Significance in Scale Space for Density Estimation

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

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

P. Chaudhuri
Indian Statistical Institute
Calcutta 700035
India

June 28, 1999

Abstract

An important problem in the use of density estimation for data analysis is whether or not observed features, such as bumps are "really there", as opposed to being artifacts of the natural sampling variability. Here we propose a solution to this problem, in the challenging two dimensional case, using the graphical technique of Significance in Scale Space. Color and dynamic graphics form an important part of the visualization method.

1 Introduction

Kernel density estimation is a smoothing method which shows structure in data that can be hard to find by other methods. See, for example, Scott (1992), Wand and Jones (1995) and Bowman and Azzalini (1997) for many interesting examples. While the method is good at finding structure, it can also miss important structure via oversmoothing, or else find unimportant spurious structure via undersmoothing.

One approach to this problem is via data based bandwidth selection, surveyed in Jones, Marron and Sheather (1996). While good methods have been successful at finding bandwidths which result in "good estimates", they are tuned for something different than understanding which features are significant. Furthermore, the best methods currently available seem to be inherently one-dimensional.

Scale space ideas were used to motivate a much different approach, called SiZer, to finding significant structure in Chaudhuri and Marron (1999). See Marron and Chaudhuri (1998a,b) and Kim and Marron (1999) for further examples and discussion. Scale space is a concept from computer vision, see Lindeberg (1994). While scale space is simply a family of Gaussian kernel smooths indexed by the bandwidth, it comes with two viewpoints that are not common in the statistical literature. The first view is that one should not try to focus on a single bandwidth, because there is usually important information available at several amounts of smoothing, i.e. at several different levels of resolution of the data. The second is that the focus of statistical inference should be shifted from the "true underlying function", to "the true underlying function viewed at the given level of resolution", i.e. to the underlying function convolved with the kernel. The last idea is very important, because it avoids the difficult problem of handling bias. See Section 6.2 of Chaudhuri and Marron (1999) for further discussion.

SiZer combines scale space ideas with a new type of visualization, to give a useful tool for finding structure in univariate data sets. The method works for both univariate regression and univariate density estimation. However, SiZer is inherently one dimensional for two reasons. One is the type of visualization used. The other is that it is based on whether the derivative is increasing or decreasing, which is not a useful concept in more than one dimension.

These problems were solved in the two dimensional setting of image analysis, by Godtliebsen, Marron and Chaudhuri (1999), where the "Significance in Scale Space" method was developed. In this paper we do a parallel development of the Significance in Scale Space concept in the different setting of bivariate density estimation.

The first challenge is to visualize the family of smooths. In one dimension, this can be done by simply overlaying the different smooths, see Marron and Chung (1997) for suggestions about this. In two dimensions we give the same insight about the family of smooths, i.e. the scale space, by considering a movie of smooths (with time indexing the log of the bandwidth). The second challenge is the display of which features are important. This is addressed via added visual cues which indicate statistical significance, of the local gradient, and/or of the local curvature.

As an illustrative example, we consider the Melbourne temperature data analyzed by Hyndman, Bashtannyk and Grunwald (1996). This is a study of how well yesterday's maximum temperature predicts today's maximum temperature. Figure 1 shows the standard lagged scatterplot, of $N = 3649$ observations, and a carefully chosen kernel smooth, where the height of the density estimation surface is represented by gray levels. There is a large white blob representing highest contours of the smooth. This blob tends to lie on the line $y = x$ (i.e. along the 45 degree line, but be careful to not confuse angular degrees here with temperature in the following), and also has a thinner extension into the higher temperatures, which is consistent with the idea that the best predictor of today's max is yesterday's max. An interesting feature which was the focus of the Analysis of Hyndman, Bashtannyk and Grunwald (1996) is the thin horizontal arm projecting out at a constant today's max of around 20 degrees (Celsius), i.e. along the line $y = 20$. They presented graphics which highlight this feature, and also explained it in terms of local meteorological knowledge, which suggested that times of high temperatures are frequently followed by a 20 degree max.

The rotationally symmetric Gaussian kernel function is used in all density estimates in this paper. The right side of Figure 1 shows only a single bandwidth $h = 5$. But it is quite useful to look at the full scale space, i.e. a broad range of bandwidths. Such figures are not shown in this paper to save space. However, it is conveniently and insightfully done by viewing a movie, that can be found in the file `ssskde1fig1b.mpg`, at the WWW address:

`http://www.unc.edu/depts/statistics/postscript/papers/marron/SSS_kde/`.

