

# Clustering Approaches

*Advanced Biostatistics*

Dean Adams

Lecture 9

EEOB 590C

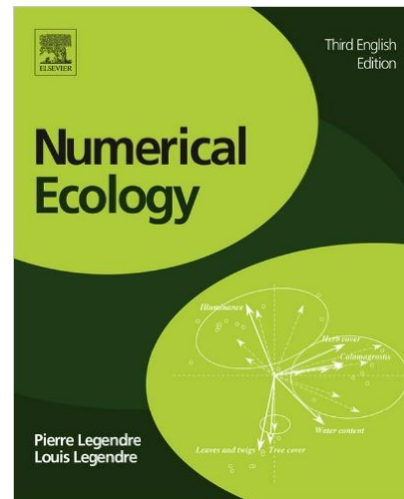
Data are points in multivariate space

- Look for patterns in high-dimensional spaces
- Generate summary plots of dataspace (ordination)
- Look for relationships of points (clustering)

**Clustering** focuses on the similarities (or differences) among observations

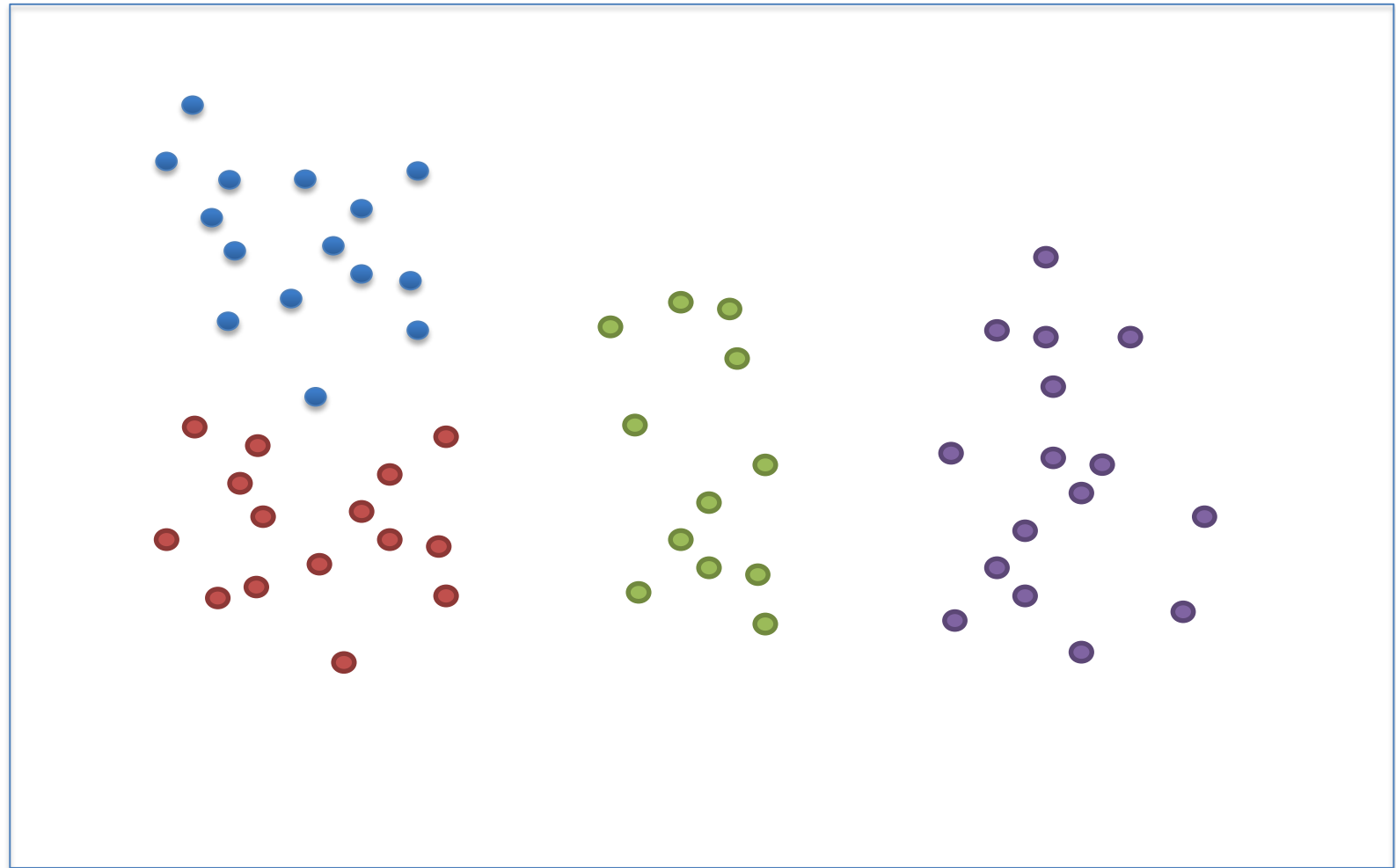
- Obtain groupings (clusters) of observations based on their similarity or difference
- Provides complementary view to ordination
- Clustering is algorithmic, not algebraic (i.e., a set of [repeated] rules for connecting observations)

