The Mode Tree: A Tool for Visualization of Nonparametric Density Features

Michael C. Minnotte [1,2]
David W. Scott [1]

Department of Mathematics and Statistics
Utah State University
Logan, Utah 84322-3900

and

Department of Statistics
Rice University
Houston, Texas 77251-1892

Research supported by
[1] the Office of Naval Research under contract N00014-90-J-1176 and
[2] the Department of Defense NDSEG Fellowship grant number DAAL03-90-G-0169.
The authors wish to thank the readers for comments that led to improvements in the paper.

Abstract

Recognition and extraction of features in a nonparametric density estimate is highly dependent on correct calibration. The data-driven choice of bandwidth h in kernel density estimation is a difficult one, compounded by the fact that the globally optimal h is not generally optimal for all values of x. In recognition of this fact, a new type of graphical tool, the *mode tree*, is proposed. The basic mode tree plot relates the locations of modes in density estimates with the bandwidths of those estimates. Additional information can be included on the plot indicating such factors as the size of modes, how modes split, and the locations of antimodes and bumps. The use of a mode tree in adaptive multimodality investigations is proposed, and an example is given to show the value in using a Normal kernel, as opposed to the biweight or other kernels, in such investigations. Examples of such investigations are provided for Ahrens' chondrite data and van Winkle's Hidalgo stamp data. Finally, the bivariate mode tree is introduced, together with an example using Scott's lipid data.

Keywords: Mode estimation; Bump hunting; Multimodality; Graphical methods; Kernel density estimation; Kernel choice.

figure=/home/minnotte/bumps/tree/revFigs/kernex.ps

Figure 1: Normal kernel density estimate of the chondrite data with $h = 1.0$, along with the data points (at bottom) and the individual kernel masses which make up the estimate.

1 Introduction

Structure in data sets reflects features of the underlying density function. Features of interest include modes, or local maxima; antimodes, or local minima; and "bumps," or regions where the second derivative is negative. In an exploratory setting, the number and location of these features is not known a priori. The parametric form of the density is generally poorly understood and any tentative choice may be too restrictive. Under such circumstances, a nonparametric density estimation technique can be highly valuable.

Kernel density estimation is a popular example of nonparametric density estimation (Rosenblatt, 1956; Parzen, 1962). Given a sample $\{X_1, \ldots, X_n\}$ of size n, the kernel density estimate at x is computed as

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

where $K_h(t) = K(t/h)/h$ and K is known as the kernel. The kernel estimator is equivalent to a mixture density with the function K equally weighted and centered at each X_i. If the kernel K itself is assumed to be a density function, then $\hat{f}_h(\cdot)$ is nonnegative and integrates to one. K is generally taken to be symmetric with mean 0 and positive variance. The statistical properties of the modes of $\hat{f}_h(\cdot)$ have been studied by Parzen (1962) and Eddy (1980).

In Figure 1, a kernel density estimate and its component kernels are displayed for the chondrite meteorite data (Ahrens, 1965). These data, the percentages of silica in 22 chondrite meteors, were first discussed in the bump-hunting context by Good and Gaskins (1980). The vertical lines below the x-axis represent the values of the individual data points. The Normal kernel,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

is used in this estimate. We shall return to the example of the chondrite data several times.

After K has been chosen, the remaining element to be specified is the bandwidth parameter h. The bandwidth is a rescaling factor which determines the extent of the region over which the probability mass for point X_i is spread. The choice of this parameter is quite crucial to the final density estimate both locally and globally. In Figure 2, we display twenty different Normal kernel density estimates of the chondrite data, with bandwidths ranging from 0.2 to 3.0. The amount of information about the number and location of pofigure=/home/minnotte/bumps/tree/revFigs/manyh.ps

Figure 2: Twenty Normal kernel density estimates of the chondrite data. Bandwidths are equally spaced on a logarithmic scale, and range from 0.2 to 3.0.

tential bumps and modes is quite large in this figure, to the point of being overwhelming. The number of modes and bumps increases rapidly as $h \to 0$.

