Cluster Analysis

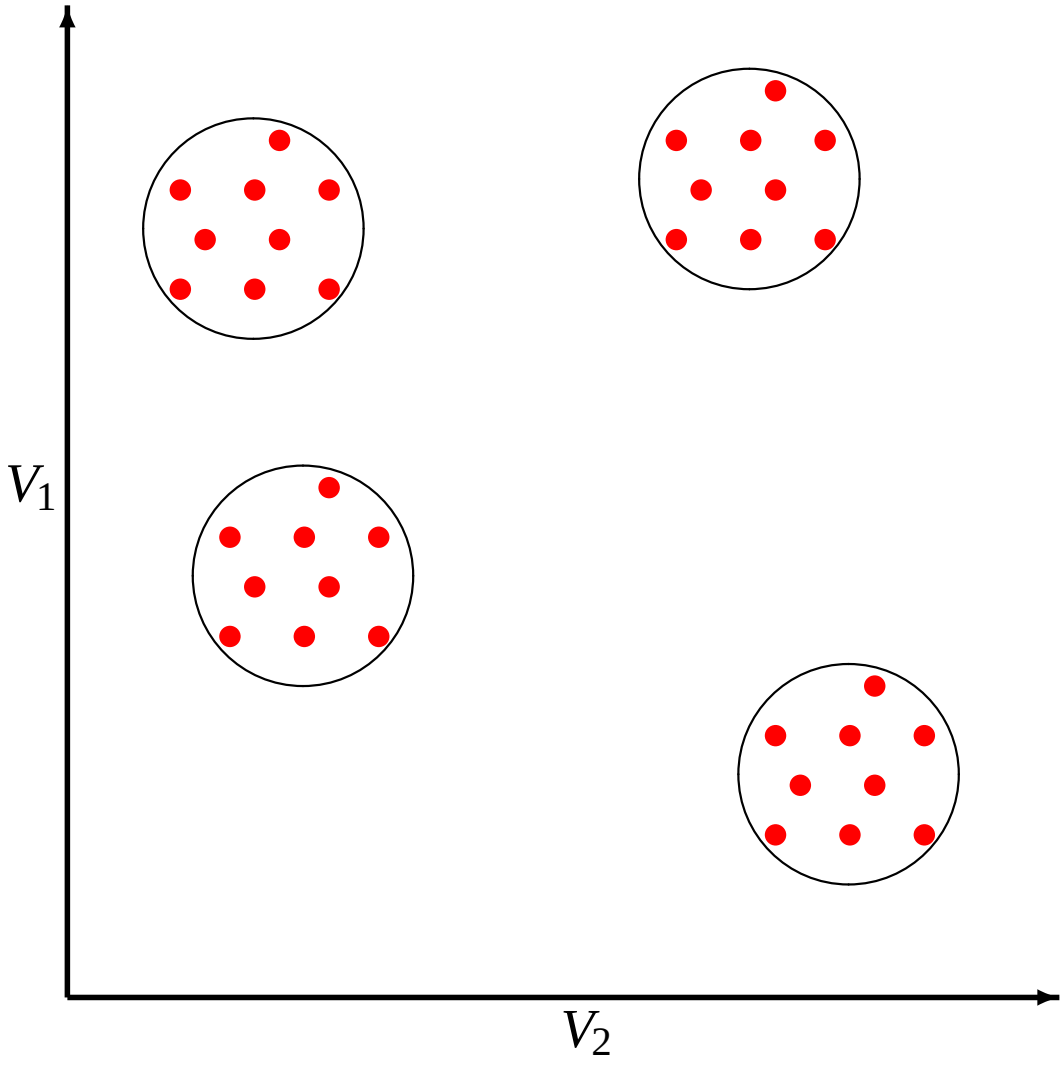
Contents:

1. Introduction and Example
2. Hierarchical methods.
3. Non-hierarchical methods.
4. Examples.

Cluster analysis

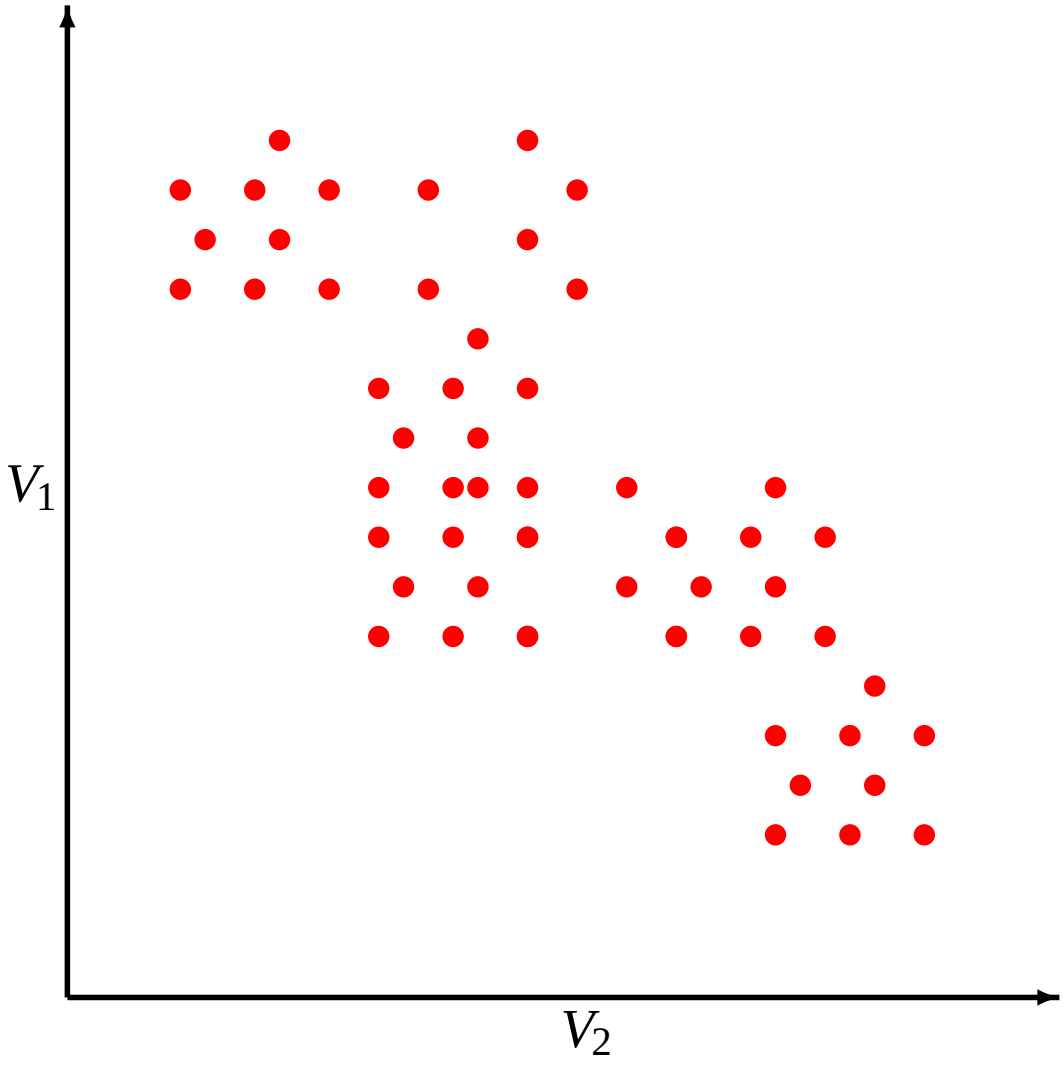
- The primary objective of cluster analysis is to classify cases into relatively homogeneous groups based on the set of variables considered. The cases in the group are relatively similar in terms of these variables and different from individuals in other groups.
- The resulting clusters of individuals should then exhibit hight internal (within-cluster) homogeneity and hight external (between-cluster) heterogeneity. Thus, if the classification is successful, the individuals within clusters will be close together when plotted geometrically, and different clusters will be far apart.
- Cluster analysis is also called *classification analysis* or *numerical taxonomy*.
- Here we are concerned with clustering procedures that assign each individuals to one only cluster.
- In cluster analysis there is no a priory information about the group or cluster membership for any of the individuals. Groups or clusters are suggested by the data and are not defined a priori.

An ideal clustering situation



- This is an ideal clustering situation in which the clusters are distinctly separated on two variables (V_1: quality of product, and V_2: price of product). Note that the consumers falls into one cluster and there are no overlapping areas.

A practical clustering situation



- This is a clustering situation which is more likely to be encountered in practice. The boundaries of the clusters are not clear cut, and the classification of some consumers is not obvious, as many of them could be grouped into one cluster or another.

Assumption in Cluster analysis

- Cluster analysis is not a statistical inference technique where parameters from a sample are assessed as possibly being representative of a population. Instead, cluster analysis is an objective methodology for quantifying the structural characteristics of a set of observations.

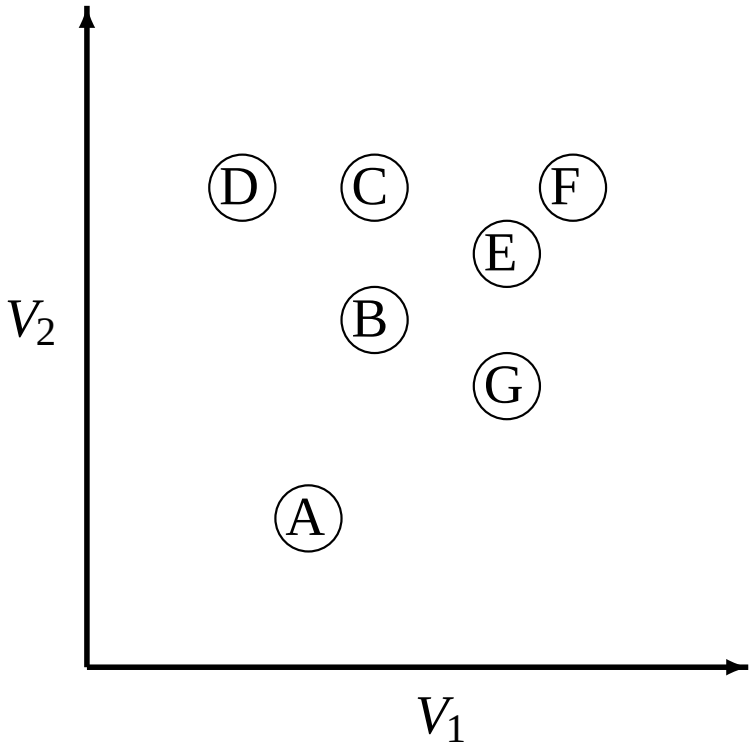
- As such it has strong mathematical properties but not statistical foundations. The requirements of normality, linearity and homoscedasticity that were so important in other techniques really have little bearing on cluster analysis. The researcher must focus, however, on the representativeness of the sample.

- The cluster analysis is only good as the representativeness of the sample. Therefore all efforts should be taken to ensure that the sample is representative and the results are generalizable to the population of interest.

How cluster analysis work: A bivariate example

- Suppose a marketing researcher wishes to determine market segments in a small community based on their patterns of loyalty to brands and stores. A small sample of seven respondents is selected. Two measures of loyalty were measured for each respondent on a 0-to-10 scale: V_1 (store loyalty) and V_2 (brand loyalty).

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4



- The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups.

- How do we measure similarity_?

  We require a method of simultaneously comparing observations on the two clustering variables (V_1 and V_2). E.g. Distance.

- How do we form clusters_?

  No matter how similarity is measured, the procedure must group the most close observations that are most similar into a cluster. That is, the procedure must determine the group membership of each observation.

- How many clusters do we form_?

  Any number of rules might be used. However, the fundamental task is to asses the *average* similarity across clusters such that:
  as the average increases, the cluster become less similar.

Trade-off: fewer clusters versus less homogeneity.

Simple structure is striving toward parsimony, is reflected in a few clusters as possible. Yet as the number of clusters decreases the homogeneity within the clusters necessarily decreases. Thus a balance must be made between defining the most basic structure (fewer clusters) that achieves the necessary level of similarity within the clusters.

Proximity Matrix

- Similarity will be measured according to the Euclidean distance (straight-line) between each pair of observations. The distance between the two points (x_1, y_1) and (x_2, y_2) is given by:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

  The smaller distance indicate greater similarity and bigger distance indicates most dissimilarity.

- To measure the homogeneity between clusters we will use the Error Sum of Squares (ESS).

Error Sum of Squares (ESS)

The ESS is the sum of the *square distances* of the observations from the mean of the clusters.

E.g. $E=(6,6)$, $F=(7,7)$ and $G=(6,4)$. The mean of E and F is given by:

$$M=(\frac{6+7}{2},\frac{6+7}{2})=(6.5,6.5).$$

The ESS of $\{E,F\}$ is given by:

$$\begin{aligned}\text{ESS}_{\{E,F\}} &= \Big((6-6.5)^2+(6-6.5)^2\Big)\\ &\quad+\Big((7-6.5)^2+(7-6.5)^2\Big)\Big)\\ &=0.5^2+0.5^2+0.5^2+0.5^2\\ &=1.\end{aligned}$$

The mean of E, F and G is given by:

$$M=(\frac{6+7+6}{3},\frac{6+7+4}{3})=(\frac{19}{3},\frac{17}{3}).$$

Also:

$$\text{ESS}_{\{E,F,G\}}=5\frac{1}{3}.$$

	Observations						
Observe.	A	B	C	D	E	F	G
A	0						
B	3.16	0					
C	5.10	2.00	0				
D	5.10	2.83	2.00	0			
E	5.00	2.27	2.24	4.12	0		
F	6.40	3.61	3.00	5.00	1.41	0	
G	3.61	2.24	3.61	5.00	2.00	3.16	0

- There are 21 distances (excluding the zero ones). Generally for n observations there are $n(n-1)/2$ distances.

- There are various methods for forming clusters. Here the agglomerative hierarchical procedure is considered.

  For the current example the following procedure is used:

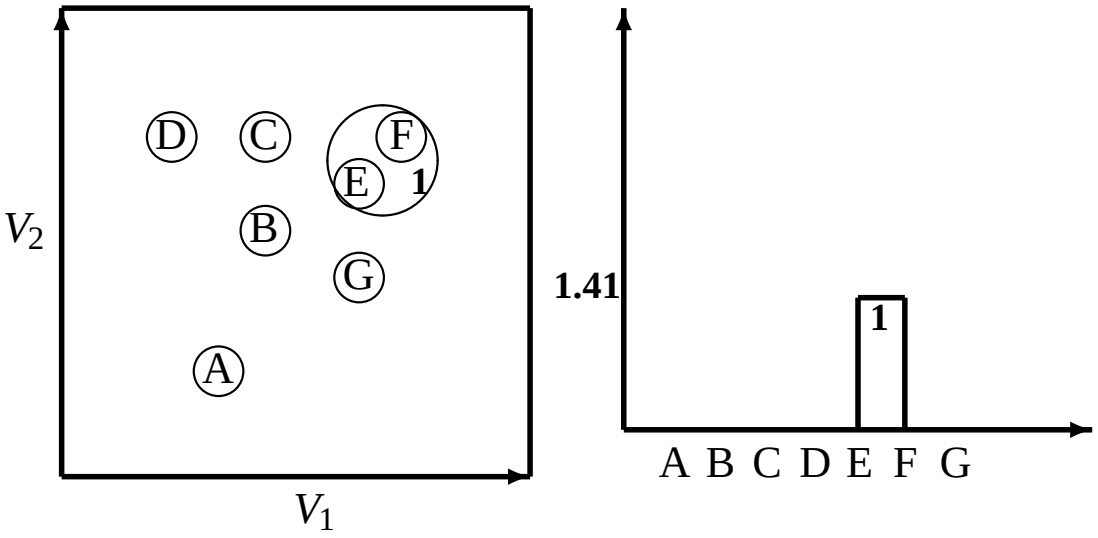
  1. Let each observation denote a cluster.
  2. Identify the two most similar (closest) observations not already in the same cluster and combine their clusters.
  3. Repeat step 2 until all observations are in the same cluster.

  Note that various methods can be used in joining two clusters together.

Step 1

Observ.	Observations						
	A	B	C	D	E	F	G
A	0						
B	3.16	0					
C	5.10	2.00	0				
D	5.10	2.83	2.00	0			
E	5.00	2.27	2.24	4.12	0		
F	6.40	3.61	3.00	5.00	1.41	0	
G	3.61	2.24	3.61	5.00	2.00	3.16	0

Identifies the two closest observations (E and F) and combines them into a cluster.



The $\text{ESS}_{\{E,F\}} = 1$.

