Generalized Single Linkage Clustering

Werner Stuetzle
Rebecca Nugent

Department of Statistics
University of Washington

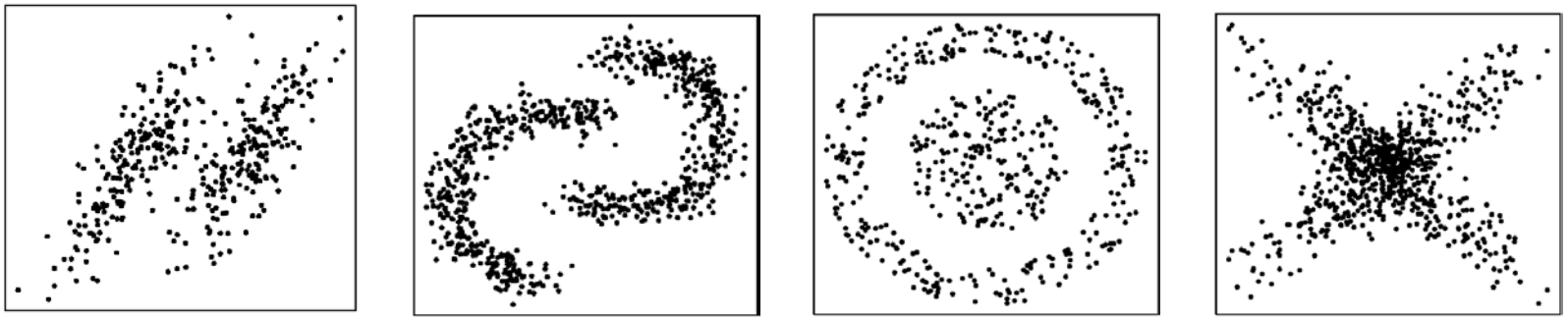
1. Goal of clustering

Detect presence of distinct groups.

Definition of *distinct groups* (Carmichael, George, and Julius):

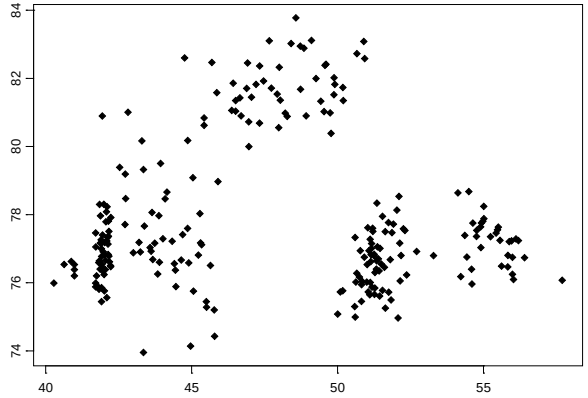
Contiguous, densely populated areas of feature space, separated by contiguous, relatively empty regions.

(a) - (c): Distinct groups in the sense of CG&J;
(d): not covered by definition.



(a) (b) (c) (d)

2. Statistical approaches to clustering



- Detect that there are 5 or 6 distinct groups.
- Assign group labels to observations.

Need to specify sampling model and population characteristic of interest.

Without sampling model, concept of "cluster validity" does not make sense.

Without well specified population characteristic it is impossible to evaluate and compare clustering methods $\Rightarrow$ no "progress".

Sampling model in this talk:

Feature vectors $\underline{x}_1, \ldots, \underline{x}_n$ are iid sample from some density $p(\underline{x})$.