Unfortunately, h must be selected by the user, and no completely satisfactory method of doing so has been found. [Scott (1992) presents a survey of available techniques that focus on minimizing L_2 error, a criterion which is only loosely related to bumps and modes.] Even assuming the best global choice for h, the fact remains that no single value of h will perform well for all points x (Terrell and Scott, 1992), as we will demonstrate in Section 4. Jones (1990) and Terrell and Scott (1992) have investigated the theoretical and practical advantages of an adaptive kernel estimate introduced by Breiman, et al. (1977) and Abramson (1982):

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(x - X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x - X_i}{h_i}\right), \tag{2}$$

where h varies with sample point X_i. Unfortunately, the original problem of choosing one smoothing parameter h in Equation 1 has been replaced by the necessity of choosing many smoothing parameters in Equation 2. In a related approach, Wand, Marron, and Ruppert (1991) have examined the use of data transformation families in estimator (1) to address the problem. Here, the difficulty of choice of bandwidth remains, plus there is the additional problem of choice of transformation. No method seems to be sufficiently general to work for all densities or data sets.

In this paper, when we focus our attention on a particular mode, we desire an adaptive estimator. We compromise by using a fixed, but different, bandwidth for each potential mode. Graphical tools help us in this investigation and in the organization of our results.

2 The Mode Tree

The difficulties inherent in the choice of a specific h suggest that another method of viewing the data might be appropriate, one which summarizes information from density estimates calculated for a large variety of values of h. One possibility is the overlaid density plot shown in Figure 2. While such a figure certainly has value, it is difficult to pull out information on any single estimate. Another option, used by Roeder (1990), treats h as a second independent variable. Here, a sequence of density estimates is viewed using a perspective plot of the function $\psi(x, h) = \hat{f}_h(x)$; see Figure 3. While we find this plot more informative than Figure 2, its use also invokes all of the disadvantages of perspective figure=/home/minnotte/bumps/tree/revFigs/chonpersp.ps

Figure 3: Perspective plot of $\psi(x,h) = \hat{f}_h(x)$ for the chondrite data. Forty density estimates are plotted, with bandwidths ranging on a logarithmic scale from 0.2 (at the rear of the plot) to 3.0 (at the front).

figure=/home/minnotte/bumps/tree/revFigs/basnormtree.ps

Figure 4: Normal kernel mode tree for the chondrite data. Solid lines represent mode locations at each bandwidth. The horizontal dashed lines indicate the "splitting" of a mode.

plots, most notably hidden features and difficulties in interpreting the complex surface. The figure seems to overemphasize estimates with small bandwidths. The eye is drawn to the high, sharp peaks at the back of the plot, to the neglect of the lower and broader, but potentially more realistic, features found at wider bandwidths.

Another possibility is a dynamic display on a computer screen, in which the estimate varies as the user interactively adjusts the smoothing parameter. Tierney (1990) describes such a procedure. While this is potentially a powerful tool with which to investigate estimates calculated using a variety of bandwidths, it is limited by the need for a computer screen to display, not merely calculate, the results. Static plots require only paper or transparency to display once their computations have been performed.

If the key information about the structure in an estimate at a given bandwidth can be condensed to a 1-dimensional plot, then the second dimension can be allocated to h, without the problems of perspective plots. As the most easily visible changes in a density estimate due to varying h have to do with the number, location, and size of modes (i.e., the "wiggliness" of the density), these features provide a good 1-dimensional surrogate for the density estimate as a whole. In addition, when there is strong reason to believe that multiple modes are present in the true density, they are often among the most important and interesting features. Thus, in lieu of presenting the entire collection of density estimates, information about their modes is a valuable contribution.

The basic *mode tree* is very simple to define. The mode locations are plotted against the bandwidth at which the density estimate with those modes is calculated. In Figure 4, the solid vertical lines represent the modes corresponding to those in the density estimates in Figure 2 for the chondrite data. A larger number of values of h (here, 200) is used in producing the mode tree than in the superimposed plot. Notice the choice of the logarithmic scale for the vertical axis. The values of h should be chosen to be equally spaced on a logarithmic scale, as large changes at high values of h have less of an effect on the density estimate than smaller changes at lower values of h. If the mode locations are plotted for all value of h, then a set of lines, or *mode traces*, will result. In the Appendix, a matching algorithm is described to determine the mode traces derived from a *finite* set of mode points figure=/home/minnotte/bumps/tree/revFigs/basbiwtree.ps

Figure 5: Biweight kernel mode tree for the chondrite data. Note the modes which appear briefly and then vanish, as well as the modes which continue after the point at which they split.

so that they can be connected between levels.

