Contents

1 Introduction to Cluster Analysis **3**
1.1 Introduction to Cluster Analysis . . . 3
1.1.1 Types of Cluster Analysis . . . 4
1.1.2 Steps to conduct a Cluster Analysis . . . 4
1.1.3 Statistical Significance Testing . . . 5
1.1.4 Dendrograms . . . 6
1.2 Summary . . . 7
1.2.1 Hierarchical Clustering . . . 7
1.2.2 Things to Watch Out For . . . 8

2 Fundamentals of Cluster Analysis **11**
2.1 Hierarchical cluster analysis . . . 11
2.2 Clustering Linkage Algorithm : The Fundamentals . . . 12

3 Distance Measures and Standardization **14**
3.1 Cluster Analysis : Proximity Matrices . . . 14
3.2 Distance measures . . . 16
3.2.1 Distance measures . . . 16
3.2.2 Euclidean Distance . . . 17
3.2.3 Euclidean Distance : Worked Example . . . 18
3.2.4 Squared Euclidean distance . . . 18
3.2.5 Manhattan (City Block) Distance . . . 19
3.2.6 Other Measures . . . 19
3.3 Standardizing the Variables . . . 20
3.4 Standardized Distance . . . 22
3.4.1 Logarithmic Transformation . . . 23
3.5 Standardizing the Variables: SPSS Implementation . . . 23

4 Linkage **24**
4.1 Linkage Methods for Cluster Analysis . . . 24
4.2 Summary of Linkage methods . . . 24
4.2.1 Centroid method . . . 25
4.2.2 Nearest neighbour method . . . 25
4.2.3 Nearest neighbour method . . . 26
4.2.4 Furthest neighbour method . . . 27
4.2.5 Average (between groups) linkage method . . . 27
4.2.6 Wards method . . . 27

4.2.7 Wards Linkage method (IMPORTANT) 27

5 Implementation **28**
5.1 SPSS Implementation and Output 28
5.1.1 Proximity matrix 28
5.1.2 Cluster Membership 28
5.1.3 Icicle Plot 28
5.1.4 SPSS Agglomeration Schedule 29

6 Kmeans Clustering **32**
6.1 Non-Hierarchical Clustering (Kmeans) 32
6.2 K-Means Clustering 33
6.2.1 Initial Cluster Centres 33
6.2.2 Demonstration of k-means 35
6.2.3 Optimal Number of Clusters 36
6.2.4 Performance of k-means clustering 36

Chapter 1

Introduction to Cluster Analysis

1.1 Introduction to Cluster Analysis

- ***A cluster is a group of relatively homogeneous cases or observations.***
- Cluster analysis is a major technique for classifying a large volumes of information into manageable meaningful piles. Cluster analysis is a **data reduction** tool that creates subgroups that are more manageable than individual data items.
- The term cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories.A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop ***taxonomies***.
- **Cluster analysis** (CA) is an exploratory data analysis tool for organizing observed data into meaningful taxonomies, groups, or clusters, based on combinations of independent variables, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown.
- Both cluster analysis (and later, discriminant analysis) are concerned with **classification**. Each cluster thus describes, in terms of the data collected, the class to which its members belong. Items in each cluster are similar in some ways to each other and dissimilar to those in other clusters.
- In this sense, CA creates new groupings without any preconceived notion of what clusters may arise, whereas ***discriminant analysis*** classifies people and items into already known groups
(*Theory topic for later on : Supervised and Unsupervised Learning*).
- In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation.

- **Important:** Cluster analysis simply discovers structures in data without explaining why they exist. It examines the full complement of inter-relationships between variables. Cluster Analysis provides no explanation as to why the clusters exist nor is any interpretation made.
- However, the latter requires prior knowledge of membership of each cluster in order to classify new cases. In cluster analysis there is no prior knowledge about which elements belong to which clusters.
- The grouping or clusters are defined through an analysis of the data. Subsequent multivariate analyses can be performed on the clusters as groups.
- Cluster analysis is a tool of discovery revealing associations and structure in data which, though not previously evident, are sensible and useful when discovered. Importantly, CA enables new cases to be assigned to classes for identification and diagnostic purposes; or find ***exemplars*** to represent classes.

1.1.1 Types of Cluster Analysis

There are three main types of cluster analysis.

- Hierarchical Clustering Analysis
- Non-hierarchical Clustering Analysis (K-means clustering)
- Two Step Clustering Analysis

Within hierarchical clustering analysis there are two subcategories:

- Agglomerative (start from n clusters,to get to 1 cluster)
- Divisive (start from 1 cluster, to get to n cluster)

1.1.2 Steps to conduct a Cluster Analysis

1. Select a distance measure
2. Select a clustering algorithm
3. Determine the number of clusters
4. Validate the analysis

Because we usually don't know the number of groups or clusters that will emerge in our sample and because we want an optimum solution, a two-stage sequence of analysis occurs as follows:

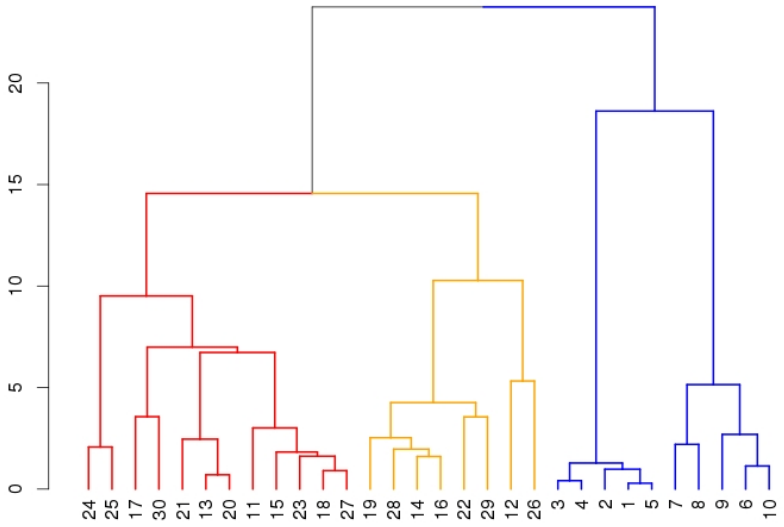
1. We carry out a hierarchical cluster analysis using **Ward' Method** applying squared ***Euclidean Distance*** as the distance or similarity measure. This helps to determine the optimum number of clusters we should work with.
2. The next stage is to rerun the hierarchical cluster analysis with our selected number of clusters, which enables us to allocate every case in our sample to a particular cluster.

1.1.3 Statistical Significance Testing

- Note that the previous discussions refer to clustering algorithms and do not mention anything about statistical significance testing.
- In fact, cluster analysis is not as much a typical statistical test as it is a "collection" of different algorithms that "put objects into clusters according to well defined similarity rules."
- The point here is that, unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any ***a priori hypotheses***, but are still in the exploratory phase of our research.
- In a sense, cluster analysis finds the "most significant solution possible." Therefore, statistical significance testing is really not appropriate here, even in cases when p-values are reported.

