# CLUSTER ANALYSIS

## 1) Introduction

Cluster analysis is a technique used to combine observations into groups or clusters such that:

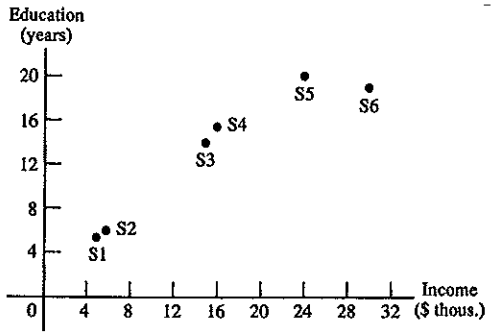
- Each cluster is homogeneous or compact with respect to certain characteristics, so that observations in each group are similar to each other.
- Each group should be different from other groups with respect to the same characteristics.

Let us take an example:

Table 7.1: Hypothetical data

| Subject ID | Income ($ thous.) | Education (years) |
|---|---|---|
| S1 | 5 | 5 |
| S2 | 6 | 6 |
| S3 | 15 | 14 |
| S4 | 16 | 15 |
| S5 | 25 | 20 |
| S6 | 30 | 19 |

Figure 7.1: Geometric view of cluster analysis




We need some measure of similarity to cluster observations. For example, we can use the squared Euclidian distance between subjects S1 and S2, $D_{12}^2 = (5-6)^2 + (5-6)^2 = 2$.

In general, with $p$ variables we can compute $D_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ (Eq 7.1)

where subjects are $i, j$. Therefore we obtain a similarity matrix table, as shown below:

***Table 7.2*** **Similarity Matrix Containing Euclidean Distances**

| | S1 | S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|---|
| S1 | 0.00 | 2.00 | 181.00 | 221.00 | 625.00 | 821.00 |
| S2 | 2.00 | 0.00 | 145.00 | 181.00 | 557.00 | 745.00 |
| S3 | 181.00 | 145.00 | 0.00 | 2.00 | 136.00 | 250.00 |
| S4 | 221.00 | 181.00 | 2.00 | 0.00 | 106.00 | 212.00 |
| S5 | 625.00 | 557.00 | 136.00 | 106.00 | 0.00 | 26.00 |
| S6 | 821.00 | 745.00 | 250.00 | 212.00 | 26.00 | 0.00 |

## 2) Hierarchical Clustering (HC)

Table 7.2 shows that subjects S1 and S2 are similar to each other, as are subjects S3 and S4. Either od these pairs could be selected (the tie is broken randomly). Let's choose S1 and S2 and merge them into one cluster. So we obtain 5 clusters (5 Cs); S1 and S2 $\in$ C1, and each other S forms the remaining Cs.

With HC, there are a number of methods to compute the similarity distance between clusters:

*a) Centroid Method*

In this method, the distance between clusters is obtained from computing the squared Euclidian distance between the centroids of the clusters.

***Table 7.3*** **Centroid Method: Five Clusters**

***Data for Five Clusters***

| Cluster | Cluster Members | Income ($ thous.) | Education (years) |
|---|---|---|---|
| 1 | S1&S2 | 5.5 | 5.5 |
| 2 | S3 | 15.0 | 14.0 |
| 3 | S4 | 16.0 | 15.0 |
| 4 | S5 | 25.0 | 20.0 |
| 5 | S6 | 30.0 | 19.0 |

***Similarity Matrix***

| | S1&S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|
| S1&S2 | 0.00 | 162.50 | 200.50 | 590.50 | 782.50 |
| S3 | 162.50 | 0.00 | 2.00 | 135.96 | 250.00 |
| S4 | 200.50 | 2.00 | 0.00 | 106.00 | 212.00 |
| S5 | 590.50 | 135.96 | 106.00 | 0.00 | 26.00 |
| S6 | 782.50 | 250.00 | 212.00 | 26.00 | 0.00 |

