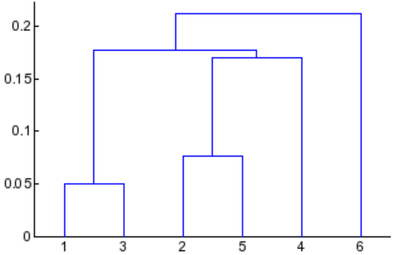
STATS 415: Hierarchical Clustering

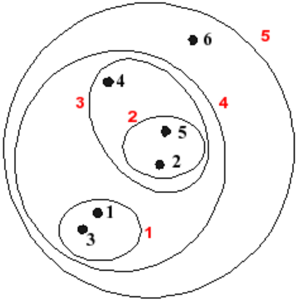
Prof. Liza Levina

Department of Statistics, University of Michigan

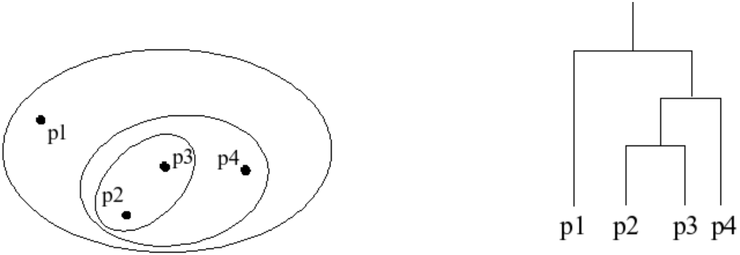
Hierarchical Clustering

- A set of "nested" clusters organized as a hierarchical tree.
- Can be visualized as a dendrogram – a tree-like diagram that records the sequences of merges or splits.

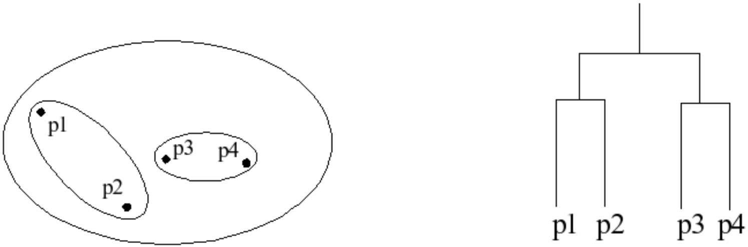




Examples



Traditional Hierarchical Clustering

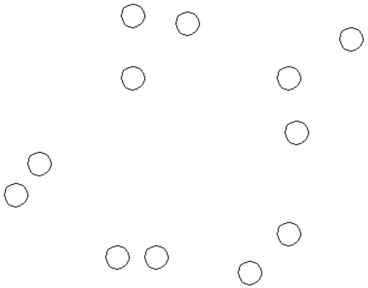


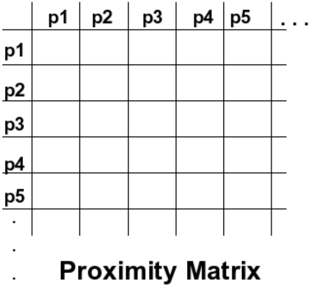
Two types of hierarchical clustering

- Agglomerative
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or K clusters) left
- Divisive
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are K clusters)

Start

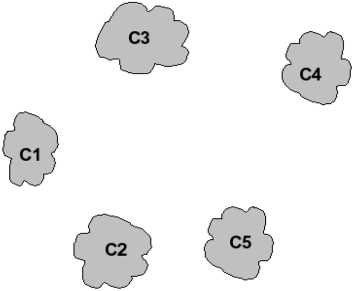
Start with clusters of individual points and a dissimilarity matrix





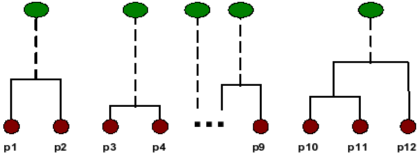
Intermediate

After some merging steps, we have clusters:



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

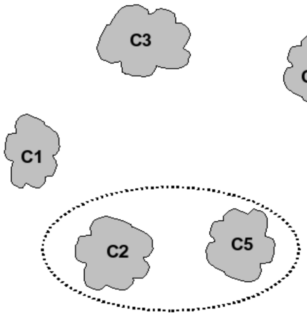
Proximity Matrix

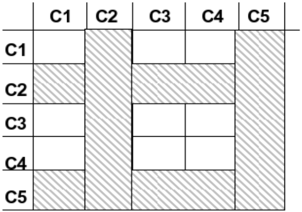


Take the next step

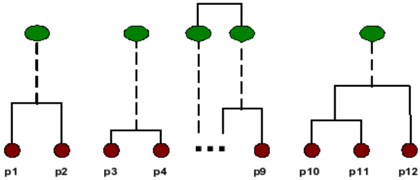
We want to merge two closest clusters:

- How do we determine which are closest?





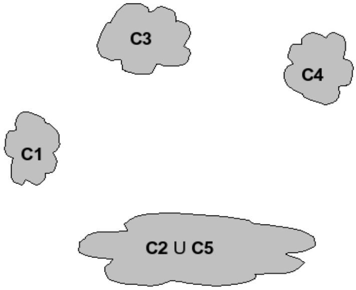
Proximity Matrix



After Merging

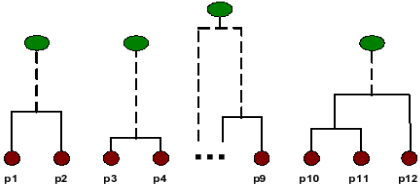
Let's say we merged C2 and C5.

- How do we update the dissimilarity matrix?

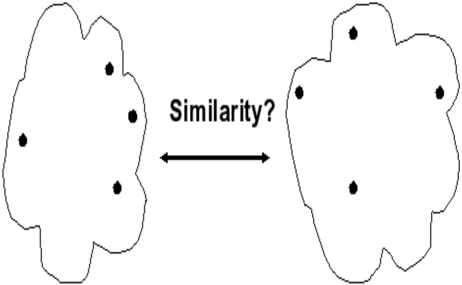


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Answering both questions requires computing dissimilarity between clusters, not just pairs of points.



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						

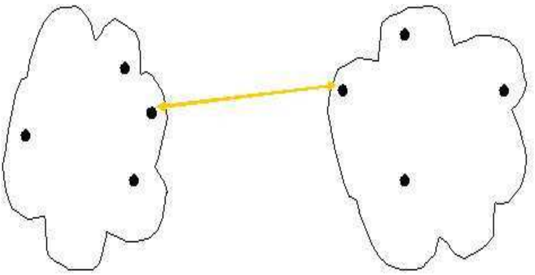
Cluster dissimilarity measures

- Single linkage (min)
- Complete linkage (max)
- Average linkage
- Distance between centroids
- Other methods driven by various loss functions

Single (min) linkage

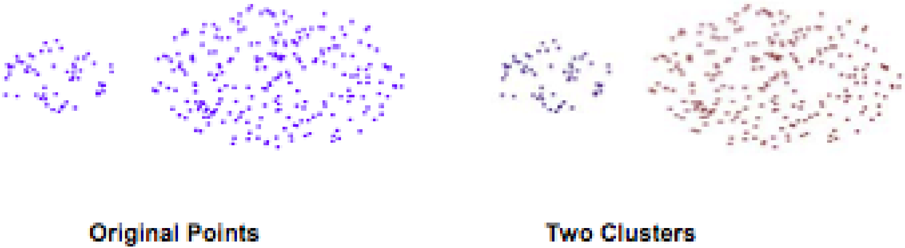
Distance between two nearest points:

$$d(C_1, C_2) = \min_{s,t}\{d(s,t) : s \in C_1,\ t \in C_2\}$$



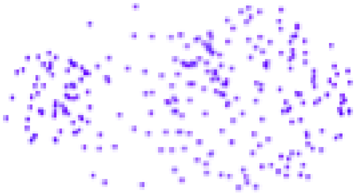
Strengths of min linkage

Can handle diverse shapes/sizes

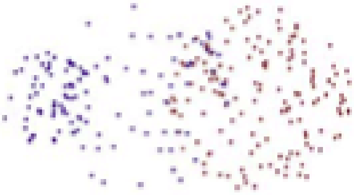


Weaknesses of min linkage

Sensitive to noise and outliers



Original Points

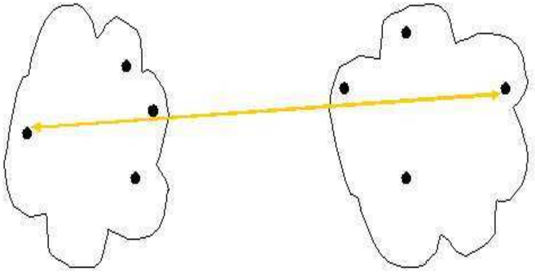


Two Clusters

Complete (max) linkage

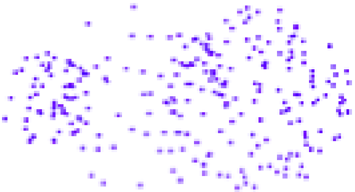
Distance between the farthest points

$$d(C_1, C_2) = \max_{s,t}\{d(s,t) : s \in C_1,\ t \in C_2\}$$

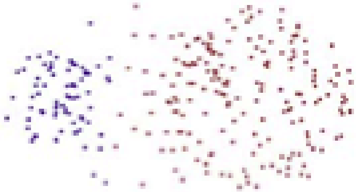


Strengths of max linkage

Robust to noise and outliers



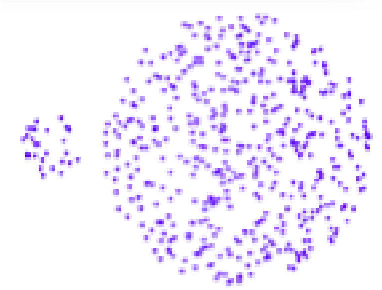
Original Points



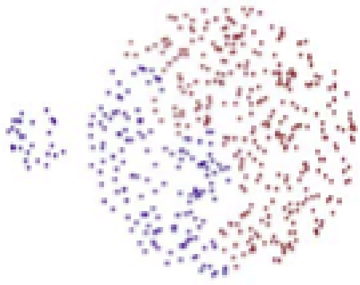
Two Clusters

Weaknesses of max linkage

- Tendency towards breaking large clusters
- Preference for spherical clusters



Original Points

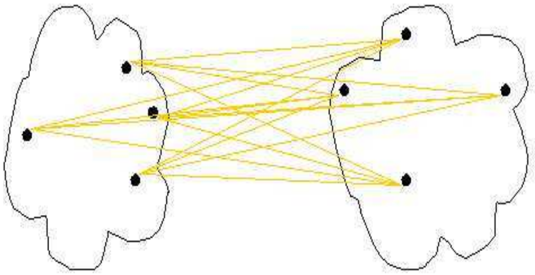


Two Clusters

Average linkage

Average of distances between all pairs of points

$$d(C_1, C_2) = \frac{\sum_{s \in C_1} \sum_{t \in C_2} d(s,t)}{n_1 n_2}$$



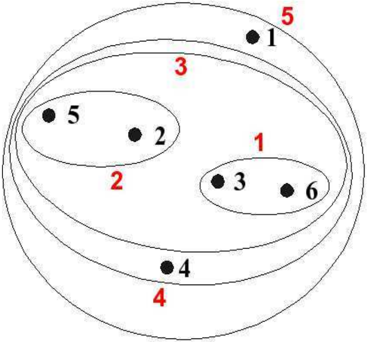
Properties of average linkage clustering

