The Stochastic Mode Tree and Clustering

David W. Scott and William F. Szewczyk*

July 31, 2000

Abstract

The problem of finding the number and location of clusters is technically difficult because the problem is reasonably ill-posed. Hierarchical clustering methods produce dendrograms which may be pruned to select one of many different possible clusterings. Dendrograms may also be derived from novel approaches such as the mode tree (Minnotte and Scott, 1993), which records and plots the locations of all of the modes of a multivariate kernel density estimator for all bandwidths. Alternative clustering approaches such as k-means and mixture modeling give answers that may depend strongly on initial guesses. The purpose of this paper is to introduce an algorithm that fits an individual component of the mixture model, and to illustrate the use of the resultant estimated *probability density component* as a mechanism for locating clusters and generating a mode tree. However, the new *stochastic mode tree* does not simply find all modes within the kernel family $\{\hat{f}(x|h),\ 0 < h < \infty\}$, but rather performs local density estimation across the data domain. The resulting figure greatly reduces the number of plausible clustering configurations. An interactive algorithm for pruning noisy features is derived and illustrated with real data. The results may be used to locate clusters or to provide initial parameter values for fitting mixture models in any dimension.

1. Introduction

The search for structure in an unknown set of data $\{\mathbf{x}_i \in \Re^d,\ i = 1, \ldots, n\}$ is one of the most challenging problems in statistics. Graphical approaches can reveal important details about individual variables or pairs of variables, such as skewness, correlation, and clumping (Swayne, Cook, and Buja 1998). However, no graphical exploration can be expected to uncover all structure in higher dimensions. Without any prior information on a parametric density form, a nonparametric approach such as the multivariate kernel estimator

$$\hat{f}(\mathbf{x}) = \frac{1}{n\,|H|} \sum_{i=1}^{n} K\left(H^{-1}(\mathbf{x} - \mathbf{x}_i)\right) , \tag{1}$$

may be contemplated (Scott 1992). Once the matrix of smoothing parameters H is calibrated, the kernel estimate can be probed for the presence of simple forms of structure. The

*David W. Scott is Professor, Department of Statistics, MS-138, Rice University, Houston, TX 77005-1892 (Email: scottdw@stat.rice.edu). William F. Szewczyk is a senior mathematician with the National Security Agency (Email: wfszewc@afterlife.ncsc.mil).

most important features on the density surface are modes. Indeed, modes may naturally be associated with clusters (Hartigan 1975; Good and Gaskins 1980; Sager and Thisted 1982).

In general, cluster analysis provides a sophisticated toolbox for the discovery of structure in data, but clustering algorithms do not always rely upon probability models. Hierarchical clustering algorithms bypass the density function to form clusters iteratively. Starting with each data point as an individual cluster, the "closest" pair is combined into a new cluster recursively. A history of the process may be presented graphically in a binary tree known as the *dendrogram*; for example, see Figure 1 for a dendrogram of the male blue crab data (Ripley 1996). While almost entirely assumption- and model-free, the tree is not unique, depending heavily upon the choice of metric, definition of distance between two sets of points (clusters), and any initial transformation of the data. Finally, the dendrogram embodies many possible clusterings, and the task of choosing where to cut the tree remains. Without some model assumptions, this last task must remain largely *ad hoc*.

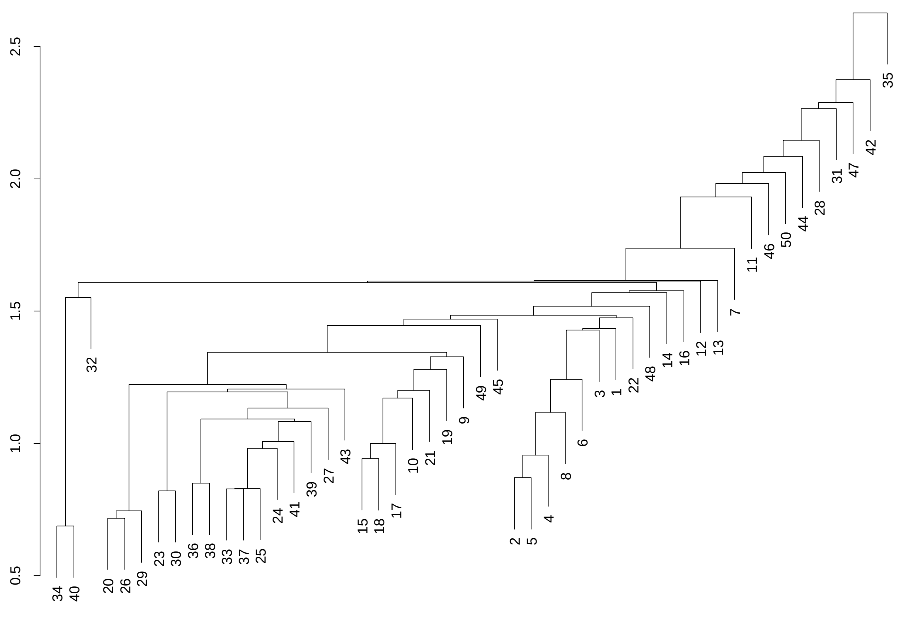

Figure 1: Single linkage dendrogram of sphered/blurred male blue *Leptograpsus* crab data.

An appealing flexible class of probability models is the mixture of Normals (MON),

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K} w_k \, \phi(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)\,, \tag{2}$$

where $\phi(\cdot|\boldsymbol{\mu}, \Sigma)$ denotes a multivariate normal density with mean $\boldsymbol{\mu}$ and covariance Σ. When fitted to data by the EM algorithm (Dempster, Laird, and Rubin 1977), this model provides a powerful and easily interpretable summary of the structure in the data by associating clusters with components. Cluster centers are given by the means, $\boldsymbol{\mu}_k$, cluster shapes by the covariance matrices, Σ_k, and cluster sizes by the weights, w_k. Observe the close relationship between MON and the kernel method in Equation (1), which is a highly constrained MON (with all the data "rounded" to one of the K values, $\boldsymbol{\mu}_k$, and all $\Sigma_k = H^2$).

In practice, a precise choice of K is not possible for high-dimensional data. In theory, any multivariate density may be represented by an infinite mixture of normals, but the representation is by no means unique in the L_2 sense (since the basis functions are not orthogonal). One cluster might be well-represented by a single multivariate normal, while another cluster may require dozens of normal components to capture skewness (and yet still be unimodal). The non-uniqueness grows exponentially with d and is yet another manifestation of the curse of dimensionality.

Furthermore, the stability of the estimates of the parameters of the mixture decreases as K increases (although the overall density estimate may be quite stable). Care must be exercised in the EM algorithm to avoid the introduction of Dirac spikes into the final estimate. An engineering solution is to take K much larger than necessary, carefully control the progress of the EM algorithm, so as to be relatively confident in the overall quality of $\hat{f}$ but with no expectation of interpretability of the estimated parameters. For example, one model of NIST speech data features $K = 2048$ mixtures in $d = 39$ dimensions (Reynolds 1995). Post-analysis (Scott and Szewczyk 2000) showed that K could be reduced by more than half with no change in the performance of $\hat{f}$ for discrimination. Of course, the parameters when $K = 1024$ are hardly more interpretable!

Banfield and Raftery (1993) have carefully constructed a hierarchy of mixture models by allowing Σ_k to be an identity matrix, I_d, an arbitrary diagonal matrix, proportional to each other, or fully unconstrained. The authors use BIC (Kass and Raftery 1995) to determine K. Such an approach generally results in a small K, but in practice has trouble with large dimensions ($d > 10$). The paradox is that as d increases, Σ_k should be less constrained, but practical issues require the Σ_k to be *more* constrained.

The relationship between the number of modes in a mixture density and K is a complicated issue; see Titterington, Smith, and Makov (1985). Obviously, the mixture density (2) may have many fewer than K modes. However, Scott and Szewczyk (1997) showed that the mixture (2) may have *more* than K modes when $d > 1$, but any extra modes tend to be very small. Boswell (1983) successfully probed the kernel surface in $d = 100$ dimensions. Numerical optimization can locate almost all sample modes in the kernel estimate. However, finding all the modes is expensive and no test exists that assures every mode has been found.