1.1.4 Dendrograms

- The dendrogram is a tree-structured graphical representation, used to visualize of the results of ***hierarchical cluster analysis***.
- This is a tree-like plot where each step of hierarchical clustering is represented as a joining (or fusion) of two branches of the tree into a single one.
- The branches represent clusters obtained on each step of hierarchical clustering.
- The result of a clustering is presented either as the ***distance*** or the similarity between the clustered rows or columns depending on the selected distance measure.



1.2 Summary

- **Cluster analysis** is a convenient method for identifying homogenous groups of objects called clusters. Objects (or cases, observations) in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster.
- There are three cluster analysis approaches: hierarchical methods, partitioning methods (more precisely, k-means), and two-step clustering, which is largely a combination of the first two methods.
- Each of these procedures follows a different approach to grouping the most similar objects into a cluster and to determining each objects cluster membership.
- Some approaches most notably hierarchical methods require us to specify how similar or different objects are in order to identify different clusters. Most software packages, such as SPSS, calculate a measure of (dis)similarity by estimating the distance between pairs of objects. Objects with smaller distances between one another are more similar, whereas objects with larger distances are more dissimilar.
- An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. This question is explored in the next step of the analysis. Sometimes, however, number of segments that have to be derived from the data will be known in advance.
- By choosing a specific clustering procedure, we determine how clusters are to be formed. (This always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables overall variance of objects in a specific cluster), or maximizing the distance between the objects or clusters). The procedure could also address the question of how to determine the (dis)similarity between objects in a newly formed cluster and the remaining objects in the dataset.

1.2.1 Hierarchical Clustering

- Hierarchical clustering procedures are characterized by the tree-like structure established in the course of the analysis. Most hierarchical techniques fall into a category called **agglomerative clustering**. In this category, clusters are consecutively formed from objects. Initially, this type of procedure starts with each object representing an individual cluster.
- These clusters are then sequentially merged according to their similarity. First, the two most similar clusters (i.e., those with the smallest distance between them) are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on. This allows a hierarchy of clusters to be established from the bottom up.
- A cluster hierarchy can also be generated top-down. In **divisive clustering**, all objects are initially merged into a single cluster, which is then gradually split up. Divisive procedures are quite rarely used in practice. We therefore concentrate on the agglomerative clustering procedures.

- This means that if an object is assigned to a certain cluster, there is no possibility of reassigning this object to another cluster. This is an important distinction between these types of clustering and partitioning methods such as ***k-means***.

1.2.2 Things to Watch Out For

- In statistics, the occurrence of several variables in a multiple regression model are **closely correlated** to one another, and carrying the same information, more or less. Multicollinearity can cause strange results when attempting to study how well individual independent variables contribute to an understanding of the dependent variable, often undermining the analysis.

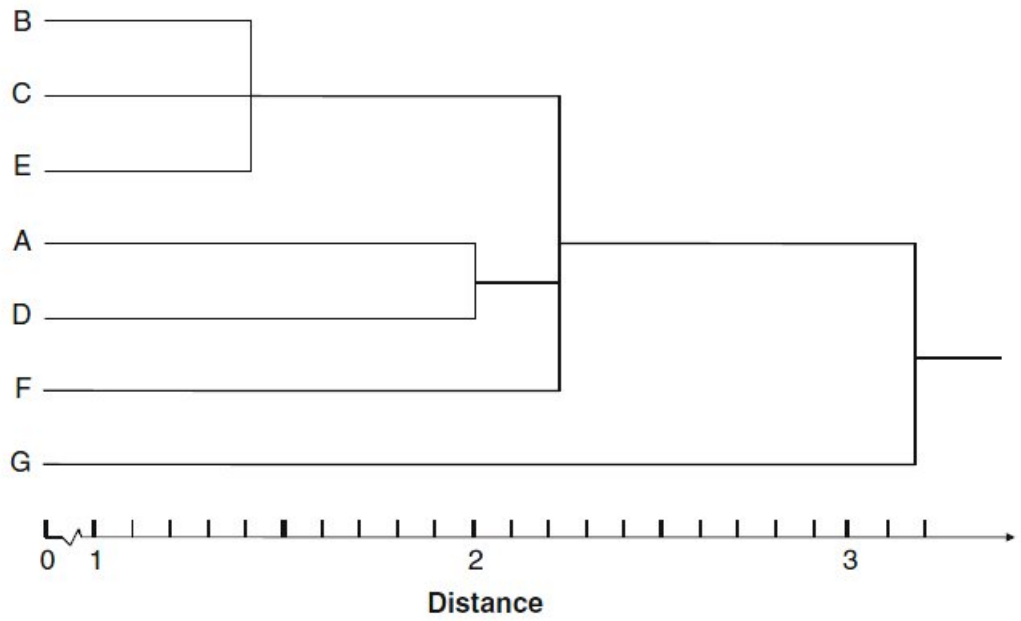
- In many analysis tasks, the variables under consideration are measured on different scales or levels. This would clearly distort any clustering analysis results. We can resolve this problem by ***standardizing*** the data prior to the analysis.

- Different standardization methods are available, such as the simple ***z standardization***, which re-scales each variable to have a mean of 0 and a standard deviation of 1.

- In most situations, however, ***standardization by range***(e.g., to a range of 0 to 1 or -1 to 1) is preferable. We recommend standardizing the data in general, even though this procedure can potentially reduce or inflate the variables influence on the clustering solution.

  An understanding of linkage method's other than than Ward method will be expected in the end of year examination.

- A common way to visualize the cluster analysiss progress is by drawing a dendrogram, which displays the distance level at which there was a combination of objects and clusters. Here is an example of a dendrogram (which corresponds to the example in the next section of material.



- An important question is how to decide on the number of clusters to retain from the data. Unfortunately, hierarchical methods provide only very limited guidance for making this decision. The only meaningful indicator relates to the distances at which the objects are combined. Similar to factor analysiss scree plot, we can seek a solution in which an

additional combination of clusters or objects would occur at a greatly increased distance. This raises the issue of what a great distance is, of course. For this purpose, we can make use of the dendrogram.

- In constructing the dendrogram, SPSS rescales the distances to a range of 025; that is, the last merging step to a one-cluster solution takes place at a (rescaled) distance of 25. The rescaling often lengthens the merging steps, thus making breaks occurring at a greatly increased distance level more obvious. Despite this, this distance-based decision rule does not work very well in all cases.

  It is often difficult to identify where the break actually occurs. This is also the case in our example above. By looking at the dendrogram, we could justify a two-cluster solution ([A,B,C,D,E,F] and [G]), as well as a five-cluster solution ([B,C,E], [A], [D], [F], [G]).

Chapter 2

Fundamentals of Cluster Analysis

2.1 Hierarchical cluster analysis

This is the major statistical method for finding relatively homogeneous clusters of cases based on measured characteristics.

Agglomerative clustering starts with each case as a separate cluster, i.e. there are as many clusters as cases, and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left.

