Cluster Analysis

Objective of Hierarchical Cluster Analysis

- Humans mentally organize information about individual objects by classifying the objects?
    - Classes summarize and condense information into manageable cogitative concepts.
    - This organization of the objects follows a purpose, which determines the features of the observations that we employ to describe and label the classes.
- These classes may even be follow a hierarchically structure. Example: Biological classification schemes.

  Note: In this hierarchical classification the individual circles are nested within each other.
- Partition the set of observations into clusters:
    - that are internally *homogeneous* (similar) with respect to some selected attributes and
    - that are as *heterogeneous* (dissimilar or distinct) as possible to the other clusters.

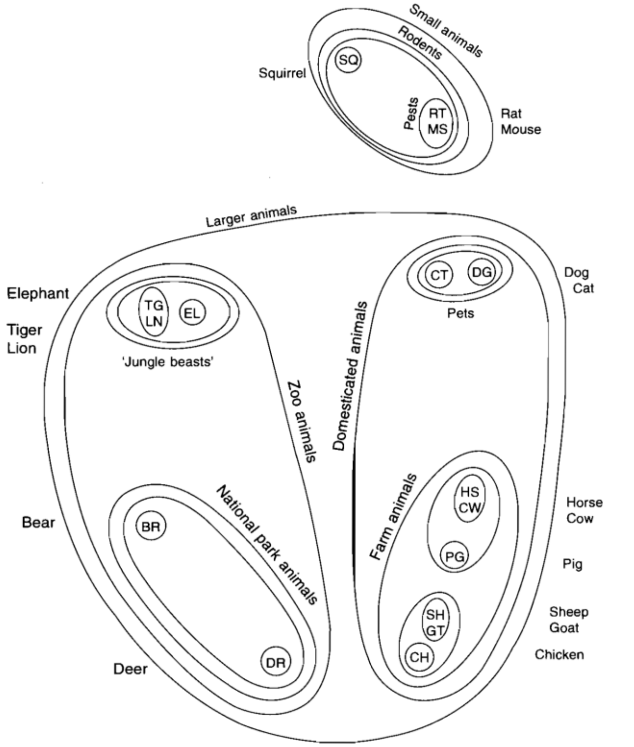
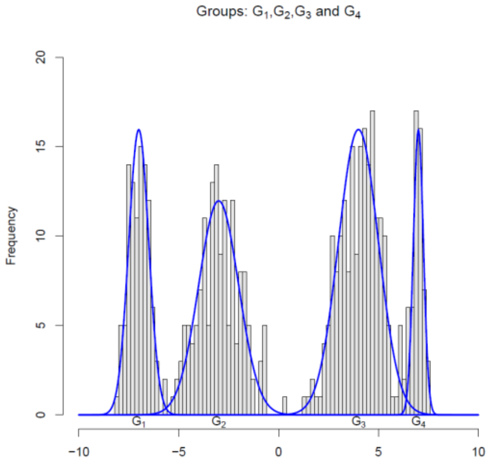
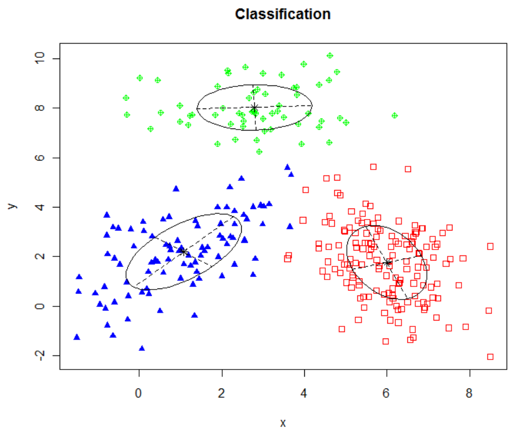
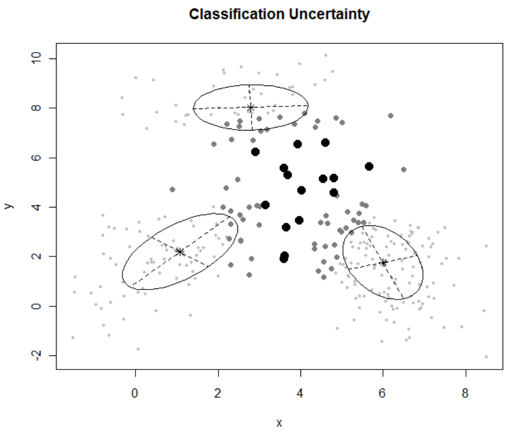


Figure 2.16 Mental associations among 16 kinds of animals. MDS solution plus hierarchical clustering solution. (Taken with permission from *Psychometrika*, 1974, **39**, 373–421.)

