# INFORME

## “Agrupamiento del aceite de oliva según sus características químicas”

Navarro Castillo, Alejandra

Análisis de Datos

Junio 2021

## Índice.



## Resumen

En este informe se aplican métodos de análisis clúster con el objetivo de diferenciar y agrupar observaciones de una muestra de aceites de oliva según sus características, recogidas en el conjunto de datos en forma de ocho variables. Finalmente se caracteriza cada grupo de aceites de oliva resultante.

## Introducción y objetivos

Dado el conjunto de datos AO.RData que contiene 572 observaciones de aceites de oliva italianos y ocho variables que describen la cantidad de algunos componentes químicos de estos aceites, el objetivo del problema es intentar saber si existen diferentes tipos de aceite de oliva.

Para ello tenemos las medidas de ocho tipos de ácidos en cada muestra. Estas variables son:

- “Palmitic”
- “Palmitoleic”
- “Stearic”
- “Oleic”
- “Linoleic”
- “Linolenic”
- “Arachidic”
- “Eicosenoic”

Para intentar lograr el objetivo del problema, vamos a mirar las características comunes de las observaciones e intentar agruparlas con técnicas de análisis clúster.

## Descripción de los datos

Como ya hemos dicho, nuestro fichero contiene 572 observaciones. Describiremos a continuación el resumen de los datos.

Resumen de los datos

| | mean | sd | IQR | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|---|---|
| palmitic | 1231.74 | 168.59 | 265.00 | 610 | 1095.00 | 1201.0 | 1360.00 | 1753 |
| palmitoleic | 126.09 | 52.49 | 81.50 | 15 | 87.75 | 110.0 | 169.25 | 280 |
| stearic | 228.87 | 36.74 | 44.00 | 152 | 205.00 | 223.0 | 249.00 | 375 |
| oleic | 7311.75 | 405.81 | 680.00 | 6300 | 7000.00 | 7302.5 | 7680.00 | 8410 |
| linoleic | 980.53 | 242.80 | 410.00 | 448 | 770.75 | 1030.0 | 1180.75 | 1470 |
| linolenic | 31.89 | 12.97 | 14.25 | 0 | 26.00 | 33.0 | 40.25 | 74 |
| arachidic | 58.10 | 22.03 | 20.00 | 0 | 50.00 | 61.0 | 70.00 | 105 |
| eicosenoic | 16.28 | 14.08 | 26.00 | 1 | 2.00 | 17.0 | 28.00 | 58 |

Las características más notables de este conjunto de datos es que hay mucha diferencia entre las cantidades de algunos ácidos. Por ejemplo, el ácido oleico “oleic” tiene unas cantidades mucho mayores que el resto de las variables. Esto es algo para tener en cuenta posteriormente en el análisis.

También es algo notable que en los ácidos linolénico “linolenic” y araquídico “arachidic” el menor valor es cero, lo cual significa que el nivel de esa sustancia está por debajo de la sensibilidad de la prueba en algunas observaciones. También se tendrá en cuenta esta observación si queremos pasar a medir las variables en una escala logarítmica ya que es una manera frecuente de medición en química.

## Preparación de los datos

La suma de las ocho componentes de cada muestra debería ser 10.000. Observamos que en nuestra base de datos no es así. Esto puede deberse a los errores en la medición de las variables.

Debido a esto, se realiza una transformación de las variables para trabajar con una medida relativa posteriormente en el análisis.

Teniendo en cuenta esto último y nuestro interés por transformar los datos a una escala logarítmica, se va a realizar la siguiente transformación:

$$x_{ij}{}_{nueva} = log\left(\frac{x_{ij}+1}{\sum_{j=1}^{8}(x_{ij}+1)}\right)$$

donde, i es el índice de las observaciones y j es el índice para las variables.

Con esta transformación hemos tipificado los datos para que no se tengan en cuenta los errores de medición (al dividir todas las variables por una cantidad fija para cada observación) y luego hemos expresado los datos en una escala logarítmica.

