# Ejercicio 1

A) Cuáles son los valores de $n$ y $p$? Cuanto vale y que indica el valor $x_{32}$ ? Y el vector $\mathbf{x}_6$ ?

El valor de $n$ es 18 y el valor de $p$ es 13. El valor de $x_{32}$ es 12.68 e indica el tamaño de la flor de la observacion 3. El vector $\mathbf{x}_6$ indica la relacion entre el ancho y largo de la hoja.

B) Cómo clasificaría las variables sobre las que se está trabajando ?

Las variables con las que se esta trabajando son de tipo continuo en todos los casos. Mas aun, todas estan medidas en escala de intervalo.

C) Encuentre el vector de medias y matriz de varianzas-covariancias asociados a la tabla de datos.

Las medias son

| TAMFLOR | LONGPET | ANCHOPET | SUPHOJA | LONANCHO | PECLIMBO | PESOF |
|---|---|---|---|---|---|---|
| 28.9 | 12.9 | 15.5 | 41 | 1 | 0.4 | 46.7 |

| LONGF | ANCHOF | ESPESORF | PESOEND | LONGEND | ANCHOEND |
|---|---|---|---|---|---|
| 42 | 43.7 | 43.9 | 2.4 | 22 | 19.7 |

Calculamos la matriz de covarianza de la siguiente manera

```
matriz_covarianza <- cov(datos)
```

Pero solo mostramos aquellos casos de mayor covarianza, ya que la tabla es muy grande como para incluirla en formato pdf.

| Variable 1 | Variable 2 | Covarianza |
|---|---|---|
| LONANCHO | PECLIMBO | -0.001 |
| PECLIMBO | PESOEND | 0.003 |
| PECLIMBO | ANCHOF | -0.004 |
| PECLIMBO | LONGEND | -0.005 |
| PECLIMBO | ANCHOEND | -0.005 |

| Variable 1 | Variable 2 | Covarianza |
|---|---|---|
| LONGF | ANCHOF | 26.314 |
| PESOF | LONGEND | 35.033 |
| PESOF | ESPESORF | 49.700 |
| PESOF | ANCHOF | 57.539 |
| PESOF | LONGF | 67.532 |

D) Podría decir cuál y cuáles variables son las más dispersas ?

Utilizando el coeficiente de variación, podemos decir que las variables más dispersas, en orden decreciente, son:

| PESOEND | PESOF | LONGEND | LONGF | SUPHOJA | ANCHOEND | ANCHOPET |
|---|---|---|---|---|---|---|
| 0.42 | 0.26 | 0.21 | 0.14 | 0.14 | 0.13 | 0.11 |

| ANCHOF | TAMFLOR | ESPESORF | PECLIMBO | LONGPET | LONANCHO |
|---|---|---|---|---|---|
| 0.11 | 0.11 | 0.1 | 0.09 | 0.08 | 0.06 |

Por lo tanto, podemos concluir que la variable mas dispersa es el peso del endocarpio, seguido por el peso de la flor. Si hubieramos observado otra medida que depende de la escala de medicion, como por ejemplo el desvio estandar, no hubieramos incluido al peso del endocarpio ya que su valor medio (2.4) es mucho mas bajo que el valor medio de otras variables, como por ejemplo tamaño de la flor, largo de la flor, etc.

A) Estandarice las variables por media y desvío. Ahora puede responder al inciso (d) ?

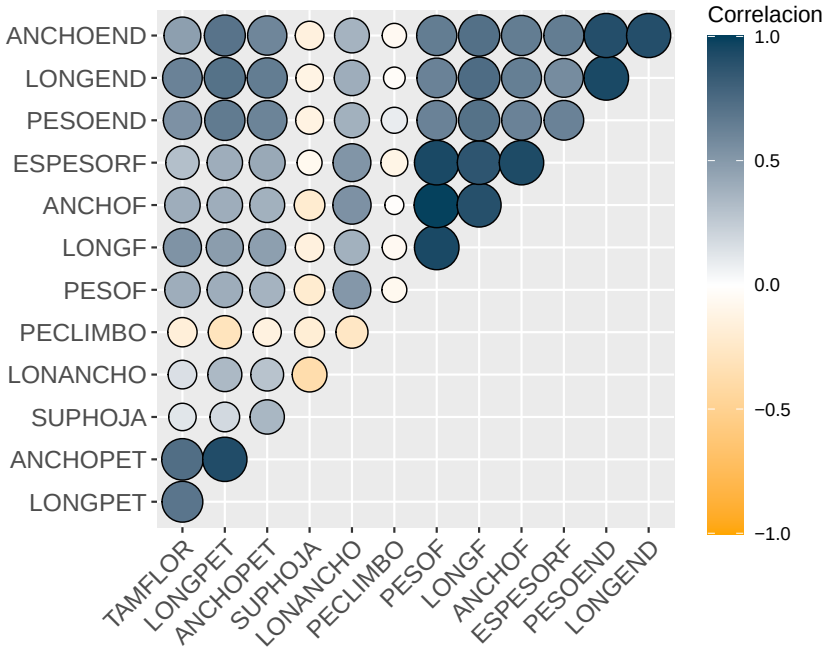
El coeficiente de variacion no existe ya que todas las variables tienen media igual a 0, por lo que no podriamos responder al inciso (d) luego de la estandarizacion. Sin embargo, lo expuesto en el inciso (d) es suficiente para concluir sobre que variables tienen mayor y menor dispersion.