- Distinct groups of observations may already be evident in a set of univariate multimodal distributions or in bivariate plots:







- What happens if the overlap between the clusters shrinks?
- Note clusters may be intertwined (e.g., concave):

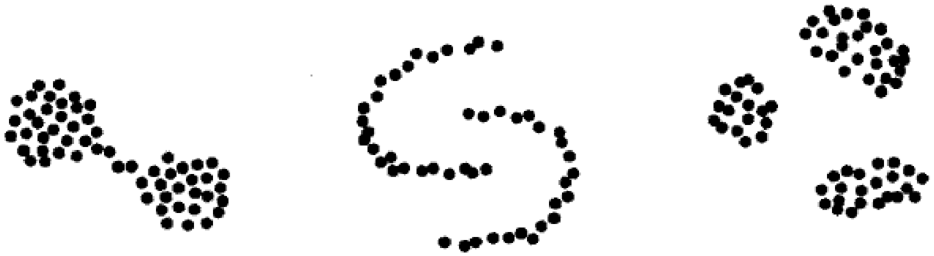
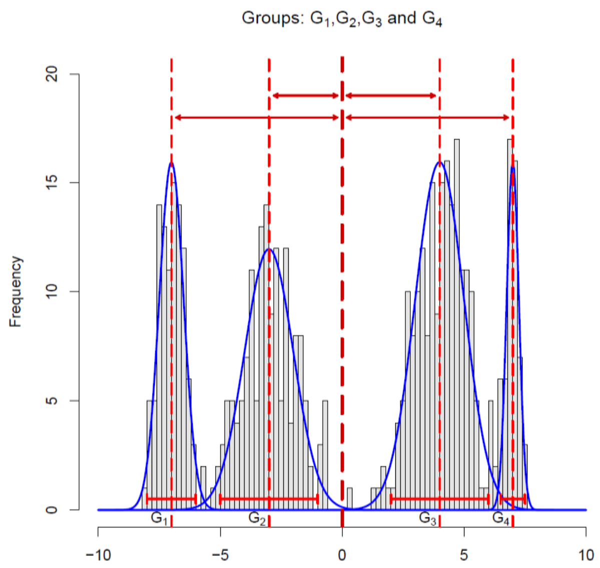


Figure 1.2 Clusters with internal cohesion and/or external isolation. (*Source:* Gordon, 1980.)

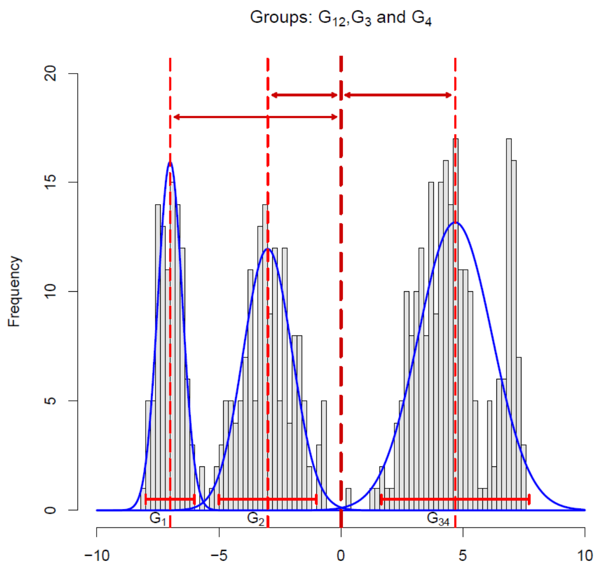
- Example: Four group a mixture of normal distributions:

Group:	n	$\bar{X}$	s
G_1	100 ($\pi = 0.2$)	-7	0.5
G_2	150 ($\pi = 0.3$)	-3	1
G_3	200 ($\pi = 0.4$)	4	1
G_4	50 ($\pi = 0.1$)	7	0.25

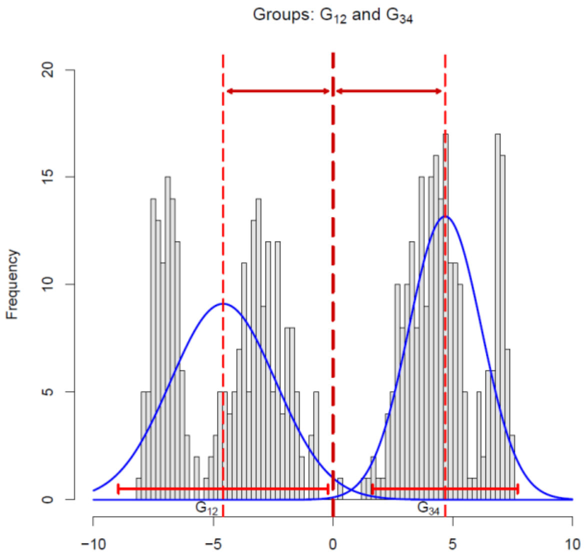
Agglomeration schedule based on between cluster distances:



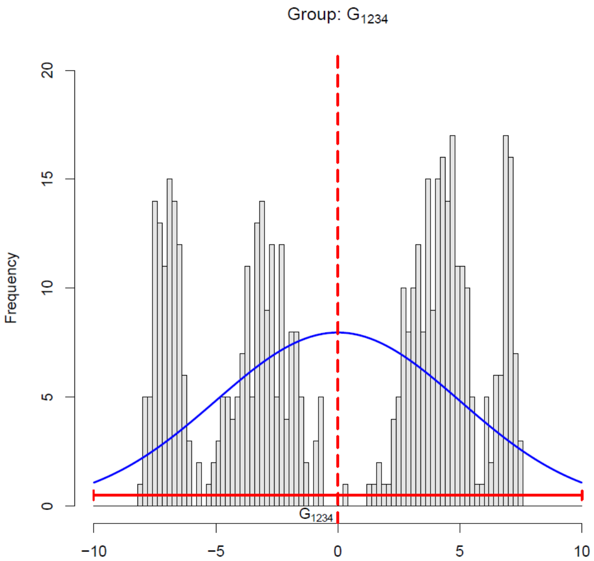
Initial clusters G_1, G_2, G_3 and G_4



Merger of G_3 with G_4 into G_{34}



Merger of G_1 and G_2 into G_{12}



Merger of G_{12} and G_{34} into G_{1234}

- Extreme ends of the classification spectrum: absolute homogeneity ⇔ full heterogeneity:
  - ***Maximum homogeneity*** within the clusters is achieved when each object establishes its individual ***single-object*** cluster.
  - ***Maximum heterogeneity*** within the set clusters is achieved once all objects are merged into just ***one*** global cluster.
  - Analogue to the variance decomposition in regression analysis TSS=ESS+RSS the total variability with regards to the global mean can be broken down into:

    TotalVariablility = SumOfVariablilityWithinClusters + VariablilityAmongClusters.

- Which decisions have to be made in order to conduct a cluster analysis:
    - Which ***set of attributes*** makes the objects ***distinct*** enough so they group into externally heterogeneous but internally homogeneous clusters?
    - What influence does ***rescaling*** or jointly ***transforming*** these variables have on the cluster analysis outcome?
    - How do we ***measure*** heterogeneity and homogeneity?
    - Into ***how many*** homogeneous clusters does the dataset break up before we combine incompatible groups of clusters?
    - Shall the ***classification scheme*** develop a hierarchical organization of the objects?
    - Which ***algorithm*** shall be employed to find clusters?
    - How do we ***interpret*** the individual clusters and the differences between the clusters? How do we ***numerically*** measure the classification and ***visually*** describe it?
- This lecture will focus on ***agglomerative hierarchical*** cluster analysis for ***quantitative*** variables:
    - ***Agglomerative*** clustering starts with the individual objects and successive merges them based on specific criteria into larger and larger cluster (or merges already generated clusters into larger joint clusters)
    - In contrast, ***divisive*** clustering starts with one super cluster that comprises all objects and subsequently breaks it up into sub-clusters.