## Análisis clúster

El objetivo de este análisis es tanto saber en cuántos grupos se dividen los datos del fichero dado un criterio de agregación, como saber qué individuos pertenecen a cada grupo.

Para empezar, se hace un análisis de componentes principales. El objetivo de utilizar las componentes principales es que, mediante este procedimiento, aunque se pierde variabilidad específica de cada grupo, se captura adecuadamente la variabilidad original entre grupos, que es nuestro principal objetivo.

El análisis de componentes principales se va a realizar utilizando la matriz de varianzas covarianzas de las observaciones debido a que las unidades de las variables son homogéneas y por ello las magnitudes son directamente comparables.

*Tabla 1. Componentes principales*

Importance of components:

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 |
|---|---|---|---|---|---|---|---|---|
| Standard deviation | 1.413557 | 0.9412609 | 0.45599769 | 0.34961726 | 0.173875262 | 0.144743343 | 0.073760130 | 6.502053e-03 |
| Proportion of Variance | 0.610876 | 0.2708610 | 0.06356993 | 0.03736906 | 0.009242767 | 0.006405066 | 0.001663297 | 1.292491e-05 |
| Cumulative Proportion | 0.610876 | 0.8817370 | 0.94530689 | 0.98267594 | 0.991918712 | 0.998323778 | 0.999987075 | 1.000000e+00 |

*Tabla 2. Pesos de las variables en cada componente principal*

```
Loadings:
            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
palmitic                   0.129  0.106  0.151         0.955  0.178
palmitoleic  0.123  0.207  0.698  0.421  0.450 -0.151 -0.229
stearic                                -0.117 -0.985
oleic                                                -0.176  0.976
linoleic                   0.415  0.234 -0.866                0.122
linolenic    0.512 -0.384 -0.388  0.662
arachidic    0.525 -0.565  0.367 -0.519
eicosenoic   0.664  0.697 -0.160 -0.202
```

Como observamos en la Tabla 1, las dos primeras componentes “Comp.1” y “Comp.2” representan el 88% de la variabilidad total de la muestra, por lo que ya nos dan una buena comprensión de cómo se relacionan los datos entre sí. La primera componente principal “Comp.1” representa una mayor presencia de los ácidos “linolenic”, “arachidic” y “eicosenoic”, mientras que la segunda componente principal “Comp.2” representa la relación de los ácidos “linolenic” y “arachidic” con respecto a “eicosenoic”.

Si representamos estas dos variables mediante una nube de puntos, podemos ir comprendiendo y observando cuánto y cómo de cerca se encuentran las distintas observaciones del conjunto de datos.

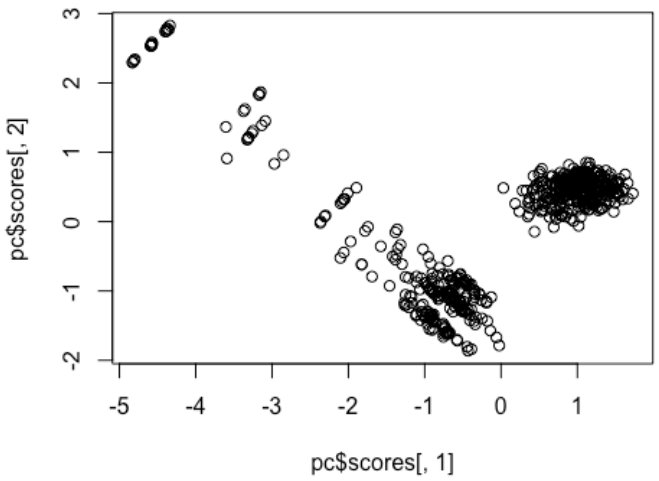
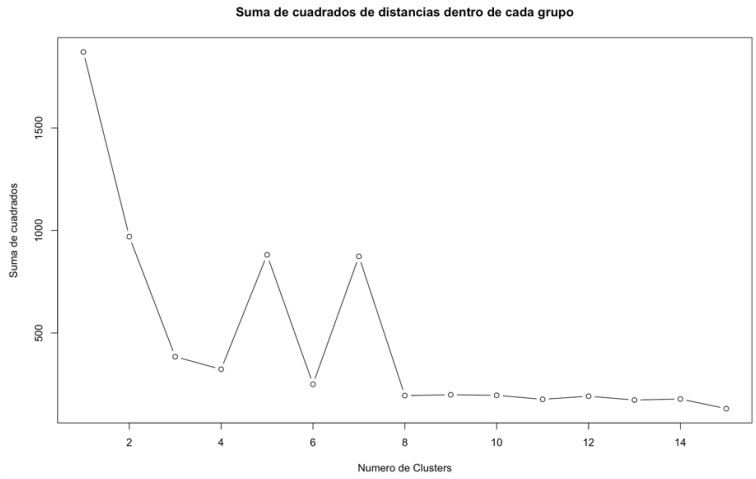



