Tema 6.- Análisis cluster (AC)

Asignatura: ESTADÍSTICA MULTIVARIANTE

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

Noviembre, 2023

¿Qué es el análisis cluster?

- El **análisis cluster** (AC) es una técnica multivariante cuyo principal objetivo es **agrupar objetos** formando conglomerados (clusters) con un **alto grado de homogeneidad interna y heterogeneidad externa**.
- En otras palabras, el AC es un **procedimiento exploratorio** que permite encontrar **estructuras con similitudes** en un conjunto de datos con cierta variabilidad.
- La **motivación** de esta técnica va a asociada a la necesidad de diseñar una estrategia que permita definir grupos homogéneos. En este sentido es un **método de clasificación**.
- El AC tiene **aplicación** en un amplio número de situaciones de **distintas áreas de las ciencias**: psicología, biología, sociología, economía, ingeniería, investigación de mercados y marketing, etc.

Objetivo del análisis cluster

Tal y como se ha dicho anteriormente, el objetivo de este análisis es **encontrar grupos** (clusters) de manera que:

- La **homogeneidad de los grupos** debe ser **alta**, es decir, la **variabilidad de las observaciones dentro de cada grupo** ha de ser **baja**.
- La **homogeneidad entre los clusters** debe ser **baja** o lo que es lo mismo, la **variabilidad entre** elementos de **distintos grupos** debe ser **alta**.

Algunas consideraciones

- **Similitud con el análisis factorial**:
  - Mientras el **análisis factorial agrupa variables** según ciertos factores latentes, el **análisis cluster agrupa objetos**.
- Es habitualmente utilizada como **técnica exploratoria**.
- **Inconvenientes**:
  - Es un procedimiento meramente **descriptivo, a-teórico y no inferencial**.
  - **No ofrece soluciones únicas**.
    Aunque existiera una estructura de clasificación 'verdadera' en los datos, las soluciones del AC **dependen de las variables consideradas y del método empleado**.
- **Ventajas**:
  - Es un **procedimiento** totalmente **objetivo**. No se tiene información acerca de los grupos de clasificación sino que estos se construyen durante el desarrollo del análisis.

Objetivos del análisis

Los objetivos que usualmente son abordados mediante AC son:

- **Descripción de una taxonomía**: clasificación de objetos realizada empíricamente (uso exploratorio o confirmatorio).
- **Simplificación de los datos:** la estructura en cluster obtenida simplifica el conjunto de observaciones.
- **Identificación de la relación**: relaciones entre las observaciones (relaciones que a priori pueden estar ocultas).

Problema a resolver: **selección de variables** para el AC, ya que introducir variables irrelevantes aumenta la posibilidad de errores.

Criterios de selección usuales:

- Se puede realizar un ACP previo y resumir el conjunto de variables.
- Seleccionar solo aquellas variables que caracterizan los objetos que se van agrupando.

Diseño de la investigación mediante AC

Existen una serie de **requisitos previos** que hay que tener en cuenta en el diseño de la investigación antes de realizar un AC.

i. **Detección de outliers** y posible exclusión.
El análisis cluster **es muy sensible a la presencia de objetos muy diferentes** del resto (outliers). Es habitual utilizar **métodos gráficos** para su identificación.

ii. Definir una **medida de similitud** entre los objetos. Una medida de similitud entre objetos se entiende como una **medida de correspondencia, o parecido**, entre los objetos que van a ser agrupados.

   - Para **datos métricos** se suelen utilizar **medidas de correlación** o de **distancia**.
   - Para datos **no métricos** se suelen usar **medidas de asociación**.

iii. **Tipificar los datos**.

   - El orden de las similitudes puede cambiar sustancialmente con tan sólo un cambio en la escala de una de las variables.
   - Sólo se tipificará cuando sea necesario.

A continuación nos detenemos brevemente en el concepto de **similaridad**.