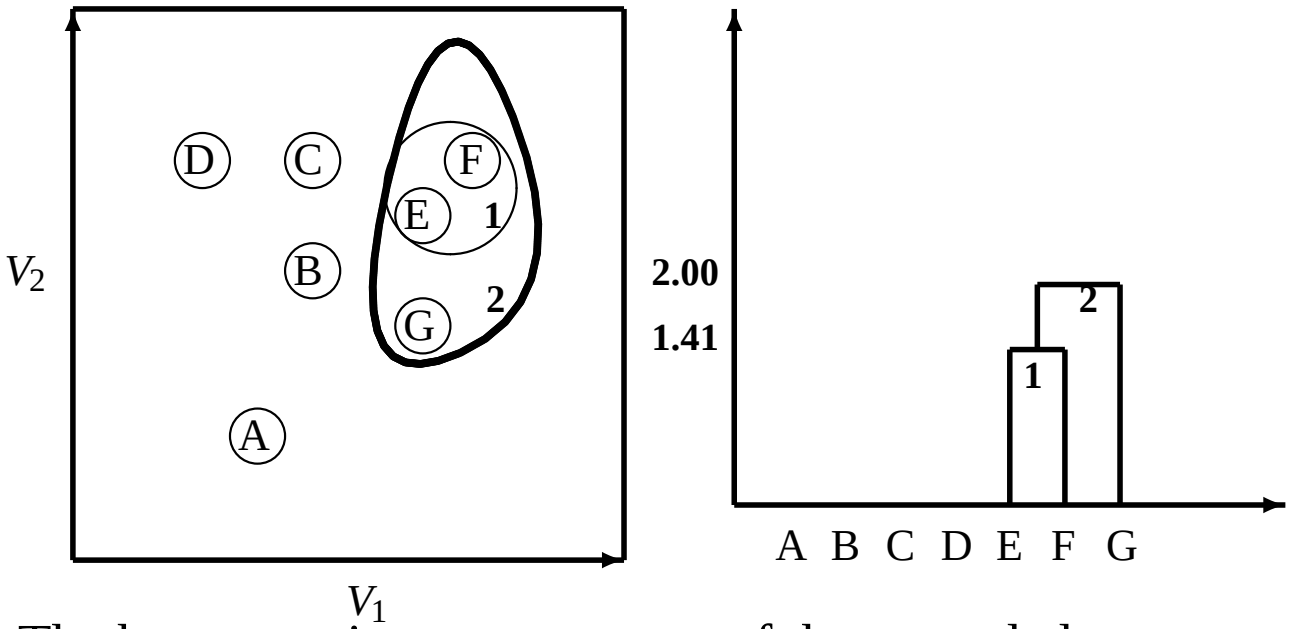
Thus, the total ESS of this clustering solution is:

$$\begin{aligned}\text{ESS} &= \text{ESS}_{\{A\}} + \text{ESS}_{\{B\}} + \text{ESS}_{\{c\}} + \text{ESS}_{\{D\}} + \text{ESS}_{\{E,F\}} + \text{ESS}_{\{G\}} \\ &= 0 + 0 + 0 + 0 + 1 + 0 = 1.\end{aligned}$$

Step 2

Observ.	Observations					
	A	B	C	D	$\{E,F\}$	G
A	0					
B	3.16	0				
C	5.10	2.00	0			
D	5.10	2.27	2.00	0		
$\{E,F\}$	5.00	2.34	2.24	4.12	0	
G	3.61	2.24	3.61	5.00	2.00	0

Finds the next closest pairs of observations (clusters): $G-\{E,F\}$, $C-D$ and $B-C$ all have distances 2.00. We chose to combine the clusters $\{G\}$ and $\{E,F\}$ to give the new cluster $\{E,F,G\}$.

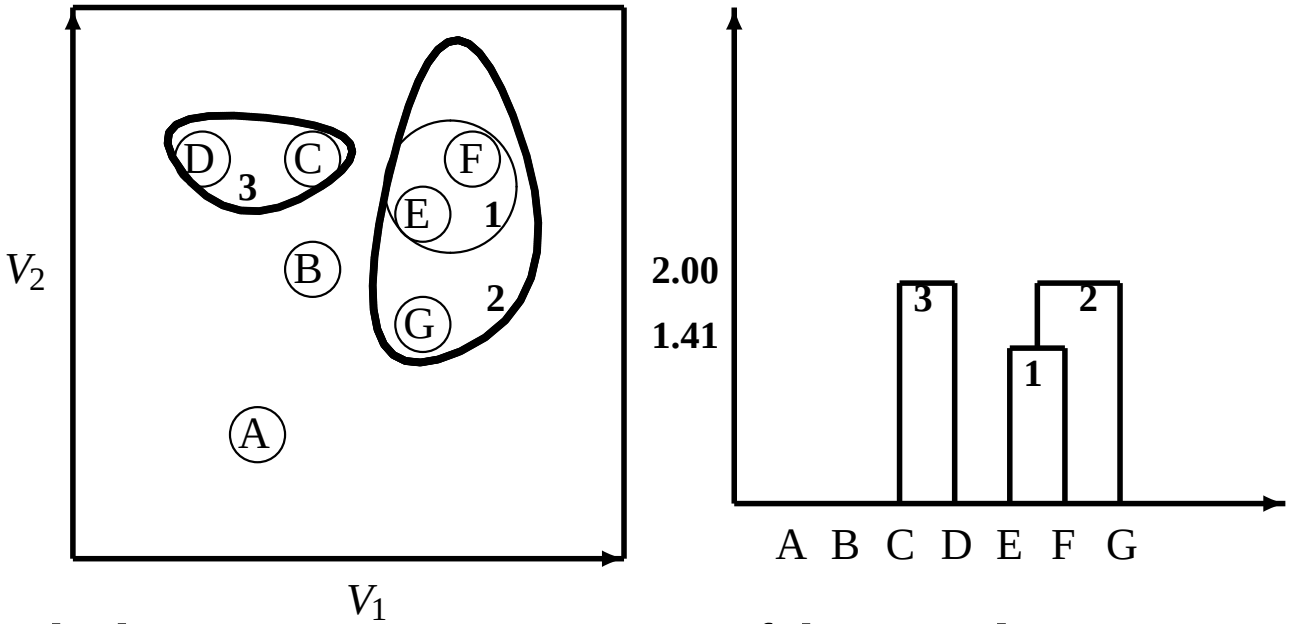


The homogeneity measurement of the second cluster $\{E,F,G\}$ is the $\text{ESS}_{\{E,F,G\}} = 5.33$.

Step 3

Observ.	Observations				
	A	B	C	D	$\{E,F,G\}$
A	0				
B	3.16	0			
C	5.10	2.00	0		
D	5.00	2.83	2.00	0	
$\{E,F,G\}$	3.61	2.24	2.24	4.12	0

Finds the next closest pairs of observations (clusters): $C-D$ and $B-C$ all have distances 2.00. Let combine the observations C and D into the cluster $\{C,D\}$.

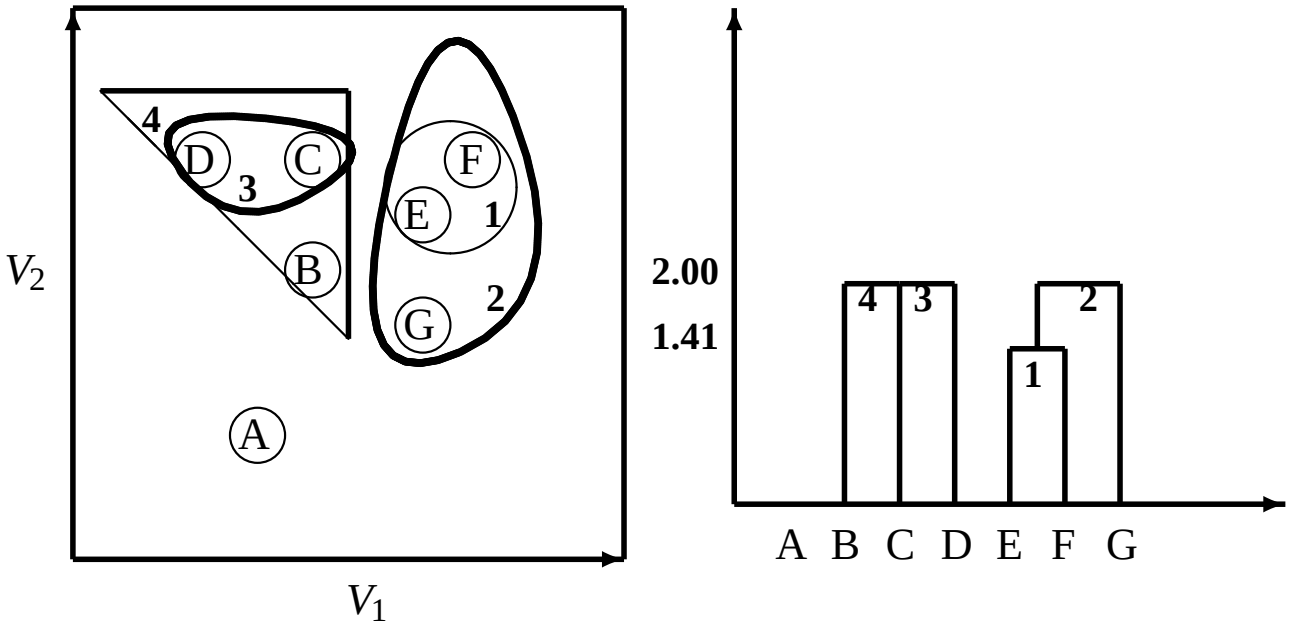


The homogeneity measurement of the new clusters $\{E,F,G\}$ and $\{C,D\}$ is the sum of their EES. That is, $\text{EES} = \text{EES}_{\{E,F,G\}} + \text{EES}_{\{C,D\}} = 5.33 + 2 = 7.33.$

Step 4

Observ.	Observations			
	A	B	$\{C,D\}$	$\{E,F,G\}$
A	0			
B	3.16	0		
$\{C,D\}$	5.00	2.00	0	
$\{E,F,G\}$	3.61	2.24	2.24	0

Finds the next closest pairs of observations (clusters): It combines B and $\{C,D\}$ into a new cluster $\{B,C,D\}$.



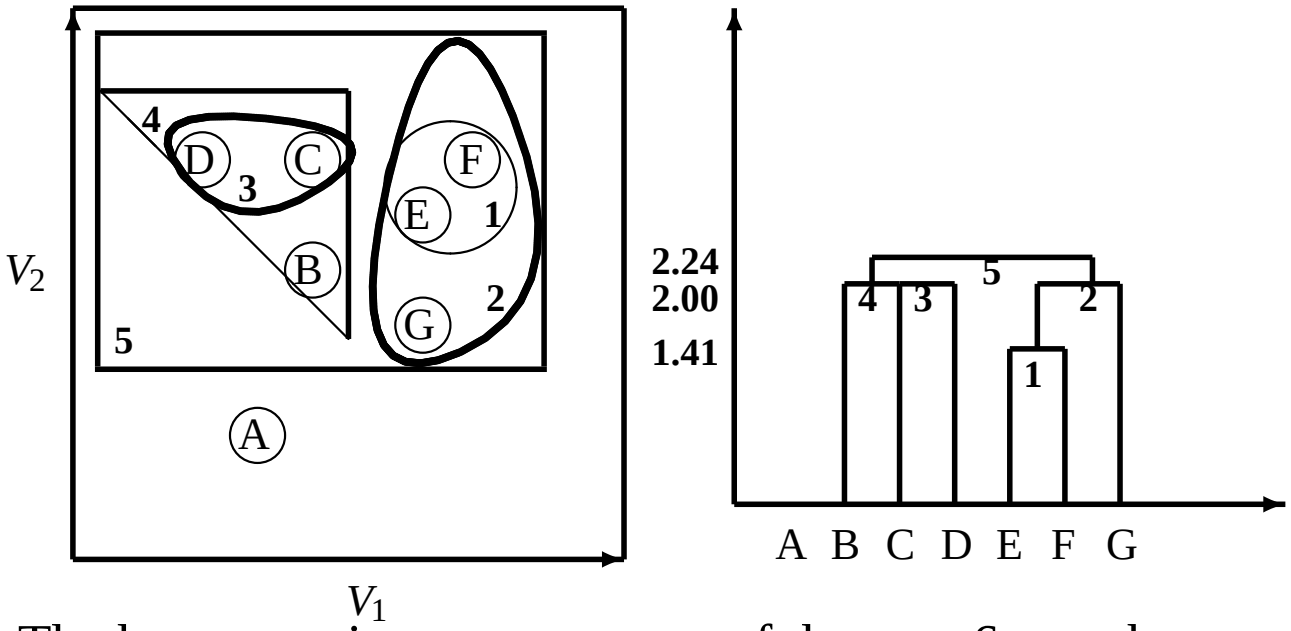
The homogeneity measurement of the new clusters $\{E,F,G\}$ and $\{B,C,D\}$ is given by

$$\text{EES} = \text{EES}_{\{E,F,G\}} + \text{EES}_{\{B,C,D\}} = 5.33 + 5.33 = 10.67.$$

Step 5

Observ.	Observations		
	A	$\{B,C,D\}$	$\{E,F,G\}$
A	0		
$\{B,C,D\}$	3.16	0	
$\{E,F,G\}$	3.61	2.24	0

The smallest distance is 2.24. Thus, it combines the two 3-member clusters $\{B,C,D\}$ and $\{E,F,G\}$ to give $\{B,C,D,E,F,G\}$.



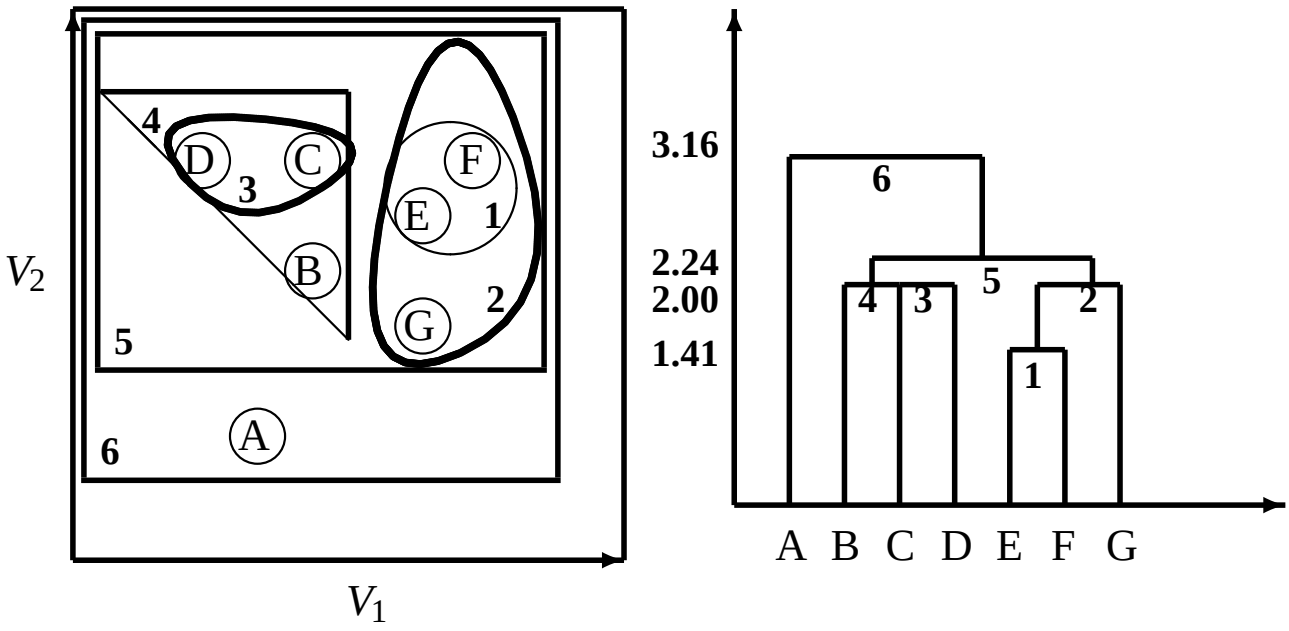
The homogeneity measurement of the new 6-member cluster $\{B,C,D,E,F,G\}$ is given by:

$$\text{EES} = \text{EES}_{\{B,C,D,E,F,G\}} = 24.83.$$

Step 6

Observ.	Observations	
	A	$\{B,C,D,E,F,G\}$
A	0	
$\{B,C,D,E,F,G\}$	3.16	0

The final step joins all the observation in a single cluster, i.e. $\{A,B,C,D,E,F,G\}$.