***Table 7.4*** **Centroid Method: Four Clusters**

***Data for Four Clusters***

| Cluster | Cluster Members | Income ($ thous.) | Education (years) |
|---|---|---|---|
| 1 | S1&S2 | 5.5 | 5.5 |
| 2 | S3&S4 | 15.5 | 14.5 |
| 3 | S5 | 25.0 | 20.0 |
| 4 | S6 | 30.0 | 19.0 |

***Similarity Matrix***

| | S1&S2 | S3&S4 | S5 | S6 |
|---|---|---|---|---|
| S1&S2 | 0.00 | 181.00 | 590.50 | 782.50 |
| S3&S4 | 181.00 | 0.00 | 120.50 | 230.50 |
| S5 | 590.50 | 120.50 | 0.00 | 26.00 |
| S6 | 782.50 | 230.50 | 26.00 | 0.00 |

*Table 7.5* **Centroid Method: Three Clusters**

*Data for Three Clusters*

| Cluster | Cluster Members | Income ($ thous.) | Education (years) |
|---|---|---|---|
| 1 | S1&S2 | 5.5 | 5.5 |
| 2 | S3&S4 | 15.5 | 14.5 |
| 3 | S5&S6 | 27.5 | 19.5 |

*Similarity Matrix*

| | S1&S2 | S3&S4 | S5&S6 |
|---|---|---|---|
| S1&S2 | 0.00 | 181.00 | 680.00 |
| S3&S4 | 181.00 | 0.00 | 169.00 |
| S5&S6 | 680.00 | 169.00 | 0.00 |

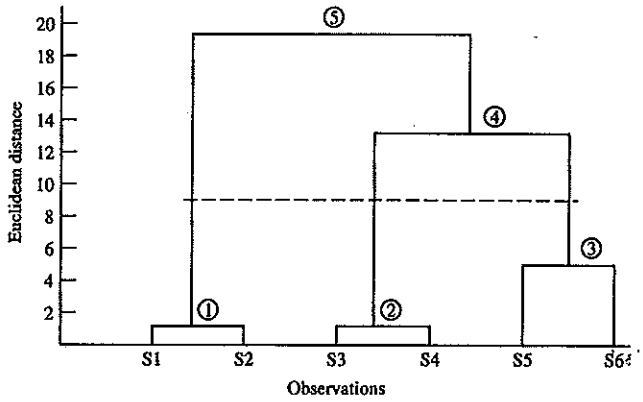



**Figure 7.2** Dendrogram for hypothetical data.

The various steps of the hierarchical clustering process are represented graphically in a dendrogram (or tree) as shown in Figure 7.2 above.

*b) Single-Linkage or Nearest-Neighbor Method*

In this method, the distance between two clusters is given by the minimum of the distance between all possible pairs of subjects in the two clusters. For example, the distance between C1 and S3 is the minimum of $D^2_{13} = 181$ and $D^2_{23} = 145$.

Similarly, the distance between cluster 1 and subject S4 is the minimum of the following distances:

$$D^2_{14} = 221 \quad \text{and} \quad D^2_{24} = 181.$$

This procedure results in the following similarity matrix of squared euclidean distances:

| | S1&S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|
| S1&S2 | 0.00 | 145.00 | 181.00 | 557.00 | 745.00 |
| S3 | 145.00 | 0.00 | 2.00 | 136.00 | 250.00 |
| S4 | 181.00 | 2.00 | 0.00 | 106.00 | 212.00 |
| S5 | 557.00 | 136.00 | 106.00 | 0.00 | 26.00 |
| S6 | 745.00 | 250.00 | 212.00 | 26.00 | 0.00 |

The next step is to merge subjects S3 and S4 to form a new cluster and develop a new similarity matrix. The squared euclidean distance between cluster 1 (consisting of subjects S1 and S2) and cluster 2 (consisting of subjects S3 and S4) is the minimum of the following distances: $D^2_{13}, D^2_{14}, D^2_{23}$, and $D^2_{24}$.

*c) Complete-Linkage or Farthest-Neighbor Method*

In this method, the distance between two clusters is given by the maximum of the distance between all possible pairs of subjects in the two clusters. For example, the distance between C1 and S3 is the maximum of $D_{13}^2 = 181$ and $D_{23}^2 = 145$. The distance between C1 and S5 is the maximum of $D_{15}^2 = 625$ and $D_{25}^2 = 557$.