*Gráfico 1. Nube de puntos de las dos primeras componentes principales*

A primera vista y representando solo las dos primeras componentes principales, se podrían diferenciar 4 grupos los cuales se separan en el espacio notablemente.

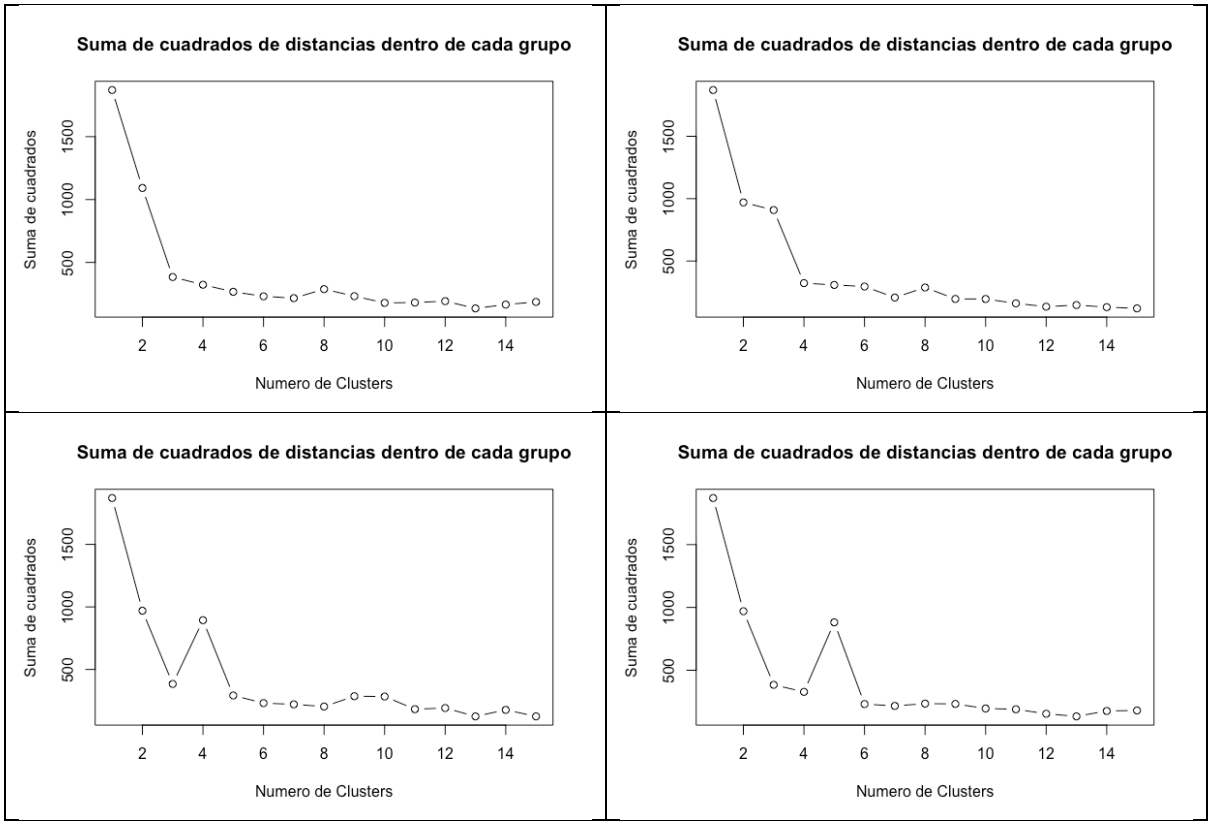
Para ser más concretos, vamos a aplicar un método para determinar cuántos grupos de aceites se pueden considerar. El método será el de la suma de cuadrados de las distancias dentro de cada grupo al hacer el algoritmo de las k-medias.