It is useful to think of modes "splitting" as h decreases, or, alternatively, "coalescing" as h increases. Between any pair of adjacent modes is a valley containing exactly one antimode. As h decreases, a critical or saddle point appears between two adjacent modes and thereafter a new mode and antimode appear. This event will be apparent from the increased number of mode locations. The new mode is the one at the smaller value of h not matched to any mode from the previous, slightly greater, value of h. Depending on the relative location of the new mode/antimode pair, either the left or the right mode may be thought of as splitting. If the new mode appears to the right of the new antimode, it is the left mode which has split; otherwise, the right. Thus the mode tree may be made more informative by adding the horizontal connections seen as dashed lines in Figure 4. These connections show the relationships between new modes and the old modes from which they split. These connections give this plot the structure which justifies the "tree" label.

The mode tree in Figure 4 was calculated using a Normal kernel, as in Figures 1, 2, and 3. This choice was quite deliberate. Silverman (1981) showed that for a Normal kernel, the number of zeroes in all derivatives of $\hat{f}_h$ is monotone decreasing in h. These zeroes include all modes, antimodes, and inflection points. Silverman's proof hinges on the fact that the Normal density is totally positive and that the convolution of a Normal kernel density estimate with bandwidth h_1 and a Normal density with standard deviation h_2 is also a Normal kernel density estimate with bandwidth $h_3 = (h_1^2 + h_2^2)^{1/2}$. The desirable behavior in $\hat{f}_h$ can be clearly seen in Figure 4. All modes found at a given level of h remain as h decreases.

Surprisingly, this result need not be true for other kernels. Figure 5 shows the mode tree generated by the same chondrite data set but using the biweight kernel,

$$K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}(x).$$

Keep in mind that smoothing parameters for different kernels (and thus the y-axes of Figures 4 and 5) are not directly comparable. Sample modes appear and disappear in an irregular fashion. The appearance of such distracting features in practical situations is not well recognized. These spurious peaks are generally quite small and are more accidents of the estimation method than true features of the data. Similar results have been observed in estimates from other members of the Beta family (of the form $c_m(1 - x^2)^m$, where m is an integer and c_m is the constant which makes the kernel integrate to 1 in the range

$[-1, 1]$), even though the family approaches a Normal kernel as $m \to \infty$. In addition, while a mode tree generated using a Cauchy kernel appeared well-behaved, indicating that the Normal kernel may not be unique, those generated using the "arctan" kernel

$$K(x) = \frac{3^{5/4}}{(2\pi)^{3/2}} \left[\frac{\pi}{2} - \arctan\left(\frac{3^{9/4}|x|^3}{(2\pi)^{3/2}} \right) \right]$$

(with infinite support but discontinuous third derivative) and the "bump" kernel

$$K(x) = 2.25228 \exp\left(\frac{-1}{1 - x^2} \right) I_{[-1,1]}(x)$$

(with finite support but infinitely continuous derivatives) did not. Thus neither infinite support nor continuous derivatives alone appear to guarantee good behavior of the mode tree, and we conclude that the Normal kernel should be used for mode tree computations.

3 Enhancements to the Mode Tree

As we have shown, the basic mode tree, in and of itself, can be quite informative and useful. Yet the possibility exists to significantly increase the information presented, thereby further improving our understanding of the data.