- Overview of approaches
- SAHN methods
- Dendrograms
- K-means partitioning



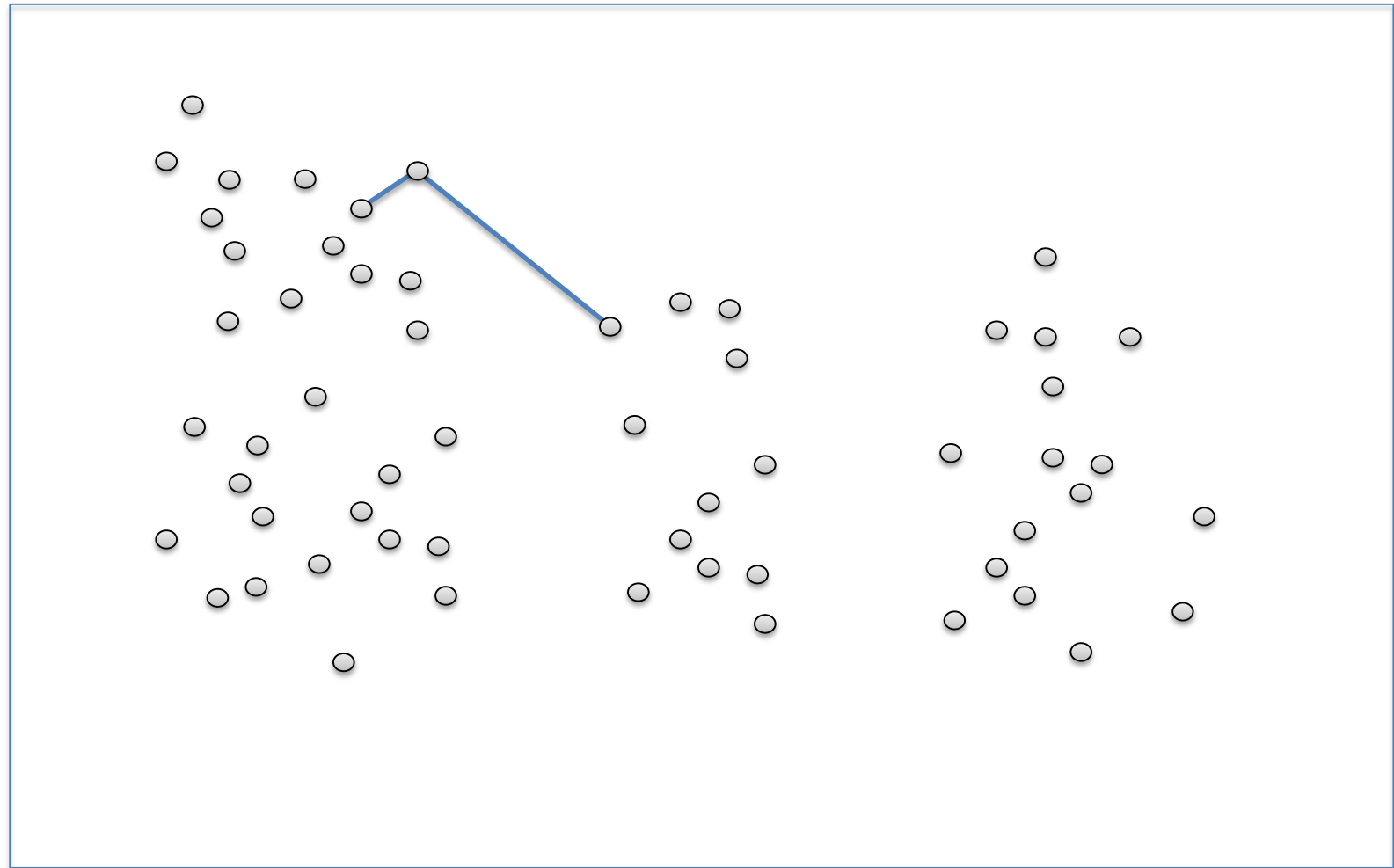
Legendre and Legendre 2012.  
Numerical Ecology, 3<sup>rd</sup>  
edition. Elsevier.

