Asignatura: **Bioestadística**
Análisis estadístico de datos multivariantes

Juana María Vivo
Dpto. Estadística e Investigación Operativa

Índice

1. Introducción

El Análisis Multivariante incluye una amplia variedad de técnicas estadísticas usadas para analizar situaciones en las que se estudian diversas variables conjuntamente, así como las relaciones entre los distintos grupos de la población, o de las poblaciones.

Primer Grupo: Técnicas estadísticas multivariantes de ajuste de modelos estadísticos
Determina un modelo estadístico que ajuste a una colección de datos observada.

Segundo Grupo: Técnicas estadísticas multivariantes de reducción de datos
Transforma variables observadas en otras no observadas. En este grupo, se encuentran el análisis de componentes principales, escalado multidimensional y el análisis de correspondencias.

Tercer Grupo: Técnicas estadísticas multivariantes de clasificación
Establece clases o familias de clases, que permiten agrupar y ordenar los individuos que se pretende describir. Se clasifican en este grupo técnicas multivariantes tales como el análisis de conglomerados.

Notación y conceptos usuales

En términos generales, se considera n individuos observados sobre m variables $X_1, X_2, ..., X_m$ y pueden presentarse como una matriz X de tamaño $n \times m$, constituida por n individuos (filas) y m variables (columnas), que se denomina *matriz de datos*

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ x_{31} & x_{32} & x_{33} & ... & x_{3m} \\ ... & ... & ... & ... & ... \\ x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{pmatrix}$$

donde x_{ij} es el valor del individuo (fila) i para la variable (columna) j.

En estadística multivariante, resulta habitual la noción de matriz de varianzas y covarianzas ($\Sigma \in M(m)$), usualmente llamada matriz de covarianzas. Se trata de una matriz simétrica. La varianza total de X es la suma de varianzas, es decir, la suma de los m elementos de la diagonal de la matriz de covarianzas. Cuando las variables originales son estandarizadas la matriz correspondiente se denomina matriz de correlaciones. Las correlación varían entre -1 y 1, indicando la dirección (directa o indirecta) y el grado de la relación entre las variables (más fuerte cuanto más próximo a 1 y a -1 y más débil cuanto más próximo a 0). En este caso, la varianza total es igual a m, i.e., igual al número de variables.

2. Componentes Principales

El análisis de componentes principales (ACP) es una técnica estadística de análisis multivariante de reducción de datos. Se aplica cuando se dispone de un número elevado de variables cuantitativas intercorrelacionadas y consiste en sustituir las m variables originales por k combinaciones lineales de las mismas no directamente observables, denominadas *componentes principales* o *factores*, que habrá que interpretar y nombrar. Obviamente se pretende que k sea menor que m, que expresen una proporción razonable de la dispersión o variación total (*inercia de la nube de puntos*) cuantificada como la traza de la matriz de covarianzas, $tr(\Sigma)$.

Esta técnica descriptiva de análisis de interdependencia proporciona componentes principales que están incorrelacionadas entre sí y pueden ordenarse de acuerdo a su varianza asociada, interpretada

en términos de información de la componente. Nótese que la suma de las varianzas de las componentes coincide con la suma de las variables originales.

En general, la extracción de componentes principales se lleva a cabo con variables tipificada, a veces con variables expresadas en desviaciones respecto a su media, para eliminar los efectos derivados de escala. Por otro lado, conviene destacar que el número de componentes que se obtiene coincide con el número de variables originales por lo que es importante abordar los métodos usuales para determinar k, i.e., el número de componentes principales que utilizaremos representando al conjunto de las variables originales.

2.1. Obtención de las componentes principales

En términos generales, se considera n individuos observados sobre m variables originales $X_1, X_2, ..., X_m$ e $Y_1, Y_2, ..., Y_m$ las variables incorrelacionadas que son combinaciones lineales de las originales:

$$Y_j = a_{j1}X_1 + a_{j2}X_2 + ... + a_{jm}X_m = a_j'X$$

para cada $j = 1, 2, ...m$; con $a_j' = (a_{j1}, a_{j2}, ..., a_{jm})$ es un vector con componentes constantes tal que $a_j'a_j = 1$ (condición de ortogonalidad). La primera componente se calcula tomando a_1 tal que Y_1 tenga la mayor varianza posible, sujeta a la restricción de que $a_1'a_1 = 1$,

$$Var(Y_1) = Var(a_1'X) = a_1'\Sigma a_1$$

a través del metodo de los multiplicadores de Lagrange: $Var(Y_1) = \lambda$, para maximizar la varianza de Y_1 se tiene que tomar el mayor autovalor, λ_1 y el correspondiente autovector a_1. La segunda componente principal se calcula obteniendo a_2 de modo que la variable obtenida, Y_2 esté incorrelada con Y_1, i.e, $Cov(Y_1, Y_2) = 0$ y razonando de manera similar, elegimos λ_2 como el segundo mayor autovalor de la matriz Σ con su autovector asociado a_2. Análogamente, se obtiene $Y_3, ..., Y_m$, en orden decreciente de varianza.

Por tanto, las m componentes Y se pueden expresar como el producto de una matriz formada por los autovectores, multiplicada por el vector X que contiene las variables originales $X_1, ..., X_m$. La matriz de covarianzas viene dada por:

$$\Lambda = diag(\lambda_1, ...\lambda_m) = Var(Y) = A'Var(X)A = A'\Sigma A$$

2.2. Porcentajes de variabilidad

Como anteriormente mencionamos, cada autovalor λ_i corresponde a la varianza del componente Y_i correspondiente al autovector a_i. Luego, la varianza total de las componentes, es la suma de todos los autovalores:

$$\sum_{i=1}^{m} Var(Y_i) = \sum_{i=1}^{m} \lambda_i = tr(\Lambda)$$

$$tr(\Lambda) = tr(A'\Sigma A) = tr(\Sigma A'A) = tr(\Sigma)$$

Por tanto, la suma de las varianzas de las variables originales y la suma de las varianzas de las componentes son iguales. Esto permite hablar del porcentaje de varianza total que recoge un componente principal:

$$\frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} = \frac{\lambda_i}{\sum_{i=1}^{m} Var(X_i)}$$

Asi, también se podrá expresar el porcentaje de variabilidad recogido por los primeros m componentes:

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i}$$

con $k < m$.

En general, si es posible, se suele tomar como máximo tres componentes principales para representarlas gráficamente.

Observación 2.1 *Si trabajamos con las variables originales estandarizadas, la extracción de las componentes principales se lleva a cabo desde la matriz de correlaciones. En la matriz de correlaciones todos los elementos de la diagonal son iguales a 1, por lo que la variabilidad total es igual al número total de variables (la suma total de todos los autovalores es m) y la proporción de varianza recogida por el autovector j -ésima componente es: λ_j/m.*

2.3. Caso práctico

Edgar Anderson's Iris Data This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. *iris* is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. *Available*: iris datasets.

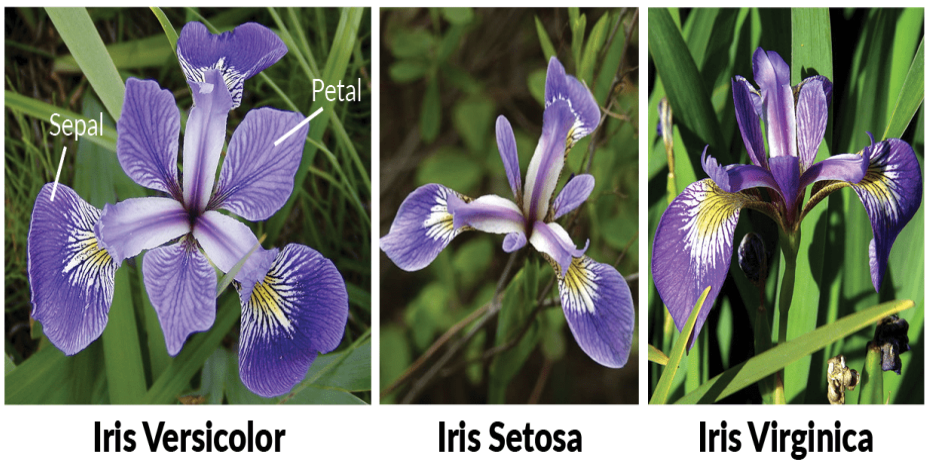


Figura 1: Especies del género Iris en *Fisher Iris data set.*

Edgar Anderson recopiló esta base de datos para cuantificar la variación morfológica de las flores de Iris de tres especies relacionadas. Basado en la combinación de las cuatro características medidas por Fisher vamos a llevar a cabo una análisis de componentes principales con el fichero iris, que teniendo en cuenta que las variables están expresadas en la misma unidad de medida no es necesario estandarizarlas.

Mediante el ACP identificaremos k variables no directamente observables a partir de las 4 medidas recogidas, denominadas componentes principales o factores, con $k < 4$, tal que:

- k sea un número pequeño.

- Se pierda la menor cantidad posible de información.
- La solución obtenida sea interpretable.

Los pasos serán los siguientes,

- Evaluación de la idoneidad del ACP.
- Extracción de los factores.
- Cálculo de las puntuaciones factoriales para cada caso.

En la evaluación de la idoneidad del ACP, es necesario que las variables presenten factores comunes. Es decir,que estén muy correlacionadas entre sí. Los coeficientes de la matriz de correlaciones deben ser grandes en valor absoluto.

Contrastar que las correlaciones entre las variables son distintas de cero de modo significativo mediante el *test de esfericidad de Barlett*, comprobando si el determinante de la matriz es distinto de uno, es decir, si la matriz de correlaciones es distinta de la matriz identidad.

Obsérvese que si las variables están correlacionadas la matriz de correlaciones presenta muchos valores altos en valor absoluto fuera de la diagonal principal, siendo el determinante siempre menor que 1, pues la unidad está asociada a la incorrelación de las variables y por tanto, a la no idoneidad de esta técnica. Concretamente, el test de Barlett realiza el contraste:

$$\begin{cases} H_0 & : & |R| = 1 \\ \\ H_1 & : & |R| \neq 1 \end{cases}$$

donde el determinante de la matriz da una idea de la correlación generalizada entre todas las variables. Esta prueba compara la matriz de correlación con la matriz identidad, en orden a comprobar si existen cierta redundancia entre las variables que podemos resumir en un número de factores. El estadístico de este test viene dado por:

$$\chi^2 = -(n - 1 - \frac{2m + 5}{6}) ln|R|$$

que bajo H_0 sigue una distribución $\chi^2_{\frac{m(m-1)}{2}}$, donde n es el tamaño muestral y m es el número de variables. Nótese que H_0 se rechaza para valores altos del estadístico, y que en caso de *no* rechazar H_0, el ACP es un procedimento estadístico no adecuado para nuestros datos. Por otro lado, cabe destacar que esta prueba está sesgada por el tamaño muestral.

Esta prueba se puede ejecutar cargando diversas librerias tales como *psych* y *REdaS*, aunque también resulta sencilla programar como una función que podemos usar en cualquier momento.

```
# Test de esfericidad de Bartlett
bartlett.sphere <- function(data.frame){
  n <- dim(data.frame)[1]
  m <- dim(data.frame)[2]
  R <- cor(data.frame,use='pairwise.complete.obs')
  chi2 <- -(n-1-(2*m+5)/6) * log(det(R))
  cat("\n","test de esfericidad de Barlett","\n","\n")
  cat('chi.square = ', round(chi2,3) ,
      ', df = ', (m^2-m)/2,
      ', p-value = ', pchisq(chi2,df=(m^2-m)/2,lower.tail=F))
```

```
}
bartlett.sphere(data.frame)
```

Por otro lado, la medida de la adecuación muestral de Kaiser-Meyer-Olkin (KMO) tiene el mismo objetivo que la prueba de esfericidad de Bartlett, pero basándose en que la correlación entre dos variables puede estar influenciada por otras restantes. Utilizando la correlación parcial para medir la relación entre dos variables eliminando el efecto de las variables restantes, el índice KMO compara los valores de las correlaciones entre las variables y los de las correlaciones parciales y viene dado por la siguiente expresión:

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

La matriz de correlación parcial $A = (a_{ij})$ se puede obtener a partir de la inversa de la matriz de correlación $R = (r_{ij})$, que denotaremos por $R^{-1} = (v_{ij})$ como sigue

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} \times v_{jj}}}$$

El índice KMO toma valores entre 0 y 1. A partir de su expresión se puede deducir que cuanto más pequeño sea el valor del KMO, mayor es el valor de los coeficientes de correlación parcialess a_{ij}, y por lo tanto, menos adecuado es realizar un ACP. En la literatura puede encontrarse tablas para la intepretación del valor del KMO obtenido para una base de datos. En particular, Kaiser, Meyer y Olkin aconsejan que si KMO$\geq$ 0.75 el ACP es muy adecuada, si $0.5 \leq KMO \leq 0.75$ es adecuada y si $KMO \leq 0.5$ es inadecuada.

KMO Value	Degree of Common Variance
0.90 to 1.00	Marvelous
0.80 to 0.89	Meritorious
0.70 to 0.79	Middling
0.60 to 0.69	Mediocre
0.50 to 0.59	Miserable
0.00 to 0.49	Don't Factor or unacceptable

Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin Test of Sampling Adequacy (KMO) are commonly used to provide more complex measures for assessing the strength of the relationships and suggesting factorability of the variables (Beavers et al., 2013). Kaiser (1974) recommends that the accepted index of KMO & Bartlett's Test of Sphericity should be over 0.5. Also, the Bartlett's Test of Sphericity relates to the significance of the study and thereby

Un código, no tan sencillo como el anterior, podría ser usado, así como, los correspodientes comandos incluidos en los R paquetes *psych* y *REdaS*.

```
kmo.test <- function( data.frame ){

  library(MASS)
  X <- cor(as.matrix(data.frame))
```

```
  iX <- ginv(X)
  S2 <- diag(diag((iX^-1)))
  AIS <- S2%*%iX%*%S2                      # anti-image covaritestance matrix
  IS <- X+AIS-2*S2                         # image covariance matrix
  Dai <- sqrt(diag(diag(AIS)))
  IR <- ginv(Dai)%*%IS%*%ginv(Dai)         # image correlation matrix
  AIR <- ginv(Dai)%*%AIS%*%ginv(Dai)       # anti-image correlation matrix
  a <- apply((AIR - diag(diag(AIR)))^2, 2, sum)
  AA <- sum(a)
  b <- apply((X - diag(nrow(X)))^2, 2, sum)
  BB <- sum(b)
  MSA <- b/(b+a)                        # indiv. measures of sampling adequacy

  AIR <- AIR-diag(nrow(AIR))+diag(MSA)  # Examine the anti-image of the
  # correlation matrix. That is the
  # negative of the partial correlations,
  # partialling out all other variables.

  kmo <- BB/(AA+BB)                # overall KMO statistic

  # Reporting the conclusion
  if (kmo >= 0.00 && kmo < 0.50){
    test <- 'The KMO test yields a degree of common variance
    unacceptable for FA.'
  } else if (kmo >= 0.50 && kmo < 0.60){
    test <- 'The KMO test yields a degree of common variance miserable.'
  } else if (kmo >= 0.60 && kmo < 0.70){
    test <- 'The KMO test yields a degree of common variance mediocre.'
  } else if (kmo >= 0.70 && kmo < 0.80){
    test <- 'The KMO test yields a degree of common variance middling.'
  } else if (kmo >= 0.80 && kmo < 0.90){
    test <- 'The KMO test yields a degree of common variance meritorious.'
  } else {
    test <- 'The KMO test yields a degree of common variance marvelous.'
  }

  ans <- list(  overall = kmo,
                report = test,
                individual = MSA,
                AIS = AIS,
                AIR = AIR )
  return(ans)

}

kmo.test(data.frame)
```

Obsérvese que estas medidas de adecuación muestral nos permiten detectar si podemos o no reducir la dimensionalidad, pero no dan indicaciones sobre el número apropiado de factores.

Existe diferentes opciones para determinar el número de componentes principales o factores que se deben de retener:

El gráfico de la varianza asociada a cada factor que refleja la ruptura entre la pronunciada pendiente de los factores más importantes y el descenso gradual de los restantes (los sedimentos)

El criterio de Kaiser que consiste en conservar aquellos factores cuyo autovalor asociado sea mayor que 1.

La comunalidad asociada a la variable j-ésima es la proporción de variabilidad de dicha variable explicada por los k factores retenidos. Equivale a la suma de la fila j-ésima de la matriz factorial, sería igual a 0 si los factores comunes no explicarán nada la variabilidad de una variable, y sería igual a 1 si quedase totalmente explicada.

```
> #Leo los datos
> data(iris)
> #Hago un análisis descriptivo de los datos
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> iris.new=iris[,1:4]
> attach(iris.new)

data(iris)
summary(iris)
   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
  Median :5.800   Median :3.000   Median :4.350   Median :1.300
  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
        Species
  setosa    :50
  versicolor:50
  virginica :50

> #Calculo de la matriz de correlaciones
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000

> #Realizo las pruebas de idoneidad de la técnica (MSA)
```

```
> bartlett.sphere(iris[,1:4])

 test de esfericidad de Barlett

chi.square =  706.959 , df =  6 , p-value =  1.92268e-149
> kmo.test(iris[,1:4])
$overall
[1] 0.5400767

$report
[1] "The KMO test yields a degree of common variance miserable."

$individual
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.5840603    0.2695746    0.5307484    0.6342065

$AIS
            [,1]        [,2]        [,3]        [,4]
[1,]  0.14138828 -0.16306512 -0.04835822  0.03183583
[2,] -0.16306512  0.47599290  0.07592356 -0.06065047
[3,] -0.04835822  0.07592356  0.03198823 -0.03882559
[4,]  0.03183583 -0.06065047 -0.03882559  0.06214973

$AIR
           [,1]       [,2]       [,3]       [,4]
[1,]  0.5840603 -0.6285707 -0.7190656  0.3396174
[2,] -0.6285707  0.2695746  0.6152919 -0.3526260
[3,] -0.7190656  0.6152919  0.5307484 -0.8707698
[4,]  0.3396174 -0.3526260 -0.8707698  0.6342065

> #install.packages("psych")
> library("psych")
> R=cor(iris[,1:4])
> cortest.bartlett(R, n = 150)
$chisq
[1] 706.9592

$p.value
[1] 1.92268e-149

$df
[1] 6

> KMO(iris[,1:4])
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = iris[, 1:4])
Overall MSA =  0.54
MSA for each item =
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
        0.58         0.27         0.53         0.63
> #install.packages("REdaS")
> library("REdaS")
> bart_spher(iris[,1:4])

Bartlett's Test of Sphericity

Call: bart_spher(x = iris[, 1:4])

     X2 = 706.959
     df = 6
p-value < 2.22e-16

> KMOS(iris[,1:4])

Kaiser-Meyer-Olkin Statistics

Call: KMOS(x = iris[, 1:4])

Measures of Sampling Adequacy (MSA):
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.5840603    0.2695746    0.5307484    0.6342065

KMO-Criterion: 0.5400767
```

A la vista de la matriz de correlaciones y del p-valor asociado al estadístico del contraste de esfericidad de Bartlett, tiene sentido llevar a cabo un análisis de componentes principales en el fichero de datos iris.

Por tanto, vamos describir el conjunto de datos *iris* a través del ACP, reduciendo la dimensionalidad y responder a varias preguntas:

En relación a los individuos ¿Hay similitudes entre los iris para todas las variables? ¿Podemos establecer diferentes perfiles de iris? ¿Podemos oponer a un grupo de iris a otro?

En relación a las variables ¿Podemos resumir las características por un pequeño número de variables?

