John Hartigan's contributions to clustering

(and some extensions)

Werner Stuetzle
Department of Statistics
University of Washington

1. Introduction

What's the goal of "cluster analysis"?

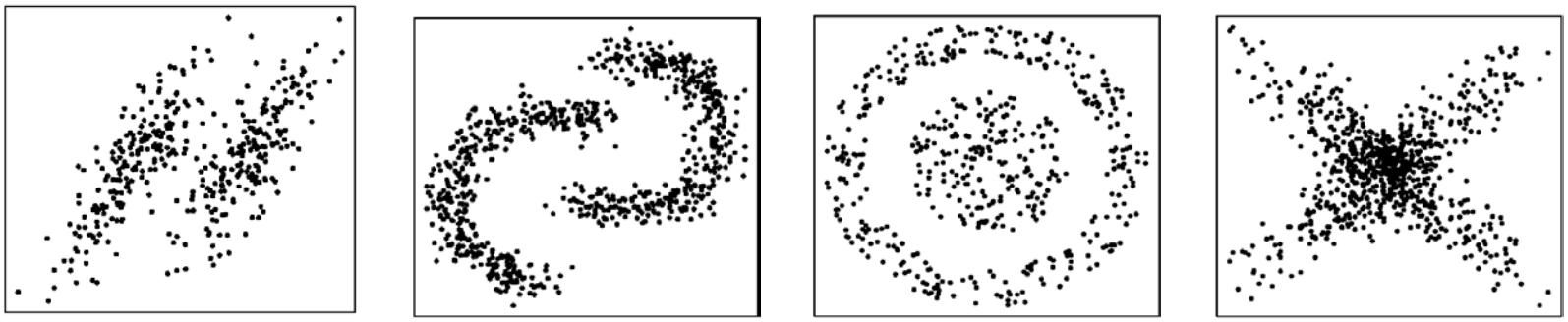
Detect presence of distinct groups in a data set

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)

Essence of JH's contribution:

Cast cluster analysis as a statistical problem.

- Data are regarded as a sample from some underlying population;
- There is a definition of "distinct groups" for the population density $p(\underline{x})$;
- The goal of clustering methods is to estimate the groups in the population from the sample.

Important because:

- Without sampling model, the concept of "cluster validity" does not make sense;
- Without well specified population characteristic it is hard to evaluate and compare clustering methods $\Rightarrow$ no "progress".

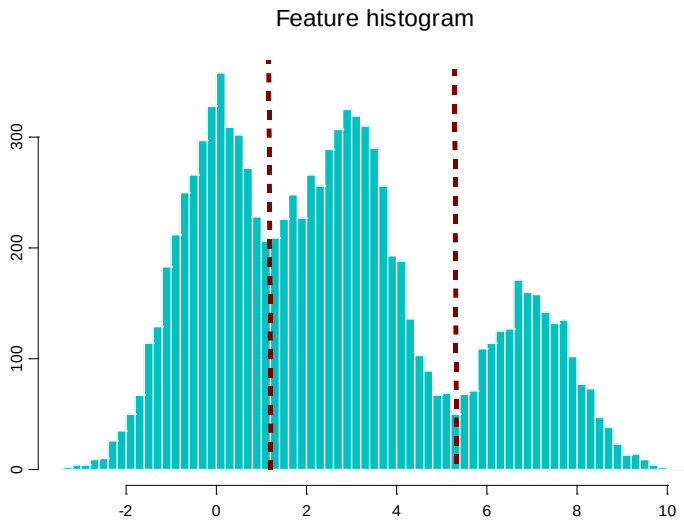
Two basic approaches, parametric (model based) and nonparametric.

Nonparametric clustering (focus of JH's work)

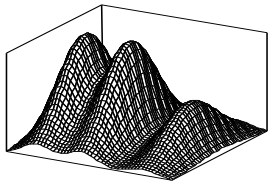
Based on premise that groups correspond to modes of density $p(\underline{x})$.

Need to

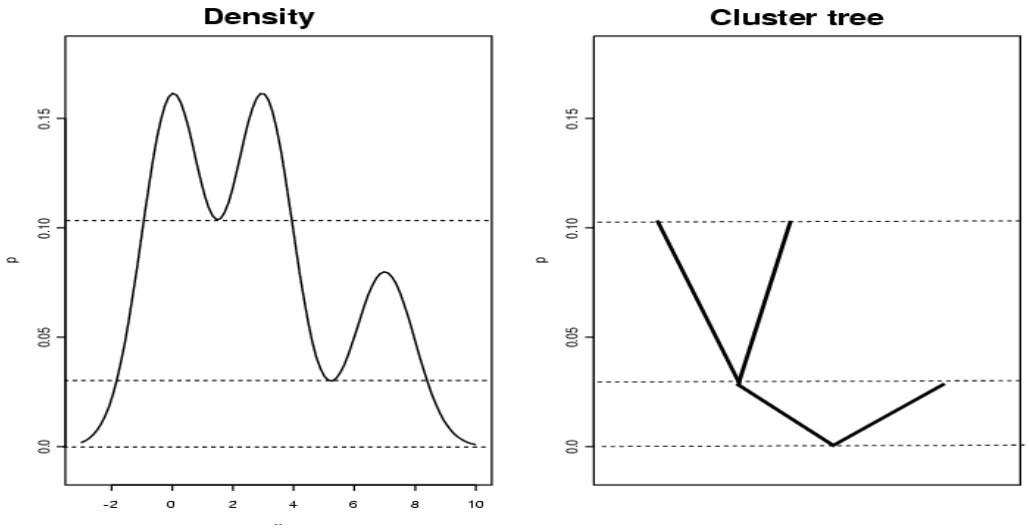
- Estimate modes;
- Assign each observation to the "domain of attraction" of a mode.