E) Halle la matriz de correlación. Que variables son las más relacionadas?

La martiz de correlacion es obtenida con la funcion `cor()`.

```
matriz_correlacion <- cor(datos)
```

A continuación una representacion grafica de la matriz de correlacion, la cual permite identificar de forma mas sencilla las variables mas relacionadas:




F) Pueden dividirse las variables en subgrupos, de modo que las variables dentro de un mismo subgrupo tengan elevadas correlaciones entre sí y que las que se encuentren en subgrupos diferentes tengan bajas correlaciones? Si es así, cuáles variables quedan en cada uno de los subgrupos?

- Los subgrupos de variables con altas correlaciones son los siguientes:

1. Peso, longitud, ancho y espesor del fruto (caracteristicas del fruto)
2. Peso, longitud y ancho del endocarpio (caracteristicas del endocarpio)
3. Tamaño de la flor, longitu y ancho del petalo (caracteristicas de la flor)

- El subgrupo de variables con bajas correlaciones son los siguientes:

1. Superficie de la hoja, relacion entre peciolo-limbo y relacion entre longitud y ancho de la hoja (caracteristicas de la hoja); tamaño de la flor, longitud y ancho del petalo (caracteristicas de la flor)

Cabe destacar que la superficie de la hoja y relacion peciolo-limbo presentan correlacion muy baja o nula con cualquiera de las otras variables.

G) Encuentre la matriz que mide el grado de similaridad entre las variedades en función de la distancia euclídea calculada sobre los datos originales.

```
matriz_distancia <- dist(datos, method = "euclidean")
```

H) Podría decir cuales son los tres pares de variedades que presentan mayor semejanza?

| Variedad 1 | Variedad 2 | Distancia |
|---|---|---|
| CORBATO | PALAU | 4.795 |
| GINESTA | MANRI | 5.712 |
| CURROT.T | CRISTALI | 5.779 |

I) Repita lo realizado en el inciso (h) pero sobre las variables estandarizadas por media y desvío estándar. Son las mismas las tres variedades más parecidas? Comente al respecto

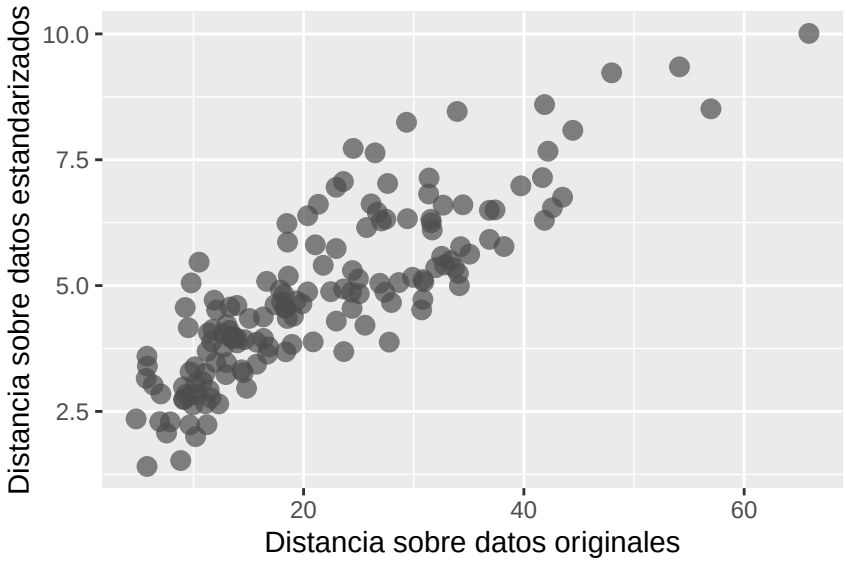
```
matriz_distancia_std <- dist(scale(datos))
```