En relación a ambos ¿Podemos caracterizar grupos de individuos por variables?

```
> iris.pca=PCA(iris[,1:4], scale.unit=T, ncp=5, graph=T)
#iris: base de datos utilizados
#scale.unit: para elegir si se debe escalar los datos o no
#ncp: número de dimensiones consideradas en el resultado
#graph: para trazar los gráficos o no


>barplot(iris.pca$eig[,1],main="Eigenvalues",names.arg=1:nrow(iris.pca$eig))
```

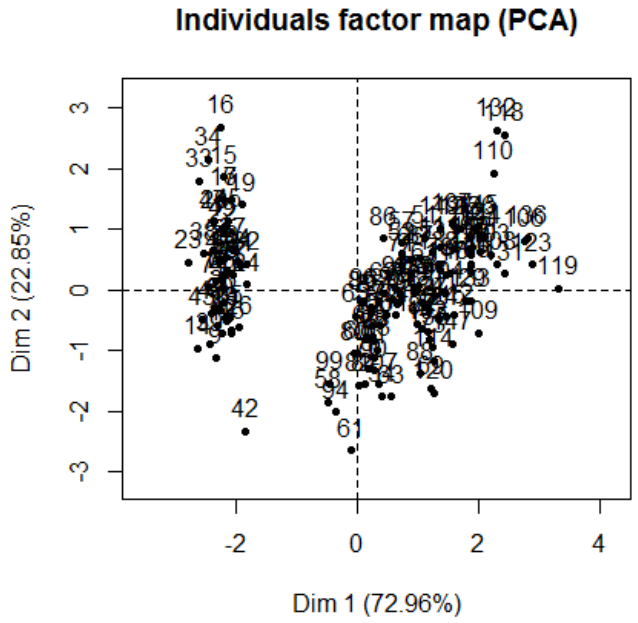


Figura 2: Puntuaciones factoriales

```
> summary.PCA(iris.pca)

 Call:

 Eigenvalues
                        Dim.1   Dim.2   Dim.3   Dim.4
 Variance               2.918   0.914   0.147   0.021
 % of var.             72.962  22.851   3.669   0.518
 Cumulative % of var.  72.962  95.813  99.482 100.000
```

Los resultados obtenidos nos indican que retener las dos primeras componentes principales (95.81 % de la inercia total (varianza total) lo que supone una pérdida de información del 4.29 %. En particular, el primer factor explica el 72.96 % de la varianza total y el segundo el 22.85 %.

```
 Individuals (the 10 first)
                   Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
 1              |  2.319 | -2.265  1.172  0.954 |  0.480  0.168  0.043 |
 2              |  2.202 | -2.081  0.989  0.893 | -0.674  0.331  0.094 |
 3              |  2.389 | -2.364  1.277  0.979 | -0.342  0.085  0.020 |
 4              |  2.378 | -2.299  1.208  0.935 | -0.597  0.260  0.063 |
 5              |  2.476 | -2.390  1.305  0.932 |  0.647  0.305  0.068 |
 6              |  2.555 | -2.076  0.984  0.660 |  1.489  1.617  0.340 |
 7              |  2.468 | -2.444  1.364  0.981 |  0.048  0.002  0.000 |
 8              |  2.246 | -2.233  1.139  0.988 |  0.223  0.036  0.010 |
 9              |  2.592 | -2.335  1.245  0.812 | -1.115  0.907  0.185 |
 10             |  2.249 | -2.184  1.090  0.943 | -0.469  0.160  0.043 |
                Dim.3    ctr   cos2
```

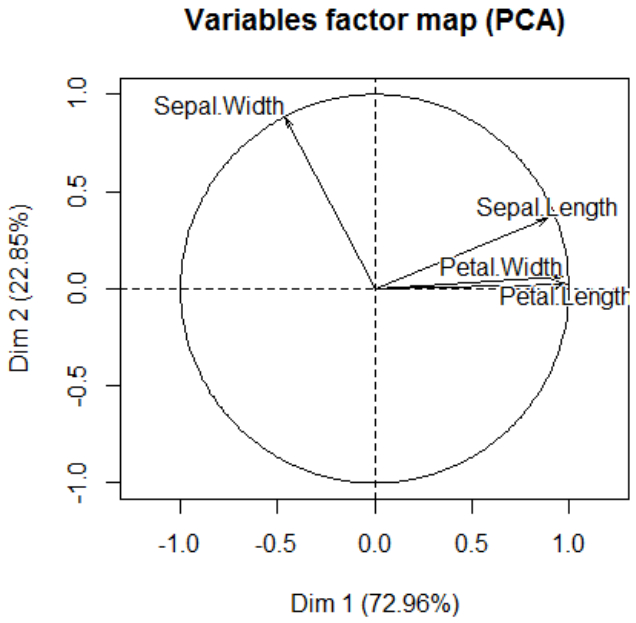


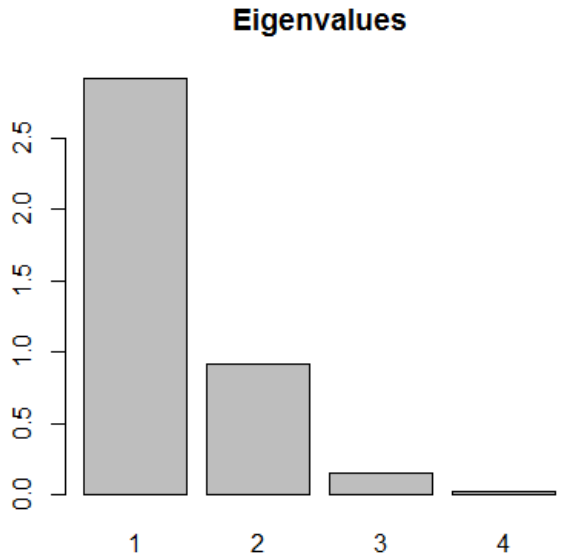
Figura 3: Cargas factoriales

```
1            -0.128  0.074  0.003 |
2            -0.235  0.250  0.011 |
3             0.044  0.009  0.000 |
4             0.091  0.038  0.001 |
5             0.016  0.001  0.000 |
6             0.027  0.003  0.000 |
7             0.335  0.511  0.018 |
8            -0.089  0.036  0.002 |
9             0.145  0.096  0.003 |
10           -0.254  0.293  0.013 |
```

El cálculo de las puntuaciones factoriales consiste en pasar de la matriz original con las variables $X_1, ..., X_m$ a la de los valores según los k factores. Obsérvese que estas puntuaciones factoriales se pueden guardar y utilizar en análisis posteriores como técnicas de regresión múltiple o en análisis de cluster.

```
Variables
                Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
Sepal.Length |  0.890 27.151  0.792 |  0.361 14.244  0.130 | -0.276 51.778
Sepal.Width  | -0.460  7.255  0.212 |  0.883 85.247  0.779 |  0.094  5.972
Petal.Length |  0.992 33.688  0.983 |  0.023  0.060  0.001 |  0.054  2.020
Petal.Width  |  0.965 31.906  0.931 |  0.064  0.448  0.004 |  0.243 40.230
                cos2
Sepal.Length  0.076 |
Sepal.Width   0.009 |
Petal.Length  0.003 |
Petal.Width   0.059 |
```

A continuación teniendo en cuenta las cargas factoriales para las dos primeras componentes podemos concluir que el primer factor explica la longitud de pétalo y sépalo y la amplitud de pétalo;



mientras que el segundo factor la amplitud de sépalo, ver Figura (3). Por lo que podemos dividir la nube de puntos en cuatro partes ver Figura (4):

Primer cuadrante alta longitud y amplitud de pétalo y sépalo,

Segundo cuadrante baja longitud de pétalo y sépalo así como baja amplitud de pétalo pero alta de sépalo,

Tercer cuadrante baja longitud y amplitud de pétalo y sépalo y

Cuarto cuadrante alta longitud de pétalo y sépalo así como alta amplitud de pétalo pero baja de sépalo.

```
> biplot(iris.pca$ind$coord[,1:2],iris.pca$var$coord[,1:2],xlim=c(-4,4))
```

La función *dimdesc* permite describir las componentes retenidas, calculando el coeficiente de correlación entre una variable y una dimensión y lleva a cabo una prueba de significación. Estas tablas dan el coeficiente de correlación y el valor p de las variables que se correlacionan de manera significativa a las principales dimensiones.

```
> dimdesc(iris.pca, axes=c(1,2))

# res.pca: el resultado de un PCA
#axes: los ejes elegidos

$Dim.1
$Dim.1$quanti
              correlation       p.value
Petal.Length    0.9915552 3.369916e-133
Petal.Width     0.9649790  6.609632e-88
```

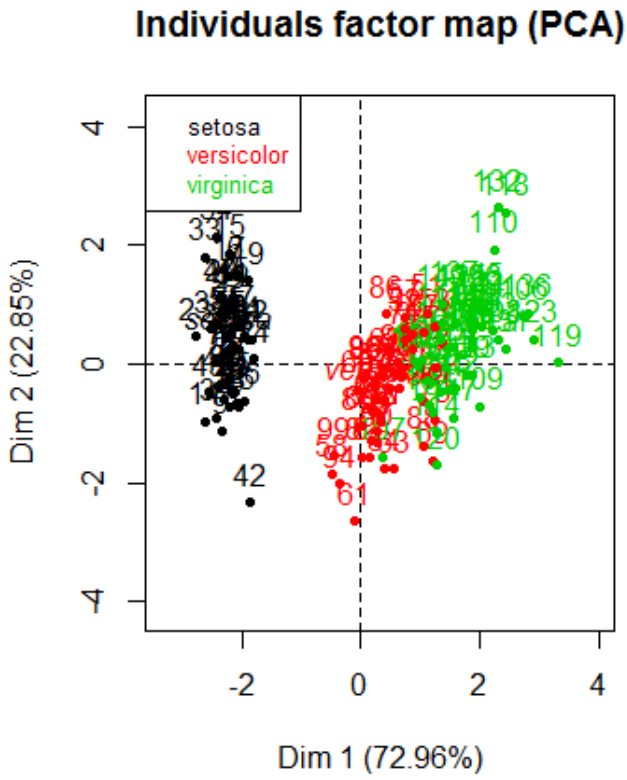


Figura 4: Puntuaciones factoriales por Especie

```
Sepal.Length   0.8901688  2.190813e-52
Sepal.Width   -0.4601427  3.139724e-09


$Dim.2
$Dim.2$quanti
             correlation      p.value
Sepal.Width    0.8827163 2.123801e-50
Sepal.Length   0.3608299 5.731933e-06
```

En nuestro caso particular, las variables que presentan un p-valor menor que 0.05 aparecen en las tablas que muestran que las variables "Petal.Lenght", "Petal.Width" y "Sepal.Lengh" son las más correlacionada con la primera dimensión y "Sepal.Width" es la más correlacionada con la segunda. Esto confirma nuestra primera interpretación.

Mayor información de los resultados obtenidos puede encontrarse con el comando *names* del siguiente modo:

```
> names(iris.pca)
[1] "eig"  "var"  "ind"  "svd"  "call"

#eig: matriz que contiene todos los valores propios, el porcentaje de varianza
#y el porcentaje acumulado de varianza

#var: matrices que contienen todos los resultados para las variables activas
```

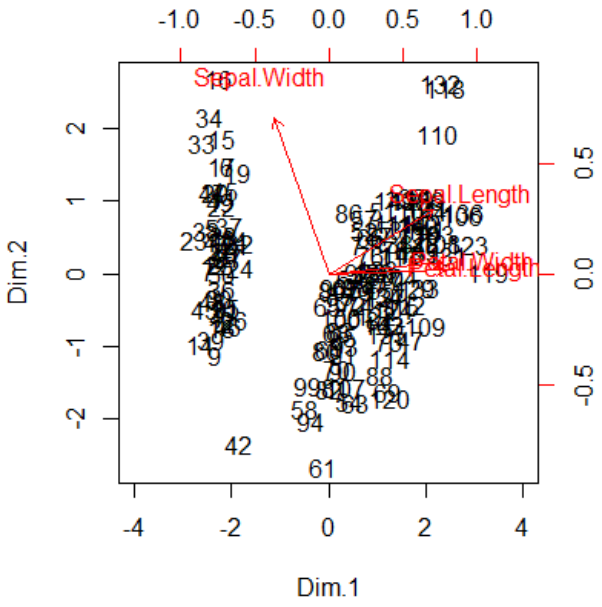


Figura 5: biplot para el ACP con iris

```
#(coordenadas, la correlación entre las variables y ejes, coseno cuadrado,
#contribuciones)

#ind: matrices que contienen todos los resultados para los individuos activos
#(coordenadas, coseno cuadrado, contribuciones)

#ind.sup: matrices que contienen todos los resultados para los individuos
#suplementarias (coordenadas, coseno cuadrado)

#quanti.sup: matrices que contienen todos los resultados para las variables
#cuantitativas suplementarias (coordenadas, la correlación entre las variables
#y los ejes)

#quali.sup: matrices que contienen todos los resultados para las variables
#categóricas suplementarios (coordenadas de cada categoría de cada variable,
#v.test que es un criterio con una distribución normal, y ETA2 que es el
#coeficiente de correlación cuadrado entre una variable cualitativa y una
#dimensión)
```

Ejercicio 2.1 *Obtener la salida de las siguientes sentencias. Comentar los resultados.*

```
> iris.pca=PCA(iris, scale.unit=T, ncp=5, graph=T,quali.sup=5)
> plot.PCA(iris.pca,axes=c(1,2), choix="ind", habillage=5)
```

O también, mediante:

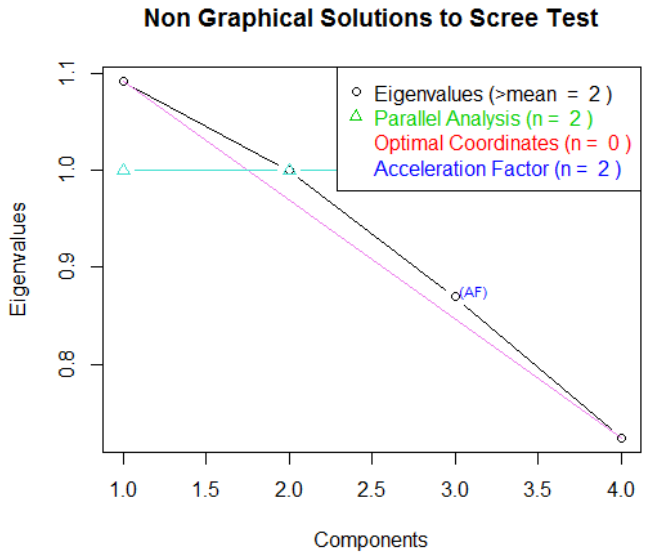
```
> iris.pca
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 150 individuals, described by 4 variables
```

```
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "summary statistics"
12 "$call$centre"     "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"      "weights for the individuals"
15 "$call$col.w"      "weights for the variables"
```

Por último, para determinar el número óptimo de factores para extraer usaremos el comando *parallel* del R package *nFactors*. Esta función proporciona la distribución de los valores propios de la matriz de correlación/covarianza de las variables aleatorias normales no correlacionadas. Se devuelven la media y un cuantil seleccionado de esta distribución.

```
#Determinación del número de factores para extraer
> install.packages("nFactors")
> library(nFactors)
> ev <- eigen(cor(iris[,1:4]))
> ap <- parallel(subject=nrow(iris[,1:4]),var=ncol(iris[,1:4]),rep=100,cent=.05)
> nS <- nScree(ev$values,ap$eigen$qevpea)
> plotnScree(nS)
```



donde *eigen$qevpea* es el cuantil de la distribución de valores propios.

3. Análisis de conglomerados

En este tema abordaremos algunas técnicas de análisis multivariante muy conocidas y aplicadas, denominadas globalmente como análisis cluster.

El análisis cluster incluye diferentes algoritmos de clasificación que se encargan de establecer clases o familias de clases, que permiten agrupar y ordenar los individuos que se pretende describir. Estas clases se denominan clusters o conglomerados y la agrupación se basa en la similaridad o disimilaridad de las individuos. El término fue utilizado por primera vez por Tryon en 1939.

El análisis cluster se utiliza en diversas disciplinas científicas (Hartigan, 1975), especialmente aquellas que requieren desarrollar esquemas de clasificación tales como una taxonomía para un conjunto de objetos, a sugerir modelos estadísticos para describir poblaciones, a asignar nuevos individuos a las clases para diagnóstico e identificación, etc... que concretaremos en aplicaciones usuales más adelante para una visión más clara. Aún con pocas observaciones, el número posible de combinaciones de grupos y de individuos que integran los posibles grupos se hace inmanejable desde el punto de vista computacional. Por lo que se hace necesario, pues, encontrar métodos o algoritmos que determinen el número y componentes de los clusters más aceptable, aunque no sea el óptimo absoluto. En este sentido, para adaptarse a diferentes tipos de enfoques, se han desarrollado distintos procedimientos. Cabe señalar que a diferencia de otros métodos estadísticos, los procedimientos de clustering no tienen hipótesis a priori, ya que se localizan en la fase exploratoria de un estudio estadístico, no siendo apropiados los contrastes de hipótesis estadística.

Como mencionamos anteriormente, se considera n individuos observados sobre m variables $X_1, X_2, ..., X_m$, que se puede representar como una matriz X de tamaño $n \times m$, constituida por n individuos (filas) y m variables (columnas), que se denomina *matriz de datos*.

En el desarrollo del presente capítulo, veremos que algunos de los métodos trabajan directamente sobre la matriz de datos y otros sobre una matriz derivada de ella, una matriz de similaridades/distancias de orden n, que determina la similitud o distancia entre cada par de individuos en términos de proximidad o lejanía.

Se distiguen en:

- **Métodos no jerárquicos**, que clasifican a los individuos en k conglomerados no anidados, estudiando todas las particiones de individuos en esos k grupos y eligiendo la mejor partición. Cluster de k-medias
- **Métodos jeráquicos**, que clasifican de los individuos en conglomerados anidados (árbol de clasificación) y se dividen en:
  - (i.) ***aglomerativos o asociativos***, comienza considerando n conglomerados, hasta obtener un único conglomerado.
  - (ii.) ***divisivos o disociativos***, comienza considerando un único conglomerado, hasta obtener n conglomerados.

A partir de la adecuada selección de las variables relevantes para identificar los clusters, cada uno de los individuos sujetos al análisis vendrán representados por los valores de la medida escogida de similaridad (o disimilaridad) entre ellos y la elección del criterio para agruparlos en cluster o conglomerados. Este es el punto de partida de la clasificación, que analizaremos detalladamente a lo largo de las sesiones dedicadas a este tema.

En general, las variables relevantes suelen ser del mismo tipo, cuantitativo y si tenemos variables cualitativas se recodifican en numéricas. Por otro lado, es evidente que sobre cualquier individuo es posible considerar un gran número de variables, pero la inclusión de variables irrelevantes puede aumentar la posibilidad de errores en los resultados. Por ello se deben eliminar las variables irrelevantes

en base al objetivo de la investigación. El análisis de componentes principales es una técnica que nos permite resolver este problema evitando la correlación entre las variables del análisis (redundancia).

Previamente al análisis, se deben estandarizar las variables cuantitativas, media 0 y desviación típica 1, cuando las variables están medidas en diferentes unidades o escalas, la comparación entre unas variables u otras será difícil, diferentes medidas entonces diferentes ponderaciones. Las variables binarias no suelen transformarse y las variables cualitativas se recodifican en binarias (presencia/ausencia) para las diferentes modalidades.

En cualquier caso, cabe destacar que los cluster obtenidos dependerán de la medida y el criterio seleccionados y no existe una técnica fiable para determinar cuál es el óptimo. Otro problema que adelantamos es la decisión del número de conglomerados, ya que no existe un procedimiento que lo determine. En su ausencia estudiaremos las distancias a las que se van uniendo los conglomerado y parar cuando la distancia llegue a un valor determinado. Generalmente se estudia la solución, y nos decantaremos por un número de cluster interesante a nuestro análisis.

Aunque generalmente se pretende agrupar individuos, Q-técnicas, existen situaciones en las que es interesante agrupar variables con el objetivo de encontrar variables de comportamiento similar, lo que simplemente supone transponer la matriz de datos y aplicar el método general. Este tipo de análisis se denominan R-técnicas.

El análisis cluster es habitualmente aplicado en psicología, biología, sociología, economía, ingeniería y administración y dirección de empresas, etc. Además, puede incluso recibir diversos nombres según la rama científica en la que se aplique, por ejemplo*análisis Q*, *taxonomía numérica* entre otros.

En general, se trata de una herramienta de reducción de datos mediante la clasificación en grupos manejables, útil en situaciones tales como la que se puede presentar tras la recogida de datos de un cuestionario, un número elevado de observaciones. También clasificación psicológica o rasgos personales o a la segmentación del mercado.

3.1. Medidas de similaridad

Para clasificar adecuadamente los individuos debemos determinar la similaridad o disimilaridad (o divergencia) entre ellos, en función de lo similares o no que resulten ser sus representaciones en el espacio de las variables. Por tanto, resulta necesario disponer de medidas numéricas. En este sentido, se cuenta con diversos índices de similaridad y de disimilaridad, con propiedades y utilidades distintas que no podemos pasar alto para su correcta aplicación. Estas medidas de asociación se expresan en términos de una distancia o una similaridad. Cuando se elige una distancia como medida de asociación (por ejemplo la distancia euclídea) cada conglomerado contendrán individuos cuya distancia entre ellos ha de ser pequeña. Cuando se elige una medida de similaridad (por ejemplo el coeficiente de correlación) los grupos formados contendrán individuos con una similaridad alta entre ellos.

Nótese que si se agrupan variables, los indicadores están basados en coeficientes de correlación en valor absoluto, o bien en criterios de posesión o no de una serie de atributos (tablas de presencia-ausencia) en caso de variables cualitativas.

Dados dos vectores $x_i, x_j \in \mathbb{R}^k$, diremos que hemos establecido una *distancia* entre ellos si definimos una función que verifique los siguientes axiomas:

- $d : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$, i.e., $d(x_i, x_j) \geq 0$. (no negativa)
- $d(x_i, x_i) = 0, \forall i$
- $d(x_i, x_j) = d(x_j, x_i)$ (simetría)
- $d(x_i, x_j) \geq d(x_i, x_p) + d(x_s, x_p)$ (propiedad triangular)

Dados dos vectores $x_i, x_j \in \mathbb{R}^k$, diremos que hemos establecido una *similaridad* entre ellos si definimos una función que verifique los siguientes axiomas:

- $s : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$, i.e., $s(x_i, x_j) \leq s_0$.
- $s(x_i, x_i) = s_0, \forall i$
- $s(x_i, x_j) = s(x_j, x_i)$

Una función s, verificando las condiciones de la definición anterior, se llama *similaridad métrica* si, además, verifica:

- $s(x_i, x_j) = s_0 \Rightarrow x_i = x_j$
- $|s(x_i, x_p) + s(x_p, x_j)|s(x_i, x_j) \geq s(x_i, x_p)s(x_p, x_j), \forall z \in \mathbb{R}^k$

3.1.1. Ejemplos de distancias entre individuos

En análisis cluster se pueden utilizar diferentes tipos de medidas de distancias para formar los conglomerados, basadas en una o varias dimensiones. En general, la distancia de geométrica entre dos conglomerados es la manera más directa de medir la distancia entre dos conglomerados en un espacio multi-dimensional. Sin embargo, cabe señalar que al algoritmo de agrupación no le "importa" si las distancias en que se basan son distancias reales o en alguna otra medida derivada de la distancia que es más significativa para la investigación, y depende de la investigación para seleccionar el método adecuado para su aplicación específica.

Distancia euclídea

La distancia euclídea es el tipo de distancia más usual. Se obtiene como:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Distancia euclídea al cuadrado

Si estamos interesados en colocar progresivamente mayor peso a los objetos que están más separados, elevando al cuadrado la distancia euclídea estándar:

$$d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$$

Nótese que la distancia euclídea y la distancia euclídea al cuadrado, se calculan a partir de los datos no estandarizados, lo que puede afectar por las diferencias de escala entre las dimensiones.