Population characteristic of interest: Cluster tree.



Structure of level sets is described by cluster tree



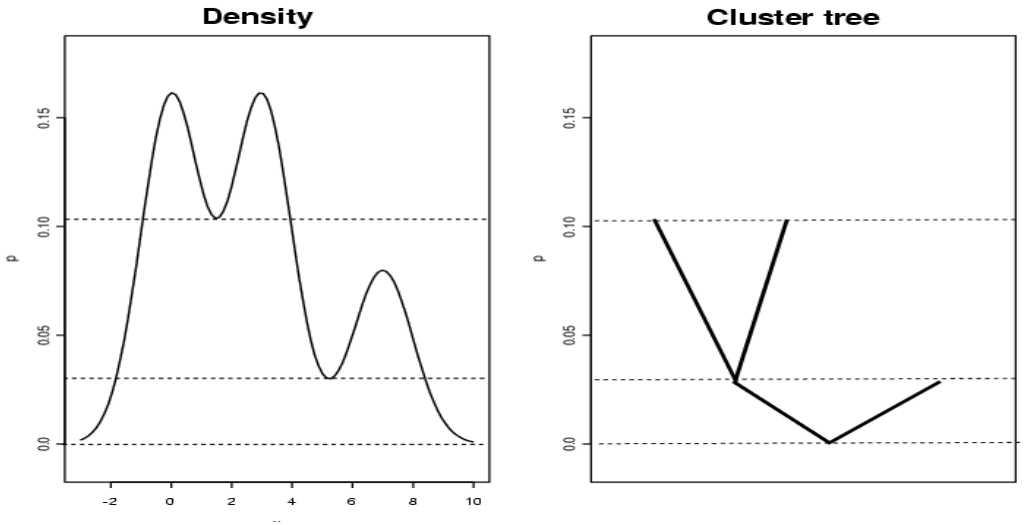
Cluster tree is easiest to define recursively:

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

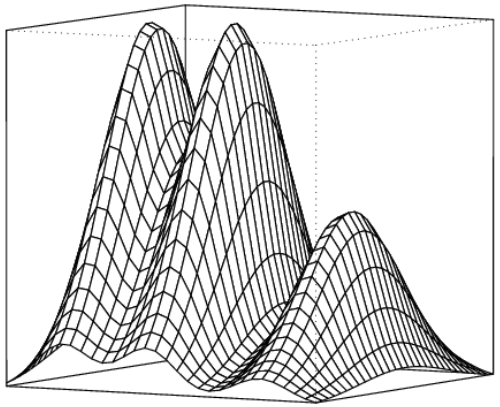
- represents the entire support of the density;
- is associated with density level $\lambda(N) = 0$.

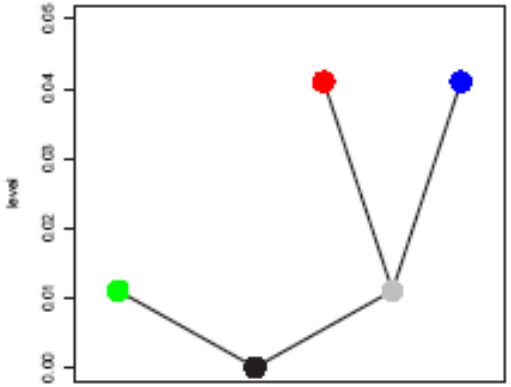


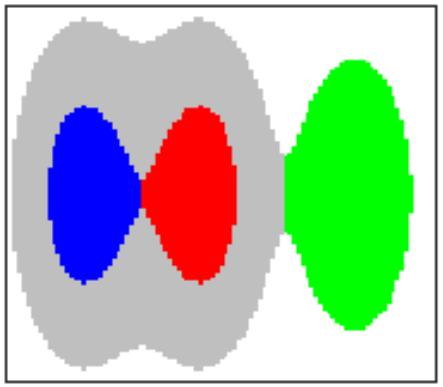
To determine descendents of node N:

- 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 target population characteristic in nonparametric clustering.

3. Single linkage clustering

Single linkage clustering is one of many hierarchical clustering methods.

General structure of hierarchical clustering methods:

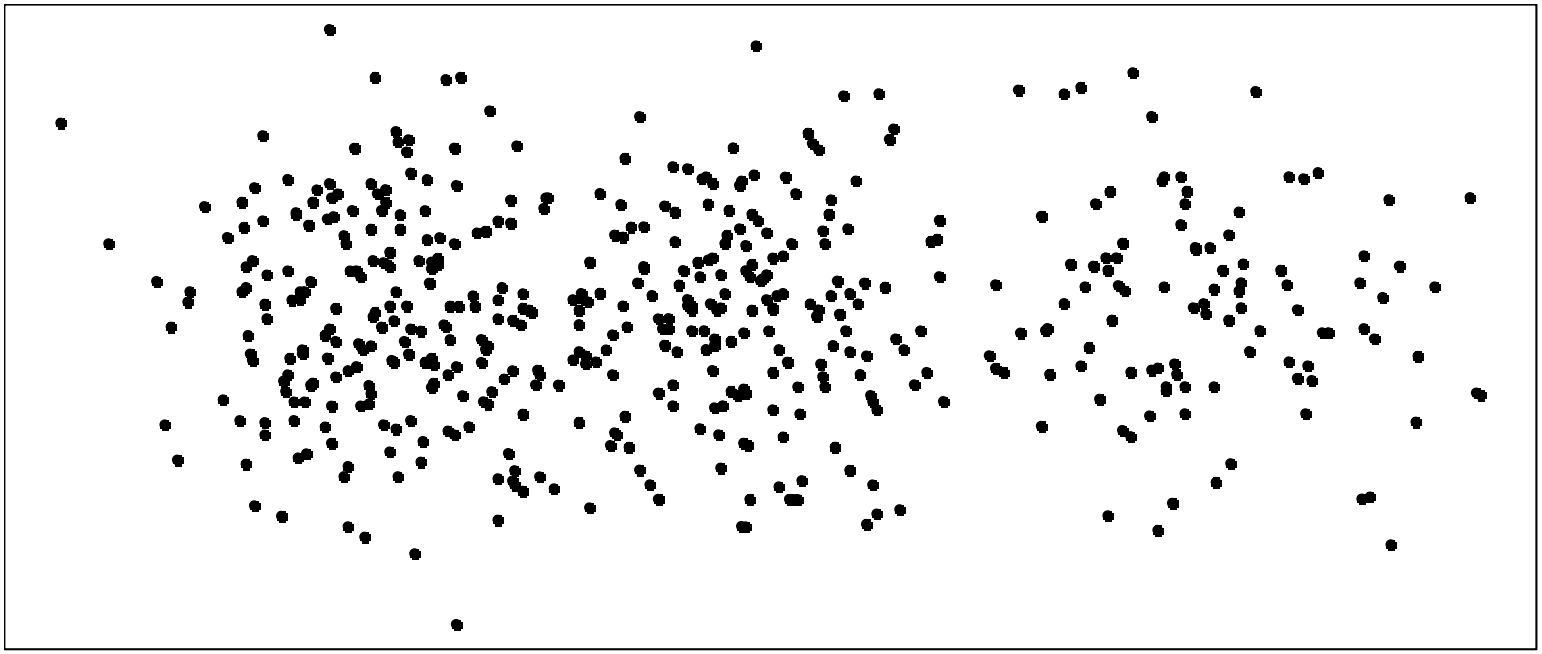
- Define a measure of distance between sets of observations ("fragments").
- Initially, every observation forms its own fragment.
- Progressively merge the two closest fragments, until only one fragment is left.