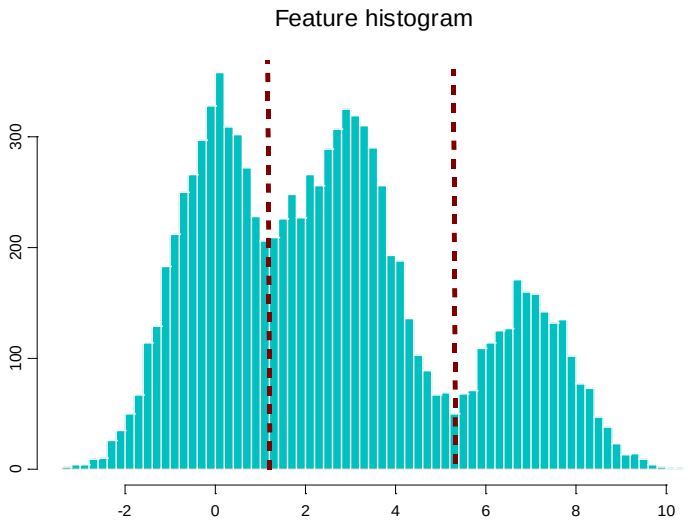
Nonparametric approach

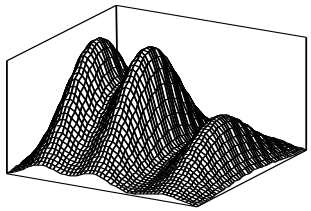
Based on premise that groups correspond to modes of density $p(\underline{x})$.

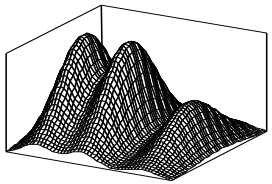
Clustering methods should be able to "detect and resolve distinct data modes, independently of their shape and variance" (Wishart 1969).

Need to

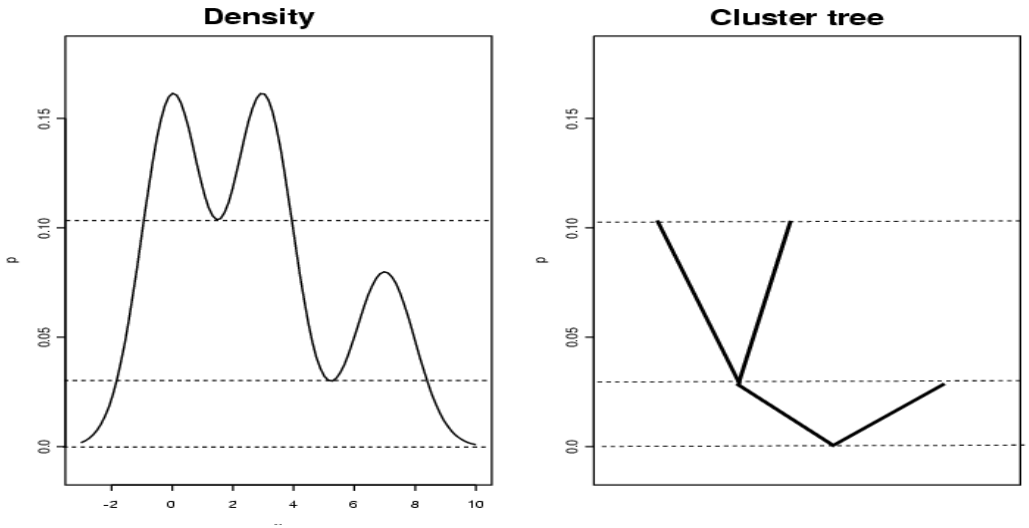
- Estimate modes;
- Assign each observation to the "domain of attraction" of a mode.