Distancia city-block (Manhattan)

Esta distancia es la diferencia media entre las dimensiones. En la mayoría de los casos, es medida de distancia produce resultados similares a la distancia euclídea. Sin embargo, nótese que en esta medida, el efecto de una única diferencia grande es reducido. Se obtiene como:

$$d_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

Distancia Chebychev

Esta medida de distancia puede ser apropiada cuando se quiere definir dos conglomerados como "diferentes" si son diferentes en una de las dimensiones y se calcula como:

$$d_{ij} = max|x_{ik} - x_{jk}|$$

Distancia potencia

En las situaciones en la que se requiere aumentar o diminuir el peso progresivo que se coloca en dimensiones en las que los respectivos conglomerados son muy diferentes, se puede utilizar la distancia potencia que puede calcularse como:

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^p \right)^{1/r}$$

siendo r y p paramétros definidos por el usuario. En particular, el primero parámetro controla el peso progresivo que tiene lugar en las diferencias grandes entre conglomerados y el segundo controla el peso progresivo que se da en diferencias sobre dimensiones individuales. Si $r = p$ obtendríamos la distancia de Minkowski. Si $r = p = 2$ obtenemos la distancia euclídea.

Distancia de Mahalanobis

Viene dada por

$$d_{ij} = (x_i - x_j)^t W^{-1} (x_i - x_j)$$

donde W es la matriz de covarianzas entre las variables. Nótese que si la correlación es nula y las variables están estandarizadas, se obtiene la distancia euclídea.

Porcentaje de desacuerdo

Esta medida es útil cuando los datos en el análisis son cualitativos y se computa como:

$$d_{ij} = \frac{(N^o x_i \neq x_j)}{i}$$

3.1.2. Ejemplos de distancias entre variables

Coeficiente de correlación de Pearson

Se define como

$$r = \frac{S_{XY}}{S_X S_Y}$$

donde S_{XY} es la covarianza entre X y Y y S_X y S_Y son las desviaciones estándar de X e Y respectivamente.

Coeficiente de correlación de rangos de Kendall

Esta medida de asociación para datos ordinales mide el grado de correspondencia que existe entre los rangos que se asignan a los valores de las variables analizadas. Este coeficiente se calculan:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n_2 - 1)}$$

siendo d_i la diferencia entre los rangos correspondientes a la observación i-ésima. El coeficiente toma valores entre -1 y +1, en caso de que tome un valor cercano a 0 las variables no están significativamente relacionadas.

3.2. Métodos jerárquicos

Los procedimientos jerárquicos aglomerativos son los más usuales. Organizan los clusters jerárquicamente mostrando relaciones y estructuras entre los datos que a priori no son evidentes. Además, los resultados se representan gráficamente mediante árboles que resumen el proceso de agrupación, denominados *dendogramas*, y que determinan el grado de similaridad o disimiliaridad entre los individuos por la posición de su enlace en el mismo.

Ejemplo 3.1 *Consideremos 5 individuos y dos variables:*

Individuo	*Variable1*	*Variable2*
1	*1*	*1*
2	*2*	*1*
3	*4*	*5*
4	*7*	*7*
5	*5*	*7*

que representamos en un plano:

```
> plot(Variable1,Variable2,type="n")
> text(Variable1,Variable2,labels=Individuo,cex=1)
```

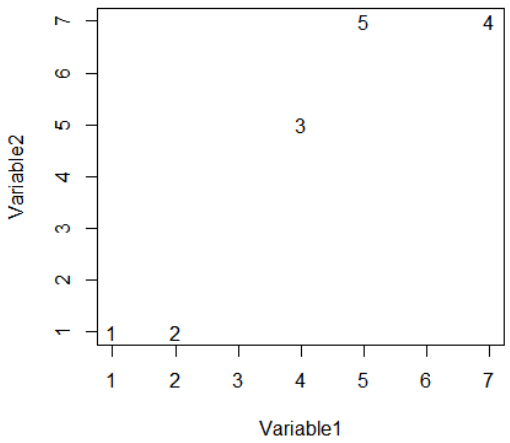


Figura 6: Representación de Individuos en el plano

Considerando su matriz de distancias euclídeas entre los individuos:

Tabla 1: Matriz de distancias euclídeas.

0	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

Paso 0 Cinco conglomerados iniciales: 1, 2, 3, 4 y 5 (coordenadas y distancias).

Paso 1 Conglomerados 1 y 2 más similares (matriz de distancias), se combinan en un nuevo cluster (C1). Y medimos las distancias del C1 de centroide (1.5,1) a los tres clusters restantes, a partir de las nuevas coordenadas.

Individuo	Variable1	Variable2
C1	1.5	1
3	4	5
4	7	7
5	5	7

$$\begin{pmatrix} * & C1 & 3 & 4 & 5 \\ C1 & 0.0 & * & * & * \\ 3 & 4.7 & 0.0 & * & * \\ 4 & 8.1 & 3.6 & 0.0 & * \\ 5 & 6.9 & 2.2 & 2.0 & 0.0 \end{pmatrix}$$

Paso 2 Clusters 4 y 5 más similares se combinan en un nuevo cluster (C2) de centroide (6,7). Recalcular las distancias.

Individuo	Variable1	Variable2
C1	1.5	1
3	4	5
C2	6	7

$$\begin{pmatrix} * & C1 & 3 & C2 \\ C1 & 0.0 & * & * \\ 3 & 4.7 & 0.0 & * \\ C2 & 7.5 & 2.8 & 0.0 \end{pmatrix}$$

Paso 3 Clusters 3 y C2 más similares se combinan en un nuevo cluster (C3) de centroide (5.3,6.3). Recalcular las distancias.

Individuo	Variable1	Variable2
C1	1.5	1
C3	5.3	6.3

$$\begin{pmatrix} * & C1 & C3 \\ C1 & 0.0 & * \\ C3 & 6.5 & 0.0 \end{pmatrix}$$

Paso 4 Sólo clusters C1 y C3 se combinan en un nuevo cluster (C4) de centroide (3.4,3.65). Termina el proceso.

El proceso completo, i.e, todos los pasos se puede visualizar a través del dendograma:

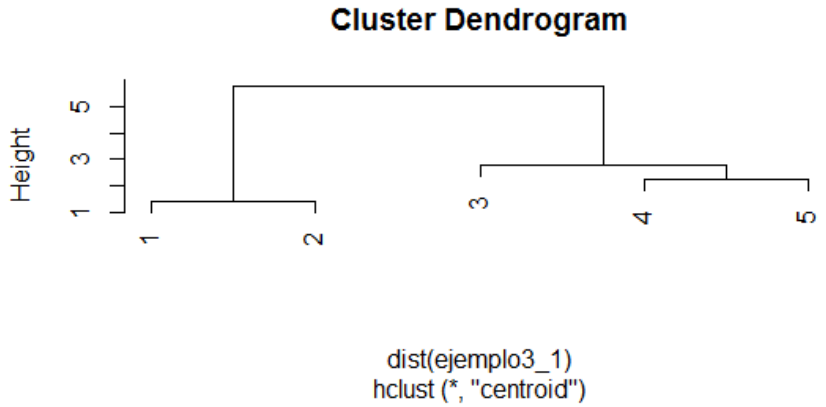


Figura 7: Dendograma Ejemplo 3.1, método del centroide

Ejercicio 3.1 *A partir de la base de datos anterior, realizar el mismo proceso anterior para obtener la clasificación de los individuos considerando diferentes métodos ("vecino más próximo", "vecino más lejano",...).*

Mediante líneas horizontales determinamos el número de conglomerados, por lo que el óptimo es subjetivo. Además, si seleccionamos un número muy bajo corremos el riesgo de que los clusters obtenidos sean demasiado hetereogéneos y si es demasiado elevado entonces resulte complicado interpretar. Nótese las distancias de las combinaciones de los clusters son menores en los primeros pasos y en el último, se registrará la mayor. Por lo que, resulta aconsejable sobre un gráfico de la evolución de las combinaciones de los conglomerados observar donde se produce el salto más brusco, conocida como la regla del codazo. En nuestro ejemplo, observamos que el salto más brusco se produce entre los pasos 3 y 4, luego el número óptimo de conglomerados es 2 (paso 3).

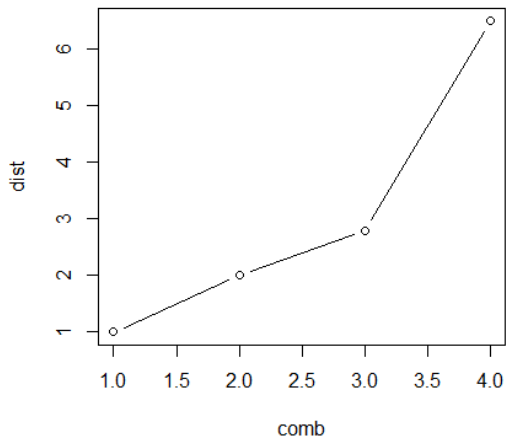


Figura 8: Evolución de la clasificación, regla del codazo

3.2.1. Agrupamiento aglomerativo

El término aglomerativo indica que el dendograma comienza separando cada individuo en un cluster por sí mismo. En cada paso sucesivo, relajando el criterio de agrupación, se combinan los dos (o más) conglomerados más similares hasta que todos los individuos están en un árbol de clasificación completa en el último paso. El criterio de agrupación se basa en la distancia, de manera que los individuos más cercanos entre sí pertenecerían al mismo conglomerado o cluster, y los más alejados entre sí pertenecerían a distintos clusters.

A partir de una base de datos, los clusters que se construyen dependen de nuestra propia especificación de los siguientes parámetros:

- El método cluster define las reglas para la formación del cluster. Por ejemplo, cuando calculamos la distancia entre dos clusters, podemos usar el par de objetos más cercano entre clusters o el par de objeto más alejados, o un compromiso entre estos métodos.

- La medida define la fórmula para el cálculo de la distancia. Por ejemplo, la medida de distancia euclídea calcula la distancia como una línea recta entre dos clusters. Las medidas de intervalo asumen que las variables están medidas en escala; las medidas de conteo asumen que son números discretos, y las medidas binarias asumen que toman dos valores.

- La estandarización permite igualar el efecto de las variables medidas sobre diferentes escalas.

El resultado es un dendograma que muestra los casos más similares vinculados más estrechamente. El nivel de las líneas verticales que unen dos casos o conglomerados indica el nivel de similitud entre

ellos. Es importante señalar que la jerarquía de ramificación y el nivel de similitud son las únicas características importantes del dendograma. El orden exacto de los casos a lo largo del eje vertical no es significativa.

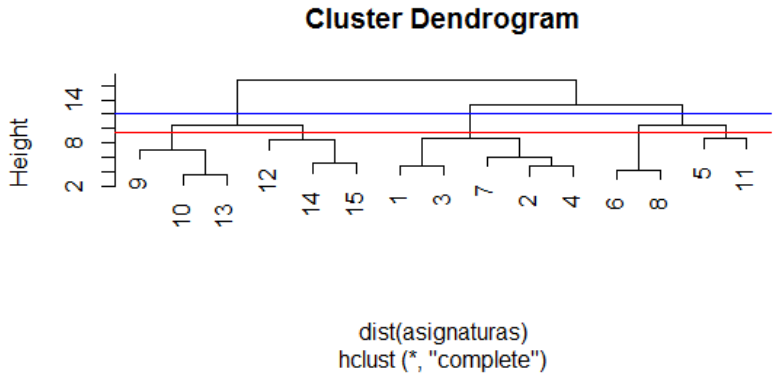


Figura 9: Dendograma asignaturas, método completo

Obsérvese que el dendrograma muestra la formación de los conglomerados, así como las distancias entre ellos. Se puede comprobar, por ejemplo, que las observaciones más distantes al resto es las de los alumnos número 5, número 11 y número 13, ya que son las últimas (mayor distancia) en incorporarse. Por el contrario, las observaciones más cercanas entre sí son la 10 y la 13, que forman el primer grupo (distancia más próxima a 0), y el 6 y el 8, que forman el segundo.

También, el dendrograma nos presenta la composición de cada cluster en cada paso. Así, si quisiéramos hacer una división en 5 conglomerados bastaría con trazar la línea roja, con la siguiente distribución:

Conglomerado	Alumnos
1	9, 10, 13
2	12, 14, 15
3	1, 2, 3, 4, 7
4	5, 6, 8, 11

Ejercicio 3.2 *Responder a las siguientes apartados, teniendo en cuenta la Figura 9.*

(1) Proporcionar la distribución correspondiente si trazamos la línea azul.

(2) Razonar el número apropiado de conglomerados que deberíamos considerar.

3.2.2. Amalgamiento o Reglas de Agrupación

Las reglas de amalgamiento o de agrupación son necesarias para determinar cuando dos conglomerados son suficientemente similares para ser combinados. Existen diferentes reglas o criterios de agrupación y cada uno puede producir distintas clasificaciones, y por tanto, no existe una única clasificación correcta. De partida

$$d_{i,j+k} = \delta_1 d(i,j) + \delta_2 d(i,k) + \delta_3 d(j,k) + \delta_4 |d(i,j) - d(i,k)|$$

donde δ_i son ponderaciones que varían según el método utilizado. Véase Tabla 2, siendo n_i, n_j y n_k el número de objetos en cada uno de los grupos y $0 \leq \lambda \leq 1$.

Método del vecino más cercano

Tabla 2: Métodos de ponderaciones.

Método	δ_1	δ_2	δ_3	δ_4
Mínimo	$1/2$	$1/2$	0	$-1/2$
Máximo	$1/2$	$1/2$	0	$1/2$
Media	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0
Centroide	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	$-n_in_j/(n_i+n_j)^2$	0
Mediana	$1/2$	$1/2$	$-1/4$	0
Ward	$(n_k+n_i)/(n_i+n_j)$	$(n_k+n_j)/(n_i+n_j)$	$-n_k/(n_k+n_i+n_j)$	0
M Flexible	$(1-\lambda)/2$	$(1-\lambda)/2$	λ	0

También conocido como enlace simple, considera que la distancia entre dos conglomerados es la mínima de distancias entre dos individuos de dos conglomerados:

$$d_{i,j+k} = \text{mín}\left\{d(i,j), d(i,k)\right\}.$$

Las agrupaciones resultantes tienden a representar cadenas largas de conglomerados anidados y sus características principales son:

- No es útil para resumir datos: conglomerados grandes y sin sentido.
- Detecta valores atípicos y outliers.
- Medidas de similaridad o disimilaridad.
- Invariante bajo transformaciones monótonas de la matriz de distancias.

Método del vecino más alejado

También conocido como enlace completo, considera que la distancia entre dos conglomerados es la máxima de distancias entre dos individuos de dos conglomerados

$$d_{i,j+k} = \text{máx}\left\{d(i,j), d(i,k)\right\}$$

y sus características principales son:

- Conglomerados pequeños y compactos.
- Detecta valores atípicos y outliers.
- Medidas de similaridad o disimilaridad.
- Invariante bajo transformaciones monótonas de la matriz de distancias.

Método del centroide

La distancia entre dos conglomerados es la distancia entre los centroides de los mismos y sus características principales son:

- Variables en escala de intervalo.
- Distancia entre conglomerados se obtiene mediante las distancias entre los vectores medios.
- Si los tamaños de los dos grupos a mezclar son muy diferentes, entonces el centroide del nuevo grupo será muy próximo al de mayor tamaño y probablemente estará dentro de este grupo.

Método de la varianza mínima o método de Ward

La distancia entre dos conglomerados es la suma de los cuadrados entre grupos en el ANOVA sumando para todas las variables. En cada paso se minimiza la suma de cuadrados dentro de los clusters sobre todas las particiones posibles obtenidas combinado dos conglomerados del paso anterior.

Y sus características principales son:

- Muy eficiente.
- Conglomerados pequeños.
- Sensible a valores atípicos y outliers.
- Matriz de distancia y tabla de contigencia.
- Invariante bajo transformaciones monótonas de la matriz de distancias.

Método de la media

También conocido como enlace promedio, considera que la media (mediana) de las distancias entre todos los individuos de dos conglomerados. El proceso sólo utiliza las distancias y sus características principales son:

- Clusters ni demasiado grandes ni demasiado paqueños.
- Tendencia a fusionar clusters con varianzas pequeñas y a proporcionar clusters con la misma varianza.
- Medidas de similaridad o disimilaridad.
- No invariante bajo transformaciones monótonas de la matriz de distancias.
- Buena representación gráfica de los resultados.

Observaciones

Algunos de los problemas que surge en la clasificación de individuos cuando se aplican métodos jerárquicos son los siguientes:

i. No se consideran fuentes de error y de variación, por lo que son muy sensibles a outliers.

ii. Si un objeto se clasifica erróneamente en un cluster al inicio del proceso, no se corrige en etapa posterior.

Una buena práctica consiste en usar diferentes distacias o similitudes y comprobar si la clasificación se mantiene, es decir, existencia de grupos naturales.

3.3. Métodos no jerárquicos

Estos procedimientos se utilizan para agrupar individuos, pero no variables, en un número k de clusters prefijado. Además, no se tiene que especificar una medida de distancias ni almacenar las iteraciones, lo que hace posible trabajar con base de datos relativamente grande. Las k clases forman una única partición y no están organizadas jerárquicamente ni relacionadas entre sí. Inicialmente, los conglomerados son elegidos aleatoriamente como *representantes* y que cambian en cada iteración.

Habitualmente, se utiliza el método de las k-medias. Computacionalmente, este algoritmo

(1) Comienza con k individuos al azar como conglomerados unitarios y se asignan el resto de individuos a los cluster con el centroide más próximo.

(2) Se recalculan los centroides de los k conglomerados tras cada asignación, reasignándose los individuos al centroide más próximo.

(3) Este método puede ser iterado en etapas sucesivas (partición óptima) hasta que ningún individuo cambie de cluster en la reasignación. En ese caso se trata del método de las k-medias convergente.

Resumiendo, es un método que permite asignar a cada individuo el conglomerado más próximo respecto al centroide (media), con el objetivo de

i. minimizar la variabilidad dentro de los clusters

ii. maximizar las variabilidad entre los clusters

En cierto sentido, esto es análogo al *ANOVA invertido*, ya que el ANOVA evalúa la variabilidad entre los grupos frente a la variabilidad dentro del grupo cuando calcula la significación del contraste para la hipótesis de las medias de cada grupo son diferentes entre sí. En el método de las k-medias, el algoritmo introduce o elimina individuos de los clusters para conseguir los resultados de ANOVA más significativos. Por ello, se considera el estadístico F de Snedecor:

$$F_{m,n} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$

Concretamente, se calcula como el cociente de las medidas de cuadrados entre los clusters y las medias de cuadrados dentro de los clusters. Respecto a la interpretación de los resultados, se debe examinar las medias para cada conglomerado en cada dimensión para valorar hasta que punto son diferentes los k clusters obtenidos, teniendo en cuenta que, teoricamente, las medias deberían ser estadísticamente diferentes. Es decir, $F > 1$, ya que entonces las distancias entre los centroides de los conglomerados son mayores que las distancias de los individuos dentro de los clusters y por tanto, los clusters están suficientemente diferenciados entre sí.

Ejemplo 3.2 *Considerar 4 individuos y dos variables, X_1 y X_2, obtener la partición óptima con k=2:*

Individuo	X_1	X_2
1	***5***	***3***
2	***-1***	***1***
3	***1***	***-2***
4	***-3***	***-2***

Obsérvese que al fijar k conglomerados iniciales.

i. Si dos centroides iniciales están poco diferenciados entre sí.

ii. La existencia de outliers produce al menos un conglomerado con alta dispersión.

iii. Fijar a priori k conglomerados puede dar lugar a clusters artificiales o bien clusters distintos combinados.

3.3.1. Ventajas e inconvenientes

Procedimentos jerárquicos	Procedimentos no jerárquicos
No se especifica el número de clusters	Necesario especificar el número de clusters
Puede ser muy lento	Más rápido, más fiable
Influencia de la decisión inicial	Necesario establecer la semilla inicial
Problemas para datos con alto nivel de error	

Obsérvese que una práctica aconsejable consistiría en ejecutar primero un método jerárquico para definir el número de clusters y posteriormente, utilizar el procedimiento k-medias para formar los clusters.

3.4. Caso práctico

Veamos un caso práctico de aplicación del análisis cluster usando el fichero de datos *leukemia* del artículo anteriormente referenciado de Golub et al. (1999).

En la presente sección, trataremos de manera diferenciada los métodos jerárquicos y los no jerárquicos. En los primeros, las funciones más usuales son hclust del paquete stats y agnes del paquete cluster para los procedimientos aglomerativos, y la función diana del paquete cluster para los divisivos. En el caso de los métodos de partición iterativa se utiliza principalmente la función kmeans del paquete stats.

Así pues, comenzaremos con la aplicación de los algoritmos jerárquico y continuaremos con los no jeráquicos. En cada una de las correspondientes secciones, exploraremos el código R adecuado para llevar a cabo el análisis estadístico requerido así como la interpretación de los resultados obtenidos de la salida que visualizamos en la consola de RStudio.