**Following the above rule the similarity matrix after the first step (i.e., the five-cluster solution) is**

| | S1&S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|
| S1&S2 | 0.00 | 181.00 | 221.00 | 625.00 | 821.00 |
| S3 | 181.00 | 0.00 | 2.00 | 136.00 | 250.00 |
| S4 | 221.00 | 2.00 | 0.00 | 106.00 | 212.00 |
| S5 | 625.00 | 136.00 | 106.00 | 0.00 | 26.00 |
| S6 | 821.00 | 250.00 | 212.00 | 26.00 | 0.00 |

*d) Average-Link Method*

In this method, the distance between two clusters is given by the average of the distance between all possible pairs of subjects in the two clusters. For example, the distance between C1 and S3 is (181+145)/2=163.

| | S1&S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|
| S1&S2 | 0.00 | 163.00 | 201.00 | 591.00 | 783.00 |
| S3 | 163.00 | 0.00 | 2.00 | 136.00 | 250.00 |
| S4 | 201.00 | 2.00 | 0.00 | 106.00 | 212.00 |
| S5 | 591.00 | 136.00 | 106.00 | 0.00 | 26.00 |
| S6 | 783.00 | 250.00 | 212.00 | 26.00 | 0.00 |

**Once again, the second cluster is formed by combining subjects S3 and S4.**

*e) Ward's Method*

Clusters are formed not by computing distances but rather by maximizing within clusters homogeneity thus minimizing the within-group sum of squares (ESS).

Initially each observation is a cluster and thus ESS is zero. Next, form 5C, one C of size 2 and the other 4C of size 1 see Table 7.6 below). For example, the ESS for the cluster with 2 subjects (S1 and S2) is $(5-5.5)^2 + (6-5.5)^2 + (5-5.5)^2 + (6-5.5)^2 = 1.0$ and the ESS for the

remaining 4 C is 0 at each cluster of 1 S. Thus the total ESS for the cluster is 1.

Table 7.6 below gives all 5C solutions with their ESS. We can select C1 and merge S1 and S2.

The next step is to form 4C. For example, the ESS for the cluster that includes S1, S2, and S3 is $(5-8.67)^2+(6-8.67)^2+(15-8.67)^2+(5-8.33)^2+(6-8.33)^2+(14-8.33)^2=109.33$, which is the ESS for the 1st four-cluster solution. Cluster solution 5 has the min ESS.

This procedure is repeated for all remaining steps (ie: all possible 3C solutions, 2C, etc.)

***Table 7.6*** **Ward's Method**

| | Members in Cluster | | | | | |
|---|---|---|---|---|---|---|
| **Cluster Solution** | **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 |

### 3) Non-Hierarchical Clustering *(k-means Method)*

To implement the k-means clustering method, we follow these steps:

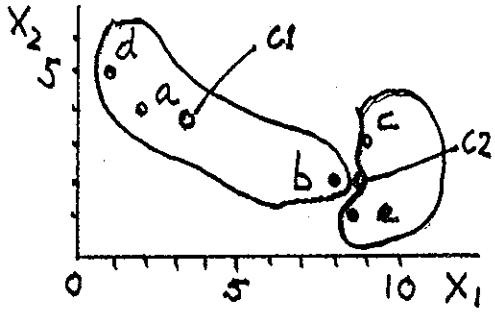
- Specify the # of clusters and, arbitrarily or deliberately, the members of each cluster.
- Calculate each cluster's centroid and the distances between each observation and centroid. If an observation is nearer the centroid of a cluster other than the one to which it currently belongs, re-assign it to the nearer cluster.
- Repeat step (2) until all obs. Are nearest the centroid of a cluster to which they belong.
- If the # of C cannot be specified with confidence, repeat steps (1) to (3) with a different number of C and evaluate the results.