The clustering method uses the dissimilarities or distances between objects when forming the clusters. The SPSS programme calculates ***distances*** between data points in terms of the specified variables.

A hierarchical tree diagram, called a ***dendrogram*** on SPSS, can be produced to show the linkage points. The clusters are linked at increasing levels of ***dissimilarity***. The actual measure of dissimilarity depends on the measure used.

2.2 Clustering Linkage Algorithm : The Fundamentals

To better understand how a clustering algorithm works, lets manually examine some of the single linkage procedures calculation steps. We start off by looking at the initial (Euclidean) distance matrix displayed previously.

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

- In the very first step, the two objects exhibiting the smallest distance in the matrix are merged. Note that we always merge those objects with the smallest distance, regardless of the clustering procedure (e.g., single or complete linkage). (N.B. In the following example, ties will be broken at random.)
- As we can see, this happens to two pairs of objects, namely B and C ($\mathbf{d(B, C)} = 1.414$), as well as C and E ($\mathbf{d(C, E)} = 1.414$). In the next step, we will see that it does not make any difference whether we first merge the one or the other, so lets proceed by forming a new cluster, using objects B and C.

Objects	A	B, C	D	E	F	G
A	0					
B, C	2.236	0				
D	2	2.236	0			
E	3.606	1.414	3	0		
F	4.123	3.162	2.236	2.828	0	
G	5.385	5.657	3.606	5.831	3.162	0

- Having made this decision, we then form a new distance matrix by considering the single linkage decision rule as discussed above. According to this rule, the distance from, for example, object A to the newly formed cluster is the minimum of $\mathbf{d(A, B)}$ and $\mathbf{d(A, C)}$. As $\mathbf{d(A, C)}$ is smaller than d(A, B), the distance from A to the newly formed cluster is equal to d(A, C); that is, 2.236.
- We also compute the distances from cluster [B,C] (clusters are indicated by means of squared brackets) to all other objects (i.e. D, E, F, G) and simply copy the remaining distances such as $\mathbf{d(E, F)}$ that the previous clustering has not affected.
- Continuing the clustering procedure, we simply repeat the last step by merging the objects in the new distance matrix that exhibit the smallest distance (in this case, the newly

Objects	A	B, C, E	D	F	G
A	0				
B, C, E	2.236	0			
D	2	2.236	0		
F	4.123	2.828	2.236	0	
G	5.385	5.657	3.606	3.162	0

formed cluster [B, C] and object E) and calculate the distance from this cluster to all other objects.

Objects	A, D	B, C, E	F	G
A, D	0			
B, C, E	2.236	0		
F	2.236	2.828	0	
G	3.606	5.657	3.162	0

- We continue in the same fashion until one cluster is left. By following the single linkage procedure, the last steps involve the merger of cluster [A,B,C,D,E,F] and object G at a distance of 3.162.

Objects	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	2.236	0	
G	3.606	3.162	0

Objects	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	3.162	0

Chapter 3

Distance Measures and Standardization

3.1 Cluster Analysis : Proximity Matrices

- A **proximity** is a measurement of the **similarity** or **dissimilarity**, broadly defined, of a pair of objects. If measured for all pairs of objects in a set (e.g. driving distances among a set of U.S. cities), the proximities are represented by an object-by-object proximity matrix
- The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items.
- A proximity is thought of as a similarity if the larger the value for a pair of objects, the closer or more alike we think they are. Examples of similarities are co-occurrences, interactions, statistical correlations and associations, social relations, and reciprocals of distances. A proximity is a dissimilarity if the smaller the value for a pair of objects, the closer or more alike we think of them. Examples are distances, differences, and reciprocals of similarities.
- Proximities are normally symmetric, so that the proximity of object a to object b is the same as the proximity of object b to object a. For example, the distance from Boston to NY is 206 miles, and the distance from NY to Boston is also 206 miles. However, in the case of one-way streets, it is possible for distances to be non-symmetric.
- For n items - the full proximity matrix is a symmetic square matrix in which the entry in cell (j, k) is some measure of the similarity (or distance) between the items to which row j and column k correspond.
- The main diagonal contains zeroes. There are $\frac{n}{2} \times n - 1$ calculations required.

Exercise: Using ***nearest neighbour*** linkage, describe how the agglomeration schedule based on the following proximity matrix. With nearest neighbour, a case is assigned to the cluster of the case with which it has the shortest distance. Cluster are also joined on this basis.

- The closest pair in terms of distance (2.84) are cases 3 and 8. So this is the first linkage.
- The next closest pair (3.18) are 8 and 9. The next linkage joins case 9 to 3 and 8.

Case	1	2	3	4	5	6	7	8	9	10
1	0.00	**4.82**	89.39	85.97	46.26	71.87	56.42	23.75	31.57	11.70
2	**4.82**	0.00	94.24	38.96	**5.55**	35.07	74.52	71.27	61.84	**4.84**
3	89.39	94.24	0.00	57.65	27.27	25.31	20.89	**2.84**	63.50	89.39
4	85.97	38.96	57.65	0.00	**22.94**	**7.13**	70.49	23.09	**12.75**	85.97
5	46.26	**5.55**	27.27	**22.94**	0.00	39.44	17.43	79.22	14.47	46.26
6	71.87	35.07	25.31	**7.13**	39.44	0.00	27.50	30.65	13.34	71.87
7	56.42	74.52	20.89	70.49	17.43	27.50	0.00	91.16	44.92	**6.42**
8	23.75	71.27	**2.84**	23.09	79.22	30.65	91.16	0.00	**3.18**	23.75
9	31.57	61.84	63.50	**12.75**	14.47	13.34	44.92	**3.18**	0.00	31.57
10	11.70	**4.84**	89.39	85.97	46.26	71.87	**6.42**	23.75	31.57	0.00

- The next closest pair (4.82) are 1 and 2. So this is the next linkage. [So far (3,8,9) and (2,10)]
- The next closest pair (4.84) are 2 and 10. The next linkage joins case 1 to 2 and 10.
- The next closest pair (5.55) are 2 and 5. The next linkage joins case 5 to 1, 2 and 10. [So far (3,8,9) and (1,2,5,10)]
- The next closest pair (6.42) are 7 and 10. The next linkage joins case 7 to 1, 2, 5 and 10.
- The next closest pair (7.13) are 4 and 6. The next linkage joins case 4 to 6. [So far (3,8,9), (4,6) and (1,2,5,10) All cases are in clusters. This is a 3 cluster solution.]
- The next closest pair (11.70) are 1 and 10. Disregard, because they are already clustered together.
- The next closest pair (19.44) are 4 and 9. This joins cluster (4,6) to cluster (3,8,9) [So far (3,4,6,8,9) and (1,2,5,10). This is a 2 cluster solution.]
- The next closest pairing is 4 and 5. This linkage joins all cases together in one cluster.

3.2 Distance measures

- Suppose in the previous example the rule for grouping a number of dinners was whether they shared the same table or not.
- These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects.
- If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the joining algorithm does not "care" whether the distances that are "fed" to it are actual real distances, or some other derived measure of distance that is more meaningful to the researcher; and it is up to the researcher to select the right method for his/her specific application. Squared Euclidean distance.