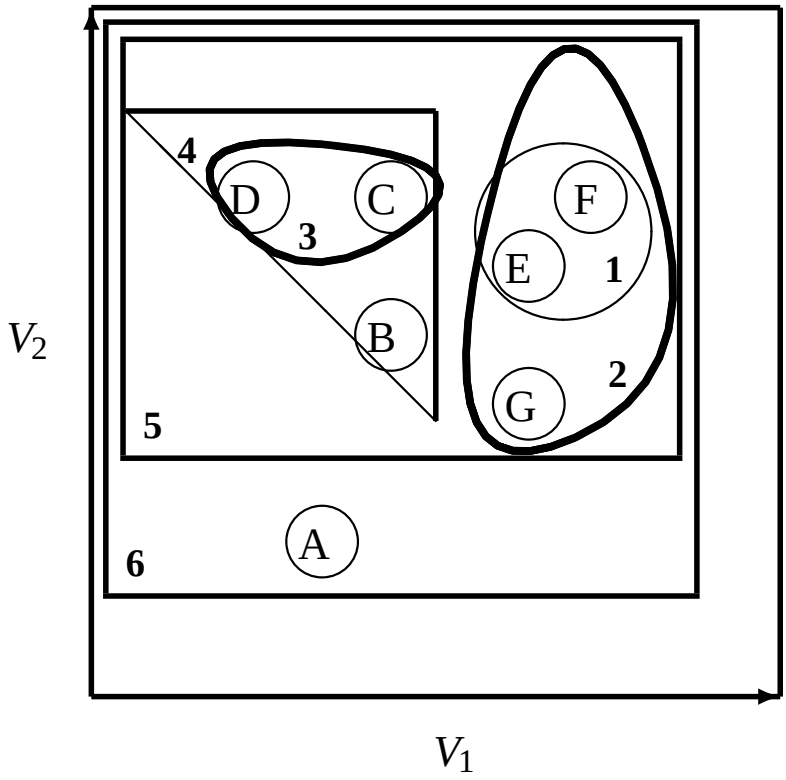


The homogeneity measurement of the final 7-member cluster $\{A,B,C,D,E,F,G\}$ is

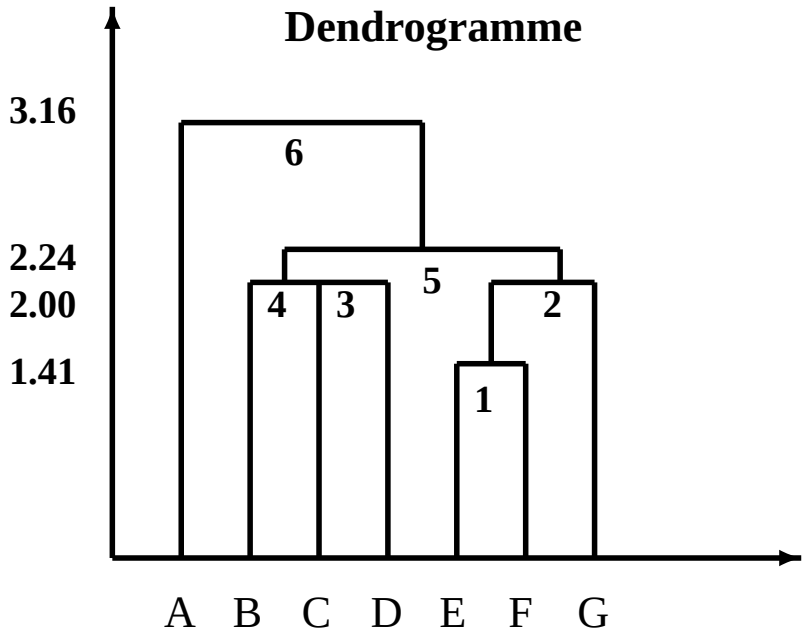
$$\text{EES} = \text{EES}_{\{A,B,C,D,E,F,G\}} = 41.43.$$

Nested grouping and Dendrogram

Nested grouping



Dendrogramme



Agglomeration Process			Cluster Solution		
Step	Minimum Distance between Unclustered Observations*	Observ. Pair	Cluster Membership	# of Clusters	Overall Similarity Measure (EES)
0	Initial Solution		$\{A\}\{B\}\{C\}\{D\}$ $\{E\}\{F\}\{G\}$	7	0
1	1.41	$E-F$	$\{A\}\{B\}\{C\}\{D\}$ $\{E,F\}\{G\}$	6	1
2	2.00	$E-G$	$\{A\}\{B\}\{C\}\{D\}$ $\{E,F,G\}$	5	5.33
3	2.00	$C-D$	$\{A\}\{B,C\}\{D\}$ $\{E,F,G\}$	4	7.33
4	2.00	$B-C$	$\{A\}\{B,C,D\}$ $\{E,F,G\}$	3	10.67
5	2.24	$B-E$	$\{A\}\{B,C,D,E,F,G\}$	2	24.83
6	3.16	$A-B$	$\{A,B,C,D,E,F,G\}$	1	41.43

*Euclidean distance between observations

The goal is to obtain the simplest structure possible that still represents homogeneous groupings. If we monitor the overall similarity measure as the number of clusters decreases, large increases in the overall measure indicate that two clusters were not that similar.

In the particular example the overall measure increases when we first joint two observations (Step 1) and then again when we make the first 3-member cluster (step-2). In the next two steps (3 and 4) the overall measure does not change substantially. This indicates that we are forming other clusters with essentially the same homogeneity. of the existing clusters. In step 5 which joins the two 3-member cluster we see a large increase. This indicates that joining these two clusters resulted in single cluster which was markedly less homogeneous. Thus, we consider the cluster solution of step 4 better than that of step 5.

We can also see that in the last step 6 the overall measure again increased slightly. This indicates that even though the last observation remained separated until the last step, when it was joined it changed the cluster homogeneity.

Thus, when reviewing the range of cluster solutions, the three-cluster solution solution of step 4 seems the most appropriate for a final cluster solution: Two equally size clusters and the single outlying observation.

- The final cluster solution requires substantial researcher judgment.
- It is always up to the researcher to make the final decision as to the number of clusters to accept as the final solution.
- Inmost marketing research studies more than two variables are measured on each object and the situation is much more complex with many more observations
- Given the vectors $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$, then the (Euclidean) distance between X and Y is given by:

$$\sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2}.$$

E.g. let $X = \begin{pmatrix} 9 \\ 3 \\ 1 \end{pmatrix}$ and $Y = \begin{pmatrix} 10 \\ 2 \\ 9 \end{pmatrix}$.

The Euclidean distance is given by:

$$\sqrt{(9-10)^2 + (3-2)^2 + (1-9)^2} = \sqrt{66} = 8.12.$$

Example: salary and number of Cars

The relationship between the salary and car ownership of three individuals (objects) A, B and C are given by:

	Original Data		Mean-Corrected Data		Standardize Data	
Objects	Salary$	cars	Salary$	cars	Salary$	cars
A	20000	1	0	-1/3	0	$-1/\sqrt{3}$
B	30000	2	10000	2/3	1	$2/\sqrt{3}$
C	10000	1	-10000	-1/3	-1	$-1/\sqrt{3}$
Mean	10000	4/3	0	0	0	0
SD	10000	$1/\sqrt{3}$	10000	$1/\sqrt{3}$	1	1

	Original		Mean-Corrected		Standardize	
Objects	Distance	rank	Distance	Rank	Distance	Rank
A-B	10000	1	10000	1	$\sqrt{4}=2$	2
A-C	10000	1	10000	1	1	1
B-C	20000	2	20000	2	$\sqrt{7}=2.65$	3

Note that for the original and standardized data the distances between A and B are given, respectively, by:

$$d(A,B)=\sqrt{(20000-30000)^2+(1-2)^2}=10000.00005$$
$$=10000.$$

and $$d(A,B)=\sqrt{(0+1)^2+(\tfrac{-1-2}{3})^2}=\sqrt{4}=2.$$

Standardizing the data

- If the variables are measured in vastly different units, then the clustering solution will be influence by the units of measurement. In such cases, before clustering the individuals, the data should be standardize by rescaling each variable to have a mean zero and a standard deviation of unity.

Example

Suppose three individuals (objects) A, B and C are measured on two variables: Probability of purchasing brand X (in percentages) and amount of time spent viewing commercials for brand X (in minutes or seconds). The following results have been obtained:

	Commercial viewing time		
	Purchase		
Object	Probability(%)	(minutes)	(seconds)
A	60	3.0	180
B	65	3.5	210
C	63	4.0	240

From the obtained information distance measures can be calculated. Here the Euclidean and squared (absolute) Euclidean distances are calculated. The distance values with smaller values indicating greater proximity and similarity are shown below:

	Distance based on **minutes** of viewing time			
	Simple Euclidean Distance		Squared or Absolute Euclidean Distance	
Object	Value	Rank	Value	Rank
A-B	5.025	3	25.25	3
A-C	3.162	2	10.00	2
B-C	2.062	1	4.25	1

The most similar objects (with the smallest distance) are B and C, followed by A and C, with A and B the least similar (or least proximal). The orderings hold for all two distance measures, but the relative similarity, or dispersion, between objects is the most pronounced in the squared Euclidean distance measure.

The ordering of similarities can change markedly with only a change in the scaling of one of the variables. If we measured the viewing time in seconds instead of minutes, then differences would emerge:

	Distance based on **seconds** of viewing time			
	Simple Euclidean Distance		Squared or Absolute Euclidean Distance	
Object	Value	Rank	Value	Rank
A-B	30.41	2	925	2
A-C	60.07	3	3609	3
B-C	30.06	1	904	1

The similarity orderings have changed substantially. While B and C are still the most similar, the pair A-B is now next most similar and is almost identical to the similarity of B-C.

What has occurred is that the scale of the viewing time variable has dominated the calculations making the purchasing probability less significant in the calculations.

Consider standardizing the data. That is, subtract the mean and then divide by the standard deviation[a]. Notice that the mean and standard devision of the purchasing probability are given, respectively, by $(0.60+0.65+0.63)/3 = 0.627$ and 0.025. Similarly the mean and standard deviation of the viewing time in minutes (seconds) are given by 3.5 (210) and 0.5 (30), respectively.

	Standardize Values		Simple Euclidean Distance		Squares Euclidean Distance	
Object Pair	Purchase Probability	Min/Sec of Viewing Time	Values	Rank	Value	Rank
A-B	-1.06	-1.0	2.22	2	4.95	2
A-C	0.93	0.0	2.33	3	5.42	3
B-C	0.13	1.0	1.28	1	1.63	1

The ordering holds for the two distance measures, but relative similarity, or dispersion between objects is the most pronounced in the squared Euclidean distance measure.

[a] $\mathrm{SD} = \sqrt{(x_i - \overline{x})^2/(n-1)}$

Hierarchical procedures: The agglomerative method

- The Hierarchical procedures involve the construction of a hierarchy of a tree like structure. There are two hierarchical clustering procedures: The agglomerative and divisive methods.
- In the **agglomerative method** each object or observation starts out as its own cluster. In subsequent steps, the two closest clusters (or individuals) are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. In some cases an individual joins the first two in a cluster. In others, two groups of individuals formed at an earlier stage may join together in a new cluster. Eventually, all individuals are grouped into one large cluster. For this reason, agglomerative procedures are sometimes refereed to as buildup methods.
- An important characteristic of the hierarchical procedures is that it results at an earlier stage are always nested within the results at a later stage, creating a similarity tree. E.g. a six-cluster solution is obtained by joining two clusters found at the seven-cluster stage. Thus, any member of a cluster can trace its membership in an unbroken path to its beginning as a single observation. The process is shown as **dendrogram**.
- The **Divisive method** proceeds in the direction opposite to agglomerative method.

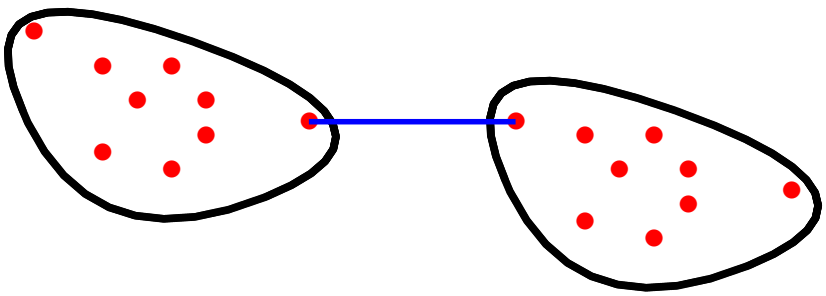
Methods for constructing the Clusters

There are five popular methods for developing the clusters during the agglomerative method. The methods differ in how the distance between clusters is computed:

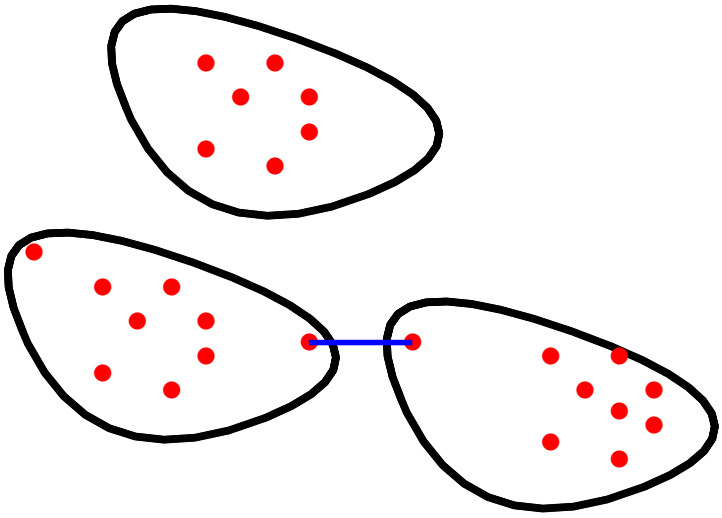
- Single Linkage

  The single-linkage procedure is based on minimum distance. It finds the two objects separated by the shortest distance and places them in the first cluster. Then, the next shortest distance it is found, and either a third object joins the first two to form a cluster, or a new two-member cluster is formed. The process continues until all objects are in one cluster.

  The distance between any two clusters is the shortest distance form any point in one cluster to any point in the other. Two clusters are merged at any stage by the single shortest link between them.



Problems may occur with the *Single Linkage* when the clusters are poorly delineated. In such cases the Single Linkage can form a long snakelike chains and eventually all individuals are places in one chain.



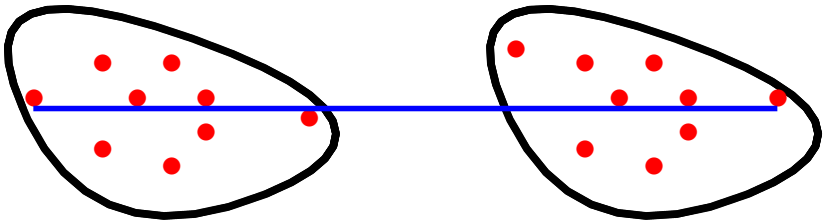
- Complete Linkage

  The complete linkage method is similar to single linkage except that the cluster criterion is based on maximum distance. The maximum distance between individuals in each cluster represents the smallest (minimum-diameter) sphere that can enclose all objects in both clusters. This method is called complete linkage because all objects in a cluster are linked to each other at maximum distance or by minimum similarity. This technique eliminates the snaking problem identified with single linkage.

  The use of shortest distance reflects only the *closest*

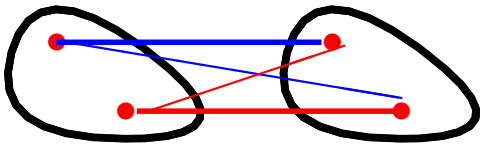
(most similar) single pair of objects and the complete linkage reflects the most extreme (least similar) pair of objects.

That is, in the complete linkage the further away objects in a cluster (outliers) are given more weight in the cluster decision.



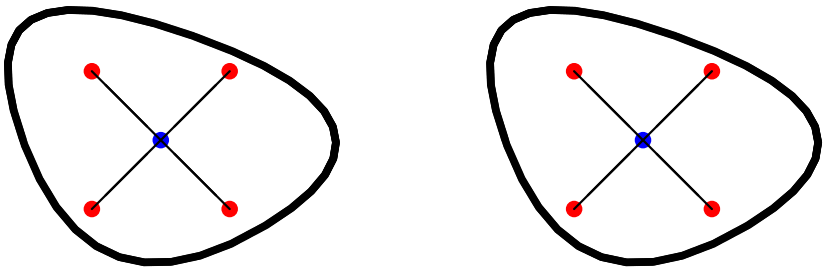
- Average Linkage

  In the average linkage method the cluster criterion is the average distance from all individuals of one cluster to all individuals in another cluster. That is, the criterion is based in all members of the clusters rather than on a single pair of extreme members. This method is one of the most popular ones, even though it is more computational expensive.



- Ward's method

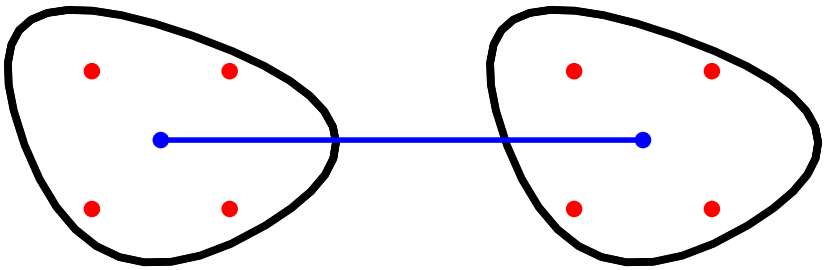
A variance methods attempt to generate clusters to minimize the within-cluster variance. A commonly used variance procedure is the Ward's method. For each cluster, the means for all the variables are computed. Then, for each object, the Euclidean distance to the cluster means is calculated. These distances are summed for all the objects. At each stage, the two clusters with the smallest increase in the overall sum of squares within-cluster distances are combined.



This method tends to combine clusters with a small number of observations. It is also biased towards the production of clusters with approximately the same number of observations.

- Centroid method

  In the centroids methods, the distances between two clusters is the distance between their centroids. Cluster centroids are the mean values of the observations on the variables in the cluster. In this method every time individuals are grouped, or clusters are merged, a new centroid is computed. One advantage of this method is that it is less affected by outliers.



- Of the hierarchical methods the average linkage and Ward's method have been shown to perform better than other procedures.

Ward's methods

The Ward's method is another hierarchical clustering method based on within variance rather than linkage. Suppose a sample of n observations has been partitioned into g groups, the ith group containing n_i observations with mean $\bar{x}_i$. The within group sum of squared deviations for this g-group partition is:

$$W = \sum_{j=1}^{g} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \ ,$$

where x_{ij} is the jth observation in the ith group.

The value of W may be calculated for any partition. The quantity W/n is the pooled within group variance for the partition.

The Ward clustering technique proceeds as follows:

1. Begin with n groups, each group consisting of one observation. At this stage $W = 0$.
2. At each stage reduce the number of groups by one through the merger of those two groups whose combination gives the least possible increase in W.
3. Continue for a total of $n - 1$ mergers until there is only one group.

Example

Consider the one-dimensional case. A sample of 6 observations gives the scores $1, 2, 5, 7, 9, 10$. The Ward clustering technique will give the sequence of groupings:

Step	# of Groups	Groups	W
0	6	$(1),(2),(5),(7),(9),(10)$	0.00
1	5	$(1,2),(5),(7),(9),(10)$	0.50
2	4	$(1,2),(5),(7),(9,10)$	1.00
3	3	$(1,2),(5,7),(9,10)$	3.00
4	2	$(1,2),(5,7,9,10)$	15.25
5	1	$(1,2,5,7,9,10)$	67.33

All combinations in step 1:

pair	centre	W	pair	centre	W
$(1,2)$	1.5	0.5	$(5,7)$	6.0	2.0
$(1,5)$	3.0	8.0	$(5,9)$	7.0	8.0
$(1,7)$	4.0	18.0	$(5,10)$	7.5	12.5
$(1,9)$	5.0	32.0	$(7,9)$	8.0	2.0
$(1,10)$	5.5	40.5	$(7,10)$	8.5	4.5
$(2,5)$	3.5	4.5	$(9,10)$	9.5	0.5
$(2,7)$	4.5	12.5			
$(2,9)$	5.5	24.5			
$(2,10)$	6.0	32.0			

All 4-cluster possibilities	W
$(1,2,5),(7),(9),(10)$	8.67
$(1,2,7),(5),(9),(10)$	
$(1,2,9),(5),(7),(10)$	
$(1,2,10),(5),(7),(9)$	
$(1,2),(5,7),(9),(10)$	
$(1,2),(5,9),(7),(10)$	
$(1,2),(5,10),(7),(9)$	
$(1,2),(5)(7,9),(10)$	
$(1,2),(5)(7,10),(9)$	$0.5+0+(7-8.5)^2+(10-8.5)^2=5$
$(1,2),(5)(7),(9,10)$	$0.5+0+0+0.5=1$

All 3-cluster possibilities	W
$(1,2,5),(7),(9,10)$	$8.67+0+0.5=9.17$
$(1,2,7),(5),(9,10)$	
$(1,2,9,10),(5),(7)$	
$(1,2),(5,7),(9,10)$	$0.5+2.0+0.5=3.0$
$(1,2),(5,9,10),(7)$	
$(1,2),(5),(7,9,10)$	5.17

All 2-cluster possibilities	W
$(1,2,5,7),(9,10)$	15.25
$(1,2,9,10),(5,7)$	
$(1,2),(5,7,9,10)$	

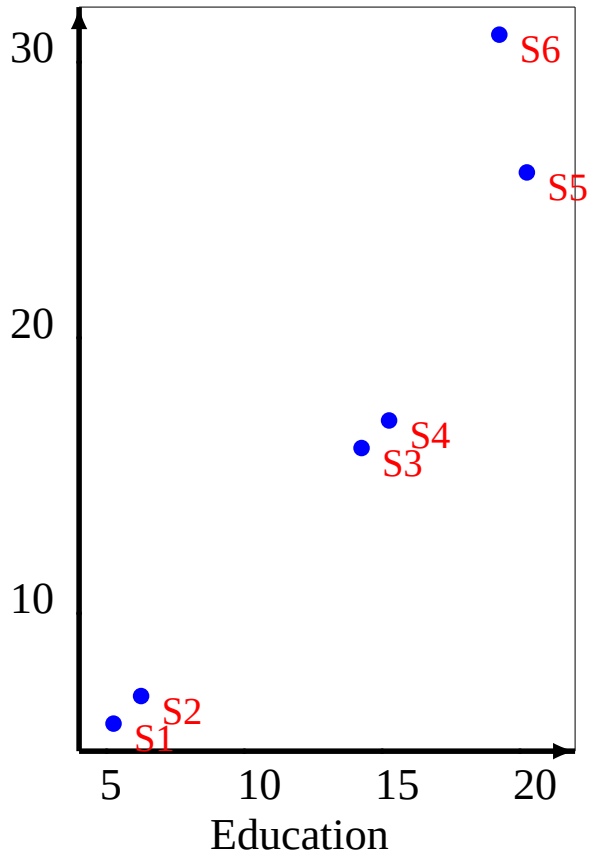
All 1-cluster possibilities	W
$(1,2,5,7,9,10)$	67.33

What is the ESS of the 2-cluster solution $\{(1,2,5),(7,9,10)\}$?

Example: Centroid and Ward's methods

Consider the following hypothetical data with sid (subject identification), income ($thousands) and education (years).

Sid	Income	education
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



Centroid

The similarity matrix is given by:

	S1	S2	S3	S4	S5	S6
S1	0					
S2	2	0				
S3	181	145	0			
S4	221	181	2	0		
S5	625	557	136	106	0	
S6	821	745	250	212	26	0

Calculating the squared Euclidean distances: E.g

$$d^2(S1,S2) = (5-6)^2 + (5-6)^2 = 2$$
$$d^2(S2,S4) = (6-16)^2 + (6-15)^2 = 181$$

a. Five clusters

The first cluster is formed by combining subjects S1 and S2. The first cluster is then represented by the centroid of the subjects of S1 and S2. The new cluster has an average education of 5.5 years (=(5+6)/2) and an average income of 5.5 thousand dollars.

The next table gives the data for the 5 clusters that have been formed and the corresponding similarity matrix: (using squared Euclidean distances).

	{S1,S2}	S3	S4	S5	S6
{S1,S2}	0				
S3	162.50	0			
S4	200.50	2	0		
S5	590.50	135.96	106	0	
S6	782.50	250	212	26	0

E.g.
$d^2(\{S1,S2\},S5) = (5.5-25)^2 + (5.5-20)^2 = 590.50.$

b. Four clusters

As can be seen, S3 and S4 have the smallest distance and are most similar. We group these two subjects into a new cluster. This cluster will be represented by the centroid of the subjects in this group. i.e. $(15.5, 14.5)$.

	{S1,S2}	{S3,S4}	S5	S6
{S1,S2}	0			
{S3,S4}	181	0		
S5	590.50	120.50	0	
S6	782.50	230.50	26	0

E.g. $d^2(\{S3,S4\},S5) = (15.5-25)^2 + (14.5-20.0)^2 = 120.5.$

c. Three clusters

S5 and S6 have smallest distance and therefore are combined to form a 3rd cluster. The centroid of this cluster is $(27.5, 19.5)$.

	{S1,S2}	{S3,S4}	{S5,S6}
{S1,S2}	0		
{S3,S4}	181	0	
{S5,S6}	680	169	0

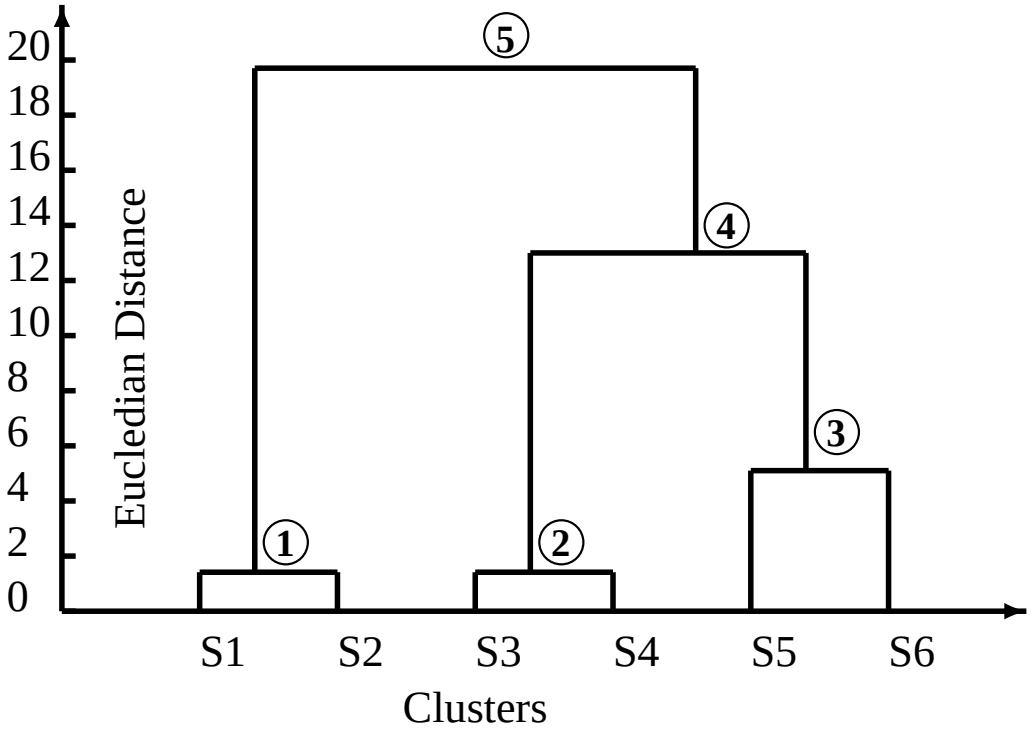
d. Two clusters

{S3, S4} and {S5,S6} have smallest distance and will be combined to form a new cluster with centroid $(21.5, 17)$.

The Square Euclidean distance between the clusters {S3, S4, S5,S6} and {S1,S2} is 388.25.

e. One cluster

Group all subjects into one cluster.



Ward's method

The Ward's method does not compute distances between clusters. It forms clusters by maximizing within-clusters homogeneity. The within-cluster sum of squares is used as the measure of homogeneity (this is also known as the error sums of squares ESS).

a. Five clusters

Initially each observation is a cluster and therefore the ESS is 0. To form 5 clusters, we need 1 cluster of size 2 and 4 clusters of size 1.

E.g. we can have one cluster consisting of subjects S1 and S2. ESS $=$
$(5-5.5)^2+(5-5.5)^2+(6-5.5)^2+(6-5.5)^2=1$

E.g. we can have one cluster consisting of subjects S2 and S5 ESS $=(6-15.5)^2+(6-13)^2+(25-15.5)^2+$ $(20-13)^2=278.5$.

The table gives all possible 5 cluster solutions with their ESS. Based on the criterion of minimizing ESS cluster solution 1 or 10 can be chosen. This is chosen at random. Let us chose cluster solution 1.

b. Four clusters

E.g. the cluster consisting of subjects S1, S2 and S6:

$$\begin{aligned}\text{ESS} &= (5-13.66)^2+(5-10)^2+(6-13.66)^2+(6-10)^2 \\ &\quad +(30-13.66)^2+(19-10)^2 \\ &= 522.66\end{aligned}$$

Cluster solution 5 is the one which minimizes the ESS.

Table 7.6 **Ward's Method**

Cluster Solution	Members in Cluster 1	2	3	4	5	ESS
(a) All Possible Five-Cluster Solutions						
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0
(b) All Possible Four-Cluster Solutions						
1	S1,S2,S3	S4	S5	S6		109.333
2	S1,S2,S4	S3	S5	S6		134.667
3	S1,S2,S5	S3	S4	S6		394.667
4	S1,S2,S6	S3	S4	S5		522.667
5	S1,S2	S3,S4	S5	S6		2.000
6	S1,S2	S3,S5	S4	S6		69.000
7	S1,S2	S3,S6	S4	S5		126.000
8	S1,S2	S4,S5	S3	S6		54.000
9	S1,S2	S4,S6	S3	S5		107.000
10	S1,S2	S5,S6	S3	S4		14.000

Agglomeration schedule

The results of hierarchical methods are usually summarized in an agglomeration schedule. Consider the five objects with the following dissimilarities were clustered with complete linkage:

	Object1	Object2	Object3	Object4	Object5
Object1	0				
Object2	1.0	0			
Object3	2.0	3.0	0		
Object4	8.0	9.0	10.0	0	
Object5	11.0	12.0	13.0	5.0	0.0

The Agglomeration schedule is represented as:

Step or Stage	Clusters combined Cluster 1	Cluster 2	Level or Coefficient
1	1	2	1.0
2	1	3	3.0
3	4	5	5.0
4	1	4	13.0

The agglomeration coefficient is the distance between the two cases of clusters being combined.

The agglomeration schedule contains the following information:

- In the first step the cluster to which object 1 belongs is combined with the cluster to which object 2 belongs. The two clusters are merged at a level v1.

- In the second step the cluster to which object 1 belongs is combined with the cluster to which object 3 belongs at a level of 3.0. Note that the schedule only enumerates the first object of a cluster. Cluster 1 in step 2 actually consists of two objects, namely object 1 and 2 that have been merged in the first step.

- In step 3 the clusters to which object 4 and 5 belong are combined at a level of 5.0.

- In step 4 the clusters to which object 1 and object 4 belong are amalgamated. All objects are now in one single cluster because cluster 1 consists of the object 1, 2 and 3 and cluster 2 contains objects 4 and 5. The clusters are merged at a level of 13.0.

The schedule does not inform us which objects belong to a cluster. However, this information may be deduced from the schedule. Its main purpose is to inform about the process and to give some hints about the number of clusters. The agglomeration levels should continuously increase (if dissimilarities are used) or decrease (if similarities are analyzed).

Some computer programmes provide additional information. SPSS shows at which step a cluster appeared first and in will be merged with another cluster:

Stage	Clusters combined		Agglom. Coeffic.	Stage Cluster First Appears		Next Stage
	Clus 1	Clus 2		Clus 1	Clus 2	
1	1	2	1.0	0	0	2
2	1	3	3.0	1	0	4
3	4	5	5.0	0	0	4
4	1	4	13.0	2	3	0

Unfortunately, SPSS does not record the number of cluster, the increase of dissimilarities or the decrease of similarities and ties. Ties occur if more than one pair of *most similar* clusters are present in a certain step.

We can use the table to identify single observations that are joint very late in the clustering process -potential outliers.

How many clusters should be formed

- One of the most important issues in the Cluster analysis is determining the final number of clusters to be formed (also known as stopping rule). Unfortunately, no standard objective selection procedure exists. There is no internal statistical criterion used for inference. A number of ad hoc procedures have been developed most of them quite complex.

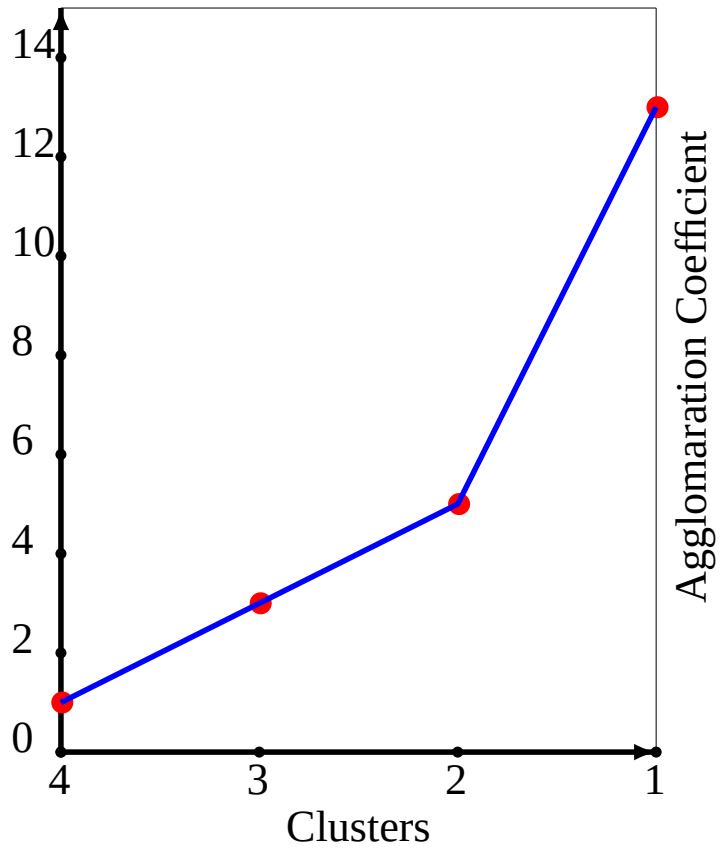
- One class of stopping rules that is relatively simple examines some measure of similarity or distance between clusters at each successive step. A cluster solution is defined when when the similarity measure exceeds a specified value or when the successive values between steps makes a sudden jump.