Diseño de la investigación mediante AC

En esta fase es esencial la selección de una medida que permita cuantificar la relación entre elementos.

- Se podrá distinguir entre **similaridades**, que indican como de parecidas son dos observaciones, y **distancias o disimilaridades** que corresponden al concepto métrico del análisis.
- Los conceptos de similaridad y distancia se denominan **anti-proporcionales**, es decir una **similaridad pequeña** se corresponde con una **gran distancia** entre observaciones mientras que un **valor alto de similaridad** se corresponde con un **valor pequeño de distancia**.

Dependiendo de los datos se elegirá una distancia o similaridad adecuada.

- **Datos de intervalo:** distancia euclídea, euclídea al cuadrado, coseno, correlación de Pearson, Chebychev, Minkowski o una personalizada (para más información consultar: Medidas de similaridad para datos de intervalo).
- **Datos binarios**: distancia euclídea, euclídea al cuadrado, varianza, dispersión, forma, concordancia simple, Lambda, D de Anderberg, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance y Williams, etc. (para más información consultar: Medidas de similaridad para datos binarios).

Supuestos para el AC

- **Representatividad de la muestra**.
  Un buen agrupamiento dependerá de la calidad de los datos considerados.
- **Análisis de la multicolinealidad** ya que las variables que están correlacionadas están implícitamente ponderadas con más fuerza.

Procedimiento

En este paso es importante tener en cuenta los siguientes aspectos:

i. **Elegir el algoritmo** para la obtención de los clusters.
   - Métodos jerárquicos.
   - Métodos no jerárquicos.

ii. Número de clusters adecuado: **regla de parada**.

iii. **Adecuación del modelo**: comprobar que el modelo no ha definido clusters con un solo objeto o de tamaños muy desiguales.

A continuación se describe la filosofía de los **métodos jerárquicos y no jerárquicos**, así como algunos métodos destacados.

Métodos jerárquicos

Los **métodos jerárquicos** se basan en la construcción de una estructura en forma de árbol denominada árbol de clasificación o **dendograma**.

Los métodos jerárquicos se dividen en dos corrientes:

- **Procesos aglomerativos**: aquellos que persiguen agrupar clusters para formar uno nuevo.
- **Procesos disociativos**: se parte de un cluster que, en etapas sucesivas, se separa para obtener clusters más pequeños y homogéneos.

Algunos ejemplos de estos métodos son:

- Vinculación intergrupos o intragrupos.
- Vecino más próximo (encadenamiento simple) o vecino más lejano (encadenamiento completo).
- Agrupación de centroides.
- Vinculación de medianas.
- **Método de Ward.**

Métodos jerárquicos: método de Ward (notación)

El método de *Ward* se caracteriza por ser un enfoque jerárquico en el cual, en cada paso del proceso, **se fusionan los dos clusters que muestren el menor aumento en el valor total de la suma de los cuadrados de las diferencias de cada individuo con respecto al centroide del cluster**. Notemos por

- x_{ij}^k se refiere al valor de la j-ésima variable para el i-ésimo individuo dentro del cluster k, considerando que este cluster contiene n_k individuos.
- m^k representa el centroide del cluster k, cuyas componentes son m_j^k.
- E_k denota la suma de cuadrados de los errores del cluster k, es decir, la distancia euclídea al cuadrado entre cada individuo en el cluster k y su centroide.

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

- E representa la suma total de los cuadrados de los errores que abarcan todos los clusters. En otras palabras, si consideramos que existen "h" clusters en total, resulta

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

Métodos jerárquicos: método de Ward (proceso)

- El proceso comienza con m clusters, cada uno de los cuales consta de un solo individuo, lo que significa que en esta etapa inicial, cada individuo coincide con el centro del cluster. Por lo tanto, en este primer paso, se tiene $E_k = 0$ para cada cluster, lo que implica que $E = 0$.
- El objetivo principal del método de *Ward* radica en encontrar, en cada etapa, los dos clusters cuya unión genere el menor aumento en la suma total de errores, E.