Single linkage clustering: Define

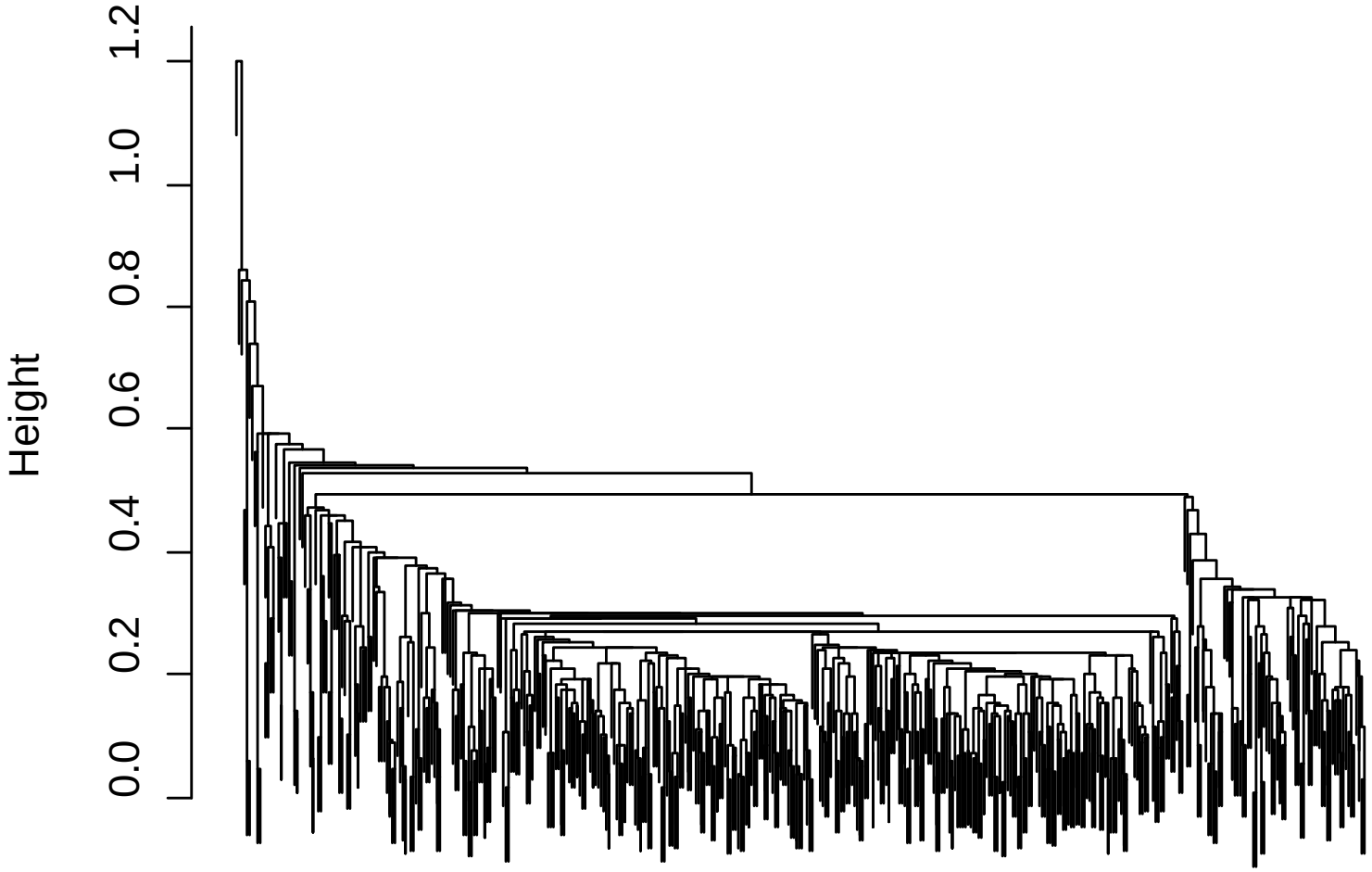
$$d(F_1, F_2) = \min_{\underline{x}_i \in F_1} \min_{\underline{x}_j \in F_2} d(\underline{x}_i, \underline{x}_j) \,.$$

Result of merge process can be represented by dendogram.

Sample from 3 component Gaussian mixture



Single linkage dendogram for sample



Extracting clusters from a dendogram

Standard procedure: dendogram cutting.

- Choose distance threshold ρ (critical step).
- Remove all dendogram nodes at height $> \rho$ and their incident edges.
- The leaves of the resulting subtrees are the clusters.

Alternative description: Stop merging once distance smallest distance between fragments is $> \rho$.

Optimal number of clusters (as measured by adjusted rand index): 93.

Cluster 1 captures 137 (of 200) observations form group 1.
Cluster 6 captures 119 (of 200) observations from group 2.
No cluster captures more than 21 (of 100) observations from group 3.

Dendogram cutting performs very poorly. Why?

4. Single linkage clustering and nearest neighbor density estimation (JH)

The nearest neighbor density estimate based on a sample $\underline{x}_1, \ldots, \underline{x}_n$ is

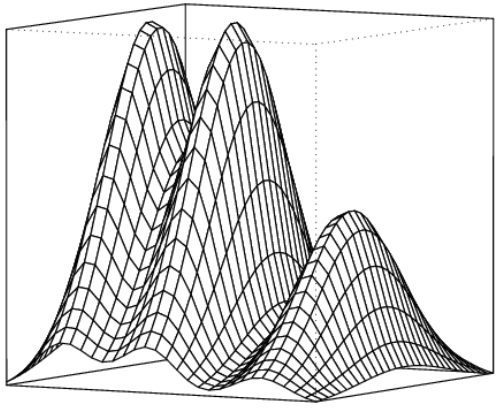
$$\hat{p}(\underline{x}) = 1/\min d(\underline{x}, \underline{x}_i)\,.$$

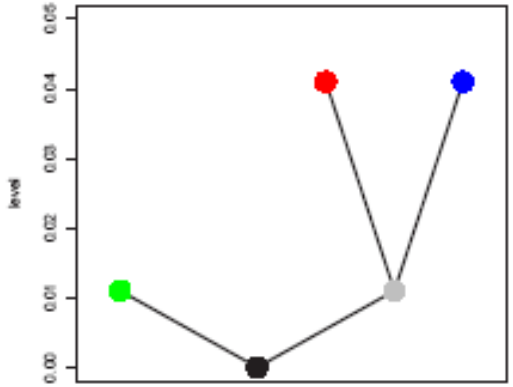
Not a particularly appealing density estimate, but hold on.

