Session:

Cluster Analysis

MODULE OBJECTIVES

At the end of this module, you will be able to:

- Explain the technique of cluster analysis
- Implement clustering techniques in R

SESSION OBJECTIVES

At the end of this session, you will be able to:

- Understand the meaning of cluster analysis, and its applications
- Describe the concepts of within-cluster and between-cluster sum of squares
- Implement similarity aggregation and agglomerative hierarchical clustering in R
- Implement R amap package and K-Means clustering
- Implement hierarchical clustering in R

"I would be better at my job if I were technical."

—Sheryl Sandberg

In Session 1 and 2 of the module, you have learned about regression analysis, which is widely used in statistical analysis. There are many other important techniques. **Clustering**, also known as segmentation, is one of them.

Clustering is one of the most widespread descriptive methods of data analysis and data mining. It is used when there is a large volume of data, and the aim is to find homogeneous subsets, which can be processed and analyzed in different ways. This is required in a wide range of contexts, especially in social sciences, medicine and marketing, where human factors mean that the data is large and difficult to interpret.

Exit of Previous Session

- Understand nonlinear regression
- Explain logistics regression and its applications
- Use R for performing nonlinear regression

Entry of this Session

- Explore the meaning of cluster analysis
- Differentiate regression and cluster analysis
- Discuss some real life examples of clustering
- Perform cluster analysis in R

Let us take an example to understand the concept of clustering.

A food product manufacturing company wants to categorize its customers on the basis of number of purchased items and cost of those items in a month. The company has collected data from its various stores. A plot for customers on the basis of number of products they purchase and cost of those products is shown in following **Figure 1**:

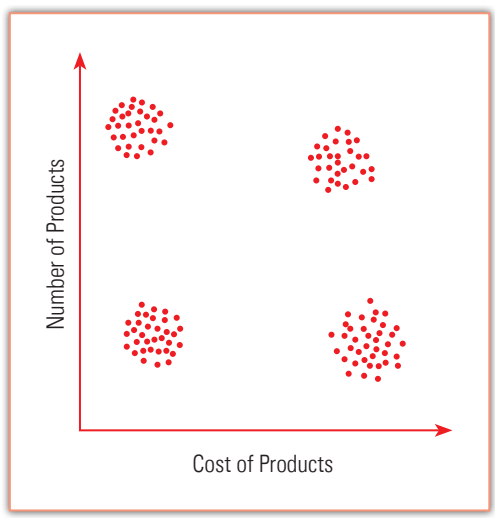


FIGURE 1
Plot for Customers

In the preceding figure, you can see that customers can be categorized into four groups. This grouping of customers is known as clustering.

Figure 2 shows four clusters of customers for **Figure 1**:

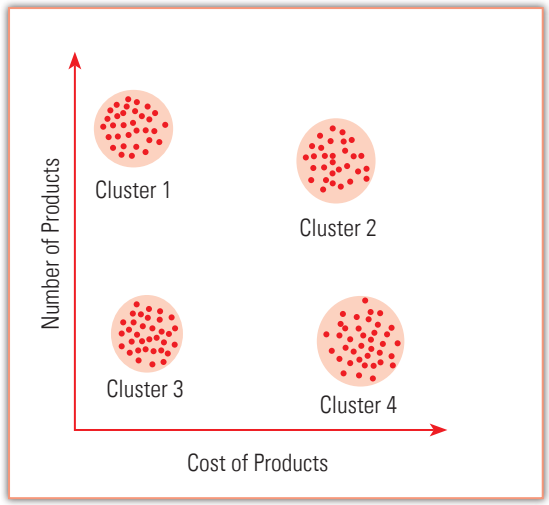


FIGURE 2
Four Clusters for Customers

From the preceding figure, it is clear that the company can categorize their customers into four groups and make a separate strategy for each group to enhance sale of their products.

In clustering, unlike regression, there is no particular dependent variable, and it is harder to compare two forms of clustering objectively.

In this session, you will learn about cluster analysis, the complexity of clustering, distance measures for clustering, and the concepts of within-cluster and between-cluster sum of squares. In addition, the session discusses agglomerative hierarchical clustering, clustering by similarity aggregation, the principle of similarity aggregation, and implementing clustering by similarity aggregation in R. In the end, the session describes the implementations of `k-means`, `expectation-maximization`, and `hierarchical` clustering in R.

Introduction to Clustering

PRE-READ CONNECT

Refer to the Pre-Read of this session to read more about segmentation and the ways to perform clustering.

Clustering is the statistical operation of **grouping objects** (individuals or variables) into a limited number of groups known as **clusters** (or segments).Clusters have the following properties:

- They are not defined in advance by the analyst, but are discovered during the operation, unlike the variables used in regression.
- They are combinations of objects having similar characteristics, which are separated from objects having different characteristics (resulting in internal homogeneity and external heterogeneity).

The essence of clustering is the distribution of objects into groups; however, this distribution is **not carried out on the basis of a predefined criterion**, and is not intended to combine the objects having the same value for such a criterion. Even the number of clusters is not always fixed in advance. This is because there is no dependent variable. Clustering is descriptive, not predictive.

DEFINITION

Clustering is a data segmentation technique that divides huge datasets into different groups on the basis of their similarity. The resulting groups are known as clusters.

Applications of Clustering

Clustering is widely used in several fields for statistical analysis. Areas where clustering techniques are commonly used are:

- **Marketing:** In the field of marketing, clustering is particularly useful in **finding customer profiles** that make up a customer base. After detecting the clusters, which **"sum up"** its customer base, a business can develop a specific strategy for each cluster base. The clusters can also be used to **keep track of customers** over months and detect the number of customers who move from one cluster to another every month. If required, businesses can also follow certain customers by setting up a cluster-driven customer panel, to ensure that all clusters are well represented.
- **Retail:** In the retail industry, apart from marketing, clustering is also used to divide all the stores of a particular company into **groups of establishments**. These establishments are homogeneous in terms of the type of customer, turnover, turnover per department (according to the type of product), size of store, and such.
- **Medical Science:** In the medical field, clustering can be used to discover **groups of patients** suitable for particular treatment protocols, each group comprising all the patients who react in the same way. The grouping of patients is done on the basis of their age, type of disease, lifestyle, and income. Clustering technique is also used in classification of protein sequence, ct-scans, finding the suitable type of drugs for a particular group of patients, and predicting the possibility of several common diseases, such as diabetes on the basis of people's lifestyle.
- **Sociology:** In sociology, clustering is used to divide a population into groups of individuals who are homogeneous in terms of social demographics, lifestyle, income opinions, expectations, and such. Generally, clustering is also useful in performing other data mining operations. To begin with, most predictive algorithms are not good at handling excessively large number of variables because of the correlations between the variables. This can affect their predictive power; however, it is difficult to describe a heterogeneous population correctly with a small number of variables. The groups formed by clustering are useful because they are **homogeneous** and can be described by a **small number of variables** specific to each group. This categorization can be used for various purposes such as polls, identifying criminals, and marketing of products.

EXAM CHECK

In your certification examination, you will be required to demonstrate your knowledge of the basics of clustering and its applications.

ADDITIONAL KNOW–HOW

In different fields, clustering is known with different names:

- **Marketing:** In marketing, clustering is often referred to as **"segmentation"** or **"typological" analyses.**
- **Medicine:** In the field of medicine, the term, **nosology**, for clustering.
- **Biology:** In the field of biology, clustering is known as **numerical taxonomy**.

Complexity of Clustering

As a general rule, the following expression is used to define the correct criteria for clustering and using efficient algorithms:

```
Bn(number of partitions for n objects)>exp(n)
```

The following formula can be used to establish the relationship between the number of partitions and number of objects:

$$B_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$$

In the preceding formula, Bn is the number of partitions, e is the exponential and n is the number of objects.

If n=4 and Bn=15, we can partition four objects a, b, c, and d into 15 different sets:

- One partition with one cluster (abcd)
- Seven partitions with two clusters (ab, cd), (ac, bd), (ad, bc), (a, bcd), (b, acd), (c, bad), and (d, abc)
- Six partitions with three clusters (a, b, cd), (a, c, bd), (ad, bc), (a, d, bc), (b, c, ad), (b, d, ac), and (c, d, ab)
- One partition with four clusters (a, b, c, d)

The complexity of a cluster depends on the number of possible combinations of objects. The preceding partitioning is performed on a small number of objects; however, if the number of objects is large, testing all possible combinations would be impossible.

Distance Measures

In clustering, objects are joined or separated on the basis of distances between them. These distances are referred to as **dissimilarity** (when objects are far from each other) or **similarity** (when objects are close to each other), and they can be calculated for single or multiple dimensions; for example, if you cluster fast foods, you will take multiple dimensions, such as, calories provided by these, price, and their rating by customers.

There are different types of methods used for calculating differences. Some common methods of calculating distances are listed as follows:

- **Euclidean Distance:** This is the most commonly used type of distance used to join or separate objects in clustering. This is the geometric measure of distance between objects in a multi-dimensional space. Let's suppose, there are two points, p $(p_1, p_2, \ldots, p_n)$ and q $(q_1, q_2, \ldots, q_n)$, in a n-dimensional space. The Euclidean distance between p and q is calculated by using the following formula:

  $$D(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 . + \ldots\ldots + (q_n - p_n)^2}$$

  Generally, the Euclidian distance is calculated on raw data without modifying it. The common advantage of this type of distance is that the addition of new objects does not affect the calculated distance.

- **Squared Euclidean Distance:** This distance is obtained by squaring the Euclidian distance. It is used to put more weight on the objects that are allocated at greater distances.

- **City–Block (Manhattan) Distance:** The City–Block distance is obtained by calculating the average difference between two points in all dimensions. In most of the cases, it is same as the Euclidian distance. Let's suppose, there are two points, p $(p_1, p_2, \ldots, p_n)$ and q $(q_1, q_2, \ldots, q_n)$, in a n-dimensional space; the City–Block distance between p and q, is calculated by using the following formula:

  $$D(p,q) = |p_1 - q_1| + |p_2 - q_2| + \ldots\ldots |p_n - q_n|$$

The Manhattan distance is sometimes used for clustering to reduce the effect of extreme individuals whose coordinates are not squared.

Within-Cluster and Between-Cluster Sum of Squares

The total sum of squares (or inertia) I of a cluster is the weighted mean of the squares of the distances of each point from the center of gravity of that cluster. The sum of squares of a cluster is calculated in the same way with respect to its center of gravity and can be written as follows:

$$\sum_{i \in I_j} p_i \left(x_i - \bar{x}_j \right)^2$$

The total sum of squares in a cluster is defined by adding between cluster sum of squares and within- cluster sum of squares. The between sum of squares and within sum of squares are defined as follows:

- **Between-Cluster Sum of Squares:** It is calculated by finding the square of the difference from the center of gravity for each cluster and then adding them, as shown in **Figure 3**:

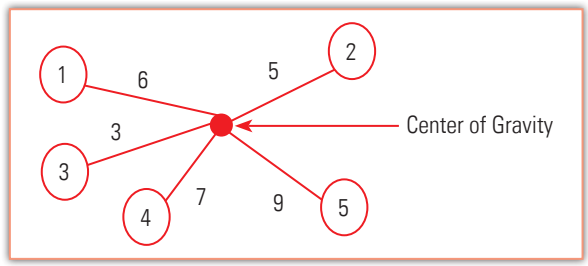


FIGURE 3
Between-Cluster Sum of Squares

In the preceding figure, you can see five clusters—1,2,3,4, and 5. The difference between these clusters from the center of gravity is 6, 5, 3, 7, and 9, respectively. You can calculate between-cluster sum of squares as follows:

```
6² + 5² + 3² + 7² + 9²      = 200
```

In this case the between-cluster sum of squares is 200.

- **Within-Cluster Sum of Squares:** It is calculated by finding the square of the difference from the center of gravity for each point and then adding them within a single cluster, as shown in **Figure 4**:

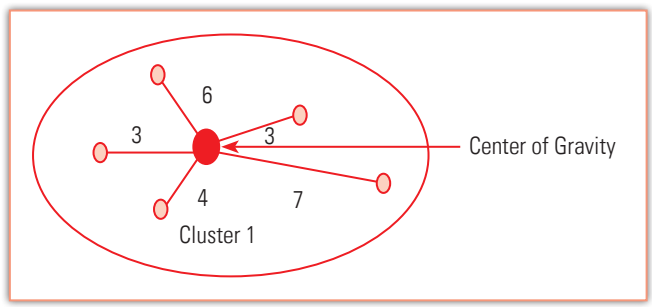


FIGURE 4
Within-Cluster Sum of Squares

In the preceding figure, you can see that a cluster that has five points. The difference of these points from the center of gravity of the cluster is 6, 3, 7, 4, and 3 respectively. You can calculate within-cluster sum of squares as follows:

```
6² + 3² + 7² + 4² + 3² =119
```

In this case, the within-cluster sum of squares is 119.

In general, the total sum of squares can be defined by the following formula:

Total Sum of Squares = Between-Cluster Sum of Squares + Within-Cluster Sum of Squares

Figure 5 shows the diagrammatic representation of the preceding formula:

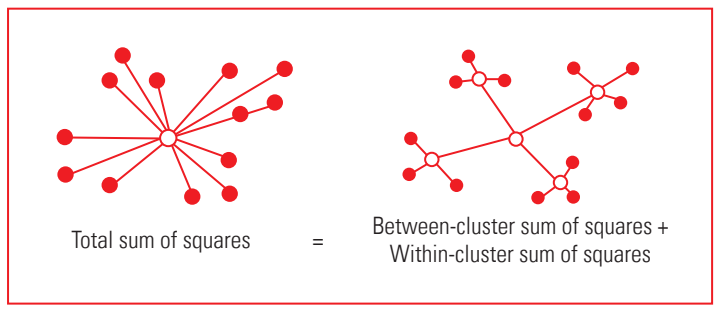


FIGURE 5 Diagrammatic Representation of the Total Sum of Squares Formula

If the objects are segmented into k clusters, with sums of squares $I_1, \ldots, I_k$, the within-cluster sum of squares is given by the following formula:

$$I_A = \sum_{j=1}^{k} I_j$$

A cluster becomes more homogeneous as its sum of squares decreases, and the clustering of the population becomes better as I_A diminishes.

The between-cluster sum of squares, I_R, of the clustering is defined as the mean of the squares of the distances of the centers of gravity of each cluster from the global center of gravity. This can be written as follows:

$$I_R = \sum_{j \in clusters} \left(\sum pi\right)\left(\bar{x}_j - \bar{x}\right)^2$$

As I_R increases, the separation between the clusters also increases, indicating satisfactory clustering.

Thus, there are two criteria for correct clustering:

- I_R should be large
- I_A should be small

Thus, the total sum of squares (or inertia) I can be given by using the following formula:

$$I = I_A + I_R$$

The preceding formula is also known as **Huygens' formula**. It shows that the total sum of squares depends on the global population only, and that the two preceding criteria (minimization of the within-cluster sum of squares and maximization of the between-cluster sum of squares) are, therefore, equivalent.

QUICK TIP

R^2 (RSQ) is the proportion of the sum of squares explained by the clusters (between-cluster sum of squares/total sum of squares). The nearer it is to 1, the better the clustering will be, but we should not aim to maximize it at all costs, because this would result in the maximum number of clusters: There would be one cluster per individual. So we need an R^2 that is close to 1 but without too many clusters. A good rule is that, if the last significant rise in R^2 occurs when we move from k to k + 1 clusters, the partition into k+1 clusters is correct.

ADDITIONAL KNOW–HOW

The Huygens' Formula has been given by a Dutch mathematician and philosopher, **Christiaan Huygens**. He was born in a Dutch family on 14th April, 1629 at The Hague, the capital city of South Holland. In his research work, Huygens mentioned that the Descartes's laws for collision elastic bodies were incorrect, and later he formulated correct laws. Huygens also extended Newton's second law of motion that is known as the quadratic form of law of motion.

Properties of Efficient Clustering

A sound clustering procedure:

- Detects the structures present in the data
- Enables easy determination of optimal number of clusters
- Yields clearly differentiated clusters
- Yields clusters that remain stable with minor changes in data
- Processes large data volumes efficiently
- Handles all types of variables (quantitative and qualitative), if required

ADDITIONAL KNOW–HOW

To distinguish true clusters in data, we often have to first interpret the data before transforming, adding, or excluding variables and then restart the clustering. Excluding a variable does not necessarily mean deleting it from the analysis base. Instead, we cease to take it into account in the clustering operation, while retaining it as an inactive variable to observe the distribution of its categories in the various clusters. It is no longer an "active" variable, but becomes an "illustrative" variable (also called a "supplementary" variable.

Knowledge Check–1

1. While performing statistical analysis in the field of sociology, you have to deal with a large amount of heterogeneous data. Which of the following is the correct reason for preferring cluster analysis over other statistical analysis techniques?

a. Cluster analysis does not include calculations.

b. Cluster analysis is simpler than other statistical techniques.

c. Cluster analysis can easily express heterogeneous data with a number of variables.

d. Cluster analysis is a technique that has been specially developed for the sociology field.

2. While performing cluster analysis, which of the following conditions you will have to achieve?

a. I_R should be large and I_A should be small.

b. I_A should be large and I_R should be small.

c. I_R and I_A should be small.

d. I_R and I_A should be large.

EXAM CHECK

In your certification examination, you will be required to demonstrate your knowledge of within-cluster and between-cluster sum of squares, and the properties of efficient clustering.

Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) is a clustering technique that produces a sequence of **nested partitions of increasing heterogeneity**, between partitions into **n** clusters. In this technique, each object is isolated and partitioned into one cluster, which includes all the objects. AHC can be used if there is a distance between partitions, which can be in either an individual space or a variable space. You must define the distance of two objects or the distance of two clusters, which will help you to efficiently categorize data.

The general form of the algorithm for this is as follows:

- **Step 1:** The observations are the initial clusters.
- **Step 2:** The distances between clusters are calculated.
- **Step 3:** The two clusters closest together are merged and replaced with a single cluster.
- **Step 4:** You start again at step 2 until there is only one cluster, which contains all the observations.

The agglomerative hierarchical clustering is represented as a tree diagram shown in **Figure 6**:

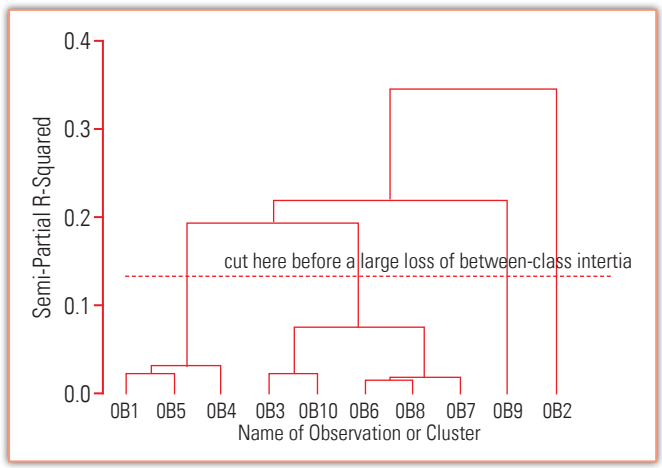


FIGURE 6
Dendrogram - Tree Diagram of Agglomerative Hierarchical Clustering

The tree generated by AHC is also known as a **dendrogram**. This tree can be cut at a greater or lesser height to obtain a smaller or a larger number of clusters. This number can be chosen by the statistician to optimize certain statistical quality criteria.

The main criterion is the loss of between-cluster sum of squares, represented in **Figure 6** by the height of the two connected branches. Because this loss must be as small as possible, the tree diagram is cut at a level where the height of the branches is large.

The Big Picture

Hierarchical clustering is used for identification of patterns in digital images, predicting the stock exchange, and in text mining and ontology. In 2005, the technique of hierarchical clustering was used for research on protein sequence classification.

Main Distances

The AHC algorithm works by searching for the closest clusters at each step and merging them. The critical point of the algorithm is the **definition of the distance between two clusters** A and B. When each of the two clusters is reduced to one element, the definition of their distance is natural, but as soon as a cluster has more than one element, the concept of the distance between two clusters is less obvious. The distance can be defined in many ways, but the most usual definitions are as follows:

- **Maximum Distance:** The maximum distance between two observations $\mathbf{a} \in A$ and $\mathbf{b} \in B$, where **a** and **b** are two elements belonging to two clusters, A and B, tends to generate clusters of equal diameter. By definition, this approach is highly sensitive to outliers, and is therefore little used. The corresponding form of AHC is called the farthest-neighbor technique, diameter criterion, or complete linkage AHC.

- **Minimum Distance:** The minimum distance between two observations **a** ∈ A and **b** ∈ B defines what is known as the nearest-neighbor technique or a single linkage AHC method. Its weak point is that it is sensitive to the chain effect (or chaining): if two widely separated clusters are linked by a chain of individual points that are close to each other, they may be grouped together.

Let's look at the example of the chain effect for two sets detected by using the average linkage and complete linkage methods, as shown in **Figure 7**:

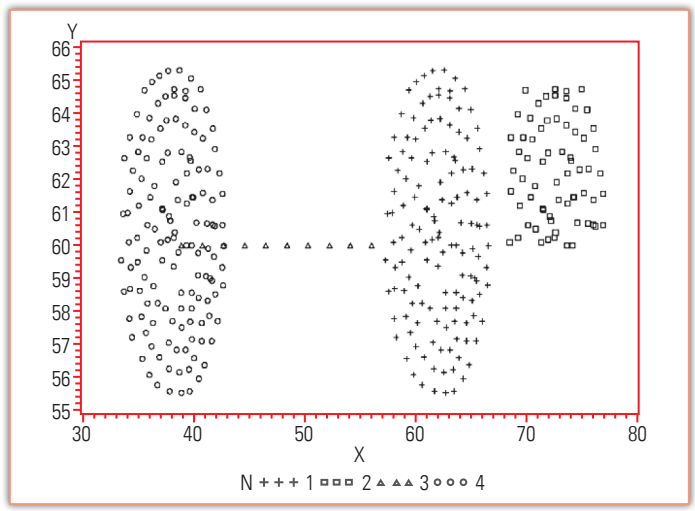


FIGURE 7
The Chain Effect

However, if the single linkage AHC method is used for the same sets to find two clusters, it isolates cluster 2 and groups the rest together, as shown in **Figure 8**:

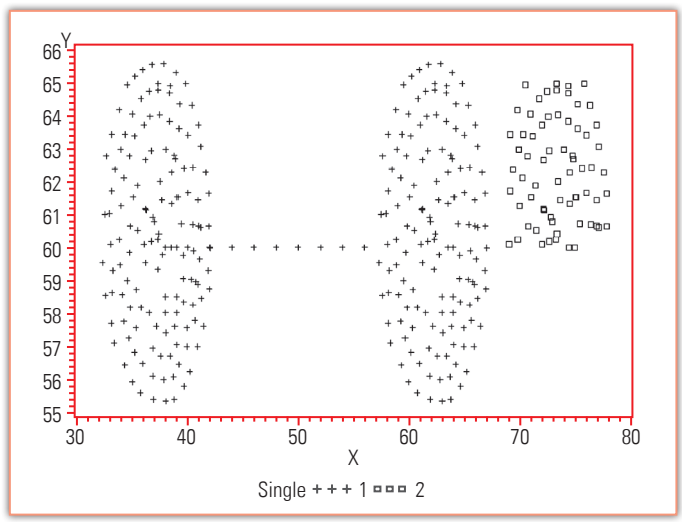


FIGURE 8
Sensitivity of Single Linkage to the Chain Effect

This is because the closest points of clusters 1 and 2 are separated by a distance greater than the shortest distance between:

- Two points of cluster 1
- Two points of clusters 1 and 3
- Two points of cluster 3
- Two points of clusters 3 and 4

Since the distance between two clusters is the shortest distance between two points in the two clusters in the single linkage method, cluster 2 is farthest from the other clusters.

This is because, although the two large clusters 1 and 4 are at a distance from each other, they are linked by cluster 3 in which each point follows the previous one at a distance shorter than the shortest distance between the two points of clusters 1 and 2: this is the chain effect. It is the main drawback of the single linkage method, but only occurs in special circumstances.

The mean distance between two observations **a** ∈ A and **b** ∈ B defines the average linkage AHC, which is intermediate between the maximum distance and minimum distance methods, and is less sensitive to cluttered data. It tends to produce clusters with the same variance.

The distance between the centers of gravity of A and B, which is used in the centroid method of AHC (the center of gravity is sometimes referred to as the centroid), is more robust to outliers but less precise. This is the simplest in terms of calculation.

The main disadvantage of AHC is its algorithmic complexity, which is non linear: To move from k + 1 clusters to k clusters, you must calculate [(k + 1)k/2] distances and combine the two closest clusters. If n is the number of individuals to be clustered, the complexity of the basic algorithm is of the order of n^3, and it will soon exceed the capacity of even a powerful computer.

TECHNICAL STUFF

Density Estimation Methods

Density estimation methods are often among the most suitable for detecting the structure of complex clusters. You can picture the data space as a landscape of peaks and valleys, where the mountains are the clusters and bottoms of the valleys are their boundaries, as shown in the following **Figure 9**:

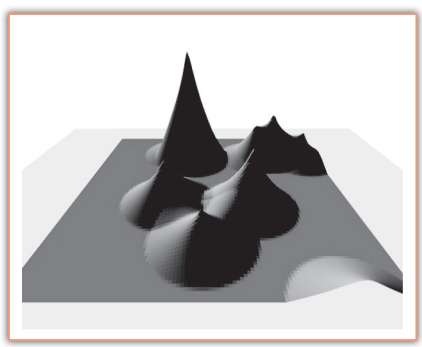


FIGURE 9
Data Space as Landscape and Valleys

As you can see, the mountains are regions of **high density**. **Density** is generally defined as the number of objects in a certain neighborhood.

In the clustering process, density is estimated by using one of the following methods:

- **The k-Nearest-Neighbors Method:** The density at a point `x` is the number `k` of observations in a sphere centered on `x`, divided by the volume of the sphere.
- **The Uniform Kernel Method:** The radius of the sphere is fixed, not the number of neighbors.
- **The Wong Hybrid Method:** This method uses the `k-means` algorithm in a preliminary analysis.

The three methods are effective for detecting all types of clusters, irregularly shaped ones, which are of unequal sizes and have variances. This arises from the principle of these methods, which does not specify any shape for the cluster in advance: A cluster grows in any direction the density is great enough. Density estimation methods operate by specifying not the number of clusters, but a smoothing **parameter**, which, depending on circumstances, can be:

- The number of clusters of the preliminary `k-means` (Wong method)
- The number `k` of neighbors of each point `x`
- The radius `r` of the sphere surrounding `x`

A distance `dP` between two clusters is defined as inversely proportional to the density in the middle of these two clusters. It is assumed that `dP` = ∞ if the two clusters are not adjacent.

TECHNICAL STUFF

EXAM CHECK

In your certification examination, you will be required to demonstrate your knowledge of agglomerative hierarchal clustering.

The problem with these methods is the difficulty in finding a good value for the smoothing parameter. Their constraints are that, on the one hand, it is better to standardize the continuous variables and exclude the outliers, and, on the other hand, they require sufficiently high frequencies.

The following example will help understand how these methods are used.

To create a cluster of data with five clusters, enter the following command:

```
> clusters_five<-kmeans(test _final,5)
```

We can create a cluster of data containing 10 clusters by using the following command:

```
> clusters_ten<-kmeans(test _final,10)
```

Clustering by Similarity Aggregation

Clustering by similarity aggregation is used to compare all the individuals in pairs at each step, thus building a global clustering, instead of local clustering as in hierarchical clustering methods. This clustering technique determines the optimum number of clusters automatically, instead of fixing them in advance.

This clustering technique is based on the work of **Pierre Michaud** and **Jean-Francois Marcotorchino**. It is is also called **relational clustering** because of the relational analysis used by the authors. Sometimes it is known as the **voting method** or the **Condorcet method**.

Principle of Similarity Aggregation

Clustering by similarity aggregation is based on the representation of data in the form of an equivalence relation.

Clustering is actually an equivalence relation R, where `iRj` if `i` and `j` are in the same cluster.

As for any binary relation defined for a set of n objects, you can associate R with an n*n matrix, which is defined by $m_{ij} = 1$ if `iRj`, and $m_{ij} = 0$ otherwise.

The three properties of an equivalence relation, namely **reflexivity**, **symmetry**, and **transitivity,** are shown by the following relations:

- $m_{ii} = 1$ (reflexivity)
- $m_{ij} = m_{ji}$ (symmetry)
- $m_{ij} + m_{jk} - m_{ik} <= 1$ (transitivity)

The clustering procedure is, therefore, a matter of finding a matrix $M = (m_{ij})$, which meets the preceding conditions.

In relational analysis, all the variables of the individuals of the population to be clustered must be qualitative; if they are not, then they must be categorized into equal intervals; for example, each `p` variable has its own natural clustering: Each cluster consists of the individuals having the same category for the variable.

The aim of relational analysis is to find a clustering that is a good compromise between the initial p natural clustering.

To do this, you can assume that m_{ij} is the number of times that the individuals `i` and `j` have been placed in the same cluster (i.e., the number of variables for which `i` and `j` have the same category), and that $M/= (m/_{ij}) = 2(m_{ij}) - p$.

Then $m/_{ij} > 0$ if i and j are in the same cluster (they "coincide") for a majority of variables, $m/_{ij} < 0$ if i and j are in different clusters for a majority of variables, and $m/_{ij} = 0$ if the number of variables for which i and j are grouped together is the same as the number of variables for which i and j are separated. It is natural to place i and j in the same final cluster if $m/_{ij}$ is positive, and to separate them if $m/_{ij}$ is negative.

But this criterion is not sufficient, because of the non-transitivity of the **majority rule** (Condorcet's paradox): there may be a majority for joining i and j, j and k, but not for joining i and k. You must, therefore, add equivalence relation constraints of the kind stated above (reflexivity, symmetry, and transitivity) to find a clustering that is closest to the majority of the p initial clustering. This brings us to a linear programming problem, which can be resolved correctly in R, as you have done in linear regression.

Implementing Clustering by Similarity Aggregation

For a better picture of the working of clustering on the basis of similarity aggregation, you need to describe the stages of clustering by using an intuitive approach rather than an absolutely rigorous approach.

For each pair of individuals (A,B), let m(A,B) be the number of variables having the same value for A, and B, and d(A,B) be the number of variables having different values for A and B, given that, for continuous variables, there are the following two conditions:

- Consider that the variables have the same value if they are in the same decile.
- Their contribution to c(A,B) is defined by using the following equation:

  $1-2(|v(A)-v(B)|)/(v_{max}-v_{min})$

  where, v_{min} and v_{max} are the outlying values of variable V.

The Condorcet criterion for two individuals A and B is defined by using the following equation:

$$c(A, B) = m(A, B)-d(A, B)$$

You can define the Condorcet criterion of an individual A and a cluster S by using the following equation:

$$c(A,S) = \sum_i c(A,B_i)$$

the summation is being over all the Bi ∈ S.

Given the preceding conditions, you start constructing the clusters by placing each individual A in the cluster S for which c(A,S) is maximum and at least 0. Sometimes, you can replace the value 0 with a larger value, to strengthen the homogeneity of the clusters. You can also have an effect on this homogeneity by introducing a factor $\sigma > 0$ into the definition of the Condorcet criterion, which becomes:

$$c(A,B) = m(A,B) - \sigma d(A,B)$$

A large value of σ will be a high cluster homogeneity factor. If c(A,S)<0 for every existing S, then A forms the first element of a new cluster.

Therefore, take the first individual A, which is compared with all the other individuals, and group it with another individual B_A, if necessary. Then take the second individual B, which is compared with the other individuals, as well as with the cluster {A, BA}, if it exists, and so on. This step is the first iteration of the clustering.

You can perform a second iteration by taking each individual again and reassigning it, if necessary, to another cluster taken from those defined in the first iteration. This way, we perform a number of iterations until one of the following two conditions is achieved:

- The specified maximum number of iterations is reached

EXAM CHECK

In your certification examination, you will be required to demonstrate your ability to implement clustering by similarity aggregation.

- The global Condorcet criterion ceases to improve sufficiently (by more than 1%, for example, a value that can be set in advance) from one iteration to the next. This global Condorcet criterion is described by the following formula:

  $\sum_A c(A, SA)$

  where, the summation is performed on all the individuals in A and the clusters S_A to which they have been assigned. In practice, two iterations (or three if absolutely necessary) will be enough to provide good results.

Use of the R amap Package

The statistical tool R provides the `amap` package for clustering by similarity aggregation. Let's use this package to perform clustering analysis.

Load the `amap` package by using the following command:

```
> library(amap)
```

If the variables of the data frame (data table) to be processed are not factors (qualitative variables), they must be transformed in advance by using the following command:

```
> for (i in 1:17) credit[,i] <- factor(credit[,i])
```

In this example, 17 variables are assumed. The variables can be numeric at the outset, but the number of categories must be small.

Now, calculate the distance between two samples, by using the `diss()` function in the package.

However, this function only processes whole numbers. You must, therefore, transform the variables in advance, by using the following command:

```
> creditn <-matrix(c(lapply(credits,as.integer),recursive=T),ncol=17)
```

You can use the following command to perform similarity aggregation clustering:

```
> pop(matrix)
```

The output of the preceding command is shown in **Table 1**:

TABLE 1 Output of the `pop()` Command:

```
Upper bound (half cost)          : 189
Final partition (half cost)      : 129
Number of classes                : 6
Forward move count               : 879424708
Backward move count              : 879424708
Constraints evaluations count    : 1758849416
Number of local optima           : 4
        Individual       class
1       1                1
2       2                2
3       3                3
4       4                3
5       5                4
6       6                3
7       7                1
8       8                2
```

9	9	3
10	10	5
11	11	2
12	12	2
13	13	2
14	14	5
15	15	2
16	16	2
17	17	1
18	18	6
19	19	2
20	20	1

You can see that the `pop()` command has detected six clusters from 1 to 6 among 20 individuals.

Knowledge Check–2

1. In clustering, any binary relation defined for a set of `n` objects can be associated as a relation R with an `n*n` matrix, which is defined by $m_{ij} = 1$ if `iRj`, and $m_{ij} = 0$ otherwise. The three properties of an equivalence relation are reflexivity, symmetry, and transitivity. Which of the following relations represents the symmetric relation?
   a. $m_{ij} = 1$
   b. $m_{ii} = 1$
   c. $m_{ij} = m_{ji}$
   d. $m_{ij} + m_{jk} - m_{ik} <= 1$
2. Density estimation methods are used to identify the structure of complex clusters. Suppose you want to estimate the density at the point `x` in a sphere centered on `x`, on the basis of the number of observations, `k`, and the volume of the sphere. In this situation, which of the following density estimation methods will you use?
   a. The k-nearest-neighbors method
   b. The uniform kernel method
   c. The Wong hybrid method
   d. The sum-of-squares method

K-Means Clustering

LAB CONNECT

In the lab session, you will practise implementing K-Means clustering.

The `k-means` is most widely used method for customer segmentation of numerical data.

This technique partitions n units into `k` ≤ `n` distinct clusters, `S = {S1, S2, . . . , Sk }`, so as to minimize the within-cluster sum of squares, by using the following formula:

$$\arg\min_{S} \sum_{j=1}^{k} \sum_{x_i \in S_j} ||x_i - \overline{m}_j||^2$$

TECHNICAL STUFF

You can use the k-means function in the R package stats. This algorithm is known to be fast and reliable. But, there is no guarantee that it converges to the global optimum. Its final result may depend on how the algorithm has been started.

Let's suppose there are observations on n units, (x1, x2, . . . , xn), with the observation on unit i representing a p-dimensional vector of attributes.

The cluster mean $\overline{m_j}$ is the mean vector of the p attributes averaged over all units in cluster S_j. The norm $||x - m||^2 = \sum_{r=1}^{p}(x_r - m_r)^2$ sums the squared differences over the p attributes. The norm assumes equal scales of the p attributes, and it does not incorporate correlations among the attributes.

The number of clusters k must be given.

Clustering is purely descriptive. Clustering groups the items according to their similarity, and it does so in an unguided (unsupervised) manner.

Two commonly used initialization approaches are followed for k-means clustering:

- Randomly choosing k units from the dataset and using these as the initial cluster means
- Randomly assigning one of the k clusters to each unit and then proceeding to the update step, and computing the initial means as the centroids of the clusters' randomly assigned units. In general, this random partition method is thought to be preferable.

EXAM CHECK

In your certification examination, you will be required to demonstrate your ability to use R amap package and implement K-Means clustering.

Applications of K-Means Clustering

The k-means clustering is the most commonly used clustering technique. Some of the common applications of k-means clustering are as follows:

- Predicting the price of products for a specific period or for specific seasons or occasions such as summers, New Year, or any particular festival
- Extracting information from electric price by time series models

An Example of Clustering in R: European Protein Consumption

Let's understand the concept of clustering as performed in R, with an example.

Consider 25 European countries (n = 25 units) and their protein intakes (in percent) from nine major food sources (p = 9). The data is listed in **Table 2**:

TABLE 2
Data for Protein Consumption

Country	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fry and Veg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6

Country	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fry and Veg
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
United Kingdom	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Table 2 may be used to learn whether the 25 countries can be separated into a smaller number of clusters. It may well be that Mediterranean countries get their protein intake from certain food categories, that are different from food staples favored by North European and German-speaking countries.

Let's start with clustering on the first two features, protein intake from red and white meat, to cluster the 25 countries into three groups.

You can create a program in R to cluster the 25 countries into three groups.

The code for creating three clusters for 25 countries is shown in **Listing 1**:

Listing 1: Code for Creating a Cluster

```
1  food <- read.csv("C:/DataMining/Data/protein.csv")
2  set.seed(1)
3  grpMeat <- kmeans(food[,c("WhiteMeat","RedMeat")], centers=3, +
   nstart=10)
4  o=order(grpMeat$cluster)
5  data.frame(food$Country[o],grpMeat$cluster[o])
```

Explanation of Listing 1	
1	Reads the data available in the `protein.csv` file
2	Fixes random starting for clusters from 1

3	Specifies the clustering on the data of red and white meat with the three clusters
4	Sorts the cluster on the basis of cluster numbering
5	Lists the clusters along with country name

The output of the **Listing 1** is shown in **Table 3**:

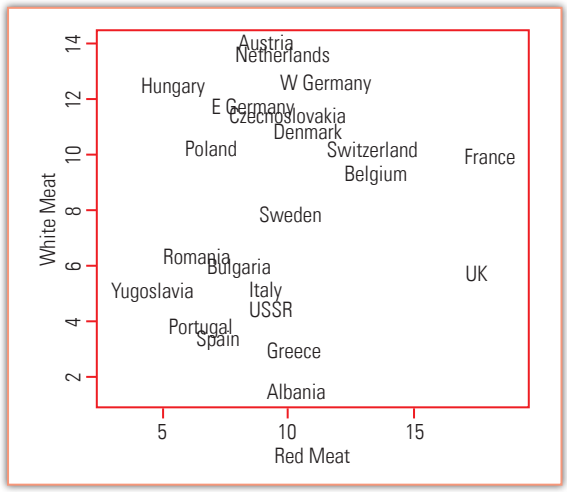
TABLE 3
Output of Listing 1

	food.Country.o.	grpMeat.cluster.o.
1	Albania	1
2	Bulgaria	1
3	Finland	1
4	Greece	1
5	Italy	1
6	Norway	1
7	Portugal	1
8	Romania	1
9	Spain	1
10	Sweden	1
11	USSR	1
12	Yugoslavia	1
13	Belgium	2
14	France	2
15	Ireland	2
16	Switzerland	2
17	UK	2
18	Austria	3
19	Czechoslovakia	3
20	Denmark	3
21	E Germany	3
22	Hungary	3
23	Netherlands	3
24	Poland	3
25	W Germany	3

Now you can use the following command to plot the scatter plot:

```
plot(food$Red, food$White, type="n", xlim=c(3,19), xlab="Red Meat",+
ylab="White Meat")
text(x=food$Red,       y=food$White,       labels=food$Country,       +
col=grpMeat$cluster+1)
```

The scatter plot generated by the preceding command is shown in **Figure 10**:



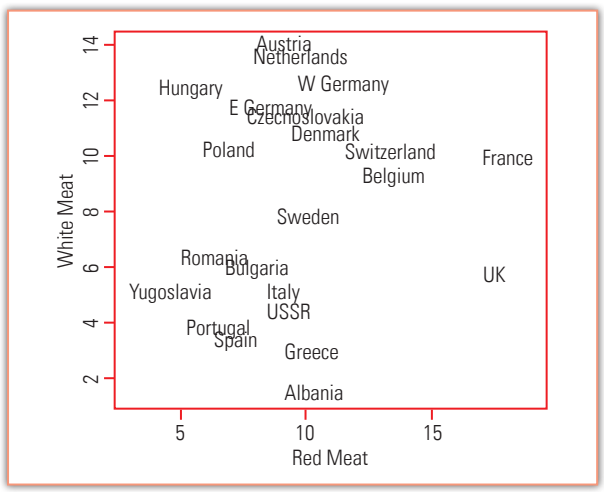
QUICK TIP

Refer to the Pre-Read of this session to learn about the use of the `k-means()` command in R for clustering analysis.

FIGURE 10
Three Cluster Assignments on Red and White Meat

In the scatter plot shown in **Figure 10**, you can see that the countries that are geographically close tend to be clustered into the same group.

In the same analysis, if you change the number of clusters from 3 to 7, you will get the scatter plot shown in **Figure 11**:



QUICK TIP

In R, you can easily generate different numbers of clusters for the same data by assigning different values to the `centers` attribute in the `k-means()` command.

FIGURE 11
Seven Cluster Assignments on Red and White Meat

Example of Clustering in R: Monthly US Unemployment Rates

Following is another example, which analyzes monthly seasonally adjusted unemployment rates covering the period **January 1976 through August 2010** for the **50 US states** (`n = 50`).

In this example, you will use summary statistics for each state, such as the average and the standard deviation of the unemployment rates, and then use these two calculated features of the monthly unemployment rates as the attributes for clustering.

The data file **unemp.csv** includes the average and the standard deviation for each state. The results for three clusters are indicated on the scatter plot of the standard deviations against the means.

Perform the following steps to perform k-means clustering on the US employment rates:

1. Read data from the unemp.csv file by using the following command:

```
unemp <- read.csv("C:/DataMining/Data/unemp.csv")
```

2. Set the seed for the cluster by using the following command:

```
set.seed(1)
```

3. Perform k-means clustering by using the following command:

```
grpunemp <- kmeans(unemp[,c("mean","stddev")], centers=3, +
nstart=10)
```

4. Sort the clusters by using the following command:

```
o=order(grpunemp$cluster)
```

5. List the clusters by using the following command:

```
data.frame(unemp$state[o],grpunemp$cluster[o])
```

6. Plot the clusters on a graph by using the following command:

```
plot(unemp$mean,unemp$stddev,type="n",xlab="mean",+ylab="stddev")
text(x=unemp$mean,y=unemp$stddev,labels=unemp$state,+
col=grpunemp$cluster+1)
```

The scatter plot generated by the preceding command is shown in **Figure 12**:

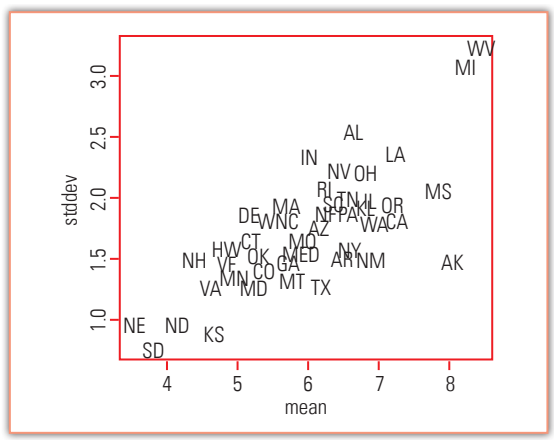


FIGURE 12
Three Cluster Assignments on US Unemployment Rates

In general, a state's standard deviation in unemployment increases with its level. In **Figure 12**, you can see groups of states with low unemployment and low variability, and states with high unemployment and high variability. Note that this approach to clustering does not incorporate differences or similarities in the state-specific time-patterns of the unemployment rates.

Knowledge Check—3

1. Which of the following statements is true about k-means clustering analysis?
   a. It minimizes the within-cluster sum of squares.
   b. It maximizes the within-cluster sum of squares.
   c. It minimizes the between-cluster sum of squares.
   d. It maximizes the between-cluster sum of squares.

2. Consider the following R command:

```
kmeans(food[,c("Var1","Var2")], centers=5, + nstart=10)
```

How many clusters will be generated by using the above command?

a. 10
b. 2
c. 5
d. 7

Implementing Hierarchical Clustering in R

You have learned about hierarchical clustering, which is an approach of clustering `n` units (or objects), each described by `p` features, into a smaller number of groups.

Hierarchical clustering creates a hierarchy of clusters that can be represented in a tree-like diagram, called a **dendrogram**.