La siguiente gráfica muestra la suma de cuadrados de las distancias dentro de cada grupo para diferentes números de clústeres.




Interpretando el gráfico anterior, debemos elegir un número de clústeres bajo y tal que haya un salto en la diferencia de la suma de cuadrados con el número anterior. En este caso se podría decir que podríamos elegir 6 u 8 clústeres.

No obstante, debemos tener en cuenta que según cómo se elijan los centros en la primera iteración del algoritmo k-means, el resultado de las gráficas puede variar y no ser concluyente. Este hecho lo constatamos al realizar varias veces este procedimiento.

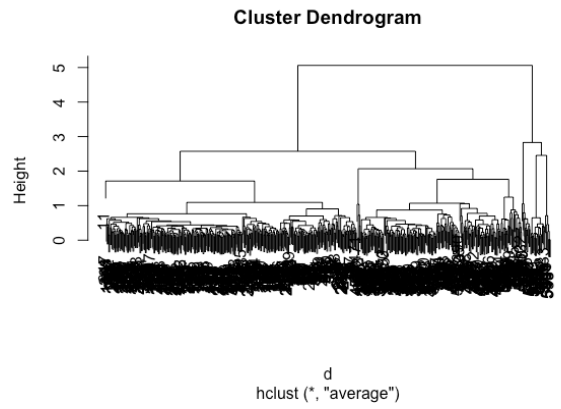


Por lo tanto, vamos a utilizar otro procedimiento. Ahora intentaremos saber el número de clústeres haciendo un **análisis jerárquico ascendente** con distintos métodos de agregación.

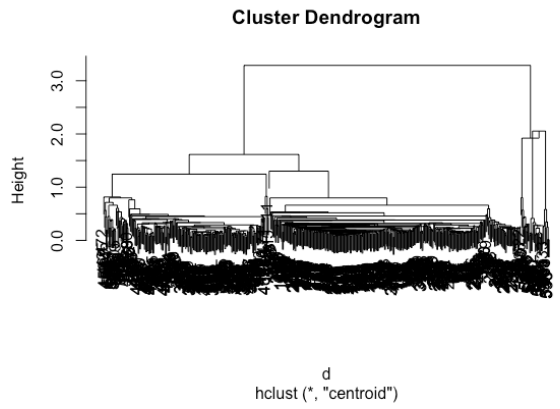
En todos los métodos se va a utilizar la distancia euclídea para medir la distancia entre puntos ya que todas las variables tienen las mismas unidades. Notar que si utilizáramos otra distancia, el resultado podría cambiar.

Vamos a ver a continuación los dendogramas para distintos criterios de agregación.

