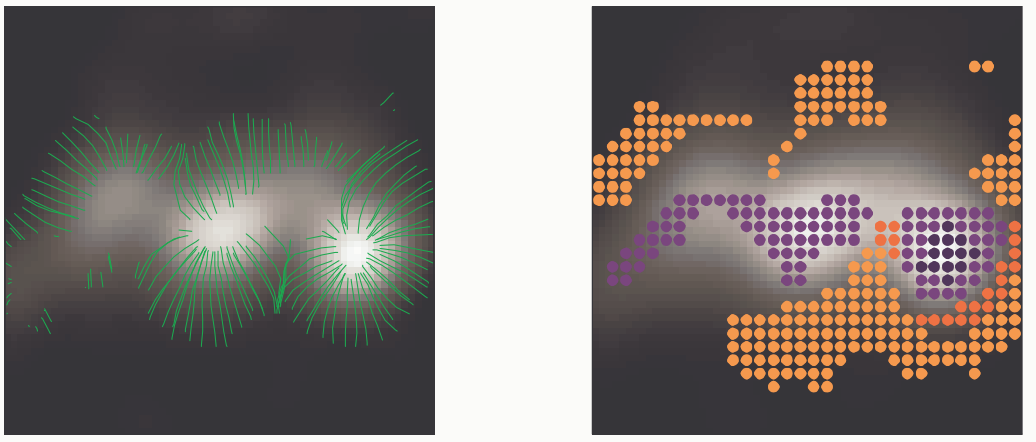

FIGURE 2B: *S^3 analysis of the Earthquake data (shown as a scatterplot in Figure 2a), using gradient streamlines (left panel) and curvature dots (right panel). The right cluster is clearly statistically significant, the middle cluster is not quite conclusive, the left cluster is less well defined.*

The lower right panel of Figure 2 shows an alternate version of S^3. This time the statistical inference is based on curvature. Curvature is conveniently

described in two dimensions using the eigenvalues of the Hessian matrix. At a grid of image locations, statistical significance of these eigenvalues is assessed. Colored dots, overlaid on top of the gray level image, provide quick visual access to this information. The following table indicates colors that are used for the various curvature cases, depending on the largest eigenvalue $\widehat{\lambda}_+$, the smallest eigenvalue $\widehat{\lambda}_-$, and the appropriate quantile $\widehat{q}_{\widehat{T}}$. See Godtliebsen et al (1999) for details of the derivation, including the choice of $\widehat{q}_{\widehat{T}}$. The latter requires substantial effort, even when the data are exactly Gaussian, because the joint distribution of the eigenvalues $\widehat{\lambda}_+$ and $\widehat{\lambda}_-$ is non-standard. Appropriate rescaling and tabulation of this distribution using simulation methods is done in Godtliebsen et al (1999).

color	**feature**	**characterization**
yellow	hole	$\widehat{\lambda}_+, \widehat{\lambda}_- > \widehat{q}_{\widehat{T}}$
orange	long valley	$\widehat{\lambda}_+ > \widehat{q}_{\widehat{T}}, \left\|\widehat{\lambda}_-\right\| < \widehat{q}_{\widehat{T}}$
red	saddle point	$\widehat{\lambda}_+ > \widehat{q}_{\widehat{T}}, \widehat{\lambda}_- < -\widehat{q}_{\widehat{T}}$
purple	long ridge	$\left\|\widehat{\lambda}_+\right\| < \widehat{q}_{\widehat{T}}, \widehat{\lambda}_- < -\widehat{q}_{\widehat{T}}$
dark blue	peak	$\widehat{\lambda}_+, \widehat{\lambda}_- < -\widehat{q}_{\widehat{T}}$