3.2.1 Distance measures

- Distance can be measured in a variety of ways. There are distances that are Euclidean (can be measured with a ruler) and there are other distances based on similarity.
- For example, in terms of geographical distance (i.e. Euclidean distance) Perth, Australia is closer to Jakarta, Indonesia, than it is to Sydney, Australia.
- However, if distance is measured in terms of the cities characteristics, Perth is closer to Sydney (e.g. both on a big river estuary, straddling both sides of the river, with surfing beaches, and both English speaking, etc).
- A number of distance measures are available within SPSS. The ***squared Euclidean distance*** is the most widely used measure.
- There are various measures to express (dis)similarity between pairs of objects. A straightforward way to assess two objects proximity is by drawing a straight line between them.This type of distance is also referred to as ***Euclidean distance*** (or straight-line distance) and is the most commonly used type when it comes to analyzing ratio or interval-scaled data.

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

The Euclidean distance is the square root of the sum of the squared differences in the variables values. Suppose B and C were positioned as $(7,6)$ and $(6,5)$ respectively.

$$d_{Euclidean}(B,C) = \sqrt{(6-5)^2 + (7-6)^2} = \sqrt{2} = 1.414$$

This distance corresponds to the length of the line that connects objects B and C. In this case, we only used two variables but we can easily add more under the root sign in the formula. However, each additional variable will add a dimension to our research problem (e.g., with ten clustering variables, we have to deal with ten dimensions), making it impossible to represent the solution graphically.

- The ***Squared Euclidean distance*** uses the same equation as the Euclidean distance metric, but does not take the square root. In the previous example, the squared Euclidean distance between B and C is 2. As a result, clustering with the Squared Euclidean distance is computationally faster than clustering with the regular Euclidean distance.

- We can compute the distance between all other pairs of objects. All these distances are usually expressed by means of a ***distance matrix***. In this distance matrix, the non-diagonal elements express the distances between pairs of objects and zeros on the diagonal (the distance from each object to itself is, of course, 0). In our example, the distance matrix is an 8×8 table with the lines and rows representing the objects under consideration.

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

- There are also alternative distance measures: The ***Manhattan distance*** or city-block distance uses the sum of the variables absolute differences. This is often called the Manhattan metric as it is akin to the walking distance between two points in a city like New Yorks Manhattan district, where the distance equals the number of blocks in the directions North-South and East-West. Using the points B and C that we used previously, the manhattan distance is computed as follows:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |6-5| + |7-6| = 2$$

3.2.2 Euclidean Distance

- The most straightforward and generally accepted way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances, an extension of Pythagoras's theorem.

- If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e. as if measured with a ruler).

- The Euclidean distance is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. The Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$$

- In a univariate example, the Euclidean distance between two values is the arithmetic difference, i.e. ***value1 - value2***. In the bivariate case, the minimum distance is the hypotenuse of a triangle formed from the points, as in Pythagoras's theorem.
- Although difficult to visualize, an extension of the Pythagoras's theorem will give the distance between two points in n-dimensional space.
- The Euclidean distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

3.2.3 Euclidean Distance : Worked Example

Compute the Euclidean Distance between the following points: $X = \{1, 5, 4, 3\}$ and $Y = \{2, 1, 8, 7\}$

x_j	y_j	$x_j - y_j$	$(x_j - y_j)^2$
1	2	-1	1
5	1	4	16
4	8	-4	16
3	7	-4	16
			49

The Euclidean Distance between the two points is $\sqrt{49}$ i.e. 7.

3.2.4 Squared Euclidean distance

The squared Euclidean distance is used more often than the simple Euclidean distance in order to place progressively greater weight on objects that are further apart.

The Squared Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k}(x_j - y_j)^2$$

The Squared Euclidean distance may be preferred to the Euclidean distance as it is slightly less computational complex, without loss of any information.

3.2.5 Manhattan (City Block) Distance

- The City-block (Manhattan) distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared).
- The City block distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k} |x_j - y_j|$$

- The City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.
- **Example**
  Compute the Manhattan Distance between the following points: $X = \{1, 3, 4, 2\}$ and $Y = \{5, 2, 5, 2\}$

x_j	y_j	$x_j - y_j$	$\lvert x_j - y_j \rvert$
1	5	-4	4
3	2	1	1
4	5	-1	1
2	2	0	0
			6

- The Manhattan Distance between the two points is 6.

3.2.6 Other Measures

- When working with metric (or ordinal) data, researchers frequently use the ***Chebychev distance***, which is the maximum of the absolute difference in the clustering variables values. This distance measure may be appropriate in cases when we want to define two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$\text{distance(x,y)} = \text{Maximum}|x_i - y_i|$$

  For B and C, this result is:

$$d_{Chebychec}(B, C) = \max(|x_B - x_C|, |y_B - y_C|) = \max(|6 - 5|, |7 - 6|) = 1$$

- **Power distance.** Sometimes we may want to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different. This can be accomplished via the power distance. The power distance is computed as:

$$\text{distance(x,y)} = (\sum_i |x_i - y_i|^p)^{1/r}$$

  Parameter p controls the progressive weight that is placed on differences on individual dimensions, parameter r controls the progressive weight that is placed on larger differences between objects. If r and p are equal to 2, then this distance is equal to the Euclidean distance.

  A few example calculations may demonstrate how this measure "behaves."

  * Parameter p controls the progressive weight that is placed on differences on individual dimensions
  * parameter r controls the progressive weight that is placed on larger differences between objects
  * If r and p are equal to 2, then this distance is equal to the Euclidean distance.

- **Percent disagreement.** This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature. This distance is computed as: distance(x,y) = (Number of xi yi)/ i

- There are other distance measures such as the Angular, Canberra or Mahalanobis distance. In many situations, the ***Mahalanobis distance*** is desirable as this measure compensates for ***multi-collinearity*** between the clustering variables. However, it is unfortunately not menu-accessible in SPSS.

3.3 Standardizing the Variables

- Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers).

- However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed.

- For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different.

Generally, it is good practice to transform the dimensions of all variables so they have similar scales.

- If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.
- In this example, both variables are measured on the same scale, so thats not much of a problem, assuming the judges use the scales similarly. But if you were looking at the distance between two people based on their IQs and incomes in dollars, you would probably find that the differences in incomes would dominate any distance measures.
- Variables that are measured in large numbers will contribute to the distance more than variables recorded in smaller numbers.
- In the hierarchical clustering procedure in SPSS, you can standardize variables in different ways. You can compute standardized scores or divide by just the standard deviation, range, mean, or maximum. This results in all variables contributing more equally to the distance measurement. Thats not necessarily always the best strategy, since variability of a measure can provide useful information.

3.4 Standardized Distance

- Let us consider measuring the distances between between two points using the three continuous variables pollution, depth and temperature. Let us suppose that a difference of 4.1 in terms of pollution is considered quite large and unusual, while a difference of 48 in terms of depth is large, but not particularly unusual.
- What would happen if we applied the Euclidean distance formula to measure distance between two cases.