Structure of level sets is described by cluster tree



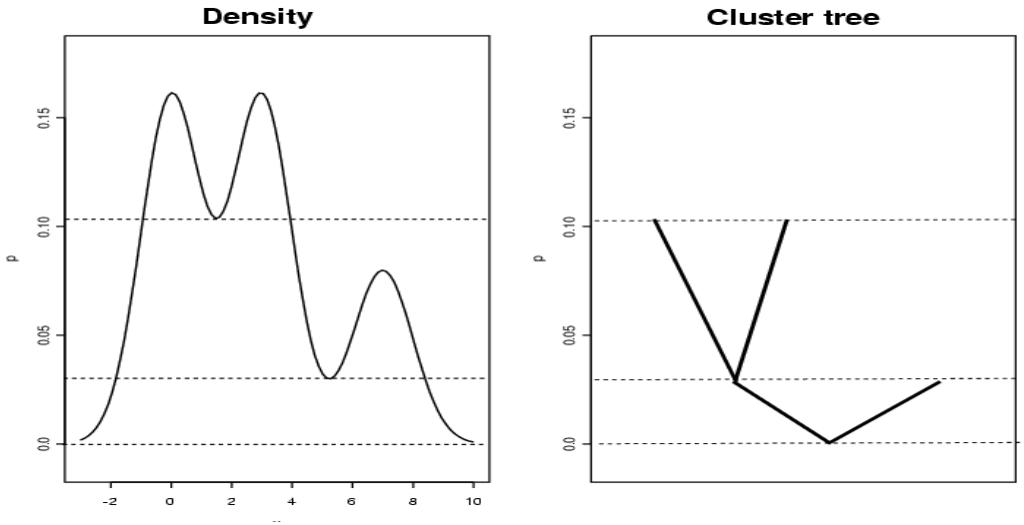
Cluster tree is easiest to define recursively:

Each node N of cluster tree

- represents a subset D(N) of feature space (high density cluster);
- is associated with a density level $\lambda(N)$.

Root node

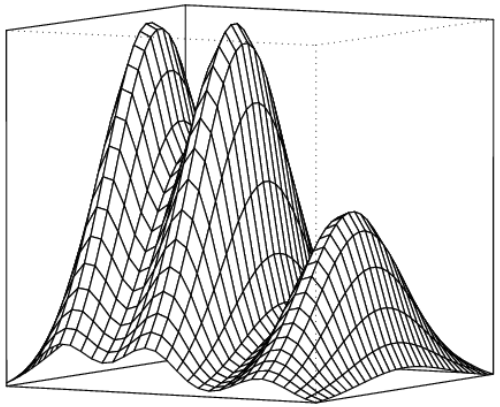
- represents the entire support of the density;
- is associated with density level $\lambda(N) = 0$.

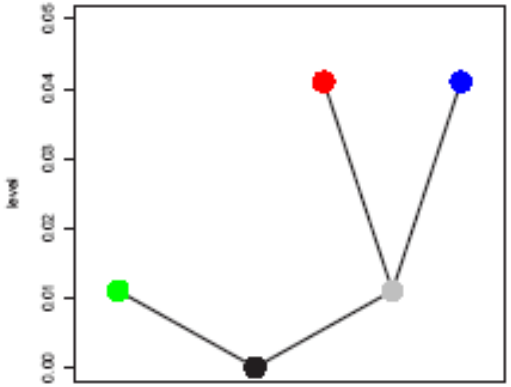


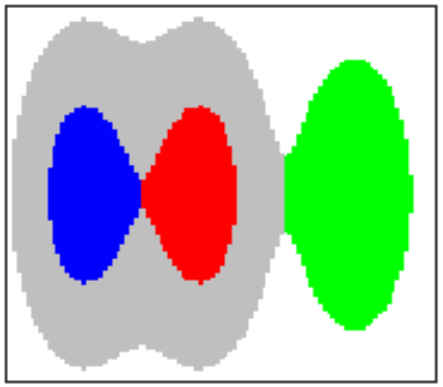
To determine descendents of node N:

- Find lowest level λ_d for which $L(\lambda; p) \cap D(N)$ has two connected components.
- If there is no such λ_d then N is leaf of the tree.
- Otherwise, create daughter nodes representing the connected components, with associated level λ_d, and recurse.

Density, cluster tree, and high density clusters







Leaves of cluster tree correspond to modes of density.

Cluster tree is invariant under non-singular affine transformations of feature space.

Cluster tree is (a) target population characteristic in non-parametric clustering.

4. Plug-in estimates of the cluster tree

Obvious idea:

- Estimate p by (typically nonparametric) density estimate $\hat{p}$;
- Estimate cluster tree of p by cluster tree of $\hat{p}$.

However, there are computational as well as statistical problems.

(i) Computational problem:

How can we compute level sets and their connected components?