El fichero de datos que utilizaremos es el dataset *leukemia* del R-package *spikeslab*.

```
> library("spikeslab")
> data(leukemia)
```

Mediante un flitrado de los datos seleccionamos lo que presentan mayores mayor variabilidad de niveles de expresión, reduciendo el conjunto de datos a 20 genes.

```
> leukemia.reorg <- leukemia[, order(apply(leukemia, 2, var), decreasing = T)]
> golub <- leukemia.reorg[, 1:20]
> golub$Y<- factor(leukemia$Y,labels=c("ALL","AML"))
```

En este punto, resulta interesante obtener las medidas descriptivas básicas de las variables del fichero.

```
> summary(golub)
    x.2404            x.3129             x.6              x.918
 Min.   :-1.5102   Min.   :-1.5102   Min.   :-1.5344   Min.   :-1.4346
 1st Qu.:-1.2745   1st Qu.:-1.1519   1st Qu.:-1.2305   1st Qu.:-1.1549
 Median :-0.8029   Median : 0.3822   Median :-0.1181   Median :-0.0112
 Mean   : 0.3716   Mean   : 0.6359   Mean   : 0.2510   Mean   : 0.4308
 3rd Qu.: 2.3283   3rd Qu.: 2.0558   3rd Qu.: 1.7127   3rd Qu.: 1.8218
 Max.   : 3.5859   Max.   : 3.7350   Max.   : 3.6552   Max.   : 3.7350
     x.979             x.973             x.1182            x.3117
 Min.   :-1.4346   Min.   :-1.5344   Min.   :-1.5344   Min.   :-1.4316
 1st Qu.:-1.2056   1st Qu.:-1.3225   1st Qu.:-1.2952   1st Qu.:-0.8991
```

```
 Median :-0.3083   Median :-1.1970   Median :-1.1682   Median :-0.4502
 Mean   : 0.3267   Mean   :-0.2172   Mean   :-0.2015   Mean   : 0.3637
 3rd Qu.: 1.6737   3rd Qu.: 0.6418   3rd Qu.: 1.4344   3rd Qu.: 2.1266
 Max.   : 3.4087   Max.   : 3.7685   Max.   : 3.5520   Max.   : 3.4431
     x.435             x.3199            x.7               x.3127
 Min.   :-1.51022   Min.   :-1.4346   Min.   :-1.5344   Min.   :-1.5344
 1st Qu.:-1.18176   1st Qu.:-0.5329   1st Qu.:-1.2484   1st Qu.:-1.3380
 Median :-0.07443   Median : 0.9947   Median :-1.0856   Median :-1.2002
 Mean   : 0.41082   Mean   : 0.8869   Mean   :-0.1081   Mean   :-0.3664
 3rd Qu.: 2.17090   3rd Qu.: 2.0111   3rd Qu.: 1.1981   3rd Qu.: 0.2259
 Max.   : 3.09050   Max.   : 3.7350   Max.   : 3.6552   Max.   : 3.8350
     x.1211            x.16              x.927             x.3389
 Min.   :-1.4235   Min.   :-1.5102   Min.   :-1.5102   Min.   :-1.5102
 1st Qu.:-0.7114   1st Qu.:-0.7020   1st Qu.:-1.0503   1st Qu.:-1.1872
 Median : 0.4253   Median : 0.2963   Median : 0.6609   Median : 0.3204
 Mean   : 0.6417   Mean   : 0.6272   Mean   : 0.5629   Mean   : 0.3081
 3rd Qu.: 1.7034   3rd Qu.: 1.9715   3rd Qu.: 1.7576   3rd Qu.: 1.6769
 Max.   : 3.7648   Max.   : 3.7940   Max.   : 3.7350   Max.   : 3.2375
     x.3116            x.3126            x.8               x.3327
 Min.   :-1.5102   Min.   :-1.43464   Min.   :-1.53441   Min.   :-1.5344
 1st Qu.:-1.2525   1st Qu.:-1.19653   1st Qu.:-1.24633   1st Qu.:-1.3153
 Median :-0.9602   Median :-0.64040   Median :-1.01832   Median :-1.2030
 Mean   :-0.1504   Mean   :-0.01142   Mean   :-0.05174   Mean   :-0.5478
 3rd Qu.: 1.3527   3rd Qu.: 0.94948   3rd Qu.: 1.23310   3rd Qu.:-1.0080
 Max.   : 3.4431   Max.   : 3.94203   Max.   : 3.33431   Max.   : 3.6552
   Y
 ALL:47
 AML:25
```

3.4.1. Ejemplo análisis cluster jerárquico mediante R

```
> hclust(d, method = "complete", members = NULL)

## S3 method for class 'hclust'
> plot(x, labels = NULL, hang = 0.1,
     axes = TRUE, frame.plot = FALSE, ann = TRUE,
     main = "Cluster Dendrogram",
     sub = NULL, xlab = NULL, ylab = "Height", ...)
```

- *d*, un matriz de distancia de los individuos.
- *method*, el método aglomerativo utilizado a elegir entre: "ward" (por defecto), "single", "complete", "average", "mcquitty", "median" o "centroid".
- *members*, nulo o un vector con el tamaño de la matriz *d*.
- *x*, un objeto del tipo producido por hclust.
- *hang*, la fracción de la altura del gráfico mediante la cual las etiquetas debería colgar por debajo del gráfico.

- *labels*, un vector de etiquetas para los hojas del árbol, por defecto se usan los nombres y números de las filas originales. Si *labels* = *FALSE* las etiquetas no se representan.
- *axes, frame.plot, ann,main, sub, xlab, ylab*, veáse en *plot.default*.

En nuestro caso particular, ejecutaremos la sentencia correspondiente y después, listaremos los objetos de *hclust*.

```
>hc=hclust(dist(golub[,-21]))
> names(hc)
[1] "merge"       "height"      "order"       "labels"      "method"
[6] "call"        "dist.method"
```

Cada uno de estos objetos nos facilitan diversa información obtenida a través de la función *hclust*, concretamente:

merge presenta una matriz de orden $n-1$ por 2, donde la fila i-ésima indica la combinación de los conglomerados en la etapa i. Si un elemento j en la fila es negativo, entonces se fusionó en esta etapa. Si j es positivo, entonces la combinación era con el clúster formado en el etapa j del algoritmo.

height suministra un conjunto de $n-1$ valores reales. La altura de la agrupación, es decir, el valor del criterio asociado con el método de agrupamiento para la aglomeración particular.

order un vector que da la permutación de las observaciones originales adecuados para el trazado, en el sentido de que una parcela clúster mediante esta ordenación y combinación de matriz no tendrá cruces de las ramas.

labels muestra las etiquetas para cada uno de los objetos que se agrupan.

call la salida corresponde a la llamada que produjo el resultado.

method indica el método de clúster que se ha utilizado.

dist.method indica la distancia que se ha utilizado para crear la matriz de distancia

Una función también incluida el paquete *stats* es *cutree* para cortar un árbol de clasificación resultante desde *hclust* en diferentes grupos especificando el número(s) de conglomerados deseados o la altura(s) de corte.

```
> cutree(tree, k = NULL, h = NULL)
```

Sobre los detalles de sus argumentos señalar,

- *tree*, un árbol producido por hclust. cutree() sólo espera una lista con las componentes merge, height y labels.
- k, un entero o vector con el número deseado de conglomerados.
- h, un escalar o vector con las alturas donde el árbol debería ser cortado.

Nótese que al menos k o h debe ser especificado y si ambos son dados k invalida h. Además, cortar árboles en una altura determinada sólo es posible en árboles ultramétricos, con alturas de agrupamiento monótonas.

```
> cutree(hc, k = 1:5) #k = 1 is trivial
      1 2 3 4 5
 [1,] 1 1 1 1 1
 [2,] 1 2 2 2 2
 [3,] 1 1 1 1 3
 [4,] 1 1 1 1 1
 [5,] 1 1 1 1 1
 [6,] 1 1 1 1 1
 [7,] 1 1 1 1 3
 [8,] 1 1 1 1 1
 [9,] 1 1 1 1 1
[10,] 1 1 1 3 4
[11,] 1 1 1 1 1
[12,] 1 1 1 1 3
[13,] 1 1 1 3 4
[14,] 1 1 1 1 1
[15,] 1 1 1 1 1
[16,] 1 1 1 3 4
[17,] 1 1 1 3 4
[18,] 1 1 1 1 1
[19,] 1 1 1 3 4
[20,] 1 1 1 3 4
[21,] 1 1 1 3 4
[22,] 1 1 1 1 3
[23,] 1 1 1 1 1
[24,] 1 1 1 3 4
[25,] 1 1 1 1 3
[26,] 1 1 1 3 4
[27,] 1 1 1 1 1
[28,] 1 2 2 2 2
[29,] 1 2 2 2 2
[30,] 1 2 2 2 2
[31,] 1 2 2 2 2
[32,] 1 2 2 2 2
[33,] 1 2 2 2 2
[34,] 1 2 2 2 2
[35,] 1 2 2 2 2
[36,] 1 2 2 2 2
[37,] 1 2 2 2 2
[38,] 1 2 2 2 2
[39,] 1 1 1 3 4
[40,] 1 1 1 3 4
[41,] 1 1 1 3 4
[42,] 1 1 1 3 4
[43,] 1 2 3 4 5
[44,] 1 1 1 3 4
[45,] 1 1 1 3 4
[46,] 1 1 1 3 4
[47,] 1 1 1 3 4
```

```
[48,] 1 1 1 1 1
[49,] 1 1 1 1 1
[50,] 1 2 2 2 2
[51,] 1 2 2 2 2
[52,] 1 2 2 2 2
[53,] 1 2 2 2 2
[54,] 1 2 3 4 5
[55,] 1 2 3 4 5
[56,] 1 2 3 4 5
[57,] 1 2 3 4 5
[58,] 1 2 3 4 5
[59,] 1 2 3 4 5
[60,] 1 2 3 4 5
[61,] 1 2 2 2 2
[62,] 1 2 2 2 2
[63,] 1 2 2 2 2
[64,] 1 1 1 3 4
[65,] 1 2 2 2 2
[66,] 1 2 2 2 2
[67,] 1 2 2 2 2
[68,] 1 1 1 3 4
[69,] 1 1 1 3 4
[70,] 1 1 1 1 3
[71,] 1 1 1 3 4
[72,] 1 1 1 3 4
```

La función *cutree* devuelve un vector con los miembros de cada grupo, si k o h son escalares; en caso contrario devuelve una matriz con los miembros de cada grupo donde cada columna corresponde a los elementos de k o de h, respectivamente (los cuales son también usados como nombres de columnas).

```
> cutree(hc, h = 12)
 [1] 1 2 3 1 1 1 3 1 1 4 1 3 4 1 1 4 4 1 4 4 4 3 1 4 3 4 1 2 2 2 2 2 2 2 2 2 2 4 4
[41] 4 4 5 4 4 4 4 1 1 2 2 2 2 5 5 5 5 5 5 5 2 2 2 4 2 2 2 4 4 3 4 4
```

Adiccionalmente, la función *cutree* permite comparar las diferentes agrupaciones obtenidas. Así, por ejemplo, la comparación del primer y cuarto clustering de nuestro análisis concreto se podría obtener a partir de las siguientes sentencias:

```
> ## Compare the 2 and 4 grouping:
> g24 <- cutree(hc, k = c(2,4))
> table(grp2 = g24[,"2"], grp4 = g24[,"4"])
    grp4
grp2  1  2  3  4
   1 14  0  0  2
   2  0 14 20  0
```

La clase *dendrogram* proporciona funciones generales para la manipulación de estructuras en forma de árbol, con la intención de que sustituya a funciones similares en la agrupación y clasificación / regresión árboles jerárquicos, y de manera que todos ellos pueden usar el mismo mecanismo para el trazado o corte de árboles.

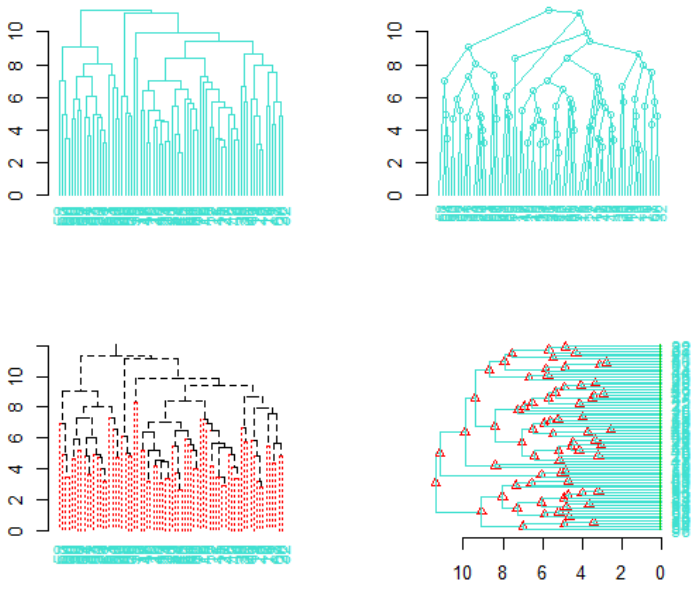


Figura 10: Dendogramas

La función *identify.hclust* lee la posición de los gráficos señalados por el puntero cuando se presiona el botón ratón, cortando el árbol en la posición vertical indicada del gráfico y destacando el conglomerado que contiene la posición horizontal del puntero. Opcionalmente se aplica una función al índice de puntos de datos contenidos en el cluster.

```
## S3 method for class 'hclust'
identify(x, FUN = NULL, N = 20, MAXCLUSTER = 20, DEV.FUN = NULL,
```

- x, un objeto del tipo producido por la función hclust.
- FUN, es una función opcional que se aplica a los números de índice de los puntos de datos de un conglomerado.
- N, el número máximo de conglomerados que se desea identificar.
- MAXCLUSTER, el número máximo de conglomerados que pueden producirse por un corte.
- DEV.FUN, es un escalar entero opcional. Si se especifica, el dispositivo de gráficos correspodiente se activa antes de aplicar FUN.

Las sentencias apropiadas para nuestra base de datos concreta serían las que aparecen abajo. Se deja como ejercicio para el lector la exploración de la correspondiente salida de resultados.

```
> require(graphics)
> hca <- hclust(dist(golub[,-21]))
> plot(hca,hang=-1,cex=0.5)
> (x <- identify(hca)) ## Terminar con 2º botón del ratón !!
```

3.4.2. Ejemplo análisis cluster no jerárquico mediante R

La ejecución del método k-medias en R se puede llevar a cabo mediante diferentes packages. Uno de ellos es *stats* que dispone de la función *kmeans*.

```
> kmeans(x, centers, iter.max = 10, nstart = 1,
  + algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)
```

El objetivo principal de este procedimiento es obtener la mejor partición en k conglomerados tal que la suma de cuadrados desde los individuos al centroide del asignados cluster sea mínima. El algoritmo por defecto es el de Hartigan y Wong en 1979, aunque también también se pueden considerar otras alternativas como los algoritmos dado por MacQueen en 1967, por Lloyd en 1957 and por Forgy en 1965 respectivamente . En raras ocasiones, cuando alguno de los individuos es extremadamente próximo, el algoritmo puede no converger, avisando y devolviendo ifault = 4, resultando aconsejable un ligero redondeo de los datos. Nótese que si $k = 1$ el resultado será el centro y la suma de cuadrado dentro del conglomerado. Salvo para el método de Lloyd-Forgy, k clusters será obtenidos si se especifica un número. Si la matriz inicial de centros es suprimida puede ocurrir que ningún individuo esté próximo a uno o más centroides.

Además la función *kmeans* devuelve un objeto de la clase "kmeans" tales como:

- *cluster*, un vector de enteros, desde 1 hasta k, indicando el cluster al cual cada individuos está localizado.
- *centers*, una matriz de los centroides de los clusters.
- *totss*, la suma total de cuadrados.
- *withinss*, vector de la suma de cuadrados dentro del cluster, uno componente por cluster.
- *tot.withinss*, total de la suma de cuadrados dentro del cluster, i.e., sum(withinss).
- *betweenss*, la suma de cuadrados entre clusters, i.e. totss-tot.withinss.
- *size*, el número de individuos en cada cluster.
- *iter*,el número de iteraciones.
- *ifault*, entero: indicador de posibles problemas.

```
> km <- kmeans(golub[,-21],3)
```

```
> km$cluster
 [1] 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 3 1 1 1 3 3 3 2 2 3 3 3 3 3 3 1 1 1 1 2 2
[56] 2 2 1 2 2 1 1 1 1 1 3 1 3 3 3 3 3
> km$centers

      x.2404     x.3129         x.6      x.918      x.979      x.973     x.1182     x.3117      x.435     x.3199
1 -1.2244043  1.4782747  0.09363294  1.5749173  2.2072110 -0.6006823  1.5782372  2.3326985 -1.0922888 1.4200661
2  0.9187424  3.1562029 -1.05381865  2.0126642  0.6928560  2.8262984  0.2503659 -0.8397041  0.9487188 1.9581791
3  1.1033638 -0.2853359  0.58201624 -0.4698614 -0.7280187 -0.5959974 -1.2198154 -0.4384181  1.0956989 0.4035406
         x.7     x.3127    x.1211       x.16      x.927     x.3389    x.3116     x.3126        x.8     x.3327
1 -0.5396924  0.3352632 0.8550848  0.2244293  1.2167893 -0.1701823  1.565261  0.7984783 -0.3019417 -0.8053451
2 -1.1857717  2.2584884 0.8689226 -0.6316053  2.8747844  1.8849892 -1.046417  2.5633752 -1.1857717 -1.1857717
3  0.3231800 -1.2338590 0.4866934  1.0779109 -0.2199108  0.2582944 -0.878439 -0.9260892  0.2953275 -0.2913221

> km$withinss
[1]  631.4231  227.8747 1423.9640
> km$tots
[1] 3581.748
```

```
> km$tot.withinss
[1] 2283.262
> km$betweenss
[1] 1298.486
> km$size
[1] 22  8 42
> km$iter
[1] 2
> km$ifault
[1] 0
```

```
> datos_pca <- rbind(golub[,-21], km$centers);datos_pca
```

```
        x.2404      x.3129         x.6       x.918       x.979       x.973     x.1182      x.3117        x.435
1    3.11211487 -1.41409516  3.11380526 -1.41409516 -0.40603987 -1.41409516 -1.4140952 -0.32018979  1.293976155
2   -1.33516303  1.97997804  0.25121538 -1.33516303  0.94302739 -1.33516303 -1.3351630  2.13963149 -0.738587927
3   -1.42349876  0.35359074  2.08052907 -1.42349876 -0.61452049 -1.42349876 -1.4234988 -0.46718045  0.100560023
4    2.27589179 -0.94746097  2.21050886 -1.36270310 -1.36270310 -1.36270310 -1.3627031 -0.55295399  2.216374131
5    2.37460208  1.44639309  0.96380802  0.67631124 -1.37341523 -1.37341523 -0.3985267 -1.37341523  1.850861298
6   -1.19283327  1.36184311  2.23036876 -1.19283327  1.06578027 -1.19283327 -1.1928333 -0.48942324 -1.192833271
7   -1.33742052 -1.25809340  3.33430954 -1.33742052 -1.33742052 -1.33742052 -1.3374205 -0.91055038  1.159642767
8    2.77759438 -1.33541394  1.96754796 -1.33541394 -1.33541394 -1.33541394 -1.3354139  0.02735239  0.934887669
9   -1.42439234  0.67533476  2.66312117 -1.42439234 -0.40674494 -1.42439234 -1.4243923 -0.17082471 -1.424392336
10  -1.18574866  0.81698415 -1.18574866  0.05816397  0.27609653 -0.03984054 -1.1857487  0.31895447 -0.006817899
11  -1.24374365 -1.14255063  0.60292122 -1.24374365 -1.24374365 -1.24374365 -1.2437437  1.03130106 -1.243743654
12  -0.02866797  2.25642715  3.24355309  1.56402980 -0.94858663  0.19047524 -1.0137741  2.46108030  0.877438809
13   2.37896718 -1.31831643 -1.31831643 -1.31831643 -1.31831643 -1.31831643 -1.3183164 -1.18759310  2.485437839
14  -1.35606060  1.64486357  0.66231234 -1.35606060  0.56521760 -1.35606060 -1.3560606 -1.35606060 -1.356060601
15   3.02273719 -1.33968610  2.34322604 -0.28969601 -1.33968610 -1.33968610 -1.3396861 -1.33968610  2.141047764
16   0.60184111 -0.93516058 -1.35738627  1.98700478 -1.35738627 -1.35738627 -1.3573863 -1.35738627  2.049761249
17  -1.51021939 -1.51021939 -1.51021939  0.71991657 -0.40535396 -1.51021939 -1.5102194 -0.64360340 -1.510219388
18   2.10745349 -1.17997570  0.84466229 -0.59727096 -1.17997570 -1.17997570 -1.1799757 -0.88613688 -1.179975696
19   2.12201426 -1.20760661 -1.20760661  0.27679479 -1.20760661 -1.20760661 -1.2076066 -1.20760661 -1.207606614
20   2.54424465 -0.96936981 -1.53441337 -0.46365884 -0.95752219 -1.53441337 -1.5344134 -1.39461598  0.396297277
21   1.10092339 -1.04576918 -1.04576918 -1.04576918 -1.04576918 -1.04576918 -1.0457692 -0.89526914  3.090499672
22   2.16548894 -0.53925028  3.65212276 -1.11159548 -1.11159548 -1.11159548 -1.1115955 -0.43324085  2.424877799
23  -1.18711813  0.34273276  3.65519390 -1.18711813  0.19341256 -1.18711813 -1.1871181 -0.82604531 -1.187118125
24   1.85098232 -1.43464528 -1.43464528 -1.43464528 -1.43464528 -1.43464528 -1.4346453 -0.86599250  2.000282639
25   1.70405152 -1.21717418  2.36626425 -0.94168238 -1.21717418 -1.21717418 -1.2171742  2.94072235  2.395600024
26   3.03128654 -1.24631406 -1.24631406 -0.62885943 -1.24631406 -1.24631406 -1.2463141 -0.69182895  2.727314454
27   2.80392759 -1.21636235  2.84693862 -1.21636235 -1.21636235 -1.21636235 -1.2163623 -0.83988241  2.156286670
28  -1.38572762  0.19740482  0.73435067 -0.08056435  0.74105773 -1.38572762 -1.3857276  3.11692690 -0.920770693
29  -1.06344516  0.15980751  0.31472494  1.80471439  2.87134458  1.06813789  2.8174608  2.12225510 -1.063445157
30  -1.27240613  2.64271140 -1.27240613 -0.23084810  2.84042110 -1.20124207  2.2786861  2.17549278 -1.272406126
31  -1.22848765  1.06236212  2.75798768  2.94780335  2.56104264 -1.22848765  2.2036089  1.46424318 -0.758878890
32  -1.40418921  1.16846518  2.09452981 -0.75141867  2.85988008 -1.40418921 -1.0999350  2.88646098 -1.069077710
33  -1.43472701  2.68237019  0.71606417  2.29852896  1.62467134  1.79718497  1.1122714  2.88023126 -0.902548547
34  -1.31425783 -0.22372550  0.14832618  2.52237891  1.15660602  0.51975110  2.1920165  2.65439383 -1.195771910
35  -1.21438538 -1.21438538  1.74338819  3.65027637  1.37582999 -1.21438538 -1.2143854 -0.47430294 -1.214385383
36  -1.36558728  2.84773233  0.09375529  1.12348957  2.84773233  0.91674345  1.8079960  2.84773233 -1.365587283
37  -1.31181697  2.30088189  1.32175861  0.65748539  2.86499807 -1.31181697  2.2870929  3.02312225 -1.087719593
38  -1.29914956 -1.29914956  0.58658432  2.32666712  1.22064999 -1.29914956  1.1695534  2.83426585 -1.299149559
39   1.62967871  0.08540474 -1.22521291  0.22139696 -1.22521291  0.14992984 -1.2252129  0.59787424  1.504395146
40  -1.25458621 -1.25458621 -1.25458621 -1.25458621 -1.25458621 -1.25458621 -1.2545862 -0.59484028  1.498089882
41   1.31386494  0.41090029 -1.10862151 -1.14543217 -1.14543217  2.61275096 -1.1454322 -1.14543217  2.288909068
42  -1.11292734  1.66294597 -1.24638365  0.81410911 -0.42247332  2.48827010 -1.2463837  0.83478083  1.844285000
43   2.77729853  3.57967433 -0.14100870  1.15208320 -1.19663299  3.57967433 -1.1966330 -1.19663299  2.288796003
44   2.85474189  0.59491970 -0.33627044 -1.07244827  0.40856046  1.35197504 -1.0724483 -1.07244827  2.299562622
45   2.78037438 -1.06247623 -1.06247623 -1.06247623 -1.06247623 -0.43758658 -1.0624762 -0.70035078  2.789692270
46   3.36335403 -0.40832494 -0.09526977 -1.12033020 -1.12033020  1.33373053 -1.1203302 -1.12033020  2.291035424
47  -0.53812635 -0.68692166 -1.27090583  0.63506511 -1.20492525 -1.27090583 -0.5804307 -0.60968886  1.612746362
48   2.96573901 -0.52716777  0.88720533 -0.83533104 -1.43155275 -1.43155275 -1.4315527 -1.43155275  2.572010240
49   2.40555076 -1.30407907  2.70859291  0.91151978 -0.11715707 -1.30407907 -1.3040791 -0.71277651  2.458365073
50  -1.28071264  1.31198142 -1.28071264  2.67876490  2.89445211 -1.28071264  2.3927636  2.99755914 -1.280712643
51  -1.27702120  1.02038171  1.48915292  3.02617395  1.82095734 -0.17706358  2.9967580  3.44306058 -1.227021204
52  -0.99086179  3.40866390 -0.99086179  0.75232663  3.40866390  1.20552884  2.7180238  0.55219122 -0.990861792
53  -1.44361853  1.15559626  0.86314370  2.49445281  2.99733306  0.07318698  2.9973331  2.41794381 -1.443618533
54  -1.32727150  2.69799211 -1.32727150  3.27867904  2.91666554  2.84089739  2.1000122 -1.32727150 -1.015448252
55   2.16581112  3.28491296 -1.15638912  1.87306070  0.08390296  3.76845304 -1.1563891 -1.15638912  1.448535446
56   3.58592665  3.12782253 -1.29220660 -0.23306249  0.37693574  3.66116015 -1.2922066 -1.29220660  2.214748898
57  -1.05800991  3.71399064 -1.05800991  3.53865526  2.39234322 -1.05800991  3.5519642 -0.81936183 -1.058009908
58  -0.61503512  3.44039685 -1.22409905  3.33264718  3.33025433 -1.22409905  2.5239554  1.58796313 -1.224099054
59   3.42957655  3.44730806 -1.09881462  1.94281255 -0.55519947  3.59496552 -1.0988146 -1.09881462  2.977308333
60  -1.11046511  3.73497632 -1.11046511  3.73497632  1.94730609  3.73497632  2.3413779 -0.66173663 -1.110465113
61  -1.21622049 -1.21622049 -1.21622049  3.34919022 -1.02999934 -1.21622049  1.5352624  3.22701256 -1.216220492
62  -0.52000998  3.16937705 -1.31109581  1.56725699  3.16937705 -1.31109581  2.1854273 -0.27874074 -1.311095810
63  -1.42036084  2.94873490  1.70249798  2.08916374  2.94873490 -1.42036084  2.3438037  2.51005276 -0.858789690
64  -1.48162766 -1.48162766 -1.48162766  1.72443812  1.57276075 -0.67374690  1.4008308  1.38846452 -0.841475935
65  -1.14315792  2.80373318 -1.10701302  3.53400465  3.32366892 -1.14315792  2.3985554  2.80182291 -1.143157921
66  -1.01867807  0.31619626  0.62624121 -0.23896099  1.86296473 -1.01867807 -1.0186781  0.70897453 -0.294501733
67  -1.18331115  2.44218678 -1.13013029 -1.18331115  1.59100683 -1.18331115 -0.8193562  2.52728218 -0.819356197
68   1.22956920  1.98894162 -1.36238133 -1.36238133 -0.94564490  2.43845981 -1.2431708 -0.55727273  2.361637893
69   2.93406135  0.42899778 -1.42246873  0.37434304 -1.42246873  1.66174810 -1.4224687 -1.06130488  2.236411789
70   2.45759412  1.27560033  3.17757073  1.06988670 -0.21119615  2.32088506 -1.0413036 -0.22323079  2.290501881
71   0.03523598  1.77837367 -1.21065485 -1.21065485 -1.21065485  1.69980957 -1.2106548  1.83900688 -0.142033143
72   2.31287414 -0.04620471 -1.29024248 -0.25114942  0.09328408  0.55011822 -1.2902425 -0.42580883  1.474536254
110 -1.22440431  1.47827465  0.09363294  1.57491735  2.20721096 -0.60068233  1.5782372  2.33269855 -1.092288758
210  0.91874237  3.15620286 -1.05381865  2.01266421  0.69285597  2.82629837  0.2503659 -0.83970406  0.948718801
310  1.10336382 -0.28533586  0.58201624 -0.46986137 -0.72801875 -0.59599741 -1.2198154 -0.43841809  1.095698864
         x.3199        x.7      x.3127      x.1211        x.16       x.927      x.3389      x.3116      x.3126
1    2.253094425  2.74940712 -1.41409516 -1.4140952  2.41614181 -1.06376240  0.51138594 -0.4396662 -1.41409516
2    2.056197299 -1.21310347 -1.33516303  0.9727421  1.37029920  1.21680834 -1.33516303  1.3077891 -1.33516303
3   -1.423498758  1.60354909 -1.42349876 -1.4234988  1.61749392  0.44592321  1.34141755 -1.4234988 -1.42349876
4   -1.362703100  1.52090115 -1.36270310  2.4683833  1.78593044 -1.36270310  0.31243867 -1.3627031 -0.86882882
```