Con los datos estandarizados, los pares de variedades mas parecidas son

| Variedad 1 | Variedad 2 | Distancia |
|---|---|---|
| CURROT.T | CURROT | 1.407 |
| CORBATO | R.CARLET | 1.525 |
| BLANCO | MARTINET | 2.000 |

Podemos ver que los tres pares de variedades mas parecidas son distintos a los que vimos en el inciso anterior donde utilizamos los datos sin estandarizar. Esto sucede porque las variables estan medidas en diferentes unidades de medicion, y al utilizar las variables sin estandarizar se le da mayor peso a las que tienen una variabilidad mayor valor en la escala de medida original.

J) Mida el grado de concordancia entre ambas matrices de distancia.

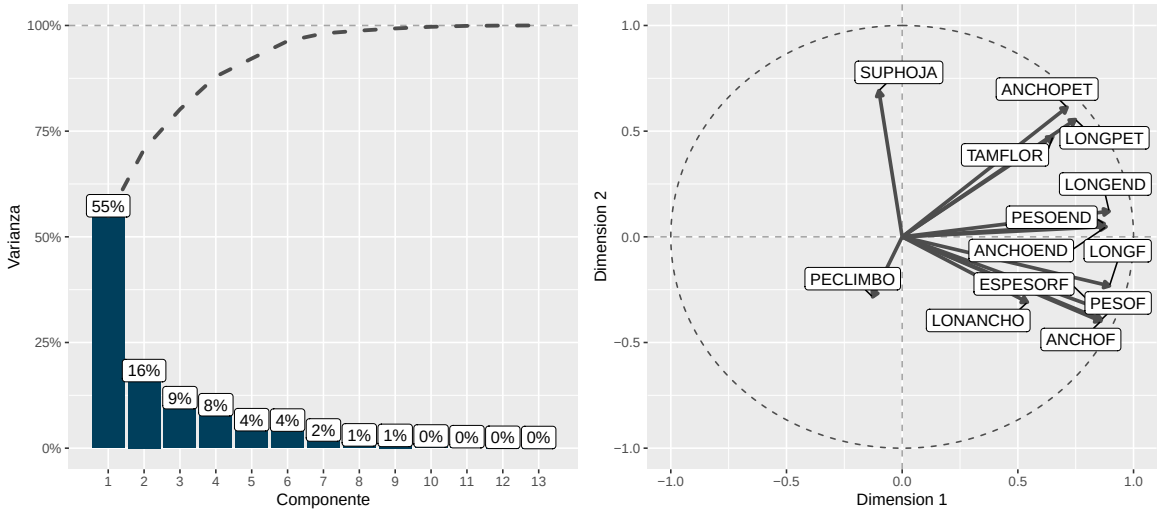



El grado de concordancia entre las matrices de distancias es 0.846.

K) Realice un Análisis de Componentes Principales utilizando de la matriz de correlaciones.

```
pca <- PCA(datos, ncp = 2, graph = FALSE)
```