(ii) Statistical problem:

How do we distinguish spurious components (modes) due to sampling variability from real components reflecting the structure of the underlying density?

Computing level sets for piecewise constant density estimates

For density estimates $\hat{p}$ that are piecewise constant over (hyper-) rectangles:

$$\hat{p}(\underline{x}) = \sum_{i=1}^{m} c_i \, I(\underline{x} \in R_i) \,,$$

level sets, their connected components, and the cluster tree can be computed exactly.

Example: Histograms, ASH estimates, piecewise constant approximations of other estimates.

More in Rebecca Nugent's talk (Session 274, Tuesday, 10:35 - 12:15)

Problem: method only viable in low dimensions (≤ 4?)

Otherwise, have to use approximations.

5. A graph based approach for approximating the cluster tree of a density (estimate)

Given: Observations $\mathcal{X} = \underline{x}_1, \ldots, \underline{x}_n \in R^m$ and density estimate $\hat{p}$.

Define graph G over observations with edge weights

$$p_{ij} = \min_{t \in [0,1]} \hat{p}((1-t)\underline{x}_i + t\underline{x}_j)$$

(minimum of estimated density along line segment $[\underline{x}_i, \underline{x}_j]$).

Sample level set $\tilde{L}(\lambda; \hat{p}, \mathcal{X})$: subgraph of G obtained by removing all edges and vertices with $p_{ij} \leq \lambda$.

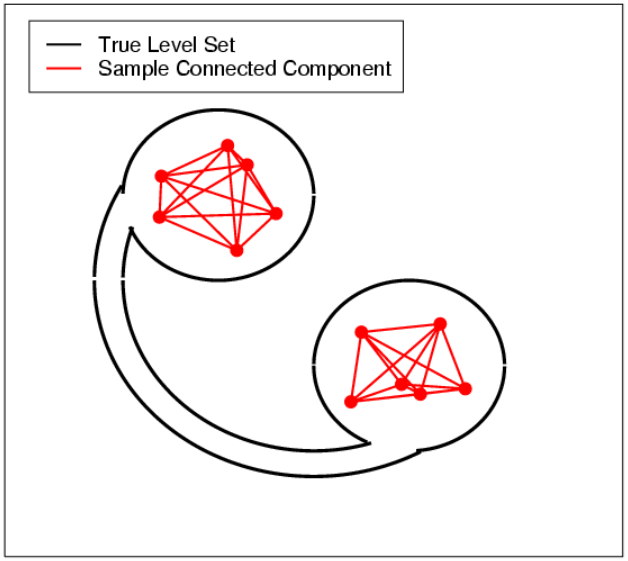
Sample connected components: Connected components of graph $\tilde{L}(\lambda; \hat{p}, \mathcal{X})$.

Note:

If $\underline{x}_i$ and $\underline{x}_j$ are in same connected component of $\tilde{L}(\lambda; \hat{p}, \mathcal{X})$ then $\underline{x}_i$ and $\underline{x}_j$ in same connected component of $\mathbf{L}(\lambda; \hat{p})$.

(There is path connecting $\underline{x}_i$ and $\underline{x}_j$ along which $\hat{p} > \lambda$.)

Reverse not necessarily true.

