EAS 509: Statistical Learning and Data Mining II

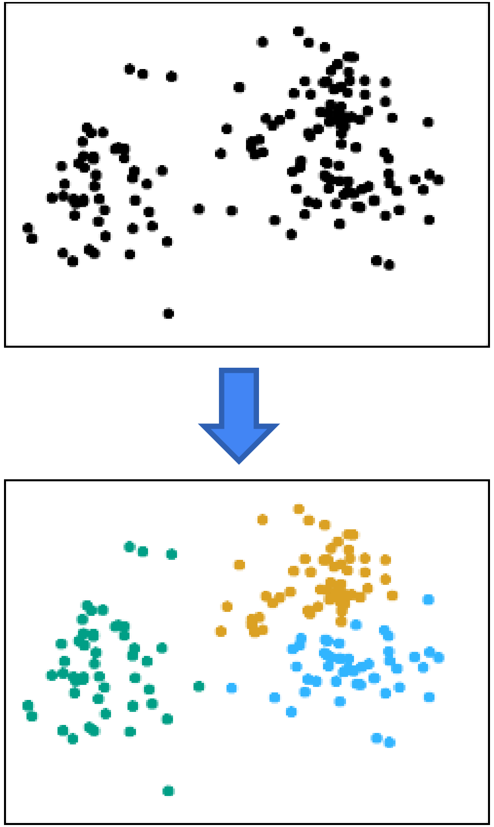
Unsupervised Methods: Clustering

Slides due to Introduction to Statistical Learning, with applications in R (2nd edition):
https://web.stanford.edu/~hastie/MOOC-Slides/unsupervised.pdf

See also videos from the book authors:
https://www.youtube.com/playlist?list=PL5-da3qGB5IBC-MneTc9oBZz0C6kNJ-f2

Textbook (Available On-line):
An Introduction to Statistical Learning 2nd Edition, by Gareth James, Daniela Witten, Trevor Hastie.
https://www.statlearning.com/ chapter 12.4 and 12.5
(Advanced) The Elements of Statistical Learning Data Mining, Inference, and Prediction, by Trevor Hastie, Robert Tibshirani, Jerome Friedman. https://hastie.su.domains/ElemStatLearn/

Clustering Recap



- *Clustering* refers to a very broad set of techniques for finding *subgroups*, or *clusters*, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- We must define what it means for two or more observations to be *similar* or *different*.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Dissimilarity Recap

Let:

$\{x_{ij}\}\, where\ i = 1, \dots, n, j = 1, \dots, p$

n observation and p features

$d_{ii'}$ distance between obs. i and i'

Suppose clustering into k – clusters:

$C_1, C_2, \dots, C_k$

$C_r = \{indexes\ of\ observations\ in\ r^{th}\ cluster\}$

$n_r = |C_r|$ - number of observations in r^th cluster

$D_r = \sum_{ii' \in C_r} d_{ii'}$ - the sum of pairwise distances for all points in cluster r

$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$ pooled within-cluster sum pf squares around cluster mean (if $d_{ii'}$ is Euclidian distance)

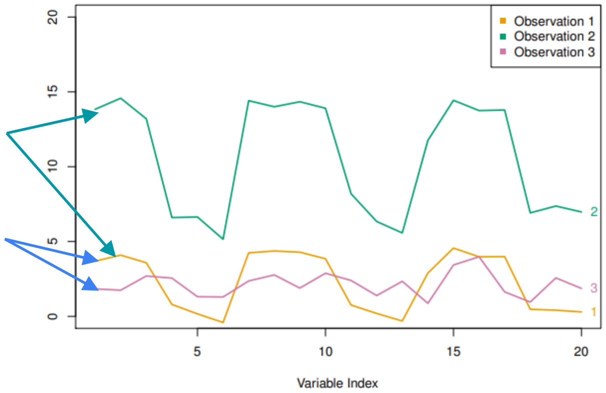
2?

Euclidian distance: $d_{ii'} = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$

L^1 distance : $d_{ii'} = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$

Pearson Correlation Distance: $d_{ii'} = 1 - r_{ii'}$

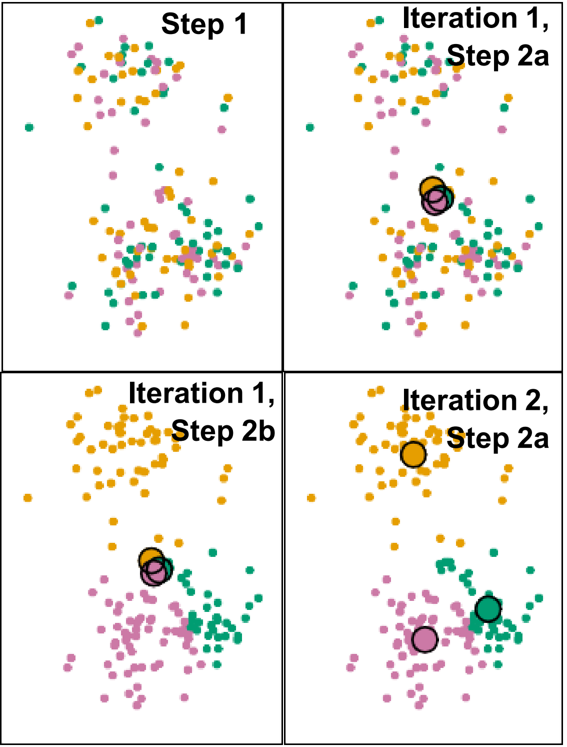
$$r_{ii'} = \frac{\sqrt{\sum_{j=1}^{p}(x_{ij} - \overline{x_i})(x_{i'j} - \overline{x_{i'}})}}{\sqrt{\sum_{j=1}^{p}(x_{ij} - \overline{x_i})^2}\ \sqrt{\sum_{j=1}^{p}(x_{i'j} - \overline{x_{i'}})^2}}$$



Correlation Distance Similar

Euclidian/L1 Distance Similar

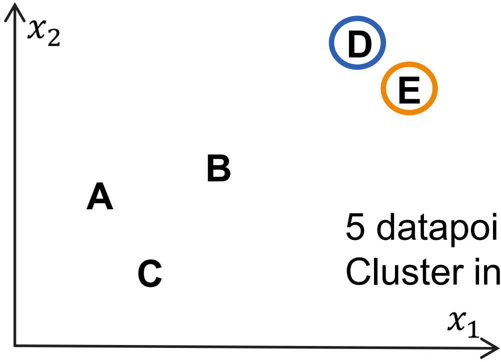
K-Means Clustering Recap: Algorithm



1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations. (Random Partition).
    1. Alternatively (Forgy): choose k-observation to be ‘centroids’ and perform 2.b

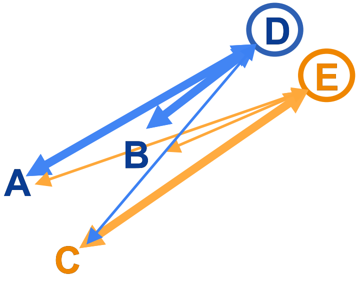
2. Iterate until the cluster assignments stop changing:
    a) For each of the K clusters, compute the cluster **centroid**. The k-th cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

1. Randomly peak *k* data points and use them as centroids (forgy algorithm)

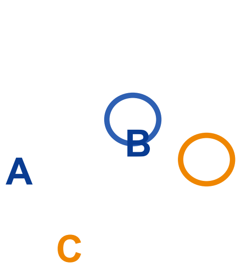


5 datapoints
Cluster into 2 groups

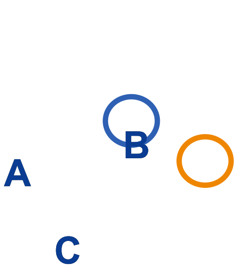
2.a Assign cluster number based on the distance to centroids



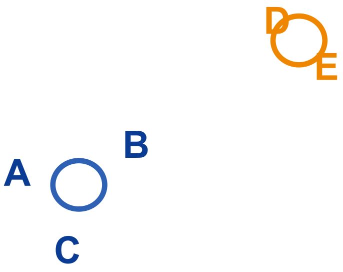
2.b Recalculate centroids as a center (mean coordinates) of each cluster



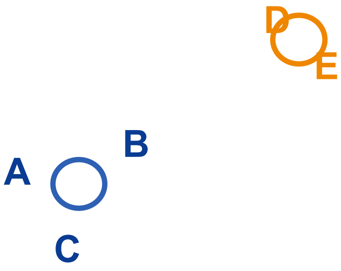
Iteration 2. Step 2.a re-Assign cluster



Iteration 2. Step 2.b Recalculate centroids



Iteration 3. Step 2.a re-Assign cluster



Clusters didn’t change -> Stop!

Properties of the Algorithm

This algorithm is guaranteed to decrease the value of the objective function (within cluster **variation**) at each step:

$$WCV(C_k) = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2$$

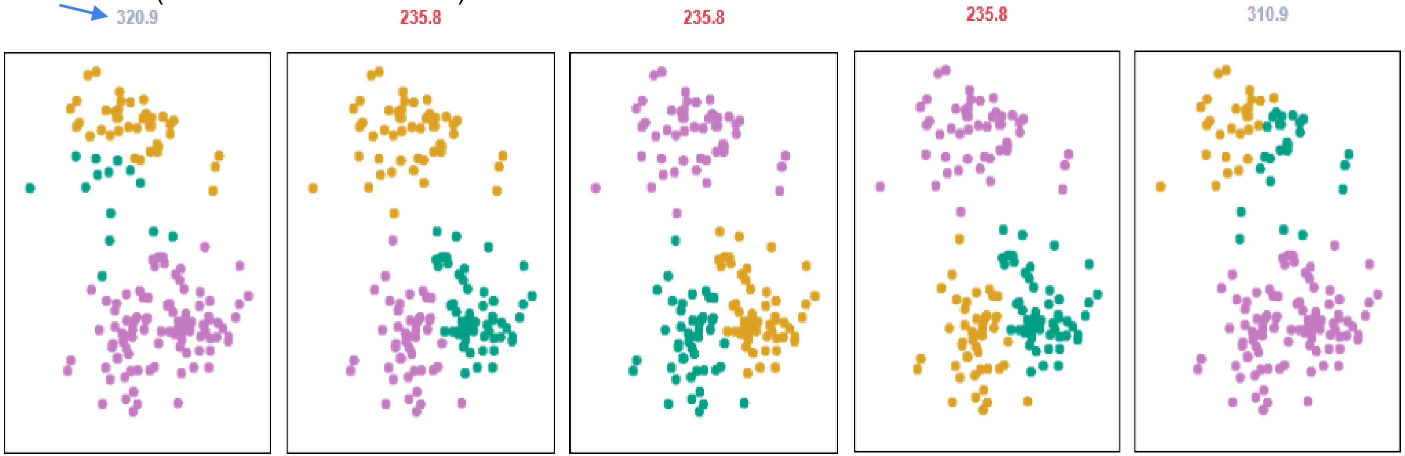
The total WCV:

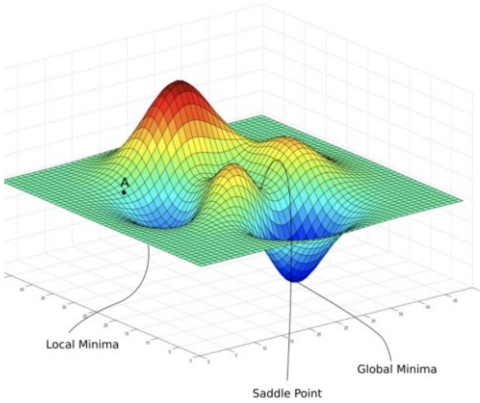
$$WCV = \sum_{k=1}^{K} WCV(C_k) = 2\sum_{k=1}^{K}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2$$

In step 2.a the point i will move from cluster k to k' only if distance is smaller to k' centroid, thus new WCV' on old centroids is $WCV' \leq WCV$. Because mean minimize sum of squares true new WCV'' will be smaller or equal to WCV' so $WCV'' \leq WCV$

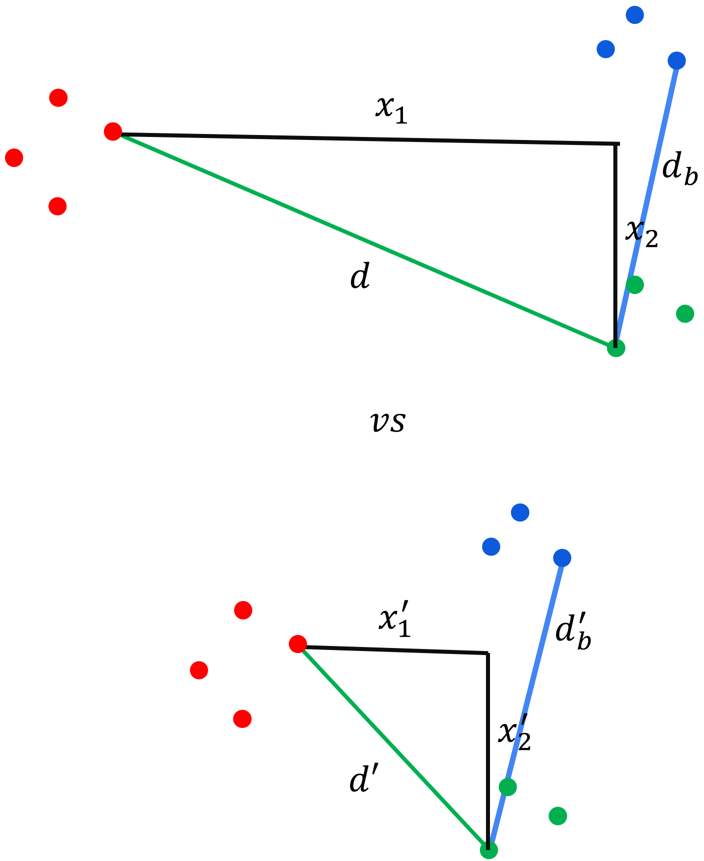
However, **it is not guaranteed to give the global minimum**

Sum of WCV (Within cluster **variation**)



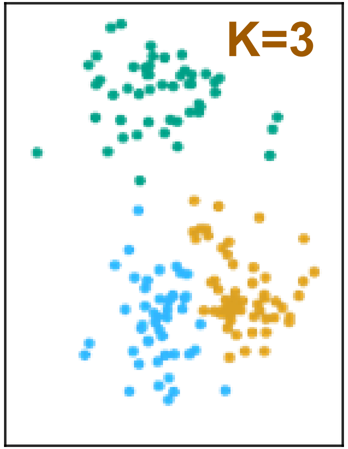


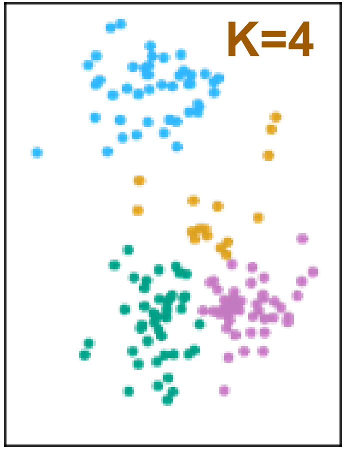
Scaling Matters



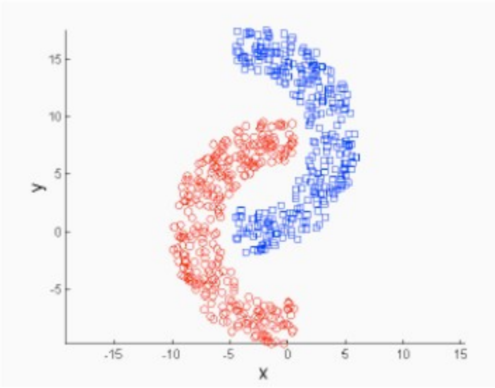
- Scaling of the variables matters!
- Dimensions with larger variance will dominate the dissimilarity.
- Typically, if the units are different, we do standardization
- If the units are same, we still can do standardization

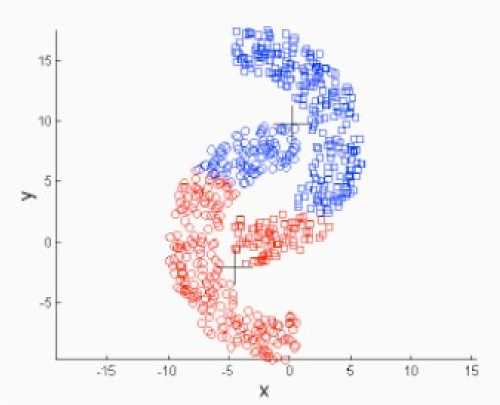
K-Means Clustering Limitations



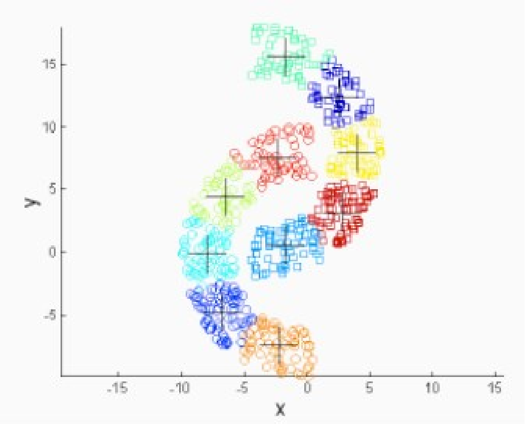


- It is not guaranteed to converge the global minimum
- When moving to higher number of clusters, it does not split the existing clusters but create a new groupings
- Favor globular clusters



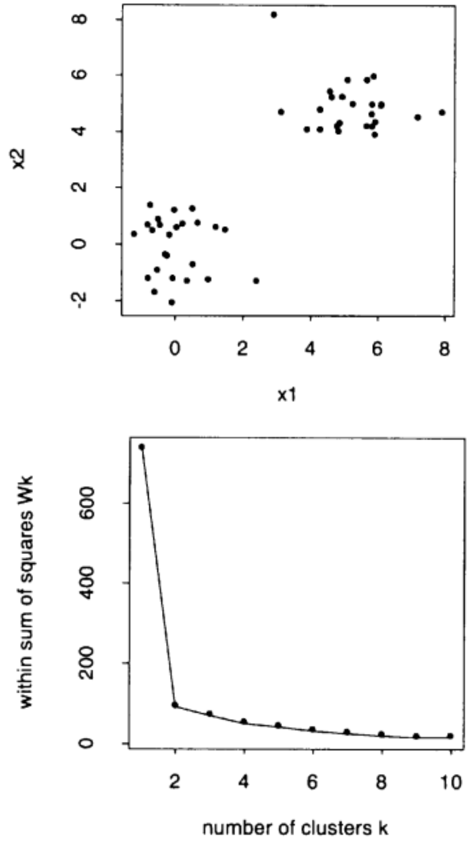


K-means (2 Clusters)



K-means (10 Clusters)

Choosing Number of Clusters



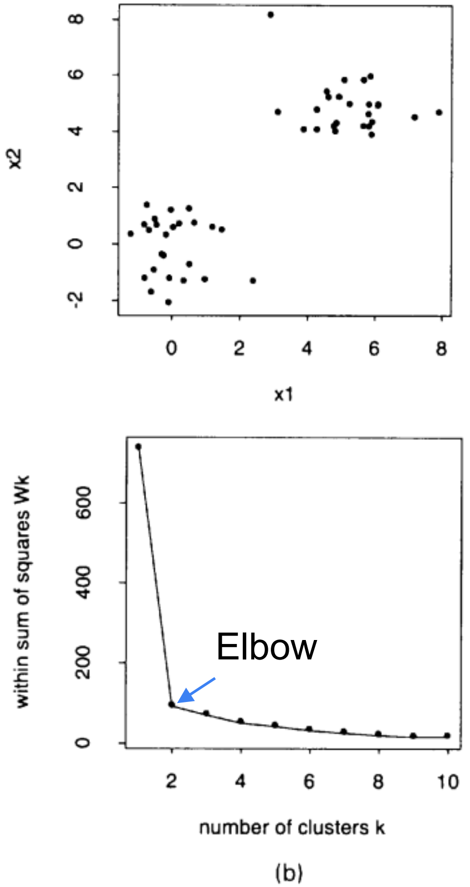
(b)

Tibshirani, Walther, G., & Hastie, T. (2001).

There is no consensus on this

- “Elbow”
- Gap Statistics
- Silhouettes

Choosing Number of Clusters: “Elbow” Method



Tibshirani, Walther, G., & Hastie, T. (2001).

Let:

$\{x_{ij}\}$ $where\ i = 1, \ldots, n, j = 1, \ldots, p$

n observation and p features

$d_{ii'}$ distance between obs. i and i'

$d_{ii'} = \sum_{j=1}^{p}\left(x_{ij} - x_{ij}\right)^2$ for Euclidian distance

Suppose clustering into k – clusters:

$C_1, C_2, \ldots, C_k$

$C_r = \{indexes\ of\ observations\ in\ r^{th}\ cluster\}$

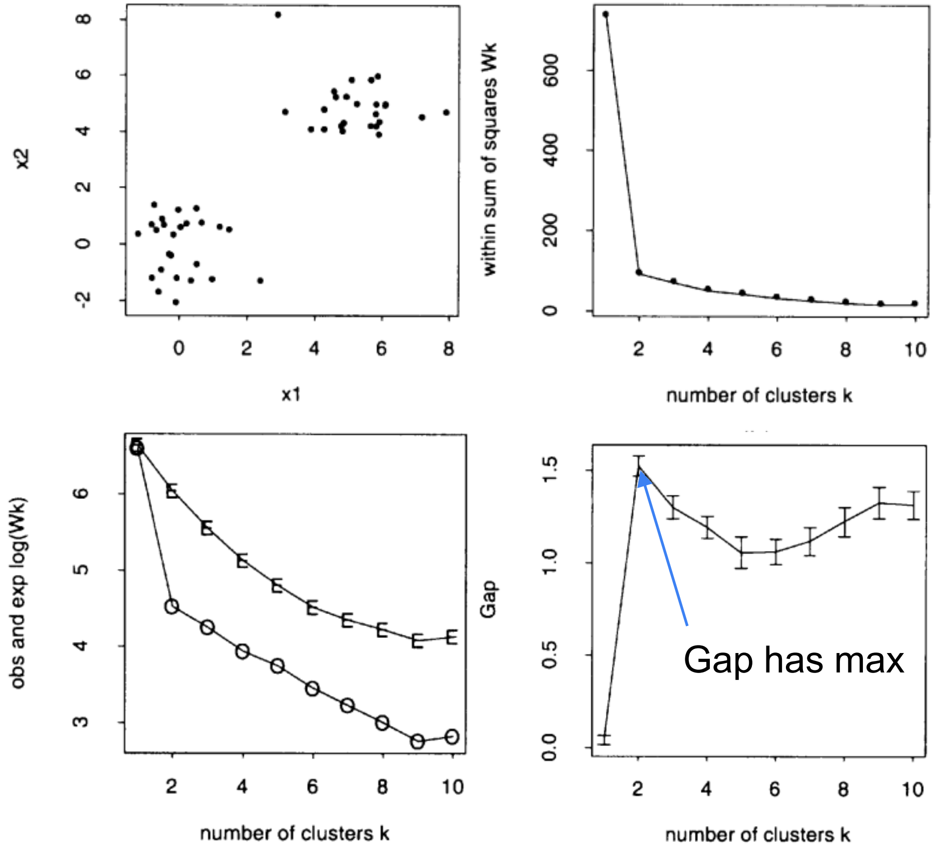
$n_r = |C_r|$ - number of observations in r^th cluster

$D_r = \sum_{ii' \in C_r} d_{ii'}$ - the sum of pairwise distances for all points in cluster r

$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$ pooled within-cluster sum of squares around cluster mean (if $d_{ii'}$ is Euclidian distance)

Choose such k which has large decrease from k-1 to k and following decreases are small.

Choosing Number of Clusters: Gap Statistics



Gap has max

The idea behind Gap Statistics is to use reference distribution and contrast within-cluster sum of squares of target system to one from reference distribution.

Gap Statistics:

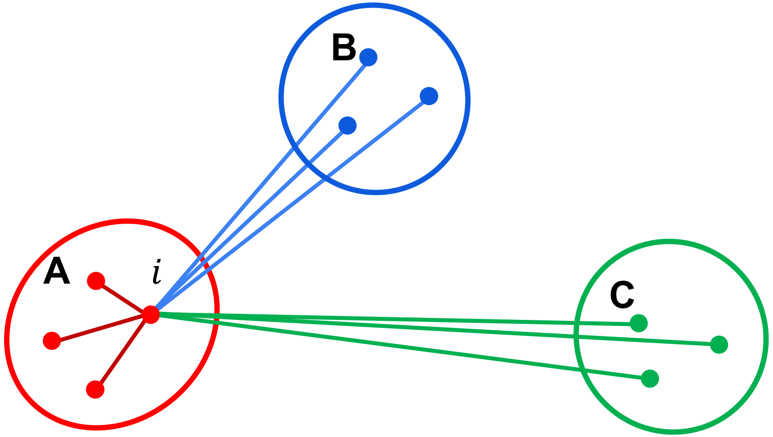
$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

where E_n^* is expectation under sample of size n from reference distribution

Selecting reference distribution as p-dimensional uniform with n observation the Gap statistic would have a peak at true k (see article for detailed explanation).

Tibshirani, Walther, G., & Hastie, T. (2001).

Silhouettes: A graphical aid to the interpretation and validation of cluster analysis



Silhouette statistics is based on comparison of **cluster tightness** and **inter-cluster separation**

Silhouette statistics:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$$-1 \leq s(i) \leq 1$$

For item i of cluster A:

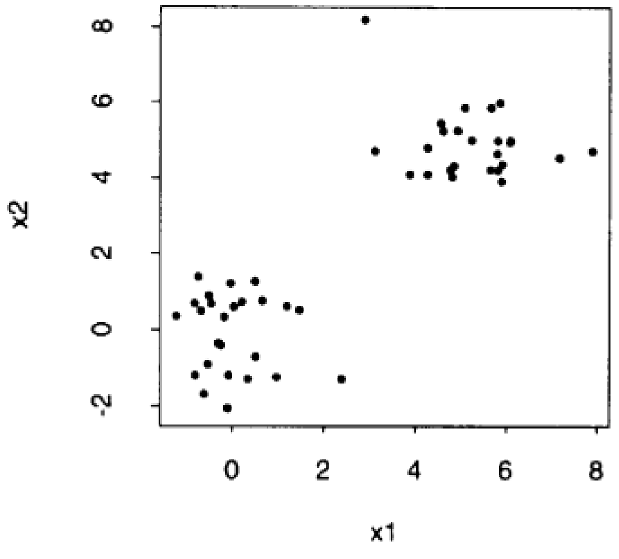
$a(i)$ - average distance (dissimilarity) between i and all other items from cluster A

$d(i, C)$ - average dissimilarity between i of A and all other items from cluster C

$b(i) = \min_{C \neq A} d(i, C)$ average dissimilarity between i of A and all other items from closest cluster

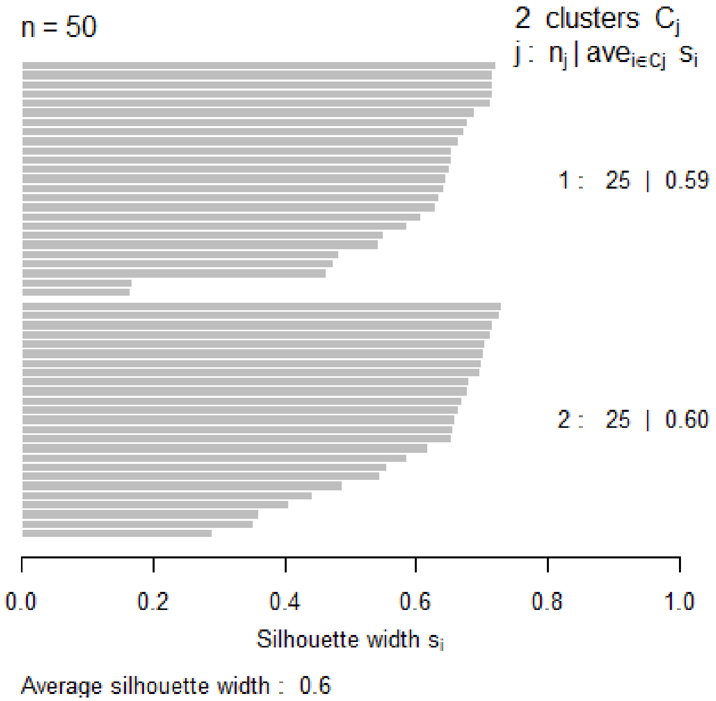
- If $s(i)$ is **positive** then separation is bigger than tightens (*good!*)
- If $s(i)$ is **zero** then separation is same as tightens (are they same cluster? may be?)
- If $s(i)$ is **negative** then separation is smaller than tightens (Hm?)

Silhouettes: A graphical aid to the interpretation and validation of cluster analysis

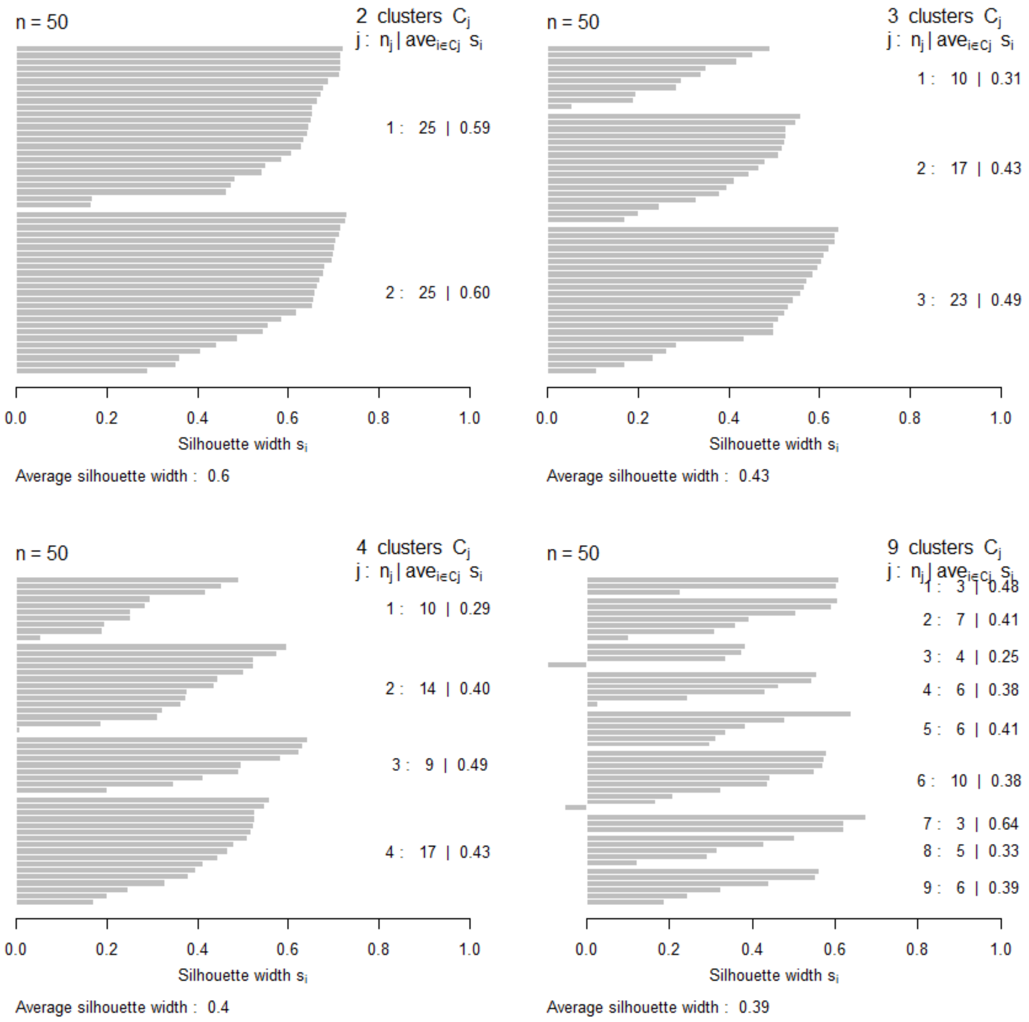


(a)

Silhouette plot

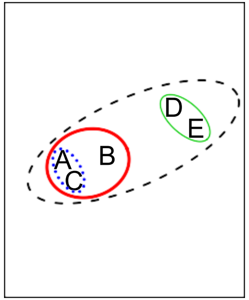


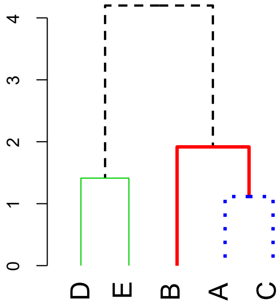
Silhouettes



Choose k with larger width and no negative items (are they outliers?)

Hierarchical Clustering

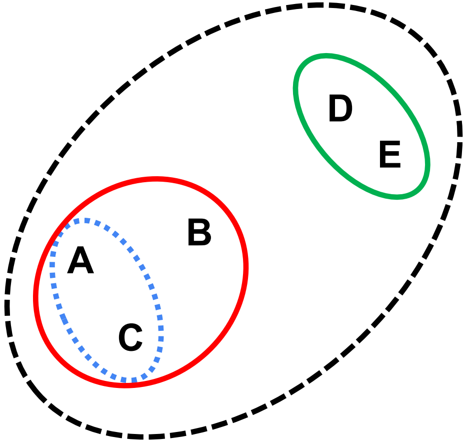


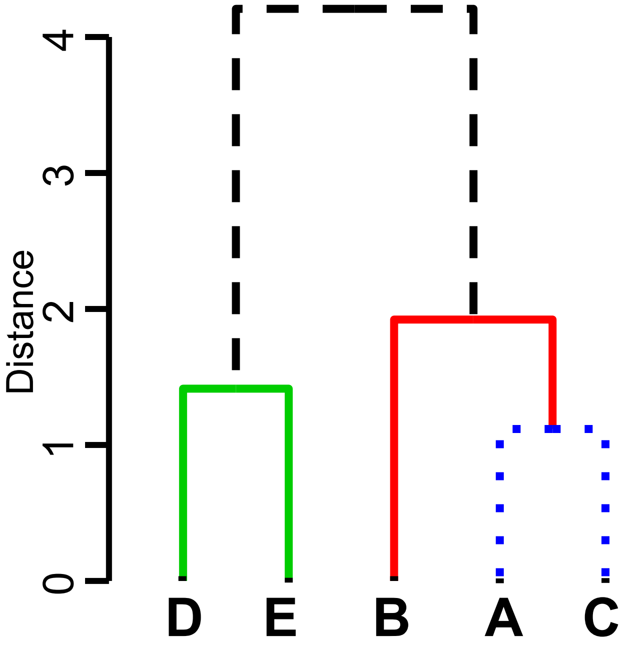


- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- *Hierarchical clustering* is an alternative approach which does not require that we commit to a particular choice of K.
- In hierarchical clustering a dendrogram is built starting from the leaves and combining clusters up to the trunk.

- Clustering methods can be grouped in:
  - bottom-up (agglomerative) methods (Hierarchical clustering)
  - top-down methods

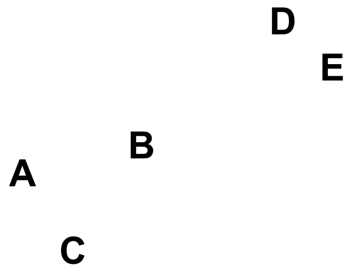
Hierarchical Clustering Algorithm

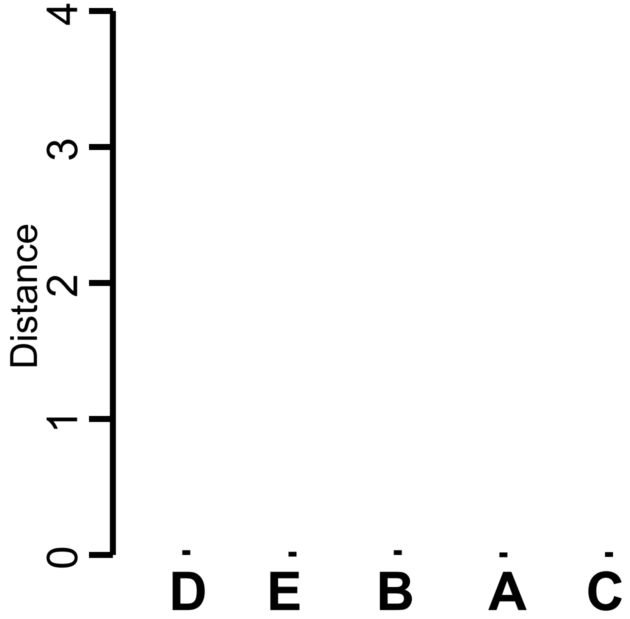




- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

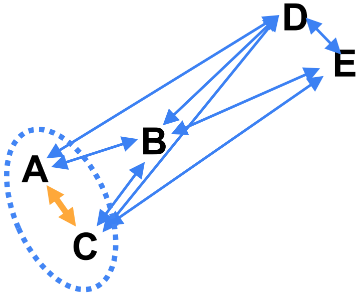
Hierarchical Clustering Algorithm

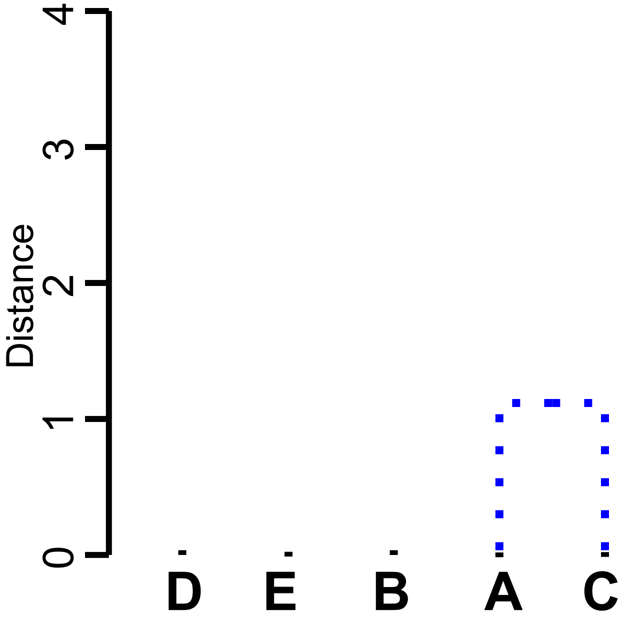




- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

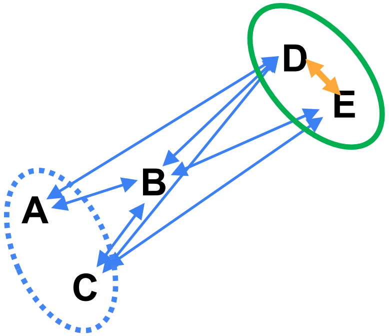
Hierarchical Clustering Algorithm

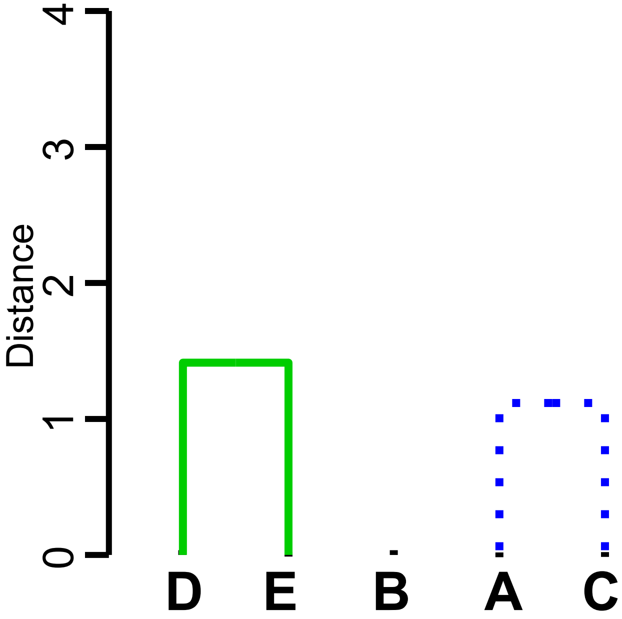




- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

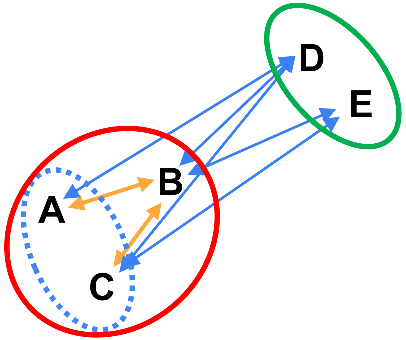
Hierarchical Clustering Algorithm

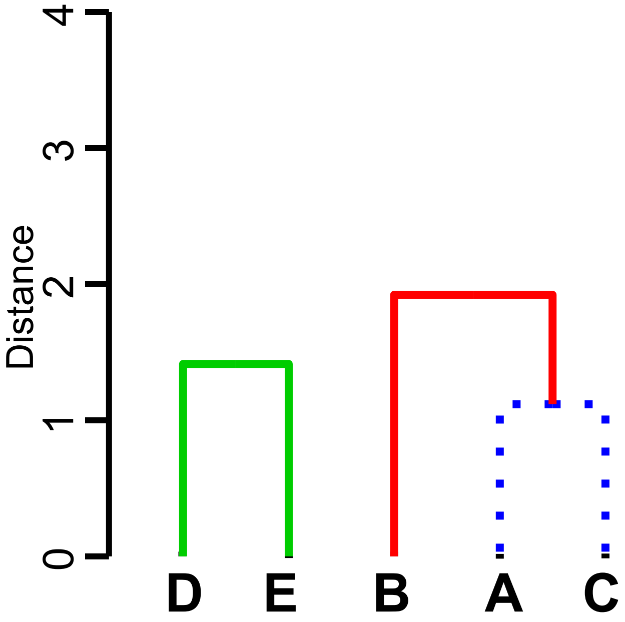




- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- **Repeat.**
- Ends when all points are in a single cluster.

Hierarchical Clustering Algorithm



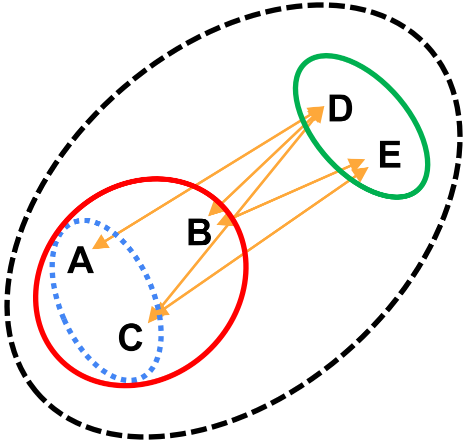


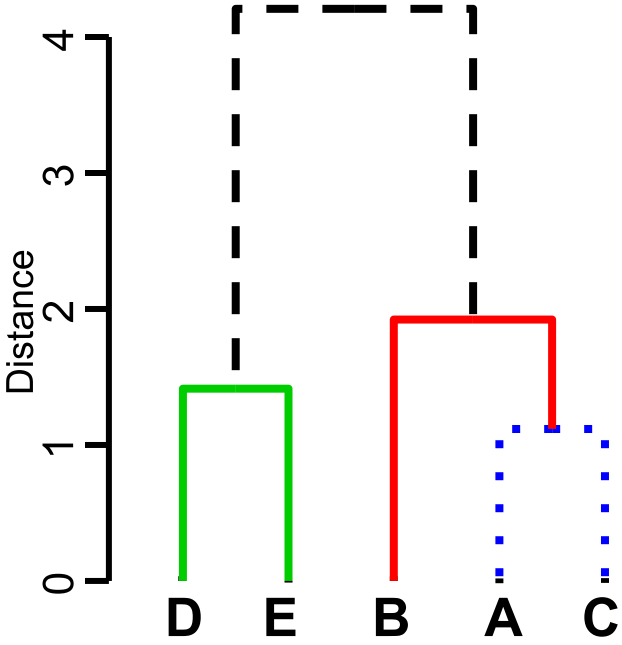
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- **Repeat.**
- Ends when all points are in a single cluster.

Linkage – defines how to calculate distance between clusters containing multiple items:

- **Complete** – largest distance
- **Single** – smallest distance
- **Average** – average dissimilarity between all elements of two clusters
- **Centroid** - Dissimilarity between the centroids

Hierarchical Clustering Algorithm



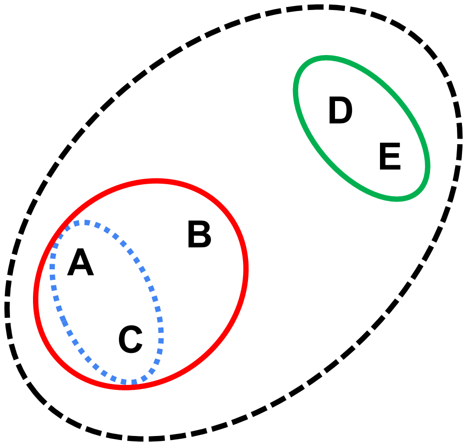


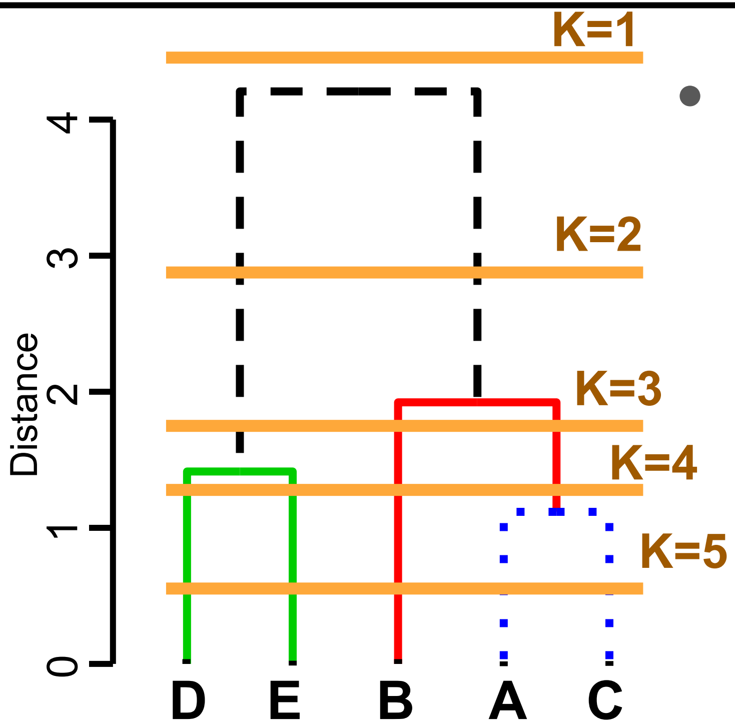
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- **Repeat.**
- Ends when all points are in a single cluster.
- **Average** – average dissimilarity between all elements of two clusters
- **Centroid** - Dissimilarity between the centroids

Linkage – defines how to calculate distance between clusters containing multiple items:

- **Complete** – largest distance
- **Single** – smallest distance
- **Average** – average dissimilarity between all elements of two clusters
- **Centroid** - Dissimilarity between the centroids

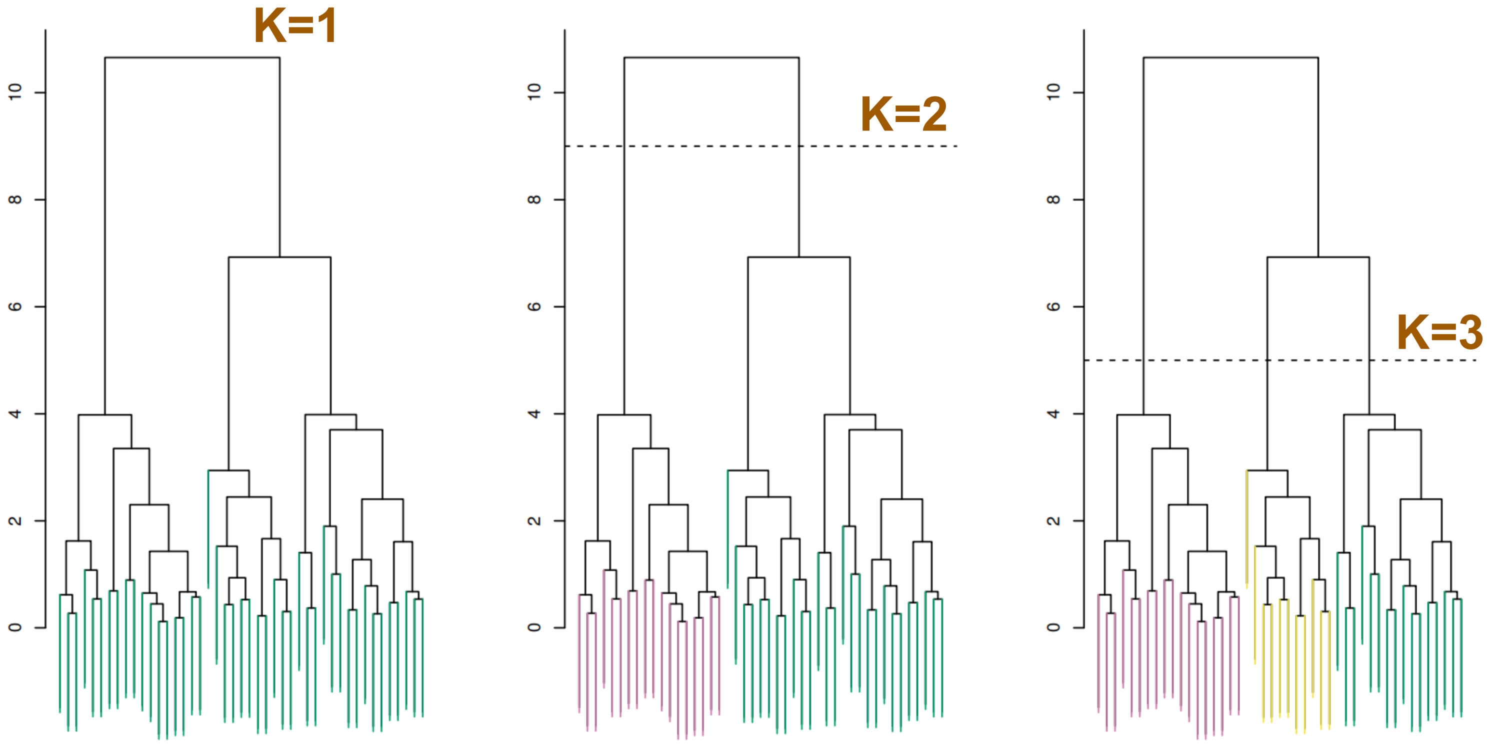
Hierarchical Clustering: Getting K-Clusters





- Cut at proper height to get desired number of clusters

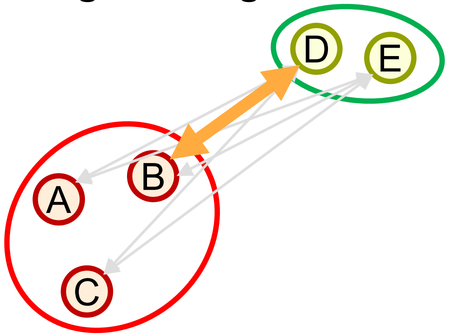
Hierarchical Clustering: Getting K-Clusters, Example



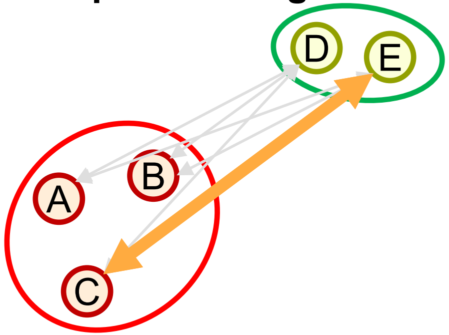
- Illustration on selecting different number of cluster

Hierarchical Clustering: Linkage

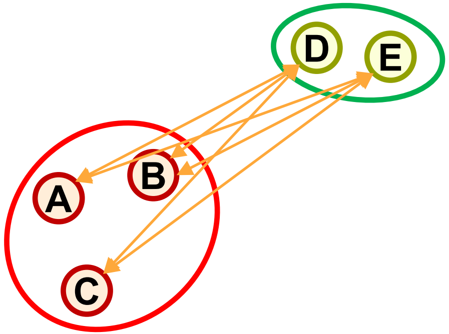
Single Linkage



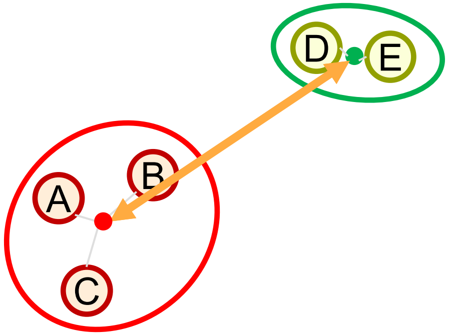
Complete Linkage



Average Linkage



Centroid Linkage



Linkage – defines how to calculate distance between clusters containing multiple items:

- **Complete** – largest distance between elements of two clusters
- **Single** – smallest distance between elements of two clusters
- **Average** – average dissimilarity between all elements of two clusters
- **Centroid** - Dissimilarity between the centroids

Hierarchical Clustering: Linkage

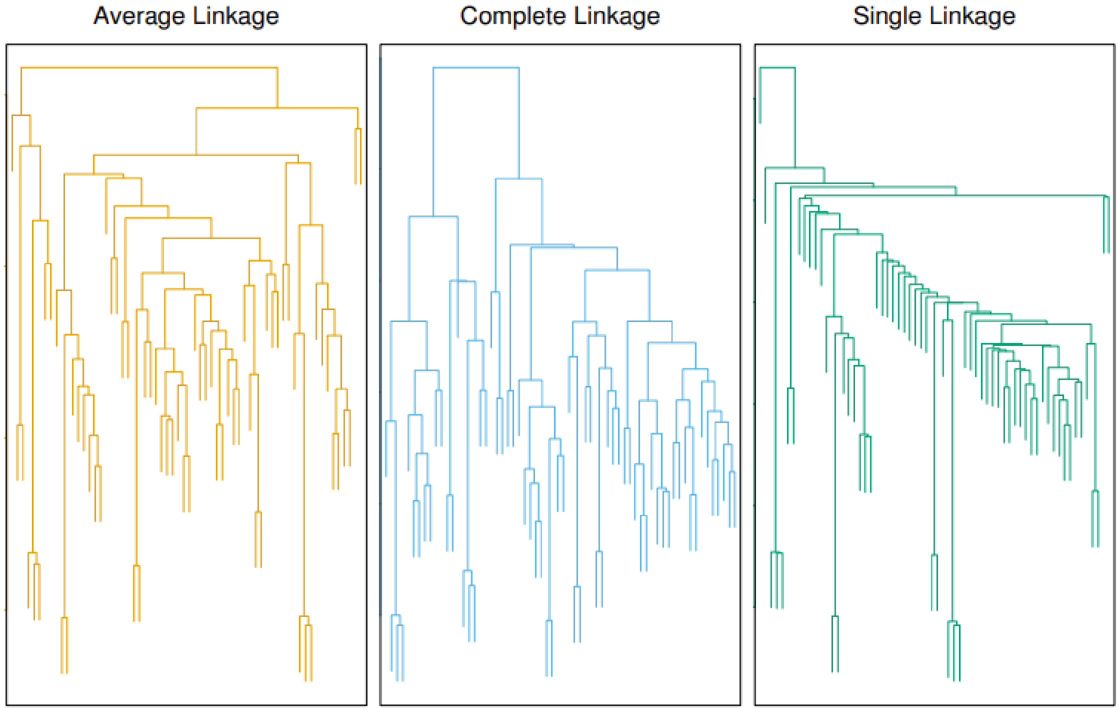


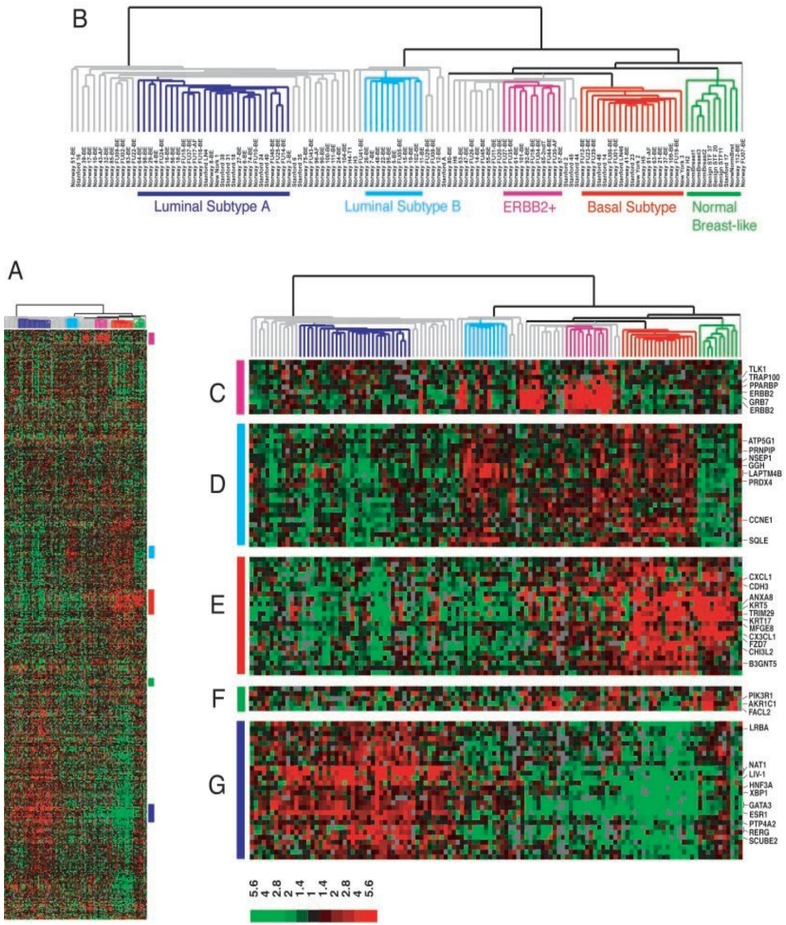
FIGURE 12.14. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

- Average and Complete Linkages are most often used
- Single linkage often produce long, stringy, clusters, i.e. one element at a time. Not balanced clusters
- Centroid are often used in genomics

Practical issues

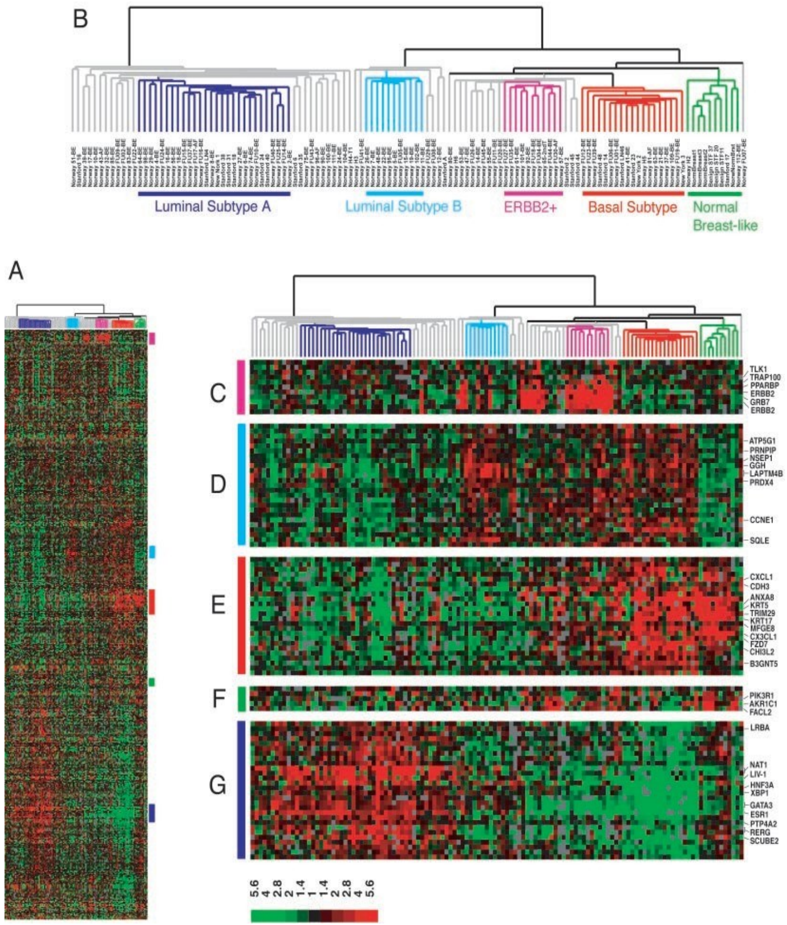
- *Scaling of the variables matters!*. Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
- How many clusters to choose? (in both K-means or hierarchical clustering). Difficult problem. No agreed-upon method.
    - Too many?
    - Too few?
- Which features should we use to drive the clustering?

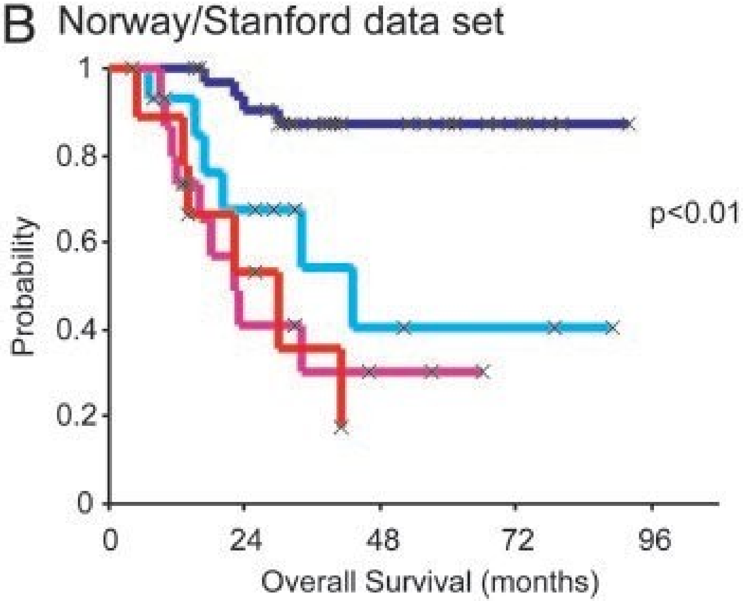
Example: breast cancer microarray study



- "Repeated observation of breast tumor subtypes in independent gene expression data sets;" Sorlie at el, PNAS 2003
- Gene expression measurements for about ~ 8000 genes, for each of 88 breast cancer patients.
- Average linkage, correlation metric
- Clustered samples using 500 *intrinsic genes:* each woman was measured before and after chemotherapy. Intrinsic genes have smallest within/between variation.

Example: breast cancer microarray study





Summary

- *Unsupervised learning* is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning
- It is intrinsically more difficult than *supervised learning* because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy).
- It is an active field of research, with many recently developed tools such as *self-organizing maps*, *independent components analysis* and *spectral clustering.*
- See *The Elements of Statistical Learning*, chapter 14.