PC 2

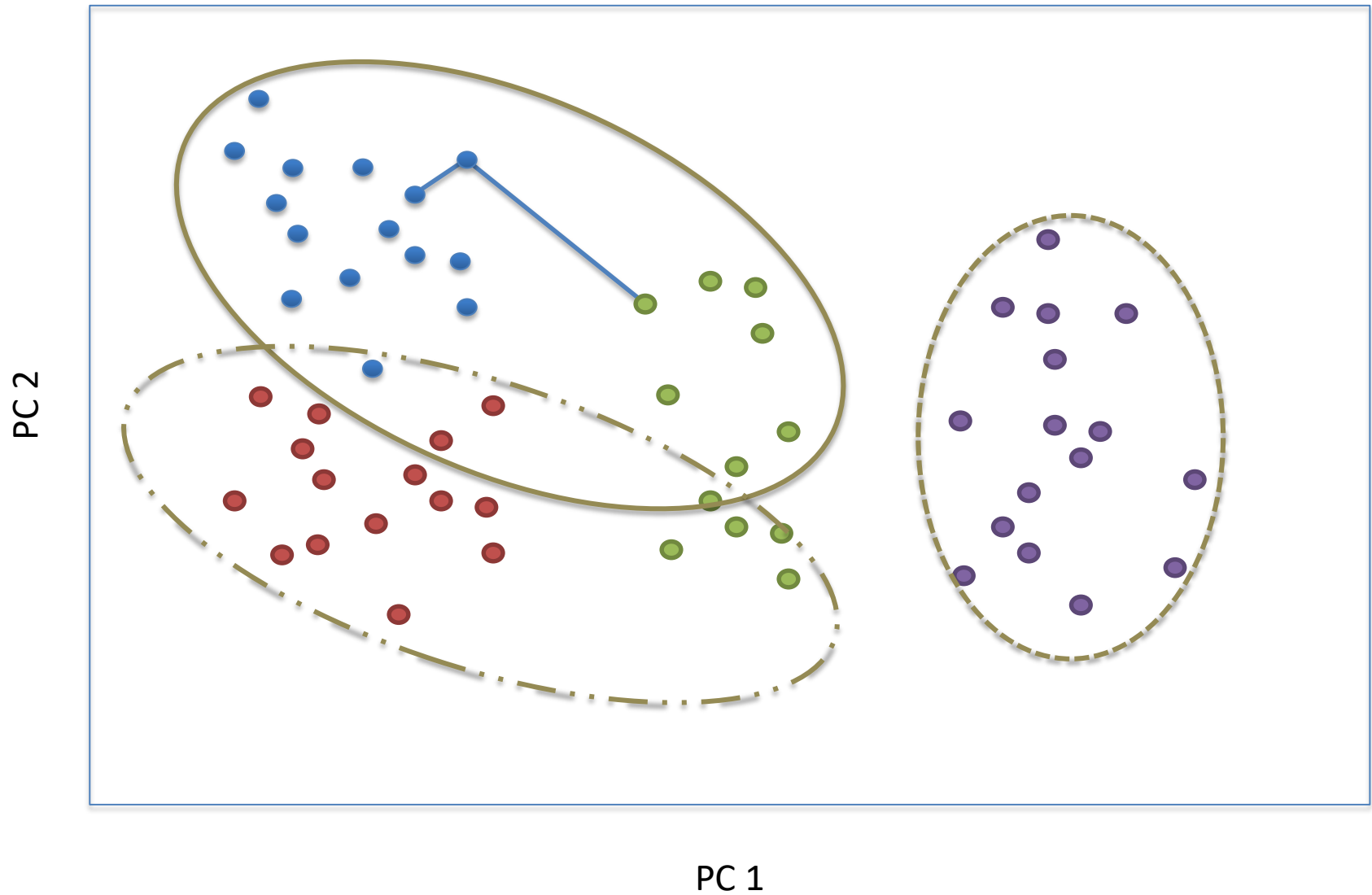


PC 1

PC 2

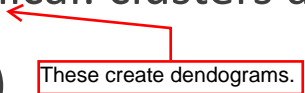


PC 1



## Two main categories

Hierarchical: clusters are nested (higher-rank clusters contain lower-rank clusters)



These create dendograms.

Non-hierarchical: optimal strategy for finding  $K$  groups

## Other ways to think of clustering method differences

Sequential (most methods) vs. Simultaneous

Agglomerative vs. Divisive (start with one versus start with all)

Monothetic vs. Polythetic (single descriptor for partitions vs. multiple descriptors) (Divisive methods only)

Probabilistic (assigns probability of homogeneity to putative within-group association matrices) vs. Non-probabilistic (most methods)

Two main categories

**Hierarchical**: clusters are nested (higher-rank clusters contain lower-rank clusters)

Non-hierarchical: optimal strategy for finding  $K$  groups

Other ways to think of clustering method differences

**Sequential** (most methods) vs. Simultaneous

**Agglomerative** vs. Divisive (start with one versus start with all)

Monothetic vs. Polythetic (single descriptor for partitions vs. multiple descriptors) (Divisive methods only)

Probabilistic (assigns probability of homogeneity to putative within-group association matrices) vs. **Non-probabilistic** (most methods)



SAHN: Sequential, agglomerative, hierarchical, nested

Commonly used family of clustering methods

General procedure:

- Start with similarity (or distance) matrix and rank-order values

- Cluster most similar objects first and recalculate if needed

- Aggregate until all objects are part of largest cluster

Equivalent to starting with most similar, and ‘connecting’ objects with lower and lower similarity as you progress (i.e., more different)

SAHN generates nested plot (dendrogram) whose axis is similarity

Different criteria used for determining **when** to include objects to clusters

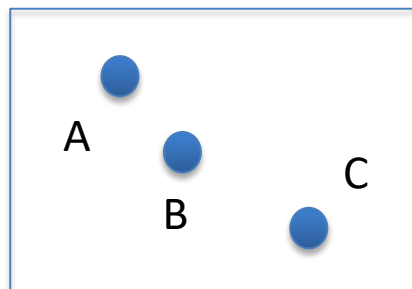
**Single linkage** (nearest neighbor clustering): add a new object to a cluster when its similarity to the *first* object is reached

**Complete linkage** (farthest neighbor clustering): add a new object to a cluster when its similarity to the *last* object is reached

Extremes of SAHN clustering

-Difference mostly amounts to ‘sliding’ nodes of dendrogram towards tips (single linkage) or towards root (complete linkage)

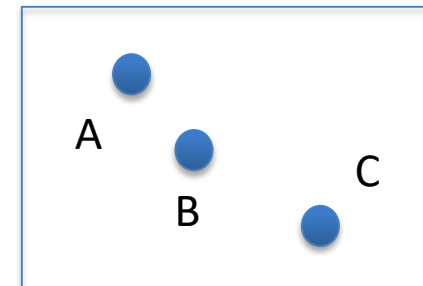
-However, cluster assignments can also change



$$d_{BC} \leq G_c \rightarrow ABC$$

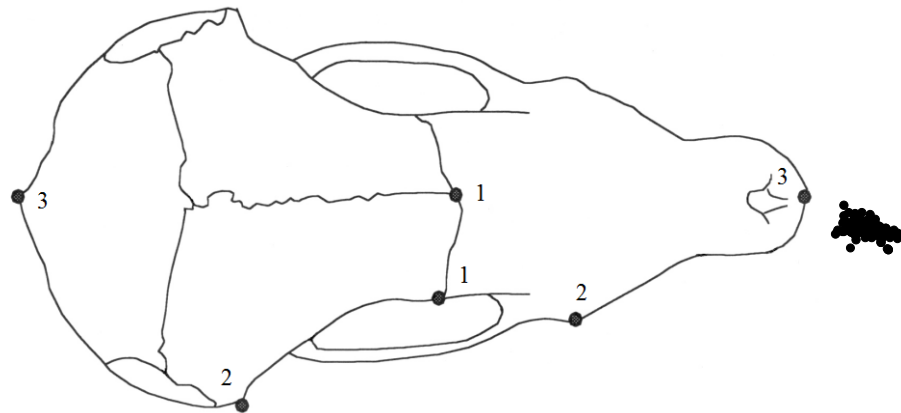
← Single linkage

Complete linkage →



$$d_{AC} \geq G_c \rightarrow AB, C$$

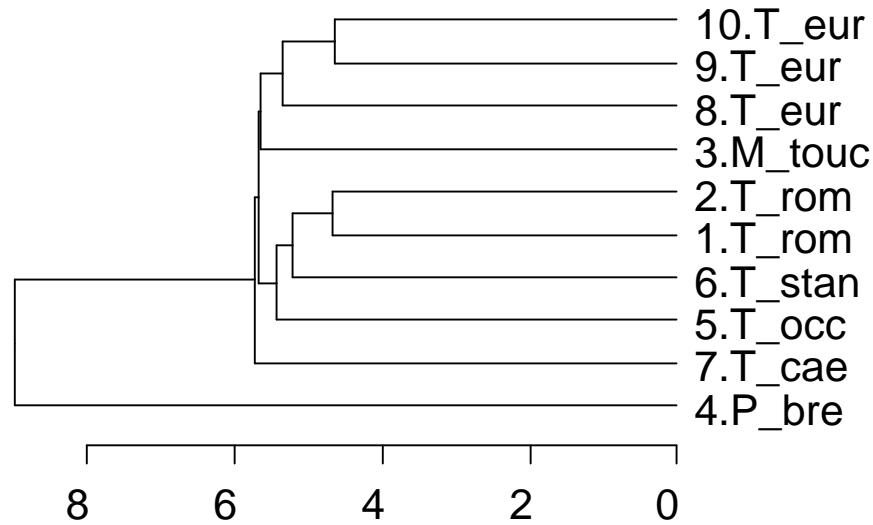
# Skull shape similarity among populations of European moles