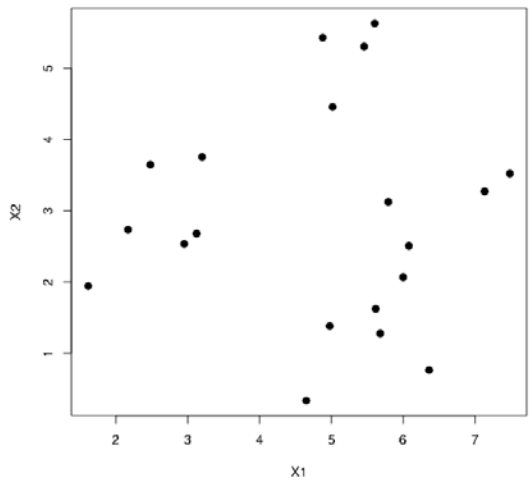


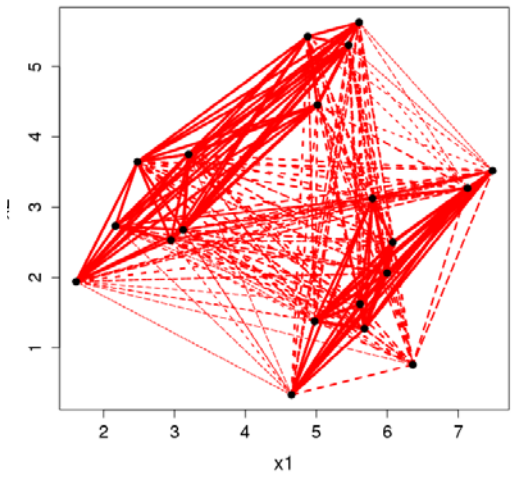
Sample connected components of $\tilde{L}(\lambda; \hat{p}, \mathcal{X})$ for different λ's have tree structure just like connected components of $L(\lambda; \hat{p})$.

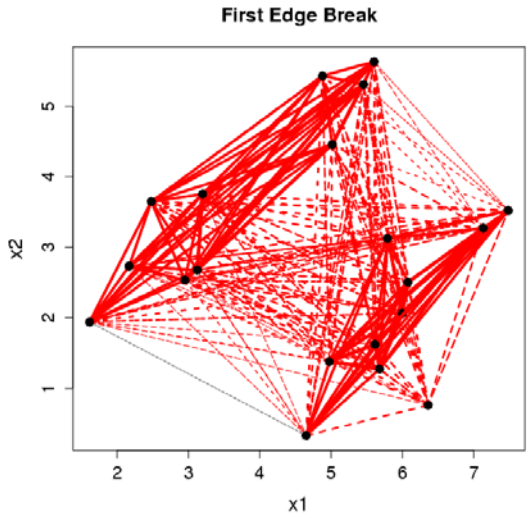
Define *sample cluster tree*. Each node N
- represents a subgraph $D(N)$ of $G = \tilde{L}(0; \hat{p}, \mathcal{X})$ (sample high density cluster);
- is associated with a density level $\lambda(N)$.

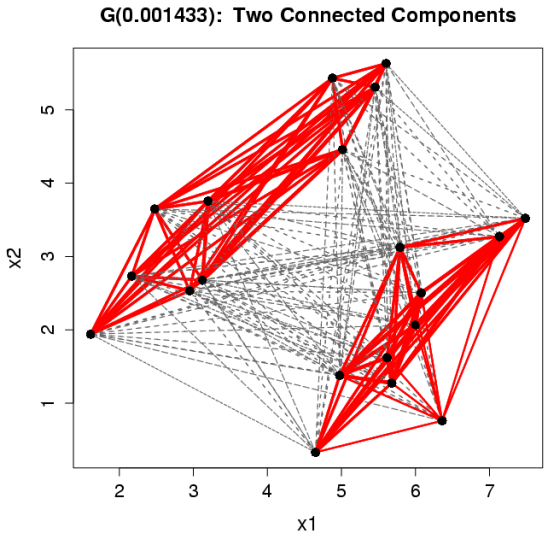
Root node
- represents the entire graph G;
- is associated with density level $\lambda(N) = 0$.

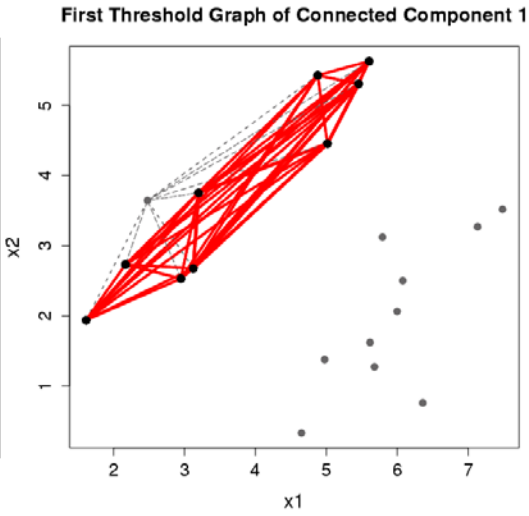
To determine descendents of node N:
- Find lowest level λ_d for which $\tilde{L}(\lambda; \hat{p}, \mathcal{X}) \cap D(N)$ has two connected components.
- If there is no such λ_d then N is leaf of the tree.
- Otherwise, create daughter nodes representing the connected components, with associated level λ_d, and recurse.