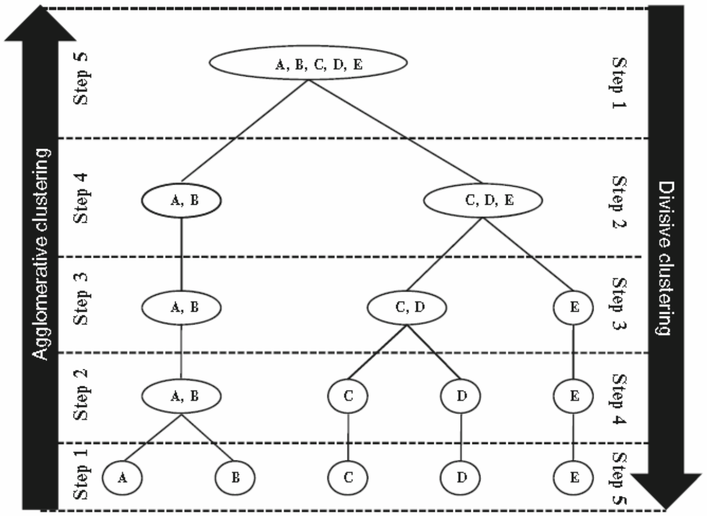


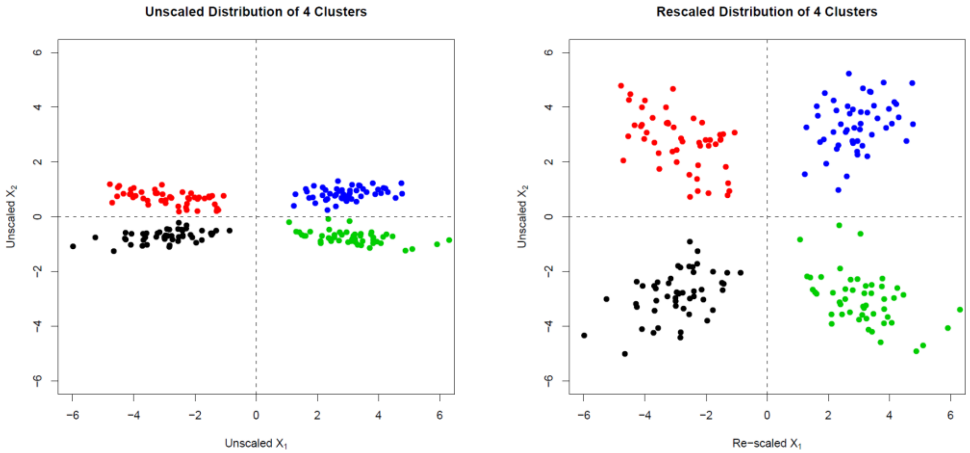
Fig. 9.3 Agglomerative and divisive clustering

  - Hierarchical cluster analysis allows identifying the ***feasible*** number of distinct clusters.

- ***Non-hierarchical*** methods like K-means allow ***reassigning objects*** that were already assigned to one cluster to another cluster in order to minimize the ***internal cluster heterogeneity***.
  For non-hierarchical methods the number of clusters needs to be given externally.
- Cluster analysis is an ***exploratory*** method mainly adopting a Q-mode (object rather than attribute focus) data analysis perspective

Scaling and transforming variables

- The underlying idea is to group observations which are similar to each other (i.e., short distance apart) in the measurement space together. Therefore the scaling of the variables becomes important.
- The ***spread*** of each variable (in particular if it is substantially different from variable to variable) or the ***correlation*** among variables can have a substantial impact on the distances among the objects:
- Scale effect: Variables with a ***large variance*** have a ***higher weight*** to the distance measures because the distances among the observations along the high variance axis are larger. Example:



  - Therefore, it is advisable to ***re-scale variables***

- The z-transformation $z(x_i) = \frac{x_i - \bar{x}}{s}$ (leads for $z(x_i)$ to a mean of zero and a standard deviation of one).
- The range transformation $r(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$ uses the range in the denominator. Its value range becomes $r(x) \in [0,1]$.