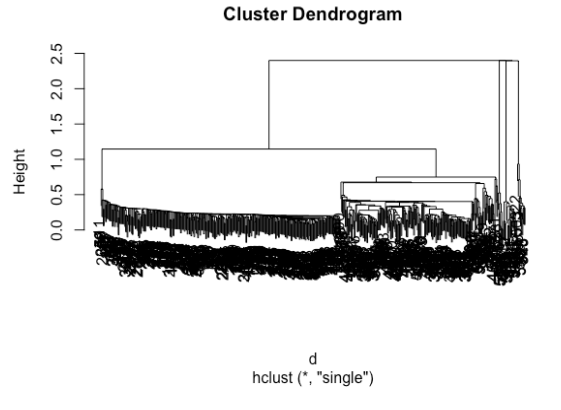
*Dendograma 1. Criterio de la media.*



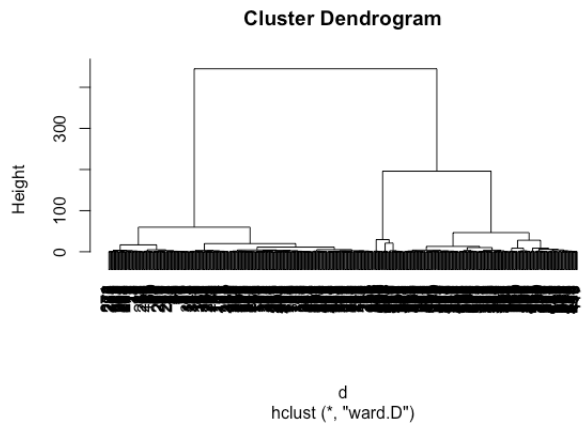

*Dendograma 2. Criterio del centroide.*




*Dendograma 3. Criterio del salto mínimo.*

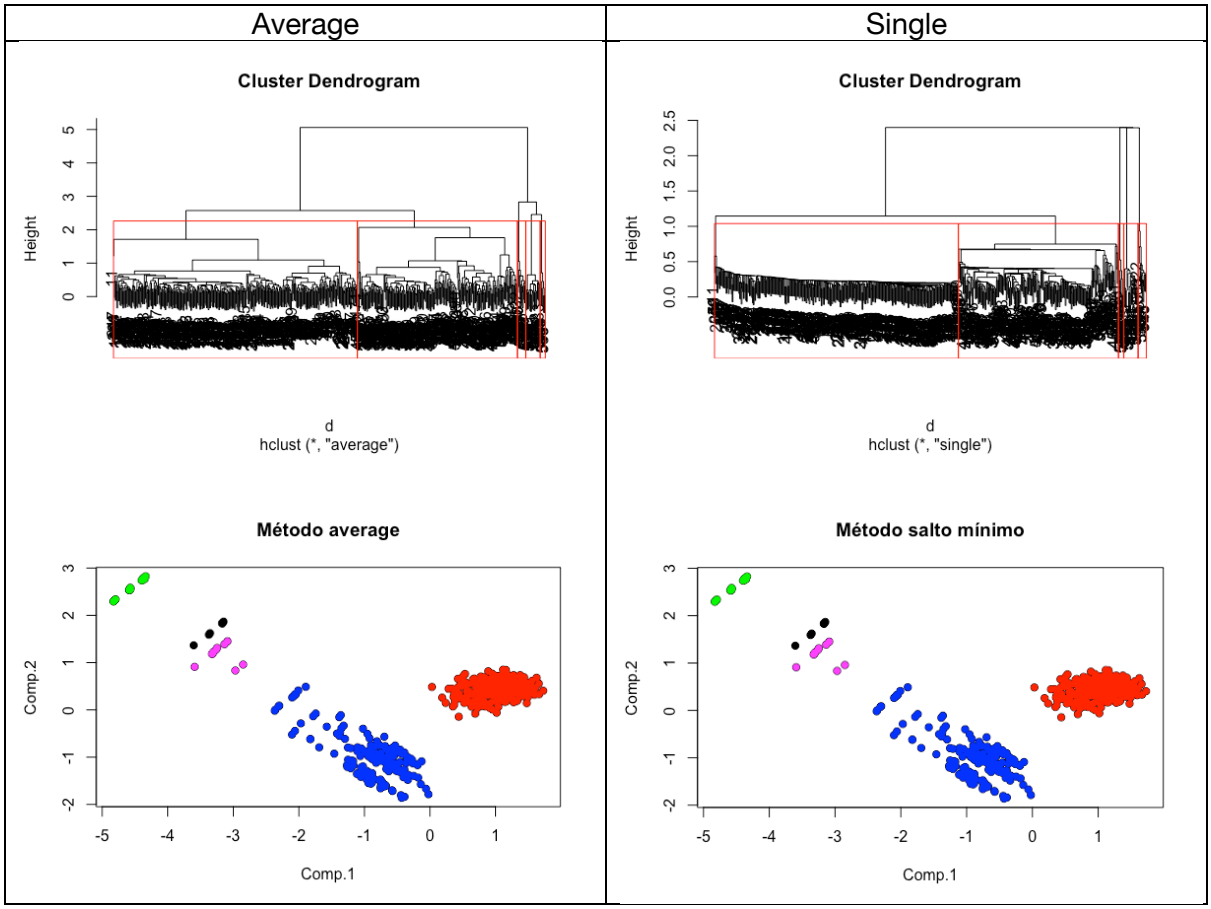



*Dendograma 4. Criterio de ward.*




De los dendogramas inferimos que un buen corte de todos los dendogramas podría ser 4, 5 o 6 clústeres.

Elegimos hacer el análisis en 5 clústeres y con los criterios “average” y “single”. En la siguiente tabla se muestran los correspondientes dendogramas de ambos criterios y a continuación la nube de puntos en las dos primeras componentes principales con los distintos grupos a los que pertenece cada individuo, representados con distintos colores.




Esta comparativa nos da mucha información ya que el hecho de que mediante estos dos criterios se llegue a la misma solución, es determinante porque estos dos criterios actúan de manera muy distinta (el de la media “average” está basado en distancias promedio entre grupos y el del salto mínimo “single” está basado en la mínima de las distancias).

Por lo tanto, podemos concluir que los aceites se pueden disociar en 5 grupos diferenciados.