Y los autovectores (cargas asociadas a cada componente) son

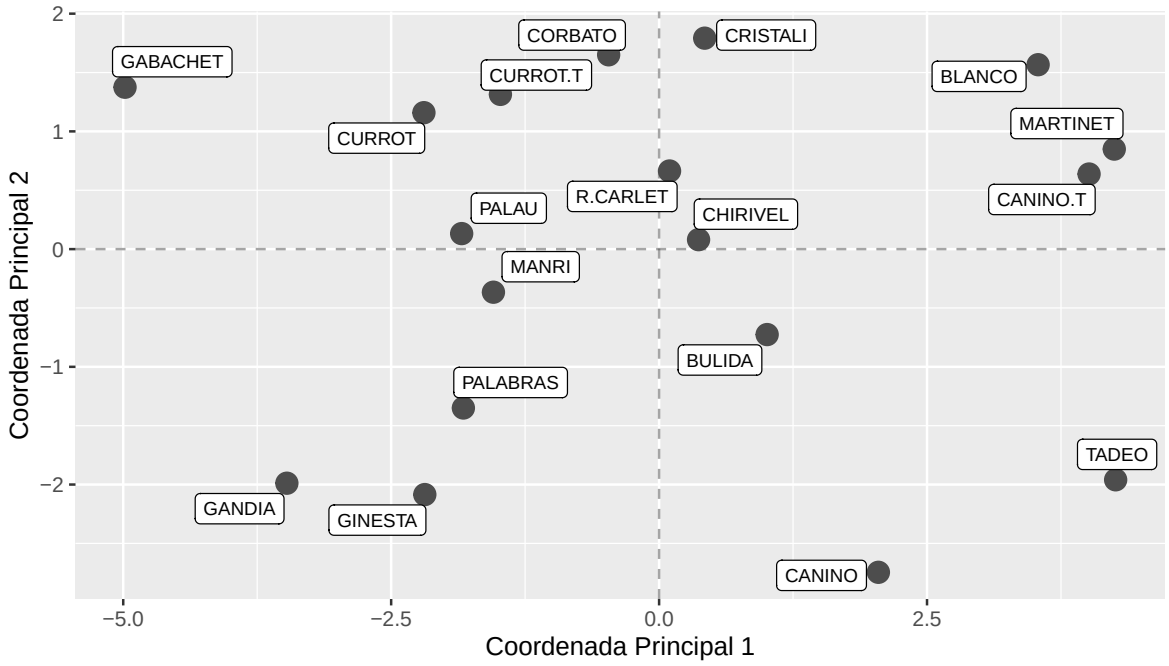



L) Analice los porcentaje de variabilidad explicada por los primeros ejes principales.

Observando los autovalores y el porcentaje de varianza explicada de cada uno, podemos decir que la primer componente explica el 54.85%, mientras que la segunda componente explica 15.86%. Luego, la varianza total explicada por estas dos componentes es 70.7%.

M) Establezca intuitivamente grupos de variedades similares según su cercanía en el plano principal.

A continuacion se presenta la representacion grafica de las variedades en el plano principal:




Tras observar el grafico podemos decir que encontramos cuatro grupos. El primer grupo contiene a MARTINET, CANINO.T y BLANCO. El segundo esta conformado por CANINO y TADEO. El tercer grupo esta conformado solamente por GABACHET, que se diferencia de todas las variedades. Y el cuarto grupo que se compone por el resto de las variedades, que estan ubicadas alrededor del comportamiento promedio, es decir, el origen del plano.

N) Encuentre e interprete gradientes de las variables originales en el plano principal en función de sus cargas sobre las dos primeras componentes.

Considerando la primer componente, los damascos que tengan flores, frutos y endocarpio grandes se van a ubicar a la derecha del grafico. Con respecto a la segunda componente, los damoscos cuyas hojas sean grandes estarán ubicados en la parte superior del grafico, lo mismo ocurre con aquellos damascos con flores grandes. Los damascos que tengan frutos pequeños, estaran ubicados en la parte inferior del grafico.

Luego:

- Damascos con hojas grandes estaran ubicados en el segundo cuadrante del grafico.
- Damascos con endocarpio y frutos grandes estaran ubicados en el cuarto cuadrante del grafico
- Damascos con flores grandes estaran ubicados en el primer cuadrante del grafico

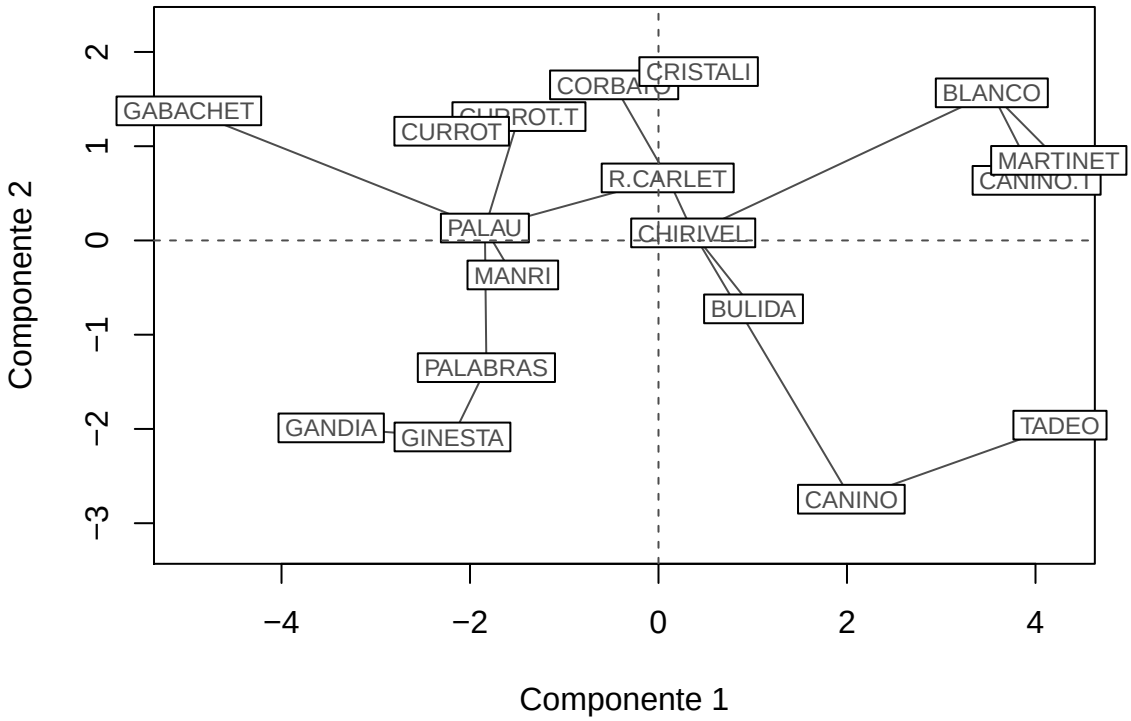
O) Caracterice los grupos determinados en el inciso (N) según los gradientes descriptos en (O).

Grupo 1: es un grupo de variedades caracterizadas por tener flores, endocarpio y hojas grandes y frutos medianos.

Grupo 2: es un grupo caracterizado por tener endoncarpio grande, frutos grandes, hojas pequeñas y flores medianas

Grupo 3: es un fruto caracterizado por tener hojas grandes, frutos y endocarpio pequeños y flores medianas.

P) Superponga en la representación del plano principal un MST. Comente al respecto, haría algún reagrupamiento ?




Q) Con el software `R` realice el ACP recurriendo a operaciones con matrices (decomposición espectral)

Realizamos la descomposicion espectral de la siguiente manera

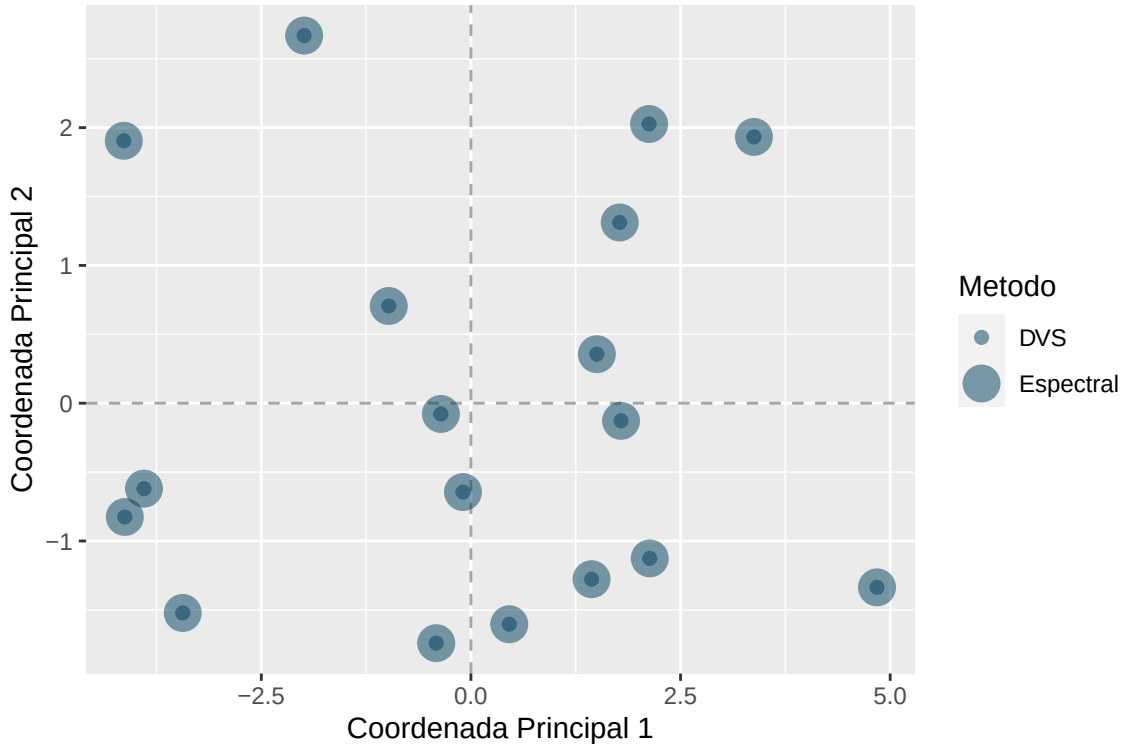
```r
X <- as.matrix(scale(datos))
eig <- eigen(cov(X))
P <- eig$vectors
Y_espectral <- X %*% P
```

R) Verifique que con el enfoque Biplot (DVS) llega a los mismos resultados

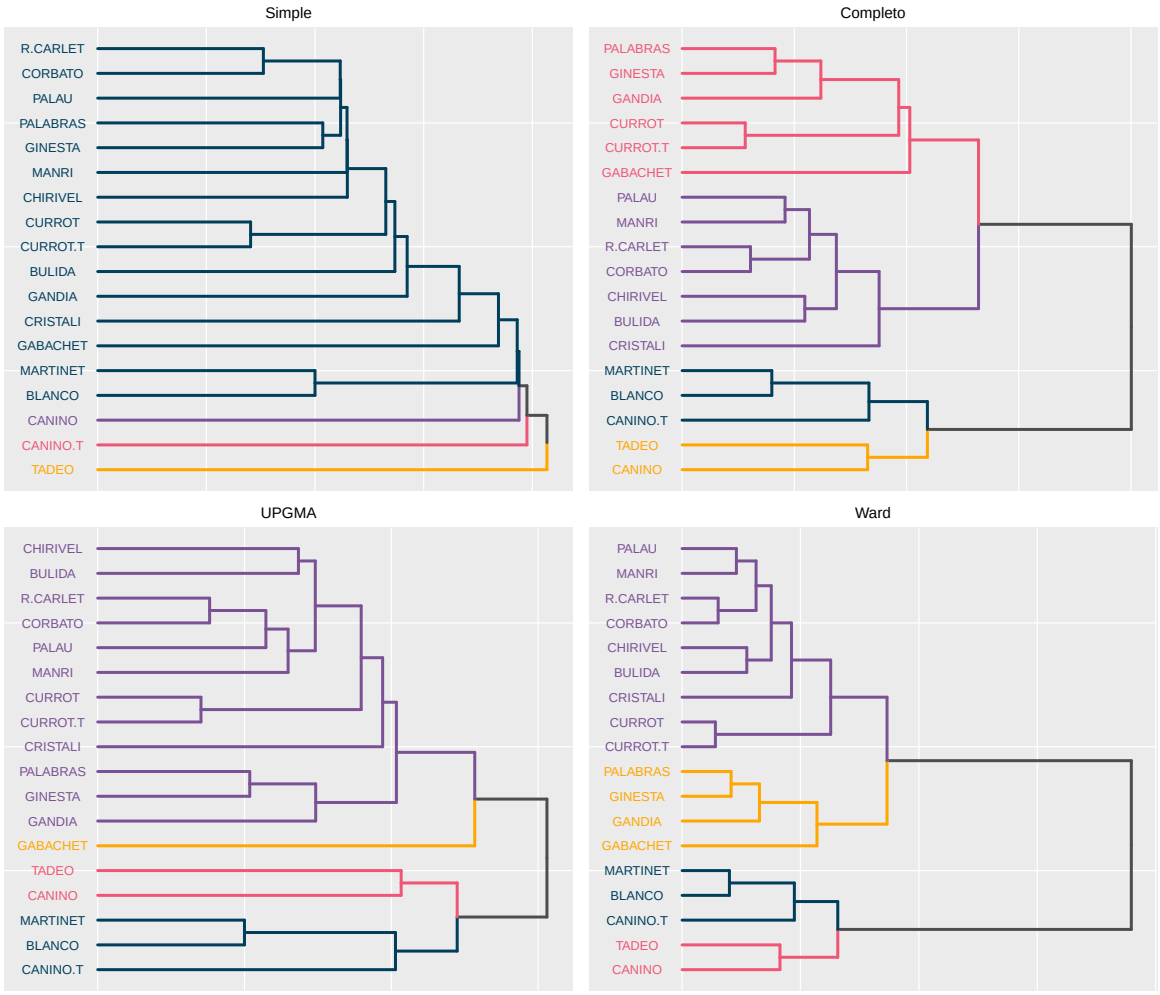
Primero calculamos las coordenadas en el plano principal mediante el enfoque Biplot.

```r
dvs <- svd(X)
U <- dvs$u
D <- diag(dvs$d)
Y_dvs <- U %*% D
```

Si hacemos la razon entre las dos primeras coordenadas obtenidas mediante descomposicion espectral y mediante DVS, vemos que la segunda razon es igual a `-1`. Esto se da porque la configuracion de los puntos mediante DVS resulta estar reflejada en el eje X respecto de la configuracion que obutivmos mediante la descomposicion espectral. En el siguiente grafico multiplicamos a la segunda componente por `-1` para mostrar mas facilmente la equivalencia de las configuraciones obtenidas mediante ambos enfoques.