5	0.391167544	1.65482795	-1.37341523	1.4311109	0.84682388	0.75573801	-0.41124406	-1.0785037	-0.32281108
6	-1.192833271	1.85032385	-1.19283327	-1.1928333	2.17512972	0.77156928	-1.19283327	-1.1928333	-1.19283327
7	-1.337420520	2.99199232	-1.33742052	1.5791344	2.66834383	-1.33742052	-0.88204946	-1.2751403	-1.33742052
8	1.124569970	1.12864643	-1.33541394	1.9403921	2.04273676	-1.33541394	0.41866237	-1.3354139	-1.33541394
9	0.429860271	2.17501074	-1.42439234	-0.8810301	2.02936745	0.48865325	-1.42439234	-1.3914593	-1.28541252
10	-1.185748659	-1.18574866	-0.75210740	-1.1857487	-1.18574866	0.99685809	0.78304620	-1.1857487	-0.06077395
11	-1.243743654	-1.06618649	-1.24374365	-0.3307806	1.19258384	-1.24374365	-1.24374365	-0.6753339	-1.24374365
12	-1.013774060	2.58829500	-1.01377406	-1.0137741	2.59628311	1.97871343	-1.01377406	1.8000261	-1.01377406
13	1.357726988	-1.31831643	-1.31831643	-1.3183164	-1.31831643	-1.31831643	-1.18759310	-1.3183164	-1.31831643
14	0.007105379	0.25374461	-1.35606060	1.0787427	0.32163608	1.20588973	-1.35606060	-1.3560606	-1.35606060
15	1.203722330	2.36176786	-1.33968610	-0.6838277	1.56030394	-1.33968610	-1.33968610	-1.3396861	-1.33968610
16	1.342385104	-1.35738627	-1.35738627	-1.3573863	-1.15603300	-1.11903048	0.49691418	-1.3573863	-0.83892658
17	1.853556153	-1.51021939	-1.51021939	0.4632358	-1.51021939	-1.51021939	-1.51021939	-1.5102194	-1.17687950
18	3.593784657	0.95415042	-1.17997570	-1.1610519	1.96320285	-1.17997570	2.12415792	-1.1799757	-0.62359827
19	-0.121923982	-1.20760661	-1.20760661	1.0324625	-0.85505281	-0.45090109	-1.11717265	-1.2076066	-1.20760661
20	-0.137100737	-1.53441337	-1.53441337	2.5261102	-0.60468703	-0.49832293	0.72287583	-0.8241745	-0.85808880
21	-1.045769178	-1.04576918	-1.04576918	-1.0457692	-1.04576918	-1.04576918	2.26684983	-0.4913176	-0.64142238
22	-1.111595482	3.48426947	-1.11159548	1.4290135	2.94318507	-1.11159548	0.18491921	-0.7002356	-0.81610419
23	-1.187118125	3.65519390	-1.18711813	-0.1452950	2.92627479	0.44976401	-1.18711813	-0.8130639	-1.18711813
24	-1.434645277	-1.43464528	-1.43464528	-0.7939280	-0.60633768	-1.02359565	-0.98302163	-1.4346453	-1.43464528
25	-0.468680751	1.79672447	-1.21717418	3.1948038	2.92221697	-1.21717418	-1.05748749	2.2293426	-1.21717418
26	0.399227967	-1.24631406	-1.24631406	0.2413573	-0.17335788	0.02566885	2.52798773	-1.2463141	-1.24631406
27	-0.104979471	2.20974107	-1.21636235	-0.8907726	2.35479154	-1.21636235	0.74749067	-1.2163623	-0.81544643
28	1.639826291	0.05310175	-1.38572762	1.6843555	1.15873072	-0.18221613	-1.38572762	2.7832257	-0.39220404
29	2.350218404	-1.06344516	3.00703677	2.8602224	-0.14579516	0.71557377	-1.06344516	1.5379327	3.27663338
30	1.581760565	-1.27240613	0.99412269	0.6198950	-0.66366876	1.80001797	-1.27240613	1.5282226	1.29352685
31	1.301027869	2.00282347	0.05204316	0.8524917	2.00900514	0.80754033	-1.22848765	0.4783499	0.88069477
32	1.235196129	1.18859141	-1.40418921	1.5969349	1.29724883	1.11553811	0.32900805	2.3714790	-0.23375333
33	1.686148194	-1.43472701	1.81615651	2.8802313	0.40454962	2.10397784	-1.43472701	2.7151277	1.74832017
34	3.253777319	-1.31425783	0.69350043	0.5740125	0.24521467	-0.02955452	0.32831901	1.9493018	1.27344929
35	1.660211864	1.53731871	-1.21438538	3.3478001	1.52683987	-1.21438538	0.33230209	-1.2143854	-0.47430294
36	2.810159183	-1.14777711	-1.36558728	0.3418359	0.27095063	2.50670026	-0.32556561	2.5747978	-0.27719179
37	3.014160249	1.03639440	0.77356083	1.1226086	1.07486151	1.65745478	-1.31181697	2.8443870	1.04887182
38	0.811944060	0.16708933	0.35699141	3.2518818	1.99656985	-1.29914956	1.57027554	2.2216771	0.47973280
39	-1.225212914	-1.22521291	-1.22521291	-1.2252129	2.56298924	0.80857635	2.94091084	-0.1247503	-1.15515871
40	0.485107460	-1.25458621	-1.25458621	-0.4912313	-1.25458621	-1.25458621	1.22198024	-0.7721910	-0.75514320
41	-0.958800396	-1.14543217	-0.38433239	2.8308044	-0.17573532	0.79085408	-1.14543217	-0.7648823	0.11491561
42	0.786411440	-1.24638365	-0.21466227	1.7148328	-0.59218647	1.74342236	2.32751723	-0.5292811	0.95906018
43	2.581536311	-1.19663299	2.74601642	3.5796743	0.46417646	3.57967433	-1.19663299	-1.1966330	2.81635106
44	1.478871196	-1.07244827	-1.07244827	0.1744703	3.58618399	0.86457056	1.74550242	-0.9933436	-0.65839439
45	0.344783158	-1.06247623	-1.06247623	1.9896846	3.50056582	-0.18782522	0.73645561	-1.0624762	-1.02686021
46	0.506878950	-1.12033020	-1.12033020	2.0542003	3.75009517	-0.39471256	0.64596588	-1.1203302	-0.64559156
47	3.678393605	-1.27090583	-1.27090583	3.0010960	-0.75343821	-0.35421261	-1.27090583	-1.2709058	-0.27732794
48	-0.494242394	0.16590412	-1.43155275	1.6996333	0.90957369	-1.08604045	-1.43155275	-1.4315527	-1.43155275
49	0.864755287	2.24409926	-1.30407907	-0.8498634	2.88340370	-1.13194971	1.78991657	-1.3040791	-1.30407907
50	-1.280712643	-1.28071264	-1.28071264	-1.2807126	-1.15814764	0.92365192	0.49325542	2.7238428	0.30075265
51	3.081815918	-0.53993266	2.56727218	1.6736753	1.04563699	1.16451166	-1.20879919	3.4430606	2.53266453
52	-0.990861792	-0.99086179	-0.99086179	-0.3987043	-0.11078682	2.89747952	-0.99086179	-0.9908618	-0.63937547
53	1.971625146	0.26906344	2.77499897	0.9467713	1.21206968	0.15433504	-0.87281157	1.6988632	2.73511123
54	3.278679036	-1.32727150	3.27657791	-1.3272715	-1.32727150	2.21284504	2.42630554	-0.9295353	3.27867904
55	-0.676346267	-1.15638912	1.85449449	2.8376356	-1.11833016	2.60417794	1.90668482	-1.1563891	1.79887529
56	0.220287918	-1.29220660	0.18225715	-0.4624598	-0.05566646	3.21156842	3.23752028	-1.2922066	0.94629138
57	3.532738612	-1.05800991	3.83495090	-0.1471349	-1.05800991	2.86264537	2.54841486	-1.0580099	3.94202717
58	2.565804754	-1.22409905	1.48207891	3.4046715	-1.22409905	2.52783962	3.15648748	0.5116623	1.64236250
59	2.207149197	-1.09881462	2.65329596	-0.1399874	-0.25508894	3.04896581	1.65405045	-1.0988146	3.03074103
60	3.734976324	-1.11046511	3.73497632	0.8960920	-1.11046511	3.73497632	2.17605358	-1.1104651	3.73497632
61	-1.216220492	-1.21622049	1.73616179	0.2890552	-1.21622049	-0.36712285	1.96724901	2.4462533	1.96121960
62	3.169377049	-1.31109581	-1.31109581	-0.2437996	-0.69476743	3.14060787	-1.31109581	-1.3110958	-0.45407125
63	2.071756813	1.22644007	1.04314736	-1.4203608	0.58732605	2.48136789	-0.78164473	1.4876172	1.24181679
64	0.670500205	-1.48162766	-1.48162766	-0.5126498	-1.15084208	-0.61994008	2.04911618	-1.4816277	-0.56925429
65	-0.648738064	-1.14315792	1.81699531	-0.4165339	-0.82000067	2.41363627	1.60558046	1.9801809	2.23569125
66	1.634034291	0.46137212	-1.01867807	3.7648195	0.16380042	-0.61018722	-0.21438669	-0.7704665	-0.87014355
67	0.106691647	-1.18331115	-1.18331115	0.0132413	-0.55068965	1.64030692	-0.72074891	1.6155647	-1.18331115
68	1.522322565	-1.36238133	-1.36238133	-0.9148119	-0.13629137	1.92125562	0.99608890	-0.5980592	-1.36238133
69	1.217267516	-1.42246873	-1.42246873	1.8953051	-0.72355093	0.65966140	-0.08697372	-1.4224687	-0.07747875
70	3.490782975	2.53006495	-1.04130363	0.3874325	3.79396763	1.70500168	2.93605420	-0.6299315	0.70921420
71	1.162755782	-1.21065485	-1.21065485	-0.4489295	-0.11500237	1.84087535	2.64078407	0.9975667	-1.21065485
72	1.996129347	-1.29024248	-1.29024248	0.6792864	-0.15348262	0.66206745	2.47990482	-1.0858865	-0.89481403
110	1.420066096	-0.53969236	0.33526319	0.8550848	0.22442932	1.21678932	-0.17018227	1.5652614	0.79847833
210	1.958179071	-1.18577169	2.25848836	0.8689226	-0.63160526	2.87478445	1.88498922	-1.0464168	2.56337518
310	0.403540573	0.32317999	-1.23385902	0.4866934	1.07791087	-0.21991075	0.25829440	-0.8784390	-0.92608921

	x.8	x.3327
1	2.62886231	-1.0273841
2	1.04030045	2.9016898
3	1.70269669	-1.4234988
4	1.62552804	-1.3627031
5	-0.31990890	-1.3734152
6	1.72910049	3.4708874
7	3.33430954	-1.3374205
8	0.97091247	-0.1708843
9	2.72135893	2.7422413
10	-1.18574866	3.4347771
11	1.20532346	3.2879615
12	2.82073050	-1.0137741
13	-1.31831643	-1.3183164
14	1.34015430	3.2780645
15	1.54985455	-0.9799845
16	-1.35738627	-1.3573863
17	-1.51021939	-1.5102194
18	-0.22024406	-1.1799757
19	-1.20760661	-1.2076066
20	-1.53441337	-1.5344134
21	-1.04576918	-1.0457692
22	3.21760326	-1.1115955
23	2.70090280	3.6551939
24	-1.43464528	-1.4346453
25	2.23491393	-1.2171742
26	-1.24631406	-1.2463141
27	1.80309357	-1.2163623
28	1.31643629	-1.3857276

```
29  -1.06344516 -1.0634452
30  -1.27240613 -1.2724061
31   1.73836589 -1.2284876
32   0.91031882  1.6158518
33  -0.50067700 -1.4347270
34  -1.19577191 -1.3142578
35   0.90143322 -1.2143854
36  -1.36558728 -1.3655873
37   0.53597666 -1.3118170
38   1.35895776 -1.2991496
39  -1.22521291 -1.1983069
40  -1.25458621 -1.2545862
41  -1.14543217 -1.1454322
42  -1.24638365 -1.2463837
43  -1.19663299 -1.1966330
44  -0.21679095 -1.0724483
45  -1.06247623 -1.0624762
46   0.01666164 -1.1203302
47  -1.27090583 -0.6096889
48  -0.31691080 -1.4315527
49   1.84605540 -1.3040791
50  -1.28071264 -1.2807126
51   0.23911206 -1.2270212
52  -0.99086179 -0.9908618
53   0.65382643 -1.4436185
54  -1.32727150 -1.3272715
55  -1.15638912 -1.1563891
56  -1.29220660 -1.2922066
57  -1.05800991 -1.0580099
58  -1.22409905 -1.2240991
59  -1.09881462 -1.0988146
60  -1.11046511 -1.1104651
61  -1.21622049 -1.2162205
62  -1.31109581 -0.7491121
63   0.62493916 -1.4203608
64  -1.48162766 -1.4816277
65  -0.97513460 -1.1431579
66  -0.77046648  0.9533110
67  -1.18331115  1.6172635
68  -1.36238133 -1.3623813
69  -1.42246873  1.5120828
70   2.98336106 -1.0413036
71  -1.21065485  2.4728911
72  -1.29024248  0.8428750
110 -0.30194169 -0.8053451
210 -1.18577169 -1.1857717
310  0.29532750 -0.2913221
```

```
> library(FactoMineR)
> pca <- PCA(datos_pca, ind.sup=73:75);pca
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 75 individuals, described by 20 variables
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$ind.sup"         "results for the supplementary individuals"
12 "$ind.sup$coord"   "coord. for the supplementary individuals"
13 "$ind.sup$cos2"    "cos2 for the supplementary individuals"
14 "$call"            "summary statistics"
15 "$call$centre"     "mean of the variables"
16 "$call$ecart.type" "standard error of the variables"
17 "$call$row.w"      "weights for the individuals"
```

```
18 "$call$col.w"      "weights for the variables"

> pca$var$coord[,1:2]
            Dim.1       Dim.2
x.2404 -0.5182234 -0.55871960
x.3129  0.7143372 -0.02965751
x.6    -0.4340978  0.73851953
x.918   0.7861929  0.06778430
x.979   0.7756237  0.40417768
x.973   0.3807579 -0.51672309
x.1182  0.8368622  0.28123192
x.3117  0.4987474  0.61252766
x.435  -0.5540988 -0.56635313
x.3199  0.4920537 -0.02585967
x.7    -0.5078035  0.64256817
x.3127  0.7935071 -0.04950105
x.1211  0.1044596  0.04454397
x.16   -0.5342648  0.47875922
x.927   0.7165126 -0.14192130
x.3389  0.1366313 -0.47906931
x.3116  0.5057214  0.55786602
x.3126  0.8430196 -0.07132648
x.8    -0.4943116  0.71359301
x.3327 -0.1901886  0.23737570
> cp <- pca$ind$coord[,1:2];cp
          Dim.1        Dim.2
1  -3.590624024  1.430748918
2  -0.120237851  1.922529484
3  -2.492507679  1.044005490
4  -3.635640826  0.322302784
5  -1.560680887 -0.539318328
6  -2.010800524  2.445578685
7  -3.964247808  2.353631274
8  -3.225838272  0.401024086
9  -2.505301175  2.759792420
10  0.041317969 -0.926364694
11 -2.223781719  1.417780461
12 -1.066682107  2.892820647
13 -2.172290112 -2.311003907
14 -1.334049470  1.170169546
15 -3.675970403  0.496697692
16 -1.117163683 -2.234473182
17 -0.573180164 -0.839664979
18 -1.886692725 -0.296010068
19 -1.298801552 -1.400130568
20 -1.472772533 -2.171230072
21 -1.809079838 -2.311242896
22 -3.983953323  1.800285906
23 -2.958546763  3.467511820
```

```
24 -2.484178485 -2.094577289
25 -2.840272390  2.432262134
26 -1.887659644 -2.732955888
27 -3.759632495  0.533306919
28  0.009922647  2.588447519
29  4.110957850  1.058665963
30  3.376100719  0.868737682
31  1.278542312  3.341556895
32  0.302067744  3.140747834
33  3.842020168  1.288536520
34  2.948162563  0.914059577
35 -0.664940402  1.618392518
36  3.222818781  1.204179739
37  2.773346052  3.128600094
38  0.998383793  2.510037810
39 -1.204197571 -1.701795343
40 -1.340668689 -1.678555224
41 -0.737045162 -2.362612709
42  1.308799950 -2.166826559
43  2.044176591 -2.797947356
44 -1.387125319 -1.701566901
45 -2.303564370 -1.718276415
46 -2.333334664 -1.594275893
47 -0.358035440 -1.304674570
48 -3.123983377 -0.978025322
49 -3.221904006  0.633053121
50  2.877723033  1.087755304
51  4.031690627  2.723965244
52  2.360125174  0.001330111
53  3.545834648  2.482612170
54  4.962703709 -1.818561841
55  1.991326743 -3.255656458
56  0.879937310 -3.856166894
57  5.331498287 -0.894518684
58  4.721499336 -0.194383099
59  2.217556344 -3.660404912
60  5.670002253 -1.804788070
61  2.507421856  0.311516605
62  2.645927962 -0.171787160
63  2.979964234  2.697739077
64  0.975737770 -0.739915311
65  4.536463269  0.762205914
66 -0.334298049  0.823254880
67  0.940436467  0.752813171
68 -0.375465266 -2.789246835
69 -0.842518172 -2.834173951
70 -1.214465247  0.515812815
71  0.295239002 -1.124731591
72 -0.635572977 -2.338605858
```

```
> cs <- pca$ind.sup$coord[,1:2];cs
        Dim.1      Dim.2
110  2.493860  1.4399978
210  3.050750 -2.5318588
310 -1.887403 -0.2720257
```