These movies can be very easily played on a PC e.g. by downloading, and then clicking on them. Gray level images are shown in most figures in this paper (and their accompanying movies). Most of them have the gray scale adjusted for maximal contrast, meaning the color black (white) is used for the minimum (maximum respectively) of the surface being represented (and this is done frame by frame in the movies). The right side of Figure 1 (and its movie version) is the only exception. Because of sparsity in gray scales, and the very high density of data in the lower left, the features of interest are invisible using the conventional full contrast gray scale. To see these features, the gray scale is modified to include only the lower 20% of the density, i.e. the color white is used for all regions where the density is higher than 20% of its maximum.

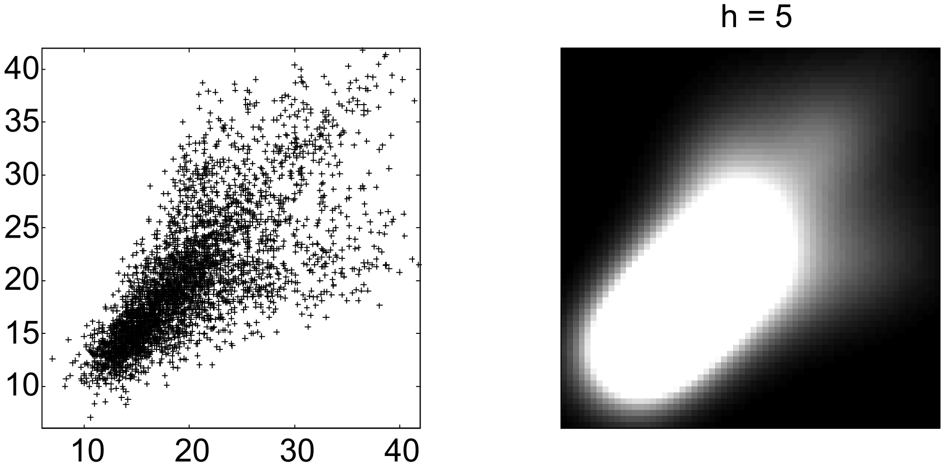

Figure 1: *Scatterplot (left) and $h = 5$ kernel density estimate (right) for Melbourne temperature data. The x axis is yesterday's maximum temperature (Celsius), and the y axis is today's maximum temperature (Celsius).*

Significance in Scale Space (S^3) methodology is useful for the part of this type of analysis where the statistician wonders if observed features, such as the arms along the lines $y = x$ and $y = 20$ on the right in Figure 1, are "really there". If so, they are worth a deeper search for causes, of the type that was done by Hyndman, Bashtannyk and Grunwald (1996). But if not, such effort could be wasted. Careful development is done in Section 2, but first the usefulness of S^3 is demonstrated.

Figure 2 shows two versions of S^3 applied to the Melbourne temperature data, at the scale $h = 5$, as considered on the right in Figure 1. The left hand plot considers an estimate of the gradient at a rectangular grid of locations. If the gradient is significantly different from 0, then an arrow is drawn in the gradient direction. Locations with no arrow, have "more noise than signal" (being in regions with very low data density, as seen in the scatterplot on the left in Figure 1), and are thus not highlighted. The texture of the arrows clearly reveals the dominant ridge along the line $y = x$. It also shows that the horizontal ridge, along the line $y = 20$ (where today's max temperature is about 20) is also "statistically significant". This would justify the search for explanations that was done by Hyndman, Bashtannyk and Grunwald (1996). Another interesting feature is some suggestion of a vertical ridge along the line $x = 20$ (where yesterday's maximum temperatures were about 20 degrees). This was not reported by Hyndman, Bashtannyk and Grunwald (1996), because their investigation was conditional on y given x, versus the full bivariate analysis done by S^3. As with the raw smooths, as shown in the right of Figure 1, it is useful to look at the full scale space, i.e. at many different smooths. Again a movie version of this can be found at the above web site, in the file `ssskde1fig2a.mpg`.