Our first enhancement is to add a dimension of information through the widths of the mode traces in the mode tree. Typically, that information relates to some feature of those modes. In Figure 6 we plot an enhanced mode tree of the chondrite data. The lines of Figure 4 have been replaced by the black regions centered on each mode location and whose horizontal width at each level of h represents the probability mass of the mode at that bandwidth. More precisely, the width is proportional to the quantity

$$M_j(h) = \int_{a_j(h)}^{b_j(h)} \left[\hat{f}_h(x) - \max\left(\hat{f}_h(a_j(h)), \hat{f}_h(b_j(h)) \right) \right]_+ dx, \tag{3}$$

where $a_j(h)$ and $b_j(h)$ are respectively the left and right antimodes surrounding mode j in $\hat{f}_h(x)$ (and which may be $-\infty$ and $+\infty$, respectively), and the "+" denotes the "positive part" of the argument. In Figure 3, the values of M_1, M_2, and M_3 when h is 1 are equal to the indicated shaded areas. This value is representative of the "size" of the mode, and in fact can be used as a statistic to test the reality of the mode in question when the bandwidth is chosen appropriately (Minnotte, 1992). M_j is the minimal L_1 distance from the density to the set of continuous functions without a local maximum between the observed antimodes in the density function. M_j is also the single-mode equivalent of Müller and Sawitzki's (1991) excess mass functional. The excess mass functional measures the amount of probability mass above the contour $f(x) \geq c$ and is not locally defined as in Equation 3.

figure=/home/minnotte/bumps/tree/revFigs/enhnormtree.ps

Figure 6: Enhanced Normal mode tree for the chondrite data. The widths of the black regions are proportional to M_j for the modes at each level. The dots represent the locations of antimodes, and the gray regions are the bumps.

A second enhancement which is also simple to include on the plot are the locations of antimodes. The enhanced mode tree for the chondrite data is shown in Figure 6 using dotted lines for the antimode traces. Observe that the additional information does not appreciably clutter the diagram as the mode and antimode traces are approximately parallel and do not cross.

In addition to modes, authors such as Good and Gaskins (1980) have investigated the presence of bumps or regions where the second derivative of a density estimate is negative. These can be easily added to the mode tree. While inflection points could simply be plotted with one or two additional line types, shading the entire area of each bump, as in the gray regions of Figure 6, is much more visually striking and suggestive. Clearly, at any given h, each mode has a surrounding bump, but the reverse is not necessarily true. A bump can occur on a slope, where the second derivative changes sign, but the first derivative does not. For an example, see the right slope of the tallest peak in Figure 12. It is interesting to note that a new mode will always start at the boundary of its shaded bump region. This is to be expected, but it is still interesting to view graphically.

Finally, important values of either x or of h can be highlighted. Values of the sample mean, median, or quartiles might be shown. Appropriate special values of h include cross-validation bandwidths or Terrell and Scott's (1985) oversmoothed Normal bandwidth given by

$$h_{\mathrm{os}} = \frac{3\hat{\sigma}}{(35)^{1/5}} \left(\int_{-\infty}^{\infty} \phi^2 \right)^{1/5} n^{-1/5} = 3\hat{\sigma}(70\sqrt{\pi}n)^{-1/5},$$

where $\hat{\sigma}$ is the sample standard deviation. The oversmoothed bandwidth h_{os} and the sample median and quartiles for the chondrite data are indicated in Figure 6.

It should be noted that the undesirable behavior observed for non-Normal kernels illustrated in Figure 5 is not limited to the mode. In Figure 8, a plot is shown of modes, bumps, and antimodes, from biweight kernel estimates of the chondrite data. Clearly, the same undesirable behavior observed in the modes is occurring in the bumps and antimodes as well. Note, for example, the behavior near $x = 31$ and $h = 1.5$. An antimode briefly becomes a mode, with surrounding bump, before quickly becoming an antimode again. Such occurrences, along with the highly erratic behavior of modes, antimodes, and inflection points, all point to how strongly inappropriate the biweight kernel is for multimodality investigations and bump hunting, especially when using a concept such as Silverman's (1981) "critical bandwidths." However, we note that the undesirable analytical behavior is limited to relatively flat regions near a mode or antimode and is relatively difficult to discern in a figure=/home/minnotte/bumps/tree/revFigs/enhbiwtree.ps

Figure 8: Enhanced biweight mode tree for the chondrite data. In comparison to Figure 6, note the erratic behavior of modes, antimodes, and bumps.

figure=/home/minnotte/bumps/tree/revFigs/falsemode.ps

Figure 9: Biweight estimate of the chondrite density with bandwidth $h = 1.5$. The indicated mode appears only within the range of bandwidths $1.45 < h < 1.67$.

plot of the density estimate. For example, Figure 9 indicates the mode mentioned above on a plot of the biweight estimate of the chondrite density with $h = 1.5$. This is the largest that this mode appears for any bandwidth.

4 An Application to Investigations of Multimodality

Minnotte (1992) demonstrates an additional use of the mode tree in the investigation of multimodality. The null hypothesis of the proposed test is that a mode in question is simply an artifact of the data, against an alternative that the mode is a true feature of the population density. Thus, the test is of the reality of individual modes, rather than of the number of modes of the density as a whole. A method is needed to choose the proper bandwidth at which to test each mode. The basic mode tree (with split-mode connections) fulfills this function admirably.

The test statistic M_j defined in Equation 3 is monotone decreasing in h. Each mode is tested at the smallest value of h for which the mode is still entirely a single entity, that is, at the value of h one step larger than the level at which the mode splits. Such bandwidths are the same as Silverman's critical bandwidths, but are associated with particular modes and not the overall number of modes observed at that level of h. Testing at these levels ensures that the test statistic will be as large as possible. The test involves recomputing M_j from bootstrap resamples to obtain an empirical p-value and is described in Minnotte (1992). The bootstrapping follows Silverman (1981), except in choice of bootstrap density. Here, in order to conform to the null hypothesis, the tested density is adjusted by moving part of the probability mass of the mode in question into one or both of the surrounding valleys.

Minnotte (1992) demonstrates that in the cases of unimodal and bimodal densities (to which all densities reduce due to the asymptotic locality of kernel estimators), the test statistics of the first two modes tested converge to zero and the equivalent masses of the true density, respectively. The proofs are based largely on results in Mammen, Marron, and Fisher (1992).

After the mode tree is used to choose the bandwidths at which to test, the results can figure=/home/minnotte/bumps/tree/revFigs/chontree1.ps

Figure 10: Mode tree used for multimodality testing of the chondrite data. Filled circles represent modes tested as significant at the $\alpha = 0.15$ level, open circles are modes which are not significant at this level.

be displayed on the mode tree. For example, with the chondrite data, the results of the test are indicated just above each split in the unenhanced mode tree of Figure 10. Filled circles indicate that the null hypothesis has been rejected at the fairly generous $\alpha = 0.15$ level (suggesting that the mode is real), while open circles indicate a failure to reject the null hypothesis at that level. The actual estimated p-values for the 3 modes are 0.061, 0.015, and 0.005 from left to right. Choosing an α-level greater than 5% has been commonly recommended for this problem; see Matthews (1983) and Izenman and Sommer (1988). Clearly, a lower choice of α will lead to a more conservative selection of possible modes. In an exploratory setting, with moderate sample sizes, a fairly high choice of α seems reasonable.

The number of significant modes estimated from the data is calculated in a recursive fashion. Each mode tested leads to two branches (which may have one or more tests themselves). If the mode is non-significant, the number of modes it passes up to its parent mode in the tree is simply the sum of the number of modes passed up by its two branches. If the mode is significant, it passes that sum or 1, whichever is greater. This means, for example, that the significant modes near $x = 33.4$ in Figure 10 are all counted as a single "real" mode, while there are three such modes in the entire distribution. Non-significant modes at the same location but different bandwidths as those found significant are disregarded, being viewed as simply being tested at inappropriate bandwidths. Either h is too large, and the mode has not yet appeared sharply, or h is too small, and too much of the mode's mass has been separated into other modes which have already split off. The mode tree is an integral part of summarizing this recursive procedure.

It should be noted, that while a single test, especially one specified in advance (by theory), satisfies the requirements of classical hypothesis testing, the simultaneous nature of the large number of tests on the mode tree does not. Thus, it may be more appropriate to consider the entire mode tree with tests as an exploratory device, useful in indicating which modes may be worthy of further study. This may include modes missed by tests such as Silverman's which may be more conceptually valid, but which are handicapped by the use of single bandwidths.

In Figure 11, we show the mode tree for the 1872 Hidalgo stamp thicknesses ($n = 485$) measured by Walton van Winkle and analyzed in Wilson (1983) and Izenman and Sommer (1988). Wilson hand-smoothed a histogram on the 437 unwatermarked stamps without gum, and came to the conclusion that the data were bimodal, with modes near .077 mm and .105 mm. Izenman and Sommer applied the critical-bandwidth test described figure=/home/minnotte/bumps/tree/revFigs/stamptree.ps

Figure 11: Mode tree used for multimodality testing of the Hidalgo stamp data. Filled circles are represent modes tested as significant at the $\alpha = 0.15$ level. Open circles represent modes not judged significant at this level.

figure=/home/minnotte/bumps/tree/revFigs/stampest.ps

Figure 12: Kernel density estimates for the Hidalgo stamp data; $h = 0.0005$. Modes found significant at the $\alpha = 0.15$ level are indicated by arrows.

in Silverman (1981) to the full set of 485 measurements. Their results supported Wilson in rejecting unimodality, but indicated *seven* modes (located at .072, .080, .090, .100, .110, .120, and .130 mm), rather than Wilson's two. When Izenman and Sommer applied a square root transformation to the data in an effort to check for spurious modes in the *right* tail, Silverman's test supported nine modes, the same seven modes, plus an additional two in the *left* tail at .060 mm and .064 mm. We note that the small mode at .060 mm is based on only a single data point.

The mode tree shown in Figure 11 indicates that the transformation attempt may be misfiring in two ways. All seven modes of the original test are indicated, as well as the larger of the two small left-tail modes found by Izenman and Sommer after transformation. In addition, however, we found two additional and tantalizing modes at .075 mm and .115 mm. The estimated p-values for the ten modes from left to right are 0.148, 0.003, 0.109, 0.003, 0.112, 0.008, 0.013, 0.120, 0.085, and 0.023.

The more impressive of the new modes is at .075 mm, and may be clearly seen in Figure 12, as the medium-sized peak between the two large modes at .072 mm and .080 mm. It is not surprising that Silverman's test missed these modes, for by the time the bandwidth is small enough to distinguish these modes from their large neighbors, numerous spurious modes have appeared in the right tail of the density. This is a prime example of the inability of a single h to work well for all x, mentioned in Section 1. Under such circumstances, any procedure testing multimodality globally using a single bandwidth is likely to fail to recognize such modes.

5 Bivariate Mode Trees

Given that the mode tree is so useful in one dimension, the potential should be even greater in two dimensions when options such as Figures 2 and 3 are not available. While including all of the information of Figure 6 is probably impossible, a three-dimensional plot of mode locations and bandwidth is both feasible and useful.

An example of a bivariate mode tree is displayed in Figure 13. The data set is the figure=/home/minnotte/bumps/tree/revFigs/bivmodetree.ps

Figure 13: Bivariate mode tree for the standardized logarithm of the lipid data. The bandwidth h ranges from 0.2 (at the bottom of the plot) to 1.0 (at the top), equally spaced on a logarithmic scale. The contour plots represent the density estimates at the extreme bandwidths, and the bivariate mode traces have been projected onto the xh- and yh-planes.

320 lipid (cholesterol and triglyceride) levels of men with heart disease from Scott, et al. (1978), who observed a strong bimodal feature. A logarithmic transformation was applied in each dimension, and then the marginal data were standardized to have sample variance 1. This standardization is necessary, as a single value of h was used for the bandwidth in both dimensions. We note that such marginal standardization and use of a single bandwidth is not the optimal procedure for pure bivariate density estimation purposes. The contour plots of the bivariate Normal kernel density estimates at the top and bottom of the figure correspond to bandwidths of 1.0 and 0.2, respectively. The thick mode traces represent the modes for 201 bivariate kernel estimates with bandwidths between these levels (equally spaced on a logarithmic scale). To aid interpretation, the bivariate mode traces are projected onto the xh- and yh-planes. The lack of simple ordering on the plane makes the concept of "splitting" far more problematical than in the univariate setting. As connections would crowd an already complicated plot, and there are problems associated with them (both practical and conceptual), no attempt was made to draw such connections.

Only a single mode was found for values of h larger than 0.59, while 26 modes were found by the time h was 0.2. Note that several modes based on outliers appear before larger modes (as indicated by the lower contour plot) near the center of the density. This indicates that a bivariate version of Silverman's test would suffer from the lack of adaptivity mentioned in the discussion of the stamp data even more severely than the univariate version does. The bivariate multimodality testing problem should be quite challenging.

6 Discussion

The increases in computing power of the past few years have vastly enlarged the potential scope of our statistical toolbox. By utilizing the capabilities of today's machines, we no longer need restrict ourselves to one or even a few bandwidths in nonparametric density estimation. The mode tree (which perhaps should be called an *acrogram*, from the Greek root *acro* for "top" or "peak") is one direction in which this potential can be realized.

In addition to uses in multimodality investigations, the mode tree is also valuable in standard density estimation, by indicating how key features in a density estimate will react to changes in the bandwidth, and identifying interesting bandwidths. It also indicates figure=/home/minnotte/bumps/tree/revFigs/matchex.ps

Figure 14: Example of the point-matching algorithm described in the Appendix and used to connect the mode traces. The arrows indicate points among the two sets which are matched.

another direction for adaptive density estimation. The choice of bandwidth for given regions can be determined by the mode tree and the multimodality tests which use it.

A related problem is the consideration of extrema in a nonparametric regression curve; see Mammen (1991). The mode tree can be adapted to that setting.

Testing bumps as well as modes is also feasible, using either the bump information in an enhanced mode tree, or using a basic mode tree based on $\hat{f}'_h$, instead of $\hat{f}_h$. In any event, we have shown that the mode tree can be a valuable graphical tool in any data analysis for which bumps and modes are items of interest.

Appendix: A Univariate Point-Matching Algorithm

The vertical connections of the mode tree require a matching algorithm of some kind. The one used by the authors works as described in the following paragraphs and Figure 14.

The matching procedure will be described for ordered sets of points $\mathbf{a} = (a_1 < a_2 < \ldots < a_{k_a})$ and $\mathbf{b} = (b_1 < b_2 < \ldots < b_{k_b})$. For each $i = 1, \ldots, k_a$, set

$$\alpha_{i1} = \min\{\ell;\ |a_i - b_\ell| \leq |a_i - b_j| \quad \forall j = 1, \ldots, k_b\}.$$

Thus α_{i1} is the index of the member of $\mathbf{b}$ which lies closest to a_i. If $a_i < b_1$ or $a_i > b_{k_b}$, set α_{i2} to 0, otherwise

$$\alpha_{i2} = \begin{cases} \min\{\ell; b_\ell > a_i\}, & \text{if } a_i > b_{\alpha_{i1}} \\ \max\{\ell; b_\ell < a_i\}, & \text{if } a_i < b_{\alpha_{i1}} \\ \min\{\ell; |a_i - b_\ell| \leq |a_i - b_j| \quad \forall j \neq \alpha_{i1}\}, & \text{if } a_i = b_{\alpha_{i1}}. \end{cases}$$

Here α_{i2} gets the index of the closest member of $\mathbf{b}$ to a_i which lies on the other side of a_i from $b_{\alpha_{i1}}$. The values β_{j1} and β_{j2} are found by reversing $\mathbf{a}$ and $\mathbf{b}$ in the above formulas.

With the above four lists of indices, α_1, α_2, β_1, and β_2, we begin the actual process of matching. For each value $i = 1, \ldots, k_a$, we see if

$$\beta_{\alpha_{i1}1} = i,$$

which would indicate that a_i and $b_{\alpha_{i1}}$ are each the closest element in the other list to each other. If this is true, a_i and $b_{\alpha_{i1}}$ are considered matched to each other and removed from further consideration. Assuming the h's are picked to be reasonably close together, most of the modes should have been matched in this step. After all such pairs are removed, the remainder are investigated further. For any i such that a_i was not matched in the first cycle, if $b_{\alpha_{i1}}$ also remains unmatched and

$$\beta_{\alpha_{i1}2} = i,$$

a_i and $b_{\alpha_{i1}}$ are matched and removed. The process is repeated on the remainder two more times, the first time matching still unmatched pairs if

$$\beta_{\alpha_{i2}1} = i$$

and the second time matching any remaining pairs if

$$\beta_{\alpha_{i2}2} = i.$$

In each of these two cases, a_i is matched with $b_{\alpha_{i2}}$.