- Correlation effect: The underlying information in ***highly correlated variables*** is redundant. Therefore, using a set of redundant variables will have a higher impact on the calculated distances.
  - Example: Assume, for instance, that two variables X and Y are perfectly correlated whereas the variable Z is uncorrelated with the other variables.
    The distance between objects i and j in the attribute space becomes $d_{ij} = \underbrace{|x_i - x_j| + |y_i - y_j|}_{2\cdot|x_i - x_j| \text{ or } 2\cdot|y_i - y_j|} + |z_i - z_j|$.
    This gives extra weight to the joint information in variables X and Y.
  - The ***Mahalanobis*** distance can adjust for these high correlation effects among the variables.

Similarly, ***component scores*** from a preceding ***principal component analysis*** can be used because component scores are uncorrelated among each other.

- Non-metric features: How is a distance between classes that are based on nominal or ordinal scaled features calculated? How is a distance based on features with mixed measurement scales calculated?
- Similarity vs Dissimilarity: Homogeneity within clusters and heterogeneity between clusters can be evaluated either in terms of similarity $s_{\{C_k \cup C_l\}}$ or dissimilarity metrics $d_{\{C_k \cup C_l\}}$. There is an inverse relationship between the similarity and dissimilarity metrics.
- Triangle Equations: Does a dissimilarity metric need to satisfy the triangle equation:

$$d_{\{C_k \cup C_l\}} \leq d_{\{C_k \cup C_h\}} + d_{\{C_l \cup C_h\}} \, .$$

Hierarchical Agglomerative Clustering Algorithms

- Iterative steps an agglomerative hierarchical clustering algorithm:
  1. Variable selection: Decide which variables shall be used to describe the objects.
  2. Transformation: Decide if and how to scale the variables.
  3. Distance Metric: Decide on the distance metric between the objects to be clustered.
  4. Object Distance Matrix: Calculate a distance matrix (or similarity matrix) among the individual objects.

5. Update Rule: Decide how to measure the dissimilarity between [a] an individual object and an existing cluster of objects or [b] between one cluster of objects and another clusters of different objects.
6. Identification Step: Identify that [a] pair of objects, [b] the object-cluster pair, or the [c] cluster-cluster pair that is the most similar (smallest distance) and therefore contributes least to an increase in the internal cluster heterogeneity once merged.
7. Merger Step: From step 6 ***merge*** the identified pairs into a new cluster.
8. Dissimilarity Update Step:
   - Place the ancestors into the newly merged cluster.
   - Update the dissimilarity matrix by calculating the dissimilarity between the newly merged cluster and already existing clusters or individual objects.
   - The updated dissimilarity matrix will have one row and column less than the prior dissimilarity matrix.
9. Repeat the sequence: The identification step 6, the merger step 6 and the update step 7. Stop once all objects are merged into one just one super cluster or when the heterogeneity within the newly merged clusters exceeds a given threshold.

- Example: Single Linkage (Nearest-Neighbor) Method
  - Aggregation rule $d_{(R)(S)} = \min(d_{rs} \mid r \in R \text{ and } s \in S)$

- Start with a dissimilarity matrix, i.e., distance matrix, among all objects:

$$\mathbf{D}_0 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & \boxed{2} & 4 & 7 & 9 \\ 2 & & 0 & 8 & 9 & 8 \\ 3 & & & 0 & 3 & 7 \\ 4 & & & & 0 & 5 \\ 5 & & & & & 0 \end{array}$$