Another version of S^3 replaces the notion of "significant gradient", by "significant curvature". Now second partial derivatives are considered, and summarized by eigenvalues of the Hessian matrix, which give a rotation invariant notion of curvature. Colored dots represent statistical significance of the eigenvalues in various ways. Details are given in Section 2. Dark blue dots are used where both eigenvalues are significantly negative, e.g. near local maxima of the density estimation surface, shown here as the bright white region. There are two clusters of dark blue dots, which may represent different seasons. Light purple is used for one significantly negative eigenvalue, with the other not significant, as along a thin ridge. This coloring appears along the line $y = x$ for higher temperatures, as expected. It also appears along the ridge found by Hyndman, Bashtannyk and Grunwald (1996), along the line $y = 20$, which is another way of seeing that this feature of the data is "really there", and worth careful investigation. The color red is used where one eigenvalue is significantly positive, and the other is significantly negative, as at a saddle point. This appears along the line $y = x$ between the two purple regions, because the density estimation surface bends upwards to get to the bright peak At these points, the negative eigenvalue shows the downwards curve of the ridge in one direction, while the positive eigenvalue reflects the general upwards curve of the ridge towards the bright peak. There is a similar red region for the ridge along the line $x = 20$, again suggesting there may be a vertical ridge. Orange dots are used where one eigenvalue is significantly positive, and the other is not significant, as in a long valley. These appear where the density estimation surface curves up from the horizontal. Again the full scale space is informative, and we strongly recommend viewing the movie version, in the file `ssskde1fig2b.mpg` at the above web address.

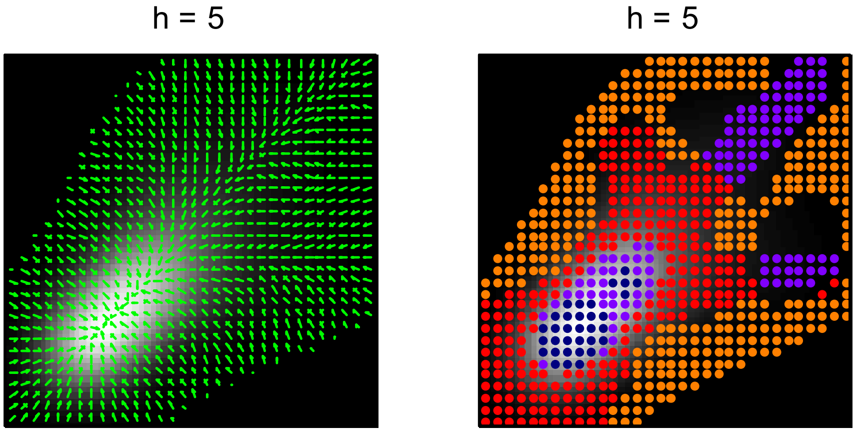

FIGURE 2: *Significant gradient (left) and curvature (right) versions of Significance in Scale Space for the $h = 5$ kernel density estimate, for the lagged Melbourne maximum temperature data.*

Figure 2 shows that the gradient and curvature versions of S^3 find different types of features. This has motivated the development of a hybrid version, which combines the two visual paradigms. This approach first finds significant arrows, as on the left of Figure 2. Then where curvature is also significant, the arrow is recolored with the curvature color (left green where curvature is not significant). And where curvature is significant, but gradient is not, a colored dot is used. Such an S^3 plot is shown in the left part of Figure 3. This is not easy to look at the first time, as there may be considerable problems with information overload. But after building some experience, we find this preferable to either approach taken separately. In particular this combines into a single plot the lessons about the features noted above. For the above reasons a movie version is available at the above web address in the file `ssskde1fig3a.mpg`.

A weakness of these versions of S^3 is that they are not rotation invariant. In particular, the arrows and dots are arranged along rectangular grid points. This results in certain "raster effects", e.g. the ridge along the line $y = x$ has a different texture in the left part of Figure 2, than the ridges along the lines $y = 20$ and $x = 20$. This problem is well understood in the area of visualization of vector fields in computer graphics, see Helman and Hesselink (1989) for an access to this literature. The right side of Figure 3 shows an adaptation of the "streamline" idea to the context of S^3. Each line starts at a randomly chosen pixel with a significant gradient (i.e. a green arrow as in the left of Figure 2). The line is step-wise extended in both gradient directions, until the gradient is no longer significant. Intuitively this corresponds to moving in the directions of steepest ascent and descent. Note that this visualization clearly highlights ridges, since the streamlines first march up the slope of the ridge, in a direction that is quite different from the ridge direction. Then when they get to the crest, they turn and follow the ridge direction. The previously discussed ridges along the lines $y = x$ and $y = 20$ are clearly shown in this way. Also the suggestion of a ridge along the line $x = 20$ is perhaps most marked with this version of S^3. Again it is well worth studying the movie version with more scales, in the file `ssskde1fig3b.mpg` at the above location.