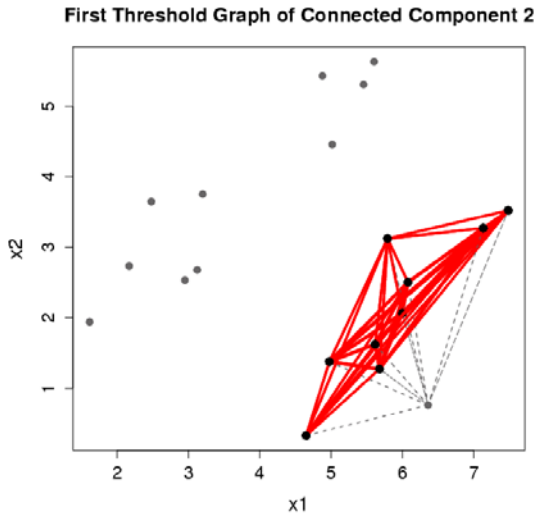

x2
x1


x1


First Edge Break
x2
x1


G(0.001433): Two Connected Components
x2
x1


First Threshold Graph of Connected Component 1
x2
x1


First Threshold Graph of Connected Component 2
x2
x1

Problem: The edge weights

$$p_{ij} = \min_{t \in [0,1]} \hat{p}((1-t)\underline{x}_i + t\underline{x}_j)$$

are solutions of optimization problem.

"Hack" approach: Evaluate $\hat{p}$ on grid along segment $[\underline{x}_i, \underline{x}_j]$.

More principled approach:

- To construct sample cluster tree we only need order of edge weights;
- For many density estimates we can obtain upper bound on 2nd derivative along line segment;
- If we have upper bound on 2nd derivative we can obtain arbitrarily tight bounds on p_{ij} at the cost of additional evaluations of $\hat{p}$ and its derivative.

6. Connection to single linkage clustering

Prop: We can construct the sample cluster tree by applying the recursive thresholding process to the maximal spanning tree of G instead of G itself.

Note:

- The maximal spanning tree of $(G, \{p_{ij}\})$ is the minimal spanning tree of $(G, \{1/p_{ij}\})$.
- We construct the sample cluster tree by recursively breaking "long" edges of the minimal spanning tree (edges with small p_{ij}).
- Applying this process to the Euclidean minimal spanning tree *is* single linkage clustering.

There is also a mathematical connection.

Prop: The sample cluster tree of the nearest neighbor density estimate

$$\hat{p}(\underline{x}) \sim \frac{1}{d(\underline{x}, \mathcal{X})}$$

is the single linkage dendogram.

6. Example

Objects: 572 olive oil samples coming from 9 different areas, grouped into 3 regions (1, 2, 3, 4) (5, 6) (7, 8, 9).

Features: Concentration of 8 different chemicals.

Question: How well can we recover the grouping into regions and areas?