With kernel estimates, the use of the same smoothing parameter everywhere in $\Re^d$ leads to heavy oversmoothing in some regions and heavy undersmoothing in other regions. Indeed, Terrell and Scott (1992) show how H should vary locally (i.e., covariance shape vary as well as size). Thus in practice, any fixed kernel estimate may be expected to suppress modes in rough areas (oversmoothing) and to display false modes near peaks or tails in smooth areas (undersmoothing). Minnotte and Scott (1993) found a simple algorithm that avoids this local adaptivity problem. They suggest finding the locations of all modes as a function of the smoothing parameter. When $d = 1$ and $K = \phi$, the resulting figure, which they called a *mode tree*, looks remarkably like a dendrogram; see Figure 2. (Note that there are only 36 unique body diameter values among the 50 cases.) The special properties of the Normal kernel guarantee that the mode tree is a binary tree (Silverman 1981). When $h \approx 0$ each unique data point is its own mode. As h increases, pairs of modes collapse into each other until eventually a single final mode exists near $x = \bar{x} = 13.35$mm.

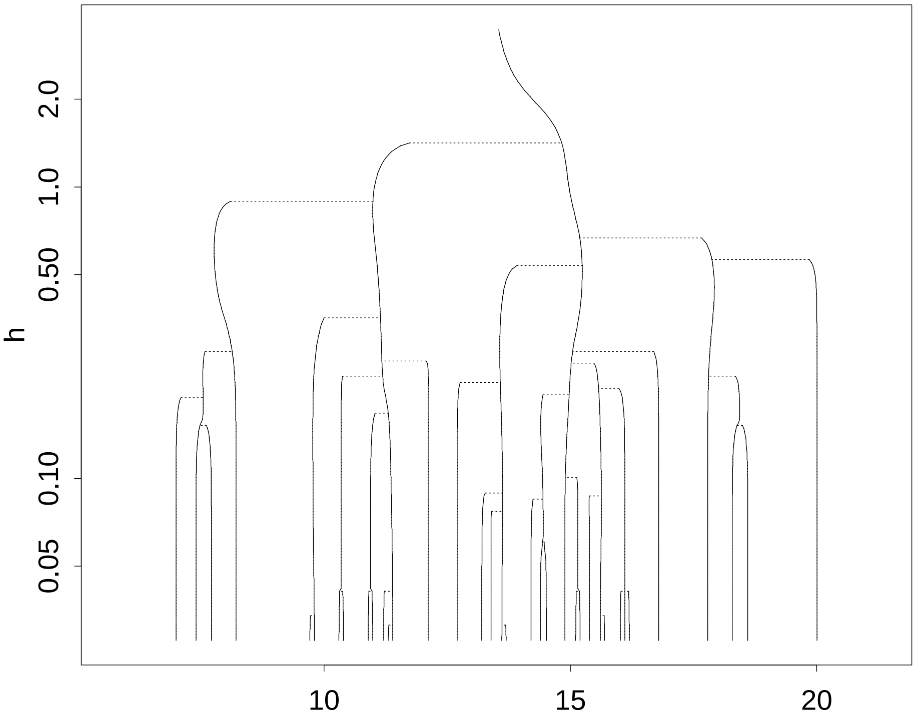

Figure 2: Mode tree of the body diameter (mm) of 50 male blue *Leptograpsus* crabs.

When $d > 1$, with $H = h \cdot I_d$, the authors show that the mode tree may still be formed. A smooth mode tree exists in $\Re^{d+1}$ but is difficult to display when $d > 2$. A simple algorithm converts the mode tree into a dendrogram (where the horizontal axis has no particular relationship to the original variables), with the "dissimilarity" measured by the bandwidth, h. For example, the mode tree using all 5 variables of the 50 male blue crabs is displayed in Figure 3. The result is somewhat similar to the single linkage clustering in Figure 1.

While no single kernel estimate with fixed bandwidth can be expected to show all of the real modes and only the real modes, the family of kernel estimates is likely to contain all of the modes. "Real" modes tend to persist (vertically) in the mode tree, while "false" modes cluster around real modes and seldom persist very long before splitting. Of course, singleton outliers ($x = 20$mm, e.g.) are exceptions, and the use of the logarithmic vertical scale greatly magnifies the finer details. Minnotte (1997) demonstrated one way to "test" the veracity of a mode at each split; see also Hartigan and Hartigan (1985), Sun (1997), Marron and Chaudhuri (1998), Scott (1998a), and Walthers (1999). The mode tree approach captures the essence of a successful density-based clustering strategy but is still limited in the testing arena beyond one dimension.

The purpose of this paper is to extend recent results on *single-component* mixture fitting by a minimum-distance criterion to a *stochastic mode tree* that allows probing of high-

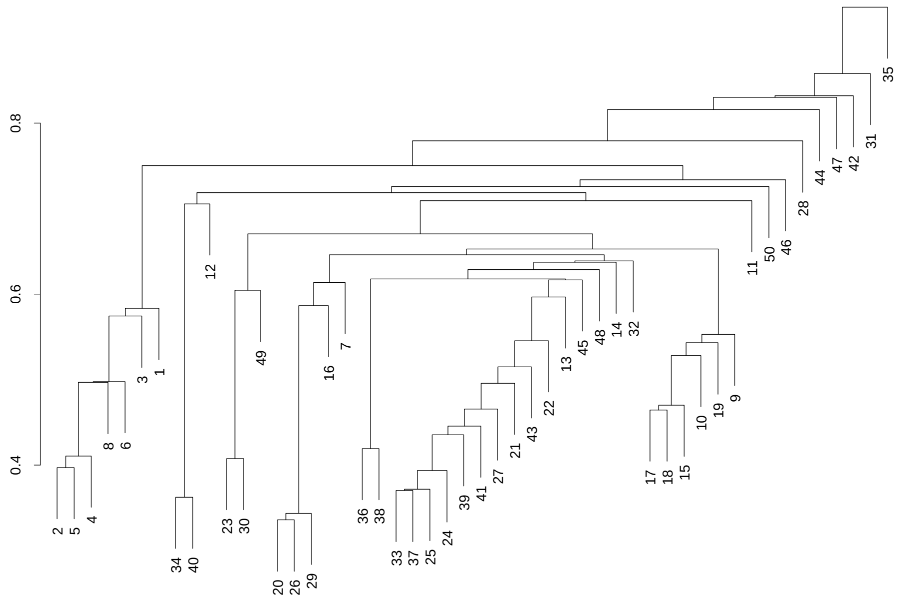

Figure 3: Mode tree of the sphered/blurred male blue *Leptograpsus* crab data.

dimensional data at multiple resolutions, without actually constructing a complicated locally adaptive estimator (Scott 1999). Given the present limitation that a fully parameterized locally adaptive estimator is nearly impossible to calibrate with today's technology for $d > 3$, the approach outlined here is of fundamental interest.

2. The Mode Tree

The mode tree is a natural attempt to link probability modeling with hierarchical clustering. The construction of the mode tree from the bottom up is easy to describe. When $h \approx 0$, the kernel estimate will have a mode at each unique data point. As the bandwidth h is sequentially increased ($1 - 2\%$ at each step), the change in the location of each mode is estimated. The resultant list of modes is examined for ties (that is, two modes have collapsed) and that information is recorded in the binary tree. The process continues until the final pair of modes collapses into a single mode.

In practice, care must be exercised. If too large a change in the bandwidth h is used, then several modes may appear to collapse at one step, and the exact linkages may be recorded in error. This may be avoided by decreasing the change in h until two and only two modes collapse. But the surprisingly subtle aspect in this step can be the difficulty of a successful search for the new mode location as the bandwidth is increased. Because one of the modes that merges first becomes a saddle point, and then "jumps" to the appropriate adjacent mode, Newton's method is likely to take too great an initial step and converge to the "wrong" mode.

Thus the search should maintain a local hill-climbing behavior. Yet, strict local hill-climbing algorithms are well-known to be very slow to converge. In independent work, Wong and Posner (1993) proposed a functional iteration algorithm to update modal locations. While this algorithm appears parameter free, a careful examination of its performance with data in two dimensions shows that the algorithm does not always converge to the correct adjacent mode (as determined by not wishing to cross gradient contours). Typically such an error does not have dire consequences, as these modes would have eventually joined to form a cluster and are simply being combined in the wrong order further down the tree. In practice, we may hope that such subtle errors do not occur very often.

For many iterations, the small change in bandwidth hardly changes the location of any modes. Much of the computational effort, however, is spent dealing with such (essentially stationary) modes. To determine which modes are not changing, one could calculate the Hessian at each mode. (All of the Hessian's eigenvalues must be negative, of course, at a local maximum.) Then one could simply observe which modes have any eigenvalues that are approaching zero (indicating that the mode is about to go "saddle" and jump to some nearby mode). Computation could then be focused only on those nearly "critical" modes. However, the amount of additional computation might be justified in difficult cases.