Imaginemos ahora que los clusters C_p y C_q se fusionan para dar lugar a un nuevo cluster C_t. El incremento en el valor de E será

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

El menor incremento en los errores cuadráticos es directamente proporcional a la distancia euclídea al cuadrado entre los centroides de los clusters que se fusionan.

Métodos jerárquicos: método de Ward (ejemplo)

Veamos como se aplica este procedimiento en un ejemplo con 5 individuos en los que se registran dos variables. A continuación, se presentan los datos:

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

Métodos jerárquicos: método de Ward (ejemplo)

Nivel 1

En primer lugar, calculamos las $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$ posibles combinaciones.

Partición	Centroides	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

Métodos jerárquicos: método de Ward (ejemplo)

A partir de los datos anteriores, podemos deducir que en esta etapa se fusionan los elementos C y D. La configuración actual es la siguiente: $(C, D), A, B, E$.

Nivel 2

Con la configuración actual, tomamos las $\binom{4}{2} = 6$ combinaciones posibles.

Partición	Centroides	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

Podemos inferir a partir de esto que en esta etapa se unen los elementos A y E. La configuración actual es la siguiente: $(A, E), (C, D), B$.

Métodos jerárquicos: método de Ward (ejemplo)

Nivel 3

Con la configuración actual, tomamos las $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ combinaciones posibles.

Partición	Centroides	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

Concluimos que en esta etapa unimos los clústeres B y (C, D). La configuración actual es $(A, E), (B, C, D)$.

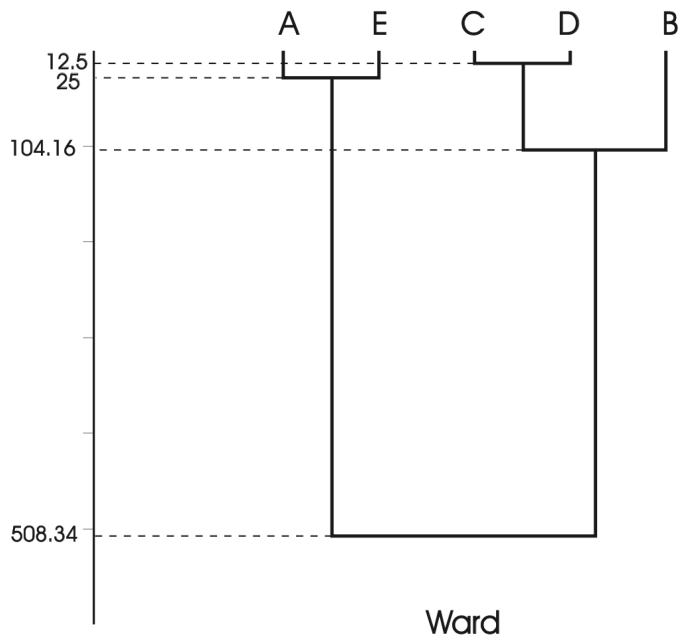
Métodos jerárquicos: método de Ward (ejemplo)

Nivel 4

Es evidente que en este paso se fusionarán los dos clústeres existentes. A continuación, se presentan los valores del centroide y los incrementos en las distancias

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

El dendrograma correspondiente se muestra en la siguiente figura



Métodos jerárquicos: método de Ward (ejemplo)

Interpretación del dendograma

- La *altura* es la clave esencial para la interpretación de cómo los elementos y los clústeres se agrupan para crear clusters.

  Cuando elementos de naturaleza similar se fusionan, sucede a una menor altura en el dendrograma, mientras que la unión de elementos con menor semejanza se manifiesta a alturas superiores.

- Cuanto **mayor sea la diferencia de alturas** en el dendrograma, **mayor claridad** aportará a la comprensión de la estructura subyacente de los datos.

- Para la identificación de los clusters, se emplea un **enfoque basado en el corte** del dendrograma.