Variables	case 1	case 2
Pollution	6.0	1.9
Depth	51	99
Temp	3.0	2.9

- Here is the calculation for Euclidean Distance:

$$d = \sqrt{(6.0 - 1.9)^2 + (51 - 99)^2 + (3.0 - 2.9)^2}$$

$$d = \sqrt{16.81 + 2304 + 0.01} = \sqrt{2320.82} = 48.17$$

The contribution of the second variable depth to this calculation is huge ,therefore one could say that the distance is practically just the absolute difference in the depth values (equal to $|51 - 99| = 48$) with only tiny additional contributions from pollution and temperature. These three variables are on completely different scales of measurement and the larger depth values have larger differences, so they will dominate in the calculation of Euclidean distances.

- The approach to take here is **standardization**, which is is necessary to balance out the contributions, and the conventional way to do this is to transform the variables so they all have the same variance of 1. At the same time we ***center*** the variables at their means, this centering is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare.
- The transformation commonly called standardization is thus as follows:

$$\text{standardized value} = \frac{\text{observed value - mean}}{\text{standard deviation}}$$

Variables	Case 1	Case 2	Mean	Std. Dev	Case 1 (std)	Case 2 (std)
Pollution	6.0	1.9	4.517	2.141	0.693	-1.222
Depth	51	99	74.433	15.615	-1.501	1.573
Temp	3.0	2.9	3.057	0.281	-0.201	-0.557

$$d_{std} = \sqrt{(0.693 - (-1.222))^2 + (-1.501 - 1.573)^2 + (-0.201 - (-0.557))^2}$$

$$d_{std} = \sqrt{3.667 + 9.449 + 0.127} = \sqrt{13.243} = 3.639$$

Pollution and temperature have higher contributions than before but depth still plays the largest role in this particular example, even after standardization. But this contribution is justified now, since it does show the biggest standardized difference between the samples.

3.4.1 Logarithmic Transformation

As as alternative to scaling or standardization, the user may opt to use the logarithm of a value, rather than the value itself.

3.5 Standardizing the Variables: SPSS Implementation

- If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.
- Variables that are measured in large numbers will contribute to the distance more than variables recorded in smaller numbers.
- In the hierarchical clustering procedure in SPSS, you can standardize variables in different ways.
- You can compute standardized scores or divide by just the standard deviation, range, mean, or maximum.
- This results in all variables contributing more equally to the distance measurement.
- Thats not necessarily always the best strategy, since variability of a measure can provide useful information.

Chapter 4

Linkage

4.1 Linkage Methods for Cluster Analysis

Having selected how we will measure distance, we must now choose the clustering algorithm, i.e. the rules that govern between which points distances are measured to determine cluster membership. There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data. This is important since it tells us that, although cluster analysis may provide an objective method for the clustering of cases, there can be subjectivity in the choice of method.

The linkage distances are calculated by SPSS. The goal of the clustering algorithm is to join objects together into successively larger clusters, using some measure of similarity or distance. SPSS provides seven clustering algorithms, the most commonly used one being ***Ward's method***.

4.2 Summary of Linkage methods

- Single linkage (minimum distance)
- Complete linkage (maximum distance)
- Average linkage

Ward's method

- Compute sum of squared distances within clusters
- Aggregate clusters with the minimum increase in the overall sum of squares

Centroid method

The distance between two clusters is defined as the difference between the centroids (cluster averages)

4.2.1 Centroid method

Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.

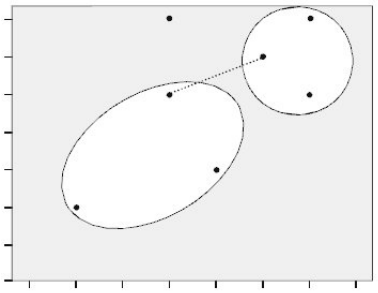
4.2.2 Nearest neighbour method

- A commonly used approach in hierarchical clustering is ***Wards linkage method***. This approach does not combine the two most similar objects successively. Instead, those objects whose merger increases the overall within-cluster variance to the smallest possible degree, are combined. If you expect somewhat equally sized clusters and the data set does not include outliers, you should always use Wards method.

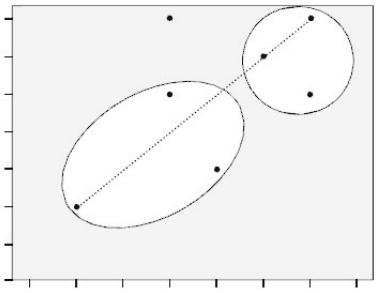
  We will use the Ward's linkage method for laboratory exercises.

- Other most popular agglomerative clustering procedures include the following:

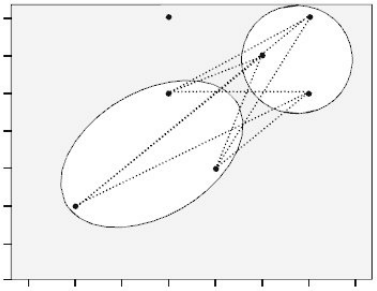
  Single linkage (nearest neighbor) : The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.



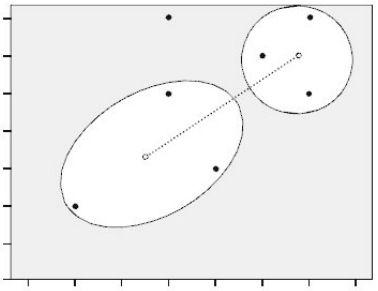
  Complete linkage (furthest neighbor) : The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.



  Average linkage : The distance between two clusters is defined as the average distance between all pairs of the two clusters members.



Centroid : In this approach, the geometric center (centroid) of each cluster is computed first. The distance between the two clusters equals the distance between the two centroids.



Each of these linkage algorithms can yield totally different results when used on the same data set, as each has its specific properties. As the single linkage algorithm is based on minimum distances, it tends to form one large cluster with the other clusters containing only one or few objects each. We can make use of this ***chaining effect*** to detect outliers, as these will be merged with the remaining objects usually at very large distances in the last steps of the analysis. Generally, single linkage is considered the most versatile algorithm.

Conversely, the complete linkage method is strongly affected by outliers, as it is based on maximum distances. Clusters produced by this method are likely to be rather compact and tightly clustered. The average linkage and centroid algorithms tend to produce clusters with rather low within-cluster variance and similar sizes. However, both procedures are affected by outliers, though not as much as complete linkage.

4.2.3 Nearest neighbour method

(Also known as the single linkage method).
In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but is often criticised because it doesnt take account of cluster structure and can result in a problem called chaining whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape.

4.2.4 Furthest neighbour method

(Also known as the complete linkage method).
In this case the distance between two clusters is defined to be the maximum distance between members i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

4.2.5 Average (between groups) linkage method

(sometimes referred to as Unweighted Pair Group Method with Arithmetic Mean (UPGMA)).
The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method.