Shape residuals for 113  
superimposed specimens

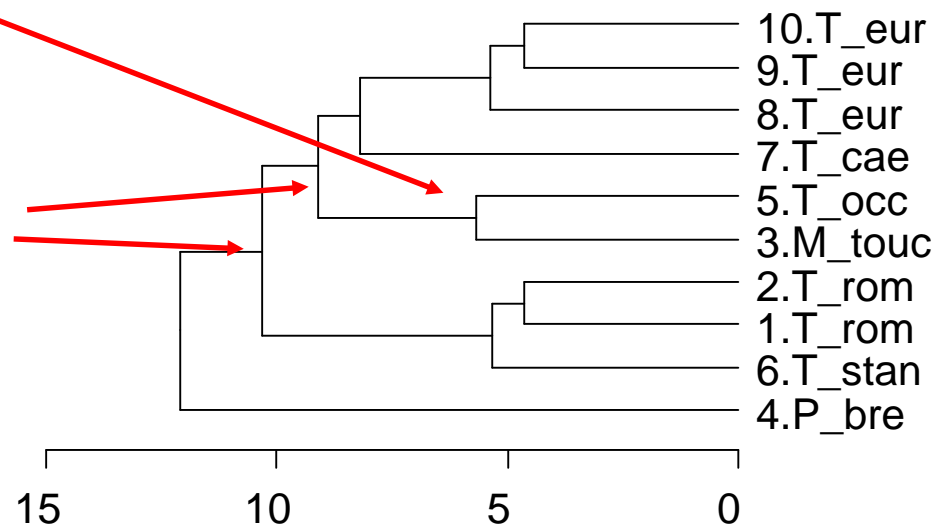


T_romana1	0									
T_romana2	4.66	0			Data expressed as shape DISTANCE					
M_touchei	8.53	8.19	0							
P_breweri	8.96	10.98	9.88	0						
T_occidentalis	6.4	5.43	5.67	9.91	0					
T_stankovici	5.35	5.2	10.28	10.32	6.45	0				
T_caeca	9.27	6.96	9.1	11.93	5.72	7.66	0			
T_europa1	8.42	7.24	8.75	12.01	7.25	8.63	7.17	0		
T_europa2	8.6	8.21	5.63	10.85	6.2	9.93	8.13	5.39	0	
T_europa3	8.82	8.26	7.58	12.09	6.57	10.32	8.21	5.35	4.64	0



Single linkage

Cluster change

Nodes shifted  
towards root

Complete linkage

Single linkage and complete linkage represent extreme criteria for clustering (include object at first instance vs. include object only when *all* instances reached)

Can be sensitive to noise in data (especially single linkage)

Can use some measure of central tendency for groups, and recalculate distances to these values

Different measures of central tendency have been used

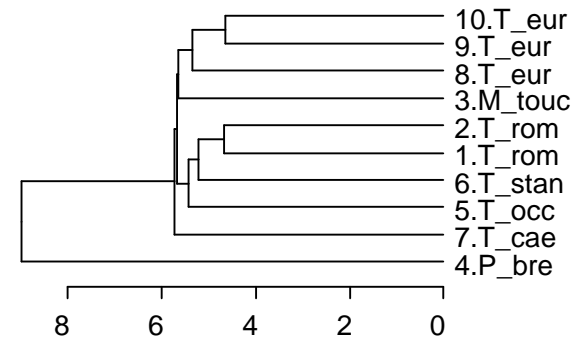
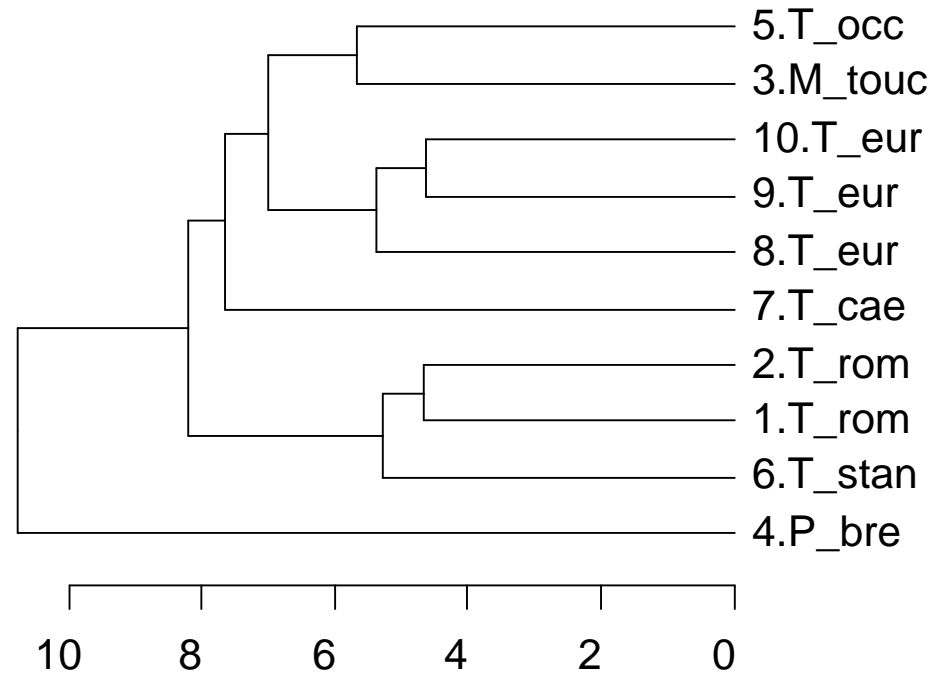
## Unweighted Pair-Group Method using Arithmetic Averages