- When a large increase occurs, the researcher selects the prior cluster solution on the logic that its combination caused a substantial decrease in similarity. This stopping rule has been shown to provide fairly accurate decisions in empirical studies.

- Frequently, the number of clusters is determined on the basis of the dendrogram. The user marks the number of 'small' hills (clusters) that combine objects at a low distance level.

  An inspection of the agglomeration levels is another frequently applied approach. The levels are read downwards starting with step 1. The user looks for a sharp increase (if dissimilarities are clustered) or decrease (if similarities are clustered) of the agglomeration levels. In the last example a considerable increase occurs between step 3 and 4. Therefore, step 3 is accepted as a best cut because step 4 increases the agglomeration level too much. Step 3 corresponds to the solution with two clusters. Cluster 1 contains object 1, 2 and 3, cluster 2 object 4 and 5.

  The method described above can be applied graphically in the so called inverse scree test. A plot is constructed in this test (it is not a test in a statistical sense!). The x-axis contains the number of clusters, the y-axis the agglomeration levels. A sharp increase in the agglomeration schedule results in an elbow knick. In our example a knick occurs at two clusters (see figure). Therefore two clusters will be selected.



- It is probably best to to compute a number of different cluster solutions (e.g. 2, 3, 4) and then decide among the alterative solutions by using poor criteria, practical judgment, common sense, or theoretical foundations. The cluster solutions will be improved by restricting the solution according to conceptual aspects of the problem.

Example

Number of Clusters	Agglomeration Coefficient	Percentage Change in Coefficient to next level
10	258.7	8.8
9	281.4	8.4
8	305.0	9.2
7	331.1	9.6
6	364.9	9.1
5	398.1	12.1
4	446.3	17.2
3	523.0	17.6
2	615.0	61.8
1	994.8	–

The clustering (agglomeration) coefficient shows rather large increase in going from four to three clusters (523.0-446.3=76.6), three to two clusters (615.0-523.0 = 92.0), and two to one cluster (994.8-615.0 = 379.8). To help identifying large relative increase in the cluster homogeneity, we calculate the percentage change in the clustering coefficient. The largest percentage increase by far occurs in going from two to one clusters. The second most noticeable change in the percentage increase occurs in combining four in to three clusters.

Disadvantages of hierarchical procedures

- Time-consuming (for a data set with n objects, the procedure offers n possible solutions).

- Not being able to amend a previous assignment. An object assigned to a cluster will never be reassigned to any other cluster. I.e. cannot correct later for erroneous decisions made earlier

Example

Consumers were asked to express their degree of agreement with the following statements on a seven-point scale (1=disagree, 7=agree):

- V1: Shopping is fun.
- V2: Shopping is bad for your budget.
- V3: I combine shopping with eating out.
- V4: I try to get the best buys while shopping.
- V5: I do not care about shopping.
- V6: You can save a lot of money by comparing prices.

V1	V2	V3	V4	V5	V6
6.00	4.00	7.00	3.00	2.00	3.00
2.00	3.00	1.00	4.00	5.00	4.00
7.00	2.00	6.00	4.00	1.00	3.00
4.00	6.00	4.00	5.00	3.00	6.00
1.00	3.00	2.00	2.00	6.00	4.00
6.00	4.00	6.00	3.00	3.00	4.00
5.00	3.00	6.00	3.00	3.00	4.00
7.00	3.00	7.00	4.00	1.00	4.00
2.00	4.00	3.00	3.00	6.00	3.00
3.00	5.00	3.00	6.00	4.00	6.00
1.00	3.00	2.00	3.00	5.00	3.00
5.00	4.00	5.00	4.00	2.00	4.00
2.00	2.00	1.00	5.00	4.00	4.00
4.00	6.00	4.00	6.00	4.00	7.00
6.00	5.00	4.00	2.00	1.00	4.00
3.00	5.00	4.00	6.00	4.00	7.00
4.00	4.00	7.00	2.00	2.00	5.00
3.00	7.00	2.00	6.00	4.00	3.00
4.00	6.00	3.00	7.00	2.00	7.00
2.00	3.00	2.00	4.00	7.00	2.00

Steps in SPSS:

Analyze/Classify/Hierarchical Cluster
/Statistics (tick on proximity matrix)
 / Plots   (tick on dendrogram)
 / Method   (select cluster method e.g
   within-groups linkage and interval
   e.g Euclidean distance)

Proximity Matrix

Case	City Block Distance 1	2	3	4	5	6	7	8	9	10	11	12	13
1	.000	16.000	6.000	13.000	17.000	3.000	5.000	5.000	12.000	16.000	14.000	5.000	17.000
2	16.000	.000	16.000	13.000	5.000	13.000	11.000	15.000	6.000	10.000	4.000	11.000	3.000
3	6.000	16.000	.000	15.000	19.000	7.000	7.000	3.000	16.000	18.000	16.000	7.000	15.000
4	13.000	13.000	15.000	.000	16.000	10.000	10.000	14.000	13.000	5.000	15.000	8.000	12.000
5	17.000	5.000	19.000	16.000	.000	14.000	12.000	18.000	5.000	13.000	3.000	14.000	8.000
6	3.000	13.000	7.000	10.000	14.000	.000	2.000	6.000	11.000	13.000	13.000	4.000	14.000
7	5.000	11.000	7.000	10.000	12.000	2.000	.000	6.000	11.000	13.000	11.000	4.000	12.000
8	5.000	15.000	3.000	14.000	18.000	6.000	6.000	.000	17.000	17.000	17.000	6.000	16.000
9	12.000	6.000	16.000	13.000	5.000	11.000	11.000	17.000	.000	10.000	4.000	11.000	9.000
10	16.000	10.000	18.000	5.000	13.000	13.000	13.000	17.000	10.000	.000	12.000	11.000	9.000
11	14.000	4.000	16.000	15.000	3.000	13.000	11.000	17.000	4.000	12.000	.000	13.000	7.000
12	5.000	11.000	7.000	8.000	14.000	4.000	4.000	6.000	11.000	11.000	13.000	.000	12.000
13	17.000	3.000	15.000	12.000	8.000	14.000	12.000	16.000	9.000	9.000	7.000	12.000	.000
14	16.000	14.000	18.000	3.000	17.000	13.000	13.000	17.000	14.000	4.000	16.000	11.000	13.000
15	7.000	15.000	9.000	10.000	14.000	6.000	8.000	8.000	13.000	13.000	15.000	6.000	16.000
16	16.000	12.000	18.000	5.000	15.000	13.000	13.000	17.000	12.000	2.000	14.000	11.000	11.000
17	5.000	15.000	11.000	10.000	14.000	6.000	6.000	8.000	13.000	13.000	15.000	6.000	16.000
18	16.000	10.000	18.000	9.000	13.000	15.000	15.000	19.000	10.000	6.000	10.000	13.000	9.000
19	16.000	16.000	18.000	5.000	19.000	15.000	15.000	17.000	16.000	6.000	18.000	11.000	15.000
20	17.000	5.000	17.000	16.000	6.000	16.000	14.000	18.000	5.000	13.000	5.000	14.000	8.000

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	2.000	0	0	3
2	6	7	2.000	0	0	7
3	10	14	3.000	0	1	8
4	2	13	3.000	0	0	14
5	5	11	3.000	0	0	10
6	3	8	3.000	0	0	15
7	6	12	4.000	2	0	9
8	4	10	4.333	0	3	11
9	1	6	4.333	0	7	13
10	5	9	4.500	5	0	12
11	4	19	5.250	8	0	17
12	5	20	5.333	10	0	14
13	1	17	5.750	9	0	15
14	2	5	6.500	4	12	18
15	1	3	6.900	13	6	16
16	1	15	7.429	15	0	19
17	4	18	8.200	11	0	18
18	2	4	13.556	14	17	19
19	1	2	14.375	16	18	0

Non-hierarchical methods

- Disadvantage: The number of clusters must be known a priori.

  1. Select k initial cluster seeds.
  2. Assign each observation to the cluster to which it is the closest.
  3. Reallocate each observation to one of the k clusters according to a rule.
  4. Stop if no reallocation. Otherwise go to step 2.

- Most of the nonhierarchical algorithms differ with respect to:

  1. Method used for obtaining initial cluster seeds.
  2. Rule for reassigning observations.

Methods for obtaining initial seeds

- Select the first k observations as initial seeds
- Take the first observation as first seed. The next observation which is at least a certain distance separated from the first seed is the second seed, the third should be at least a certain distance separated from the previous separated from the previous seeds, ...
- Take k observations at random.
- Refine selected seeds by some rule.
- Use a heuristic that identifies cluster centers such that they are as far apart as possible.
- Supplied by the researcher.

Rules for re-assigning observations

1. Calculate the centroids
   - Reallocate the objects to the cluster with the nearest centroid.
   - Recalculate the centroids after reassigning all observations
   - if the change in the centroids is larger than some specific value, then reallocate and so on.
2. The same as the previous method but now the centroids are recomputed after each reallocation

The first method is known as k-means.

K-means method

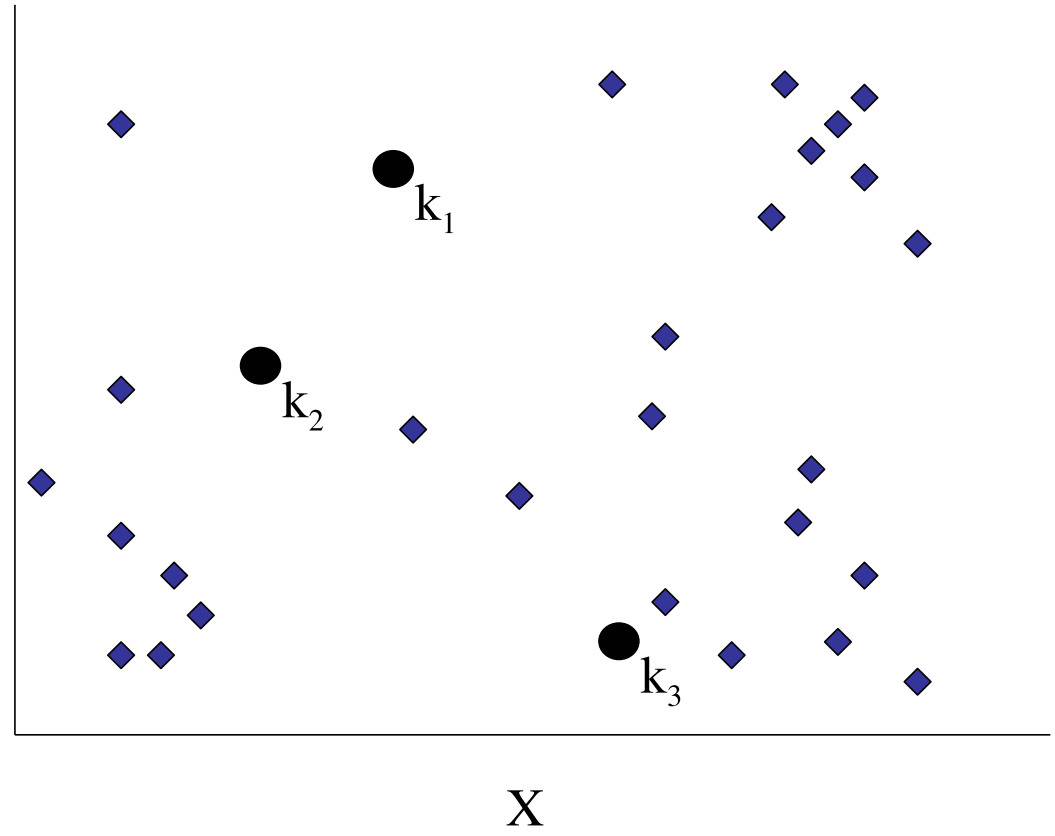
1. Start with randomly chosen cluster centers.
2. Allocate objects to the closest clusters (e.g. closer to the centroid of the cluster).
3. Recalculate the centroids of the clusters.
4. Reassign objects to clusters.
5. Repeat (step 2) until there is no change.

- There are n^k possible allocations of n objects to k clusters.
- In practice the optimal allocation is not obtained.
- Allocation depends on the initial clusters.
- Try the method with different starting clusters.

k-Means Example (I)

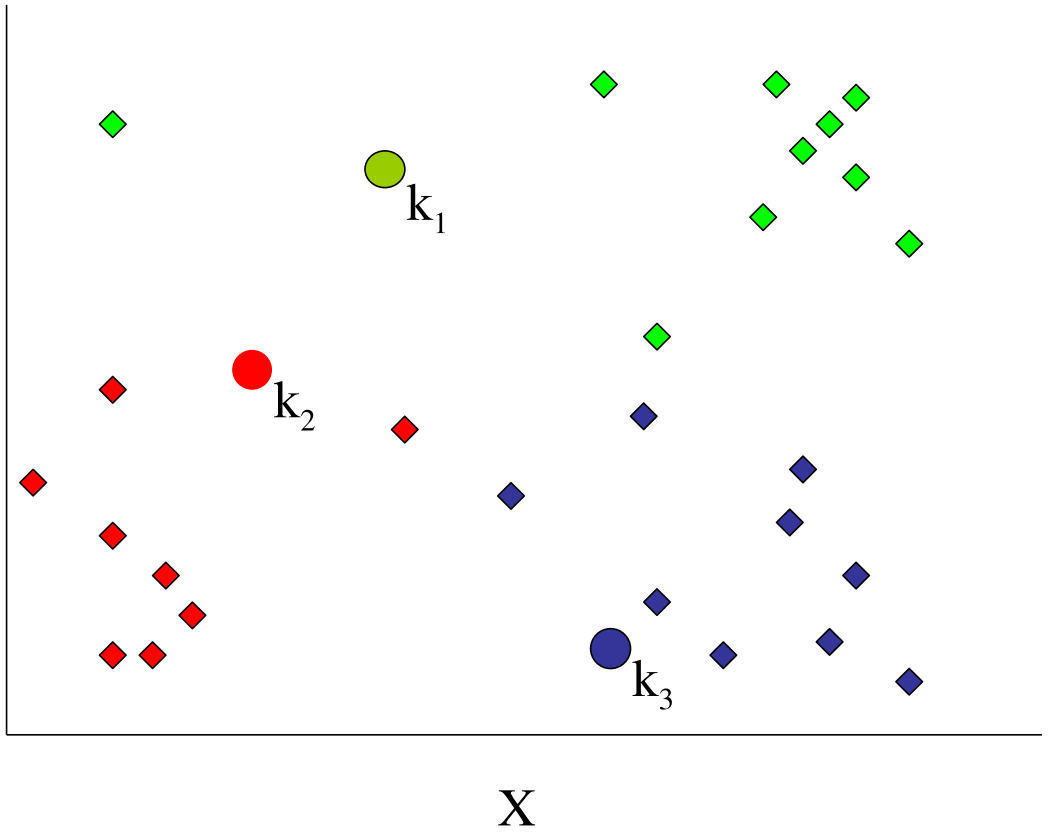
K-means: Example 1

Pick 3 initial cluster centers (randomly)



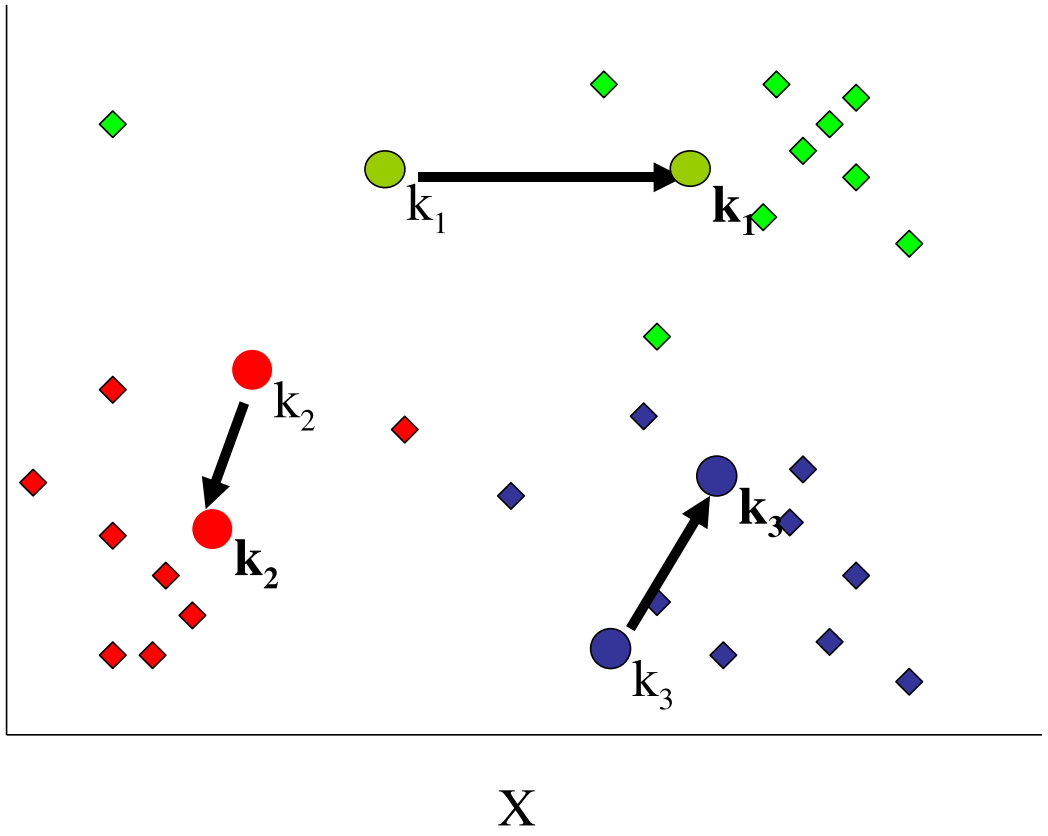
k-Means Example (II)