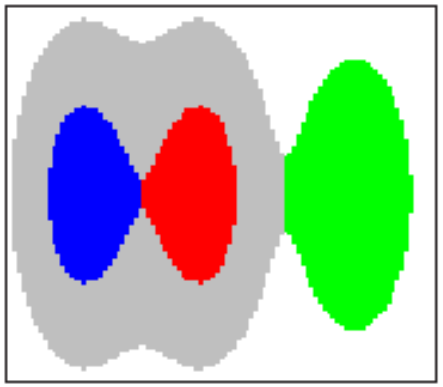
JH: The clusters obtained by cutting the single linkage dendogram at level ρ are the connected components of the level set $L(1/2\rho; \hat{p})$.

Cor: The single linkage dendogram is isomorphic to the cluster tree of the nearest neighbor density estimate.

This connection explains why dendogram cutting is not a viable method for extracting clusters from a single linkage dendogram.







Think of the left-most frame as a (highly idealized) picture of the nearest neighbor density estimate.

Low cut level separates group 3 from 1 and 2, but doesn't separate 1 from 2.

High cut level separates group 1 from 2, but misses group 3.

Need better way of pruning the cluster tree of the nearest neighbor density estimate.

5. The Runt test for unimodality (JH, Mohanty)

Single linkage merge process starts, and fragments grow, in high density regions, around modes (see plots on following slides).

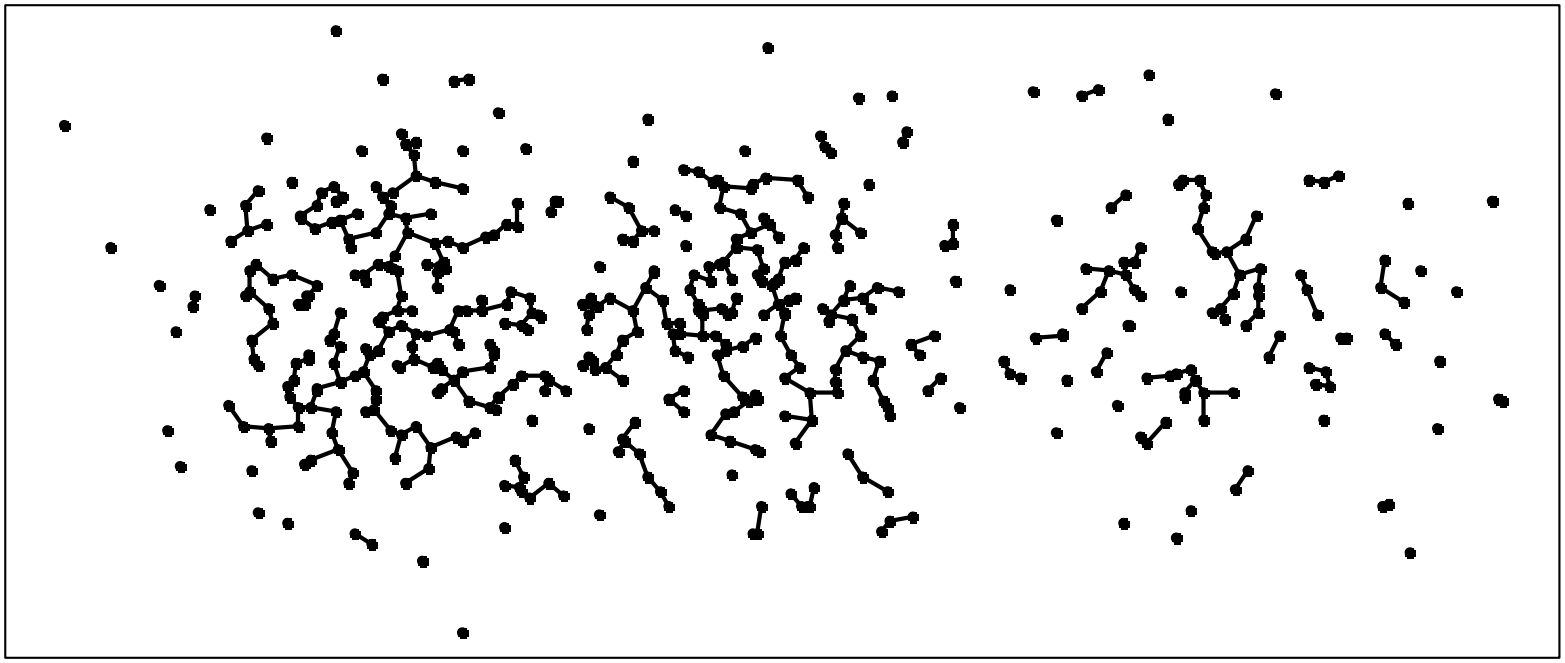
Eventually, the fragments growing around the various modes will be joined $\Rightarrow$ dendogram nodes rooting two large subtrees.