Most of these colors appear in the right side of Figure 2. Again only a single scale has been selected, $h = 5.19$, after viewing the full scale-space. This scale is larger than that chosen for the streamline analysis in the left panel, because curvature estimates feel noise more strongly than slope estimates, so more smoothing is needed for similar inference. Viewing the full scale-space is again recommended, using the movie file `SSScntr1Fig2b.avi` in the above web directory. As in the above analysis, the right cluster in the data comes through very strongly. In particular there are a number of dark blue dots. The center cluster is less clear, showing all purple dots, which only show existence of a ridge, not a cluster. Some suggestion that the middle cluster is separate is provided by the red saddle point dots between the clusters, but this is not conclusive, because these also appear on a ridge that then slopes upwards. The potential third cluster on the side shows up less strongly here than in the streamline analysis, because at this coarser scale (needed for adequate noise reduction) it is nearly smoothed away. The orange dots highlight locations where "clusters emerge from regions of no data".

A disappointing aspect of the analysis of Figure 2 is that only the right hand cluster is statistically significant in this sense. A possible approach to

investigating the significance of the other clusters is to adjust the statistical inference, via the level of significance, α. In Figure 2, and all other examples in this paper, the standard $\alpha = 0.05$ is used. Results for the less stringent case of $\alpha = 0.2$ are viewable in the movie files `SSScntr1Fig2c.avi` and `SSScntr1Fig2d.avi`. These are not shown here to save space. The most interesting result was that at the scale $h = 6.17$, at least one dark blue dot appeared at the top of each of the three clusters, providing evidence that all 3 clusters were significant (at this lower level).

All smoothing methods used in the paper are kernel based methods. For more background information on these, see for example Scott (1992), Wand and Jones (1995) and Fan and Gijbels (1996). The Gaussian kernel function is used everywhere in this paper, as this has the most appealing scale-space properties, see Lindeberg (1994).

An important technical component of these methods is correct simultaneous statistical inference. In particular, many inferences are performed here at one time. In such situations, naive implementations of hypothesis tests will result in a number of false positives. Care has been taken to avoid this problem in the construction of SiZer and S^3. See Chaudhuri and Marron (1999) and Godtliebsen et al (1999, 2001) for details.

3 New Method

The streamline approach to S^3, shown in the left panel of Figure 2, highlights "statistically significant gradient directions" in two dimensions quite well. However, gradients alone are not a particularly intuitive vehicle for understanding "surface shape". This is shown in the left hand panel of Figure 3, which is based on a simulated data set, composed of a surface with additive i.i.d. Gaussian noise. The underlying surface is an "asymmetric volcano", that is formed from the volume of revolution of a Gaussian probability density, with mean between the center and edge, with an off-center (towards the right) cylinder lowered to height 0 near the middle. One frame of the scale-space is shown, and it is recommended that others be viewed in the movie `SSScntr1Fig3a.avi` from the above web directory.

After careful contemplation, using visual clues from the underlying gray-level image, it becomes clear that the central streamlines start in the low black region near the center and climb the inner cone in a radial fashion. When they reach the circular crest, they then follow the gradient of the

crest towards the left. Finally at the top of the crest (on the left side) the gradient is no longer significant, and the streamlines stop. Streamlines coming from the outer edge climb the outer cone in a radial way, and join the inner streamlines at the crest.