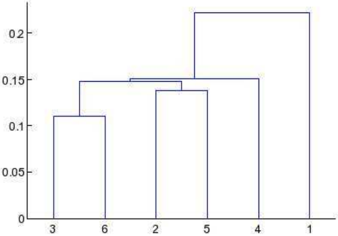
- Compromise between min and max linkage clustering
- Less susceptible to noise and outliers than min linkage; but more so than max linkage
- Still has a preference for spherical clusters, but less so than max linkage
- In practice often ends up similar to one of them

Toy example: dissimilarity matrix

	P1	P2	P3	P4	P5	P6
P1	0.000	0.234	0.216	0.368	0.342	0.235
P2	0.234	0.000	0.145	0.194	0.143	0.243
P3	0.216	0.145	0.000	0.158	0.285	0.102
P4	0.368	0.194	0.158	0.000	0.284	0.220
P5	0.342	0.143	0.285	0.284	0.000	0.386
P6	0.235	0.243	0.102	0.220	0.386	0.000

Toy example: min linkage

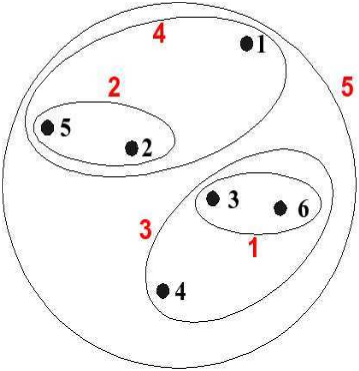


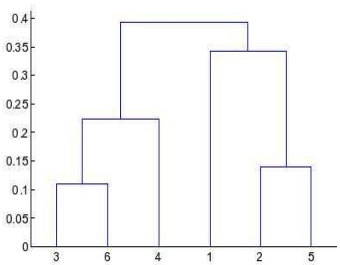


Toy example: dissimilarity matrix

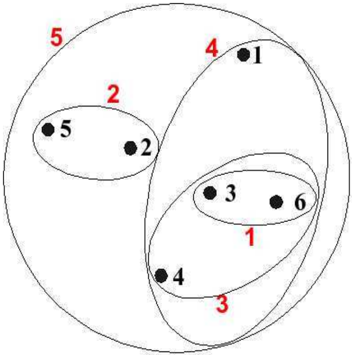
	P1	P2	P3	P4	P5	P6
P1	0.000	0.234	0.216	0.368	0.342	0.235
P2	0.234	0.000	0.145	0.194	0.143	0.243
P3	0.216	0.145	0.000	0.158	0.285	0.102
P4	0.368	0.194	0.158	0.000	0.284	0.220
P5	0.342	0.143	0.285	0.284	0.000	0.386
P6	0.235	0.243	0.102	0.220	0.386	0.000

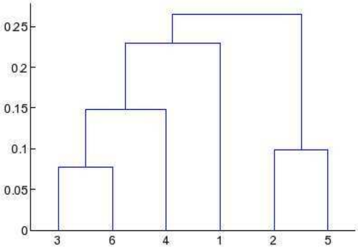
Toy example: complete linkage





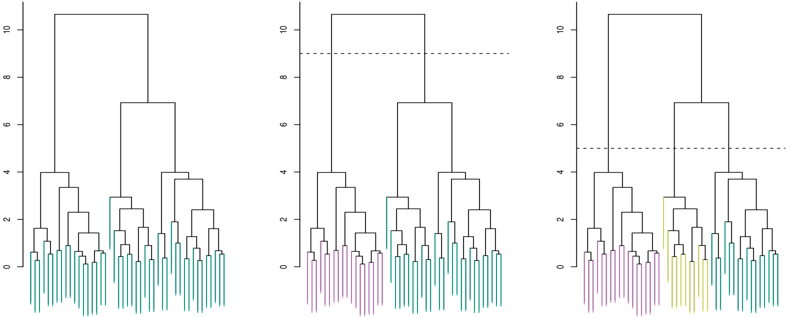
Toy example: average linkage





Choosing Clusters

- To choose clusters we draw lines across the dendrogram
- We can form any number of clusters depending on where we draw the line, from $K = 1$ to $K = n$