Uses averages of clusters to join additional objects

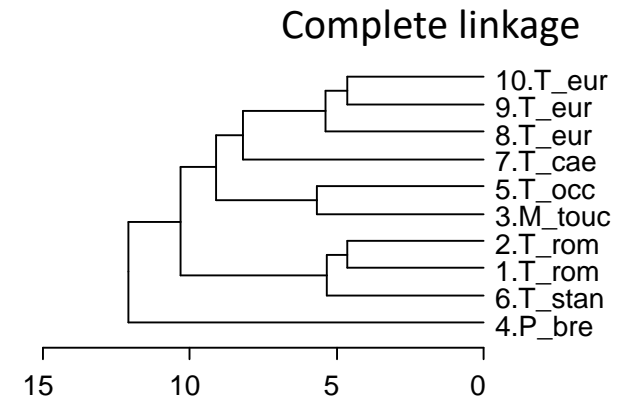
- Connect closest 2 objects in a cluster
- Calculate their average similarity to each object *not* in cluster
- Replace original similarity scores with averages
- Add new object to cluster when distance to average is reached
- Recalculate average for cluster and continue

Method unweighted because it gives same weight to *original* similarity scores

(e.g., when 3<sup>rd</sup> object added, new average found by dividing by 3, etc.)



Single linkage

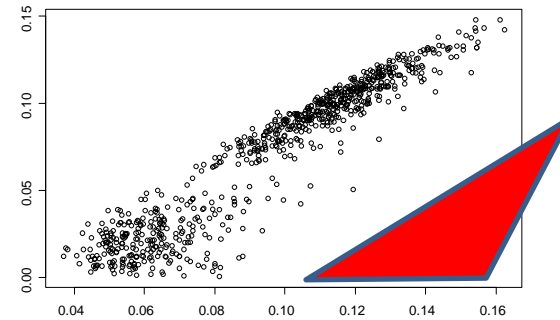
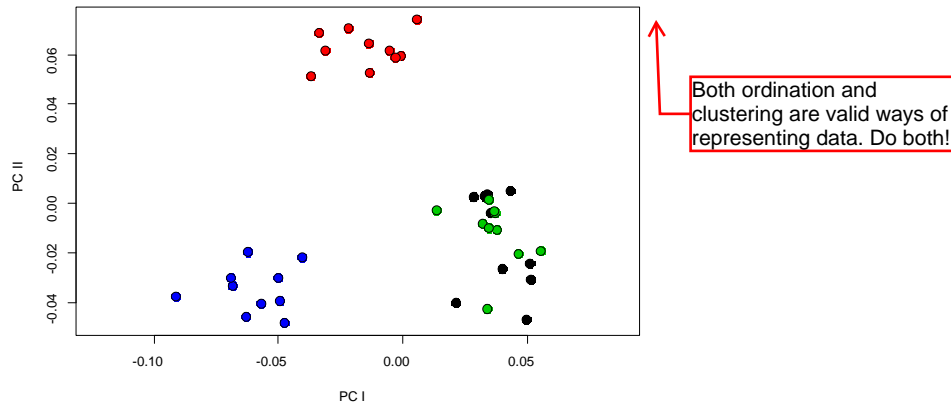


Complete linkage

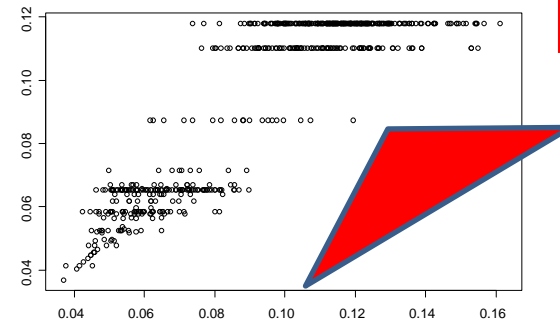
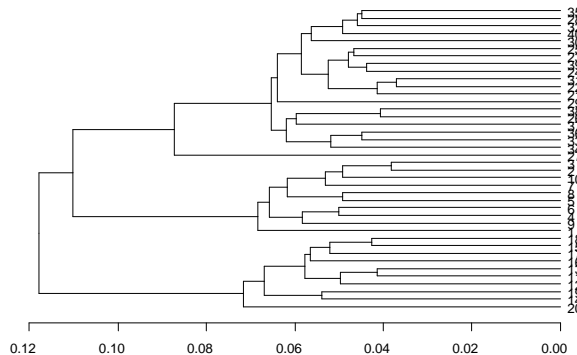
What are we doing?

Both ordination and clustering represent patterns in multivariate space

Ordination: large distances more accurately depicted, small distances less-so



Clustering: the opposite: small distances more accurately depicted



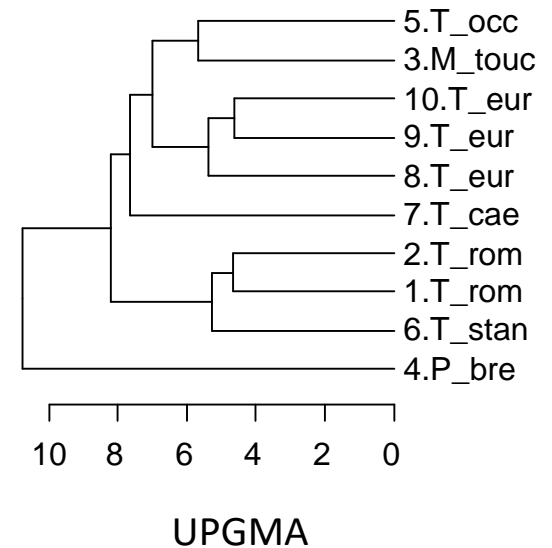
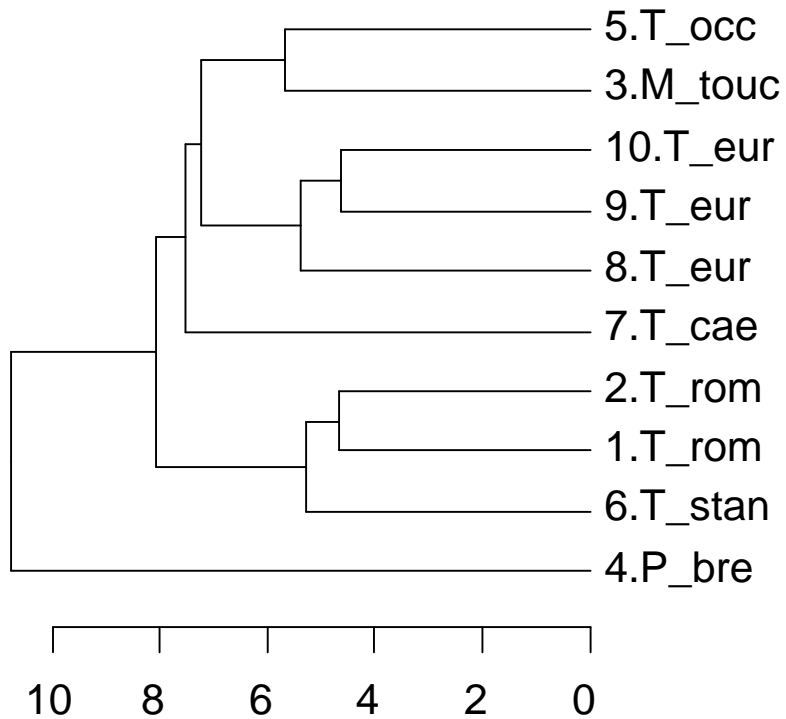
Want to see more clustering/ordination in molecular phylogenetics. Other applications?