Runt size of dendogram node = smaller of the number of leaves of the two subtrees rooted at the node (JH, Mohanty)

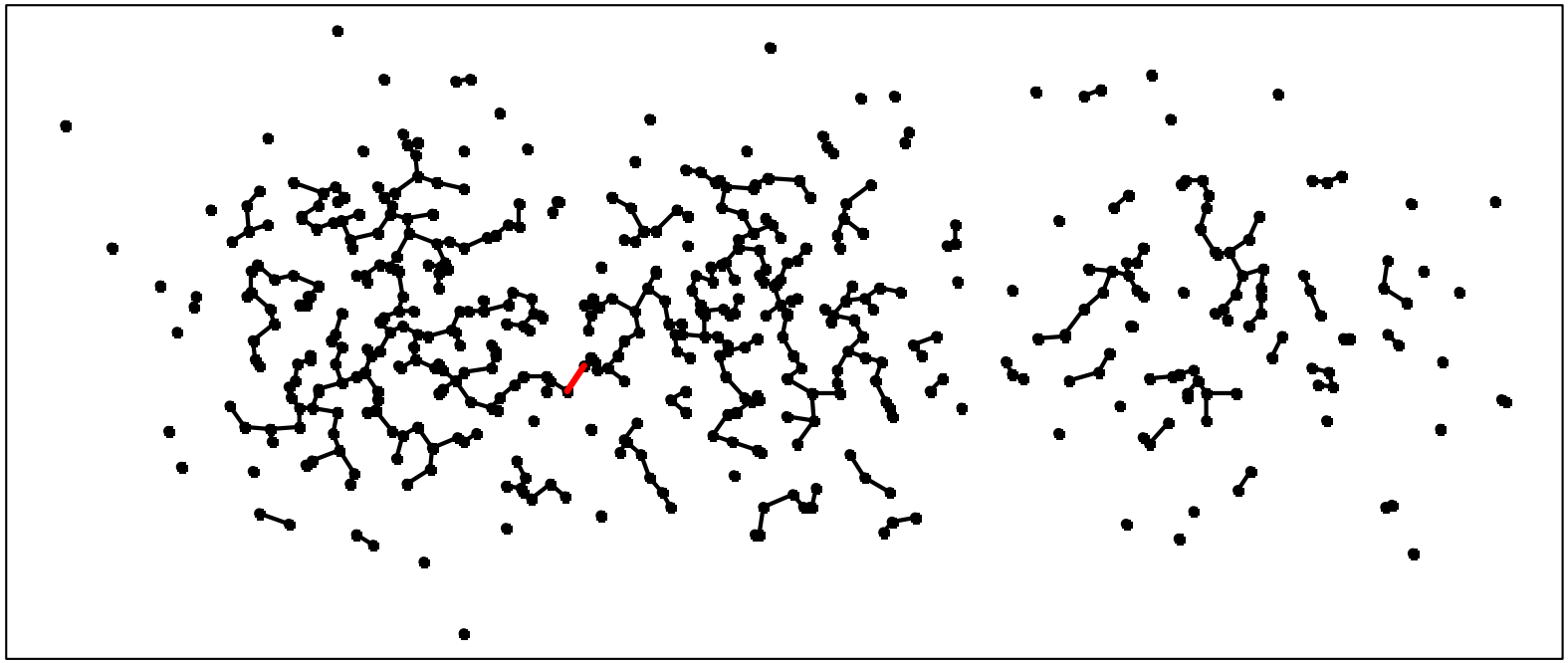
JH, Mohanty used maximum runt size as test statistic for testing unimodality.

Samples from unimodal distributions tend to have small maximum runt size.