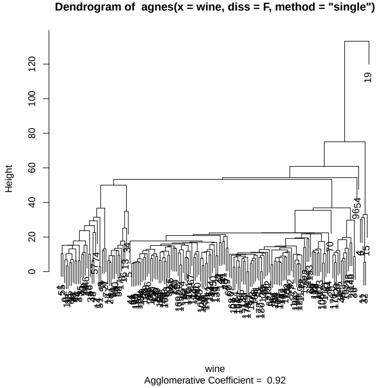


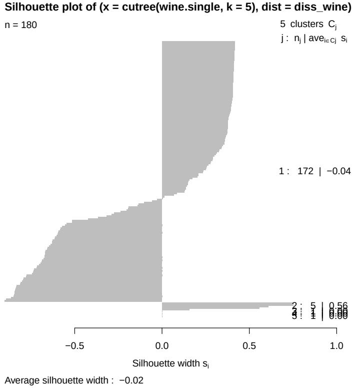
Example: the wine data

180 wines, 13 variables

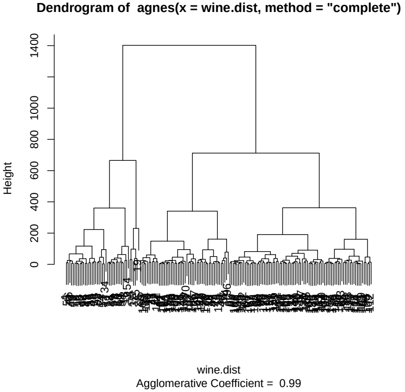
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

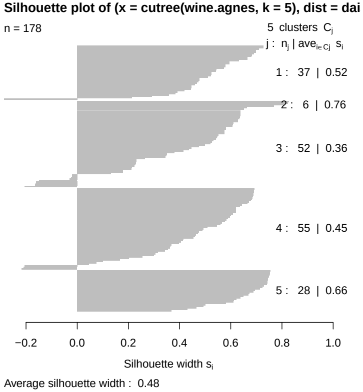
Wine data: min linkage clustering



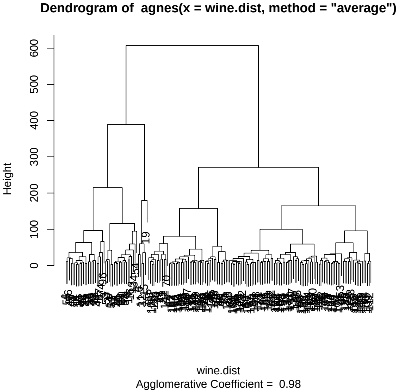


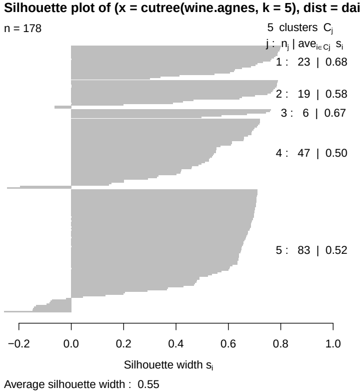
Wine data: max linkage clustering





Wine data: average linkage clustering





Wine data example summary

- All methods but single linkage are similar; this is not uncommon
- The silhouette plot tends to give better visual assessment than the dendrogram
- The agglomerative coefficient tends to be not as informative as the silhouette width

Advantages of hierarchical clustering

- Gives a family of possible solutions; any number of clusters can be obtained by “cutting” the dendrogram at the appropriate level
- Computationally fast
- Does not require raw data, only distances (or dissimilarities) between points

Disadvantages of hierarchical clustering

- No global optimization criterion, greedy algorithm
- No objective way to choose where to stop ("cut" the tree)
- Different ways of measuring distance between clusters can give rise to very different solutions