S) Obtenga 4 dendogramas ultramétricos según diferentes criterios de encadenamiento (SIMPLE, COMPUESTO, UPGMA y WARD).

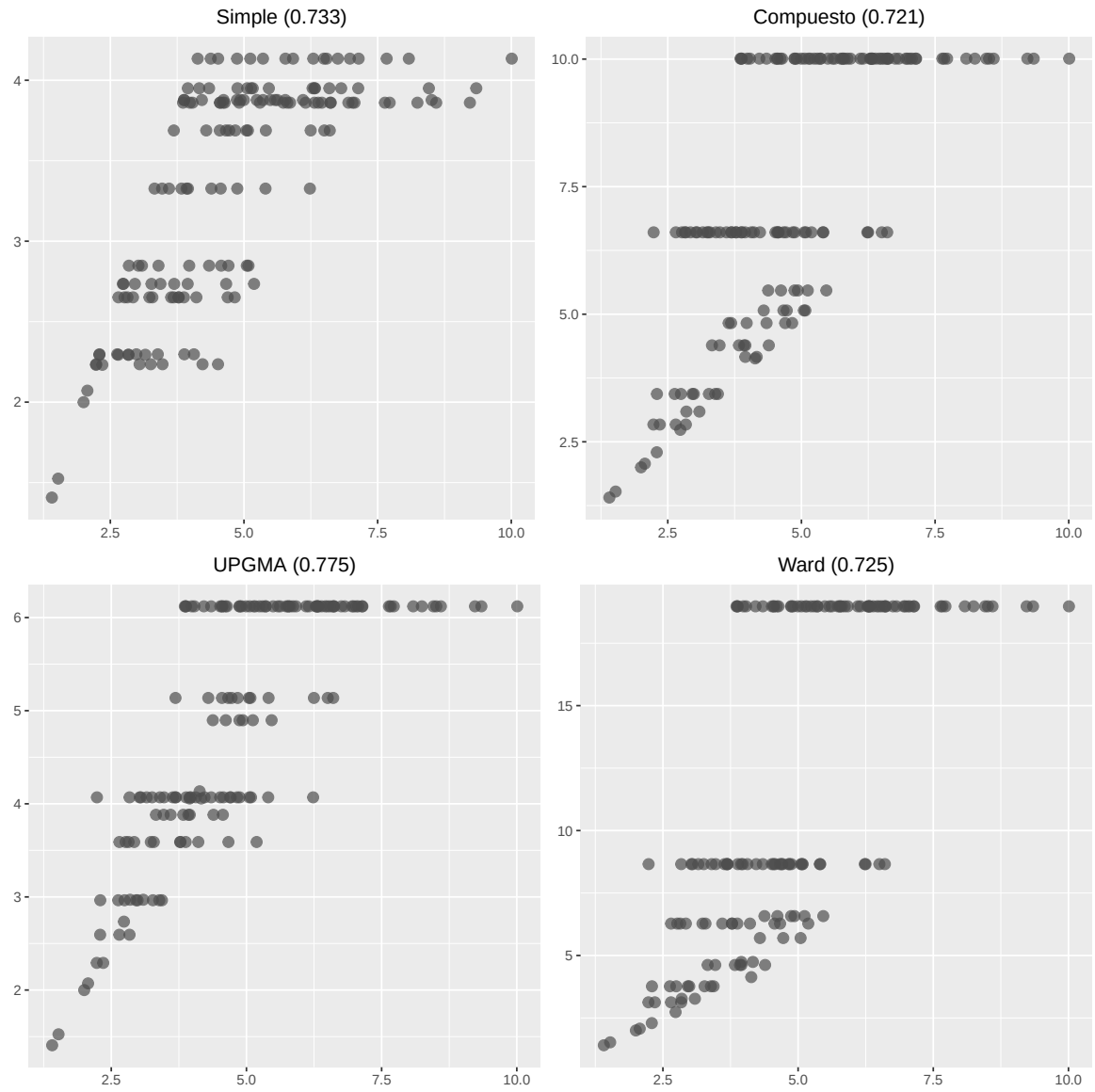



T) Asocie a cada uno de los árboles obtenidos en el inciso anterior la matriz cofenética correspondiente. Que miden los elementos de estas matrices ?

```
cophenetic_min <- cophenetic(cluster_min)
cophenetic_comp <- cophenetic(cluster_com)
cophenetic_upgma <- cophenetic(cluster_upgma)
cophenetic_ward <- cophenetic(cluster_ward)
```