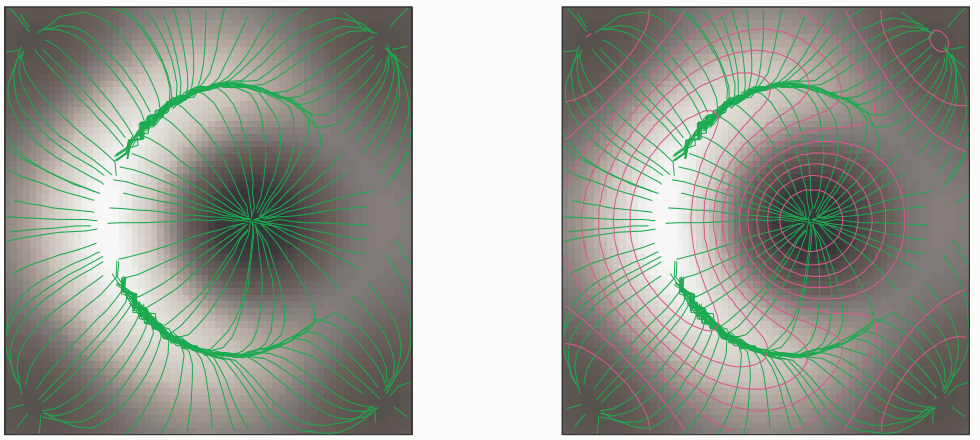

FIGURE 3: *S^3 analysis of the assymetric volcano simulated example. Streamline only (left panel) and streamline and contour (right panel) versions. This shows that the addition of contours enhances the interpretability.*

While the streamlines describe the statistically significant aspects of the shape of the surface, substantial thought is required for complete understanding. The goal of this section is to provide additional visual clues, which assist this process, in particular by the addition of contour lines to the S^3 graphics. The result of this is shown in the right hand panel of Figure 3. Note that the contour lines, shown in purple, are orthogonal to the green gradient lines. Statistical significance of the feature being illustrated (e.g. a cluster), is shown by only drawing contours in regions where the gradient is statistically significant. This notion of statistical significance is particularly well suited to finding clusters. In particular a significant cluster will be a hill of high density (i.e. light gray), that is highlighted by a purple circle surrounding it. The interpretation of the circle is that everywhere around, the slope is significant, which is a useful notion of "cluster".

The contour lines in the right part of Figure 3 provide quicker intuitive understanding of the shape of the surface. The central circular contours clearly show the inner cone of the volcano. The banana shaped higher level

contours immediately reveal the shape of the light colored curved ridge on the left side.

These new significant contours are constructed using the same statistical inference methods as for the streamline version of S^3. In particular, each pixel location is flagged as having either a significant, or an insignificant gradient. Streamlines are drawn by using a step-wise procedure, following gradient directions, with some random starting values. This methodology could be used for the contours, by simply stepping orthogonally to the gradient. However, contours are not easy to draw in this way, because they generally form closed curves, which are not simple to construct using step-wise procedures involving accumulating numerical errors. But contours have the advantage that many ready made functions to draw them are available. We use the generic contour subroutine in Matlab, with deletion of parts of the contours in regions where the gradient is not significant.

Figure 4 shows some S^3 analyses of the Melbourne temperature data. The raw data here is a lag one scatterplot of daily maximal temperatures in Melbourne Australia, over a 10 year span. The x axis shows yesterday's maximum, and the y axis shows today's maximum. A simple weather prediction idea is to use yesterday's maximum to predict today's maximum. If that were exactly correct, all the data points would lie on the diagonal 45° line. Using a conditional density estimation analysis, Hyndman et al (1996) showed visually that there is a "horizontal ridge of high density", i.e. a ridge where today's maximum is 20 degrees (Celsius). Verification of the statistical significance of this ridge (and evidence for a related vertical ridge) were provided by Godtliebsen et al (2001), using an analysis similar to that shown in the left panel of Figure 4. Both the horizontal and vertical ridges have a physical explanation, discussed in Godtliebsen et al (2001).

The left hand panel shows the streamline only analysis at the scale $h = 4.36$. Again all scales should be viewed, and are available in the movie file `SSScntr1Fig4a.avi` at that above web address. The horizontal ridge is visible as a coalescence of streamlines. The diagonal ridge does not show up well because near 30 degrees the gradient along the ridge is no longer significant. The vertical ridge does not appear at this scale, but does show up for smaller scales.

