Contents

1 Cluster Analysis : Introductory Concepts

- Cluster analysis is a convenient method for identifying homogenous groups of objects called clusters.
- A cluster is a group of relatively homogeneous cases or observations. 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, *essentially 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.
- Software packages, such as `R` 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.
- **Number of Clusters:** An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. 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.

2 Statistical Considerations

2.1 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.

2.2 Multicollinearity

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. Multi-collinearity 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.

3 Types of 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 this 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.
- A consequence of the Hierarchical Clustering method is that 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****.*

Summary 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)

4 Hierarchical agglomerative clustering methods

- Another class of clustering methods, known as ***hierarchical agglomerative clustering methods***, starts out by putting each observation into its own separate cluster. It then examines all the distances between all the observations and pairs together the two closest ones to form a new cluster.
- This is a simple operation, since hierarchical methods require a distance matrix, and it represents exactly what we want - the distances between individual observations.
- So finding the first cluster to form simply means looking for the smallest number in the distance matrix and joining the two observations that the distance correspnds to into a new cluster. Now there is one less cluster than there are observations.

4.1 Agglomeration Methods

- To determine which observations will form the next cluster, we need to come up with a method for finding the distance between an existing cluster and individual observations, since once a cluster has been formed, we'll determine which observation will join it based on the distance between the cluster and the observation. Some of the methods that have been proposed to do this are to take the minimum distance between an observation and any member of the cluster, to take the maximum distance, to take the average distance, or to use some kind of measure that minimizes the distances between observations within the cluster. Each of these methods will reveal certain types of structure within the data.
- Using the *minimum* tends to find clusters that are drawn out and "snake"-like, while using the *maximum* tends to find compact clusters.
- Using the mean is a compromise between those methods.

- One method that tends to produce clusters of more equal size is known as ***Ward's method***. It attempts to form clusters keeping the distances within the clusters as small as possible, and is often useful when the other methods find clusters with only a few observations.

5 More on Agglomeration Methods (SPSS)

The following was written for SPSS users
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***.

5.1 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.

5.2 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.

5.3 Average (between groups) linkage method

(sometimes referred to as 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.

5.4 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.

5.5 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.

6 Hierarchical Clustering: Implementation with R

- Agglomerative Hierarchical cluster analysis is provided in R through the `hclust` function.
- Notice that, by its very nature, solutions with many clusters are nested within the solutions that have fewer clusters, so observations don't "jump ship" as they do in k-means or the pam methods.
- Furthermore, we don't need to tell these procedures how many clusters we want - we get a complete set of solutions starting from the trivial case of each observation in a separate cluster all the way to the other trivial case where we say all the observations are in a single cluster.
- Traditionally, hierarchical cluster analysis has taken computational shortcuts when updating the distance matrix to reflect new clusters. In particular, when a new cluster is formed and the distance matrix is updated, all the information about the individual members of the cluster is discarded in order to make the computations faster.

6.1 The `agnes` function (cluster package)

- The **cluster** library provides the `agnes` function which uses essentially the same technique as `hclust`, but which uses fewer shortcuts when updating the distance matrix.
  For example, when the mean method of calculating the distance between observations and clusters is used, `hclust` only uses the two observations and/or clusters which were recently merged when updating the distance matrix, while agnes calculates those distances as the average of all the distances between all the observations in the two clusters.
- While the two functions will usually agree quite closely when minimum or maximum updating methods are used, there may be noticeable differences

when updating using the average distance or Ward's method.

7 Partitioning around Medoids

- The R ***cluster*** library provides a modern alternative to k-means clustering, known as pam, which is an acronym for "Partitioning around Medoids".
- The term ***medoid*** refers to an observation within a cluster for which the sum of the distances between it and all the other members of the cluster is a minimum. **pam** requires that you know the number of clusters that you want (like k-means clustering), but it does more computation than k-means in order to insure that the medoids it finds are truly representative of the observations within a given cluster.
- Recall that in the k-means method the centers of the clusters (which might or might not actually correspond to a particular observation) are only recaculated after all of the observations have had a chance to move from one cluster to another. With **pam**, the sums of the distances between objects within a cluster are constantly recalculated as observations move around, which will hopefully provide a more reliable solution.
- Furthermore, as a by-product of the clustering operation it identifies the observations that represent the medoids, and these observations (one per cluster) can be considered a representative example of the members of that cluster which may be useful in some situations. **pam** does require that the entire distance matrix is calculated to facilitate the recalculation of the medoids, and it does involve considerably more computation than k-means, but with modern computers this may not be a important consideration.
- As with k-means, there's no guarantee that the structure that's revealed with a small number of clusters will be retained when you increase the number of clusters.

8 Distance Measures

8.1 Euclidean Distance

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

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

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.

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.

8.2 Squared Euclidean Distance

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.

8.3 Standardized Distances

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.

8.4 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.

8.5 `R` - Distance Measures supported by the `dist` Function

- The `dist` function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.
- The distance measures supported by `dist` are
  - `euclidean` (but not squared euclidean directly)
  - `maximum`
  - `manhattan`
  - `canberra`
  - `binary`
  - `minkowski`.

8.6 Manhattan (City Block) Distance

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.

9 Cluster Analysis : Proximity Matrices

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.

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 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.
- 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.

10 Linkage Methods

-
-