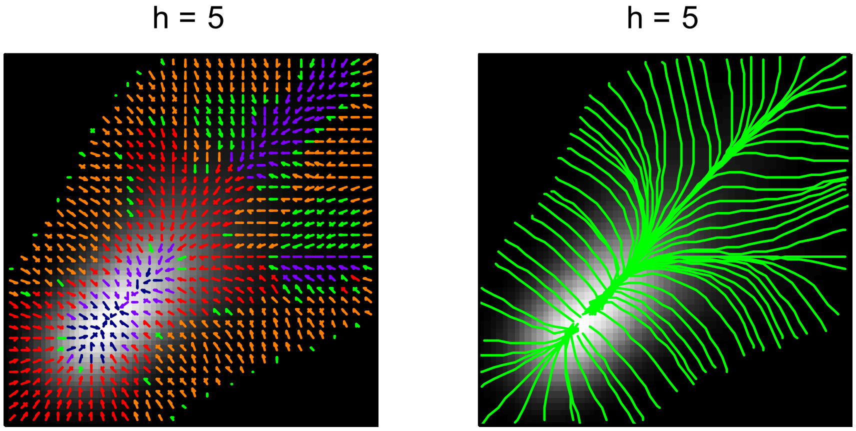

Figure 3: *Combined gradient and curvature (left) and streamline (right) versions of Significance in Scale Space for the $h = 5$ kernel density estimate, for the lagged Melbourne maximum temperature data.*

Readers who have been actually viewing the movies as suggested above, have probably noticed why it is useful to study more than one scale. This is that the vertical auxiliary ridge, i.e. along the line $x = 20$, which was not mentioned by Hyndman, Bashtannyk and Grunwald (1996), shows up as being clearly statistically significant at scales closer to $h = 3.3$, as shown in Figure 4.

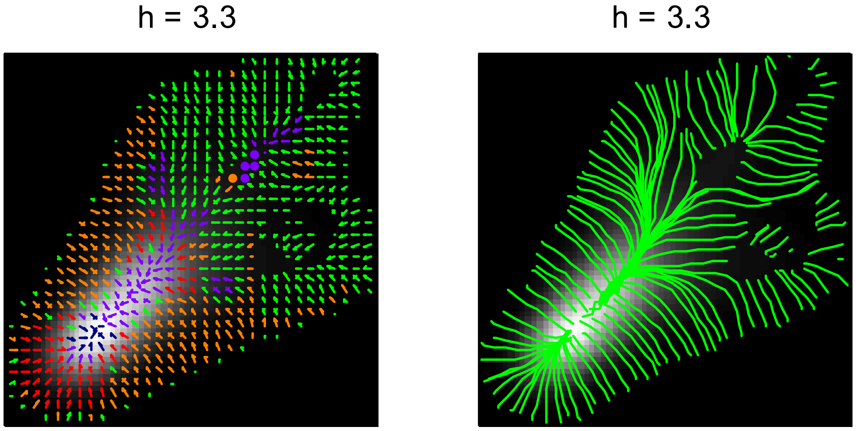

FIGURE 4: *Combined gradient and curvature (left) and streamline (right) versions of Significance in Scale Space for the $h = 3.3$ kernel density estimate, for the lagged Melbourne maximum temperature data.*

The scale, i.e. bandwidth, in Figure 4 has been carefully chosen from the full scale space, because it now highlights the statistical significance of the vertical ridge in the data along the line $x = 20$. This appears in the left part as the light purple vertical ridge. It appears in the right part as the confluence of streamlines. Hence, this feature is also worth deeper investigation. L. E. Chambers, from the Australian Bureau of Meteorology Research Centre, has confirmed that frequent occurrence of a maximum temperature of 20 degrees, followed by a higher temperature can be explained as temperatures driven by sea breezes on one day, followed by Northerly winds bringing high temperatures the next day.

Section 2 gives the details of the development of S^3 for density estimation. Section 3 gives more applications to real data sets.

2 Development of the methods

As seen in the right panel of Figure 1, the structure in a bivariate data set $\{(X_k, Y_k) : k = 1, ..., N\}$ may be understood from a kernel density estimate. This is defined as

$$\widehat{f_h}\left(x, y\right) = N^{-1} \sum_{k=1}^{N} K_h\left(x - X_k, y - Y_k\right),$$

where K is the kernel function, and h is the bandwidth, i.e. window width. See, for example Scott (1992), Wand and Jones (1995) and Bowman and Azzalini (1997), for discussion of many properties and variations of this estimator. In this paper, K is taken to be a spherically symmetric Gaussian density, and h is the common marginal standard deviation, for reasons given in Lindeberg (1994), Chaudhuri and Marron (1997) and Chaudhuri and Marron (1999). Thus the kernel has the product form

$$K_h\left(x - X_k, y - Y_k\right) = \varphi_h\left(x - X_k\right) \cdot \varphi_h\left(y - Y_k\right),$$

where φ_h denotes the rescaling

$$\varphi_h\left(\cdot\right) = \frac{1}{h}\varphi\left(\frac{\cdot}{h}\right).$$

Rapid calculation of $\widehat{f}_h\left(x, y\right)$, and also its derivatives as needed for S^3, can be done by first binning the data to an equally spaced grid. Details of binning are given in Section 2.1. This allows fast computation by simple convolution, as described in Section 2.2. The distribution theory needed for the statistical inference of S^3 is described in Section 2.3.