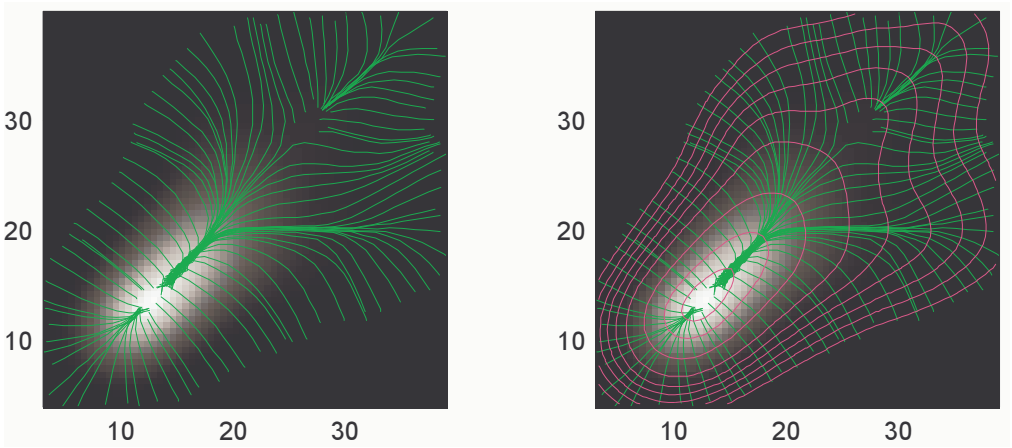

FIGURE 4: *Shows streamline and contour analysis is more effective than streamlines only, for the Melbourne temperature data.*

The right panel of Figure 4 shows that adding contours again enhances the analysis, for the same scale (see also the full scale-space movie in `SSScntr1Fig4b.avi`), $h = 4.36$. Curves in the contours provide a stronger visual impression of "ridges" than is available from the streamlines. Even a hint at the vertical ridge is available in this way. The contour plot provides much more immediate visual insight about the diagonal and horizontal ridges.

An issue in the construction of contours is their spacing. Most of the examples in this paper use equal height spacing. However, this is sometimes inappropriate. This is demonstrated in a simulated example in Figure 5. Here the underlying surface is two elongated peaks, as can be seen from the gray level plot. The left panel of Figure 5 shows equal height spacing of the contours. Here the scale is $h = 4$, but other scales revealed similar effects, see the movie version in the file `SSScntr1Fig5a.avi` in the same web directory. Note that in the large flat areas on the upper left and the lower right, the contours are somewhat deficient, in the sense that no contour appears for long stretches of the significant green streamlines.

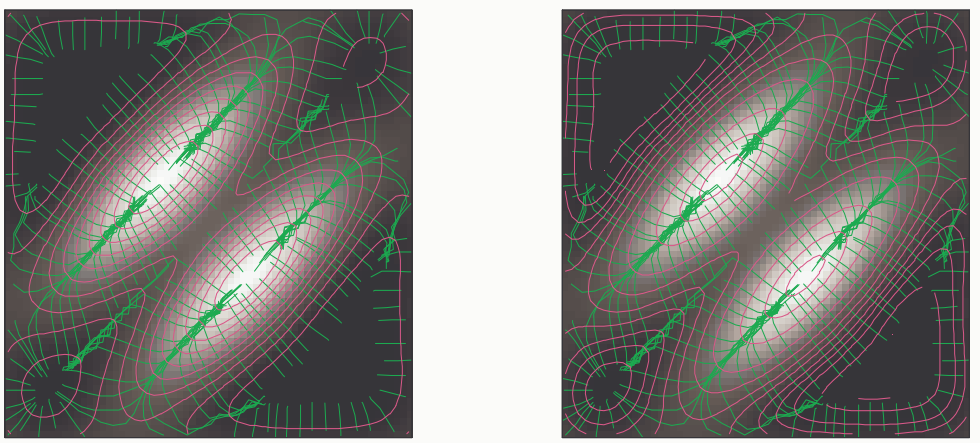

FIGURE 5: *Simulated example showing the difference between equal height and modified quantile contour spacing.*

A simple solution to this problem is to add more contours. However, this is not satisfactory, because the contours then become too dense in other regions. A better solution is to use different types of contour spacing. An alternate approach to contour spacings is explained using the gray level histogram shown in Figure 6. The gray bars in Figure 6 appearing in the top half of the plot show how equally spaced height contours relate to the population of gray levels in the image. Note that a quite large number of pixels near the left edge of the population (nearly black) are represented by only a few contour lines. This explains the poor contour performance in the dark areas of the left panel of Figure 5.