## Resultados

### Caracterización de cada grupo resultante

Primero veamos cuantos elementos hay de cada grupo.

| Grupo 1 | Grupo 2 | Grupo 3 | Grupo 4 | Grupo 5 |
|---|---|---|---|---|
| 323 | 212 | 11 | 19 | 7 |

Ahora procedemos a analizar la composición química de cada grupo para ver lo que lo caracteriza con respecto a los demás. Lo vemos a través de los estadísticos resumen de las primeras dos componentes principales en cada grupo.

```
, , Variable = Comp.1

     Statistic
Group   mean    sd   IQR     0%    25%    50%    75%   100%
    1  1.007 0.309 0.410  0.030  0.809  1.017  1.219  1.727
    2 -0.850 0.473 0.444 -2.368 -0.981 -0.734 -0.538 -0.024
    3 -3.219 0.200 0.202 -3.591 -3.314 -3.263 -3.112 -2.850
    4 -4.553 0.168 0.201 -4.829 -4.589 -4.577 -4.388 -4.338
    5 -3.284 0.170 0.197 -3.605 -3.364 -3.171 -3.167 -3.149

, , Variable = Comp.2

     Statistic
Group   mean    sd   IQR     0%    25%    50%    75%  100%
    1  0.448 0.181 0.249 -0.145  0.332  0.465  0.581 0.851
    2 -1.031 0.450 0.473 -1.861 -1.335 -1.082 -0.862 0.488
    3  1.175 0.196 0.222  0.834  1.070  1.211  1.292 1.451
    4  2.581 0.175 0.210  2.294  2.536  2.555  2.747 2.827
    5  1.706 0.187 0.225  1.365  1.607  1.832  1.832 1.867
```