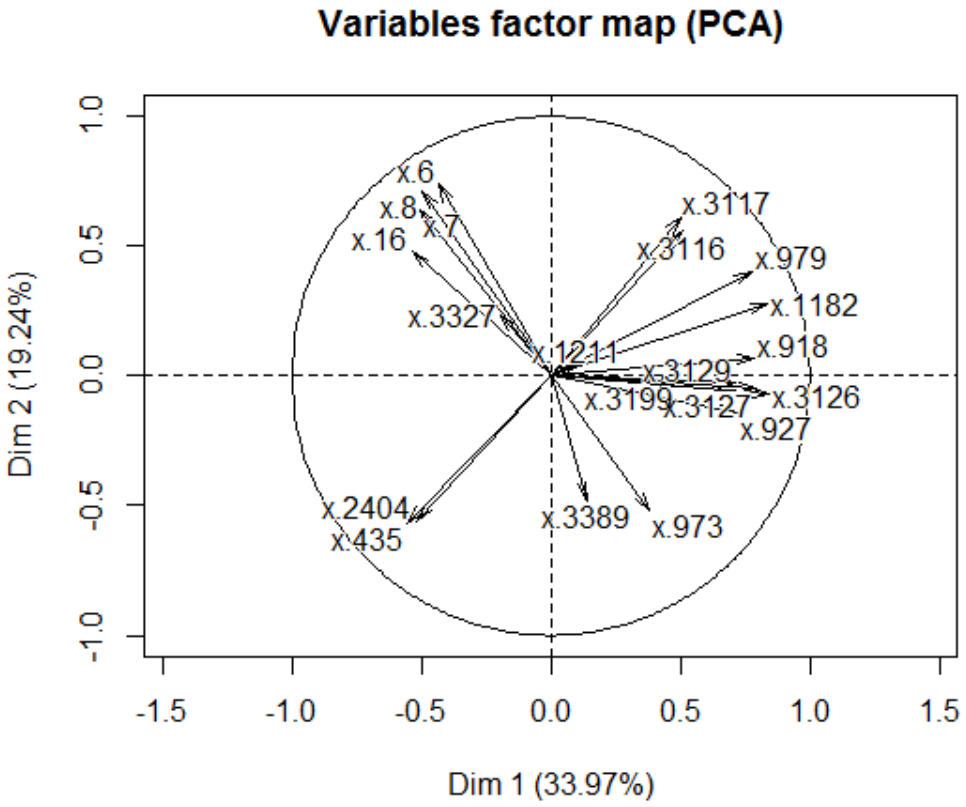


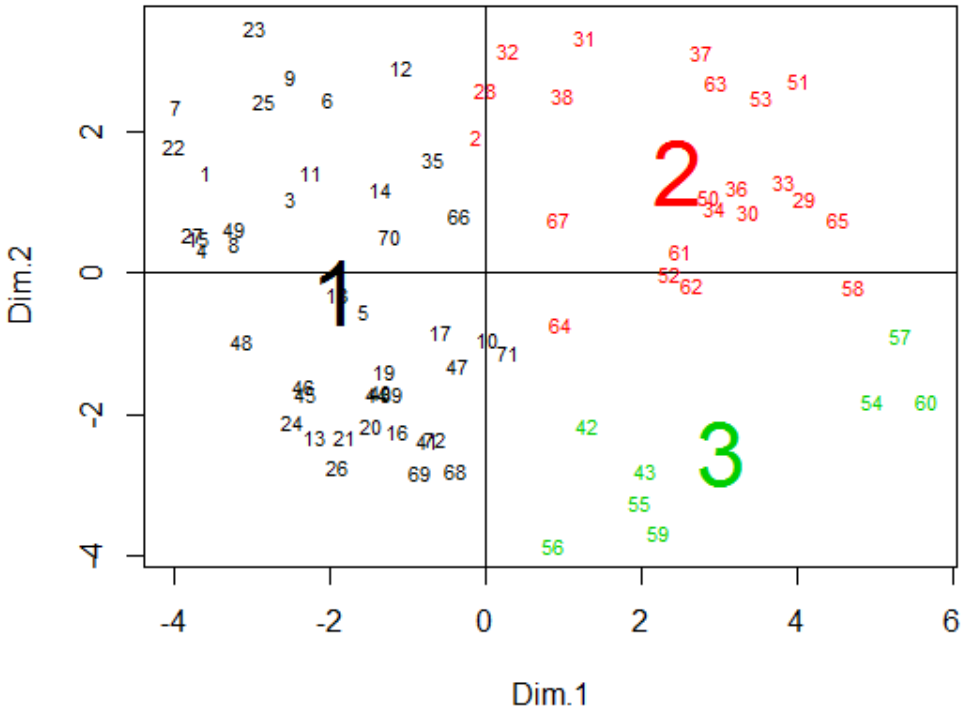
Figura 11: Componentes principales

Representación gráfica

```
> plot(cp,type="n")
> abline(h=0,v=0)
> text(cp, labels=row.names(cp),col=km$cluster,cex=0.7)
> text(cs, labels=row.names(km$centers),col=1:3,cex=3)
```

A continuación, una agrupación jerárquica se realiza en las dos primeras componentes principales. Nótese que las variables suplementarias no intervienen en el cálculo de distancias y por tanto, en el agrupamiento, pero pueden ser útiles para describir la clasificación.

La función *HCPC* proporciona la agrupación jerárquica sobre los resultados *pca* de la función *PCA*, teniendo en cuenta las dos primeras dimensiones mantenidas en la función *PCA* (*ncp=2*). La forma del dendrograma que se muestra en la figura 12 sugiere la solución óptima consiste en dividir en tres grupos y lo representa mediante una línea negra sólida. Nótese que hemos considerado que el número de clusters se encuentre entre 3 y 10, pero el usuario debe especificar el número de conglomerados, ya que por defecto *nb.clust=0*. En el caso, en que se especifique *nb.clust=-1* el número óptimo de



conglomerados queda fijado y *nb.clust* es un número entero. El argumento *conso=0* significa que no se utiliza ningún agrupamiento particional para consalidar la partición obtenida por el dendrograma.

```
library(FactoMineR)
pca <- PCA(datos_pca, ind.sup=73:75, scale.unit=T,ncp=2,graph=F);pca
hcpc <- HCPC(pca,nb.clust=0,conso=0,min=3,max=10)
```

Obsérvese que los individuos se clasifican según la primera componente principal, usando *order=T*. La clasificación en tres grupos junto con sus baricentros se representa en el mapa suministrado a partir de las dos primeras componentes principales y los individuos se colorean de acuerdo a su grupo, ver Figura 13.

Ejercicio 3.3 *A partir de la base de datos practica1.txt en la que se dispone de 11 individuos sobre los que se han observado dos variables ficticias X e Y, obtener la clasificación óptima de estos individuos en función de ambas variables, siguiendo los siguientes pasos:*

(1) Visualizar las observaciones mediante una representación gráfica previa.

(2) Explorar la matriz de distancias entre las observaciones.

(3) Obtener el análisis cluster jerárquico (considerar al menos dos o tres métodos diferentes).

(4) Representar la agrupación mediante un dendograma (o un bannerplot).

(5) ¿Encuentras diferencias entre usar un modelo u otro en relación a los dos apartados anteriores?

(6) Discutir las diferentes agrupaciones.

(7) Interpretar los resultados.

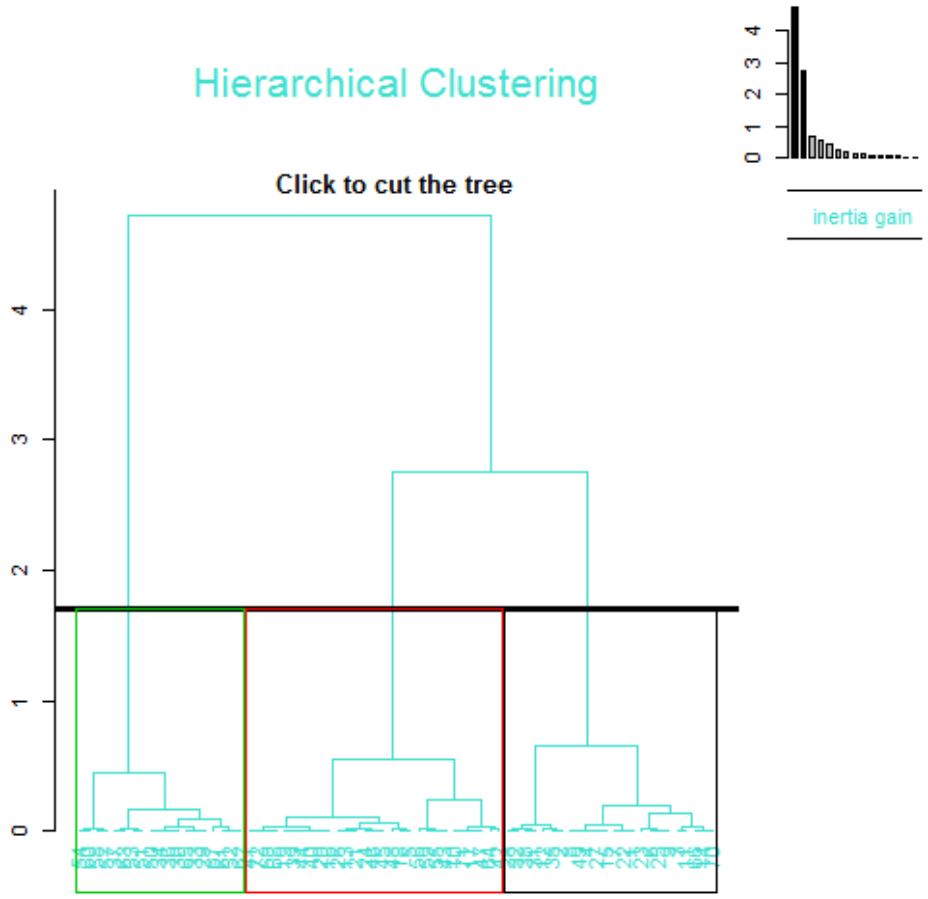


Figura 12: Dendrograma

4. Ejercicios: ACP y AC

A continuación, se facilita la descripción de algunas de las bases de datos que utilizaremos. Concretamente, las que se encuentran en alguno de los paquetes de R. Esta información es imprescindible tenerla en cuenta en el momento de llevar a cabo el correspondiente análisis estadístico y puede obtenerse mediante *help*.

Edgar Anderson's Iris Data

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

Available: iris {datasets}.

Ruspini Data

The Ruspini data set, consisting of 75 points in four groups that is popular for illustrating clustering techniques. A data frame with 75 observations on 2 variables giving the x and y coordinates of the points, respectively.

Available: ruspini {cluster} as data(ruspini).

Ejercicio 4.1 *Análisis Cluster Jerárquico*

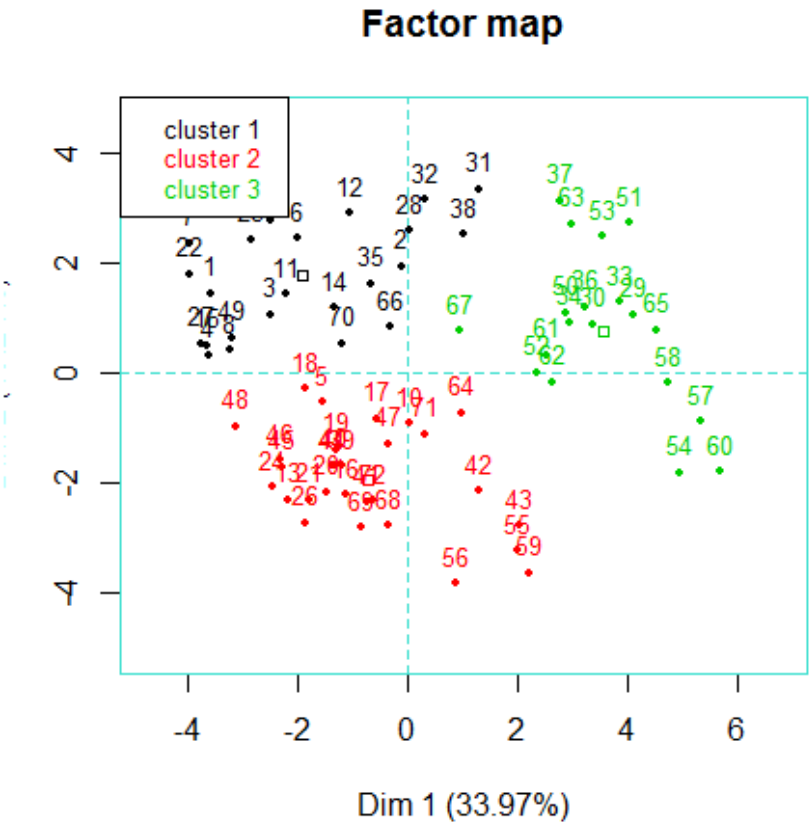


Figura 13: Dendrograma

(1) Leer los datos en R.

(2) Explorar mediante un análisis descriptivo los datos con los que estamos trabajando. Empezar con un resumen de estadísticos descriptivos y un gráfico de dispersión de las variables relevantes. A partir de estos resultados preliminares, contestar a las siguientes cuestiones:

i. *¿Hay algún tipo de relación entre las variables?*

ii. *¿Detectas la existencia de grupos de individuos?.*

(3) Encontrar clasificación jerárquica que muestre posibles similitudes entre clases de individuos, en base a las variables observadas. Sugerencia: Usar la función scale para estandarizar las variables. Obtener y comentar el correspondiente dendrograma.

(4) Comparar los resultados anteriores con los obtenidos mediante otro método asociativo diferente al anterior.

(5) A partir de los resultados de la clasificación, contestar las siguientes cuestiones:

i. *¿Número de conglomerados apropiado?*

ii. *¿Cuántos individuos están incluidos en cada uno de los coglomerados?*

(6) Representar de nuevo los datos mediante gráficos de dispersión identificando los conglomerados. Comentar la relación entre las variables que se visualiza incorporando esta información.

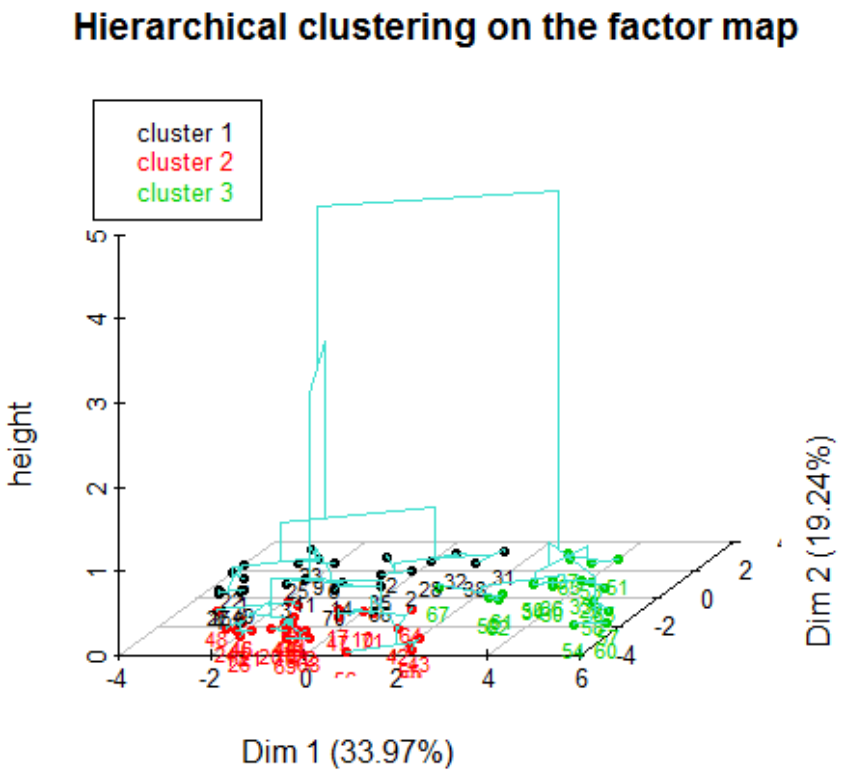


Figura 14: Dendrograma

Ejercicio 4.2 *Análisis Cluster No Jerárquico*

(1) Leer los datos en R.

(2) Explorar mediante un análisis descriptivo los datos con los que estamos trabajando. Empezar con un resumen de estadísticos descriptivos y un gráfico de dispersión de las variables relevantes. A partir de estos resultados preliminares, contestar a las siguientes cuestiones:

i. *¿Hay algún tipo de relación entre las variables?*

ii. *¿Detectas la existencia de grupos de individuos?.*

(3) Realizar el análisis cluster no jerárquico de las muestras utilizando el método k-means. Intentar obtener el correspondiente dendrograma, ¿problemas?

(4) A partir de los resultados de la clasificación, contestar las siguientes cuestiones:

i. *¿Cuántos individuos están incluidos en cada uno de los coglomerado?*

ii. *Los conglomerados de la partición obtenida ¿están bien diferenciados?. (Obtener la media, mínimo y máximo de los variables para cada conglomerado y comentar los resultados).*

(5) Representar de nuevo los datos mediante gráficos de dispersión identificando los conglomerados. Comentar la relación entre las variables que se visualiza incorporando esta información.

5. Escalamiento multidimensional

Multidimensional Scaling o escalamiento multidimensional (MDS) son técnicas de análisis multivariante que, como el Análisis de Componentes Principales, tienen el objetivo de hallar una configuración de puntos que los represente en un espacio de reducida dimensión, reteniendo al máximo la información inicial, cuando la información que disponemos de los n individuos es una matriz de distancias o proximidades.

Una representación espacial de dimensión k de una matriz de distancias D consiste en un conjunto de n puntos de $\mathbb{R}^k$, cuyas distancias reproduzcan lo mejor posible las originales. Concretamente, se trata de encontrar una matriz Y de n filas representando a los individuos y k columnas, denominadas *coordenadas principales* ($k << n$), de manera que las distancias euclídeas entre los puntos fila sean lo más próximas posibles a las distancias originales. El cálculo de las coordenadas principales, por orden de importancia, se centra en visualizar la estructura de interdependencias entre los individuosde mediante una representación gráfica bidimensional o tridimensional. El MDS también puede entenderse como una técnica complementaria al ACP, pues mientras el ACP investiga sobre las interrelaciones entre un conjunto de variables observadas a partir de una matriz de covarianzas, o de correlaciones de orden m, el MDS lo hace sobre la configuración o estructura de un conjunto de individuos desde una matriz de orden n de distancias o proximidades entre los mismos.

Las técnicas MDS pueden ser métricas y no métricas, dependiendo de la matriz de distancias D, si es una matriz de distancias o no lo es.

5.1. Escalamiento multidimensional métrico: Análisis de Coordenadas Principales

El Escalamiento multidimensional clásico de una matriz de datos, también conocido como análisis de coordenadas principales (Gower, 1966), se basa en la descomposición espectral de la matriz Q que reproduce la matriz de productos cruzados de filas y que puede interpretarse como la matriz de similitudes entre individuos. Dicha matriz Q de elementos q_{ij} puede ser obtenida a partir de la matriz de distancias D haciendo

$$q_{ij} = -\frac{1}{2}(d^2_{ij} - d^2_{i.} - d^2_{.j} + d^2_{..})$$

indicando por $d^2_{i.}$ la suma de la fila i, por $d^2_{.j}$ la suma de la columna j y por $d^2_{..}$ la suma de todos los elementos de la matriz D^2, d^2_{ij}.

Si la matriz Q es semidefinida positiva de rango m, podemos representar Q de la forma $Q = V\lambda V'$, donde V es la matriz formada por los m vectores propios asociados a los autovalores no nulos de Q, contenidos en la matriz diagonal λ. Si tomamos la matriz

$$Y = V\lambda^{\frac{1}{2}}$$

se obtiene una matriz de orden nxm con m variables centradas e incorreladas, cuyas distancias euclídeas entre los puntos fila reproducen la métrica inicial.