Matlab software for both the image and density estimation version of S^3 is available at the web site:

`http://www.unc.edu/depts/statistics/postscript/papers/marron/SSS_software/`.

The whole collection of files should be downloaded, e.g. to a single directory, because many of them call each other. The Matlab subroutine `conv2.m`, in the Signal Processing toolbox is required. The main call is to the subroutine `sss1.m`. The Matlab command "`» help sss1`" gives information about how to use the various versions of S^3. A difference between the image and density estimation versions of S^3 is that a "typically reasonable" range of bandwidths for the "full scale space" for images is $h \in [1, 8]$, while $h \in [2, 16]$ is generally more appropriate for density estimation (since more smoothing is often required before significant features appear). These bandwidth ranges are thus the defaults in our software.

2.1 Binning

There are several methods for binning data to an equally spaced rectangular grid, of the form

$$\left\{\left(x_i, y_j\right) : x_i = L_x + i\Delta_x,\ y_j = L_y + j\Delta_y,\ i = 0, ..., n,\ j = 0, ..., m\right\},$$

i.e. a rectangular lattice, where the x_i are equally spaced over $[L_x, L_x + n\Delta_x]$ and the y_j are equally spaced over $[L_y, L_y + m\Delta_y]$. Here we describe "simple binning" and "linear binning". See Appendix D of Wand and Jones (1995) for much more discussion about binning, and access to the literature.

Simple binning, also called "nearest neighbor binning", is best viewed as moving each data point to the grid point that is its nearest neighbor. For an equally spaced grid, this can be done rapidly by replacing the obvious iterated minimization approach by a rounding approach. In particular, a generic data point (X, Y), in the interior of the grid, should be mapped to the bin center indexed by i, j where the data are first mapped to an integer lattice, and then appropriately rounded,

$$\begin{aligned} i &= \mathit{ceil}((X - L_x) / \Delta_x), \\ j &= \mathit{ceil}((Y - L_y) / \Delta_y), \end{aligned}$$

where *ceil* is the function which rounds its argument up to the next integer. Then the mapped points are counted to give a matrix C of bin counts, whose i, j-th entry is

$$c_{i,j} = \#\,(\text{data points assigned to bin } i, j)\,.$$

The idea behind linear binning is to split the unit mass of each data point, among the grid points that are its four nearest neighbors. This is done in a way that properly reflects distance to each grid point. In particular, first find the location of the nearest bin center,

$$\begin{aligned} i &= \mathit{round}((X - L_x) / \Delta_x), \\ j &= \mathit{round}((Y - L_y) / \Delta_y), \end{aligned}$$

where *round* is the function which rounds each observation to the nearest integer. The unit mass is split according to the weights

$$\begin{aligned} \delta_x &= (X - L_x) / \Delta_x - (i - .5), \\ \delta_y &= (Y - L_y) / \Delta_y - (j - .5), \end{aligned}$$

and then assigned as

bin	mass
i, j	$(1 - \delta_x)(1 - \delta_y)$
$i + 1, j$	$\delta_x (1 - \delta_y)$
$i, j + 1$	$(1 - \delta_x) \delta_y$
$i + 1, j + 1$	$\delta_x \delta_y$

Good visual insight into these weights comes from Figure D.3 of Wand and Jones (1995).

Some account needs to be made for data which lie outside the bivariate interval $[L_x, L_x + n\Delta_x] \times [L_y, L_y + m\Delta_y]$. Two approaches may be appropriate, depending on the context. One is to simply ignore points that are outside, i.e. not count them in the binning process. The other is to move them so that they lie at the nearest boundary point, and then proceed with binning.