For large datasets, it would be more convenient to construct the mode tree from the top down, since detail near the bottom of the mode tree is not important. However, as the bandwidth is decreased, finding the location of the newly emerging mode can be difficult. Typically, when a mode "splits," one of the new modes remains at the current location. The other mode "jumps" somewhere in the vicinity, but finding it among the dozens of other modes nearby may or may not succeed, especially when $d >> 1$. Furthermore, if the increment on the bandwidth h is too great, then several new modes may have appeared, and there is no way to know or prove that all new modes have been found. Thus, while the bottom-up approach may be slow, its behavior is much more understandable and controllable.

Once a "correct" mode tree is computed, the decision about which modes are "real" and which are just noise begins. Minnotte (1997) proposed a local bootstrap procedure to test the mode-pairing at each branch of the mode tree. Minnotte's idea is related to Silverman's (1981) bootstrap test, but while Silverman counted the total number of modes in each bootstrap sample, Minnotte computes and compares the size of the "excess mass" in the region of the two modes that are about to be combined. This idea works well in one dimension, but its extension to two or more dimensions will be challenging and computationally expensive. A related approach that does not require bootstrapping was given by Müller and Sawitzki (1991). Marron and Chaudhuri (1998) highlight regions in the mode tree where the derivative of the kernel estimate is significantly positive or negative to identify likely modal regions.

We believe the mode tree is a useful framework in which to attack the clustering problem. However, the mode tree suffers from the fact that all possible modes (and hence clusters) are displayed without a figure of merit. Marchette and Wegman (1997) proposed a clever enhancement of the mode tree by replacing the kernel estimator with a locally adaptive MON they call a filtered kernel estimator. Their algorithm usually, but not always, elevates real modes closer to the top of the mode tree.

In the next section, we propose a new variety of mode tree which is not based on all possible modes of a kernel estimate, but only those that satisfy a local optimization problem.

3. Partial Mixture Fitting and the PDC

Using probability density functions which are not densities is common practice. For example, Bayesians employ improper priors, which integrate to infinity. Negative probabilities are observed in orthogonal series and higher-order kernel methods. Here, the idea is to fit a nonnegative "density" which integrates to a number (usually) less than 1. Specifically, consider the partial probability density model,

$$\hat{f}(x) = w \cdot \phi(\mu, \sigma^2)\,,$$

where w is considered a free parameter on the interval $(0, 1]$. This model comprises one component of the full mixture model in Equation (2) and will be denoted by PDC for *partial density component.*

Consider fitting a partial probability model to the mixture density shown in Figure 4a where $K = 3$ and

$$(\mu_k) = (-1, 0, 3)^T \qquad (\sigma_k) = \left(\frac{1}{9}, 1, \frac{1}{3}\right)^T \qquad (w_k) = \left(\frac{3}{51}, \frac{32}{51}, \frac{16}{51}\right)^T . \tag{3}$$

First, observe that no meaningful partial probability model can be fit by maximum likelihood, because the weight w will always take the value 1 when the likelihood is maximized. This follows since the logarithm of the density separates the weight w from the other parameters. Such is not the case, however, if a minimum distance criterion is employed. In this paper, integrated squared error (ISE) or the squared L_2-norm is used:

$$\mathrm{ISE} = \int_{-\infty}^{\infty} \left[\hat{f}(x) - f(x)\right]^2 dx\,. \tag{4}$$

Integrated squared error is a commonly used goodness-of-fit measure in nonparametric density estimation.

When f is a normal mixture, a simple closed-form expression exists for the theoretical ISE, which is of interest as the solution to a consistent data-based approach as $n \to \infty$. First expand Equation (4) to

$$\mathrm{ISE}(\mu, \sigma, w) = \int \left[\hat{f}(x)^2 - 2\hat{f}(x)\, f(x) + f(x)^2\right] dx\,. \tag{5}$$

Recall the identity

$$\int_{-\infty}^{\infty} \phi(x|\mu_1, \sigma_1^2)\, \phi(x|\mu_2, \sigma_2^2)\, dx = \phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)\,.$$

Then for a normal PDC, the first integral in Equation (5) equals $w^2\phi(0|0, 2\sigma^2)$ or $w^2/(2\,\sigma\sqrt{\pi})$. The third integral equals

$$\int_{-\infty}^{\infty} f(x)^2 dx \equiv R(f) = \sum_{k=1}^{K}\sum_{\ell=1}^{K} w_k\, w_\ell\, \phi(0|\mu_k - \mu_\ell, \sigma_k^2 + \sigma_\ell^2)\,.$$

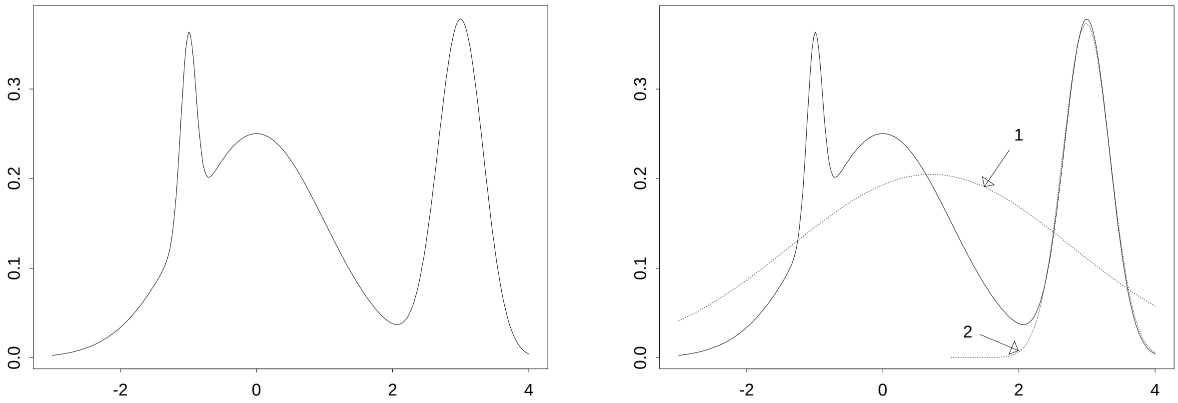

Figure 4: (a) Mixture of 3 normals. (b) Two partial normal fits according to integrated squared error (dashed lines). Parameter values are given in the text.

A similar result exists for the second integral. Combining gives

$$\mathrm{ISE}(\mu, \sigma, w) = \frac{w^2}{2\,\sigma\sqrt{\pi}} - 2w \sum_{k=1}^{K} w_k\, \phi(0|\mu - \mu_k, \sigma^2 + \sigma_k^2) + R(f)\,. \tag{6}$$

Since the last term, $R(f)$, does not depend on any of the parameters of the PDC, $R(f)$ may be ignored when finding the best set of parameters.

With the parameters in Equation (3), minimization of Equation (6) over (μ, σ, w) leads to one of two partial densities: $(.707, 2.066, 1.061)$ or $(2.99, .348, .326)$; see Figure 4b. The second PDC matches the third component values of $(3, .333, .314)$. The first PDC is a "best" fit to the entire density, but the parameters differ from the theoretical moments of the mixture ($\mu = .882$ and $\sigma = 1.663$). Also note that for the unconstrained best L_2 fit, w^* turns out to be greater than 1. There is no local PDC solution that fits either of the left two bumps alone. As Scott (1999) has noted, PDC's are not attracted to mixture components with mass below a certain threshold, unless a component is sufficiently isolated. Thus fitted PDC's may or may not correspond to individual components in the underlying mixture density.

Can the PDC be modified to find all three modes in the mixture? An affirmative answer lies in recognizing that the minimization of ISE does not preclude optimizing over a *subset* of the three PDC parameters, (μ, σ, w). Since greatest interest lies with a cluster's location, the parameter μ is always optimized. By fixing either σ or w or both, can all three components be resolved? In Figure 5, partial PDC fits were computed by minimizing criterion (6), first with σ fixed and optimizing over (μ, w), and then with w fixed and optimizing over (μ, σ), for the same 1,000 random initial values. (The sampling ranges for the initial values of (μ_j, σ_j, w_j), $j = 1, \ldots, 1000$, were $(-4, 3)$, $(.05, 2)$, and $(.05, 1)$, respectively. The sampling distributions were not uniform for σ and w, but biased towards smaller values.)

In Figure 5a, which displays the 1000 points, (μ_j^*, σ_j), all three components are clearly indicated when $\sigma < 0.2$. When $0.4 < \sigma < 1$, the left two components are not separated. Only one mean location is found when $\sigma > 1.05$. Observe that the estimates of the mean

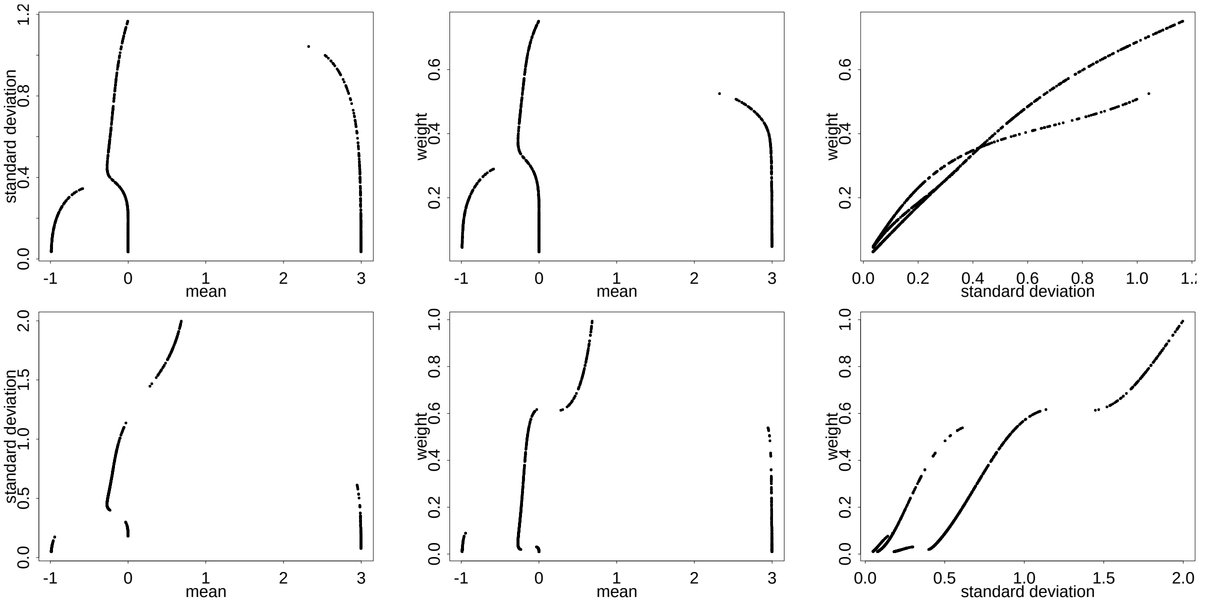

Figure 5: PDC parameter estimates for 1,000 random starting points. In the top three frames, the minimization is with respect to μ and w only, with σ fixed. In the bottom three frames, the minimization is with respect to μ and σ only, with w fixed.

locations vary smoothly with σ, except at several critical values where the "branch" of the curve ends. The same structure is apparent with the corresponding (μ_j^*, w_j^*) points shown in Figure 5b, which is easily explained by noting the high correlation between (σ_j, w_j^*) in Figure 5c for each "branch."

A similar but less compelling story is apparent in Figures 5(d–f), where the optimization is over μ and σ with w fixed at its initial random value. In frame 5d, only one mean location is found when $\sigma_j^* > 0.61$. The expected branch near 0 appears when $\sigma_j^* < .30$, but disappears when $\sigma_j^* < .18$. The final mode at $x = -1$ appears when $\sigma_j^* < 0.17$. The pair (σ_j^*, w_j) are still highly correlated in 5f, but the view of (μ_j^*, w_j) in 5e is much less clear as the mode at 0 barely appears at all, and the "combined" mode at $-.26$ persists after the modes at 0 and -1 have appeared. Empirically, we prefer $(\mu_j^*, \sigma_j, w_j^*)$ to $(\mu_j^*, \sigma_j^*, w_j)$.

Since greatest interest lies with the location parameter, μ, the simulation was repeated with initial values of the three parameters chosen at random, but optimization limited to μ alone. In Figure 6a, the tree-like structure of (μ_j^*, σ_j) is apparent as before. The 1000 random standard deviation and weight inputs are shown in Figure 6c. A 3-D view of (μ_j^*, σ_j, w_j) confirms that Figures 6a and 5a are the same, modulo different random starting values. Apparently, the initial value of w does not affect the value of μ^*. A closer look at Equation (6) reveals that when σ is fixed, w and μ are uncoupled in the second term. Hence, the best value of μ is independent of w. In particular, any fixed or random value may be used for w during the optimization. Given μ^*, Equation (6) takes the simple form $c_1 w^2 - 2c_2 w$, and w^* is easily seen to equal c_2/c_1, yielding Figure 5b again. Numerical optimization over μ alone (i.e., without w) significantly reduces the computational effort.

While each may be useful in practice, after reviewing many figures such as 5 and 6, we

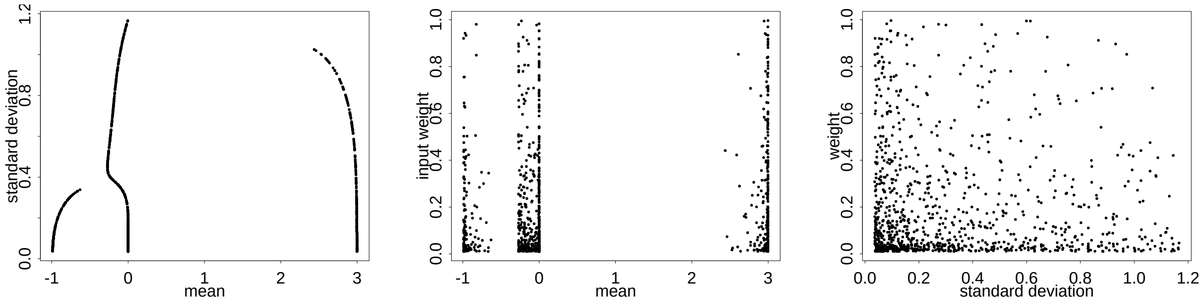

Figure 6: PDC parameter estimates for 1,000 random starting points. The minimization is with respect to μ only. (a) (μ^*, σ). (b) (μ^*, w). (c) (σ, w).

find that optimizing over μ and w with σ fixed is preferable. In particular, we find the plot of (μ^*, w^*) is more interpretable than (μ^*, σ). In practice, these diagrams must be estimated from data, and they will be noisier. A data-based version of PDC is the next topic.

4. Stochastic Mode Tree

The ISE criterion for the PDC may be estimated with data $\{x_1, \ldots, x_n\}$ from a *general* density $f(x)$ by noting that the second integral in Equation (5) is simply minus two times the expectation $\mathrm{E}\hat{f}(X)$ at a random point X. A simple unbiased estimator may be obtained by averaging $\hat{f}(\cdot)$ over the sample data. As before, the third integral in Equation (5) may be omitted as it is constant with respect to the parameters. Thus we obtain the following data-based criterion for numerical optimization:

$$\mathrm{L_2E}(\mu, \sigma, w) = \frac{w^2}{2\,\sigma\sqrt{\pi}} - \frac{2\,w}{n}\sum_{i=1}^{n} \phi(x_i|\mu, \sigma^2)\,. \tag{7}$$

The notation $\mathrm{L_2E}$ for parametric estimation was introduced by Scott (1998b) to denote the L_2 estimation error criterion. Note that when σ is fixed in Equation (7), then w and μ are uncoupled just as in the theoretical version in Equation (6). Thus μ^* may be computed with w fixed at any value, and then w^* computed as c_2/c_1, using the sample version of c_2.

A sample of size $n = 1275$ was generated from the mixture density described in Equation (3). Various combinations of the parameters were estimated using random starts. In Figure 7, the stochastic mode tree (SMT) for this simulated dataset is displayed. While the 3 main modes are clearly evident, so are a number of small features not in the true density but only in the sample.

Figure 8 displays both the mode tree based on the kernel estimator as well as the stochastic mode tree. Note how the stochastic mode tree suppresses a number of "potential" modes in these data. Specifically, the SMT acts as a partial inferential tool (filter) for gauging the weight of each mode or cluster of a particular size. For example, the leftmost data point at $x = -3.28$ exists as a potential cluster in both figures, but its significance is obviously nil in the SMT version. Conversely, the cluster at $x = -1$ is more prominent in the SMT.

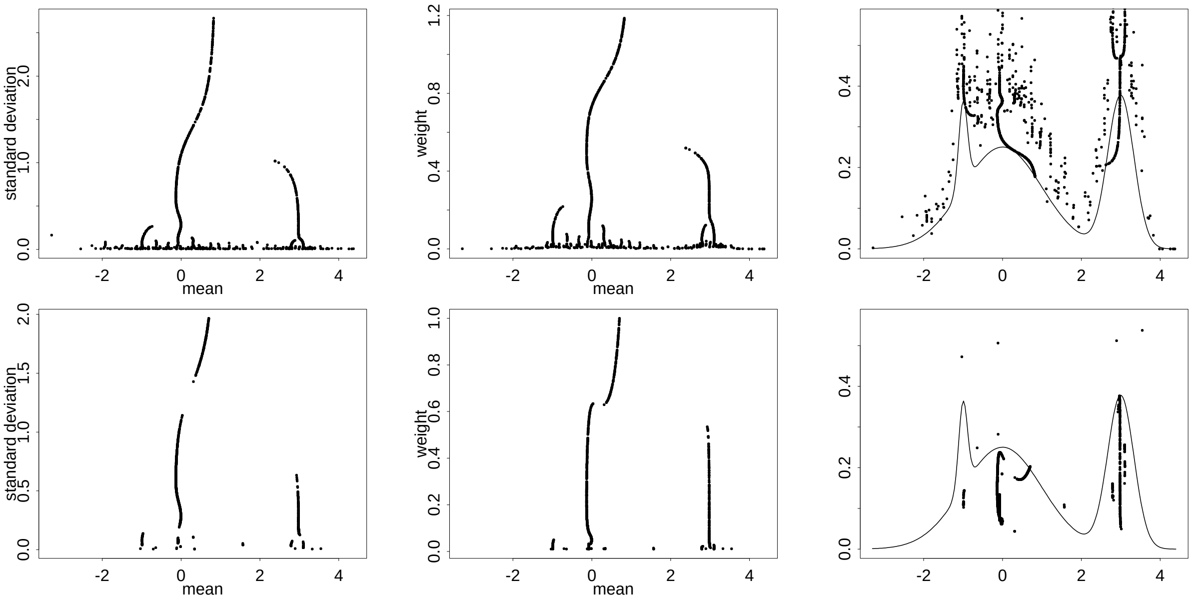

Figure 7: (a–c) 1275 partial density fits $(\mu_j^*, \sigma_j, w_j^*)$ to simulated data, with optimization over μ and w but not σ: (a) (μ_j^*, σ_j) (b) (μ_j^*, w_j^*) (c) The estimates are plotted on the density scale at $\left(\mu_j^*, w_j^* \phi(\mu_j^* | \mu_j^*, \sigma_j^2)\right)$. (d–f) 1275 partial density fits $(\mu_j^*, \sigma_j^*, w_j)$ with the same random starts: (d) (μ_j^*, σ_j^*) (e) (μ_j^*, w_j) (f) $\left(\mu_j^*, w_j \phi(\mu_j^* | \mu_j^*, \sigma_j^{*2})\right)$.

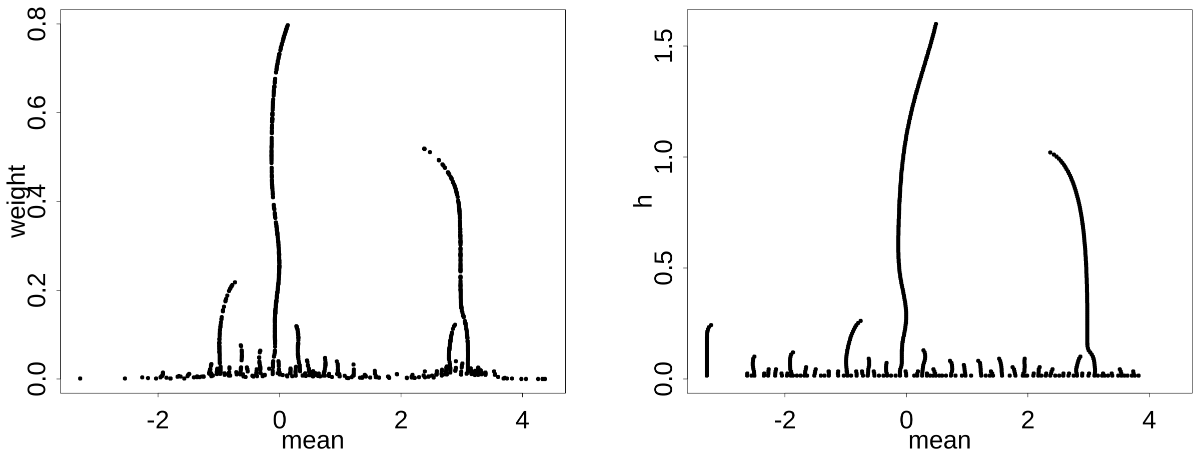

Figure 8: (a) Stochastic mode tree of simulated data. (b) Original mode tree, but drawn without horizontal connections and on a linear scale to facilitate comparison.

5. Special Cases

In this section, we briefly examine the consequences of using a partial normal component to fit non-Gaussian mixtures. In the multivariate case, the use of $\Sigma = h \cdot I_d$ is examined when the local Hessian is not spherical.

A sample of size 1000 was generated from the beta mixture $\frac{2}{5}\mathrm{B}(4,8) + \frac{3}{5}\mathrm{B}(10,3)$; see Figure 9. Fitting partial normal components (not the true density) gives good location estimates for the smaller values of σ_j and w_j, i.e., for the more local PDC fits.

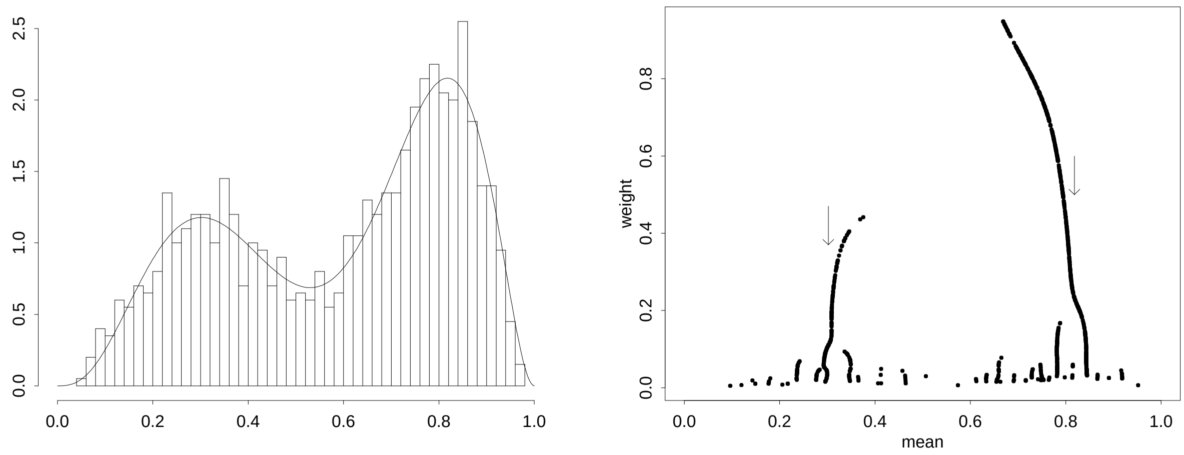

Figure 9: (a) Beta mixture and sample histogram with $n = 1000$. (b) Stochastic mode tree of (μ_j^*, w_j^*) from Normal partial density fits, with arrows indicating the true modal locations.

In the multivariate case, consider a partial density component model that is Gaussian with covariance matrix proportional to the identity matrix. Here it is demonstrated that this model can find the correct location of Gaussian components for which the covariance matrix is not proportional to the identity. For an equal-weight three-component bivariate normal mixture, a sample of size 1200 was generated and is shown in Figure 10a. (The mean vectors from left to right are $\binom{-3}{3}$, $\binom{0}{0}$, and $\binom{5.0}{-1.5}$. The correlation for the first is .75 and 0 for the others. The variances are all 1 except for $\sigma_y^2 = 4$ in the last.) The estimated partial normal components shown in Figure 10(b,c) give good location estimates. Note that the optimal weight estimates are less interpretable when the parametric form is not correct. In particular, fitting $w \cdot \mathrm{MN}(\mu, \sigma^2 \cdot I_2)$ to data with very large σ often results in $w^* > 1$.

6. Examples

Ripley (1996) uses five measurements on crabs to illustrate various discrimination algorithms, comparing blue and orange crabs as well as male and female sexes. One of the techniques employed is linear discriminant analysis, which is optimal if the data are normal, $MN(\mu_k, \Sigma)$. Here, attention is focused on the group of male blue crabs ($n = 50$). The stochastic mode tree was computed on the blurred and sphered data; see Figure 11. The figure suggests there may be 2–4 clusters in these data. The inference of more than one cluster is supported by examining an averaged shifted histogram (ASH,Scott (1992)) of the first two principal components (explaining 93.8% of the variance) of the raw data. In general,

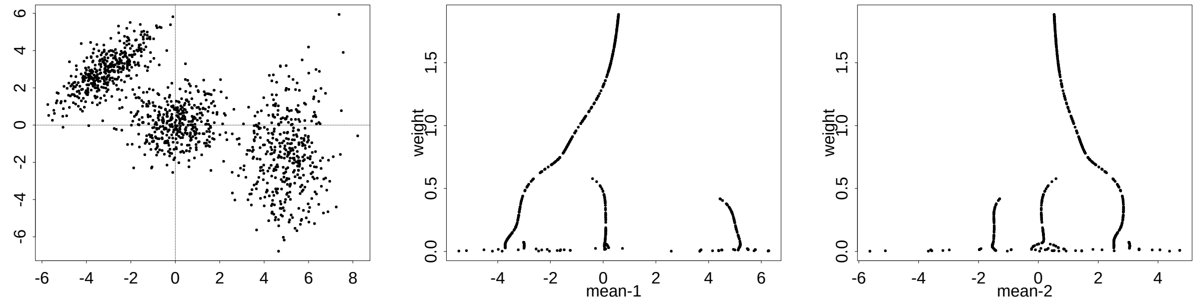

Figure 10: (a) Sample of size 1200 from a bivariate mixture of three Normals. (b) Stochastic mode tree of the bivariate Normal partial density fits. A view of the points (μ_1^*, w^*) is shown. (c) A view of the points (μ_2^*, w^*) is shown.

the number of clusters will not decrease as other variables are added. (An explanation for the observed "gap" in Figure 11a for $w^* \in (0, .113)$ may be found in Section 7.)

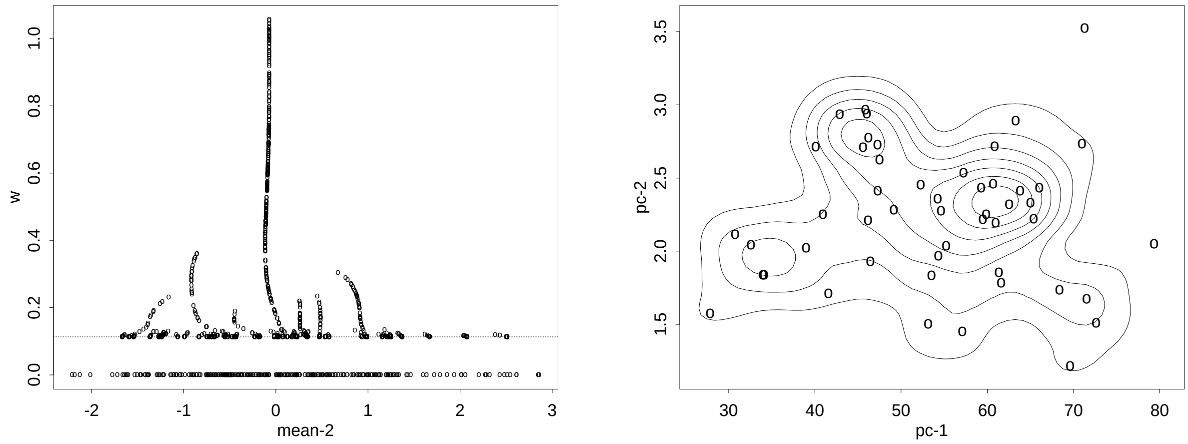

Figure 11: (a) Projection of the stochastic mode tree of 5-D male blue crab data onto the petal width variable. The horizontal dotted line is at $w^* = 0.113$. (b) ASH estimate of the first two principal components of the raw male blue crab data.

Figure 12 displays a portion of the stochastic mode tree along the petal length variable for the well-known 4-D Fisher iris data, which were also blurred. Also shown is a bivariate ASH of that variable and another. The three species are indicated, but other modes are apparent. We return to this and other examples in the next section.

7. Inferential Possibilities

When σ is so small that a data point x_i is "distant" from its nearest neighbor, then one PDC solution is $\mu^* = x_i$; hence, from equation (7),

$$\mathrm{L_2E}(\mu^*, \sigma, w) = \frac{w^2}{2\,\sigma\sqrt{\pi}} - \frac{2\,w}{n\sigma\sqrt{2\pi}}\,.$$

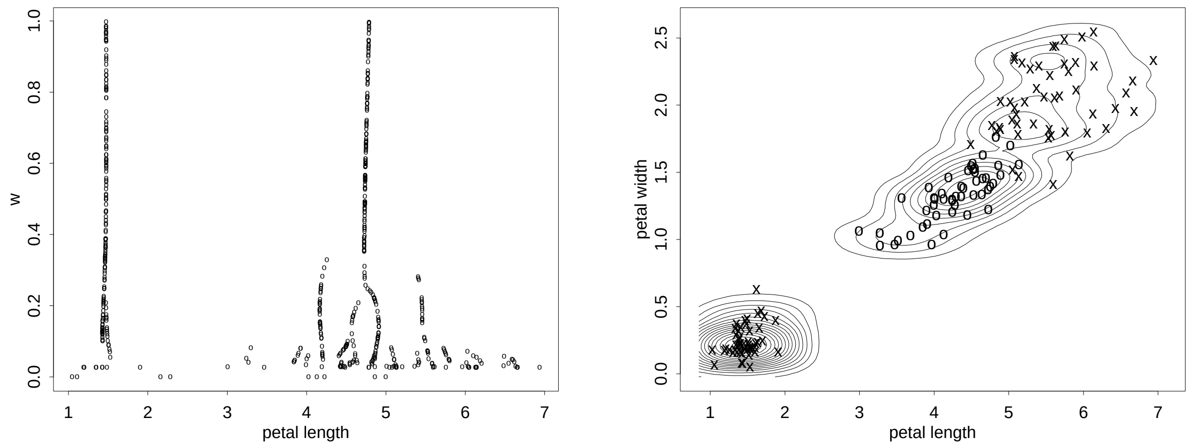

Figure 12: (a) Portion of the stochastic mode tree of iris data along the petal length. (b) ASH estimate of the petal length and width for the *Setosa*, *Versicolor*, and *Virginica* species.

It follows that $w^* = \sqrt{2}/n$, a constant irrespective of the particular (small) value of σ. Thus, if one wished to exclude isolated clumps of points with fewer than 3 or 4 members, a simple threshold rule can be developed for excluding any PDC solution with $w^* \leq 3\sqrt{2}/n$. For $\mathbf{x} \in \Re^d$, a similar argument shows that

$$w^* = 2^{d/2}/n \tag{8}$$

for isolated points. This formula is quite accurate in practice. For example, if $d = 5$ and $n = 50$, then $w^* = 0.113$; cf. the vertical gap in Figure 11a. Note that $w^* = 0$ is also a stationary point for this problem. The gap for the Iris data is given by $w^* = 2^{4/2}/150 = 0.027$, which is (barely) visible in Figure 12a.

A more sophisticated line of reasoning explaining "noise" in the SMT may be developed. Suppose a PDC solution, (μ^*, σ), exists and σ is small. Consider the original $\mathrm{L_2E}$ criterion:

$$\int \hat{f}(x)^2 dx - 2\int \hat{f}(x)\, f(x)\, dx\,. \tag{9}$$

Taking a Taylor's series about the point $x = \mu^*$ for a Normal PDC, Equation (9) becomes

$$\frac{w^2}{2\sqrt{\pi}\sigma} - 2w \int_x \phi(x|\mu^*, \sigma^2) \left[f(\mu^*) + (x - \mu^*) f'(\mu^*) + \frac{1}{2}(x - \mu^*)^2 f''(\mu^*) + \cdots \right] dx$$

or, since σ is small and μ^* and σ^2 are the moments of the normal PDC,

$$\frac{w^2}{2\sqrt{\pi}\sigma} - 2w \left[f(\mu^*) + \frac{1}{2}\sigma^2 f''(\mu^*) + O(\sigma^4) \right] .$$

Retaining only the leading terms, the optimal value of the weight approximately satisfies

$$w^* = 2\sigma\sqrt{\pi} f(\mu^*) + \sigma^3 \sqrt{\pi} f''(\mu^*)\,.$$

Again, this argument may be extended to the multivariate case. For $\mathbf{x} \in \Re^d$, let $\hat{f}$ be a normal PDC with $\Sigma = \sigma^2 I_d$. Then a similar argument leads to

$$w^* = \left(2\,\sigma\sqrt{\pi}\right)^d f(\mu^*) + O(\sigma^{d+2})\,.$$

Intuitively, a "false" mode in the SMT gives rise to points (μ^*, w^*) which approximately satisfy the leading term $w^* = 2\sigma\sqrt{\pi} f(\mu^*)$. For true modes, on the other hand, the linear approximation will not be sufficient. To examine this hypothesis, we re-examined the SMT results in Figure 8. Using the true density, we computed $f(\mu_j^*)$, then a predicted optimal weight for each solution by $\hat{w}_j = 2\sigma\sqrt{\pi} f(\mu_j^*)$. Using XGobi (Swayne, Cook, and Buja 1998), we plotted $(w_j^*, \hat{w}_j)$, and brushed all the points falling on the 45-degree line near the origin. In Figure 13, the highlighted points as well as the correspondingly highlighted stochastic mode tree are displayed. Clearly, this approach successfully eliminates most of the background clutter while retaining all and only the 3 components.

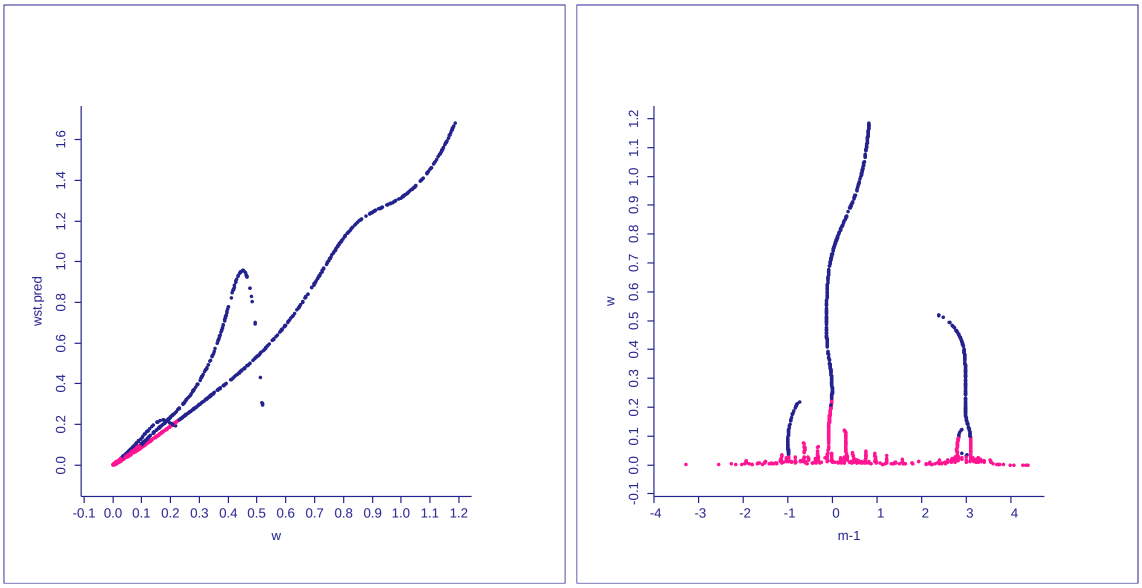

Figure 13: (a) The highlighted (red) points satisfy the local Taylor's Series prediction for w^* given σ, using the true density for $f(\mu^*)$. (b) The corresponding stochastic mode tree for the 1275 points with color linked to the left frame.

In practice, an estimate of the density is required to find $\hat{w}$. Either a kernel estimator, a mixture estimator, or a nearest-neighbor estimator may be used for this task. While we eventually hope to use a mixture estimate for this task, we chose a k-th nearest-neighbor estimator with fairly small k. While such an estimate is noisier than a kernel estimate, it is better able to cope with different sized bumps in high dimensions with only a little tweaking.

For the (blurred) iris data, we chose $k = 8$ and used XGobi to brush points in the $(w_j^*, \hat{w}_j)$ plot near the origin on the 45-degree line; see Figure 14a. When the corresponding points in the stochastic mode tree are plotted in Figure 14b, the three species are clearly highlighted, together, perhaps, with a hybrid species of *Versicolor* and *Virginica* (Thompson (2000), pp. 250–256). (Compare to Figure 12b.)

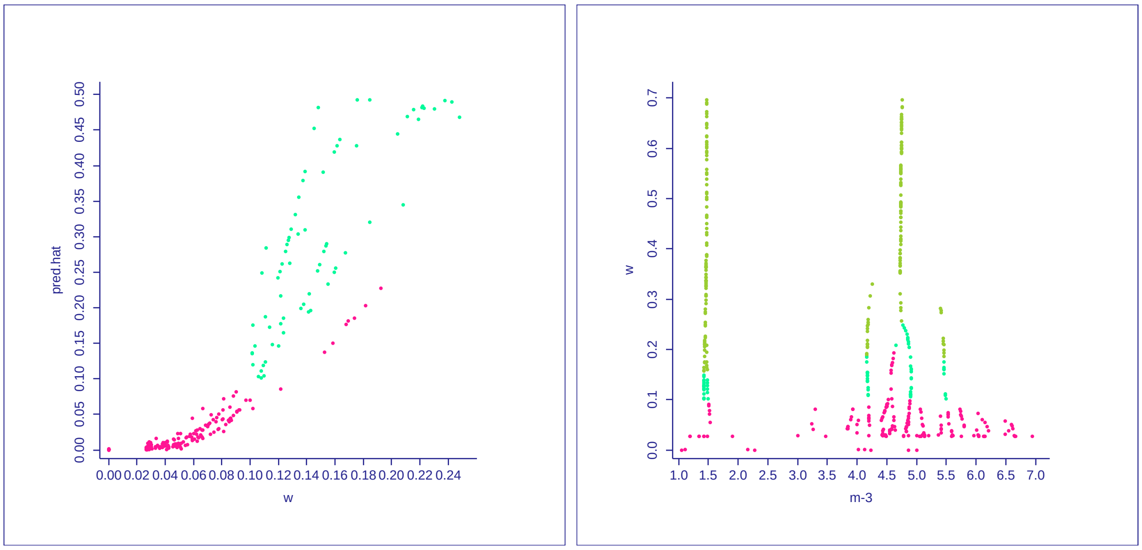

Figure 14: (a) For the blurred 4-D Iris data, highlighted (red) points approximately satisfy the local Taylor's Series prediction for w^* given σ, using an 8-NN density estimate for $f(\mu^*)$. (b) Corresponding brushed stochastic mode tree projected onto petal length.

For the simulated 3-component bivariate example, we computed $\hat{f}$ using an 8-th nearest-neighbor estimate, rather than the true density. The brushed $(w_j^*, \hat{w}_j)$ plot and the highlighted stochastic mode tree are displayed in Figure 15.

Finally, we re-examined the blurred male blue crab data. We computed $\hat{f}$ using an 4-th nearest-neighbor estimate. The brushed $(w_j^*, \hat{w}_j)$ plot and the highlighted stochastic mode tree are displayed in Figure 16. Note that all the points at $w^* = 0$ fall at the origin in the left frame. Three groups seem to be supported by this result.

Results from other well-known univariate datasets are not shown here, but are available from our web site.

8. Conclusions

Modes are an excellent summary of data. A kernel estimator can have many modes, but as the smoothing parameters change, so do the number and location of the modes. Determining which of these modes are "real" is not an easy task. The ordinary mode tree displays all possible modal locations without criticism. In one or two dimensions, some tests and procedures are available. In more dimensions, mixture models are of particular interest, but begin to experience practical problems as numerical fitting produces more singular components.

The stochastic mode tree "probes" the (multivariate) data density surface for normal patches and modes that are located and scaled to provide best local fits for a given (random) scale. Alternatively, choosing a polynomial model (e.g. centered Beta) for the partial density component may make sense from a Taylor's series' point of view. Plotting w^* against μ^* reveals actual modal structure more clearly. Questions of whether an individual mode exists may be tested by fitting mixtures with components located at the indicated positions (i.e.

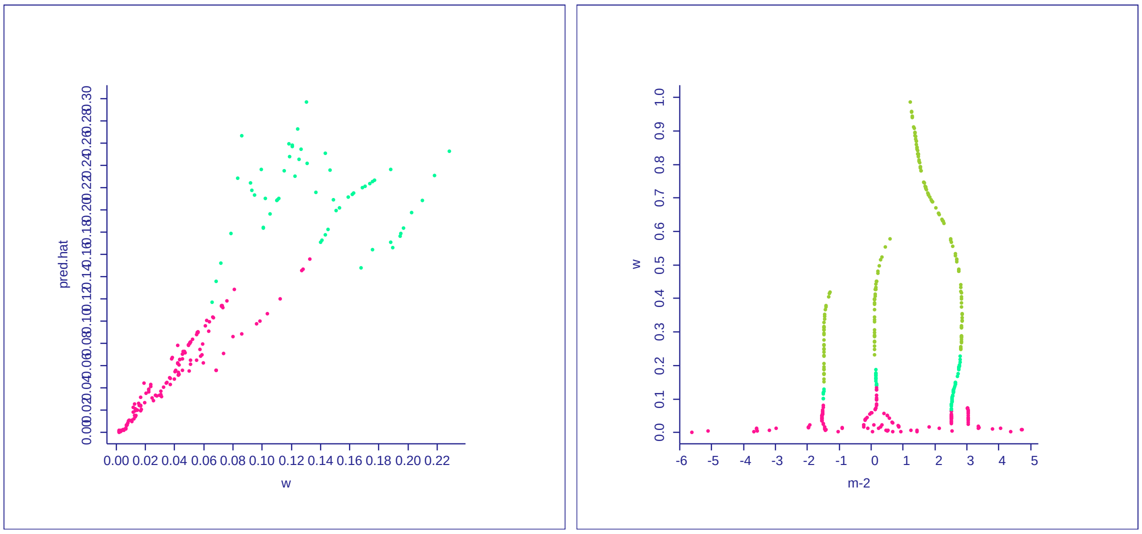

Figure 15: (a) For the 3-component bivariate simulated dataset, highlighted (red) points approximately satisfy the local Taylor's Series prediction for w^* given σ, using 8-NN density estimate for $f(\mu^*)$. (b) The corresponding stochastic mode tree against μ_2.

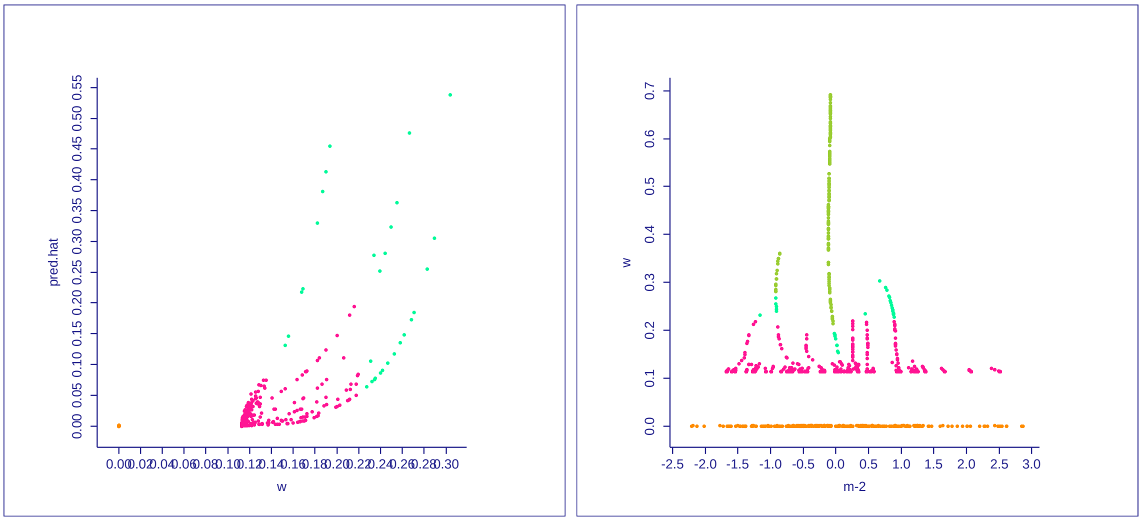

Figure 16: (a) Highlighted (red) points satisfy the local Taylor's Series prediction for w^* given σ, using 4-NN density estimate for $f(\mu^*)$. (b) The corresponding stochastic mode tree against μ_2.

don't optimize over the locations); see Scott and Szewczyk (1999). Another possible method is to examine mode trees from bootstrap samples; see Minnotte, Marchette, and Wegman (1998) for details. The interactive techniques described in Section 7 give very promising results as well.

The algorithm as stated is rotationally invariant, not because the data are, but because the covariance model of the PDC is spherical. A local estimate of the shape of the density there may be obtained by computing the negative of the Hessian matrix at each sample mode. An improved PDC might use that matrix for the covariance, and iterate several times. Finally, the estimates could be used to initialize a MON fit that does not optimize over the mean locations.

The techniques described here are exploratory, as is cluster analysis generally, so the approach in Section 7 is quite reasonable. We also have attempted fitting 2-component PDC's, with random starts. This line of inquiry is promising but much more work is required to understand how to extract the meaningful interpretations from the collection of solutions.

9. Acknowledgments

This research was supported in part by NSF Grant 99-71797 and was performed while the first author was on sabbatical at the NSA.

References

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*, 803–822.

Boswell, S. B. (1983). Nonparametric mode estimation for higher dimensional densities. Technical Report Ph.D. thesis, Dept. of Math Sciences, Rice University.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood estimation from incomplete data via the em algorithm. *J. R. Statist. Soc. B 39*, 1–38.

Good, I. J. and R. A. Gaskins (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by the scattering and meteorite data (with discussion). *J. Amer. Statist. Assoc. 75*, 42–73.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons, Inc.

Hartigan, J. A. and P. M. Hartigan (1985). The dip test of unimodality. *Ann. Statist. 13*, 70–84.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *J. Amer. Statist. Assoc. 90*, 773–795.

Marchette, D. and E. J. Wegman (1997). The filtered mode tree. *J. Comp. Graph Stat 6*, 143–159.

Marron, J. S. and P. Chaudhuri (1998). When is a feature really there: The sizer approach. *Proceedings of SPIE 3371*, 306–312.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *Ann. Statist. 25*, 1646–1660.

Minnotte, M. C., D. Marchette, and E. J. Wegman (1998). The bumpy road to the mode forest. *J. Comp. Graph Stat 7*, 239–251.

Minnotte, M. C. and D. W. Scott (1993). The mode tree: A tool for visualization of nonparametric density features. *J. Comp. Graph. Stat. 2*, 51–68.

Müller, D. W. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc. 86*, 738–746.

Reynolds, D. A. (1995). Automatic speaker recognition using gaussian mixture speaker models. *Lincoln Laboratory Journal 8*, 173–192.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Sager, T. W. and R. A. Thisted (1982). Maximum likelihood estimation of isotonic modal regression. *Ann. Statist. 10*, 690–707.

Scott, D. W. (1992). *Multivariate Density Estimation*. New York: John Wiley & Sons, Inc.

Scott, D. W. (1998a). On fitting and adapting of density functions. *Computing Sciences 30*, 116–123.

Scott, D. W. (1998b). Parametric modeling by minimum L_2 error. Technical Report 98-3, Department of Statistics, Rice University.

Scott, D. W. (1999). Remarks on fitting and interpreting mixture models. *Computing Sciences 30*, 104–109.

Scott, D. W. and W. F. Szewczyk (1997). Bumps along the road towards multivariate mode trees. Presented at NSF Workshop: Bumps, Jumps, Clustering and Discrimination, Houston Texas.

Scott, D. W. and W. F. Szewczyk (1999). Simplifying mixture models with applications. *Computing Sciences 30*, 118–122.

Scott, D. W. and W. F. Szewczyk (2000). From kernels to mixtures. Technical Report submitted, Rice University.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. B 43*, 97–99.

Sun, J. (1997). Bump hunting problems: A new test. Presented at NSF Workshop: Bumps, Jumps, Clustering and Discrimination, Houston Texas.

Swayne, D. F., D. Cook, and A. Buja (1998). XGobi: Interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics 7*, 113–130.

Terrell, G. R. and D. W. Scott (1992). Variable kernel density estimation. *Ann. Statist. 20*, 1236–1265.

Thompson, J. R. (2000). *Simulation: A Modeler's Approach*. New York: John Wiley & Sons.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons, Inc.

Walthers, G. (1999). Multiscale analysis of a semiparametric model via penalized maximum likelihood. Technical report, Department of Statistics, Stanford University.

Wong, Y.-F. and E. C. Posner (1993). A new clustering algorithm applicable to multispectral and polarimetric SAR images. *IEEE Trans Geoscience Remote Sensing 31*, 634–644.

Keywords

Kernel density; Minimum distance estimation; Mixture models; Modes; Number of clusters