Por otro lado, la solución obtenida tiene la ventaja de que las coordenadas principales se ordenan en orden de importancia y podemos usar las k primeras para una representación sintética de la configuración de los individuos en un plano o en un espacio tridimensional. En este caso

$$Y_k = V_k\lambda_k^{\frac{1}{2}}$$

donde V_k es la matriz con los k vectores propios asociados a los k mayores propios de Q, es una representación espacial de dimensión $k < m$, que no reproduce exactamente la métrica inicial, pero tiene la característica de ser la que mejor representa la estructura de los puntos en dimensión k.

Este método puede aplicarse en los casos en que la matriz D es compatible con la métrica euclídea, para ello sera necesario comprobar que la matriz Q no tiene valores propios negativos. Habitualmente, el coeficiente de bondad de la representación, dado por

$$P_k = 100\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_m}$$

sugiere un ajuste razonable por encima del 80 %. Usualmente, también se valora la máxima diferencia entre las distancias observada y las reproducidas mediante la representación espacial de dimensión k.

En otro caso, esto es, si la matriz de distancias D no es compatible con la distancia euclídea, la matriz Q anterior tiene autovalores negativos. Si no hay un considerable número de tales autovalores y la magnitud de los mismos sea no significativa, es posible aplicar los métodos clásicos de MDS modificando la matriz D mediante una constante c que sumada a los elementos no diagonales la convierten a ésta en una matriz compatible con la métrica euclídea. Si la matriz Q posee autovalores negativos de magnitud apreciable es conveniente aplicar los métodos no métricos de MDS.

5.2. Escalamiento multidimensional no métrico

El análisis de proximidades, escalado no métrico tiene por objetivo construir una configuración de puntos de los que se dispone cierta información sobre sus parecidos o disimilaridades. Supongamos, ahora, que dicha información sobre disimilaridades entre individuos está recogida en la matriz simétrica Δ y la disimilaridad entre los individuos i y j es el término δ_{ij} .

En ciertas aplicaciones, se mide el grado de proximidad o similaridad entre cada par de individuos. Si S es la matriz de similaridades o proximidades entre individuos, podemos transformar ésta en una matriz de disimilaridades Δ haciendo:

$$\delta_{ij} = s_{ii} + s_{jj} - 2s_{ij}$$

El procedimento del MDS no métrico consiste en deformar las similaridades originales entre individuos δ_{ij} contenidas en una matriz Δ mediante una función monótona creciente, de forma que se conserven las relaciones de orden de proximidad entre los n individuos y que la matriz resultante D de elementos d_{ij} sea compatible con la distancia euclídea.

Así, fijada la dimensión k para la representación de los puntos, el algoritmo comienza con una configuración inicial de la que obtenemos las distancias euclídeas d_{ij} y éstas se relacionan con las originales mediante un modelo de regresión

$$d_{ij} = \varphi(\delta_{ij}) + \varepsilon$$

donde la transformación φ es monótona creciente (normalmente una función lineal o polinómica) y ε es un término de error. Si llamamos disparidades a las distancias ajustadas

$$\widehat{d}_{ij} = \varphi(\delta_{ij})$$

la configuración final es la que hace mínima la expresión $\sum_{i<j}(d_{ij} - \widehat{d}_{ij})^2$.

En este caso, la bondad de la representación se determina a través del stress, dado por:

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \widehat{d}_{ij})^2}{\sum_{i<j}\widehat{d}_{ij}^2}} \cdot 100\,\%$$

que se considera buena si $S < 5\,\%$.

5.3. Casos Prácticos

En este apartado, aplicaremos los técnicas de escalamiento multidimensional métrico y no métrico en distintos subapartados.

En el caso de la MDS métrico, usaremos la función *cmdscale* de package *stats*, a partir de una matriz de distancias entre pares de elementos o individuos, y obtendremos una representación en un espacio de baja dimensión, k.

```
cmdscale(d, k = 2, eig = FALSE, add = FALSE, x.ret = FALSE)

#d: matriz simétrica $D$ de distancias o disimilitudes.

#k: dimensión máxima del espacio de representación.

#eig: devuelve (TRUE) o no (FALSE) los autovalores.

#add: calcula (TRUE) o no (FALSE) la constante c.

#x.ret: devuelve (TRUE) o no (FALSE) la matriz doblemente centrada.
```

Observación 5.1 *Obsérvese que la matriz doblemente centrada* $\widehat{\widehat{D}}$ *verifica que* $Q = -\frac{1}{2}\widehat{\widehat{D}}$.

5.3.1. Ejemplos escalamiento métrico: la función cmsdcale

Distintos ejemplos ilustrarán la aplicación del MSD métrico. Principalmente, mostrando las siguiente situaciones:

- **Configuración de puntos en un espacio de reducida dimensión**.
- **Configuración de puntos a partir de una matriz de distancias**.

Ejemplo 5.1 *Consideramos diez puntos sobre un espacio de dimensión cinco. ¿Podemos encontrar un espacio de dimensión menor para estos diez puntos que reproduzcam los mejor posible los datos originales?*

A continuación realizamos un análisis de coordenadas principales sobre las coordenadas de los 10 puntos con el fin de obtener una representación de los mismo en un espacio tridimensional con una bondad de la representación del 81 %.

```
> x1 <- c(3,5,6,1,4,2,0,0,7,2)
> x2 <- c(4,1,2,1,7,2,4,6,6,1)
> x3 <- c(4,1,0,1,3,5,1,4,5,4)
> x4 <- c(6,7,2,0,6,1,1,3,1,3)
> x5 <- c(1,3,6,3,2,0,1,5,4,1)
> X <- data.frame(x1,x2,x3,x4,x5)
> D <- dist(X)
> cmds5 <- cmdscale(D,k=5,eig=T)
> cmds5$points
           [,1]        [,2]       [,3]       [,4]        [,5]
 [1,] -1.6038325  2.38060903 -2.2301092 -0.3656856  0.11536476
 [2,] -2.8246377 -2.30937202 -3.9523782  0.3419185  0.33169405
 [3,] -1.6908272 -5.13970089  1.2880306  0.6503227 -0.05133897
```

```
 [4,]  3.9527719 -2.43233961  0.3833746  0.6863995 -0.03460933
 [5,] -3.5984894  2.75538195 -0.2551393  1.0783741 -1.26125237
 [6,]  2.9520356  1.35475175 -0.1899027 -2.8211220  0.12385813
 [7,]  3.4689928  0.76411068  0.3016531  1.6369166 -1.94209512
 [8,]  0.3545235  2.31408566  2.2161772  2.9240116  2.00450379
 [9,] -2.9362323 -0.01279597  4.3117385 -2.5122743 -0.18911558
[10,]  1.9256952  0.32526941 -1.8734445 -1.6188611  0.90299062
> cmds5$eig
 [1]  7.518716e+01  5.880560e+01  4.960516e+01  3.042789e+01
 [5]  1.037419e+01  1.086006e-14  5.381235e-15 -5.316641e-15
 [9] -8.539862e-15 -1.050854e-14
> cmds5$x
NULL
> cmds5$ac
[1] 0
> cmds5$GOF
[1] 1 1
> cumsum(cmds5$eig)/sum(cmds5$eig)
 [1] 0.3350586 0.5971157 0.8181726 0.9537692 1.0000000 1.0000000
 [7] 1.0000000 1.0000000 1.0000000 1.0000000
> cmds3 <- cmdscale(D,k=3,eig=T)
> cmds3$points
            [,1]        [,2]       [,3]
 [1,] -1.6038325  2.38060903 -2.2301092
 [2,] -2.8246377 -2.30937202 -3.9523782
 [3,] -1.6908272 -5.13970089  1.2880306
 [4,]  3.9527719 -2.43233961  0.3833746
 [5,] -3.5984894  2.75538195 -0.2551393
 [6,]  2.9520356  1.35475175 -0.1899027
 [7,]  3.4689928  0.76411068  0.3016531
 [8,]  0.3545235  2.31408566  2.2161772
 [9,] -2.9362323 -0.01279597  4.3117385
[10,]  1.9256952  0.32526941 -1.8734445
> cmds3$eig
 [1]  7.518716e+01  5.880560e+01  4.960516e+01  3.042789e+01
 [5]  1.037419e+01  1.086006e-14  5.381235e-15 -5.316641e-15
 [9] -8.539862e-15 -1.050854e-14
> cmds3$x
NULL
> cmds3$ac
[1] 0
> cmds3$GOF
[1] 0.8181726 0.8181726
```

Ejemplo 5.2 *En este ejemplo, ejecutaremos MDS métrico desde la matriz de las distancias por carretera (en km) entre 21 ciudades de Europa. Disponible en: eurodist datasets.*

```
> require(graphics)
> library("datasets")
> data("eurodist")
```

```
>
> loc <- cmdscale(eurodist,k=2,eig=T,add=T)
> loc
$points
                      [,1]        [,2]
Athens          -2683.21958  3149.75394
Barcelona        1448.32785   734.87726
Brussels         -234.92183  -735.20436
Calais            -18.89811  -821.81942
Cherbourg         447.29935  -588.49619
Cologne          -610.07448  -765.02148
Copenhagen      -1284.52022 -1645.10194
Geneva             76.87490   433.22109
Gibraltar        3068.22174   586.65681
Hamburg         -1071.27454 -1217.42287
Hook of Holland  -463.20236 -1040.25930
Lisbon           2870.79381   -59.70814
Lyons             422.93917   289.39491
Madrid           2299.01430   262.98439
Marseilles        577.30944   611.87702
Milan            -346.38045   788.77859
Munich           -939.74443   256.49547
Paris             176.90762  -473.49514
Rome             -904.75548  2004.88469
Stockholm       -1505.31353 -2317.08258
Vienna          -1325.38318   544.68728

$eig
 [1]  4.227188e+07  2.953910e+07  9.553423e+06  8.377974e+06  6.299475e+06
 [6]  5.615219e+06  5.227436e+06  4.239826e+06  4.029548e+06  3.673937e+06
[11]  3.441726e+06  3.248488e+06  2.816048e+06  2.684931e+06  2.620954e+06
[16]  2.254576e+06  1.706717e+06  1.594186e+06  1.182005e+06  9.313624e-09
[21] -4.246525e-09

$x
NULL

$ac
[1] 2132.678

$GOF
[1] 0.5115564 0.5115564
```

Los valores propios y vectores propios de la matriz D. Las dos primeras coordenadas principales son también los dos primeros vectores propios de D. Obsérvese que la matriz D tiene un valor propio negativo, por el hecho de que las carreteras no van en línea recta. Pero como es inapreciable, podemos realizar la representación tomando las dos primeras coordenadas principales.

```
> cumsum(loc$eig)/sum(loc$eig)
 [1] 0.3011301 0.5115564 0.5796117 0.6392934 0.6841687 0.7241695 0.7614080
```

```
 [8] 0.7916110 0.8203161 0.8464879 0.8710056 0.8941467 0.9142072 0.9333338
[15] 0.9520045 0.9680653 0.9802234 0.9915798 1.0000000 1.0000000 1.0000000
```

El primer eje principal explica el 33.11 % de la variabilidad, y lo podemos entender como una dimensión que ordena las ciudades de Oeste-Este. El segundo eje principal explica 21.04 %. y se podría interpretar como una dimensión que ordena las ciudades de Norte-Sur.

```
> x <- loc$points[, 1]
> y <- -loc$points[, 2]
> plot(x, y, type = "n", xlab = "", ylab = "", asp = 1, axes = FALSE,
  + main = "cmdscale(eurodist)")
> text(x, y, rownames(loc$points), cex = 0.6)
> x1 <- -x
> y1 <- y
> plot(x1, y1, type = "n", xlab = "", ylab = "", asp = 1, axes = FALSE,
  + main = "cmdscale(eurodist)")
> text(x1, y1, rownames(loc$points), cex = 0.6)
>
```

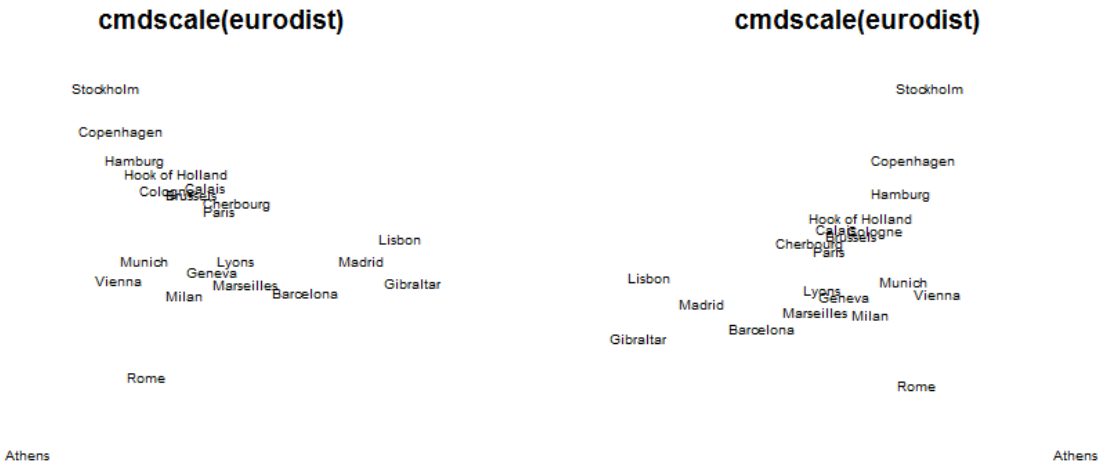


Figura 15: Representaciones por análisis de coordenadas principales y rotadas para eurodist

5.3.2. Ejemplos escalamiento no métrico: la función isoMDS

En este caso, utilizaremos la función de escalamiento multidimensional no métrico de Kruskal *isoMDS* disponible desde el paquete MASS

```
isoMDS (d, y = cmdscale (d, k), k = 2, maxit = 50, vestigios = TRUE,tol = 1e-3, p = 2)

#d: Matriz de disimilaridades D.
#y: Una configuración inicial.
#k: La dimensión deseada para la solución.
#maxit: El número máximo de iteraciones.
#rastro: Lógico para el rastreo de optimización.
#tol: tolerancia de convergencia.
```

```
#p: Alimentación para Minkowski distancia en el espacio de configuración.
#x: Una configuración final.
```

Observación 5.2 *La matriz R de correlaciones r_{ij} entre pares de variables puede considerarse como una matriz de similaridades entre las mismas. Mediante la transformación $d_{ij} = 1 - r_{ij}$ construimos una matriz de disimilaridades D.*

```
> library(MASS)
> data(swiss)
> summary(swiss)
   Fertility      Agriculture     Examination      Education
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
    Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60
> swiss
             Fertility Agriculture Examination Education Catholic
Courtelary        80.2        17.0          15        12     9.96
Delemont          83.1        45.1           6         9    84.84
Franches-Mnt      92.5        39.7           5         5    93.40
Moutier           85.8        36.5          12         7    33.77
Neuveville        76.9        43.5          17        15     5.16
Porrentruy        76.1        35.3           9         7    90.57
Broye             83.8        70.2          16         7    92.85
Glane             92.4        67.8          14         8    97.16
Gruyere           82.4        53.3          12         7    97.67
Sarine            82.9        45.2          16        13    91.38
Veveyse           87.1        64.5          14         6    98.61
Aigle             64.1        62.0          21        12     8.52
Aubonne           66.9        67.5          14         7     2.27
Avenches          68.9        60.7          19        12     4.43
Cossonay          61.7        69.3          22         5     2.82
Echallens         68.3        72.6          18         2    24.20
Grandson          71.7        34.0          17         8     3.30
Lausanne          55.7        19.4          26        28    12.11
La Vallee         54.3        15.2          31        20     2.15
Lavaux            65.1        73.0          19         9     2.84
Morges            65.5        59.8          22        10     5.23
Moudon            65.0        55.1          14         3     4.52
Nyone             56.6        50.9          22        12    15.14
```

Orbe	57.4	54.1	20	6	4.20
Oron	72.5	71.2	12	1	2.40
Payerne	74.2	58.1	14	8	5.23
Paysd'enhaut	72.0	63.5	6	3	2.56
Rolle	60.5	60.8	16	10	7.72
Vevey	58.3	26.8	25	19	18.46
Yverdon	65.4	49.5	15	8	6.10
Conthey	75.5	85.9	3	2	99.71
Entremont	69.3	84.9	7	6	99.68
Herens	77.3	89.7	5	2	100.00
Martigwy	70.5	78.2	12	6	98.96
Monthey	79.4	64.9	7	3	98.22
St Maurice	65.0	75.9	9	9	99.06
Sierre	92.2	84.6	3	3	99.46
Sion	79.3	63.1	13	13	96.83
Boudry	70.4	38.4	26	12	5.62
La Chauxdfnd	65.7	7.7	29	11	13.79
Le Locle	72.7	16.7	22	13	11.22
Neuchatel	64.4	17.6	35	32	16.92
Val de Ruz	77.6	37.6	15	7	4.97
ValdeTravers	67.6	18.7	25	7	8.65
V. De Geneve	35.0	1.2	37	53	42.34
Rive Droite	44.7	46.6	16	29	50.43
Rive Gauche	42.8	27.7	22	29	58.33

	Infant.Mortality
Courtelary	22.2
Delemont	22.2
Franches-Mnt	20.2
Moutier	20.3
Neuveville	20.6
Porrentruy	26.6
Broye	23.6
Glane	24.9
Gruyere	21.0
Sarine	24.4
Veveyse	24.5
Aigle	16.5
Aubonne	19.1
Avenches	22.7
Cossonay	18.7
Echallens	21.2
Grandson	20.0
Lausanne	20.2
La Vallee	10.8
Lavaux	20.0
Morges	18.0
Moudon	22.4
Nyone	16.7
Orbe	15.3

```
Oron                    21.0
Payerne                 23.8
Paysd'enhaut            18.0
Rolle                   16.3
Vevey                   20.9
Yverdon                 22.5
Conthey                 15.1
Entremont               19.8
Herens                  18.3
Martigwy                19.4
Monthey                 20.2
St Maurice              17.8
Sierre                  16.3
Sion                    18.1
Boudry                  20.3
La Chauxdfnd            20.5
Le Locle                18.9
Neuchatel               23.0
Val de Ruz              20.0
ValdeTravers            19.5
V. De Geneve            18.0
Rive Droite             18.2
Rive Gauche             19.3
> swiss.dist <- dist(swiss)
> swiss.mds <- isoMDS(swiss.dist)
initial  value 5.463800
iter   5 value 4.499103
iter   5 value 4.495335
iter   5 value 4.492669
final  value 4.492669
converged
> swiss.mds
$points
                   [,1]        [,2]
Courtelary    38.850496 -16.1546743
Delemont     -42.676573 -13.7209890
Franches-Mnt -53.587659 -21.3357627
Moutier        6.735536  -4.6041161
Neuveville    35.622307   4.6339724
Porrentruy   -44.739479 -25.4957015
Broye        -55.301247   2.9985892
Glane        -61.510950  -0.5029742
Gruyere      -56.196434 -11.5873817
Sarine       -47.880261 -18.4937959
Veveyse      -60.573600  -3.3177231
Aigle         28.500730  18.4040743
Aubonne       31.622253  26.0543764
Avenches      31.955939  19.3455733
Cossonay      32.951993  27.2866822
```

```
Echallens     11.653211  24.5294932
Grandson      39.623322  -0.1906417
Lausanne      40.455512 -24.2790922
La Vallee     51.099610 -23.2691859
Lavaux        30.753053  29.7236322
Morges        32.051544  18.1638440
Moudon        33.349605  17.2202105
Nyone         26.363999   7.9625625
Orbe          35.822440  15.4595563
Oron          29.301157  31.3756933
Payerne       30.448866  19.5104430
Paysd'enhaut  30.389346  26.4350474
Rolle         29.595391  18.6942289
Vevey         30.316991 -16.0544171
Yverdon       33.168755  11.4999792
Conthey      -67.045836  16.9000059
Entremont    -66.130908  14.2235838
Herens       -67.831773  19.3460319
Martigwy     -63.493801   8.8769860
Monthey      -59.675844  -1.3044352
St Maurice   -63.678801   7.2356724
Sierre       -69.462428  17.6354948
Sion         -57.385309  -4.8572223
Boudry        37.667244   0.0118818
La Chauxdfnd  40.842274 -29.0069374
Le Locle      38.285582 -17.6212453
Neuchatel     35.745340 -30.5746402
Val de Ruz    37.226824   2.1006842
ValdeTravers  41.086622 -15.3626392
V. De Geneve  24.329270 -73.1278621
Rive Droite   -4.756696 -17.5026420
Rive Gauche   -3.887613 -37.2642199

$stress
[1] 4.492669

> swiss.mds$points
                   [,1]        [,2]
Courtelary    38.850496 -16.1546743
Delemont     -42.676573 -13.7209890
Franches-Mnt -53.587659 -21.3357627
Moutier        6.735536  -4.6041161
Neuveville    35.622307   4.6339724
Porrentruy   -44.739479 -25.4957015
Broye        -55.301247   2.9985892
Glane        -61.510950  -0.5029742
Gruyere      -56.196434 -11.5873817
Sarine       -47.880261 -18.4937959
Veveyse      -60.573600  -3.3177231
```

```
Aigle         28.500730  18.4040743
Aubonne       31.622253  26.0543764
Avenches      31.955939  19.3455733
Cossonay      32.951993  27.2866822
Echallens     11.653211  24.5294932
Grandson      39.623322  -0.1906417
Lausanne      40.455512 -24.2790922
La Vallee     51.099610 -23.2691859
Lavaux        30.753053  29.7236322
Morges        32.051544  18.1638440
Moudon        33.349605  17.2202105
Nyone         26.363999   7.9625625
Orbe          35.822440  15.4595563
Oron          29.301157  31.3756933
Payerne       30.448866  19.5104430
Paysd'enhaut  30.389346  26.4350474
Rolle         29.595391  18.6942289
Vevey         30.316991 -16.0544171
Yverdon       33.168755  11.4999792
Conthey      -67.045836  16.9000059
Entremont    -66.130908  14.2235838
Herens       -67.831773  19.3460319
Martigwy     -63.493801   8.8769860
Monthey      -59.675844  -1.3044352
St Maurice   -63.678801   7.2356724
Sierre       -69.462428  17.6354948
Sion         -57.385309  -4.8572223
Boudry        37.667244   0.0118818
La Chauxdfnd  40.842274 -29.0069374
Le Locle      38.285582 -17.6212453
Neuchatel     35.745340 -30.5746402
Val de Ruz    37.226824   2.1006842
ValdeTravers  41.086622 -15.3626392
V. De Geneve  24.329270 -73.1278621
Rive Droite   -4.756696 -17.5026420
Rive Gauche   -3.887613 -37.2642199
> summary(swiss.mds$points)
       V1               V2
 Min.   :-69.46   Min.   :-73.128
 1st Qu.:-54.44   1st Qu.:-16.105
 Median : 29.30   Median :  2.101
 Mean   :  0.00   Mean   :  0.000
 3rd Qu.: 34.49   3rd Qu.: 17.900
 Max.   : 51.10   Max.   : 31.376
> #plot(swiss.mds$points, type = "n")
> #text(swiss.mds$points, labels = as.character(1:nrow(swiss)))
> plot(swiss.mds$points, type = "n")
> segments(-75, -0, 55, 0, lty="dotted")
> segments(0, -75, 0, 35, lty="dotted")
```

```
> text(swiss.mds$points, labels = row.names(swiss), col = "red",cex=0.7)
```

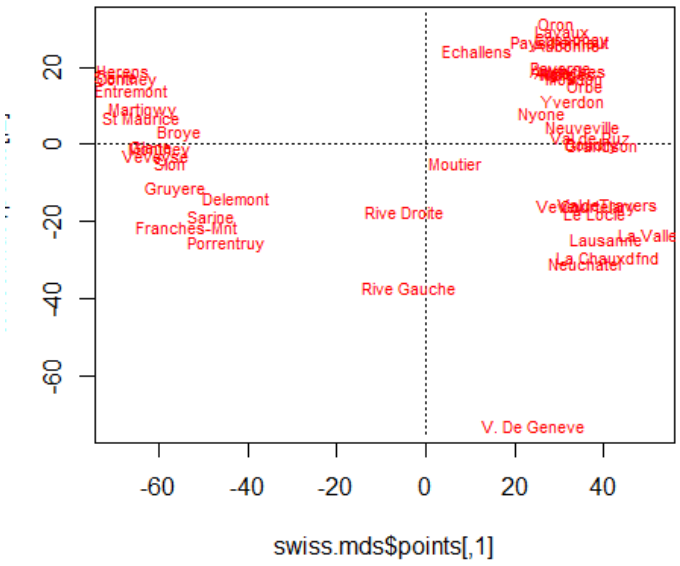


Figura 16: Representaciones por MDS no métrico para swiss

6. Análisis de correspondencias

El análisis de correspondencias es una técnica descriptiva de análisis multivariante de tablas de contingencia que describe la relación existente entre las categorías de las filas y las columnas en un espacio de dimensión reducida, y por tanto, la relación existente entre dos o más variables cualitativas nominales. Así, las distancias sobre un gráfico entre los *puntos* de categorías reflejan las relaciones entre las categorías, mayor proximidad mayor similitud.