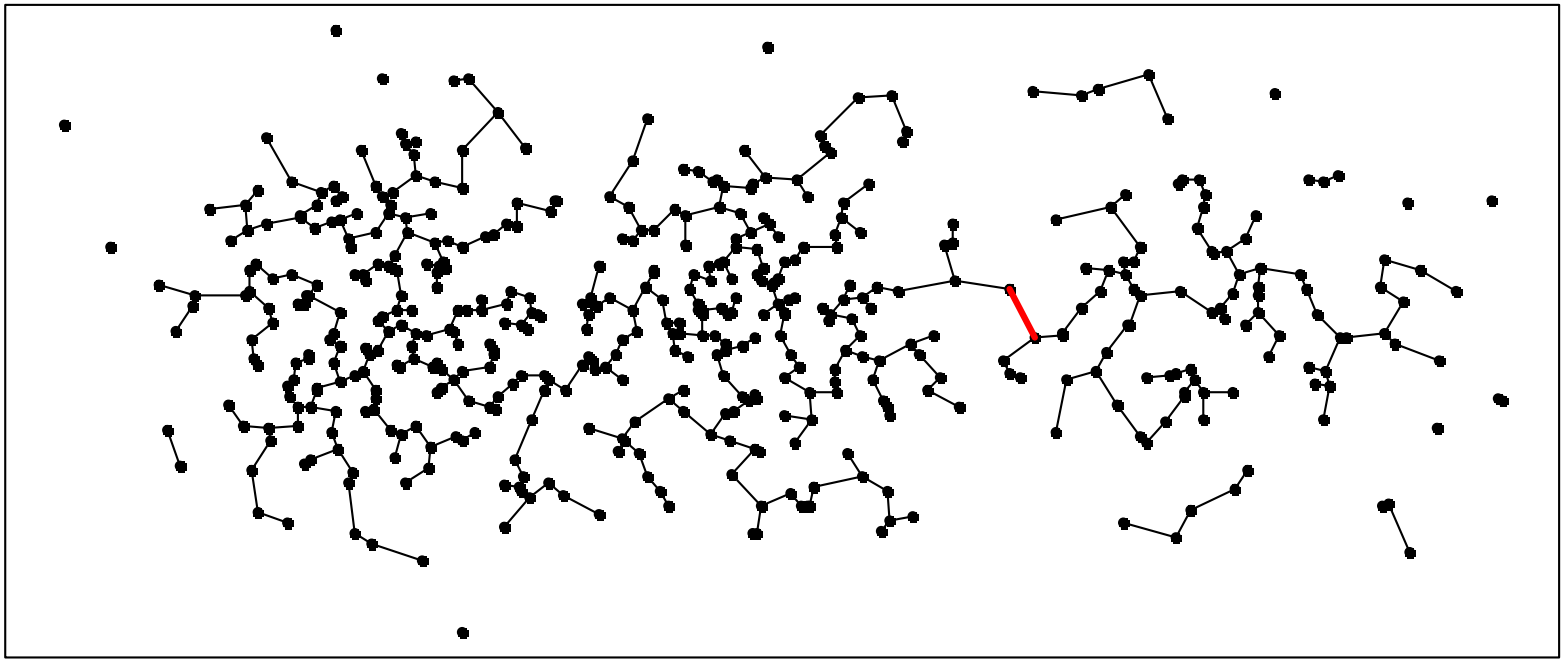
First 400 merges



First merge of large fragments; runt size = 13



Second merge of large fragments; runt size :



6. Runt pruning

Obvious approach to estimating the cluster tree of a density:

- Estimate the underlying density p by a density estimate $\hat{p}$.
- Estimate the cluster tree of p by the cluster tree of $\hat{p}$.

Cluster tree of nearest neighbor density estimate easy to compute.

However, nearest neighbor density estimate is noisy and has singularity at every observation.

$\Rightarrow$ Cluster tree of nearest neighbor density estimate is poor estimate for cluster tree of underlying density.

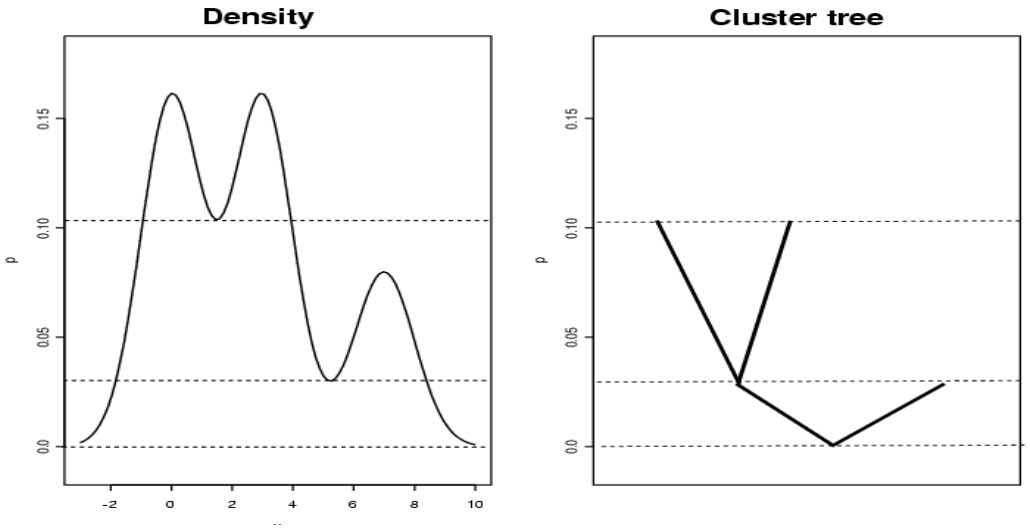
It needs to be pruned.

Standard pruning method: tree cutting (bad)

Better approach: runt pruning.

Choose runt size threshold r.

Sllightly modify recursive definition of cluster tree.



