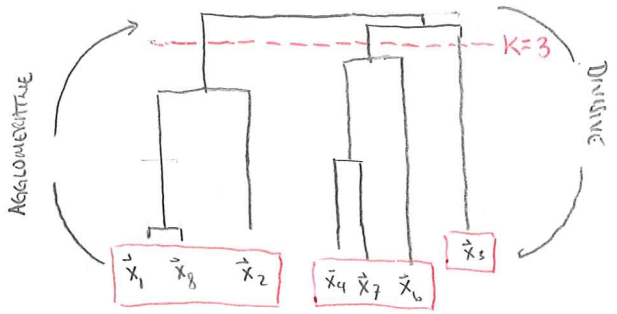
Hierarchical Clustering

Setting: N observations $\vec{x}_1, \dots \vec{x}_N \in \mathbb{R}^d$

Goal: Find clusters (groups/subsets) of points which are similar

Output: Dendrogram (Hierarchical Tree Diagram)



- **Agglomerative:** k - cluster solution generated by merging two clusters from $k+1$ cluster solution (bottom up)
- **Divisive:** $(k+1)$ - cluster solution generated by splitting one cluster from k - cluster solution (top down)
- Given a dendrogram, we can draw a horizontal line crossing K vertical lines to generate K clusters of data (branches)
  - Items which are similar are combined at low heights
    - Differences in heights define how close items are together
      - $\vec{x}_1, \vec{x}_8$ are more similar than $\vec{x}_4, \vec{x}_7$
      - $\vec{x}_1, \vec{x}_8$ cluster is more similar to $\vec{x}_2$ than $\vec{x}_4, \vec{x}_7$ is to $\vec{x}_6$
  - Horizontal distances are irrelevant

Dissimilarities / Distance

Clusters are merged / split based on dissimilarities of items in cluster

- For $\vec{x}_i, \vec{x}_j \in \mathbb{R}^d$ distance is a natural choice of dissimilarity
  - Minkowski: $\|x_i - x_j\|_p = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^p \right)^{1/p}$ $p \geq 1$
    - $p=1$: Manhattan
    - $p=2$: Euclidean
  - 1 - Correlation

$$d(x_i, x_j) = 1 - \rho_{ij} = 1 - \frac{s_{ij}}{s_i s_j}$$

(s_{ij}: sample covariance; ρ_{ij}: sample correlation of observations of x_i, x_j; s_i, s_j: sample std. dev.)

* Distance measures computed over variables, not observations

- Choice of distance / dissimilarity gives rise to distance matrix $D \in \mathbb{R}^{N \times N}$
  - How do we define distances between clusters

Agglomerative Nesting

1) Input: Distance / dissimilarity matrix D. Initially have N "clusters"
2) Find smallest dissimilarity, say D_{IJ} in $D = D^{(1)}$. Merge I, J into a new cluster IJ
3) Compute dissimilarities between IJ and all other clusters $K \neq IJ$

   Single linkage: $d_{IJ,K} = \min_{s \in IJ, t \in K} D_{st}$

   Complete: $d_{IJ,K} = \max_{s \in IJ} \max_{t \in K} D_{st}$

   Average: $d_{IJ,K} = \frac{1}{N_{IJ} N_K} \sum_{s \in IJ} \sum_{t \in K} D_{st}$

4) Form a new $(N-1) \times (N-1)$ distance matrix $D^{(2)}$ by removing rows and columns I and J, and adding new row / column for IJ
5) Repeat 2-4 $(N-1)$ times
6) Output: list of merged clusters at each step w/ value (height) of dissimilarity at merge

EXAMPLE:

$x_1 = (1,3)^T$

$x_2 = (2,4)^T$

$x_3 = (1,5)^T$

$x_4 = (5,5)^T$

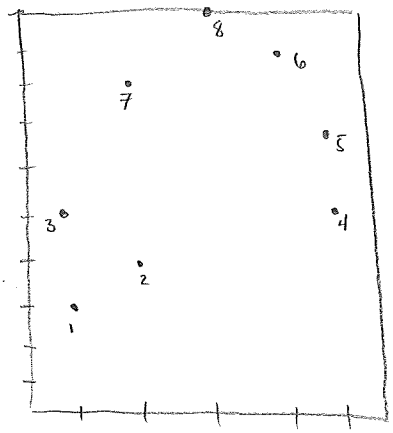
$x_5 = (5,7)^T$

$x_6 = (4,9)^T$

$x_7 = (2,8)^T$

$x_8 = (3,10)$

$$D^{(1)} = \begin{bmatrix} 0 & 1.4 & 2 & 4.8 & 5.7 & 6.7 & 5.1 & 7.3 \\ & 0 & 1.4 & 3.2 & 4.2 & 5.4 & 4 & 6.1 \\ & & 0 & 4 & 4.5 & 5 & 3.2 & 5.4 \\ & & & 0 & 2 & 4.1 & 4.2 & 5.4 \\ & & & & 0 & 2.2 & 3.2 & 3.6 \\ & & & & & 0 & 2.2 & 1.4 \\ & & & & & & 0 & 2.2 \\ & & & & & & & 0 \end{bmatrix}$$