Assign each point to the closest cluster center



k-Means Example (III)

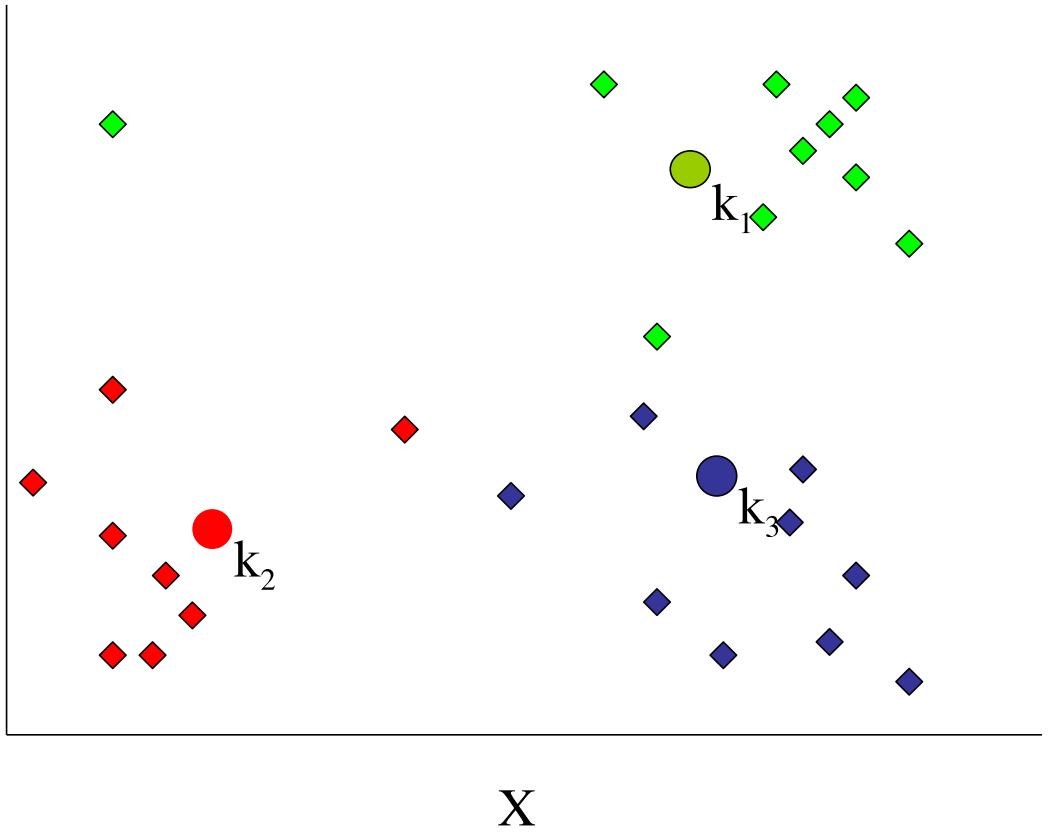
Move
each cluster
center
to the mean
of each cluster



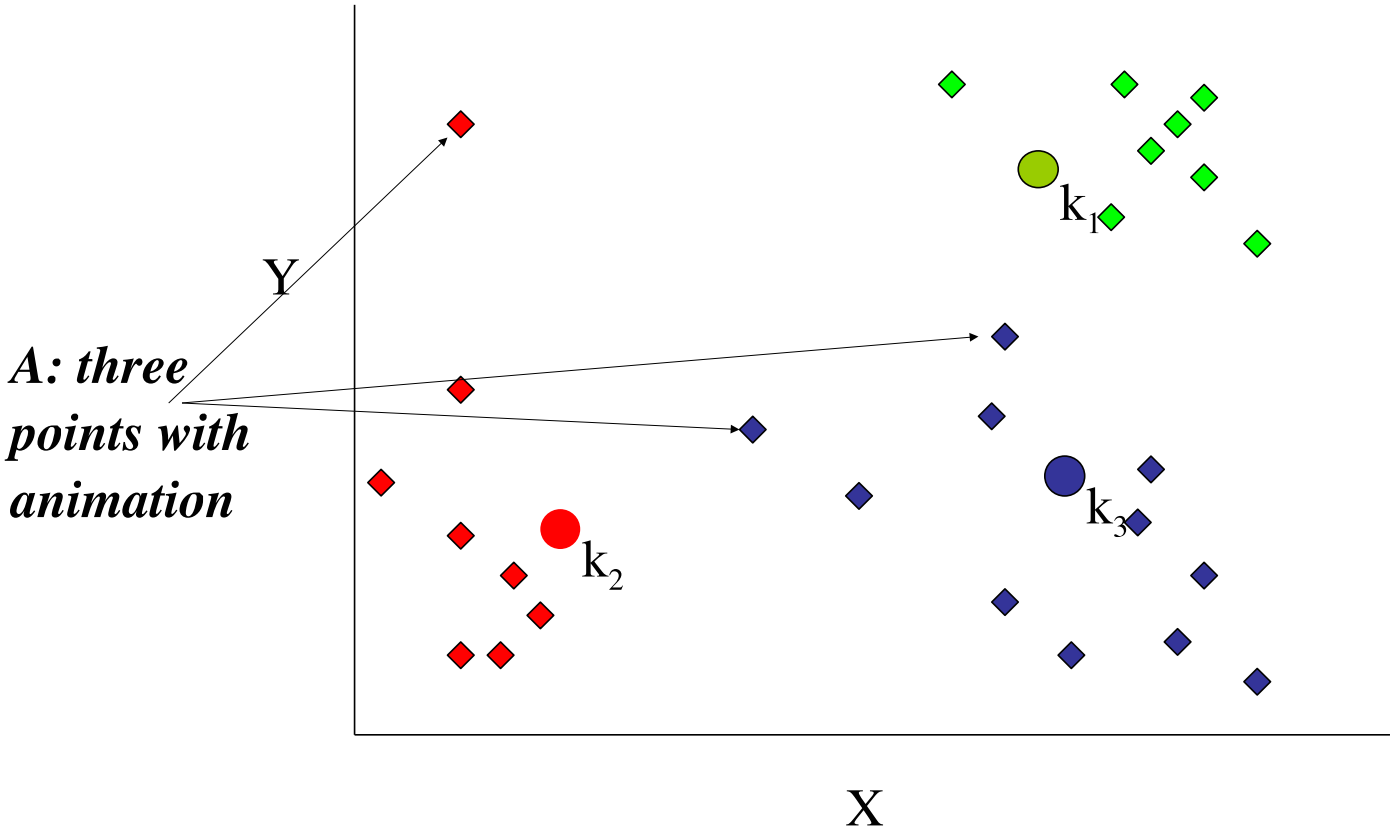
k-Means Example (IV)

Reassign points closest to a different new cluster center

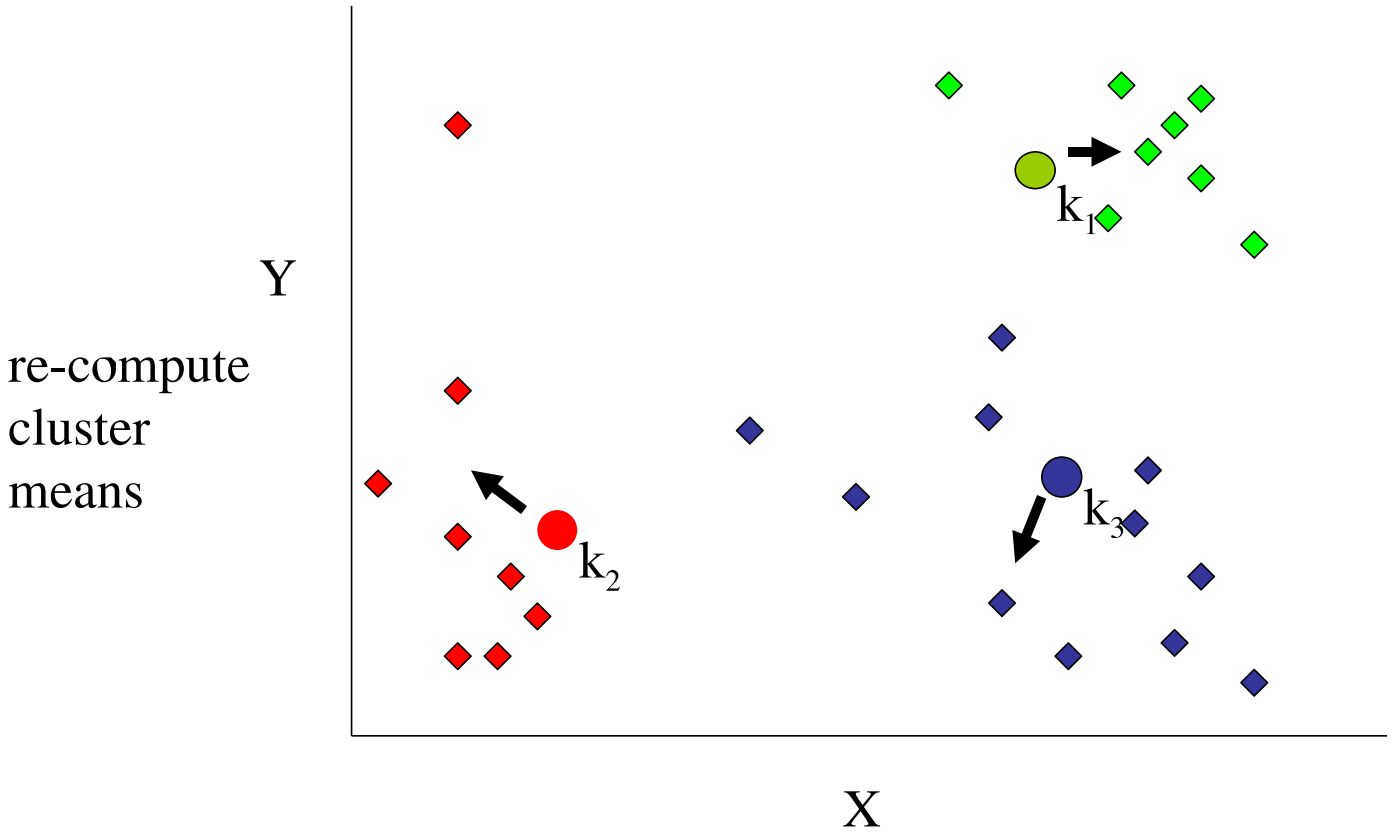
Q: Which points are reassigned?



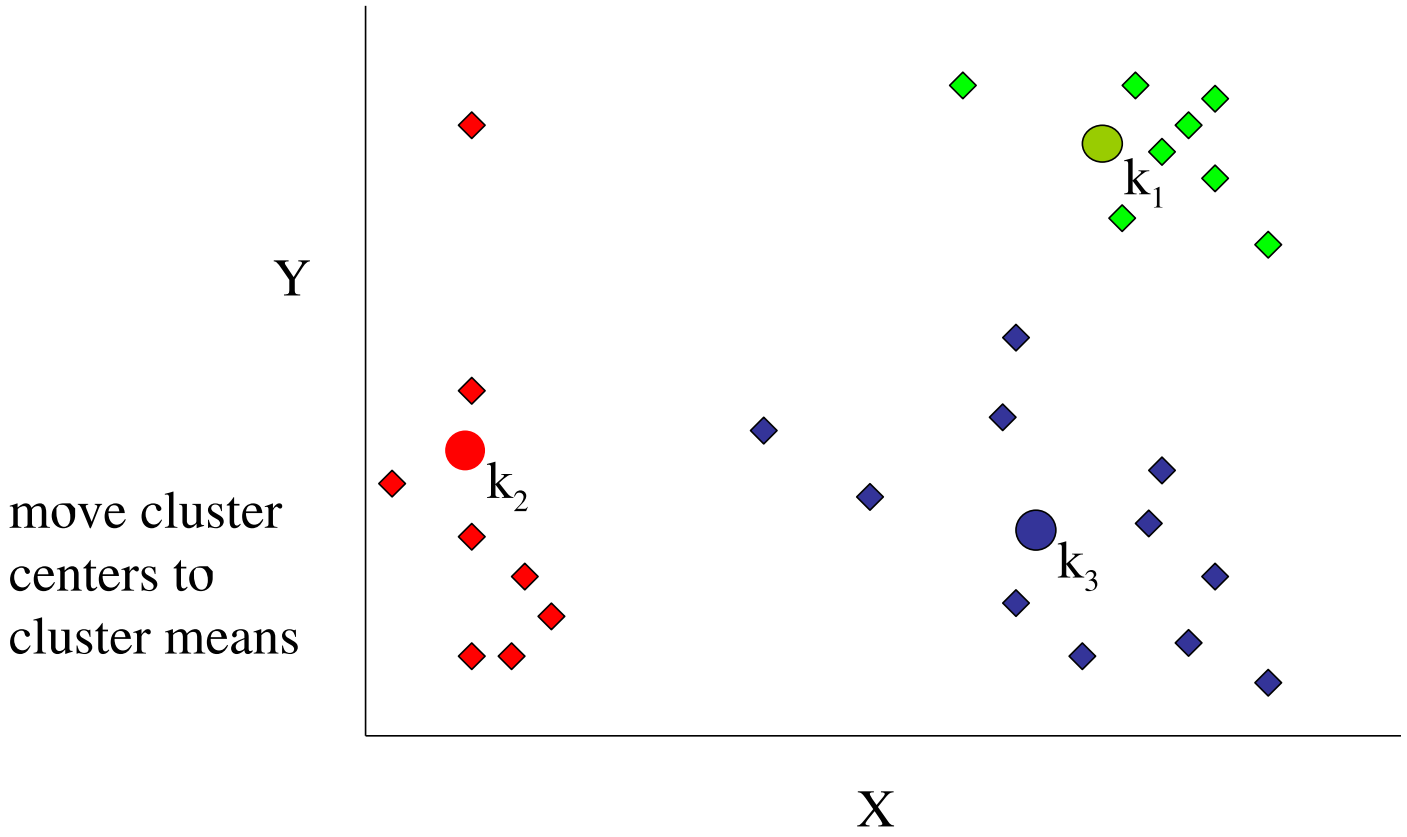
k-Means Example (V)



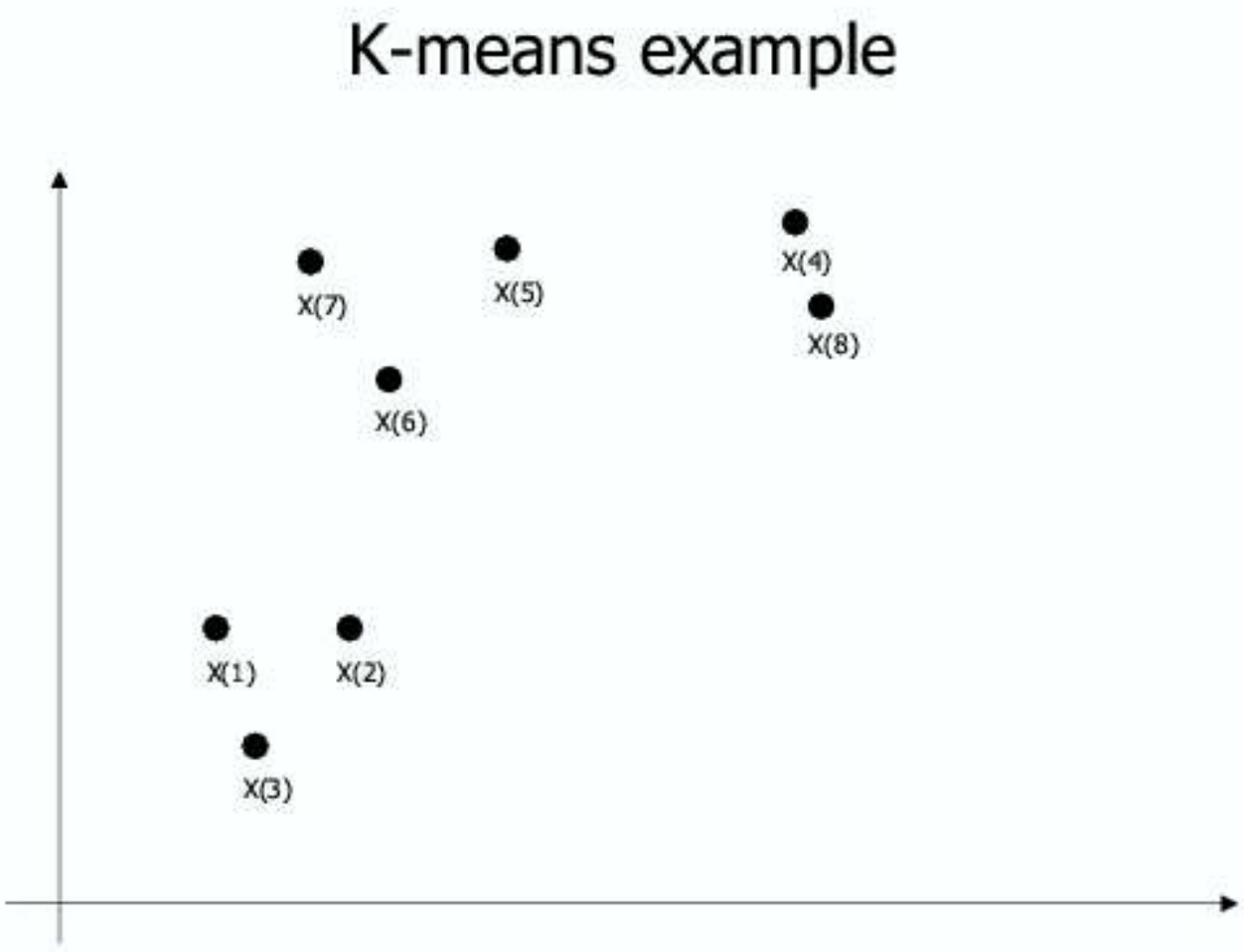
k-Means Example (VI)



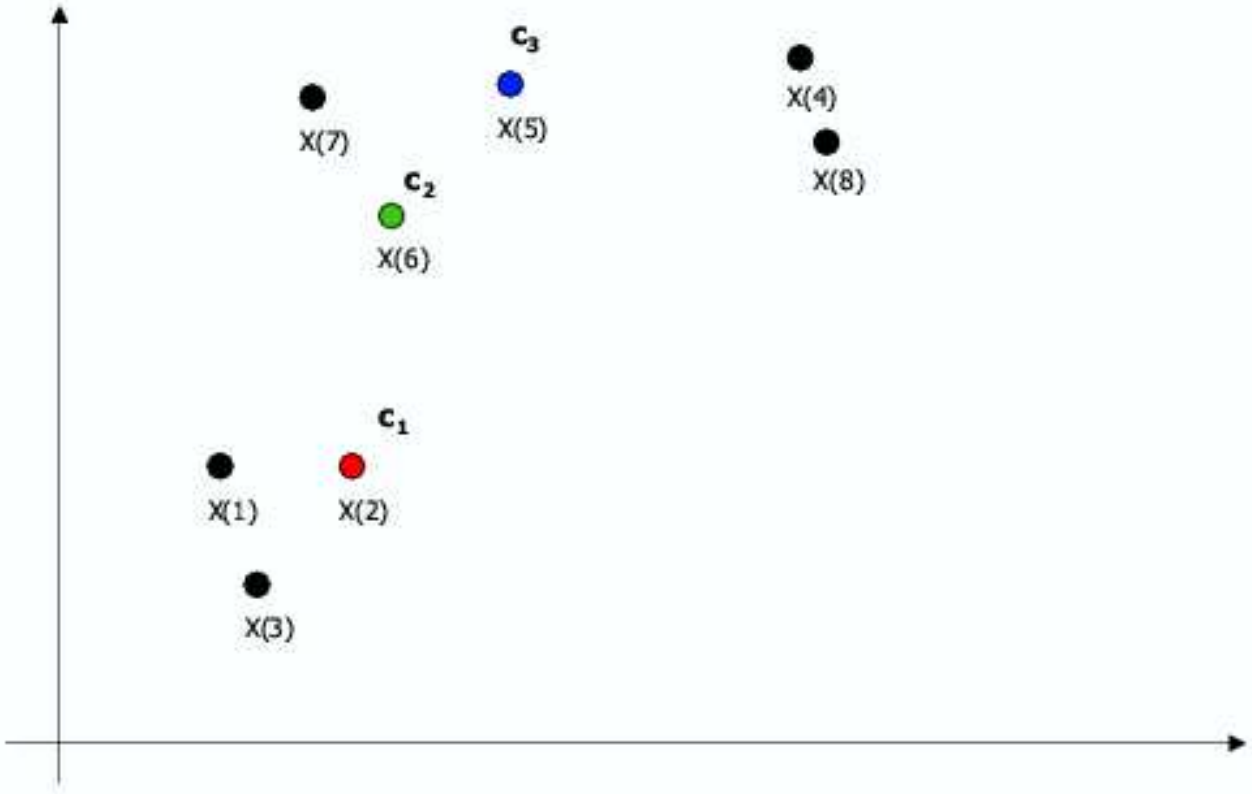
k-Means Example (VII)



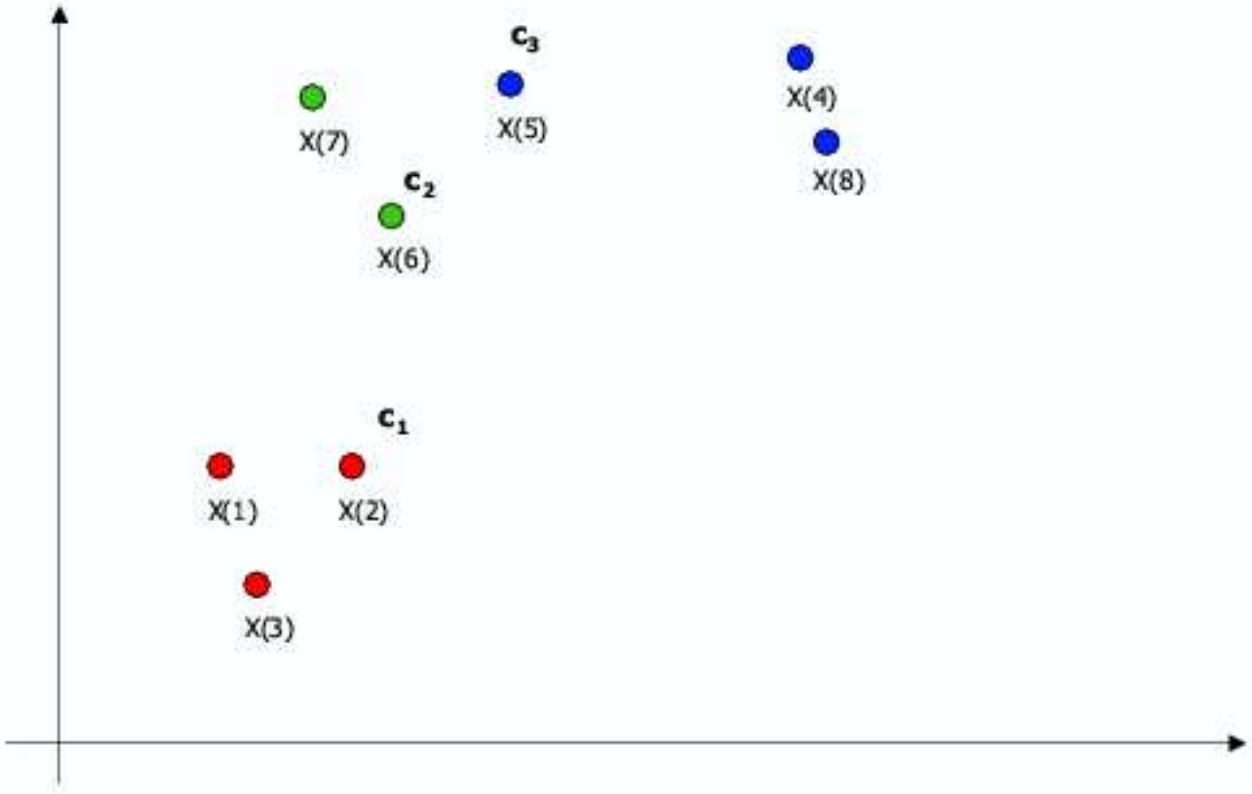
K-means: Example 2


K-means example
X(7)
X(5)
X(4)
X(8)
X(6)
X(1)
X(2)
X(3)

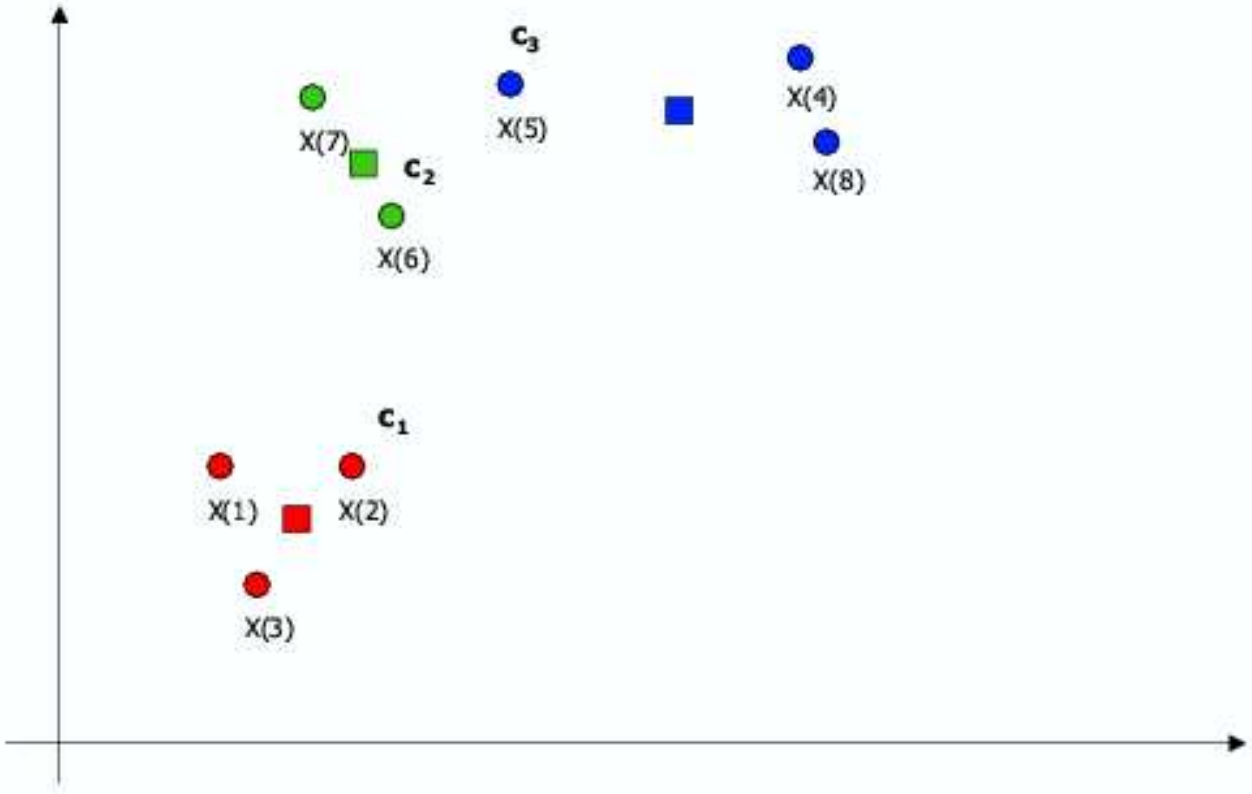
K-means example



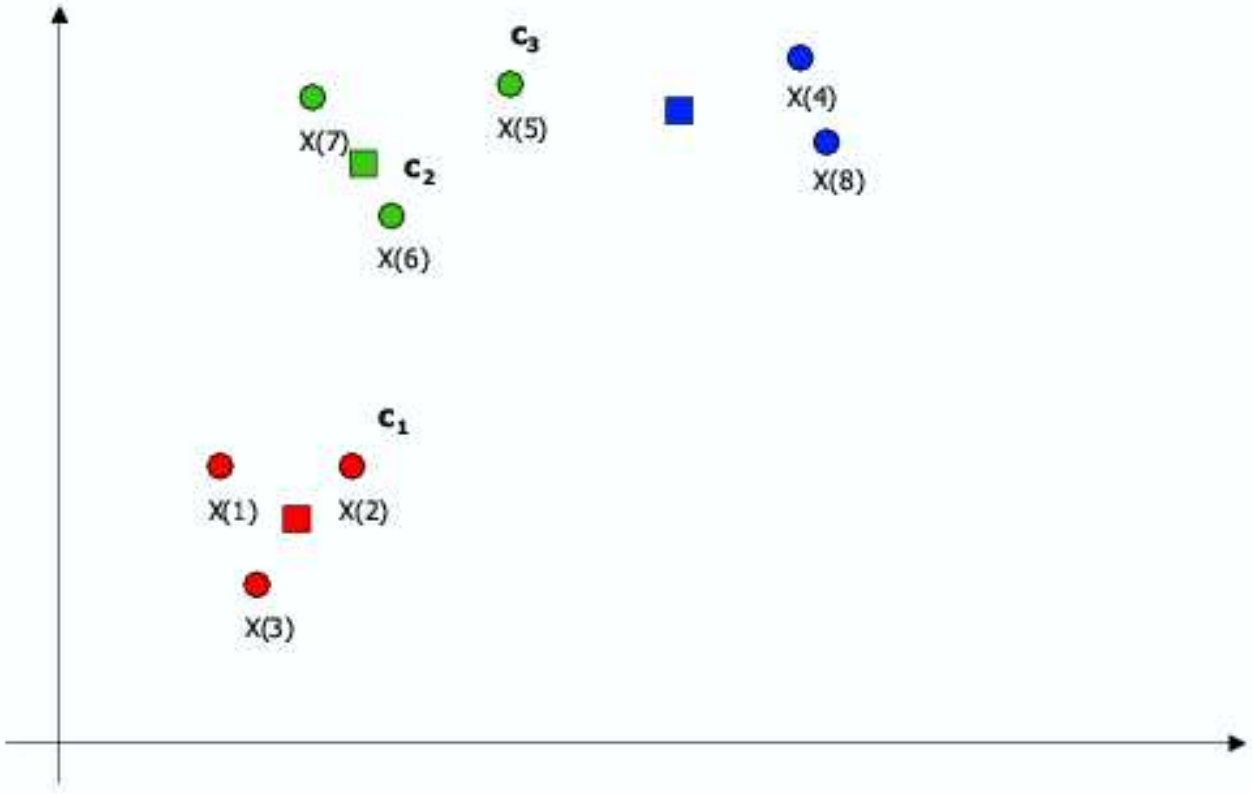
K-means example



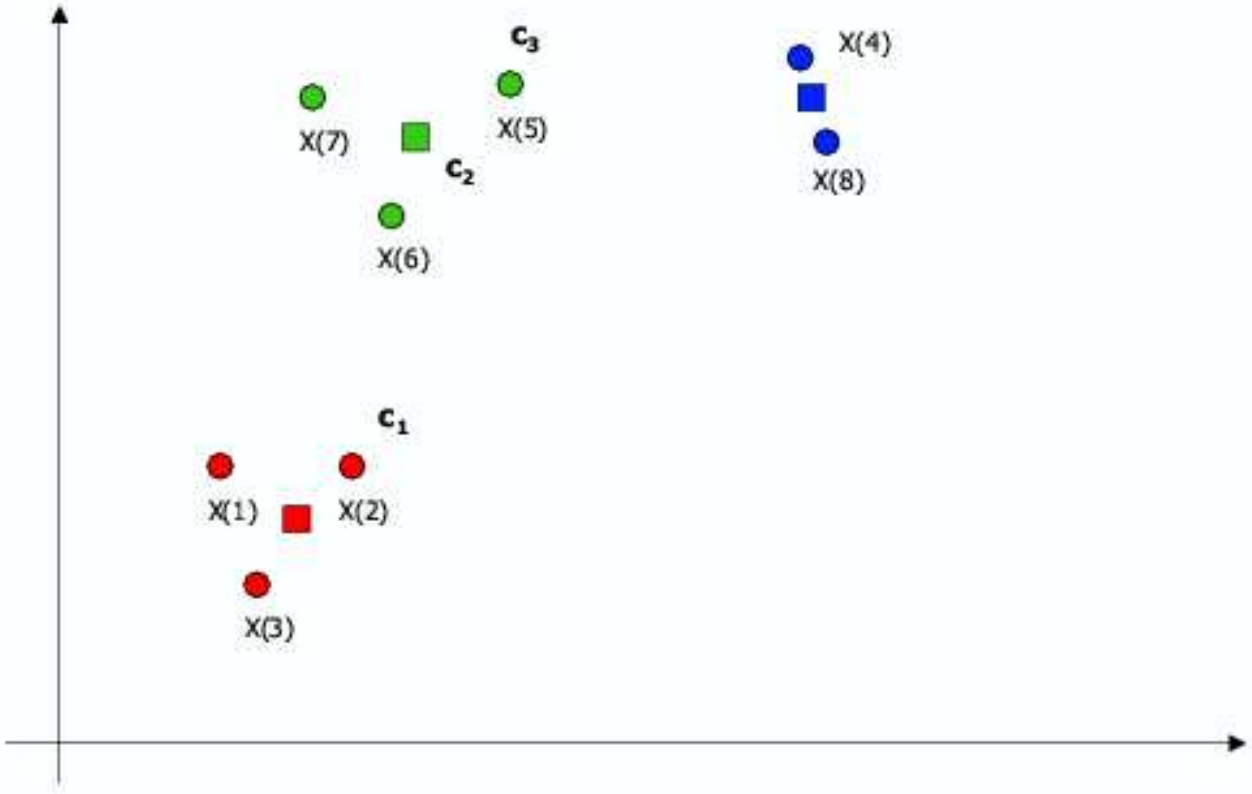
K-means example



K-means example



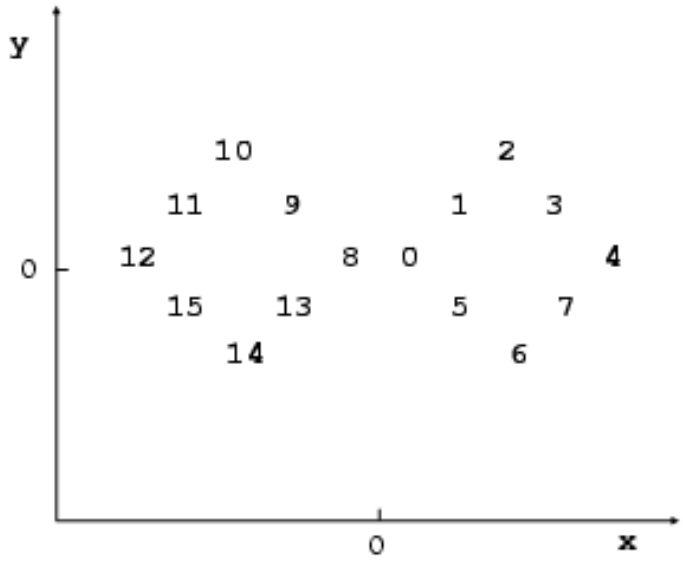
K-means example



K-means: Example 3

k-means: Example

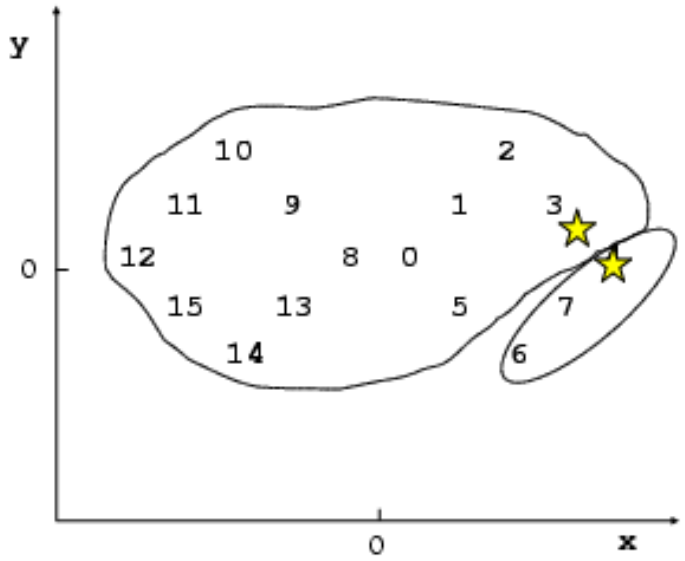
Id	x	y
0:	1.0	0.0
1:	3.0	2.0
2:	5.0	4.0
3:	7.0	2.0
4:	9.0	0.0
5:	3.0	-2.0
6:	5.0	-4.0
7:	7.0	-2.0
8:	-1.0	0.0
9:	-3.0	2.0
10:	-5.0	4.0
11:	-7.0	2.0
12:	-9.0	0.0
13:	-3.0	-2.0
14:	-5.0	-4.0
15:	-7.0	-2.0



- find the best 2 clusters

Seed: (9 0) (8 1)

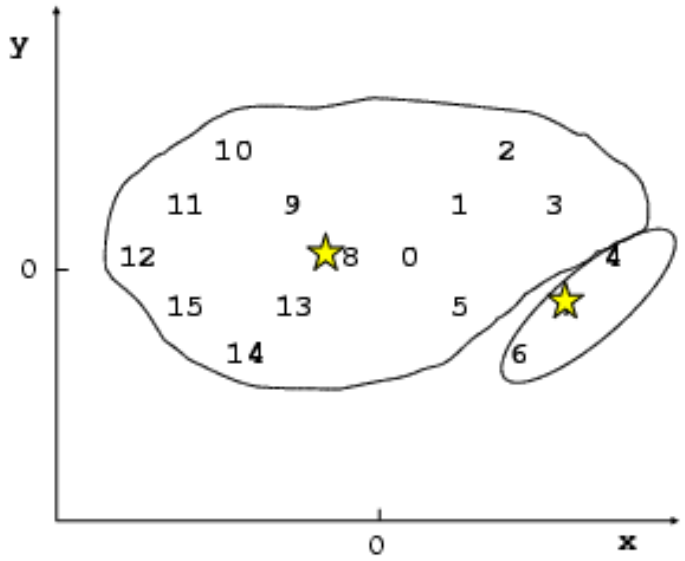
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887



Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

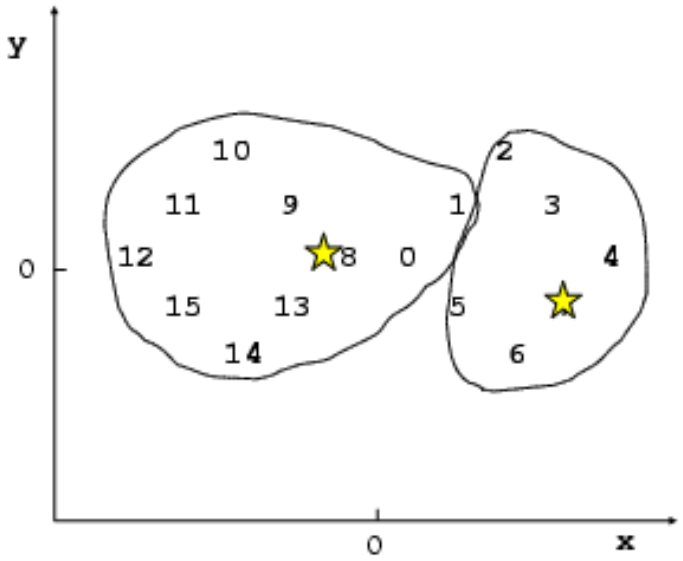
Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)



Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

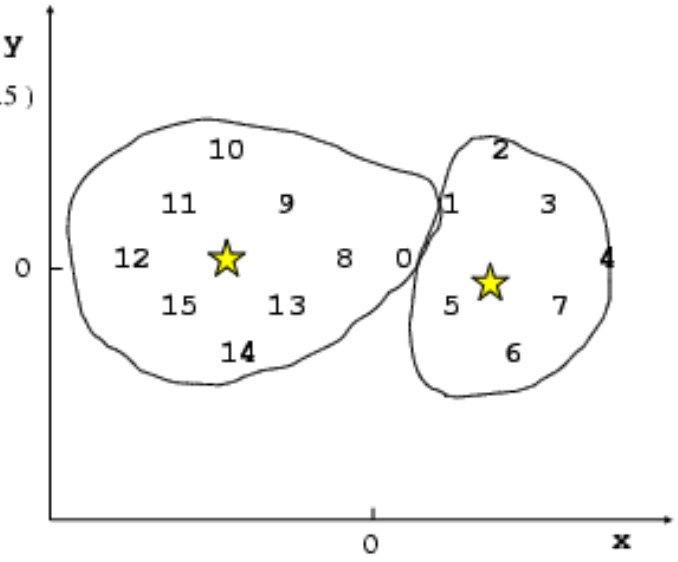


Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)

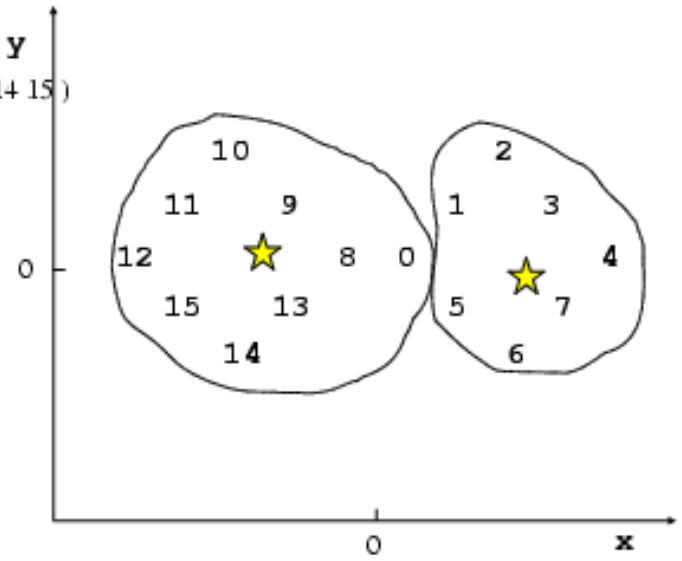


Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115



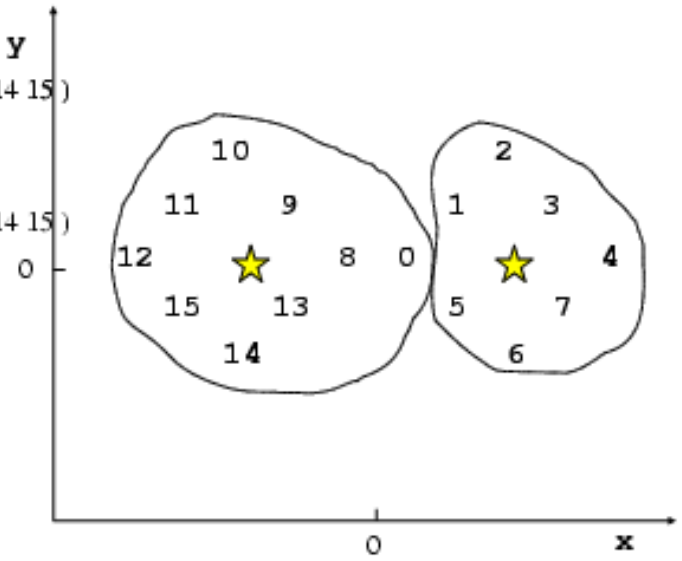
Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)



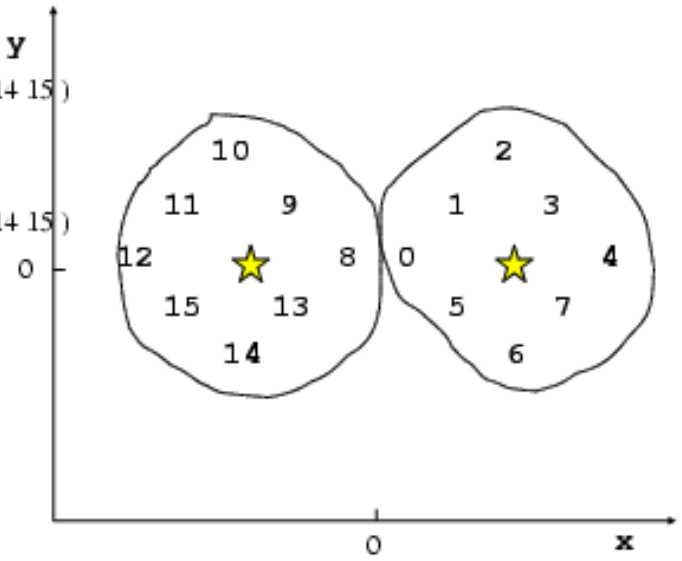
Seed: (9 0) (8 1)

Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
Cluster Centers: (5.0 0.0) (-5.0 0.0)
Average Distance: 3.41421



Seed: (9 0) (8 1)

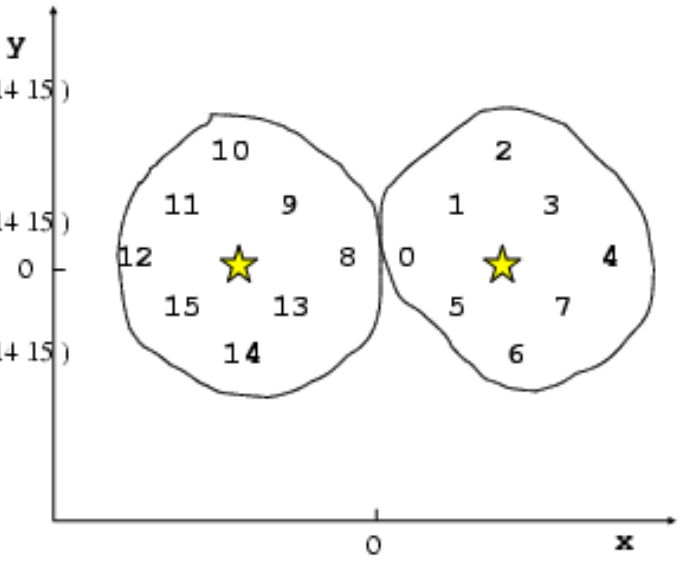
Clustering: (4 6 7) (0 1 2 3 5 8 9 10 11 12 13 14 15)
Cluster Centers: (7.0 -2.0) (-1.61538 0.46153)
Average Distance: 4.35887

Clustering: (2 3 4 5 6 7) (0 1 8 9 10 11 12 13 14 15)
Cluster Centers: (6.0 -0.33334) (-3.6 0.2)
Average Distance: 3.6928

Clustering: (1 2 3 4 5 6 7) (0 8 9 10 11 12 13 14 15)
Cluster Centers: (5.57143 0.0) (-4.33334 0.0)
Average Distance: 3.49115

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
Cluster Centers: (5.0 0.0) (-5.0 0.0)
Average Distance: 3.41421

Clustering: (0 1 2 3 4 5 6 7) (8 9 10 11 12 13 14 15)
No improvement.



Discussion

- Hierarchical methods.

  In the past hierarchical techniques were more popular with Ward's and average linkage probably being the best avaliable. Hierarchical methods have the advantage of being fast and therefore taking less computer time. However, with increasing computer power the personal computers can handle large datasets quite easily.

  Hierarchical methods are not amenable to analyzing very large samples. As sample size increases, the data storage requirement increases dramatically. E.g. a sample of 400 cases requires storage of approximately 80000 similarities. This increases to 125000 for a sample size of 500. This size exceed the capacity of most personal computers and thus, limiting the application in many instances. In such cases, a random sample of the original observations might be taken in order to reduce sample size. However, the question arise as to whether the new sample represents the original one.

  Hierarchical methods can be misleading because undesirable early combinations may persist throughout

the analysis and lead to artificial results. Of specific concern is the substantial impact of outliers particularly with the complete linkage method. To reduce this possibility the researcher may wish to cluster analyze the data several times, each time deleting problematic observations or outliers. The deletion of cases, however, even those not found to be outliers, can many times distort the solution. Thus, extreme care should be employed in the deletion of observations for any reason.

- Nonhierarchical methods.

  Nonhierarchical methods have gained increase acceptability and are applied increasingly. Their use, however, depends on the ability of the researcher to select the seed points according to some practical objective or theoretical basis. In these instances the Nonhierarchical methods have several advantages over hierarchical techniques. The results are less susceptible to the outliers in the data, the distance measure used, and the inclusion of irrelevant or inappropriate data. These benefits are realized only with the use of nonrandom seed points. Thus, the use of random seeds makes the nonhierarchical methods inferior to the

hierarchical techniques. Different seed points can yield different clusters. Which cluster should be selected_? Only through analysis and validation can the researcher select what is considered the best representation of structure, realizing there are many alternatives that may be as acceptable.

- Combination of both methods.

  Another approach is to use both methods (hierarchical and nonhierarchical) to gain the benefits of each. First, hierarchical methods can establish the number of clusters, profile the cluster centers, and identify any obvious outliers. After outliers are eliminated, the remaining observations can then be clustered by a nonhierarchical method with the cluster centers from the hierarchical results as the initial seed points. In this way, the advantages of the hierarchical methods are complemented by the ability of the nonhierarchical methods to *fine-tune* the results by allowing the switching of cluster memberships.