Se puede interpretar la caracterización de cada uno de los grupos de las tablas anteriores, pero como hemos visto ya lo que representa cada componente principal, vamos a calcular los estadísticos resumen en las variables originales que más influyen en las componentes principales: “linolenic”, “arachidic” y “eicosenoic”.

```
, , Variable = linolenic

     Statistic
Group   mean    sd   IQR     0%    25%    50%    75%   100%
    1 -5.565 0.206 0.311 -6.166 -5.713 -5.570 -5.401 -4.891
    2 -5.934 0.422 0.435 -6.815 -6.120 -5.844 -5.685 -4.949
    3 -9.212 0.001 0.000 -9.214 -9.211 -9.211 -9.211 -9.210
    4 -9.211 0.003 0.000 -9.221 -9.211 -9.211 -9.211 -9.201
    5 -6.814 0.002 0.002 -6.817 -6.816 -6.814 -6.813 -6.812

, , Variable = arachidic

     Statistic
Group   mean    sd   IQR     0%    25%    50%    75%   100%
    1 -5.063 0.173 0.211 -5.715 -5.168 -5.065 -4.957 -4.576
    2 -5.215 0.549 0.526 -6.815 -5.381 -5.021 -4.855 -4.548
    3 -6.696 0.262 0.001 -6.817 -6.813 -6.813 -6.813 -6.167
    4 -9.211 0.003 0.000 -9.221 -9.211 -9.211 -9.211 -9.201
    5 -9.212 0.002 0.002 -9.214 -9.213 -9.211 -9.211 -9.210

, , Variable = eicosenoic

     Statistic
Group   mean    sd   IQR     0%    25%    50%    75%   100%
    1 -5.910 0.306 0.365 -6.813 -6.075 -5.879 -5.710 -5.131
    2 -8.173 0.258 0.407 -8.526 -8.515 -8.113 -8.109 -7.818
    3 -8.071 0.198 0.143 -8.518 -8.113 -8.113 -7.970 -7.825
    4 -8.092 0.262 0.293 -8.518 -8.118 -8.113 -7.825 -7.815
    5 -8.007 0.262 0.285 -8.518 -8.112 -7.828 -7.827 -7.825
```

De las tablas anteriores, podemos interpretar las siguientes caracterizaciones para los 5 grupos.

El grupo 1 se caracteriza por tener una mayor presencia de la primera componente principal, es decir mayor presencia de los ácidos “linolenic”, “arachidic” y “eicosenoic” que los demás grupos.

El grupo 2 se caracteriza por tener poca presencia del ácido “eicosenoic” en comparación con los ácidos “linolenic” y “arachidic”.

Los grupos 3,4 y 5 son bastante parecidos entre sí (además son los grupos con menos individuos por grupo).

Lo que caracterizaría a los tres es que tienen poca cantidad de los ácidos “linolenic”, “arachidic” y “eicosenoic” y lo que diferenciaría a cada uno de ellos con respecto a los otros dos es que: el grupo 3 tiene más cantidad de ácido “arachidic”, el grupo 4 tiene las cantidades más bajas y el grupo 5 tiene más cantidad de ácido “linolenic”.

## Discusión

Hemos visto que la elección del número de grupos es bastante poco rigurosa. Lo único que tenemos claro es que hay 3 macro grupos que corresponderían al Grupo 1, Grupo 2 y los grupos 3,4 y 5 formarían el tercer macro grupo.

Sin embargo, en algunos criterios de agregación hemos visto que se podrían hacer aún más grupos (aunque menos exactos) para llegar a diferenciar mejor entre más tipos de aceites de oliva.

## Bibliografía


- Quick-R Guide *statmethods.net* https://www.statmethods.net/advstats/cluster.html

- Apuntes de la asignatura Análisis de Datos (Universidad de Oviedo) por Norberto Corral Blanco y Beatriz Sinova Fernández.