AVG. LINKAGE (PRIORITIZES AVERAGE DISTANCE BETWEEN CLUSTERS)

$$D^{(2)} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccc} 12 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 0 & 1.7 & 3.8 & 5.0 & 6.0 & 4.6 & 6.7 \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix} \end{array}$$

MERGED 12 DISTANCE = 1.4

$$D^{(3)} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \\ 5 \\ 68 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccc} 12 & 3 & 4 & 5 & 68 & 7 \end{array} \\ \begin{bmatrix} 0 & & & & 5.2 & \\ & 0 & & & 4.8 & \\ & & 0 & & 3.0 & \\ & & & 0 & & \\ & & & & 0 & 2.2 \\ & & & & & 0 \end{bmatrix} \end{array}$$

MERGED 68 DISTANCE = 1.4

$D^{(4)}$ =

	123	4	5	68	7
123	0	3.9	4.8	6.0	4.1
4					
5					
68					
7					

MERGED 12, 3 DISTANCE 1.7

$D^{(5)}$ =

	123	45	68	7
123			6.0	
45		0	3.8	3.7
68				
7				

MERGED 4,5 DISTANCE 2

$D^{(6)}$ =

	123	45	678
123	0	4.3	5.3
45		0	3,8
678			0

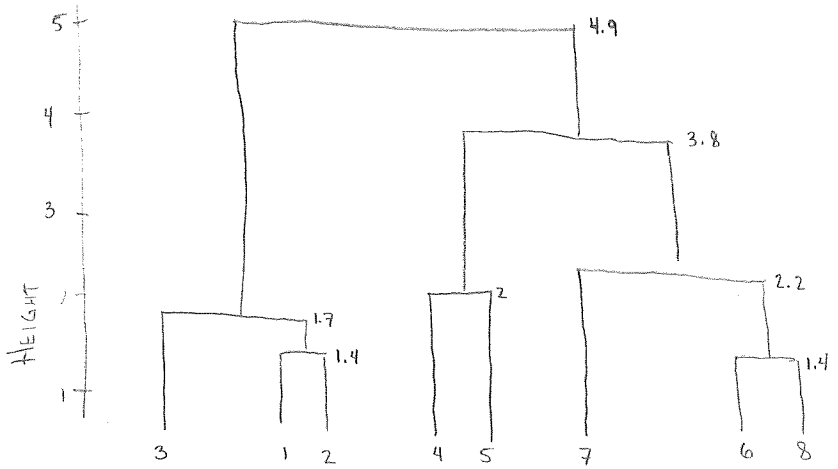
MERGED 68,7 DISTANCE 2.2

$D^{(7)}$ =

	123	45,678
123	0	4.94
45,678		0

MERGED 45, 678 DISTANCE 3.8

MERGE 123, 45678 DISTANCE

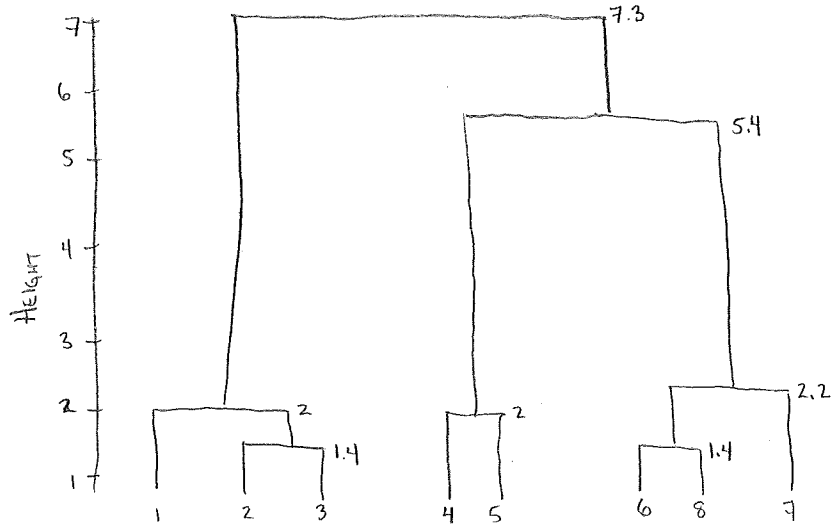


COMPLETE (REQUIRES ALL ELEMENTS BETWEEN CLUSTERS TO BE CLOSE

$D^{(1)} =$

0	~~1~~	2	~~4.5~~	~~5.8~~	~~6.7~~	~~5.1~~	7.3
	0	1.4	~~3.2~~	~~4.2~~	~~5.4~~	~~4~~	~~4.1~~
		0	~~4~~	~~4.5~~	~~5~~	~~3.2~~	~~5.4~~
			0	2	~~4.1~~	~~4.2~~	5.4
				0	~~3.2~~	~~3.2~~	~~3.6~~
					0	2.2	1.4
						0	~~2.2~~
							0

- MERGE 23 AT HEIGHT 1.4
- MERGE 68 AT HEIGHT 1.4
- MERGE 1, 23 AT HEIGHT 2
- MERGE 45 AT HEIGHT 2
- MERGE 68, 7 AT HEIGHT 2.2
- MERGE 45, 678 AT HEIGHT 5.4
- MERGE 123, 45678 AT HEIGHT 7.3

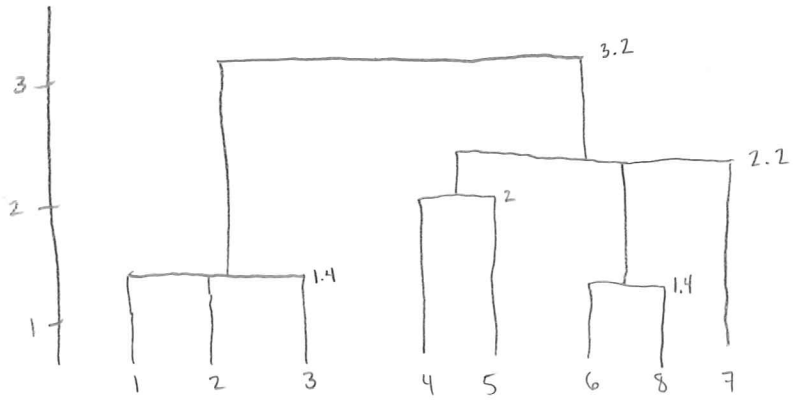


SINGLE LINKAGE (ACCEPTS LOCAL, SMALL CONNECTIONS)

$D^{(1)} =$

0	1.4	2	4.8	5.7	6.7	5.1	7.3
	0	1.4	3.2	4.2	5.4	4	6.1
		0	4	4.5	5	3.2	5.4
			0	2	4.1	4.2	5.4
				0	2.2	3.2	3.6
					0	2.2	1.4
						0	7.2
							0

- MERGE 23 AT 1.4
- MERGE 1,23 AT 1.4
- MERGE 68 AT 1.4
- MERGE 45 AT 2
- MERGE 68,7 AT 2.2
- MERGE 45,678 AT 2.2
- MERGE 123, 45678 AT 3.2



Observations

Single	Average	Complete
Joins if ANY pair of points between clusters are close	Intermediate	Joins if ALL pairs of points between clusters are close
Shorter heights		Taller heights
Long chains of clusters joined by nearby singletons		Many, small compact clusters
Ind. of size of clusters	Depends on size of clusters	Ind. of size of clusters
Invariant under monotone transformations of pairwise distances (structure not height)	May be altered by nonlinear transformations to pairwise dist	Invariant under monotone change to pairwise distances (structure not height)

Comparing Dendrograms

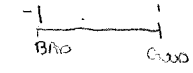
Let h_{ij} be the height that $\vec{x}_i$ and $\vec{x}_j$ were merged into the same cluster

$$H = h_{ij} = \begin{bmatrix} 0 & & & \\ h_{21} & \ddots & & \\ h_{31} & \ddots & \ddots & \\ \vdots & \ddots & \ddots & 0 \end{bmatrix}$$

Cophenetic correlation is the (Pearson) correlation between the $\frac{N(N-1)}{2}$ pairs

$$(h_{ij}, D_{ij}) \quad 1 \le i < j \le N$$

1) Can be used to compare dendrograms generated by different algorithms



2) Quality of this measure is limited by usefulness of D_{ij}

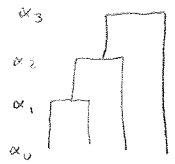
* Consider 2 non-intersecting but close manifolds

D based on Euclidean distance

Choosing Number of Clusters ("Best Tree Cut")

- **Mojena's Upper Tail Rule**

Let $\alpha_0, \ldots \alpha_{N-1}$ are the heights corresponding to stages w/ $n, n-1, \ldots, 1$ clusters. Select the number of clusters, j, corresponding to the first stage in the dendrogram satisfying



- $\alpha_{j+1} > \bar{\alpha} + k s_\alpha$
  - $\bar{\alpha} = \frac{1}{N} \sum_j \alpha_j$
  - $s_\alpha^2 = \frac{1}{N-1} \sum (\alpha_j - \bar{\alpha})^2$
  - k is user specified cutoff
    - 2.75 – 3.5 (Mojena 1977)
    - 1.25 (Milligan, Cooper 1985)
    - t-distribution (assumes normality of $\alpha_0, \ldots \alpha_{N-1}$)

* can also use moving averages (Mojena Rule 2)

- Plot $(\alpha_j - \bar{\alpha}) / s_\alpha$ and look for delineation

Ex. **NCI Microarray Data**

See Canvas > Files > Numerical Examples > Clustering