- <u>First iteration:</u> Merge objects 1 and 2 into a new cluster because they are the most similar.
  - Evaluate the distances new:

$$d_{(12)(3)} = \min(d_{13}, d_{23}) = 4$$

$$d_{(12)(4)} = \min(d_{14}, d_{24}) = 7$$

$$d_{(12)(5)} = \min(d_{15}, d_{25}) = 8$$

- Update the distance matrix and select the next pair:

$$\mathbf{D}_1 = \begin{array}{c} \\ (1,2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 4 & 7 & 8 \\ & 0 & \boxed{3} & 7 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix} \end{array}$$

- <u>Second iteration:</u> Merge objects 3 and 4 into a new cluster because they are the most similar.
  - Evaluate the distances new:

    $d_{(34)(12)} = \min(d_{(3)(12)}, d_{(4)(12)}) = 4$

    $d_{(34)(5)} = \min(d_{35}, d_{45}) = 5$
  - Update the distance matrix and select the next pair:

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ (34) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (34) & 5 \end{array} \\ \begin{pmatrix} 0 & \boxed{4} & 8 \\ & 0 & 5 \\ & & 0 \end{pmatrix} \end{array}$$

- <u>Third iteration:</u> Merge the clusters for the first iteration [object pair (12)] with the cluster from the second iteration [object pair (34)] because they are most similar
  - Evaluate the distances new

    $d_{((12)(34))(5)} = \min(d_{(12)(5)}, d_{(34)(5)}) = \min(8,5) = 5$
- <u>Last and fourth iteration:</u> merge object 5 into cluster (1,2,3,4)
- The aggregation history can be displayed in a ***dendrogram***

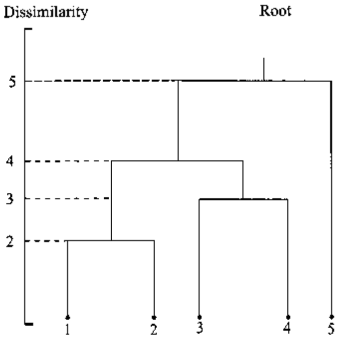


FIGURE 9.3.2. Dendogram for Single Link Example

The dissimilarity index denotes the value of the updated distances at which the merger happens.

- Different dissimilarity matrix updating rules for the interclass distances.
  Important: at the ***identification step*** always the ***shortest distance*** must be used.

Fig. 9.5 Single linkage

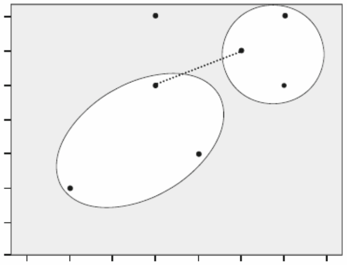


Fig. 9.7 Average linkage

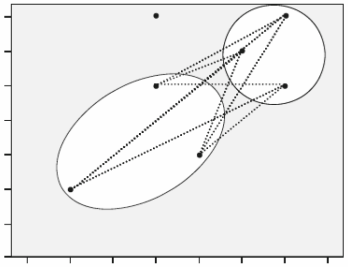


Fig. 9.6 Complete linkage

Use Max distance

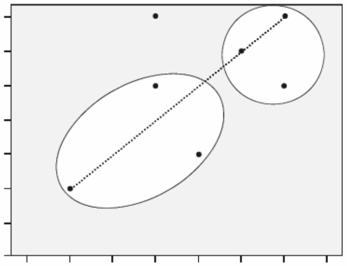
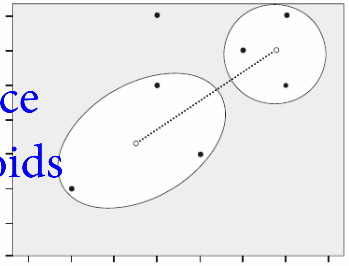
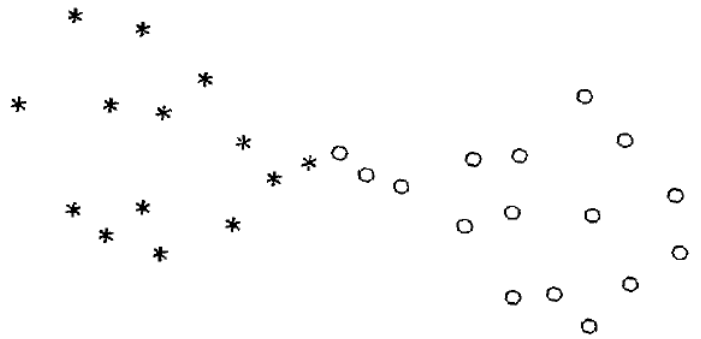