2.2 Estimation

In addition to computational speed, another advantage of binning is that then density estimation can be done with nearly the same algorithms as for nonparametric regression (with an "equally spaced design"). This happens via replacement of the matrix of regression data values, with the matrix of bin counts, denoted as C above. In particular, the density is estimated by the matrix

$$\widetilde{f}_h = N^{-1}\left(C * \widetilde{K}_h\right),$$

where $*$ denotes bivariate discrete convolution, and where $\widetilde{K}_h$ is a matrix of evaluations of the kernel function K_h. This should be viewed as an approximation of $\widehat{f}_h$. Estimates of partial derivatives have a similar simple convolution form,

$$D\widetilde{f}_h = N^{-1}\left[C * \left(D\widetilde{K}_h\right)\right], \tag{1}$$

where D denotes various partial derivative operators, including $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, $\frac{\partial^2}{\partial x^2}$, $\frac{\partial^2}{\partial x \partial y}$ and $\frac{\partial^2}{\partial y^2}$.

This formulation allows nearly direct application of some aspects of the image analysis version of S^3, as developed in Godtliebsen, Marron, and Chaudhuri (1999). A very important difference is that estimation of the local variance of the partial derivative estimates is different (because the c_{ij} are counts). The basis for our variance estimate is the fact that $\widehat{f}_h(x,y)$ is a simple average of i.i.d. random variables. Thus a sensible estimate is the usual sample variance,

$$\begin{aligned}
\widehat{var}\left[D\widehat{f}_h(x,y)\right] &= \widehat{var}\left[N^{-1}\sum_{k=1}^{N} DK_h(x-X_k, y-Y_k)\right] = \qquad (2)\\
&= N^{-1}s^2\left\{DK_h(x-X_k, y-Y_k) : k = 1, ..., N\right\} = \\
&= N^{-1}\left\{\frac{1}{N-1}\left[\sum_{k=1}^{N}\left(DK_h(x-X_k, y-Y_k)\right)^2 - N\left(D\widehat{f}_h(x,y)\right)^2\right]\right\} = \\
&= \frac{1}{N-1}\left\{N^{-1}\sum_{k=1}^{N}\left(DK_h(x-X_k, y-Y_k)\right)^2 - \left(D\widehat{f}_h(x,y)\right)^2\right\}.
\end{aligned}$$

The argument of the square of the second term inside the braces is approximated by $D\widetilde{f}_h$ as at (1) above. The first term inside the braces needs the new binned approximation

$$N^{-1}\left\{C * \left[\left(D\widetilde{K}_h\right)^2\right]\right\}.$$

The resulting approximated version of the variance estimate (2) is then used directly as the local variance in the formulas for the image analysis version of S^3.

Another important difference between density estimation and image analysis is that the former often has large regions with no data, e.g the upper left and the lower right of the scatterplot shown in Figure 1. Hence data sparsity issues need much more attention. The approach taken is the same as in Chaudhuri and Marron (1999) via the concept of "Effective Sample Size". The idea is to take a "kernel weighted count" of the number of points in each window. This motivates the definition

$$ESS_{i,j} = \frac{\sum_{k=1}^{N} K_h\left(x_i - X_k, y_j - Y_k\right)}{K_h(0,0)} \approx \frac{C * \widetilde{K}_h}{K_h(0,0)}.$$

Using the standard Binomial rule of thumb, we say that the Gaussian approximation on which S^3 is based is inadequate when "$np < 5$". Thus in the present case, we call the data "too sparse for inference" at the location i, j when $ESS_{i,j} < 5$. In the imaging version of S^3 such points were marked with green circles. This was visually effective, because there were either no such locations, or else there were some only in strips near the boundary, or else the data was sparse everywhere. But this approach was not effective for density estimation, because the much larger regions of data sparsity yielded distractingly large regions of green circles. A better approach was to plot no symbols (i.e. no arrows and no circles of the type in Figure 2) in regions of data sparsity, as this gave less distraction.

2.3 Distributions and Significance

The statistical inference of S^3 is based on the fact that derivatives of smooths satisfy central limit theorems, i.e. have limiting Gaussian distributions. An important issue is multiple comparisons, since there are essentially a large number of simultaneous tests being performed. This is addressed via the "number of independent blocks" approach developed in Section 2.4 of Godtliebsen, Marron and Chaudhuri (1999).

The statistical underpinning of the arrows shown in the left part of Figure 2 is a hypothesis test about the statistical significance of the magnitude of the gradient. The square of the magnitude has a limiting χ^2 distribution so such testing is straightforward. See Section 2.4 of Godtliebsen, Marron and Chaudhuri (1999) for details. Depending on the size of the desired binning grid (we have found 64×64 to be generally reasonable as a trade off between resolution and computation time), the arrows drawn by S^3 may be too short for good visual impression. To address this, we allow "pooling across pixels" by combining them into 2×2 blocks. On each block, the 4 hypothesis tests are performed, and an arrow representing the coordinate-wise average direction, whose length is proportional to the number of significant results, is drawn.