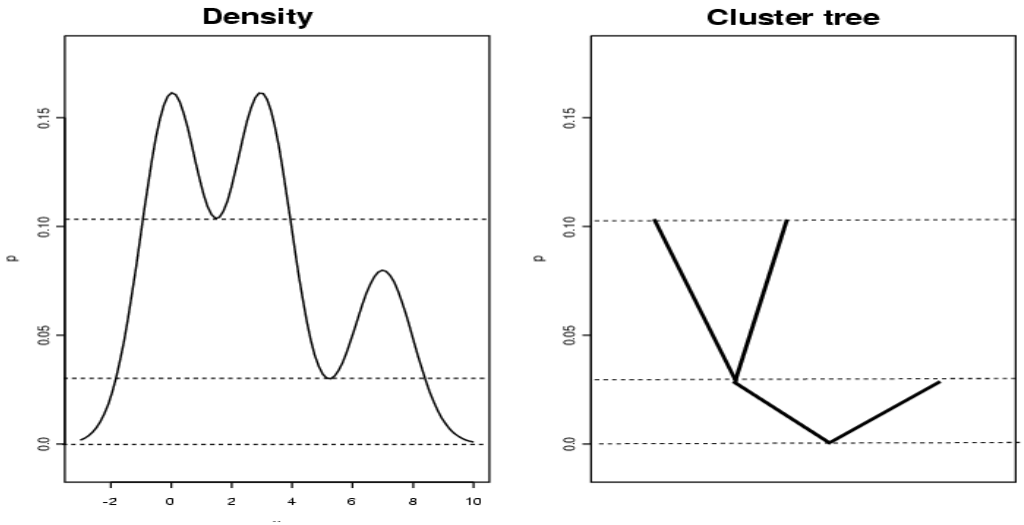
Cluster tree is easiest to define recursively:

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:

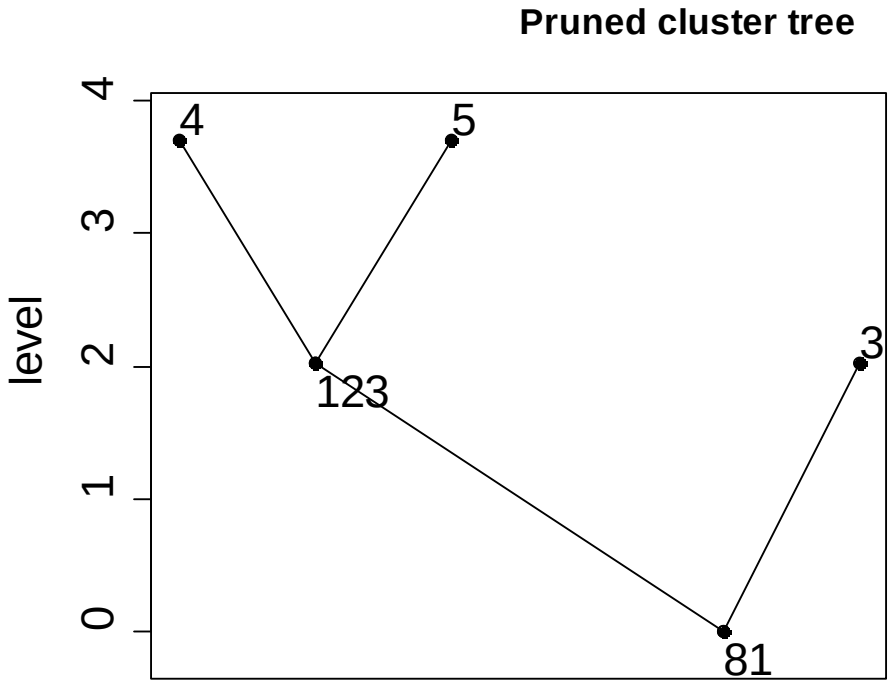
- Find lowest level λ_d for which $L(\lambda; p) \cap D(N)$ has two connected components with size (number of observations) $\geq r$.
- If there is no such λ_d then N is leaf of the tree.
- Otherwise, create daughter nodes representing the connected components, with associated level λ_d, and recurse.

Application to our 2-d example

Runt sizes of cluster tree nodes (in decreasing order):
123, 81, 30, 22, 22, 21, 17, 16, 14, 13, ...

Two splits with large runt size $\Rightarrow$ suggests three groups.

Choose runt size threshold $r = 81$.



	cluster.id		
group.id	3	4	5
1	0	191	9
2	6	18	176
3	96	0	4

Almost perfect!

7. Summary

Clustering by runt pruning the cluster tree of the nearest neighbor density estimate builds on three of JH's contributions:

- The cluster tree as the population quantity of interest.
- Connection between single linkage clustering and nearest neighbor density estimation.
- Large runt sizes as an indicator for multimodality.

Message:

The problem with single linkage clustering is not reliance on the nearest neighbor density estimate.

The problem is the use of tree cutting to prune the cluster tree of the density estimate.

Runt pruning is vastly superior and is hard to beat.

Further work not covered today:

- Generalize plug-in approach to other density estimates
  $\Rightarrow$ Generalized Single Linkage Clustering (RN, WS).
- Clustering with Confidence (RN, WS)

Thank you for your attention