  Este enfoque implica **trazar una línea horizontal** en el dendrograma a una **altura específica**. La **cantidad de líneas verticales intersectadas por esta línea horizontal determina el número de grupos** que se formarán.

En este ejemplo, para **cortes en distancias entre** 25 **y** 104.16 **se obtienen** 3 **clusters**, mientras que para **distancias superiores a** 104.16 **se obtienen** 2 **clusters**.

Métodos no jerárquicos

Los **métodos no jerárquicos**, siguen una estructura completamente diferente a la de los métodos jerárquicos. Con estos métodos se pretende clasificar las observaciones en K clusters donde K ha sido fijado previamente.

Características:

- No implican la construcción de una estructura de árbol.
- Los objetos se asignan a los clusters una vez que se ha decidido cuando se van a formar.

Destaca el método de las K medias (**K-means**):

- Debido a **MacQueen, 1967** es considerado uno de los mejores y más extendidos métodos no jerárquicos de clasificación por clusters.
- Se engloba dentro de las técnicas de **aprendizaje no supervisado** en el ámbito de la Minería de Datos.
- Parte la suposición de K clusters iniciales para, en sucesivas iteraciones, clasificar las observaciones en los K clusters fijados.

A continuación se da un breve esbozo del algoritmo K-means.

Métodos no jerárquicos: algoritmo K-means (pasos)

- En las distintas iteraciones de este algoritmo juega un papel crucial el valor del **centroide**.
- Este elemento no es más que el punto del espacio que minimiza la suma de los cuadrados de las distancias de las observaciones al centroide que se utiliza en cada etapa.

Métodos no jerárquicos: algoritmo K-means (pasos)

1. Input: Observaciones $\mathcal{L} = \{x_i, i = 1, 2, \ldots, n\}$, $K =$ número de clusters.

2. Tomar una de las decisiones siguientes:
   - Realizar una asignación aleatoria inicial de los elementos –observaciones– en K clusters y, para el cluster k, calcular su centroide actual, $\bar{x}_k$, $k = 1, 2, \ldots, K$.
   - Especificar previamente los centroides de los K clusters, $\bar{x}_k$, $k = 1, 2, \ldots, K$.

3. Error Cuadrático Medio de cada observación con respecto al centroide de su cluster actual:

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

donde $\bar{x}_k$ es el centroide del cluster k-ésimo y $c(i)$ es el cluster que contiene a x_i.

4. Reasignar cada elemento al cluster cuyo centroide se encuentre más cerca de dicho elemento, de esta manera el *ECM* se verá reducido en gran magnitud.
Se deben actualizar los centroides de cada cluster después de esta reasignación.

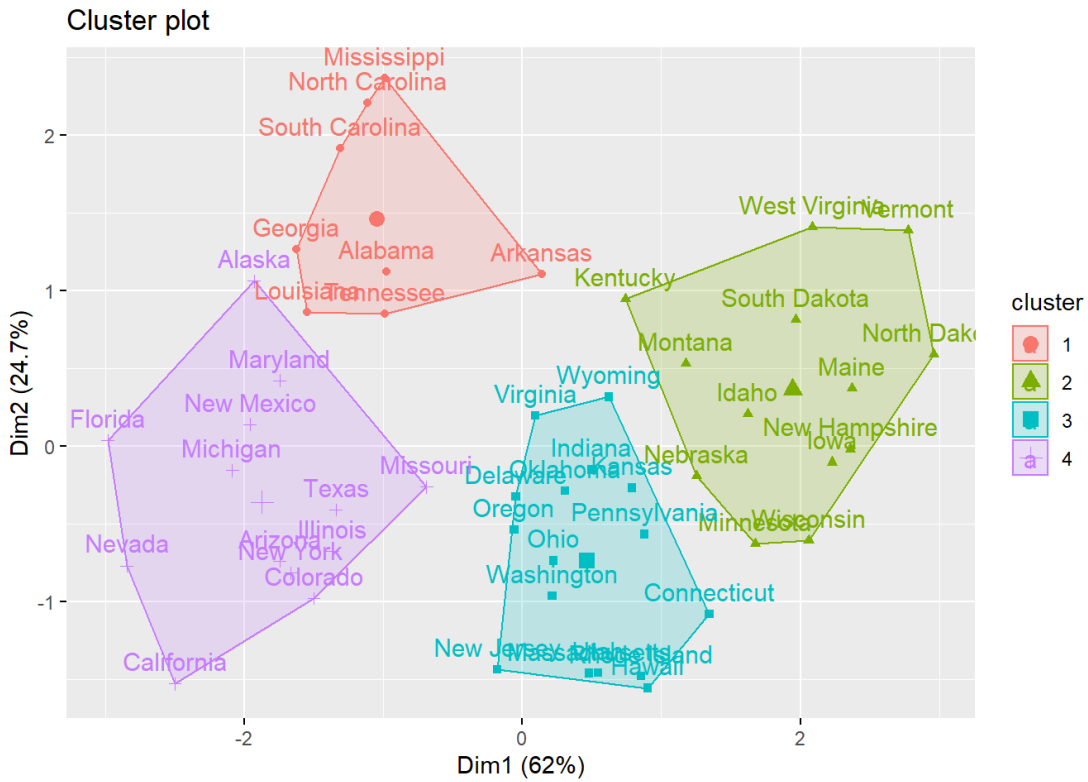
5. Repetir los pasos 3 y 4 hasta que ninguna reasignación reduzca el valor del *ECM*.

Métodos no jerárquicos: algoritmo K-means (inconvenientes)

- Es un método **bastante sensible a la elección del conjunto de valores iniciales** (puntos semilla).
- Un **enfoque** destacado **para la elección de estos puntos semilla** es el propuesto por Forgy, 1965, que considera K particiones iniciales mutuamente excluyentes de manera que es posible calcular sus centroides y son distintos. **Estos centroides se consideran los puntos semilla.**
- **Presenta problemas de robustez frente a datos outlier.** La solución es excluirlos o utilizar métodos más robustos com el **K-medoids**.
- Requiere que se **especifique previamente el número de clusters**.

  Esto puede ser complicado si no se tiene suficiente información de los datos, aunque **hay estrategias** que resuelven este problema de encontrar el **número óptimo de clusters**.

Salida gráfica final de un AC con el algoritmo K-means



Interpretación de los conglomerados

Una vez identificados los distintos clusters es importante **asignar a cada uno una etiqueta precisa** que describa la naturaleza de los mismos. Existen distintas herramientas, entre las que destacan:

- **Examen minucioso de los centroides.** Esto es útil si se trabaja sobre datos no tipificados, y sólo si no provienen de una reducción mediante ACP.
- Si el objetivo del análisis era confirmatorio, **contrastar la clasificación obtenida** con los datos preconcebidos.
- Etc.

Validación y perfil de los grupos

Como paso final ha de confirmarse que la solución es representativa de la población en general. Existen distintas herramientas, entre las que destacan:

- **Correlación cofenética**.
  Esto es la correlación entre las distancias iniciales y las finales.
- Estabilidad de la solución desde distintos procedimientos dentro del análisis cluster.
- Etc.

Práctica 3 de AC

En esta práctica se ilustran dos ejemplos de análisis cluster mediante un **método jerárquico** y mediante el **método no jerárquico K-means**.

Para la realización de esta práctica hay que **descargar y ejecutar** el archivo **Practica_3_AC.Rmd** disponible en la plataforma PRADO.

Aspectos tratados:

- Paquetes de R necesarios.
- Preparación de los datos.
- Algunas distancias para análisis cluster.
- Algoritmo de agrupamiento jerárquico.
- Algoritmo de agrupamiento no jerárquico.

[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.