4.2.6 Wards method

In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

4.2.7 Wards Linkage method (IMPORTANT)

In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

Chapter 5

Implementation

5.1 SPSS Implementation and Output

- **Hierarchical Cluster Analysis** is implemented by the **classify** option on the **analyse** menu. Three options shall appear. Select **Hierarchical**.
- We performed a hierarchical cluster analysis in SPSS, selecting all the variables (except categorical variables) in the **Variable(s)** box. We can label the cases by a categorical variable.
- We shall further requested the Dendrogram in the output. We changed all variables to z-scores to yield equal metrics and equal weighting, selected the **Squared Euclidean distance** (the default) method of determining distance between clusters and the **Ward's method** for clustering, and saved a 3-cluster solution as a new variable.

5.1.1 Proximity matrix

The output will print distances or similarities computed for any pair of cases. We will not be covering this in detail.

5.1.2 Cluster Membership

- This box allows you to specify a set number of clusters.
- If you have a hypothesis about how many clusters there are, you can specify a set number of clusters, or create a number of clusters within a range.

5.1.3 Icicle Plot

- Default choice by SPSS.
- Icicle plots visually represent information on the agglomeration schedule. You can select that all clusters are included in the icicle plot, or restrict it to a range of clusters.
- Also, you can read the plot from bottom up (vertical orientation) or from left to right (horizontal orientation).

5.1.4 SPSS Agglomeration Schedule

The procedure followed by cluster analysis at Stage 1 is to cluster the two cases that have the smallest squared Euclidean distance between them. Then SPSS will recompute the distance measures between all single cases and clusters (there is only one cluster of two cases after the first step). Next, the 2 cases (or clusters) with the smallest distance will be combined, yielding either 2 clusters of 2 cases (with 17 cases unclustered) or one cluster of 3 (with 18 cases unclustered). This process continues until all cases are clustered into a single group.

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	**1**	**3**	**,002**	**0**	**0**	**17**
2	13	18	2,708	0	0	9
3	12	17	4,979	0	0	14
4	20	21	5,014	0	0	7
5	11	14	8,509	0	0	10
6	5	8	11,725	0	0	8
7	19	20	11,871	0	4	14
8	2	5	13,174	0	6	13
9	13	15	14,317	2	0	12
10	**9**	**11**	**19,833**	**0**	**5**	**15**
11	6	7	22,901	0	0	15
12	10	13	23,880	0	9	16
13	2	4	28,378	8	0	17
14	**12**	**19**	**31,667**	**3**	**7**	**16**
15	6	9	40,470	11	10	18
16	10	12	44,624	12	14	19
17	1	2	47,720	1	13	20
18	6	16	49,963	15	0	19
19	6	10	64,785	18	16	20
20	1	6	115,781	17	19	0

Figure 5.1: SPSS Agglomeration Schedule

For the sake of clarify, we will explain Stages 1, 10, and 14.

Stage 1

- At Stage 1, Case 1 is clustered with Case 3. The squared Euclidean distance between these two cases is .002.
- Neither variable has been previously clustered (the two zeros under Cluster 1 and Cluster 2), and the next stage (when the cluster containing Case 1 combines with another case) is Stage 17.
- (Note that at Stage 17, Case 2 joins the Case-1 cluster.)

Stage 10

- At Stage 10, Case 9 joins the Case-11 cluster (Case 11 was previously clustered with Case 14 back in Stage 5, thus creating a cluster of 3 cases: Cases 9, 11, and 14).
- The squared Euclidean distance between Case 9 and Case-11 cluster is 19.833. Case 9 has not been previously clustered (the zero under Cluster 1), and Case 11 was previously clustered at Stage 5.
- The next stage (when the cluster containing Case 9 clusters) is Stage 15 (when it combines with the Case-6 cluster).

Stage 14

- At Stage 14, the clusters containing Cases 12 and 19 are joined, Case 12 has been previously clustered with Case 17, and Case 19 had been previously clustered with Cases 20 and 21, thus forming a cluster of 5 cases (Cases 12, 17, 19, 20, 21).
- The squared Euclidean distance between the two joined clusters is 31.667. Case 12 was previously joined at Stage 3 with Case 17. Case 19 was previously joined at Stage 7 with the Case- 20 cluster.
- The next stage when the Case-12 cluster will combine with another case/cluster is Stage 16 (when it joins with the Case-10 cluster).

The branching-type nature of the Dendrogram allows you to trace backward or forward to any individual case or cluster at any level. It, in addition, gives an idea of how great the distance was between cases or groups that are clustered in a particular step, using a 0 to 25 scale along the top of the chart. While it is difficult to interpret distance in the early clustering phases (the extreme left of the graph), as you move to the right relative distance become more apparent. The bigger the distances before two clusters are joined, the bigger the differences in these clusters. To find a membership of a particular cluster simply trace backwards down the branches to the name.

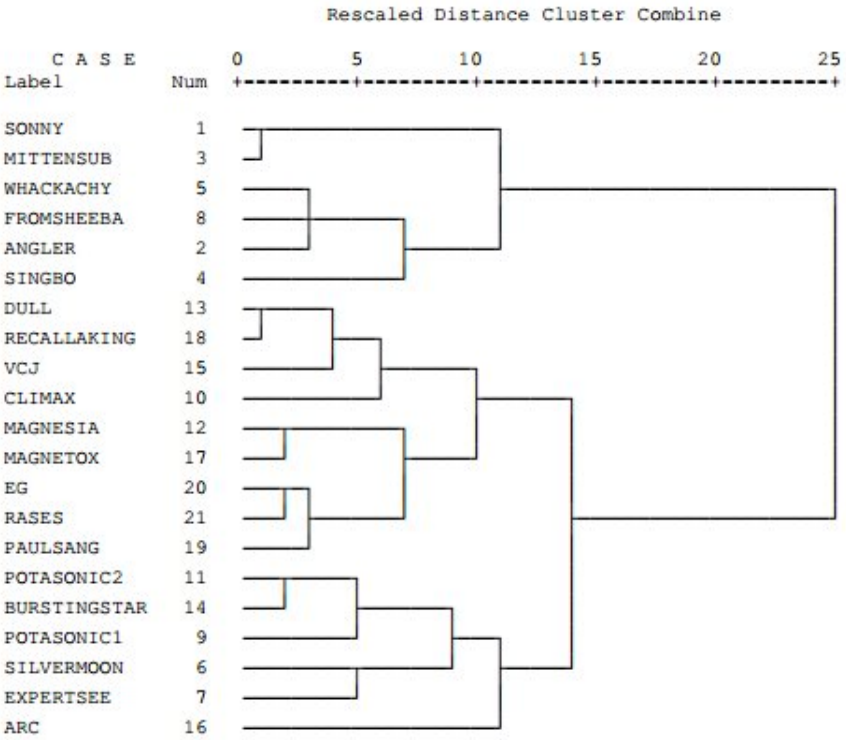


Figure 5.2: Corresponding Dendrogram

Chapter 6

Kmeans Clustering

6.1 Non-Hierarchical Clustering (Kmeans)

This method of clustering is very different from the hierarchical clustering and Ward method, which are applied when there is no prior knowledge of how many clusters there may be or what they are characterized by. The k-means clustering approach is used when you already have hypotheses concerning the number of clusters in your cases or variables. For example, you may want to specify exactly three clusters that are to be as distinct as possible.

