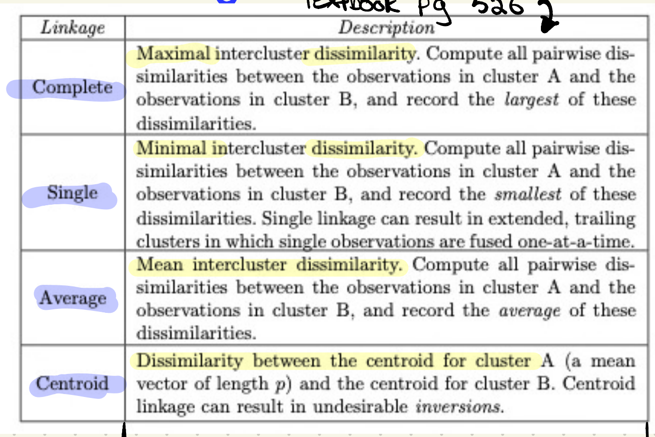
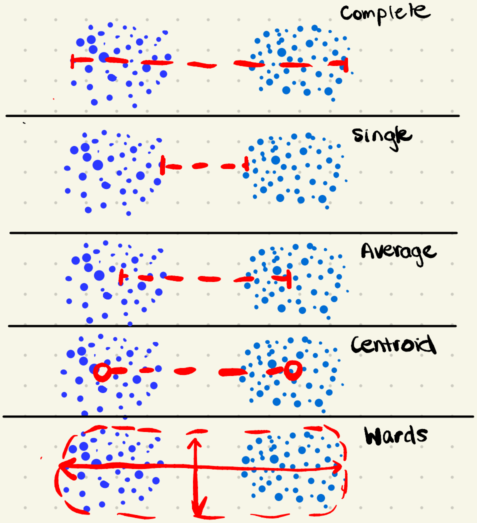
④ Linkage

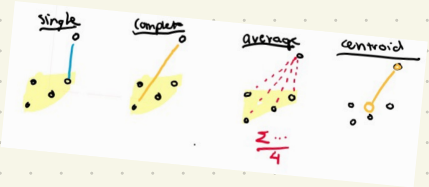
Textbook pg 526



Linkage	Description
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.
Wards	minimises the total within cluster variance

looks like



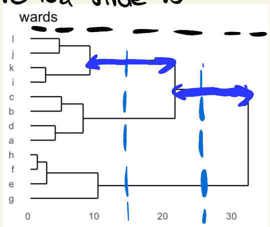


← Drawing found in my lecture notes.

* Specific for adding a single observation in

* Dendrogram

Lecture 10a slide 15



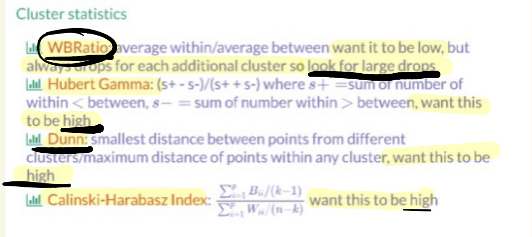
cut for 3 clusters

Cut for 2 clusters

x axis = distance

* usually cut when there is a reasonable distance without a new cluster being identified.

* Cluster stats



Cluster statistics

- WBRatio: average within/average between want it to be low, but always drops for each additional cluster so look for large drops
- Hubert Gamma: (s+ - s-)/(s+ + s-) where $s+$ =sum of number of within < between, $s-$ = sum of number within > between, want this to be high
- Dunn: smallest distance between points from different clusters/maximum distance of points within any cluster, want this to be high
- Calinski-Harabasz Index: $\frac{\sum_{i=1}^{p} B_{ii}/(k-1)}{\sum_{i=1}^{p} W_{ii}/(n-k)}$ want this to be high

These stats will likely be in next weeks lecture but because they are used in this weeks tutorial I added them in.

Pearson gamna max