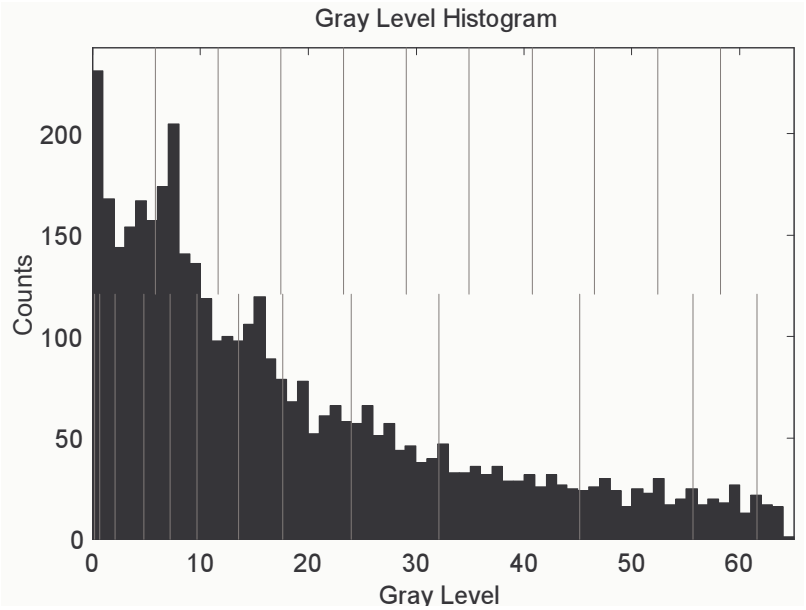

FIGURE 6: *Gray level histogram for image shown in Figure 5. This contrasts equal height contour spacing (represented as gray bars in the top half), with modified quantile spacing (represented as gray bars in the bottom half).*

Figure 6 suggests that this problem can be solved by taking the contour heights according to quantiles of the gray level population, as indicated by the gray bars in the lower half. An initial experiment (see the movie `SSScntr1Fig5c.avi` in the same web directory) with equally spaced quantiles suggested it was worth including more contours at each end. Some experimentation lead us to suggest including 0.1 and 0.4 times the quantile spacing at the lower end, and making a symmetrical inclusion at the upper end. This is done for the gray bars on the bottom of Figure 6, and for the contours in the right panel of Figure 5, see the movie version in `SSScntr1Fig5b.avi` in the same web directory. The equally spaced quantile version is quite similar the right panel of Figure 5, except that the two contours nearest the peak are missing. An advantage of quantile spacing is an additional interpretation. When 9 equally spaced quantiles are used, the region between two consecutive contours encloses ten percent of the pixels.

The Melbourne Temperature data, shown in Figure 4 is a case where equal spacing does not work, so the modified quantile spacing was used there. In general, we suggest choosing between height spacing and modified quantile

spacing, by starting with height spacing, and switching when that is unsatisfactory (which is visually apparent).

In Figure 7, the Earthquake data from Figure 2 are analyzed by adding contours to the streamline analysis from the left side of Figure 2 (at the same scale $h = 4$). The full scale-space movie is available in the file `SSScntr1Fig7a.avi` in the same web directory.