Fig. 9.8 Centroid

calculate the distance between two centroids



  - ○ *Single linkage*: Select the *smallest* distance from all objects in a cluster to all remaining objects not in that cluster (or objects within another cluster). It has the tendency to build large clusters at the initial stage and utilizes bridges between clusters:



- ○ *Complete linkage* updates with the *largest* distance. At the initial stages it has the tendency to build compact clusters. This is the opposite method to single linkage.
- ○ *Average linkage* distances between all objects in one cluster to all the objects in another clusters. Its properties are in-between single and complete linkage.
- ○ *Centroid linkage* requires that objects can be represented in the Euclidean space. Squared Euclidian distance are measured between cluster centroids:
  The new centroid of a merged cluster becomes ***weighted average*** of its predecessor clusters where the weights are the number of objects within a cluster (here shown for the j^{th} variable):

$$\bar{x}_{\{\mathcal{C}_k,\mathcal{C}_l\},j} = \frac{n_{\mathcal{C}_k} \cdot \bar{x}_{\mathcal{C}_k,j} + n_{\mathcal{C}_l} \cdot \bar{x}_{\mathcal{C}_l,j}}{n_{\mathcal{C}_k} + n_{\mathcal{C}_l}}$$

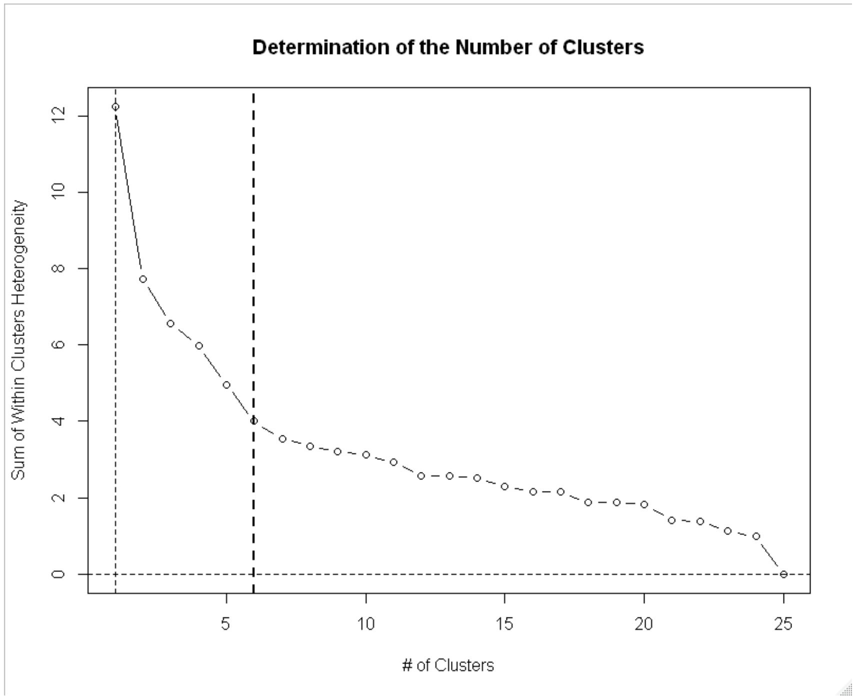
  - *Median linkage* same as the centroid linkage but without weighting the centroids by their number of objects with a cluster.
  - *Ward's method* requires that the object's features can be represented in the Euclidean space. It merges those two clusters whose joint spread around the ***new cluster centroid increases the least***. It therefore minimizes the increase in the heterogeneity within clusters.