This is the type of research question that can be addressed by the k-means clustering algorithm. In general, the k-means method will produce the exact k different clusters demanded of greatest possible distinction. Very often, both the hierarchical and the k-means techniques are used successively.

- Ward's method is used to get some sense of the possible number of clusters and the way they merge as seen from the dendrogram.
- Then the clustering is rerun with only a chosen optimum number in which to place all the cases (i.e. k means clustering).

In these methods the desired number of clusters is specified in advance and the 'best' solution is chosen. The steps in such a method are as follows:

1 Choose initial cluster centres (essentially this is a set of observations that are far apart each subject forms a cluster of one and its centre is the value of the variables for that subject).

2 Assign each subject to its 'nearest' cluster, defined in terms of the distance to the centroid.

3 Find the centroids of the clusters that have been formed

4 Re-calculate the distance from each subject to each centroid and move observations that are not in the cluster that they are closest to.

5 Continue until the centroids remain relatively stable.

Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster). Two disadvantages of non-hierarchical cluster analysis are:

1 it is often diffcult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times

2 it can be very sensitive to the choice of initial cluster centres. Again, it may be worth trying di?erent ones to see what impact this has.

6.2 K-Means Clustering

- Hierarchical clustering requires a distance or similarity matrix between all pairs of cases. That's an extremely large matrix if you have tens of thousands of cases in your data file.
- A clustering method that doesn't require computation of all possible distances is k-means clustering. It differs from hierarchical clustering in several ways. You have to know in advance the number of clusters you want. You can't get solutions for a range of cluster numbers unless you rerun the analysis for each different number of clusters.
- The algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis. In agglomerative hierarchical clustering, on the other hand, cases are added only to existing clusters. They are forever captive in their cluster, with a widening circle of "neighbours".
- The algorithm is called **k-means**, where **k** is the number of clusters you want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest.
- The k-means algorithm follows an entirely different concept than the hierarchical methods discussed before. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the ***within-cluster variation*** as a measure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such away that the within-cluster variation isminimized.Consequently,we do not need to decide on a distance measure in the first step of the analysis.
- The action in the algorithm centers around finding the k-means. You start out with an initial set of means and classify cases based on their distances to the centers.
- Next, you compute the cluster means again, using the cases that are assigned to the cluster; then, you reclassify all cases based on the new set of means. You keep repeating this step until cluster means don't change much between successive steps.
- Finally, you calculate the means of the clusters once again and assign the cases to their permanent clusters.

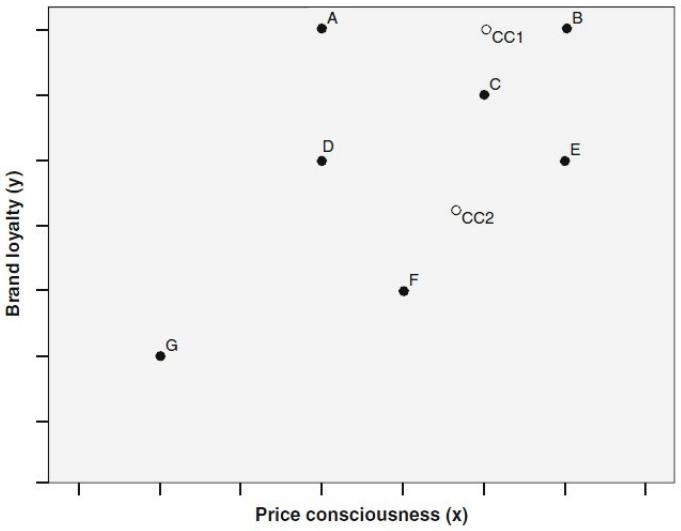
6.2.1 Initial Cluster Centres

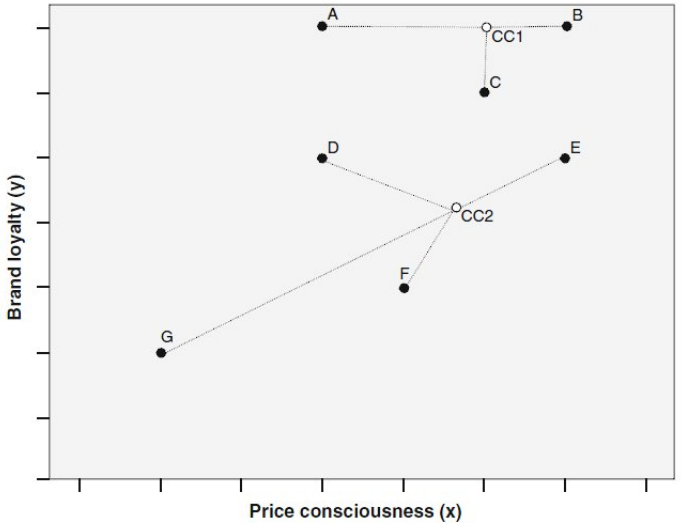
- The first step in k-means clustering is finding the k centres. This is done iteratively. You start with an initial set of centres and then modify them until the change between two iterations is small enough.

- If you have good guesses for the centres, you can use those as initial starting points; otherwise, you can let SPSS find k cases that are well separated and use these values as initial cluster centers. (i.e. The clustering process starts by randomly assigning objects to a number of clusters).
- K-means clustering is very sensitive to outliers, since they will usually be selected as initial cluster centers. This will result in outliers forming clusters with small numbers of cases. Before you start a cluster analysis, screen the data for outliers and remove them from the initial analysis. The solution may also depend on the order of the cases in the data.
- After the initial cluster centers have been selected, each case is assigned to the closest cluster, based on its distance from the cluster centers. After all of the cases have been assigned to clusters, the cluster centers are recomputed, based on all of the cases in the cluster.
- The cases are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an case to another cluster decreases the within-cluster variation, this case is reassigned to that cluster.
- Case assignment is done again, using these updated cluster centers. You keep assigning cases and recomputing the cluster centers until no cluster center changes appreciably or the maximum number of iterations (10 by default) is reached.

6.2.2 Demonstration of k-means

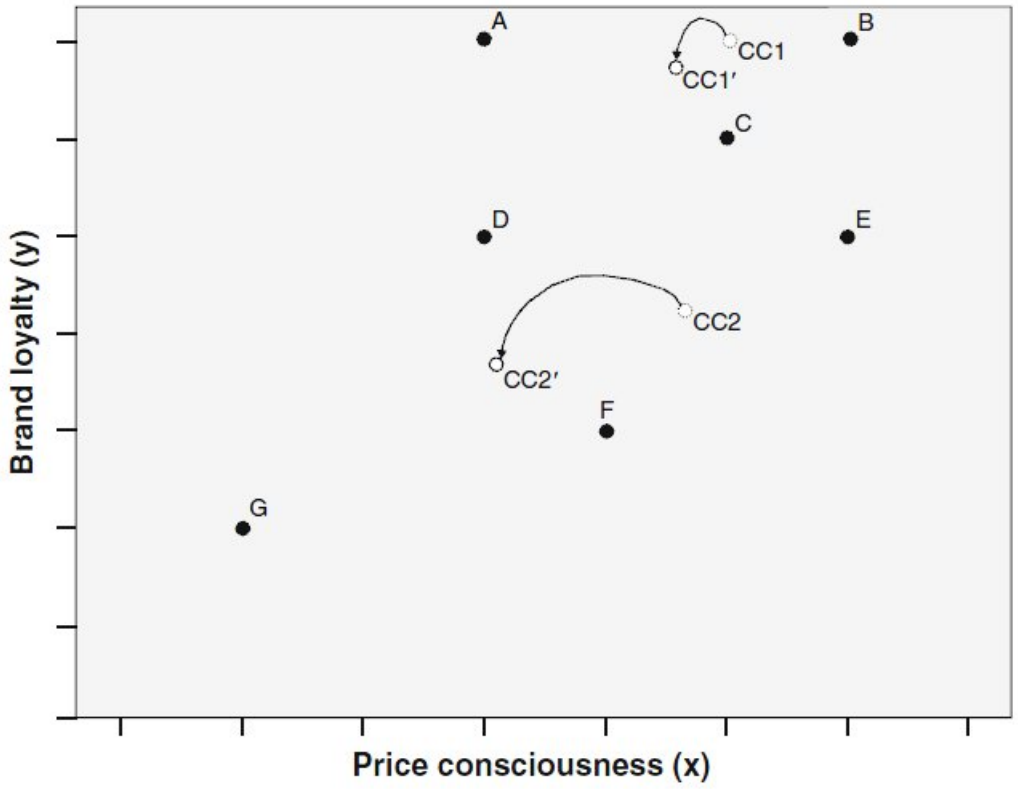
In this example, two cluster centers are randomly initiated, which CC1 (first cluster) and CC2 (second cluster). Euclidean distances are computed from the cluster centers to every single





object. Each object is then assigned to the cluster center with the shortest distance to it.

In this example, objects A, B, and C are assigned to the first cluster, whereas objects D, E, F, and G are assigned to the second. We now have our initial partitioning of the objects into two clusters. Based on this initial partition, each clusters geometric center (i.e., its centroid) is computed (third step). This is done by computing the mean values of the objects contained in the cluster (e.g., A, B, C in the first cluster) regarding each of the variables (in this example: price consciousness and brand loyalty). Both clusterscenters now shift into new positions (CC1 for the first and CC2 for the second cluster). In the fourth step, the distances from each object



to the newly located cluster centers are computed and objects are again assigned to a certain cluster on the basis of their minimum distance to other cluster centers (CC1 and CC2).

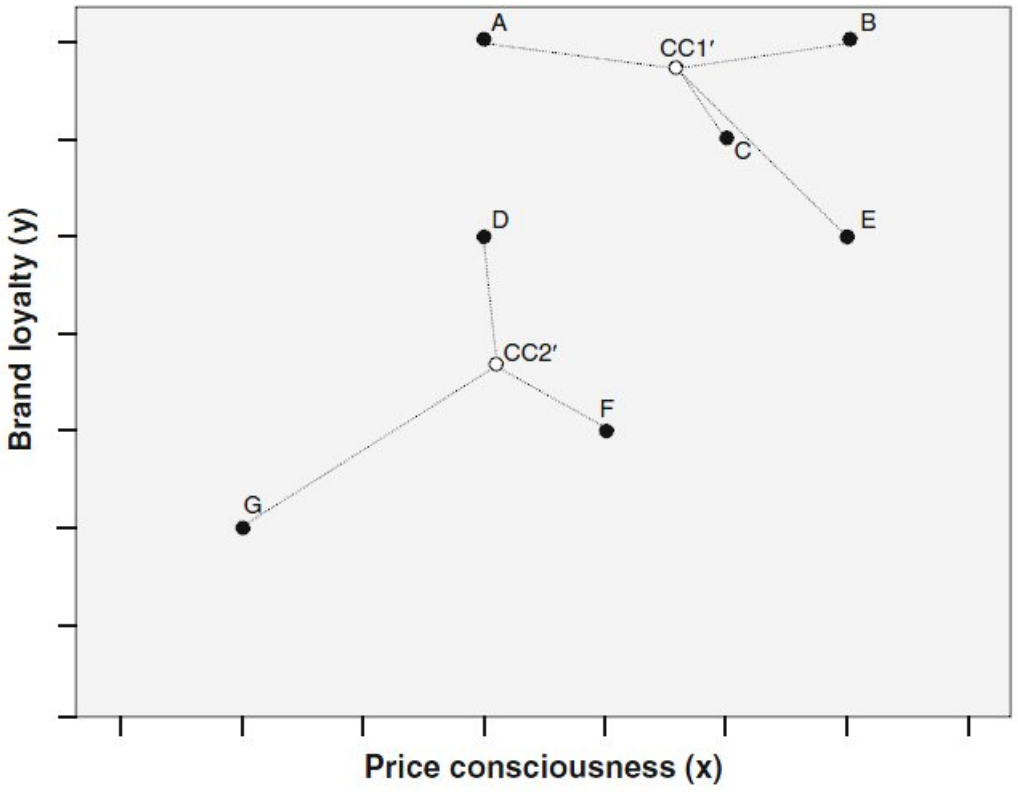
Since the cluster centers position changed with respect to the initial situation in the first step, this could lead to a different cluster solution. This is also true of our example, as object E is now unlike in the initial partition closer to the first cluster center (CC1) than to the second (CC2). Consequently, this object is now assigned to the first cluster. The k-means procedure now repeats the third step and re-computes the cluster centers of the newly formed clusters, and so on. In other words, steps 3 and 4 are repeated until a predetermined number of iterations are reached, or convergence is achieved (i.e., there is no change in the cluster affiliations).

6.2.3 Optimal Number of Clusters

One of the biggest problems with cluster analysis is identifying the optimum number of clusters. As the joining process continues, increasingly dissimilar clusters must be joined. i.e. the classification becomes increasingly artificial. Deciding upon the optimum number of clusters is largely subjective, although looking at a dendrogram would help.

6.2.4 Performance of k-means clustering

- Generally, k-means is superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant clustering variables. Furthermore, k-means can be applied to



very large data sets, as the procedure is less computationally demanding than hierarchical methods.

- In fact, we suggest definitely using k-means for sample sizes above 500, especially if many clustering variables are used. From a strictly statistical viewpoint, k-means should only be used on interval or ratio-scaled data as the procedure relies on Euclidean distances. However, the procedure is routinely used on ordinal data as well, even though there might be some distortions.
- One problem associated with the application of k-means relates to the fact that the researcher has to pre-specify the number of clusters to retain from the data. This makes k-means less attractive to some and still hinders its routine application in practice.