A través del análisis de correspondencias obtenemos medidas de correspondencia, perfiles de fila y de columna, valores propios, puntuaciones de fila y de columna, inercia, masa, estadísticos de confianza para las puntuaciones de fila y de columna, estadísticos de confianza para los valores propios, gráficos de transformación, gráficos de los puntos de fila, gráficos de los puntos de columna y diagramas de dispersión biespaciales.

6.1. Independencia

Dos variables aleatorias, X e Y, son independientes si

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \tag{6.1}$$

para todo i, j.

En el caso de una tabla de contingencia, si se aproxima la probabilidad de que sucedan x_i e y_j como la frecuencia relativa en un experimento con N número de casos posibles (regla de Laplace), entonces:

$$p_{ij} = \frac{n_{ij}}{n_{..}} p_{i.} = \frac{n_{i.}}{n_{..}} p_{.j} = \frac{n_{.j}}{n_{..}}$$

Equivalentemente, (6.1) puede expresarse como:

$$P(X = x_i, Y = y_j) = p_{ij} = p_{i.}p_{.j}$$

para todo i, j, las variables X e Y son independientes y la tabla es homogénea. Así, bajo la hipótesis de independencia, la frecuencia esperada:

$$e_{ij} = n_{..}f_{ij} = n_{..}p_{i.}p_{.j} = \frac{n_{i.}n_{.j}}{n_{..}}$$

El contraste o test de la chi-cuadrado mide si las diferencias entre las frecuencias observadas y las frecuencias esperadas bajo la hipótesis de independencia, son estadísticamente significativas. El estadístico asociado al constraste se define como sigue:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Habitualmente, se usa este contraste de independencia en tablas de contingencia.

6.2. Distancia chi-cuadrado

En general, una tabla de correspondencia o de contingencia donde hay r filas y c columnas se puede expresar a través de sus frecuencias absolutas conjunta y como en este caso, completar con las distribuciones de frecuencias marginales

n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..}$

A partir de la cual se derivan las distribuciones de frecuencias condicionadas, denominadas tablas de perfiles fila y perfiles columna:

$p_{11} = \frac{n_{11}}{n_{1.}}$	$p_{12} = \frac{n_{12}}{n_{1.}}$	$\cdots$	$p_{1c} = \frac{n_{1c}}{n_{1.}}$
$p_{21} = \frac{n_{21}}{n_{2.}}$	$p_{22} = \frac{n_{22}}{n_{2.}}$	$\cdots$	$p_{2c} = \frac{n_{2c}}{n_{2.}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p_{r1} = \frac{n_{r1}}{n_{r.}}$	$p_{r2} = \frac{n_{r2}}{n_{r.}}$	$\cdots$	$p_{rc} = \frac{n_{rc}}{n_{r.}}$

$q_{11} = \frac{n_{11}}{n_{.1}}$	$q_{12} = \frac{n_{12}}{n_{.2}}$	$\cdots$	$q_{1c} = \frac{n_{1c}}{n_{.c}}$
$q_{21} = \frac{n_{21}}{n_{.1}}$	$q_{22} = \frac{n_{22}}{n_{.2}}$	$\cdots$	$q_{2c} = \frac{n_{2c}}{n_{.c}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$q_{r1} = \frac{n_{r1}}{n_{.1}}$	$q_{r2} = \frac{n_{r2}}{n_{.2}}$	$\cdots$	$q_{rc} = \frac{n_{rc}}{n_{.c}}$

Nótese que la distancia chi-cuadrado entre dos columnas i y j (similar entre dos filas i y j) dada por:

$$d_{ij}^{\text{col}} = \sum_{k=1}^{r}\frac{1}{p_{k.}}(p_{ki} - p_{kj})^2$$

$(d_{ij}^{fil} = \sum_{k=1}^{c}\frac{1}{q_{.k}}(q_{ik} - q_{jk})^2)$

donde $p_{k.} = \frac{n_{k.}}{n_{..}}$ $(p_{.k} = \frac{n_{.k}}{n_{..}})$ puede considerarse como una distancia euclídea ponderada basada en las

proporciones de las columnas (filas), que será igual a cero si las dos columnas (filas) tienen los mismos valores para esas proporciones. También, puede observarse que las diferencias al cuadrado anteriores están multiplicadas o ponderadas mediante el factor $\frac{1}{p_k}$, de modo que las categorías de la variable que están en la columna (fila) con pocos valores tienen una mayor influencia en el cálculo de la distancia que las categoría comunes.

La distancia chi-cuadrado cumple la propiedad de equivalencia distribucional: Si dos categorías de los perfiles fila tienen el mismo valor de perfil, entonces al agruparlas en una única categoría no se modifican las distancias entre el resto de categorías de la tabla que forman las columnas. Análogamente, si se juntan o separan columnas, esto no afecta a las distancias entre los perfiles fila.

La masa de una fila o una columna de una tabla de correspondencia es la proporción de observaciones de la fila (o columna) respecto al total de observaciones (n_i/n).

El perfil medio de las filas (la fila media de perfiles) es el centroide de los perfiles fila cuando se calcula la media ponderando cada perfil por su masa. Todo esto mismo, obviamente, se puede considerar para las columnas.

La inercia total de la tabla de contingencia viene dada por: $\frac{\chi^2}{N}$ y se interpreta como la media ponderada de las distancias chi-cuadrado entre los perfiles fila y su perfil medio. Alternativamente, se puede definir para los perfiles columna.

6.3. Reducción de dimensiones

En general, los perfiles están situados en espacios de altas dimensiones de modo que no se pueden observar directamente. Se pueden determinar subespacios de dimensión menor al número mínimo entre filas y columnas menos uno, donde se puede aproximar la posición original de los perfiles. La calidad de representación en subespacios de dimensión menor se mide en porcentajes de inercia con respecto a la total.

El cálculo matemático de los subespacios se basa en minimizar las sumas de las distancias entre los perfiles y el subespacio, ponderadas por las masas de los puntos. Es decir, se calcula por el método de los mínimos cuadrados ponderados. Se pueden proyectar perfiles fila y perfiles columna de modo equivalente en el subespacio extraído.

Una manera de hacer lo anterior es mediante una aplicación directa del multidimensional scaling (MDS) en cada matriz de distancias (por filas o por columnas). Luego, se consideran y se dibujan las dos primeras coordenadas para las categorías de las filas y de las columnas en la misma gráfica etiquetadas de modo conveniente para que se puedan distinguir ambas variables.

Cuando las coordenadas de las categorías de ambas variables son grandes y positivas se deduce una asociación positiva entre las columnas y las filas correspondientes. Del mismo modo se razona en el caso de coordenadas negativas. La conclusión es que los valores de la tabla n_{ij} son mayores que los esperados bajo la hipótesis de independencia entre ambas variables. Cuando las coordenadas de las categorías de ambas variables son grandes en valor absoluto, pero tienen signos opuestos las filas y columnas correspondientes tienen asociación negativa; así los valores de la tabla n_{ij} son menores que los esperados bajo la hipótesis de independencia entre ambas variables. Finalmente, cuando el producto de las coordenadas está próximo a 0, la asociación entre las variables es baja, de modo que n_{ij} se encuentra cerca del valor esperado bajo la hipótesis de independencia.

6.4. Caso Práctico

En un estudio sobre la enfermedad de Hodgkin, un cáncer de los nodos linfáticos, cada uno de los 538 pacientes con la enfermedad fue clasificado según el *Tipo de histología* y por su *Respuesta al tratamiento* después de tres meses de iniciado éste. Los "valores"de la variable Histología considerados fueron, Predominancia de Linfocitos (PL), Esclerosis Nodular (EN), Celularidad Mixta (CM) y Agotamiento

de los Linfocitos (AL). Realizar un análisis de correspondencias con los datos recogidos que se muestran en la siguiente tabla:

Respuesta / Histología	Positiva	Parcial	Ninguna	$n_{i.}$
PL	74	18	12	104
EN	68	16	12	96
CM	154	54	58	266
AL	18	10	44	72
$n_{.j}$	314	98	126	538

Nuestro objetivo se centra en ver si existe algún tipo de relación entre la respuesta al tratamiento y el tipo de histología, i.e., si el tratamiento es más eficaz con algún tipo de tumor. Por lo que realizaremos un análisis de correspondencias.

```
> X1 <- c(74,68,154,18)
> X2 <- c(18,16,54,10)
> X3 <- c(12,12,58,44)
> X <- data.frame(X1,X2,X3)
> X
   X1 X2 X3
1  74 18 12
2  68 16 12
3 154 54 58
4  18 10 44
> rownames(X) <- c("PL","EN","CM","AL")
> colnames(X) <- c("positiva","parcial","ninguna")
> X
   positiva parcial ninguna
PL       74      18      12
EN       68      16      12
CM      154      54      58
AL       18      10      44
> chisq.test(X)

Pearson's Chi-squared test

data:  X
X-squared = 75.8901, df = 6, p-value = 2.517e-14
```

A la vista de los resultados del test de independencia de la χ^2, $p - value <<< 0.05$, se concluye que rechazamos la hipótesis nula de independencia entre ambas variables.

Ejecutamos sobre la tabla de contingencia el análisis de correspondencias utilizando la función *ca* del package *ca*

```
> ca(X, nd=d)

#obj, formula: matrices, data.frame, "xtabs" o "table" y "~ F1 + F2", con F1 y F2
  factores.
#nd: Número de dimensiones que se incluye en la salida; si NA se incluyen las
  dimensiones máximas posibles.
```

La salida contiene los valores propios y los porcentajes de inercia explicada para todas las dimensiones posibles. Además, los valores de las filas y columnas (masas, distancias chi-cuadrado de puntos a su promedio, inercias y coordenadas estándar). Sin embargo, estos valores están restringidos a dos dimensiones.

```
>library(ca)
> ca(X)

 Principal inertias (eigenvalues):
           1       2
Value      0.13839 0.00267
Percentage 98.11%  1.89%


 Rows:
               PL        EN        CM        AL
Mass     0.193309  0.178439  0.494424  0.133829
ChiDist  0.297921  0.280844  0.059490  0.898658
Inertia  0.017158  0.014074  0.001750  0.108078
Dim. 1  -0.790844 -0.736320 -0.078243  2.413154
Dim. 2  -0.908411 -1.199689  1.004090 -0.797822


 Columns:
         positiva   parcial   ninguna
Mass     0.583643  0.182156  0.234201
ChiDist  0.247372  0.125697  0.661451
Inertia  0.035715  0.002878  0.102467
Dim. 1  -0.660941 -0.167593  1.777457
Dim. 2  -0.525864  2.112276 -0.332396

#Mass: Importancia relativa de la modalidad condicionante (frecuencia marginal).
?#ChiDist: Distancia Chi-Cuadrado del correspondiente perfil al perfil medio.
?#Inertia: Contribución del perfil a la inercia total.
?#Dim. 1: Coordenada en el primer eje.
?#Dim. 2: Coordenada en el segundo eje.

> summary(ca(X))

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.138390  98.1  98.1  ************************
 2      0.002670   1.9 100.0
        -------- -----
 Total: 0.141060 100.0


Rows:
```

```
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   PL |  193 1000  122 | -294 975 121 |  -47  25 160 |
2 |   EN |  178 1000  100 | -274 951  97 |  -62  49 257 |
3 |   CM |  494 1000   12 |  -29 239   3 |   52 761 498 |
4 |   AL |  134 1000  766 |  898 998 779 |  -41   2  85 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | pstv |  584 1000  253 | -246 988 255 |  -27  12 161 |
2 | prcl |  182 1000   20 |  -62 246   5 |  109 754 813 |
3 | nngn |  234 1000  726 |  661 999 740 |  -17   1  26 |


?#name: Modalidad condicionante.
?#mass: Frecuencia marginal.
?#qlt: Calidad de la representación del perfil.
?#k1: Coordenada del perfil en el primer eje.
?#k2: Coordenada del perfil en el segundo eje.
?#cor: Contribución relativa del correspondiente factor
para representar el perfil.
?#ctr: Contribución absoluta del perfil para la construcción
del correspondiente factor.

> plot(ca(X))
```

Inercia Total mide el grado de dependencia existente entre las variables X e Y . A partir de ella se calculan las proporciones de inercia explicada por cada uno de los factores usados en la representación. Contribuciones absolutas de las modalidades miden la importancia de cada una de las modalidades de las variables analizadas en la construcción de los ejes factoriales obtenidos por el Análisis de Correspondencias. Contribuciones relativas de los factores miden la importancia de cada factor para explicar la posición, en el diagrama cartesiano, de cada una de las modalidades de las variables analizadas, representando la parte de la distancia al origen de coordenadas, explicada por dicho factor. Los resultados obtenidos, muestran que elegir un par de coordenadas es adecuado, ya que recoge el 100 % de la inercia total. Obsérvese que con una sola coordenada se obtiene el 98.11 % de la inercia. Considerando como coordenadas unidimensionales, la proyecciones sobre el eje de abcisas de las coordenadas representadas en la Figura (17), nos permiten tener en cuenta las cercanías entre los valores de las dos variables. Así, tenemos que -0.790844 y -0.736320 están cercanas a -0.660941 y que -0.078243 está cerca de -0.167593 y por último, que 2.413154 y 1.777457 pueden considerarse cercanas.

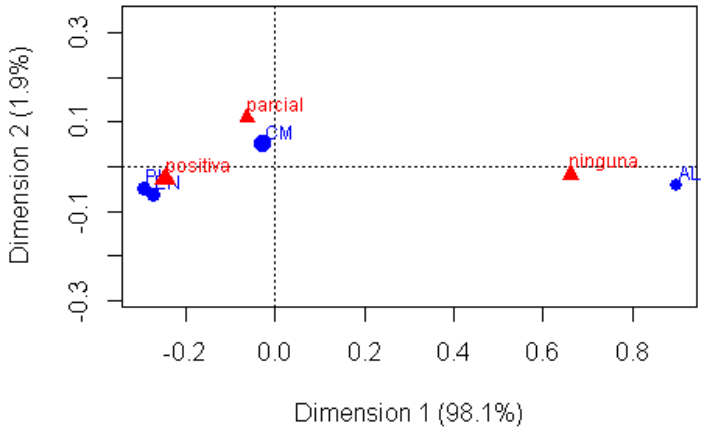


Figura 17: Relaciones entre las diferentes Histologías y la efectividad del Tratamiento

7. Ejercicios: Escalamiento Multidimensional y Análisis de Correspondencias

Ejercicio 7.1 *Considere las siguientes distancias entre nueve ciudades de Estados Unidos. ¿Es posible representar estas ciudades en un espacio bidimensional?.*

	BOS	*CHI*	*DC*	*DEN*	*LA*	*MIA*	*NY*	*SEA*	*SF*
BOS	0								
CHI	963	0							
DC	429	671	0						
DEN	1949	996	1616	0					
LA	2979	2054	2631	1059	0				
MIA	1504	1329	1075	2037	2687	0			
NY	206	802	233	1771	2786	1308	0		
SEA	2976	2013	2684	1307	1131	3273	2815	0	
SF	3095	2142	2799	1235	379	3053	2934	808	0

Ejercicio 7.2 *En 1976, Nanny Wermuth presentó un trabajo sobre 6851 nacimientos, incluyendo las dos variables siguientes:*

(1) Madre = Características de la madre, con cuatro categorías:

- *jnf = madre joven que no fumó durante el embarazo*
- *jf = madre joven que fumó durante la gestación*
- *mnf = madre mayor que no fumó durante la gestación*
- *mf = madre mayor que fumó durante la vestación*

(2) Bebe = Estado del bebé, con cuatro categorías:

- *pm = prematuro que murió antes de finalizar el primer año*
- *pv = prematuro que vivió al menos el primer año*

- *gcm = gestación completa que murió antes de finalizar el primer año*
- *gcv = gestación completa que vivió al menos el primer año*

La tabla de nacimientos contados en cada categoría se puede ver a continuación:

	pm	*pv*	*gcm*	*gcv*
jnf	50	315	24	4012
jf	9	40	6	459
mnf	41	147	14	1594
mf	4	11	1	124

- *Analizar si están relacionadas las características de la madre con el estado del bébe*
- *En caso afirmativo, estudiar cómo están relacionadas.*

Sugerencia: El primer punto requiere la aplicación de un test Chi-cuadrado y el segundo mediante un Análisis de Correspondencias.

Ejercicio 7.3 *Realizar en MSD no métrico con la matriz de eurodist.*

Ejercicio 7.4 HairEyeColor*, disponible en el paquete* datasets *contiene la distribución de color del pelo y de los ojos por sexo para 592 estudiantes de estadística. Realizar un análisis de correspondencias.*

Referencias

- Ayala, G. (2022). Bioinformática Estadística. Estadística de datos ómicos. https://www.uv.es/ayala/docencia/tami/
- Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press: New York.
- Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth and BrooksCole.
- Cox, T. F. and Cox, M. A. A. (2001). Multidimensional Scaling. Second edition. Chapman and Hall.
- Crystal, D. Ed. (1990) The Cambridge Encyclopaedia. Cambridge: Cambridge University Press.
- Cuadras, C. M. (2014). Nuevos Métodos de Análisis Multivariante.CMC Editions, Barcelona.
- Everitt, B. (1974). Cluster Analysis. London: Heinemann Educ. Books.
- Everitt, B. and Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media, 2011.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (2), 179-188.
- Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21,768-769.
- Franco, M. and Vivo, J.M. (2019). Cluster Analysis of Microarray Data. In: Bolón V. and Alonso A. (eds) Microarray Bioinformatics. Methods in Molecular Biology, vol 1986. Chapter 7, 153-183. Springer. https://link.springer.com/protocol/10.1007/978-1-4939-9442-7_7
- Gordon, A. D. (1999). Classification. Second Edition. London: Chapman and Hall / CRC.
- Gower, J. C. and Hand, D. J. (1996). Biplots. Chapman & Hall.
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-328.
- Greenacre, M. (2008). La práctica del análisis de correspondencias. Fundación BBVA. https://www.fbbva.es/wpcontent/uploads/2017/05/dat/DE_2008_practica_analisis_correspondencias.pdf
- Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28,100-108.
- Lê, S., Josse J. and Husson, F. (2008). FactoMineR: An R Packge for Multivariate Analysis. Journal of Statistical Software, 25.

- Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory 28,128-137.

- MacLean, D. (2019) R Bioinformatics Cookbook. Packt.
https://www.packtpub.com/product/rbioinformaticscookbook/9781789950694
https://github.com/PacktPublishing/RBioinformaticsCookbook

- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam and J. Neyman, 1, pp.281-297. Berkeley, CA: University of California Press.

- McNeil, D. R. (1977). Interactive Data Analysis. New York: Wiley.

- Nenadic, O. and Greenacre, M. (2007). Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. Journal of Statistical Software, 20.

- McQuitty, L.L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. Educational and Psychological Measurement, 26, 825-831.

- Peña, D. (2013). Análisis de datos multivariantes. Cambridge: McGraw-Hill España.
https://www.researchgate.net/profile/Daniel-Pena/publication/40944325_Analisis_de_Datos_Multivariantes/

- Torgerson, W. S. (1958). Theory and Methods of Scaling. New York: Wiley.

- Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer.