Unit 6.- Cluster analysis (CA)

Course: MULTIVARIATE STATISTICS

©Prof. Dr. José Luis Romero Béjar - Carlos Francisco Salto Díaz

November, 2023

What is cluster analysis?

- **Cluster analysis** (CA) is a multivariate technique whose main objective is **grouping objects** forming conglomerates (clusters) with a **high degree of internal homogeneity and external heterogeneity**.
- In other words, CA is an **exploratory procedure** that makes it possible to find **structures with similarities** in a data set with certain variability.
- The **motivation** of this technique is associated with the need to design a strategy that allows defining homogeneous groups.

  In this sense it is a **classification method**.
- CA has **application** in a wide number of situations in **different areas of science**: psychology, biology, sociology, economics, engineering, market research and marketing, etc.

Objective of cluster analysis

As said before, the objective of this analysis is to find groups (clusters) so that:

- The **homogeneity of the groups** should be **high**, that is, the **variability of the observations within each group** must be **low**.
- The **homogeneity between clusters** must be **low** or, in other words, the **variability between** elements of **different groups** must be **high**.

Some considerations

- **Similarity to factor analysis**:
  - While **factor analysis groups variables** according to certain latent factors, **cluster analysis groups objects**.
- It is usually used as **exploratory technique**.
- **Drawbacks**:
  - It is a merely **descriptive, a-theoretical and non-inferential** procedure.
  - **Does not offer unique solutions**.
    Even if there is a 'true' classification structure in the data, the CA solutions **depend on the variables considered and the method used**.
- **Advantages**:
  - It is a completely **objective procedure**. There is no information about the classification groups but rather these are constructed during the development of the analysis.

Objectives of the analysis

The objectives that are usually addressed by CA are:

- **Description of a taxonomy**: classification of objects carried out empirically (exploratory or confirmatory use).
- **Data simplification:** the cluster structure obtained simplifies the set of observations.
- **Relationship identification**: relationships between observations (relationships that may be hidden a priori).

Problem to be solved:

- **Variable selection** for the CA, since introducing irrelevant variables increases the possibility of errors.

Common selection criteria:

- You can perform a prior PCA and reduce the set of variables.
- Select only those variables that characterize the objects that are being grouped.

Research design using CA

There are a series of **prerequisites** that must be taken into account in the research design before carrying out a CA.

i. **Outlier detection** and possible exclusion.
The cluster analysis **is very sensitive to the presence of objects that are very different** from the rest (outliers). It is common to use **graphical methods** for identification.

ii. Define a **measure of similarity** between objects. A measure of similarity between objects is understood as a **measure of correspondence, or similarity**, between the objects that are to be grouped.

- For **metric** data, **correlation measures** or **distance** are usually used.
- For **non-metric** data, **measures of association** are often used.

iii. **Data standardization**.

- The order of similarities can change substantially with just a change in the scale of one of the variables.
- It is only standardized when necessary.

Below we briefly stop at the concept of **similarity**.

Research design using CA

In this phase, the selection of a measure that allows quantifying the relationship between elements is crucial.

- It will be possible to distinguish between **similarities**, which indicate how similar two observations are, and **distances or dissimilarities** that correspond to the metric concept of the analysis.
- The concepts of similarity and distance are called **anti-proportional**, that is, a **small similarity** corresponds to a **large distance** between observations while a **high similarity value** corresponds with a **small distance value**.

Depending on the data, an appropriate distance or similarity will be chosen.

- **Interval data:** Euclidean distance, Euclidean squared, cosine, Pearson correlation, Chebychev, Minkowski or a custom one (for more information consult: Similarity measures for interval data).
- **Binary data**: Euclidean distance, Euclidean squared, variance, dispersion, shape, simple reliability, Lambda, Anderberg's D, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance and Williams, etc. (for more information consult: Similarity measures for binary data).

Assumptions for the CA

- **Sample representativeness**.
  A good grouping will depend on the quality of the data considered.
- **Multicollinearity analysis** since variables that are correlated are implicitly weighted more strongly.

Procedure

In this step it is important to take into account the following aspects:

i. **Select the algorithm** to obtain the clusters.
   - Hierarchical methods.
   - Non-hierarchical methods.

ii. Appropriate number of clusters: **stopping rule**.

iii. **Model adequacy**: check that the model has not defined clusters with a single object or of very unequal sizes.

The philosophy of **hierarchical and non-hierarchical methods**, as well as some remarkable methods, is described below.

Hierarchical methods

Hierarchical methods are based on the construction of a tree-shaped structure called a classification tree or **dendrogram**.

Hierarchical methods are divided into two approaches:

- **Agglomerative processes**: those that seek to group clusters to form a new one.
- **Dissociative processes**: the starting point is a cluster that, in successive stages, is separated to obtain smaller and more homogeneous clusters.

Some examples of these methods are:

- Intergroup or intragroup linkage.
- Nearest neighbor (simple chaining) or furthest neighbor (full chaining).
- Centroid grouping.
- Linking medians.
- **Ward's method.**

Hierarchical methods: Ward's method (notation)

The *Ward* method is characterized by being a hierarchical approach in which, at each step of the process, **the two clusters that show the smallest increase in the total value of the sum of the squares of the differences of each individual with respect to the centroid of the cluster**. Let us note for

- x_{ij}^k refers to the value of the j-th variable for the i-th individual within cluster k, considering that this cluster contains n_k individuals.
- m^k represents the centroid of the cluster k, whose components are m_j^k.
- E_k denotes the sum of squares of the errors of cluster k, that is, the squared Euclidean distance between each individual in cluster k and its centroid.

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k - m_j^k \right)^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k \right)^2 - n_k \sum_{j=1}^{n} \left(m_j^k \right)^2$$

- E represents the total sum of the squares of the errors covering all clusters. In other words, if we consider that there are h clusters, then

$$E = \sum_{k=1}^{h} E_k$$

Hierarchical methods: Ward's method (process)

- The process starts with m clusters, each of which consists of a single individual, meaning that at this initial stage, each individual matches the center of the cluster. Therefore, in this first step, we have $E_k = 0$ for each cluster, which implies that $E = 0$.
- The main objective of the *Ward* method is to find, in each stage, the two clusters whose union generates the smallest increase in the total sum of errors, E.

Let us now suppose that the clusters C_p and C_q merge to give rise to a new cluster C_t. The increase in the value of E will be

$$\Delta E_{pq} = E_t - E_p - E_q = \frac{n_p n_q}{n_t} \sum_{j=1}^{n} \left(m_j^p - m_j^q \right)^2$$

The smallest increase in the squared errors is directly proportional to the squared Euclidean distance between the centroids of the merging clusters.

Hierarchical methods: Ward's method (example)

Let's see how this procedure is applied in an example with 5 individuals in which two variables are recorded. Below is the data:

Individual	X_1	X_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Hierarchical methods: Ward's method (example)

Level 1

First, we calculate the $\binom{5}{2} = 10$ possible combinations.

Partition	Centroids	E_k	E	ΔE
$(A,B),C,D,E$	$C_{AB} = (15, 12.5)$	$E_{AB} = 162.5$ $E_C = E_D = E_E = 0$	162.5	162.5
$(A,C),B,D,E$	$C_{AC} = (20, 7, 5)$	$E_{AC} = 212.5$ $E_B = E_D = E_E = 0$	212.5	212.5
$(A,D),B,C,E$	$C_{AD} = (20, 10)$	$E_{AD} = 250$ $E_B = E_C = E_E = 0$	250	250
$(A,E),B,C,D$	$C_{AE} = (7.5, 7.5)$	$E_{AE} = 25$ $E_B = E_C = E_D = 0$	25	25
$(B,C),A,D,E$	$C_{BC} = (25, 15)$	$E_{BC} = 100$ $E_A = E_D = E_E = 0$	100	100
$(B,D),A,C,E$	$C_{BD} = (25, 17.5)$	$E_{BD} = 62.5$ $E_A = E_C = E_E = 0$	62.5	62.5
$(B,E),A,C,D$	$C_{BE} = (12.5, 15)$	$E_{BE} = 162.5$ $E_A = E_C = E_D = 0$	162.5	162.5
$(C,D),A,B,E$	$C_{CD} = (30, 12.5)$	$E_{CD} = 12.5$ $E_A = E_B = E_E = 0$	12.5	12.5
$(C,E),A,B,D$	$C_{CE} = (17.5, 10)$	$E_{CE} = 312.5$ $E_A = E_B = E_D = 0$	312.5	312.5
$(D,E),A,B,C$	$C_{DE} = (17.5, 12.5)$	$E_{DE} = 325$ $E_A = E_B = E_C = 0$	325	325

Hierarchical methods: Ward's method (example)

From the data above, we can deduce that at this stage the elements C and D are merged. The current configuration is as follows: $(C, D), A, B, E$.

Level 2

With the current configuration, we take the $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ possible combinations.

Partition	Centroids	E_k	E	ΔE
$(A, C, D), B, E$	$C_{ACD} = (23.33, 10)$	$E_{ACD} = 316.6$ $E_B = E_E = 0$	316.66	304.16
$(B, C, D), A, E$	$C_{BCD} = (26.66, 15)$	$E_{BCD} = 116.66$ $E_A = E_E = 0$	116.66	104.16
$(C, D, E), A, B$	$C_{CDE} = (21.66, 11.66)$	$E_{CDE} = 433.33$ $E_A = E_B = 0$	433.33	420.83
$(A, B), (C, D), E$	$C_{AB} = (15, 12.5)$ $C_{CD} = (30, 12.5)$	$E_{AB} = 162.5$ $E_{CD} = 12.5$ $E_E = 0$	175	162.5
$(A, E), (C, D), B$	$C_{AE} = (7.5, 7.5)$ $C_{CD} = (30, 12.5)$	$E_{AE} = 25$ $E_{CD} = 12.5$ $E_B = 0$	37.5	25
$(B, E), (C, D), A$	$C_{BE} = (12.5, 15)$ $C_{CD} = (30, 12.5)$	$E_{BE} = 162.5$ $E_{CD} = 12.5$ $E_A = 0$	175	162.5

We can infer from this that at this stage the elements A and E are joined. The current configuration is as follows: $(A, E), (C, D), B$.

Hierarchical methods: Ward's method (example)

Level 3

With the current configuration, we take the $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ possible combinations.

Partition	Centroids	E_k	E	ΔE
$(A, C, D, E), B$	$C_{ACDE} = (18.75, 10)$	$E_{ACDE} = 568.75$ $E_B = 0$	568.75	531.25
$(A, B, E), (C, D)$	$C_{ABE} = (11.66, 11.66)$ $C_{CD} = (30, 12.5)$	$E_{ABE} = 233.33$ $E_{CD} = 12.5$	245.8	208.3
$(A, E), (B, C, D)$	$C_{AE} = (7.5, 7.5)$ $C_{BCD} = (26.66, 15)$	$E_{AE} = 25$ $E_{BCD} = 116.66$	141.66	104.16

We conclude that in this stage we join the clusters B and (C, D). The current configuration is $(A, E), (B, C, D)$.

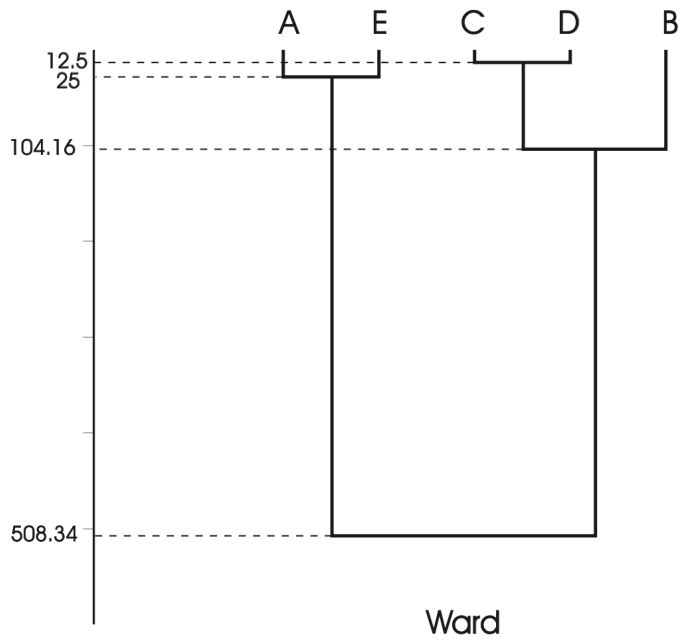
Hierarchical methods: Ward's method (example)

Level 4

It is evident that in this step the two existing clusters are merged. Below are the centroid values and the increments in the distances

Partition	Centroid	E	ΔE
(A, B, C, D, E)	$C_{ABCDE} = (19, 12)$	650	508.34

The corresponding dendrogram is shown in the following figure



Hierarchical methods: Ward's method (example)

Interpretation of the dendrogram

- *Height* is the essential key to interpreting how elements and clusters are grouped together to create clusters.

  When elements of a similar nature merge, it happens at a lower height in the dendrogram, while the union of elements with less similarity occurs at higher heights.

- The greater the height difference in the dendrogram (the more **clarity**), it will contribute to a better understanding the underlying structure of the data.

- For cluster identification, a **cut-based approach** of the dendrogram is used.

  This approach involves **drawing a horizontal line** on the dendrogram at a specific **height**. The **number of vertical lines intersected by this horizontal line determines the number of groups** that will be formed.

In this example, for **cuts at distances between** 25 **and** 104.16, 3 **clusters are obtained**, whilst for **distances greater than** 104.16, 2 **clusters**.

Non-hierarchical methods

Non-hierarchical methods follow a completely different structure than hierarchical methods. With these methods the aim is to classify the observations into K clusters where K has been previously set.

Characteristics:

- They do not involve building a tree structure.
- Objects are assigned to clusters once it has been decided when they will be formed.

The K means method stands out (**K-means**):

- Due to **MacQueen, 1967** it is considered one of the best and most widespread non-hierarchical cluster classification methods.
- It is included within the **unsupervised learning** techniques in the field of Data Mining.
- We start from the assumption of K initial clusters to, in successive iterations, classify the observations within the fixed K clusters.

A brief summary of the K-means algorithm is given below.

Non-hierarchical methods: K-means algorithm (steps)

- In the different iterations of this algorithm the value of the **centroid** plays a crucial role.
- This element is nothing more than the point in space that minimizes the sum of the squares of the distances of the observations to the centroid used in each stage.

Non-hierarchical methods: K-means algorithm (steps)

1. Input: Observations $\mathcal{L} = \{x_i, i = 1, 2, \ldots, n\}$, K = number of clusters.

2. Make one of the following decisions:
   - Perform an initial random assignment of the elements –observations– into K clusters and, for cluster k, calculate its current centroid, $\bar{x}_k$, $k = 1, 2, \ldots, K$.
   - Pre-specify the centroids of the K clusters $\bar{x}_k$, $k = 1, 2, \ldots, K$ (**the seed problem**).

3. Mean Square Error of each observation with respect to the centroid of its current cluster:

$$\text{MCE} = \sum_{k=1}^{K} \sum_{c(i)=k} (x_i - \bar{x}_k)^\tau (x_i - \bar{x}_k),$$

where $\bar{x}_k$ is the centroid of the k-th cluster and $c(i)$ is the cluster containing x_i.

4. Reassign each element to the cluster whose centroid is closest to said element, in this way the *MCE* will be greatly reduced.
The centroids of each cluster must be updated after this reassignment.

5. Repeat steps 3 and 4 until no remapping reduces the *MCE* value.

Non-hierarchical methods: K-means algorithm (drawbacks)

- It is a method **quite sensitive to the choice of the set of initial values** (seed points).

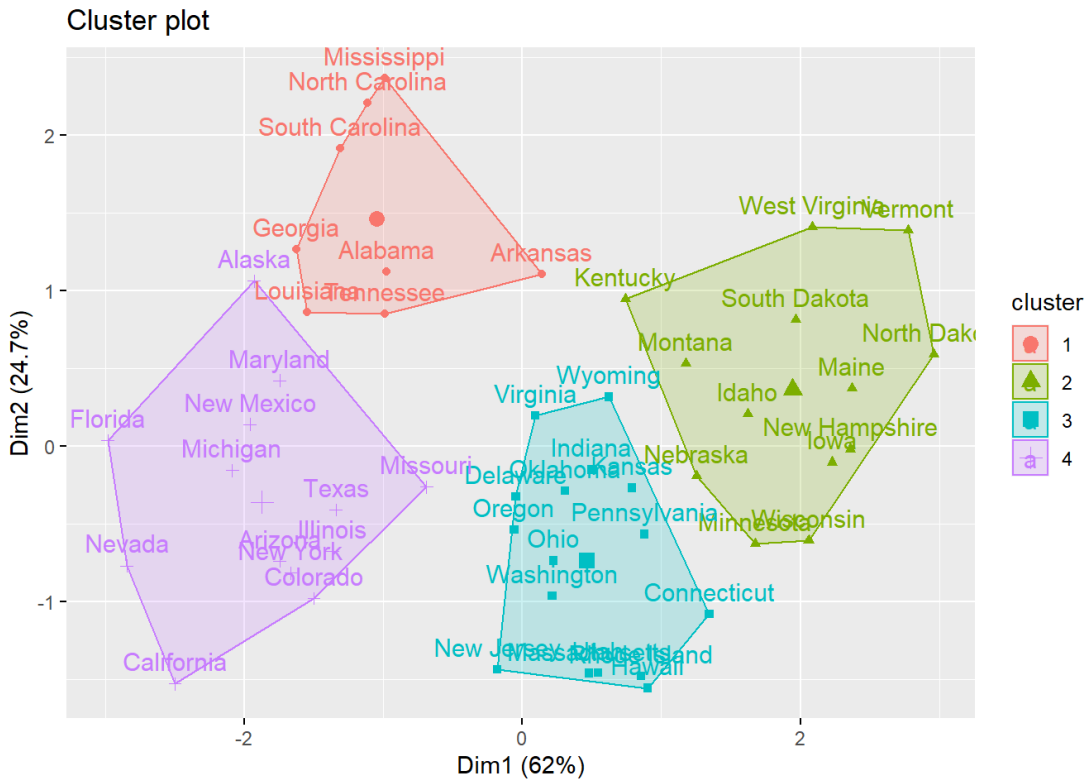
- A prominent **approach for choosing these seed points** is that proposed by Forgy, 1965, which considers K initial partitions mutually exclusive so that their centroids can be calculated and they are distinct. **These centroids are considered the seed points.**

- **It presents robustness problems against outlier data.** The solution is to exclude them or use more robust methods such as **K-medoids**.

- Requires the number of clusters to be **pre-specified**.

  This can be complicated if you do not have enough data information, although **there are strategies** that solve this problem of finding the **optimal number of clusters**.

Final graphical output of a CA with the K-means algorithm



Interpretation of the clusters

Once the different clusters have been identified, it is important to assign each one a precise label that describes their nature. There are different tools, among which stand out:

- **Close examination of centroids.** This is useful if you are working on non-standardised data, and only if it does not come from a PCA reduction.
- If the goal of the analysis was confirmatory, **contrast the classification obtained** with the preconceived data.
- Etc.

Validation and group profile

As a final step, it must be confirmed that the solution is representative of the general population. There are different tools, among which stand out:

- **Cophenetic correlation**.

  This is the correlation between the initial and final distances.

- **Stability** of the solution from different procedures within the cluster analysis.
- Etc.

CA Practice 3

In this practice, two examples of cluster analysis are illustrated using a **hierarchical method** and using the **non-hierarchical K-means method**.

To carry out this practice you must **download and execute** the file **Practice_3_CA.Rmd** available on the PRADO platform.

Topics covered:

- R packages required.
- Data preparation.
- Some distances for cluster analysis.
- Hierarchical clustering algorithm.
- Non-hierarchical clustering algorithm.

[1] Anderson, T.W. (2003, 3ª ed.). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons.

[2] Gutiérrez, R. y González, A. (1991). Estadística Multivariable. Introducción al Análisis Multivariante. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada.

[3] Härdle, W.K. y Simar, L. (2015, 4ª ed.). Applied Multivariate Statistical Analysis. Springer.

[4] Johnson, R.A. y Wichern, D.W. (1988). Applied Multivariate Analysis. Prentice Hall International, Inc.

[5] Rencher, A.C. y Christensen, W.F. (2012, 3ª ed.). Methods of Multivariate Analysis. John Wiley & Sons.

[6] Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. Online en http://www.5campus.com/leccion/aed.

[7] Timm, N.H. (2002). Applied Multivariate Analysis. Springer.

[8] Vera, J.F. (2004). Análisis Exploratorio de Datos. ISBN: 84-688-8173-2.