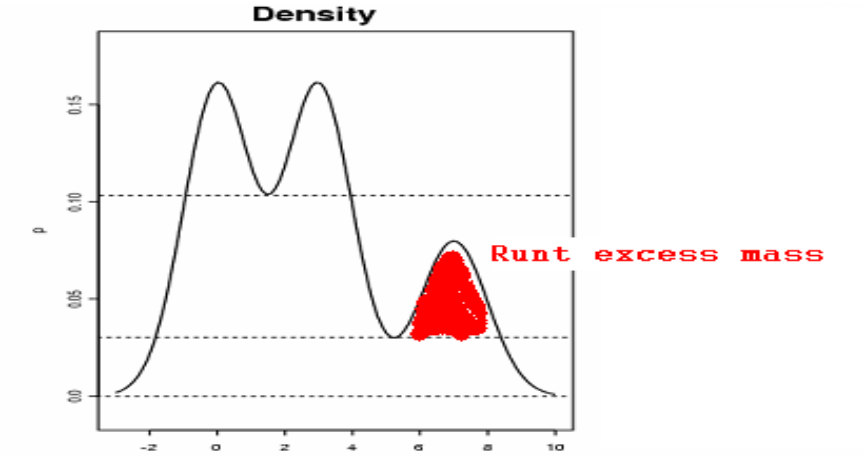
Note: To empirically evaluate performance of clustering methods, need labeled data.

Gaussian kernel density estimate; bandwidth determined by CV.

Problem: Sample cluster tree has dozens of leaves: many spurious modes.

Need diagnostic for assessing "significance" of modes.

For each split estimate *runt excess mass*



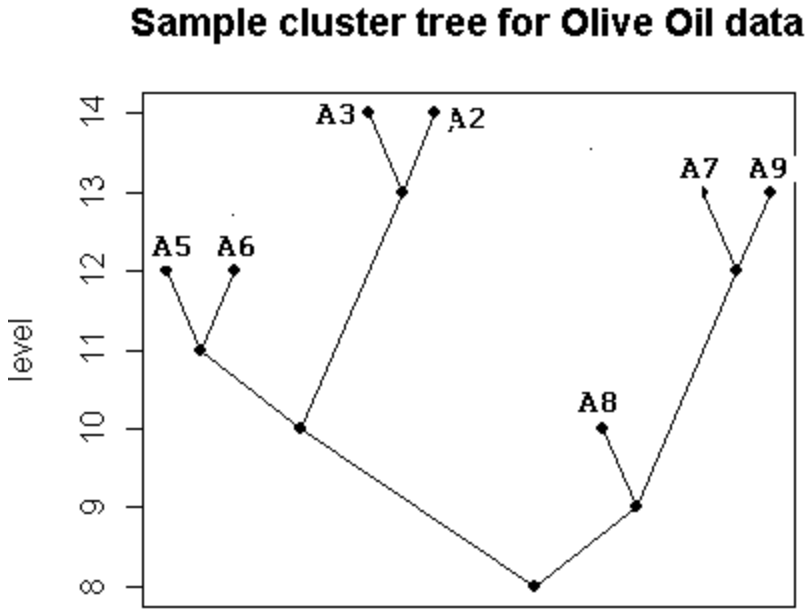
Excess mass for interior nodes of sample cluster tree, in decreasing order, expressed in number of observations:

162, 94, 43, 41, 40, 29, 12, 10, 8, 8, 7, 7 ,. . .

Gap between 29 and 12 $\Rightarrow$ prune all branches with mass ≤ 12.

Resulting tree has 7 leaves.

Sample cluster tree after pruning; leaves labelled with predominant area.



Interpretation:

- Bottom split separates region 3 from regions 1, 2.
- Next split on left separates region 1 from region 2
- Not able to identify areas 1 and 4

7. Summary

Goal of clustering: detect presence of distinct groups.

Premise of nonparametric clustering: groups $\sim$ modes of feature density.

Structure of collection of level sets is described by cluster tree; modes $\sim$ leaves.

Cluster tree is defined recursively — suggests recursive partitioning method for its computation.

For some density estimates, cluster tree can be computed exactly.

For others, cluster tree has to be approximated by solution of a graph problem $\Rightarrow$ Generalized single linkage clustering.

Generalized single linkage clustering can be applied to any density estimate.

Number of modes $\approx$ leaves of sample cluster tree depends on "bandwidth" of density estimate.

We have diagnostics to measure "size" of modes.

More principled approaches to pruning sample cluster tree using Bootstrap are under investigation.

Thanks for your interest.