Ordination/Clustering are complementary views of the high-dimensional space




## Weighted **P**air-**G**roup **M**ethod using **A**rithmetic **A**verages

- Same procedure as UPGMA, but averages are weighted
- New averages always found by dividing by 2 (regardless of # objects in cluster)
- Is equivalent to giving different weights to original similarity values



Very slight changes in this example,  
in terms of the length of inner  
branches



Unweighted Pair-Group  
Method using Centroids

Use centroids of clusters to join additional objects

Connect closest 2 objects in a cluster

Calculate centroid (from similarity scores\*) and find similarity scores from centroid to each object *not* in cluster

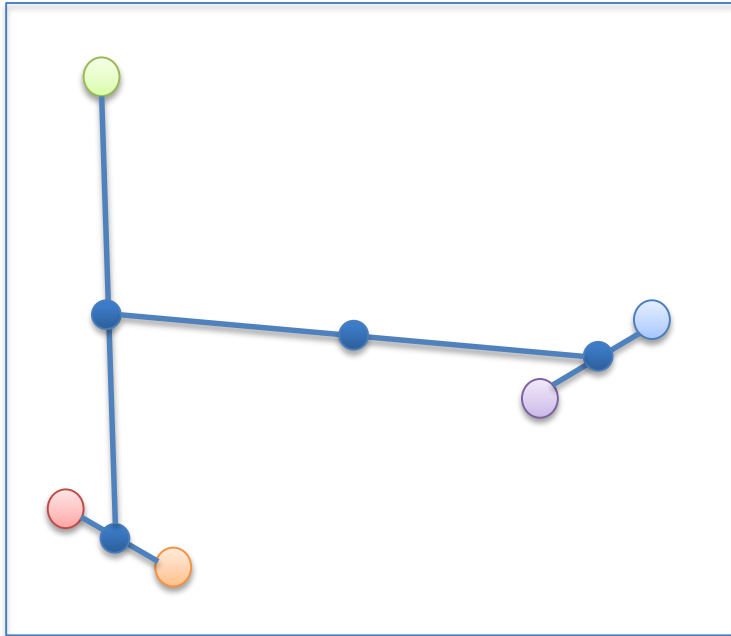
Replace original similarity scores with distances to centroid

Add new object and repeat

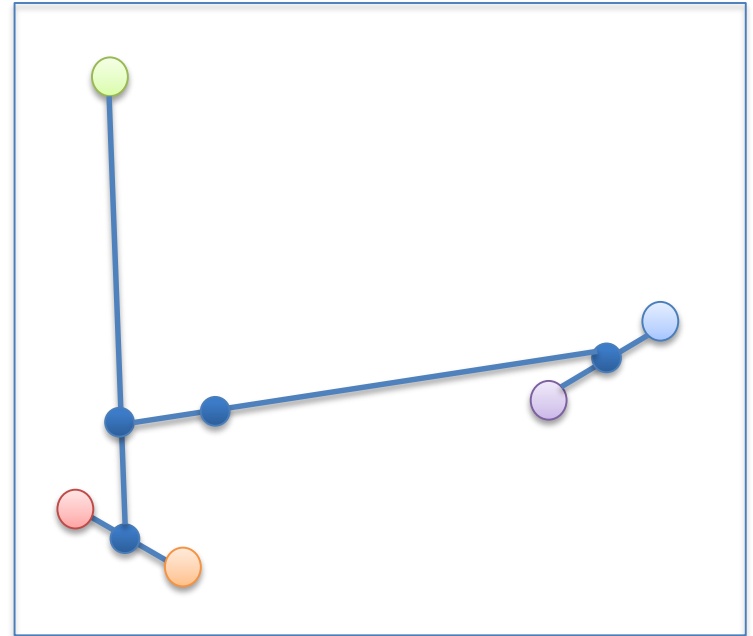
Centroid methods can find ‘reversals’ or negative branch lengths (e.g., when distance of 3<sup>rd</sup> object to centroid is smaller than distance between original pair)

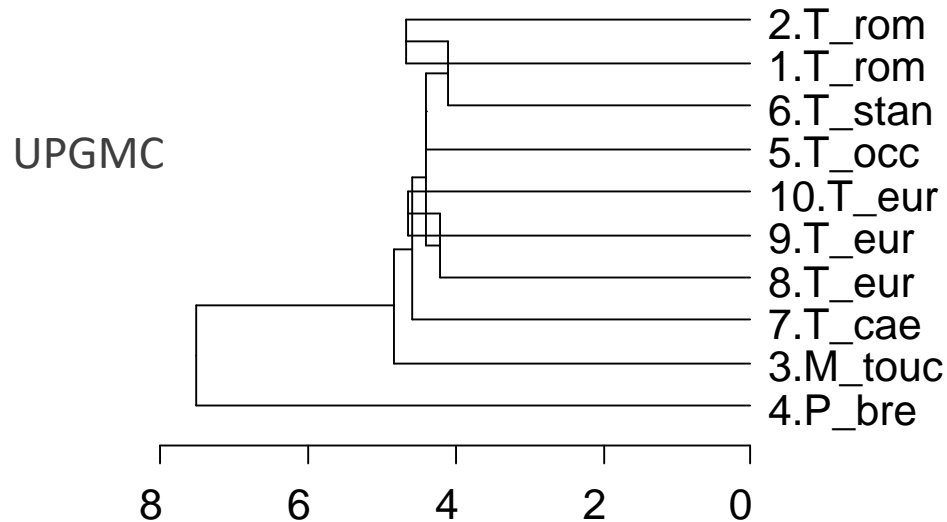
\* For details see Legendre and Legendre (1998). *Numerical Ecology*.

UPGMC

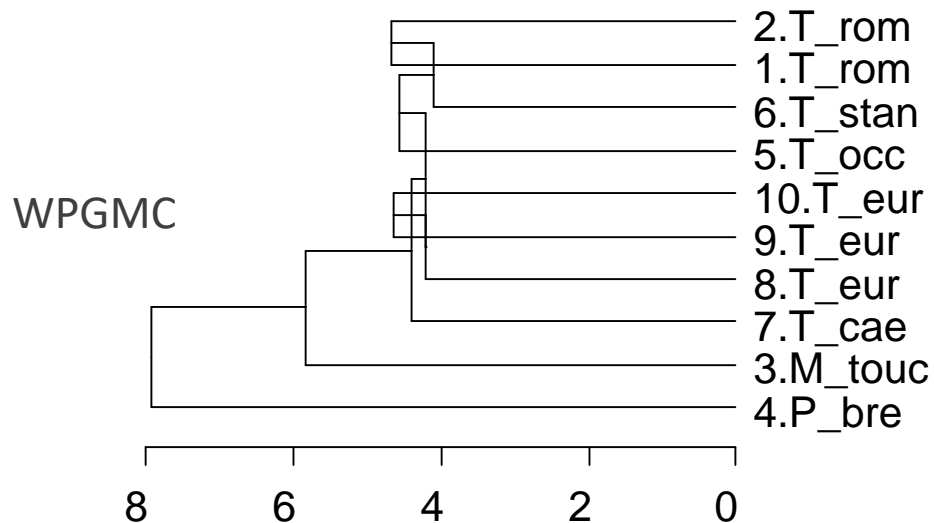


WPGMC





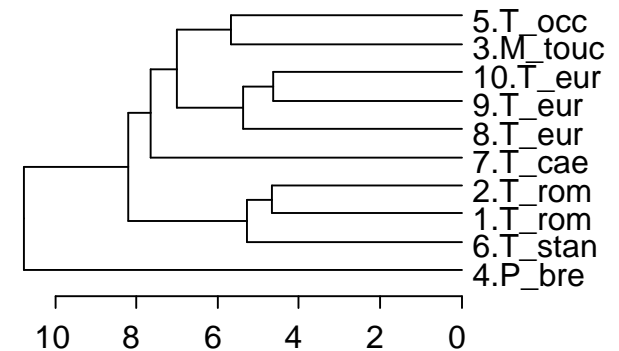
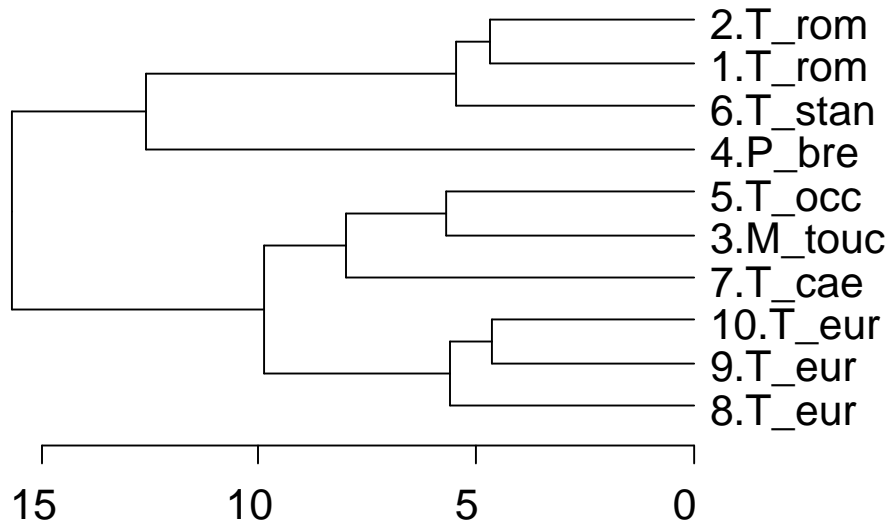
NOTE: Reversals often truncated to branch lengths = 0.0



\*These methods can find 'reversals' or negative branch lengths (e.g., when distance of 3<sup>rd</sup> object to centroid is smaller than distance between original pair)


Use cluster variance (TESS: total error sum of squares)

Add object that increases TESS the least



UPGMA

Many methods (more in Legendre and Legendre [1998])



UPGMA is the most widely used.

Method choice can influence both branch length and cluster composition.

Divisive hierarchical methods, probabilistic methods, most non-hierarchical clustering methods, plus other methods are covered in Legendre and Legendre (1998), in some detail (with references for further detail).

SAHN algorithms seek to place all  $n$  subjects into a hierarchical explanation of relatedness

**K-means** asserts a priori that an optimal number of clusters should be found and seeks an optimal solution to put  $n$  subjects into  $K$  non-hierarchical clusters.

Two main categories

Hierarchical: clusters are nested (higher-rank clusters contain lower-rank clusters)

**Non-hierarchical**: optimal strategy for finding  $K$  groups

Other ways to think of clustering method differences

**Sequential** (most methods) vs. Simultaneous

Agglomerative vs. **Divisive** (start with one versus start with all)

**Monothetic** vs. Polythetic (single descriptor for partitions vs. multiple descriptors) (Divisive methods only)

Probabilistic (assigns probability of homogeneity to putative within-group association matrices) vs. **Non-probabilistic** (most methods)



Partitions data space into groups that minimize Total Error Sums of Squares (TESS)

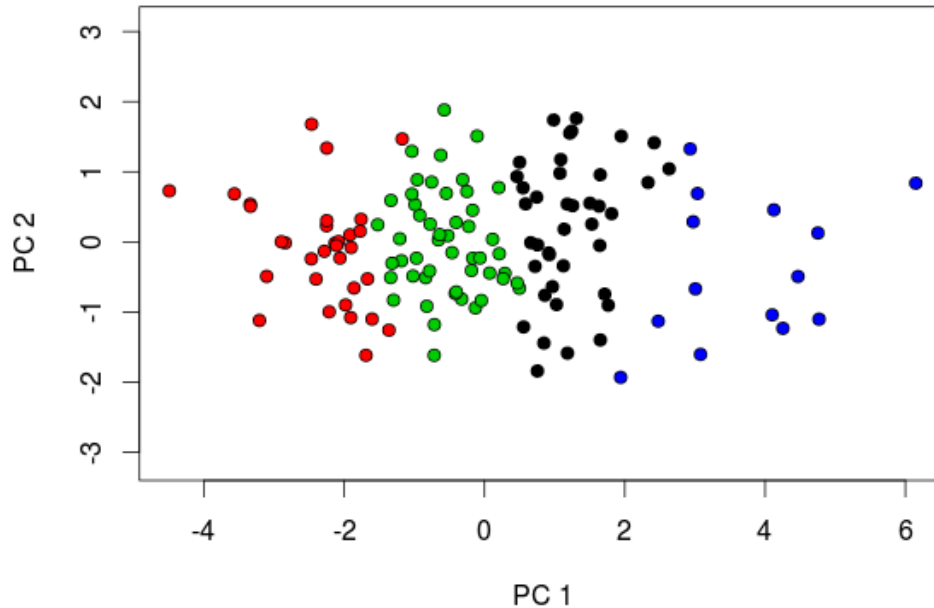
(Pooled within-group variation)

- Define # groups ( $K$ )
- Assign specimens to groups, calculate centroid, and TESS
- Repeat many times and choose solution with minimal TESS

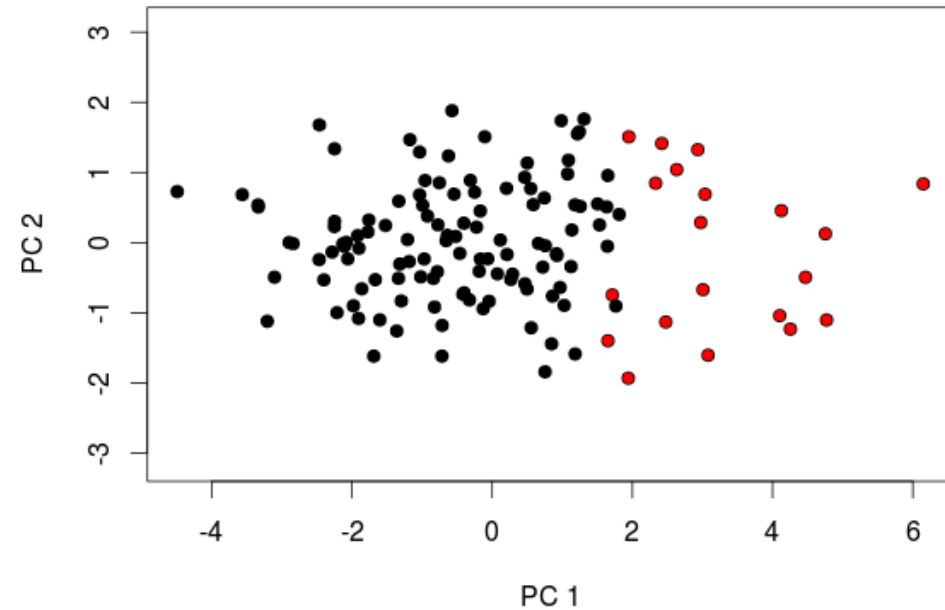
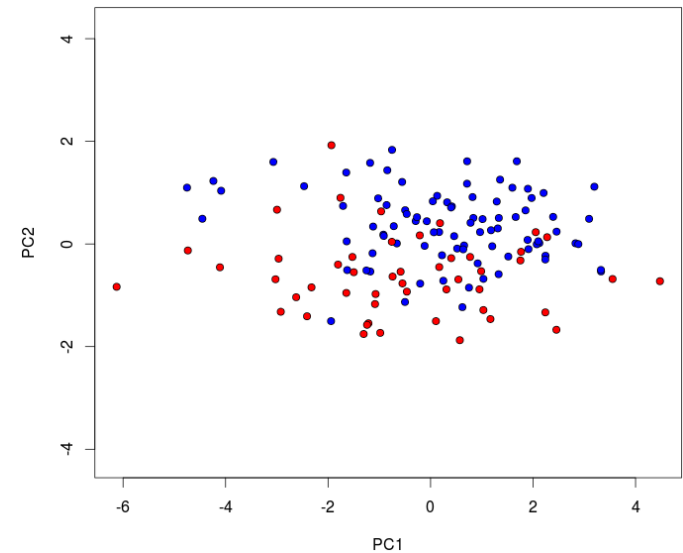
Entire process can be repeated for  $K = 2, 3, 4$  etc. to find optimal # groups

Does not yield dendrogram (not hierarchical); only yields # groups and group membership

```
> k.means4 <- kmeans(d, 4)
> k.means2 <- kmeans(d, 2)
```



```
> k.means4$totss
[1] 26664.84
> k.means4$tot.withinss
[1] 7695.846
> k.means2$totss
[1] 26664.84
> k.means2$tot.withinss
[1] 16066.81
```



- Clustering provides ‘connected’ view of object similarity
- Change of metric/distance measure may alter results
- VERY useful when combined with ordination (provide complementary views of data space)

Other methods exist: minimum spanning tree, flexible-link clustering, probabilistic clustering, evolutionary model-based ‘clustering’ (parsimony, neighbor-joining, etc.)

**Methods do NOT assume process!!** Careful in interpretation

One must make sure clustering method is commensurate with ordination method

E.g., *K*-means uses Euclidean distance to measure TESS. Corresponds to PC plot.