The statistical underpinning of the dots shown in the right part of Figure 2 are based on a hypothesis test about the eigenvalues of the Hessian matrix. The distribution theory here is more complicated, but can be expressed in terms of a single distribution that has been tabulated by simulation. See Sections 2.5 and

6.3 of Godtliebsen, Marron and Chaudhuri (1999) for details. Different colors are used to represent different types of significant curvature according to the following scheme. Denote the eigenvalues of the Hessian matrix as $\lambda_- \leq \lambda_+$, and assign colors to dots as:

Idea	Color	λ_-	λ_+
hole	Yellow	sig. > 0	sig. > 0
long valley	Orange	not sig.	sig. > 0
saddle point	Red	sig. < 0	sig. > 0
long ridge	Purple	sig. < 0	not sig.
peak	Blue	sig. < 0	sig. < 0

We have found that beginners to the S^3 method find that combined arrows and dots present too much information at once. While this effect appears to mitigate with experience (we personally prefer seeing the added information), it seems important to first start with the simpler versions. Hence our Matlab software uses a default of arrows only, as shown in the right of Figure 2. Other versions are easily employed through changing parameters.

The streamline version of S^3 uses significant pixels as calculated for the arrow approach (but no combining into 2×2 blocks). Details are given in Section 3.2 of Godtliebsen, Marron and Chaudhuri (1999). An unfortunate feature of the movie version is that the streamlines are computed frame by frame. This means that different random locations are used in different frames, which results in substantial "jitter" in the movies. The development of a version of S^3 that computes streamlines properly through the whole scale space (i.e. the *same* streamline is used at all scales for which it appears) is an interesting open problem.

3 More Examples

Additional real data examples are presented in this section, which show additional aspects of S^3.

Figure 5 shows S^3 applied to the Old Faithful Geyser data, from Table B6 of Scott (1992), who references Weisberg (1985). The y coordinate is the time of duration of one eruption of the geyser, and the x coordinate is the duration of the previous eruption. The bright spots suggest three modes in this bivariate distribution. The gradient & curvature version of S^3, in Figure 5a, shows that all three modes are statistically significant. In particular, each mode has at least some light purple dots on the top, and there are dark blue dots on the largest mode (upper right). Dark blue dots also show up for the upper left mode at some different scales, see the movie version in the file `ssskde1fig5a.mpg` at the above location. Perhaps more conclusive evidence in favor of the three modes is the occurrence of red saddle points at the ridges connecting the modes.

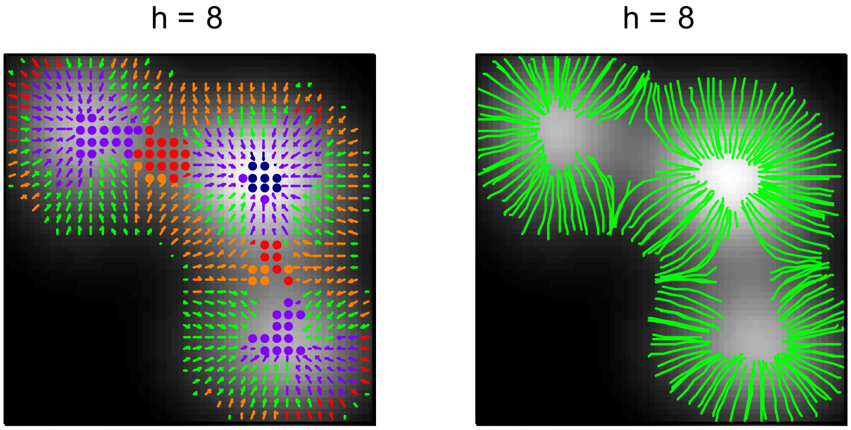

Figure 5: *Combined gradient and curvature (left) and streamline (right) versions of Significance in Scale Space for the $h = 8$ kernel density estimate, for the Old Faithful Geyser data.*

The streamline version of S^3 in Figure 5b, is less conclusive about the modality, since it does not clearly separate the modes from ridges. This is because the streamlines only use gradient information, while the important locations in Figure 5a had colored dots, indicating that the curvature, but not the gradient, was significant. Of course it is well worth studying other scales. The movie version, in the file `ssskde1fig5b.mpg` at the above location, shows at somewhat smaller bandwidths, that the upper left mode can be distinguished by streamlines alone, but not the lower right mode. The situation where the gradient - curvature version of S^3 found "vague structure" more often than the streamline version was fairly typical of our experience with other examples in density estimation (but there are exceptions, as shown in Figures 3 and 4). This contrasts with our experience for images, as reported in Godtliebsen, Marron and Chaudhuri (1999).