Los elementos de las matrices anteriores permiten tener una medida de la coherencia del criterio de agrupamiento jerarquico. Como pudo observase el criterio de encadenamiento UPGMA es el que brinda, para este conjunto de datos, un mejor criterio de agrupamiento.

U) Cuantifique la concordancia entre la matriz de distancia que dio origen a los dendogramas y las 4 matrices cofenéticas. A que conclusión llega ?




Observando las correlaciones obtenidas entre la matriz de distancia y las diferentes matrices cofeneticas puede decirse que la obtenida a traves del metodo de encadenamiento UPGMA es la mas alta. El metodo de encadenamiento UPGMA es el que

QUE MAS??

V) Existe algún punto de corte sobre el índice de jerarquización del dendograma UPGMA que origine los mismos agrupamiento de variedades obtenidos en Análisis de Componentes Principales ?

Observando el grafico (ver num q le corresponde) podemos ver con los diferentes colores que se puede realizar un corte el cual permite obtener los mismos agrupamientos que se obtuvieron a traves del Analisis de Componentes Principales

W) Mida el grado de concordancia entre los resultados obtenidos por Componentes y por Cluster UPGMA

Para hallar el grado de concordancia entre los resultados obtenidos por Componentes Principales y por CLuster UPGMA se calcula la correlacion entre las matrices de distancias, obteniendo una correlacion de 0.76. Dicha correlacion es alta y esto se debe a que

X) Halle el dendograma aditivo Neighbor Joining, representa mejor la configuración de variedades que el Cluster UPGMA ?

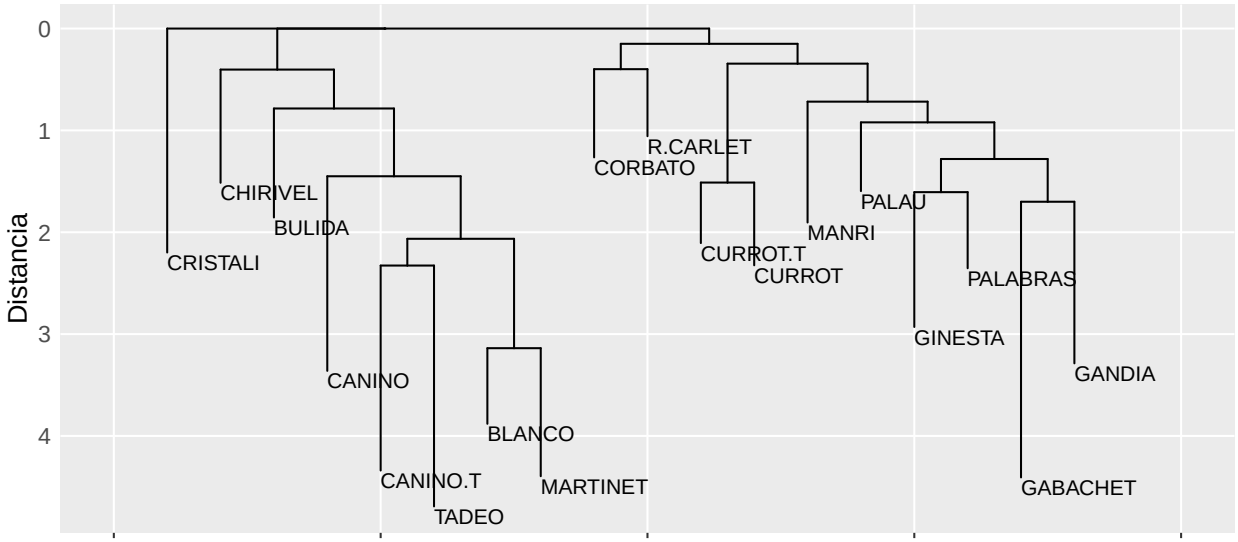



Figure 1: Representacion jerarquizada del arbol aditivo Neighbor-Joining.