## Apéndice

Código de R para desarrollar el análisis:

```
load("~/Desktop/AnalisisDatos/Informes/AO.rdata")

## DESCRIPCIÓN DE LOS DATOS
dim(AO)
round(numSummary(AO[-1])$table, 2)

## PREPARACIÓN DE LOS DATOS
AO <- AO[-1]
AO <- AO + 1
AO <- AO/apply(AO, 1, sum)
AO <- log(AO)

## ANÁLISIS CLUSTER
pc <-
princomp(~palmitic+palmitoleic+stearic+oleic+linoleic+linolenic+arachidic+e
icosenoic, data=AO, cor=FALSE)
summary(pc)
pc$loadings #coef de las variables
plot(pc$scores[,1], pc$scores[,2])

## Decidir cuántos grupos tomo
pc <- data.frame(pc$scores)
wss <- (nrow(pc)-1)*sum(apply(pc,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(pc,centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Numero de Clusters", ylab="Suma de
cuadrados", main="Suma de cuadrados de distancias dentro de cada grupo")

## CRITERIO AVERAGE
ng <- 5 #5 grupos
d <- dist(pc, method = "euclidean") # matriz distancias
fit <- hclust(d, method="average")
plot(fit)
CP$g <- cutree(fit, k=ng)
rect.hclust(fit, k=ng, border="red")

plot(CP$Comp.1, CP$Comp.2, xlab="Comp.1", ylab="Comp.2", main= "Método
average")
colores<- c("red", "blue", "magenta", "green", "black", "yellow", "orange",
"brown", "pink")
for (j in c(1:ng)) {
  points(CP$Comp.1[CP$g==j], CP$Comp.2[CP$g==j], col=colores[j], pch=16)
}

## Descripción por grupos
table(CP$g)
round(numSummary(CP[,c("Comp.1","Comp.2")], groups=CP$g)$table, 3)
round(numSummary(AO, groups=CP$g)$table, 3)

## CRITERIO SALTO MÍNIMO
ng <- 5
d <- dist(pc, method = "euclidean") # matriz distancias
fit <- hclust(d, method="single")
plot(fit)
CP$g <- cutree(fit, k=ng)
rect.hclust(fit, k=ng, border="red")
```

```
plot(CP$Comp.1, CP$Comp.2, xlab="Comp.1", ylab="Comp.2", main= "Método
salto mínimo")
colores<- c("red", "blue", "magenta", "green", "black", "yellow", "orange",
"brown", "pink")
for (j in c(1:ng)) {
  points(CP$Comp.1[CP$g==j], CP$Comp.2[CP$g==j], col=colores[j], pch=16)
}

## CRITERIO DEL CENTROIDE
ng <- 5
d <- dist(pc, method = "euclidean") # matriz distancias
fit <- hclust(d, method="centroid")
plot(fit)
CP$g <- cutree(fit, k=ng)
rect.hclust(fit, k=ng, border="red")

plot(CP$Comp.1, CP$Comp.2, xlab="Comp.1", ylab="Comp.2", main= "Método
salto mínimo")
colores<- c("red", "blue", "magenta", "green", "black", "yellow", "orange",
"brown", "pink")
for (j in c(1:ng)) {
  points(CP$Comp.1[CP$g==j], CP$Comp.2[CP$g==j], col=colores[j], pch=16)
}

## CRITERIO DE WARD
ng <- 5
d <- dist(pc, method = "euclidean") # matriz distancias
fit <- hclust(d, method="ward.D")
plot(fit)
CP$g <- cutree(fit, k=ng)
rect.hclust(fit, k=ng, border="red")

plot(CP$Comp.1, CP$Comp.2, xlab="Comp.1", ylab="Comp.2", main= "Método
ward")
colores<- c("red", "blue", "magenta", "green", "black", "yellow", "orange",
"brown", "pink")
for (j in c(1:ng)) {
  points(CP$Comp.1[CP$g==j], CP$Comp.2[CP$g==j], col=colores[j], pch=16)
}
```