Visualization of scales of multivariate density estimates

Jussi Klemelä
Department of Statistics, Economics Faculty
University of Mannheim
L 7 3-5 Verfügungsgebäude
68131 Mannheim, Germany
Email: klemela@rumms.uni-mannheim.de
Fax +49 621 1811931

October 18, 2006

Abstract

We present graphical tools for visualizing scales of multivariate density estimates. The graphical tools visualize the shapes of the estimates and they may be applied in mode detection. We consider *multiframe mode graphs* which visualize the locations of the modes of multivariate density estimates corresponding to an interval of smoothing parameter values. We define a *branching map of level set trees* which shows how the level sets of the estimates in a scale of estimates are decomposing to separated regions as function of the level. In addition it visualizes the excess masses associated with the separated regions of the level sets. A *scale and shape visualization table* contains besides a multiframe mode graph and a branching map also 4 additional windows. With this table we may zoom in to the scale of estimates by choosing an estimate for a closer inspection. Second, we may zoom in to the estimate by choosing a level set of the estimate for a more detailed visualization. The visualization tools which we present may be applied with a number of different density estimators. We give examples of the application of the tools with kernel estimates and with multivariate adaptive histograms.

Key Words: Level set tree, Mode tree, Nonparametric density estimation, Scale space analysis.
Short title: Visualization of density estimates

1 Introduction

We are interested in the shape of a multivariate density function. In particular, we are interested to detect the modes of the density $f : \mathbf{R}^d \to \mathbf{R}$. We are given a sample $X_1, \ldots, X_n \in \mathbf{R}^d$ of identically distributed random vectors with density f. We will use nonparametric density estimators to shape detection.

In one and two dimensional cases an inspection of plots and perspective plots of nonparametric density estimates, for example kernel estimates, gives a powerful tool for shape detection. When the dimension is greater or equal to three, then one needs more sophisticated visualization tools to make nonparametric density estimation useful in shape detection.

In one and two dimensional cases the "art of smoothing" has typically consisted from the inspection of the change of the estimates as the smoothing parameter changes. Thus we need visualization tools which would make it possible conveniently to scan through a scale of smoothing parameters. Such tools may be used in the spirit of *scale space analysis*, discussed by Chaudhuri and Marron (2000).

A *mode tree* is a useful visualization tool to help scanning through estimates. For one and two dimensional cases the mode tree was introduced in Minnotte and Scott (1993) to visualize the number and the locations of the modes as the smoothing parameter is changed. Marchette and Wegman (1997), Minnotte, Marchette and Wegman (1998), Scott and Szewczyk (2000) develop further the one dimensional mode tree.

We define a *multiframe mode graph* to be used as a road map directing our scanning through a scale of estimates. A multiframe mode graph makes a mode graph separately for each coordinate and uses colors to identify the modes across different windows. We want to apply multiframe mode graphs not only for kernel estimators but for any estimator whose smoothness is controlled with a real valued smoothing parameter. We use the term "graph" and not the term "tree" because we consider such scales of estimates where the number of modes is not monotonic as the function of the smoothing parameter, unlike in the case of univariate kernel estimates with the standard Gaussian kernel.

A mode graph does not visualize the relative importance of the modes. For this purpose we use *a branching map*. A branching map is a perspective plot of a 2D function whose surface is colored. With the colors and with the values of the function we visualize both the levels where the level sets of the estimates are decomposing to separate regions and also the *excess masses* associated with those separate regions, simultaneously for a scale of estimates. The definition of the branching map is based on the concept of a

level set tree of a function. Level set trees were introduced in Klemelä (2004). A level set tree is a tree of the separated components of the level sets of a function.

The excess mass associated with a separated region of a level set is the volume of the area which the density function delineates over the given level, in the given separated region. Usually excess masses have been associated with the modes of a density function and not with the separated regions of level sets. For example Hartigan (1987), Müller and Sawitzki (1991) and Minnotte (1997) have applied excess masses in cluster analysis and mode testing. Minnotte and Scott (1993) proposed to visualize the probability mass of the mode through the widths of the mode traces. We argue that it is fruitful to change the viewpoint from the mode testing to the testing of the branching of the level set tree. Thus we visualize the excess masses of such nodes of the level set tree of the estimate which are a result of the branching of the level set tree. The branching map may be seen as a first step for developing a mode testing approach based on level set trees. Chaudhuri and Marron (1999) and Godtliebsen, Marron and Chaudhuri (2002) present SiZer for inference and visualization based on scales of one and two dimensional kernel estimates. We may view a branching map as a multivariate excess mass based version of SiZer, but a branching map does not contain a formal mode testing component.

Mode graphs and branching maps give an overview of a scale of estimates. We need tools which would enable us to conveniently choose estimates from the scale and visualize those estimates. A *scale and shape visualization table* is a dynamic tool to fulfill this purpose. It consists of 6 windows. The first 2 windows show one frame of a multiframe mode graph and a branching map. We may choose an estimate from the scale and show in the second 2 windows a volume plot of the estimate, and one frame of a barycenter plot of the estimate. A volume plot shows a plot of a 1D function which is mode isomorphic to the original function. A barycenter plot shows the barycenters of the level sets of the function. Furthermore, one may choose a level set of the estimate, and visualize in the last 2 windows the shape of the level set with a radius plot and the location and orientation of the level set with a location plot. A volume plot and barycenter plot were defined in Klemelä (2004) and a radius plot and location plot were defined in Klemelä (2006). A short description and examples of these plots are given in Section 4.

The density estimation based approach to the mode and shape detection requires the availability of efficient density estimators. Kernel estimators are one of the most efficient estimators in moderate dimensional cases. However, kernel estimates are based on local averaging and suffer from the curse of dimension. For high dimensional data we have to apply different estimators.

Thus we give examples of the application of the tools also in the case of multivariate adaptive histograms.

In Section 2 we define and illustrate multiframe mode graphs. In Section 3 we define and illustrate branching maps. Section 4 introduces the scale and shape visualization table. Section 5 illustrates the application of the graphical tools for the case of kernel estimators and adaptive histograms. Section 6 contains a discussion. The appendix gives details of the construction of a multiframe mode graph.

Computations and graphics in this article have been made with R-packages "denpro" and "delt" which may be downloaded from http://denstruct.net.

2 Multiframe mode graph

2.1 Definition of a multiframe mode graph

In one and two dimensional cases the mode tree was introduced by Minnotte and Scott (1993). In a one dimensional mode tree the locations of the modes of kernel estimates are plotted when the smoothing parameter ranges over an interval. A two dimensional mode tree was defined to be a three-dimensional plot of the mode locations and bandwidth. A multivariate mode tree may also be defined as a tree which shows how the number of modes of a kernel estimate is increasing as a function of the smoothing parameter, without any spatial information. This kind of multivariate mode tree was considered in Scott and Szewczyk (2000), where it was applied to clustering.

We will define *multivariate mode graphs* as plots where we plot a one dimensional mode graph separately for each coordinate. A tree in the d-dimensional space is a 1D-structure which can be visualized with d projections. In order the projections to be useful one needs to identify the same node in different windows. One needs only to label the leaf nodes to identify uniquely all the nodes in different windows. We may however considerably ease the identification with a coloring scheme. It would not be feasible to choose a separate color for each node, but we get a useful coloring by choosing a separate color for each branch of the tree.

A multiframe mode graph is associated to a collection of density estimates

$$\hat{f}_h : \mathbf{R}^d \to \mathbf{R}, \qquad h \in H \tag{1}$$

where $H \subset \mathbf{R}$ is a finite set of smoothing parameters. Denote with

$$M_1^{(h)}, \ldots, M_{m_h}^{(h)} \in \mathbf{R}^d, \qquad h \in H, \tag{2}$$

the locations of the modes of the estimates, where m_h is the number of modes of the estimate with smoothing parameter h, $M_j^{(h)} = (M_{j,1}^{(h)}, \ldots, M_{j,d}^{(h)}) \in \mathbf{R}^d$ is the jth mode of the estimate with smoothing parameter h, $j = 1, \ldots m_h$. We interpret h as if it were the smoothing parameter of the kernel estimator: small values of h correspond to undersmoothed estimates and large values of h correspond to oversmoothed estimates.

Definition 1 A multiframe mode graph*, associated to a scale of density estimates (1) with mode locations (2), consists of d windows.*

- *The x-axis of the i:th window corresponds to the i:th coordinates of the modes and the y-axis of the windows corresponds to the scale H. That is, the i:th window, $i = 1, \ldots, d$, consists of the plot of points*

$$\left(M_{k,i}^{(h)}, h\right), \qquad h \in H, \;\; k = 1, \ldots, m_h,$$

- *To identify the same mode between different windows we use the same color to plot the same mode in different windows, but different colors to plot the different modes with the same h-value. That is, for each $h \in H$, $k = 1, \ldots, m_h$, points*

$$\left(M_{k,i}^{(h)}, h\right), \qquad i = 1, \ldots, d,$$

have the same color and for each $h \in H$, $i = 1, \ldots, d$, points

$$\left(M_{k,i}^{(h)}, h\right), \qquad k = 1, \ldots, m_h,$$

have different colors.

Tree or graph structure. In Definition 1 we did not define a tree or a graph. By adding to the mode graph parent-child connections we make the plot more easily interpretable. However, there does not seem to exist a distinguished choice for the parent-child relations, and thus we define these relations separately, instead of defining them in the proper definition of a multiframe mode graph. We give a rule for determining parent-child relations in Appendix A.

Coloring. Definition 1 gave a minimal condition for the coloring. We use the parent-child relations to enhance coloring. The coloring of the nodes is determined so that we choose first distinct colors for the modes at the root level, that is, for the modes corresponding to the largest smoothing parameter. For one of the children we choose the same color as that of the parent. For the other children we choose new distinct colors. The precise rule is given in Appendix A.

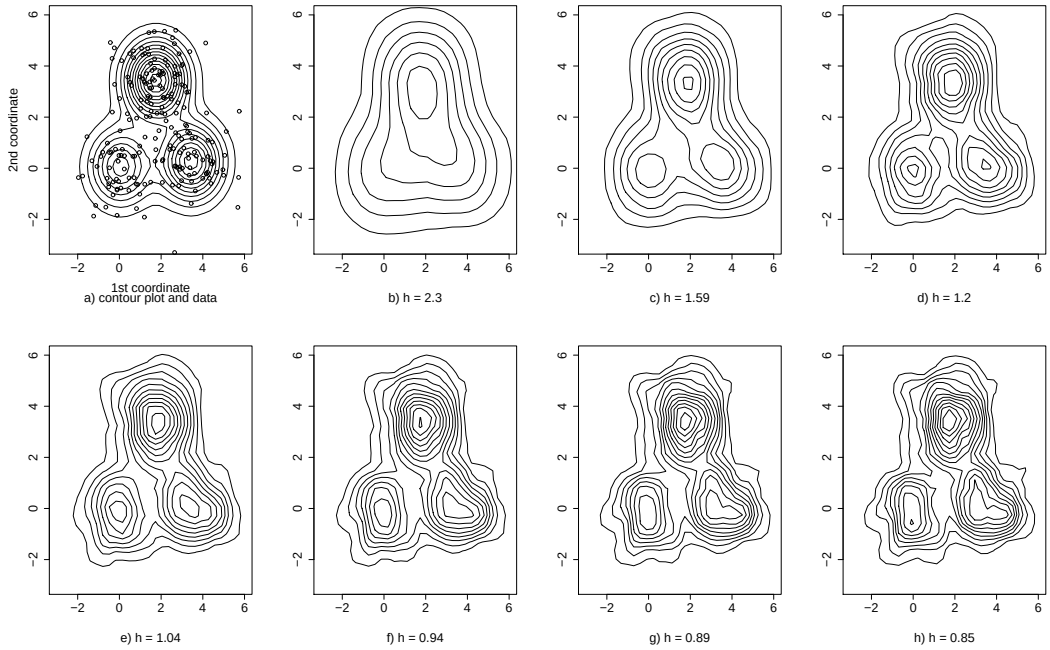

Figure 1: Frame a) shows the sample together with a contour plot of the density from which we generated the sample. Frames b-h) show contour plots of 7 kernel estimates with decreasing smoothing parameters.

2.2 Example

2.2.1 Scale of estimates

Figure 1a shows a contour plot of the density which we use for the illustration, and a sample of size 200 generated from this density. We construct a scale of kernel estimates. We apply Bartlett-Epanechnikov product kernel defined by $(x_1, \ldots, x_d) \mapsto (3/4)^d \Pi_{i=1}^d \max\{0, 1 - x_i^2\}$. Figure 1b-h show 7 kernel estimates corresponding to smoothing parameter values $(2.30, 1.59, 1.20, 1.04, 0.94, 0.89, 0.85)$.

2.2.2 Mode graph

Figure 2 illustrates a multiframe mode graph, associated to a scale of two dimensional kernel estimates. The kernel estimates were constructed from a sample of size 200 from the 3-modal density shown in Figure 1a. We applied a grid of 100 smoothing parameters in interval $[0.85, 2.3]$. The grid was not equally spaced but we used a logarithmic spacing and the h-axis has a logarithmic scale.

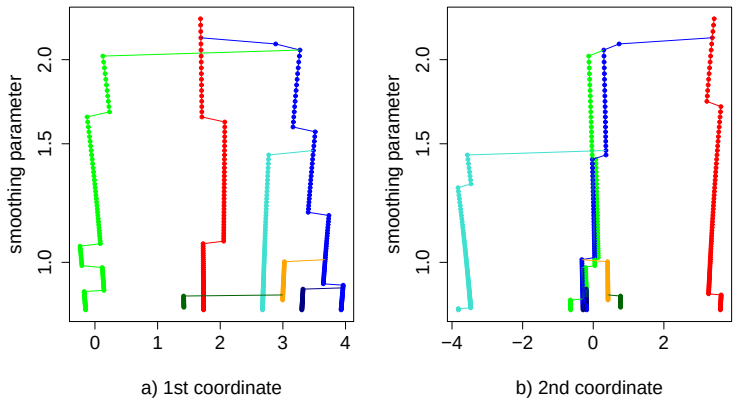

Figure 2: A multiframe mode graph of the scale of kernel estimates defined in Section 2.2.1.

Figure 2 shows that when we use smoothing parameter $h = 2.3$, then the estimate has one mode. The node corresponding to this mode is labeled with red. The multiframe mode graph shows that the location of this mode is $\approx (1.8, 2.8)$. When the smoothing parameter is decreased, then the blue and green branches appear. The red, blue, and green branches correspond indeed to the true modes of the density. The blue mode is at the location $\approx (3.5, 0)$ and the green mode is at the location $\approx (0, 0)$. The fourth branch is turquoise and it appears at a tail region, when $h \approx 1.5$. Finally, when the smoothing parameter is $h = 0.85$, there are 7 branches and thus 7 modes.

3 Branching map of a scale of level set trees

One may think about a function $f : \mathbf{R}^d \to \mathbf{R}$ in two ways: as a mapping $x \mapsto f(x)$ which assigns value $f(x)$ to a vector $x \in \mathbf{R}^d$, or as a mapping $\lambda \mapsto \{x \in \mathbf{R}^d : f(x) \geq \lambda\}$ which assigns a level set to a level. A mode graph considers a function in the first way but a map of branches considers a function in the second way. Since a level is real valued the change of the viewpoint makes it possible to visualize much information from a scale of functions in a concentrated way.

A branching map of level set trees of estimates (in a scale of estimates) is a perspective plot of a 2D function whose arguments are the level (density value) and the smoothing parameter, and the values are the excess masses of the nodes of the level set trees of estimates for a given smoothing parameter and at a given level. A map of branches shows for each estimate the levels

where the level sets of the estimate are decomposing to separated components, as we move to the higher levels, and it visualizes also the probability masses associated with these separated components. The probability masses (excess masses) associated to the separated regions of level sets measure the relative importance of the modes and bundles of modes. To define a branching map we define 4 preparatory concepts: (1) level set tree, (2) excess mass, (3) branching node, and (4) branching profile.

(1) Level set tree. A *level set tree* is a tree whose nodes represent the separated regions of the level sets of the function. The root nodes of the level set tree correspond to the separated regions of the lowest level set of the function. The child nodes of a given parent node correspond to the separated regions of the level set with one step higher level than the level of the parent node. A level set tree is defined in Klemelä (2004). Figure 3a shows a level set tree of the estimate in Figure 1d. The nodes corresponding to the 3 modes of the estimate are labeled as M1-M3.

(2) Excess mass. We want to condense the information contained in a level set tree in order to make it possible to represent information concerning a scale of multivariate estimates with a single 2D function. The first concept we use for the condensation of the information is the *excess mass associated with a node of a level set tree*, as defined in Klemelä (2004). This is the volume of the area which the density delineates over the level of the node, on the separated region of the level set associated with the node: $\int_A (f - \lambda)$, where f is the density (estimate) and A is a separated component of the level set with level $\lambda \geq 0$. In Figure 3a we have annotated 5 of the nodes with their excess masses.

(3) Branching node. The second concept we need for the condensation of the information contained in a level set tree is the concept of a *branching node of a level set tree*.

Definition 2 Branching nodes of a level set tree of a function $\mathbf{R}^d \to \mathbf{R}$ *are the nodes which have more than one child.*

Figure 3a shows the 2 branching nodes as red rectangles. The children of the branching nodes and the root node are shown as blue triangles. The tree which consists only of the root nodes of a level set tree, from the branching nodes, and from the children of the branching nodes is closely connected to the cluster tree as defined in Stuetzle (2003).

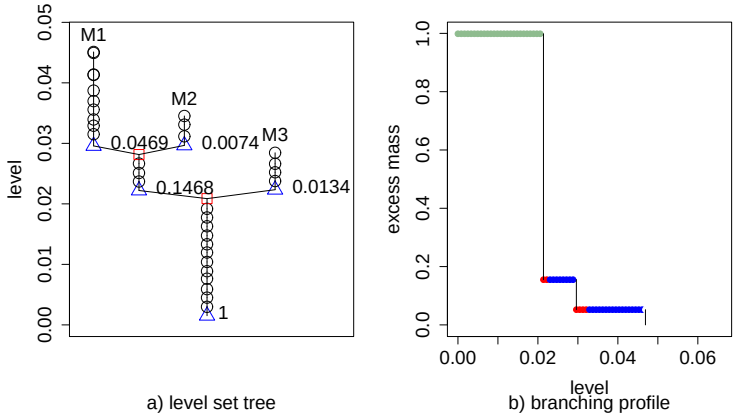

Figure 3: Frame a) shows a level set tree of the density estimate shown in Figure 1d. Frame b) shows the branching profile corresponding to the level set tree in frame a). The y-axis in the frame a) is the same as the x-axis in the frame b). These axes show the levels (density values) of the density estimate.

(4) Branching profile. Next we define a *branching profile of a level set tree*. It is a 1D plot which visualizes the number and the levels of branching of a level set tree of a function. It visualizes also the excess masses associated with the branching nodes and with the children of the branching nodes of a level set tree. The values of the plotted function are equal to the excess masses of the branching nodes. We divide the graph of the function between two levels of branching to two bands, which are colored in red and blue. The lengths of the two bands are proportional to the excess masses of the children of the branching nodes, that is, to the excess masses of the separated regions which are separating at this level.

Definition 3 *A* branching profile of a level set tree *of density* $f : \mathbf{R}^d \to \mathbf{R}$ *is a plot of the colored excess mass function, defined in the following.*

Let $b_1, \ldots, b_M$ *be the branching nodes of the level set tree of* f*. Let* λ_m *be the level of* b_m*,* $m = 1, \ldots, M$*. We assume that the branching nodes are ordered so that* $0 < \lambda_1 < \cdots < \lambda_M < \|f\|_\infty$*, where* $\|f\|_\infty = \sup_{x \in \mathbf{R^d}} f(x)$*. Denote with* excmas$(b)$ *the excess mass of node* b*.*

- *Define the* excess mass function $e : [0, \infty) \to [0, 1]$ *to be the function which gives for every level of branching the excess mass of the branching node. The function remains constant until the next level of branching. The function is equal to 1 at the origin. Thus we define the excess mass*

function by

$$e(t) = \sum_{m=0}^{M} \text{excmas}(b_m) I_{[\lambda_m, \lambda_{m+1})}(t),$$

where we denote $\lambda_0 = 0$, $\lambda_{M+1} = \|f\|_\infty$, *and* $\text{excmas}(b_0) = 1$.

- *We color each constant segment of the graph of the excess mass function so that the colors give information on the excess masses of the children of the branching nodes. We need to take into account that a level set tree may have several root nodes. Let* $r_1, \ldots, r_L$ *be the root nodes of the level set tree of* f. *Define 2 vectors of colors:* $\text{rootpaletti} = (seagreen, violet, \ldots)$ *and* $\text{paletti} = (red, blue, green, \ldots)$.
  - *Divide interval* $[0, \lambda_1)$ *to* L *subintervals so that* l*-th subinterval* I_{0l} *has length* $\text{excmas}(r_l)/\lambda_1$, $l = 1, \ldots, L$. *Choose color* $\text{col}(I_{0l}) = \text{rootpaletti}(l)$ *for each interval, assuming that the intervals are ordered so that* $\text{length}(I_{01}) < \cdots < \text{length}(I_{0L})$.
  - *Let node* b_m *has* N *children* $c_1, \ldots, c_N$. *Divide interval* $[\lambda_m, \lambda_{m+1})$, $m = 1, \ldots, M$, *to* N *subintervals so that* i*-th subinterval* I_{mi} *has length* $\text{excmas}(c_i)/(\lambda_{m+1} - \lambda_m)$, $m = 1, \ldots, M$, $i = 1, \ldots, N$. *Choose colors* $\text{col}(I_{mi}) = \text{paletti}(i)$ *for each interval, assuming that the intervals are ordered so that* $\text{length}(I_{m1}) < \cdots < \text{length}(I_{mN})$.
- *The branching profile is a plot of the graph* $(\lambda, e(\lambda))$, $\lambda \in [0, \|f\|_\infty]$, *where* $e(\lambda)$ *has color* $\text{col}(I)$ *when* $\lambda \in I$.

Figure 3b shows the branching profile corresponding to the level set tree in Figure 3a. The red color indicates always a new branch in the level set tree. Note that the absolute lengths of the color bands do not contain information, but the relative length of a red and blue band tells how the excess mass is distributed over the two branches.

Note the delicateness in Definition 3. We should not define the excess mass function $e : [0, \infty) \to [0, 1]$ so that it gives for a level λ the probability mass of the level set with level λ: $e(\lambda) \neq P_f(\{x \in \mathbf{R}^d : f(x) \geq \lambda\})$, where P_f is the probability measure corresponding to density f. Indeed, we have to take into account that the level set tree has several branches corresponding to the various modes and the total excess mass does not give information about the individual branches.

(5) Branching map. Now we are ready to define a branching map. By combining together the branching profiles of level set trees of estimates in a scale of estimates we get the branching map.

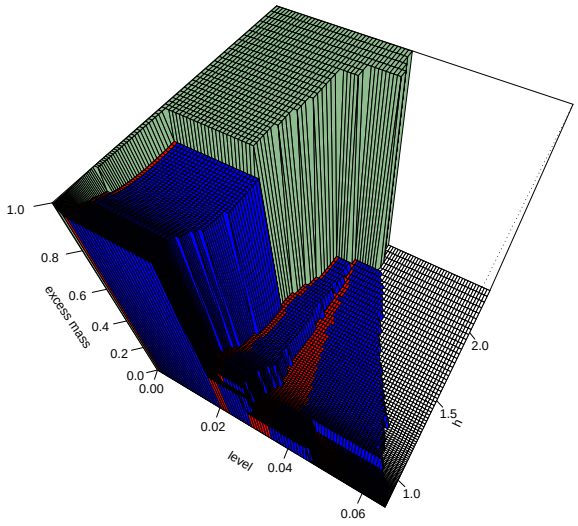

Figure 4: A map of branches of the scale of kernel estimates defined in Section 2.2.1.

Definition 4 A branching map of a scale of level set trees*, associated with the scale of estimates in (1), is the perspective plot of the 2D function* $B : [0, \infty) \times H \to [0, 1]$ *whose slices* $B(\cdot, h) : [0, \infty) \to [0, 1]$ *are the branching profiles of the estimates, defined in Definition 3*

Figure 4 shows a map of branches, constructed from the same scale of kernel estimates as the mode graph of Figure 2. The map of branches has color seagreen when h is large and this means that the estimates are unimodal for large h. The appearence of red bands signals the appearence of modes: a red band shows where the level set is decomposing to separate regions. One branching increases the number of modes by one. There appears two red bands almost simultaneously (when h is decreased) and this means that there appears three modes (1st branching implies two modes, 2nd branching implies 3 modes, ...). When h is further decreased, a third red band appears at a low level, and this means that a 4th mode appears. The relative widths of red and blue bands tell how the excess mass is distributed among the two modes (among the "old" mode and the "new" mode). The height of the surface shows the total remaining excess mass at this level, to be distributed among the modes.

One sees that the 3 modes, which appear when $h \approx 2$, emerge at relative high levels $0.02 - 0.03$, and the roots of the branches leading to these modes have non-negligible excess masses. The 4th mode, which appears when $h \approx 1.5$, emerges at a low level and it has a small excess mass. The rest of the modes appear at moderate levels and they have small excess masses.

Mode detection. In branching maps we focus on the branching structure of the level set tree. Thus branching maps are useful when we apply such an approach to mode detection where we consider whether the branches of the level set tree of a density estimate are really a true feature of the underlying density. See the discussion on mode testing in page 20.

4 Scale and shape visualization table

Definition. We define a dynamic tool for visualization of a scale of multivariate density estimates. We call this tool a *scale and shape visualization table*. The tool accompanies a mode graph and a branching map with visualizations of level set trees of the estimates and with visualizations of the shapes of the level sets of the estimates, allowing us to zoom into the shape of the estimates.

A scale and shape visualization table contains 6 windows and a control window. The first window shows one frame of a multiframe mode graph and the second window shows a branching map. The 3rd window shows a volume plot of one of the estimates in the scale and the 4th window shows one frame of a barycenter plot of the estimate. The 5th window shows a radius plot of a level set of the estimate and the 6th window shows a location plot of the level set of the estimate.

Figure 5 shows a screenshot of a scale and shape table corresponding to the scale of kernel estimates which was visualized with a mode graph im Figure 2 and with a map of branches in Figure 4.

Components of the table. Multiframe mode graphs were defined in Section 2 and branching maps were defined in Section 3.

A *volume plot* is a plot of the volume transform of a function. A volume transform of a multivariate function is a one dimensional function which is mode isomorphic to the original multivariate function. This means roughly that the volume transform has the same number of modes of the same size as the original function. The size of the modes is measured with their excess masses. A *barycenter plot* visualizes the barycenters of the level sets of a multivariate function, by showing the d projections to the coordinate axis

of a skeleton of the function. A skeleton of a function is a branching curve which goes through the barycenters of the level sets. A volume plot and barycenter plot were defined in Klemelä (2004).

A *radius plot* is a plot of the radius transform of a connected set. A radius transform of a connected set is a one dimensional function which is shape isomorphic to the original multivariate set. This means roughly that the radius transform has the same number modes of the same size as the original set has extensions. A radius transform of a set is roughly equivalent to a volume transform of a boundary function of the set. A *location plot* visualizes the location and the orientation of the set, by showing the d projections to the coordinate axis of a skeleton of the set. Note that location plots and barycenter plots use the same kind of projection and coloring technique as mode graphs. A radius plot and location plot depend on the chosen reference point, which is typically the mode of the density or the barycenter of the level set. A radius plot and location plot were defined in Klemelä (2006).

Dynamics. We may change the windows of the table through mouse clicks. We choose from the d coordinates the coordinate whose frame is shown in the window of the mode graph, we choose from the scale of estimates the estimate whose volume plot and barycenter plot are shown, we choose from the d coordinates the coordinate of the barycenter plot, we choose the level of the level set whose radius plot and location plot are shown, and we choose from the d coordinates the coordinate of the location plot. In addition, we may rotate the map of branches, zoom into the volume plot, and change the reference point for the radius and location plot. Interactivity is implemented with the "locator" function of R. For example, the rotation is implemented by defining a grid on the polar coordinates and through a mouse click one moves one step on that grid.

Example. Figure 5 shows a scale and shape visualization table. The mode graph shows the 1st coordinate and the mode labels $M1$, $M2$, $M3$ are at the height of the smoothing parameter $h = 1.562$, which is the smoothing parameter of the kernel estimate visualized in frames $III - VI$. The correspondence of the modes between the volume plot and the mode graph is shown with the coloring and with labels $M1$, $M2$, $M3$. The coloring and the labeling connects also the barycenter plot to the volume plot and to the mode graph. The colors at the lower levels of the volume plot and the barycenter plot are not connected to the coloring of the mode graph. The barycenter plot shows the first coordinate. The radius plot visualizes the level set with level 0.005, the location plot shows the first coordinate, and the reference

point in these plots is the barycenter of the level set. The colors connect the radius plot and the location plot together but these colors are not related to the colors in the other plots.

5 Examples

5.1 Kernel estimator

As an example we consider the two dimensional lipid data. The data set consists of the 320 lipid (cholesterol and triglyceride) levels of men with heart disease. This data was analyzed by Scott, Gotto, Cole and Gorry (1978) and Minnotte and Scott (1993). Logarithmic transformation is applied for the both variables and marginal data is standardized to have sample variance 1. We applied the standard Gaussian kernel in the kernel estimates.

Figure 6 shows the 2 frames of the mode graph. One sees that the two first new modes (blue and green nodes) appear at tail regions, for smoothing parameter values $h \approx 0.58$ and $h \approx 0.5$. Then the 4th and the 5th modes (orange and turquoise nodes) appear simultaneously, when $h \approx 0.38$. The orange mode is in the central region and the turquoise node is at a tail region.

Figure 7 shows a map of branches for the lipid data. One sees that the two first new modes appear at low levels, and have very small excess masses. One sees that from the 4th and 5th modes one of these modes emerges at a low level and has a very small excess mass, but the other mode appears at a higher level and although its excess mass is small, the excess mass is more evenly distributed among the emerging mode and the old mode.

Figure 8 shows perspective plots of the estimates for the smoothing parameters $h = 0.36$ and $h = 0.46$. Note that the small modes at the tail regions are difficult to notice from the perspective plot.

5.2 CART histogram

We consider multivariate adaptive histograms, which work in some high dimensional cases where the kernel estimator fails. These CART histograms are otherwise similar to regression trees defined in Breiman, Friedman, Olshen and Stone (1984), but the regressogram is replaced by the histogram and the sum of squared errors is replaced by the negative log-likelihood. The CART histograms are based on data dependent partitions. We find the data dependent partition by first constructing a collection of partitions by minimizing the negative log-likelihood in a myopic fashion, and then choosing the final partition from this collection by minimizing a complexity penalized

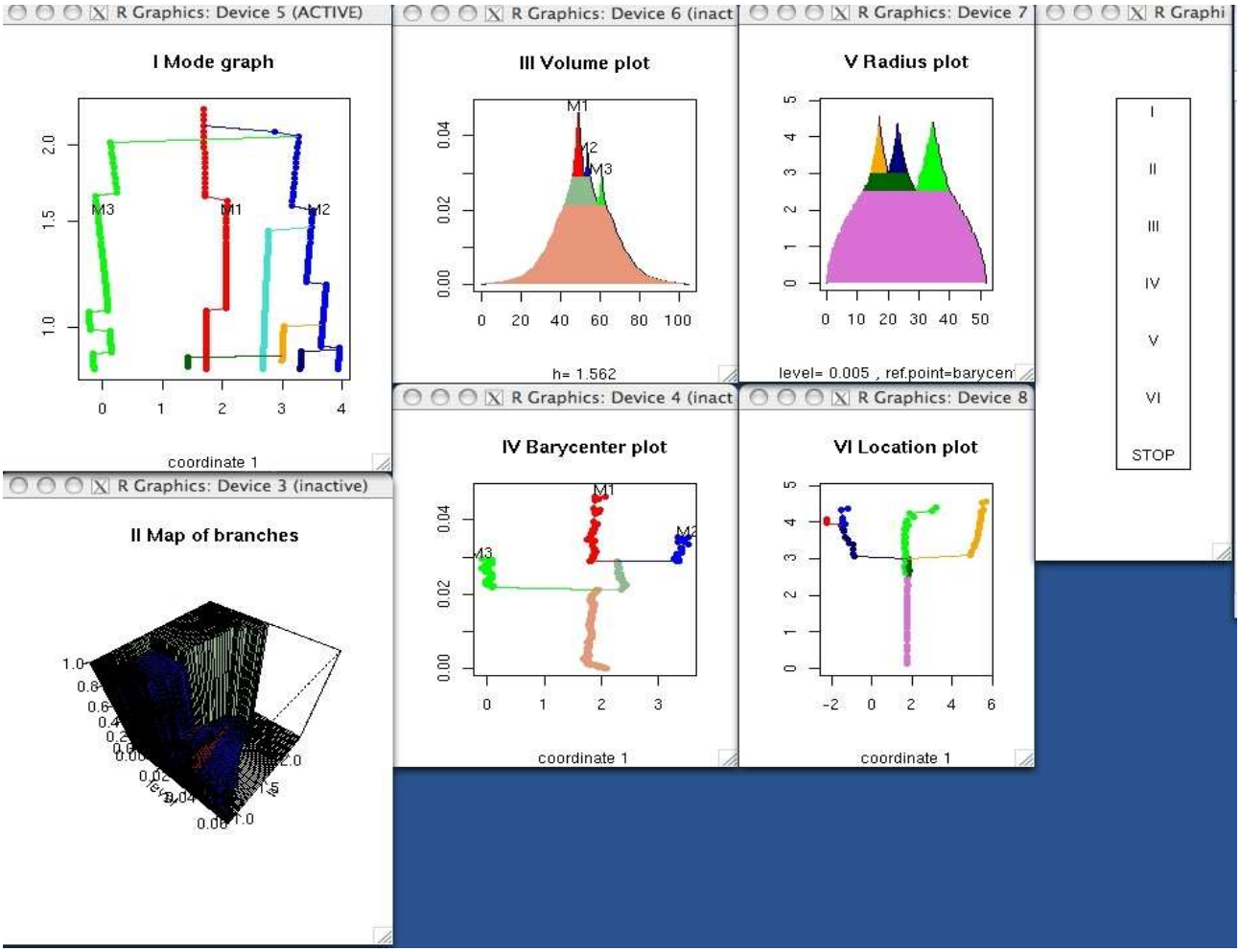

Figure 5: A screenshot of a scale and shape visualization table corresponding to the scale of estimates in the mode graph of Figure 2 and in the map of branches of Figure 4.

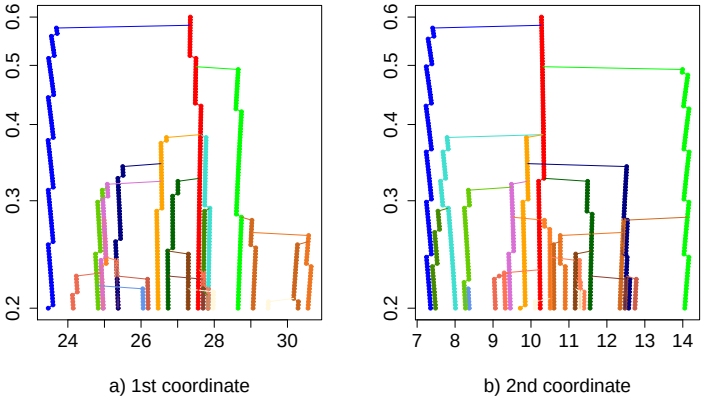

Figure 6: A multiframe mode graph for the lipid data; a) Cholesterol-frame; b) Triglyceride-frame.

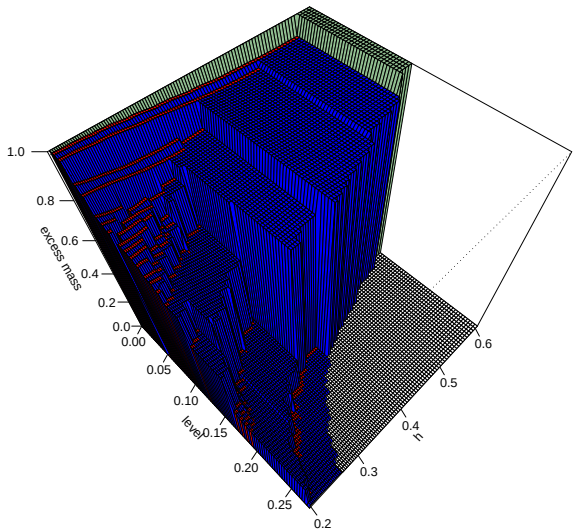

Figure 7: A map of branches for the lipid data.

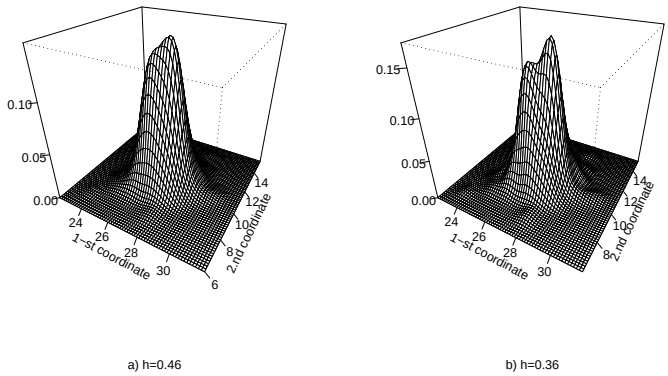

Figure 8: Perspective plots of kernel estimates for the lipid data.

likelihood criterion, where the complexity is taken to be the number of sets in the partition. The smoothing parameter of the estimator is taken to be the number of sets in the partition determining the histogram.

We consider a modification of the simulation example which was introduced by Friedman, Stuetzle and Schroeder (1984) in the connection of projection pursuit density estimation. We try to estimate the density which is the equal mixture of 5 dimensional densities $N(\mu_i, \Sigma)$, $i = 1, 2, 3$, where

$$\mu_1 = (0, c, 0, 0, 0),\ \mu_2 = (3, -c, 0, 0, 0),\ \mu_3 = (-3, -c, 0, 0, 0),$$

$c = 3^{3/2}/2 \approx 2.6$, and $\Sigma = \mathrm{diag}(1, 1, 7, 7, 7)$. Thus we add 3 pure noise dimensions, that are independent normal random variables with zero mean and variance 7, to a two dimensional random vector whose density function is the equal mixture of three two-dimensional standard Gaussian distributions with means $(0, c)$, $(3, -c)$, $(-3, -c)$. The number c was chosen in such a way that also the variances of the first two marginal distributions are 7. We will generate a sample of size 225 from this density. See also Scott and Wand (1991) who made simulations with this example, applying kernel estimators.

Figure 9 shows a multiframe mode graph. The y-axis of the mode graph shows the opposite number of the number of sets in the partitions defining the histograms. The mode graph shows that the estimates have 3 modes (red, blue, light green) over a large range of smoothing parameter values. In fact, when the partition has $\approx 10 - 40$ sets, then the histograms have 3 modes. The locations of these 3 modes are close to the locations of the true modes, when the number of sets is $\approx 10 - 30$. When the partition has ≈ 60 sets, then the histogram has 6 modes.

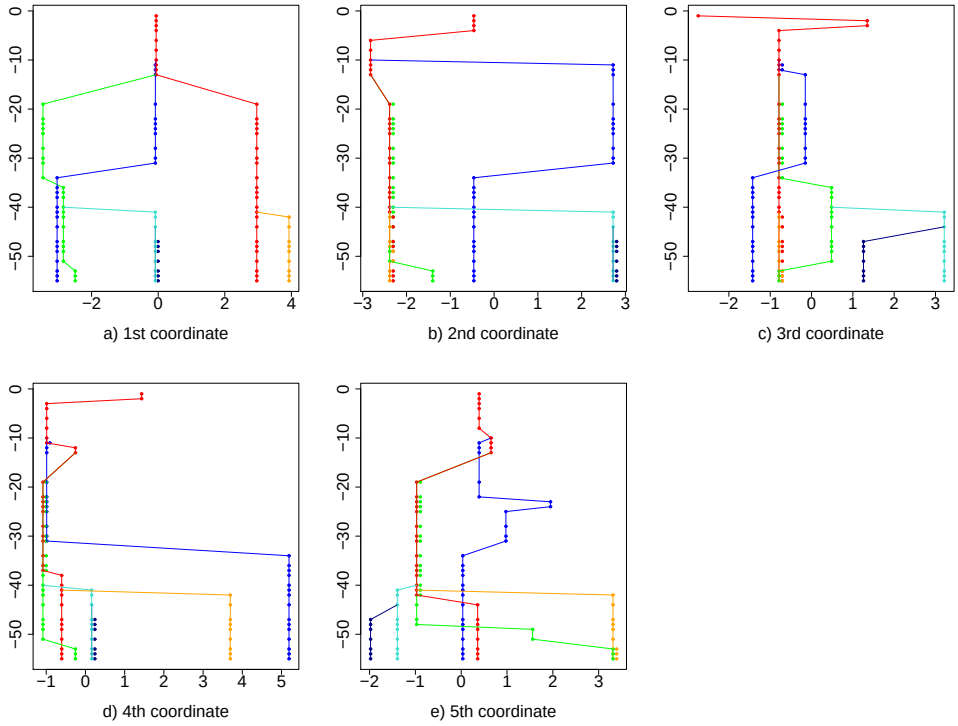

Figure 9: CART-histogram: a multiframe mode graph for the projection pursuit example. The first two frames are most important because they show the signal dimensions.

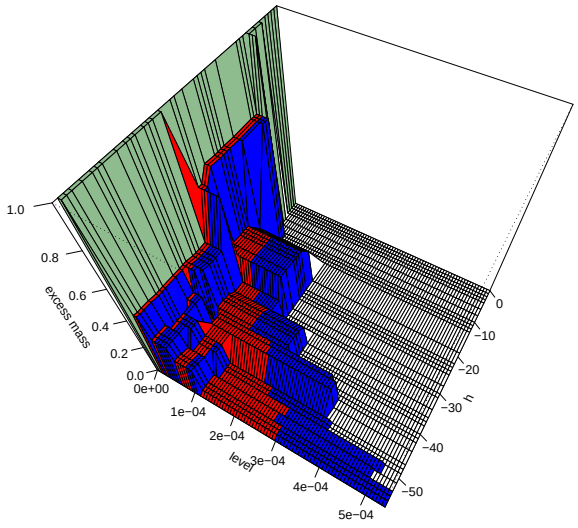

Figure 10: CART-histogram: a map of branches for the projection pursuit example.

Figure 10 shows a branching map corresponding to the mode graph in Figure 9. From the branching map one sees that the excess masses related to the modes appearing when the number of sets in the partitions is larger than 40 is much smaller than the excess masses of the 3 previous modes.

6 Discussion

Mode trees. Minnotte and Scott (1993) introduces mode trees and visualizes excess masses with these trees. They draw the enhanced mode tree by plotting black regions centered at each mode location and whose horizontal width at each level of the smoothing parameter h represents the excess mass of the mode. We have introduced visualization tools with two improvements to the classical mode tree. First, we introduce multiframe mode trees which make it possible to draw mode trees in arbitrary dimension. Secondly, we visualize excess masses with branching maps and volume plots which contain more information than merely the excess masses associated with the modes of the estimate.

Density estimators. Previously mode trees have been preferred to be drawn for kernel estimates with the standard Gaussian kernel. This choice guarantees in the univariate case that the number of modes is monotonically increasing when the smoothing parameter is decreasing. We have illustrated that one does not have to restrict oneself to this particular estimator. We may use kernel estimates with the Bartlett-Epanechnikov product kernel for computational reasons and we may use adaptive histograms to make high dimensional density estimation more efficient. For reasonable estimators the range of smoothing parameter values where the number of modes is not increasing monotonically is typically small and the non-monotonicity of the number of modes does not substantially decrease the interpretability of the mode graphs.

Mode testing. Minnotte (1997) introduces a formal testing procedure associated with a mode tree. The testing procedure relies on critical smoothing parameter values, which are such values where new modes appear. Thus the testing procedure tends to be restricted to the application of univariate kernel estimates with the standard Gaussian kernel, since this is the only known estimator where the number of modes is growing monotonically as the smoothing parameter is decreasing.

A natural alternative is to consider each single density estimate separately, to look at the levels where the level sets are splitting, and to test whether the splitting is a real feature of the underlying density. At each branching node of the level set tree we calculate the excess masses of the children and make a judgment whether the excess masses are so large that one would not expect them to arise due to random fluctuations, for the available sample size. The testing is started at the branching nodes closest to the root nodes and one proceeds recursively towards the upper levels. Only if the null-hypothesis of no branching is rejected, one needs to proceed further towards the upper levels. Figure 11 illustrates the approach. Frames a) and b) show a case where we first test the existence of mode bundle A and mode B, and if these exist, then we proceed to test the existence of modes C and D. Frames c) and d) show a case where it could happen that mode bundle A has so small excess mass that it is not detected, and then one does not proceed to test the existence of modes C and D.

With this approach we are free to apply estimators which are such that the number of modes is not monotonic with respect to the smoothing parameter value. We have not considered in this article formal mode testing procedures. We have only visualized the branching structure; associating a formal mode testing procedure to a branching map would give us a multivariate excess

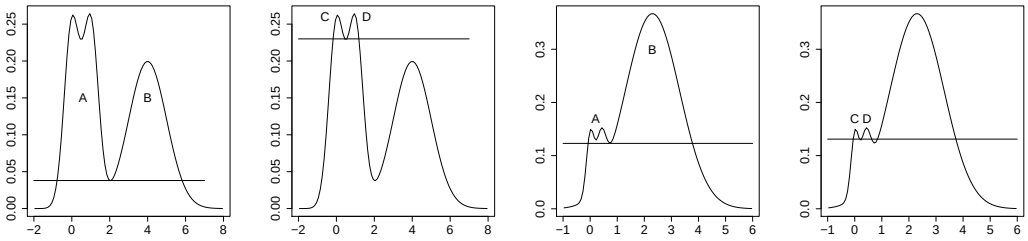

Figure 11: A principle of mode testing. Frames a) and b) show a case where one could easily detect mode bundle A and mode B, and proceed to test the existence of modes C and D. Frames c) and d) show a case where the detection of mode bundle A and mode B is difficult and one might not proceed to test the existence of modes C and D.

mass version of the SiZer.

Single bandwidth. We have assumed in this article that the density estimator has a single real valued smoothing parameter. One has suggested more flexible kernel estimators with a vector or a matrix of smoothing parameters. These improvements of kernel estimates are useful when one needs to choose one estimate (say, for presentation purposes), but since we go through all values of the smoothing parameter, we can detect many features with a real valued smoothing parameter. Even in the univariate case one would like to use spatially adaptive smoothing parameters, but this is not needed when one goes through the complete scale of estimates, as pointed out in Chaudhuri and Marron (2000). Standardization of the scales should however be used in many cases.

On the other hand, when one uses methods of complexity penalization, like CART histograms, then the spatial adaptivity and adaptivity across the dimensions is handled by choosing a flexible class over which a minimizer of the complexity penalized empirical risk is searched. In the case of CART histograms this class is the class of histograms whose partition can be obtained by a recursive splitting of the sample space. The amount of penalization is the real valued smoothing parameter. (In the case of CART histograms this is equivalent to taking the number of sets in the partition to be the smoothing parameter).

Acknowledgments

Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-2. I wish to thank the Editors and the Referees for helpful comments.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

Chaudhuri, P. and Marron, J. S. (1999), 'Sizer for exploration of structures in curves', *J. Amer. Statist. Assoc.* **94**, 807–823.

Chaudhuri, P. and Marron, J. S. (2000), 'Scale space view of curve estimation', *Ann. Statist.* **28**, 408–428.

Friedman, J. H., Stuetzle, W. and Schroeder, A. (1984), 'Projection pursuit density estimation', *Amer. Statist. Assoc.* **79**, 599–608.

Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002), 'Significance in scale space for bivariate density estimation', *J. Comput. Graph. Statist.* **11**, 1–22.

Hartigan, J. A. (1987), 'Estimation of a convex density cluster in two dimensions', *J. Amer. Statist. Assoc.* **82**, 267–270.

Klemelä, J. (2004), 'Visualization of multivariate density estimates with level set trees', *J. Comput. Graph. Statist.* **13**(3), 599–620.

Klemelä, J. (2006), 'Visualization of multivariate density estimates with shape trees', *J. Comput. Graph. Statist.* **15**(2), 372–397.

Marchette, D. J. and Wegman, E. J. (1997), 'The iterated mode tree', *J. Comput. Graph. Statist.* **6**, 143–159.

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

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

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

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

Scott, D. and Szewczyk, W. F. (2000), 'The stochastic mode tree and clustering'. To appear.

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', *J. Chronic Diseases* **31**, 337–345.

Stuetzle, W. (2003), 'Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample', *J. Classification* **20**(5), 25–47.

A The parent-child relations for a mode graph

Let $H = \{h_1, \ldots, h_L\}$, where $h_1 < \cdots < h_L$, be the set of smoothing parameters. Denote with $\mathcal{M}_h$ the set of modes corresponding to smoothing parameter $h \in H$:

$$\mathcal{M}_h = \{M_1^{(h)}, \ldots, M_{m_h}^{(h)}\}.$$

We assume given a procedure *vectormatch* which finds for two finite sets of vectors $\mathbb{X}, \mathbb{Y} \subset \mathbf{R}^d$, $\#\mathbb{X} \leq \#\mathbb{Y}$, an injective mapping $vm : \mathbb{X} \to \mathbb{Y}$. The injectivity means that $vm(x_1) \neq vm(x_2)$, when $x_1 \neq x_2$.

1. The modes in $\mathcal{M}_{h_L}$, corresponding to the largest smoothing parameter h_L, are the root nodes. The modes in $\mathcal{M}_{h_1}$ are leaf nodes.

2. We define child nodes for the modes in $\mathcal{M}_{h_i}$, $i = L, \ldots, 2$.

   (a) Assume that $\#\mathcal{M}_{h_i} \leq \#\mathcal{M}_{h_{i-1}}$: the number of modes is at this step increasing as the smoothing parameter is decreasing. (This is the usual case.)

   Let $\mathbb{X} = \mathcal{M}_{h_i}$ and $\mathbb{Y} = \mathcal{M}_{h_{i-1}}$.

   If $vm(x) = y$, then y is a child of x. The color of y is the same as that of x.

   Let M be a mode which had not a parent assigned to it: $M \in \mathbb{Y} \setminus \{vm(x) : x \in \mathbb{X}\}$. We let the closest member in $\mathbb{X}$ to be the parent of M. We choose a new color for M.

   (b) Assume that $\#\mathcal{M}_{h_i} > \#\mathcal{M}_{h_{i-1}}$: the number of modes is at this step decreasing as the smoothing parameter is decreasing. (This is the unusual case.)

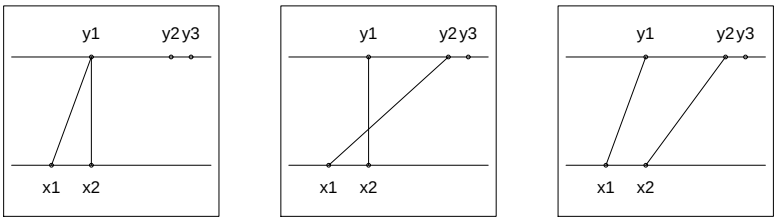

Figure 12: Finding an injection $vm : \mathbb{X} \to \mathbb{Y}$, where $\mathbb{X} = \{x_1, x_2\}$ and $\mathbb{Y} = \{y_1, y_2, y_3\}$. Frame a) shows a conflict where x_1 has y_1 as its closest and x_2 has y_1 as its closest. Frame b) shows a suboptimal resolution of the conflict. Frame c) shows a better resolution of the conflict.

> Let $\mathbb{X} = \mathcal{M}_{h_{i-1}}$ and $\mathbb{Y} = \mathcal{M}_{h_i}$.
> If $vm(x) = y$, then y is the parent of x. The color of x is the same as that of y.
> Let M be a mode which had not a child assigned to it: $M \in \mathbb{Y} \setminus \{vm(x) : x \in \mathbb{X}\}$. Mode M is a leaf node of the mode tree.

It is left to describe the procedure *vectormatch* for finding injection vm. Function vm should be such that x and $vm(x)$ are close. The number of all injections is

$$\#\mathbb{Y} \cdot (\#\mathbb{Y} - 1) \cdots (\#\mathbb{Y} - \#\mathbb{X} + 1),$$

which is so large number that one has to find a suboptimal solution. When we find for each $x \in \mathbb{X}$ the closest $y \in \mathbb{Y}$, then we do not get an injective mapping vm in all cases. One has to find a way to resolve the conflicts. If x_1 and x_2 are competing over y, that is, x_1 has y as its closest in $\mathbb{Y}$ and x_2 has y as its closest in $\mathbb{Y}$, then a simple may to resolve the conflict would be to take $vm(x_1) = y$ if $\|x_1 - y\| \leq \|x_2 - y\|$, and $vm(x_2) = y$ otherwise. This might lead to a bad overall matching, see Figure 12. We have defined the procedure *vectormatch* so that one resolves a conflict by looking such matching pairs that both x_1 and x_2 find a relatively good match. One fixes the better of these matches and continues by finding for each remaining $x \in \mathbb{X}$ the closest remaining $y \in \mathbb{Y}$. If there is a conflict, one resolves it as before, otherwise we are done. The precise algorithm is given in the technical report.

B Procedure *vectormatch*

We give a pseudo code of the procedure *vectormatch* for finding vm.

1. For each $x \in \mathbb{X}$ find the closest $y \in \mathbb{Y}$ in the Euclidean metric, denote $vm_1(x) = y$.

2. If $\#\{x \in \mathbb{X} : vm_1(x) = y\} = 1$ for each $y \in \mathbb{Y}$, then we return $vm = vm_1$. (If vm_1 is injective then we are done.)

3. Else

   (a) Set $\mathbb{X}_0 = \mathbb{X}$ and $\mathbb{Y}_0 = \mathbb{Y}$ to be the sets of available vectors.

   (b) Repeat until $\#A(y) = 1$ for each $y \in \mathbb{Y}_0$, where $A(y) = \{x \in \mathbb{X}_0 : vm_1(x) = y\}$.

      i. Set $B = \cup\{A(y) : y \in \mathbb{Y}_0,\ \#A(y) > 1\}$ to be the set of vectors $x \in \mathbb{X}_0$ which have competitors.
      For each $x \in B$, let $vm_2(x)$ be the 2nd closest to x in $\mathbb{Y}_0$, after $vm_1(x)$.

      ii. We go through all ordered subsets (x, z) of size 2 from B, and calculate

      $$crit(x, z) = \|vm_1(x) - x\|^2 + \|vm_2(z) - z\|^2.$$

      That is, when $x, z \in B$, $x \neq z$, we calculate $crit(x, z)$ and $crit(z, x)$.

      iii. We find the minimal value of $crit(x, z)$ over all ordered subsets of B of size 2.
      When (x_0, z_0) achieves the minimum, set

      $$vm(x_0) = vm_1(x_0).$$

      iv. Set

      $$\mathbb{X}_0 = \mathbb{X}_0 \setminus \{x_0\}, \qquad \mathbb{Y}_0 = \mathbb{Y}_0 \setminus \{vm_1(x_0)\}.$$

      and for each $x \in \mathbb{X}_0$ we find the closest $y \in \mathbb{Y}_0$, and set $vm_1(x) = y$.
      The idea is that we do not simply choose from B the vector which is closest to y but take into account whether the choice allows further good choices: we take into account the distance to the second best choice to guarantee that there remains potential for further good choices.

4. Set $vm(x) = vm_1(x)$ for those x for which $vm(x)$ was not yet determined and return vm.