Agglomerative hierarchical clustering uses a bottom–up approach, where each unit starts in its own cluster and pairs of clusters are merged as you move up the hierarchy. For agglomerative clustering, you can use the `agnes()` function in the R package `cluster`. Alternatively, you can use the `hclust()` function in the `stats` package.

Example 1: European Protein Consumption Revisited

Let's reconsider the examples of European protein consumption and US monthly unemployment rates in the context of hierarchical clustering.

Perform the following steps for agglomerative hierarchical clustering in R on protein consumption data:

1. Add the library named cluster by using the following command:

```
library (cluster)
```

2. Read the data from the protein.csv file by using the following command:

```
food <- read.csv("C:/DataMining/Data/protein.csv")
```

3. Perform agglomerative hierarchal clustering by using the following command:

```
foodagg=agnes(food,diss=FALSE,metric="euclidian")
```

In the preceding command, you have used the `agnes` command, which is available in the package named cluster. The argument `diss=FALSE` indicates that you have used the dissimilarity matrix that is being calculated from raw data. The argument `metric="euclidian"` indicates that Euclidian distance is being used. No standardization is used as the default is average linkage.

4. Generate a dendrogram by using the following command:

```
plot(foodagg)
```

The dendrogram generated by the preceding command is shown in **Figure 13**:

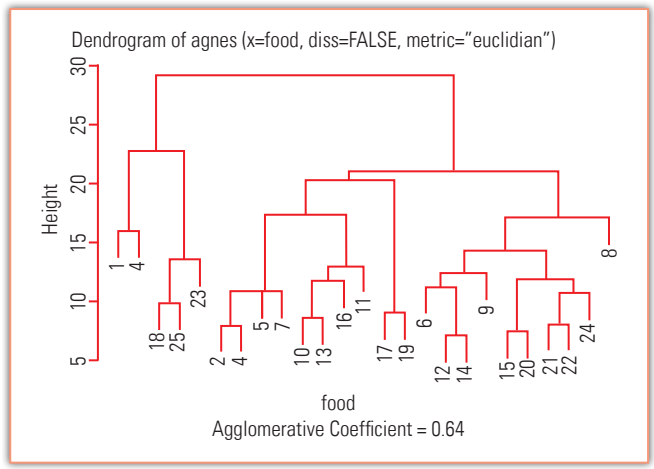


FIGURE 13
Dendrogram for Protein Consumption

Figure 13 shows the result of agglomerative clustering of the European protein data.

Now, let's revisit the example of the monthly US unemployment in the context of agglomerative hierarchical clustering.

Example 2: Monthly US Unemployment Rates Revisited

To implement agglomerative hierarchical clustering, let's revisit the data on US unemployment that we discussed earlier. Here you can use a 50 × 50 distance matrix, which is constructed as follows:

- The $(i,j)^{th}$ element of the distance matrix is defined as 1 minus the correlation coefficient of the first temporal differences (monthly changes) in states `i` and `j`.
- All pairwise correlations of the differenced time series are positive.
- For a correlation of one, the distance is zero and the two states are closely (in fact, perfectly) related.
- For a correlation close to zero, the distance is 1 and the states are different.

You may want to adjust a few outliers in the unemployment dataset. States AZ, LA, and MS have outliers. An adjusted dataset is also available. LA and MS data was adjusted to smooth out the impacts of Hurricane Katrina in the fall of 2005. The gap in AZ (we really do not know the reason for the sudden drop) was also smoothed out. The adjusted data is in the file `adj3unempstates.csv`.

Knowledge Check—4

1. You want to cluster n units (or objects), each described by p features, into a smaller number of groups to create a hierarchy of clusters that can be represented in a tree-like structure. In R, which of the following commands will you use to do this?
   a. `agnes`
   b. `order`
   c. `kmeans`
   d. `round`

EXAM CHECK
In your certification examination, you will be required to demonstrate your ability to implement hierarchical clustering in R.

- Clustering refers to the statistical operation of grouping objects (individuals or variables) into a limited number of groups known as clusters (or segments).
- Clusters have the following properties:
  - They are not defined in advance by the analyst, but are discovered during the operation.
  - They are combinations of objects with similar characteristics, which are separated from objects having different characteristics (resulting in internal homogeneity and external heterogeneity).
- Clustering technique is widely used in several fields for statistical analysis. Some of the areas where the clustering technique is used are as follows:
  - Marketing
  - Retail
  - Sociology
- As a general rule, the following expression is used to define the correct criteria for clustering and using efficient algorithms: `Bn(number of partitions for n objects)>exp(n)`
- The total sum of squares (or inertia) I of a population is the weighted mean (usually weighted by the inverse of the total frequency) of the squares of the distances of the individuals from the center of gravity of the population.
- Agglomerative Hierarchical Clustering (AHC) produces sequences of nested partitions of increasing heterogeneity, between partitions into n clusters, where each object is isolated and partitioned into one cluster, which includes all the objects. The general form of the algorithm is:
  - Step 1: The observations are the initial clusters.
  - Step 2: The distances between clusters are calculated.
  - Step 3: The two clusters closest together are merged and replaced with a single cluster.
  - Step 4: You start again at step 2 until there is only one cluster, which contains all the observations.
- Density estimation methods are most suitable for detecting the structure of complex clusters. Density is generally defined as the number of objects in a certain neighborhood. In the clustering process, density is estimated by using one of the following methods:
  - The k-nearest-neighbors method
  - The uniform Kernel method
  - The Wong Hybrid method
- Clustering by similarity aggregation is used to compare all the individuals in pairs at each step, thus building up a global clustering, instead of local clustering as in hierarchical clustering methods. This clustering technique determines the optimum number of clusters automatically, instead of fixing them in advance.
- The statistical tool R provides the amap package for clustering by similarity aggregation.
- k-means clustering is the most commonly used clustering technique. Common applications of k-means clustering are:
  - Predicting the price of products for a specific period or for specific seasons or occasions
  - Extracting information from electric price by time series models
- Two commonly used initialization approaches are followed for k-means clustering:
  - Randomly choosing k units from the dataset and using these as the initial cluster means.

- Randomly assigning one of the k clusters to each unit and then proceeding to the update step, and computing the initial means as the centroids of the clusters' randomly assigned units. In general, this random partition method is thought to be preferable.

- Hierarchical clustering creates a hierarchy of clusters that can be represented in a tree-like diagram, called a dendrogram. Agglomerative hierarchical clustering uses a bottom–up approach, where each unit starts in its own cluster and pairs of clusters are merged as you move up the hierarchy.

- For agglomerative clustering, you can use the `agnes()` function in the R package cluster. Alternatively, you can use the `hclust()` function in the stats package.