Example: Daily expenditures on food (X1) and clothing (X2)

| Person | X1 | X2 |
|---|---|---|
| *a* | 2 | 4 |
| *b* | 8 | 2 |
| *c* | 9 | 3 |
| *d* | 1 | 5 |
| *e* | 8.5 | 1 |

Suppose we start with forming 2 C, by arbitrarily assigning *a, b* and *d* to Cluster 1 and *c* and *e* to Cluster 2. The cluster centroids are calculated in the following table:

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| Obs. | $X_1$ | $X_2$ | Obs. | $X_1$ | $X_2$ |
| *a* | 2 | 4 | *c* | 9 | 3 |
| *b* | 8 | 2 | *e* | 8.5 | 1 |
| *d* | 1 | 5 | | | |
| Ave. | 3.67 | 3.67 | Ave. | 8.75 | 2 |




Note: Cluster centroid is the point with coordinates equal to the average values of the variables for the obs. In that cluster (eg: centoid of cluster one is the point with X1=3.67 and X2=3.67).

Now let's calculate the distance between $a$ and the two centroids:

$$D(a,abd) = sqrt((2-3.67)^2 + (4-3.67)^2) = 1.702$$

$$D(a,ce) = sqrt((2-8.75)^2 + (4-2)^2) = 7.040$$

Since $a$ is closer to centroid of C1 (to which it currently belongs), thus $a$ will not be reassigned.

Next, we calculate the distance between $b$ and the two centroids:

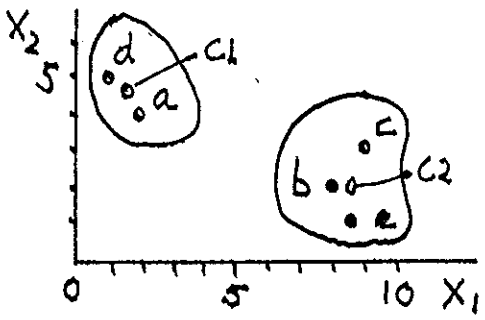
$$D(b,abd) = sqrt((8-3.67)^2 + (2-3.67)^2) = 4.641$$

$$D(b,ce) = sqrt((8-8.75)^2 + (2-2)^2) = 0.750$$

Since b is closer to Cluster 2's centroid than to that of C1, it is reassigned to C2. The new cluster centroids are calculated below.

Table 3.2 shows the final results where every obs. belongs to the C to which it is nearest and the *k-mean* method stops.

| Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|
| Obs. | $X_1$ | $X_2$ | Obs. | $X_1$ | $X_2$ |
| $a$ | 2 | 4 | $c$ | 9 | 3 |
| $d$ | 1 | 5 | $e$ | 8.5 | 1 |
| | | | $b$ | 8 | 2 |
| Ave. | 1.5 | 4.5 | Ave. | 8.5 | 2 |




| Obs. | Distance from Cluster 1 | Distance from Cluster 2 |
|---|---|---|
| $a$ | 0.707* | 6.801 |
| $b$ | 6.964 | 0.500* |
| $c$ | 7.649 | 1.118* |
| $d$ | 0.707* | 8.078 |
| $e$ | 7.826 | 1.000* |

Table 3.2

## 4) Minitab Examples

## Example 3

Manly (1986) reports 6 measurements made on prehistoric goblets from Thailand for 25 Goblets

X1 = "Mouth Width" X2 = "Total Width" X3 = "Total Height"
X4 = "Base Width" X5 = "Stem Width" X6 = "Stem Height";

1 13 21 23 14 7 8
2 14 14 24 19 5 9
3 19 23 24 20 6 12
4 17 18 16 16 11 8
5 19 20 16 16 10 7
6 12 20 24 17 6 9
7 12 19 22 16 6 10
8 12 22 25 15 7 7
9 11 15 17 11 6 5
10 11 13 14 11 7 4
11 12 20 25 18 5 12
12 13 21 23 15 9 8
13 12 15 19 12 5 6
14 13 22 26 17 7 10
15 14 22 26 15 7 9
16 14 19 20 17 5 10
17 15 16 15 15 9 7
18 19 21 20 16 9 10
19 12 20 26 16 7 10
20 17 20 27 18 6 14
21 13 20 27 17 6 9
22 9 9 10 7 4 3
23 8 8 7 5 2 2
24 9 9 8 4 2 2
25 12 19 27 18 5 12