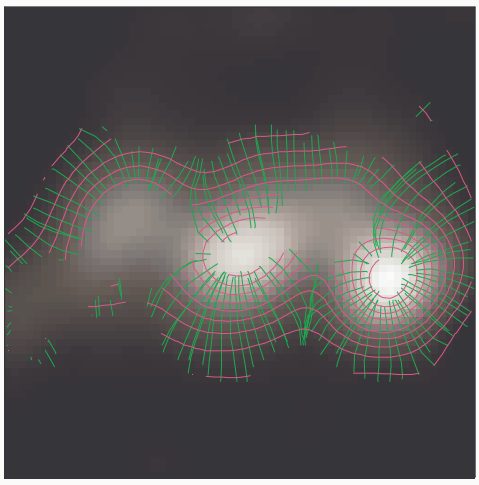

FIGURE 7: *Contour and streamline analysis of the Earthquake data.*

The main lessons about statistical significance of clusters is the same as above, but again the contours make the information more easily accessible.

4 Future Directions

While the contour version of S^3 provide an improvement to earlier versions, there are many other possible improvements, and interesting directions for further research.

One such area is to combine the information present in the dot version of S^3 with significant contours. A straightforward approach is to color the contours using the same dot colors. This could provide more immediate interpretation of some of the contours, as well as incorporating additional useful information. Figure 8 shows the curvature information available for the Melbourne Temperature data. Note that this provides a different compelling evidence for the statistical significance of the ridges. The full scale-space movie version is available in `SSScntr1Fig8.avi` in the same web directory.

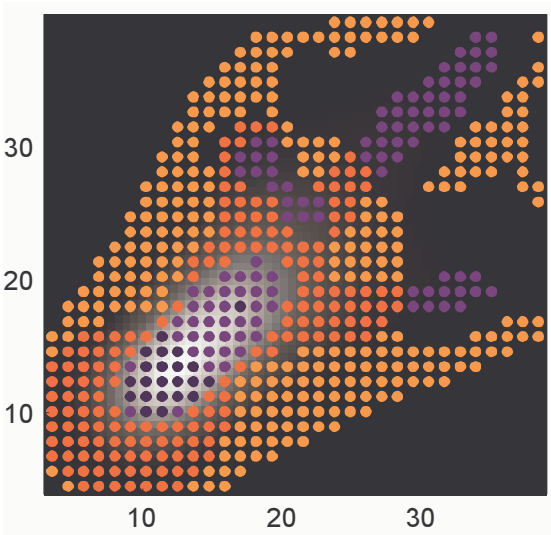

FIGURE 8: *Curvature dots analysis of the Melbourne Temperature data.*

A more complicated, but perhaps more useful extension is to visually represent the statistical significance of the curvature of the contour lines themselves. For example with the contour analysis of the Melbourne Temperature data shown in Figure 4, it appears that a number of the contours are not convex. Is this concavity statistically significant?

Perhaps the most challenging extension of S^3 is from two dimensions to three. As with the extension from SiZer (dimension 1) to S^3 (dimension 2), the main challenge is the visualization. For example, the scale-space is easily presented as an overlay in one dimension, and a movie in two dimensions, but it is less clear how to present a three dimensional scale-space. After this problem is solved, then careful attention needs to be given to which quantities (e.g. gradient and curvature) should have their statistical significance displayed, and how they should be visualized.

References

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

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

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.

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

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

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. (2001) Mode Testing Via the Excess Mass Estimate, to appear in *Biometrika*.

Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (1999) Significance in Scale Space, unpublished manuscript.

Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2001) Significance in Scale Space for Bivariate Density Estimation, to appear in *Journal of Computational and Graphical Statistics*.

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.

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

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

Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996) Estimating and visualizing conditional densities, *Journal of Computational and Graphical Statistics*, 5, 315-336.

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

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

Minnotte, M. C. (1997) Nonparametric testing of the existence of modes, *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.

Scott, D. W. (1992) *Multivariate density estimation, theory, practice and visualization*, Wiley Interscience, New York.

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

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

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