It is now known that the three modes in the Geyser data correspond to two either "long" or "short" eruptions (with little in between), and a short eruption never follows a short eruption, and there is a physical explanation of this.

Figure 6 shows the performance of S^3 on the aircraft data discussed in Section 1.3 of Bowman and Azzalini (1997). The original data are 6 variables reflecting features of aircraft, that were summarized by principal component analysis. The x axis is the first principal component, which turned out to represent those variables reflecting "size". The y axis is the second principal component, which represents "speed, adjusted for size". The bright spot at the lower left shows that most aircraft are neither large nor fast. There is a high density "arm" extending in the direction of small but fast aircraft, and also in the direction of very large, and somewhat faster aircraft. Bowman and Azzalini (1997) suggest that the data are trimodal (with the arms as modes), but S^3 does not quite find the three modes. This could well be because S^3 is an "omnibus" type of hypothesis test, which attempts to be "powerful in all directions", which entails some trade-off in power in specific directions. Methods which might provide stronger evidence of the trimodality would be based on formal "mode tests", although the literature for that mostly focuses on the one dimensional case.

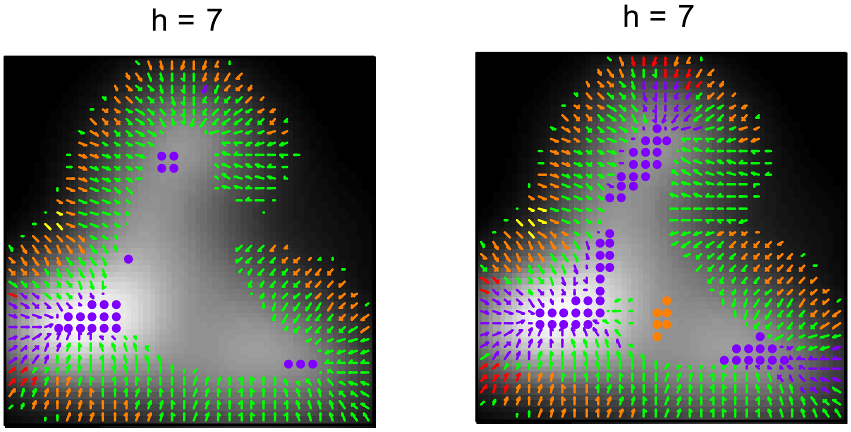

FIGURE 6: *Combined gradient and curvature versions of Significance in Scale Space for the* $h = 7$ *kernel density estimate, at the level of significance* $\alpha = 0.01$ *(left) and* $\alpha = 0.2$ *(right), for the aircraft data.*

The scale in both parts of Figure 6 is $h = 7$, chosen by studying the movie version (available in the files `ssskde1fig6a.mpg` and `ssskde1fig6b.mpg` at the above location) and trying to maximize the impression of trimodality. Figure 6a shows S^3 using the low significance level of $\alpha = 0.01$, while Figure 6b shows the much higher level of $\alpha = 0.20$. As expected, the less stringent hypothesis tests underlying Figure 6b result in more features being flagged as significant, although still not enough to conclude trimodality.

An aspect of Figure 6 that is not representative of our experience with varying α is that there are quite a few more significant features in Figure 6b. For other data sets, there tends to be less difference, even for such a large range of α values. The reason this happens here, is because for many locations, the features just happen to be near the boundary between significant and insignificant.

References

[1] Bowman, A. W. and Azzalini, A. (1997) *Applied smoothing techniques for data analysis, the kernel approach with S-plus illustrations*, Oxford Science Publications, Oxford.

[2] Chaudhuri, P. and Marron, J. S. (1997) Scale space view of curve estimation, Institute of Statistics, Mimeo Series # 2357.

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

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

[5] Helman, J. and Hesselink, L. (1989) Representation and display of vector field topology in fluid flow data sets, *IEEE Computer*, 22, 27-36.

[6] 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.

[7] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association*, 91, 401-407.

[8] Kim, C. S. and Marron, J. S. (1999) SiZer for jump detection, unpublished manuscript.

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

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

[11] 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.

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

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

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

[15] Weisberg, S. (1985) *Applied Linear Regression*, Wiley, New York.