    Squared Euclidian distances among the objects are a good input choice to Ward's method because the within cluster spread and the distances are measured in squared units

- In hierarchical cluster analysis, once an object is assigned to a group it indefinitely remains within that group.
- The within group dissimilarity (that is, the heterogeneity) is ***usually*** increasing monotonically at each step of the clustering algorithm. I.e., there are no ***inversions*** in the dendrogram.
  The *centroid* and the *median* linkage methods can potentially generate inversions.

Stopping rules for the number of clusters and cluster description

- Usually one stops joining existing clusters, when the sum of within clusters heterogeneity stops increasing only rapidly.
- A scree-plot can be used to subjectively identify the critical number of clusters:



- Clusters can be described by the means of their features within their clusters and by their within cluster spreads (e.g., standard deviations)
- A clear indication of clustering is when object features that were initially withheld from the cluster analysis, indicate the underlying grouping of the objects.

Clustering under Spatial Constraints

- In spatial analysis, the ***aggregation of areas*** into contiguous ***regions*** has to obey specific constraints such as the aggregated areas being adjacent.
- If multiple, perhaps conflicting ***constraints*** (e.g., racial composition and critical population bounds), need to be satisfied while ***minimizing the objective*** function of within region heterogeneity, then ***spatial optimization techniques*** are required to obtain a satisfactory regionalization.
  - Homogeneity constraint: The aggregated areas should constitute homogenous regions. This criterion is inherited from standard cluster analysis.
    - That is, with regards to specific attributes the variability within a region should be minimal.
    - E.g., census tracts should comprise of similar population sub-groups.
    - Classified remotely sensed pixels need to belong to the same land-use category.
  - Contiguity constraint: The aggregated areas should be contiguous. That is, they can result only into ***one region*** and not several unconnected sub-region spread over the map.
    - ***Redistricting*** of electoral districts has additional requirements, e.g., the resulting district needs to be ***compact***.
  - Capacity constraint: Particular external variables are only allowed to vary within specific bounds. For examples:

  - an electoral district needs to reflect a specific number of residents and/or mix of racial groups;
  - the population of a census tract (an aggregate of block groups which in turn are aggregates of blocks) should have between 4000-8000 inhabitants;
  - the number of students and their racial composition within the catchment area of school need to vary within specific bounds.
- Example: Clustering the 50 US States with regards to population density perhaps limiting the region (cluster) size to maximal 8 individual states.

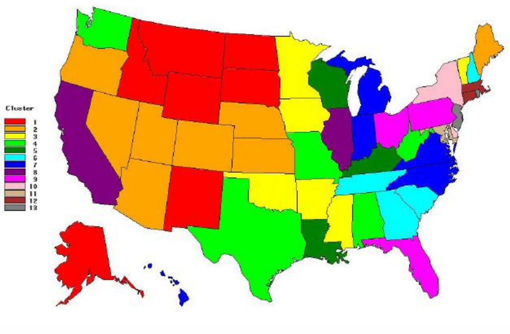


Figure 4: Example clustering without contiguity constraint.

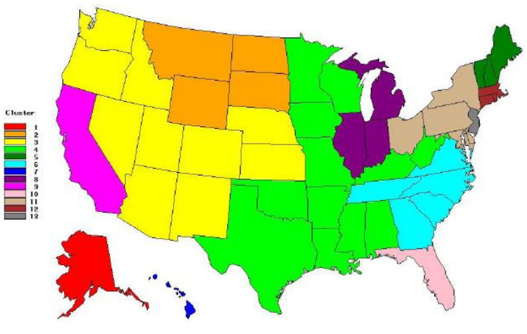


Figure 2: Example clustering with contiguity constraint.

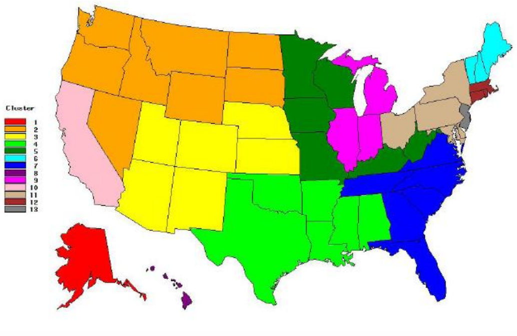


Figure 3: Example clustering with contiguity constraint and cluster size limit.