For the example in Figure 14, consider $a = (0.1, 0.3, 0.45, 0.6, 0.95)$ and $b = (0.15, 0.5, 0.8, 0.9)$. Then $\alpha_1 = (1, 1, 2, 2, 4)$, $\alpha_2 = (0, 2, 1, 3, 0)$, $\beta_1 = (1, 3, 5, 5)$, and $\beta_2 = (2, 2, 4, 4)$. The first round matches a_1 with b_1, a_3 with b_2, and a_5 with b_4. (For example, $\alpha_{31} = 2$ and $\beta_{21} = 3$.) Removing these pairs leaves a_2, a_4, and b_3. No pairs are matched in the second and third rounds, but since $\alpha_{42} = 3$ and $\beta_{32} = 4$, a_4 and b_3 are matched in the final cycle. The point a_2 remains unmatched.

By using a matching algorithm such as this, we can quickly match modes to those of the previous (next-highest) value of h. Any unmatched modes of the new set are considered new and attached horizontally as described in Section 2.

References

Abramson, I.S., (1982), "On Bandwidth Variation in Kernel Estimates—a Square Root Law," *The Annals of Statistics*, **10**, 1217–1223.

Ahrens, L.H., (1965), "Observations on the Fe-Si-Mg Relationship in Chondrites," *Geochimica et Cosmochimica Acta*, **29**, 801–806.

Breiman, L., Meisel, W., and Purcell, E., (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, **19**, 135–144.

Eddy, W.F., (1980), "Optimal Kernel Estimates of the Mode," *The Annals of Statistics*, **8**, 870–882.

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.

Izenman, A.J. and Sommer, C.J., (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, **83**, 941–953.

Jones, M.C., (1990), "Variable Kernel Density Estimates," *Australian Journal of Statistics*, **32**, 361–371.

Mammen, E., (1991), "Nonparametric Regression Under Qualitative Smoothness Assumptions," *The Annals of Statistics*, **10**, 1217–1223.

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.

Matthews, M.V., (1983), "On Silverman's Test for the Number of Modes in a Univariate Density Function", unpublished B.A. honors thesis, Harvard University, Dept. of Statistics.

Minnotte, M.C., (1992), "A Test of Mode Existence With Applications to Multimodality", unpublished Ph.D. thesis, Rice University, Dept. of Statistics.

Müller, D.W., and Sawitzki, G., (1991), "Excess Mass Estimates and Tests for Multimodality," *Journal of the American Statistical Association*, **86**, 738–746.

Parzen, E., (1962), "On Estimation of Probability Density Function and Mode," *Annals of Mathematical Statistics*, **33**, 1065–1076.

Roeder, K., (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association*, **85**, 617–624.

Rosenblatt, M., (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, **27**, 832–837.

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

Scott, D.W., Gotto, A.M., Cole, J.S., and Gorry, G.A., (1978), "Plasma Lipids as Collateral Risk Factors in Coronary Heart Disease—A Study of 371 Males With Chest Pain," *Journal of Chronic Diseases*, **31**, 337–345.

Silverman, B.W., (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society, Ser. B*, **43**, 97–99.

Terrell, G.R. and Scott, D.W., (1985), "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, **80**, 209–214.

Terrell, G.R. and Scott, D.W., (1992), "Variable Kernel Density Estimation," *The Annals of Statistics*, in press.

Tierney, L., (1990), *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, New York: John Wiley.

Wand, M.P., Marron, J.S., and Ruppert, D., (1991), "Transformations in Density Estimation," (with discussion), *Journal of the American Statistical Association*, **86**, 343–361.

Wilson, I.G., (1983), "Add a New Dimension to Your Philately," *The American Philatelist*, **97**, 342–349.