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) ?

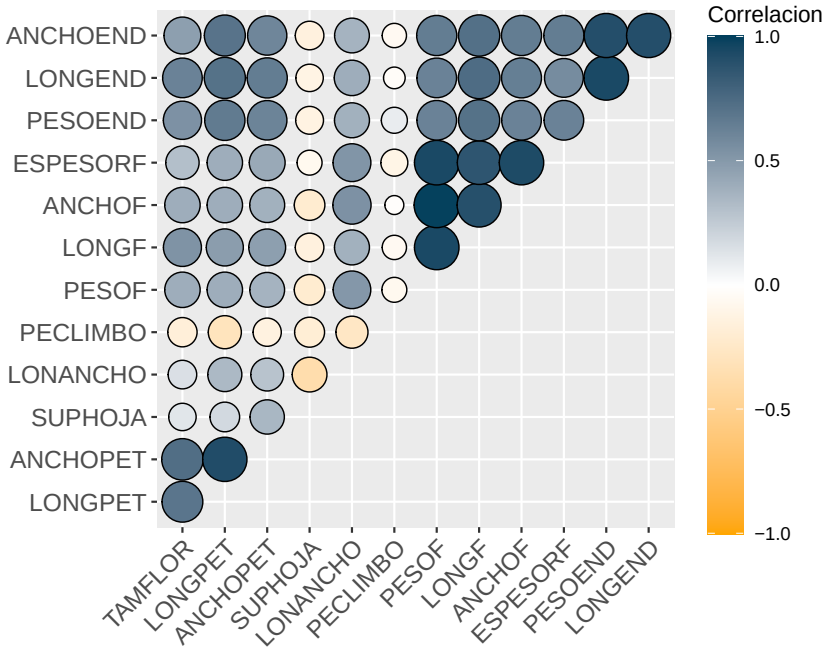
El coeficiente de variacion no existe ya que todas las variables tienen media igual a 0, por lo que no podriamos responder al inciso (d) luego de la estandarizacion. Sin embargo, lo expuesto en el inciso (d) es suficiente para concluir sobre que variables tienen mayor y menor dispersion.

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

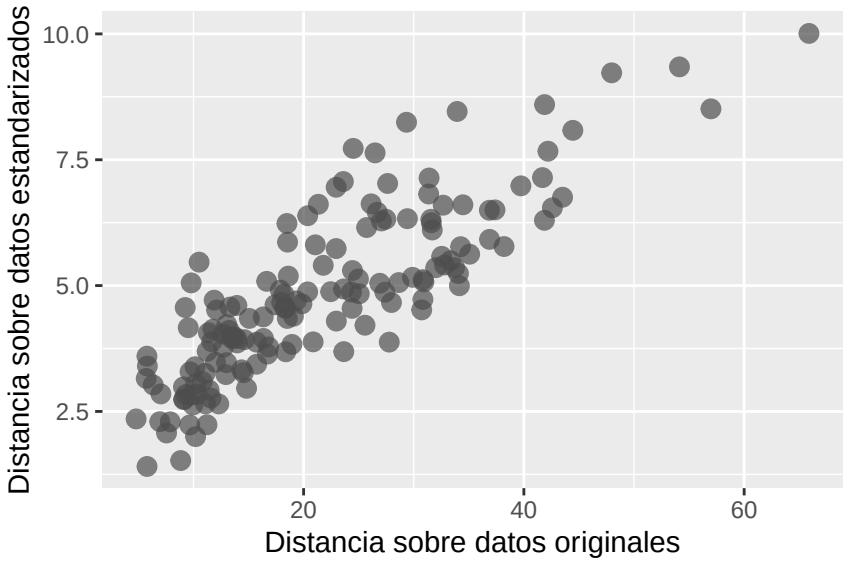
```
matriz_distancia_std <- dist(scale(datos))
```

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.

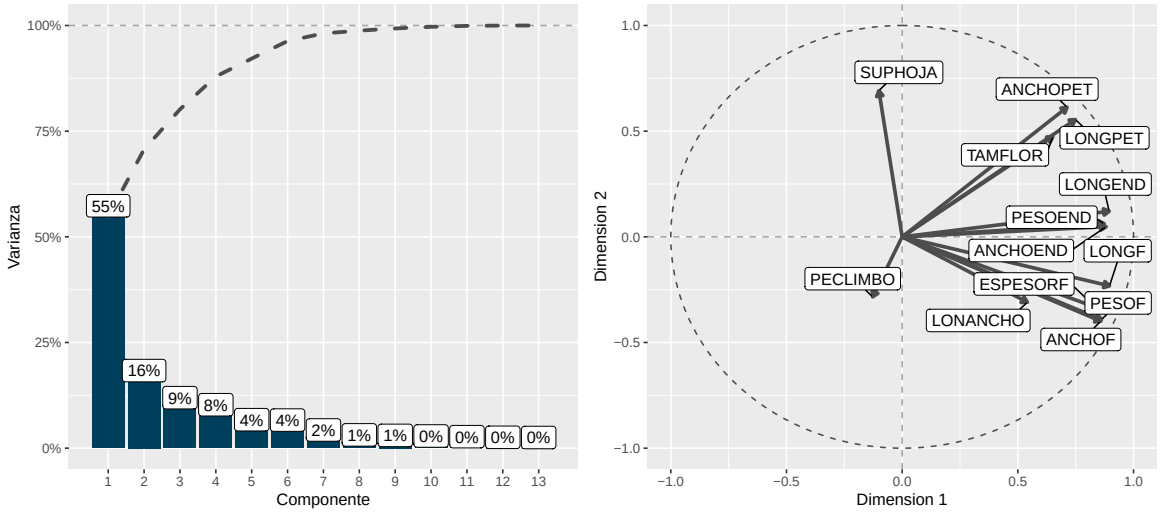


El grado de concordancia entre las matrices de distancias es 0.846.

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

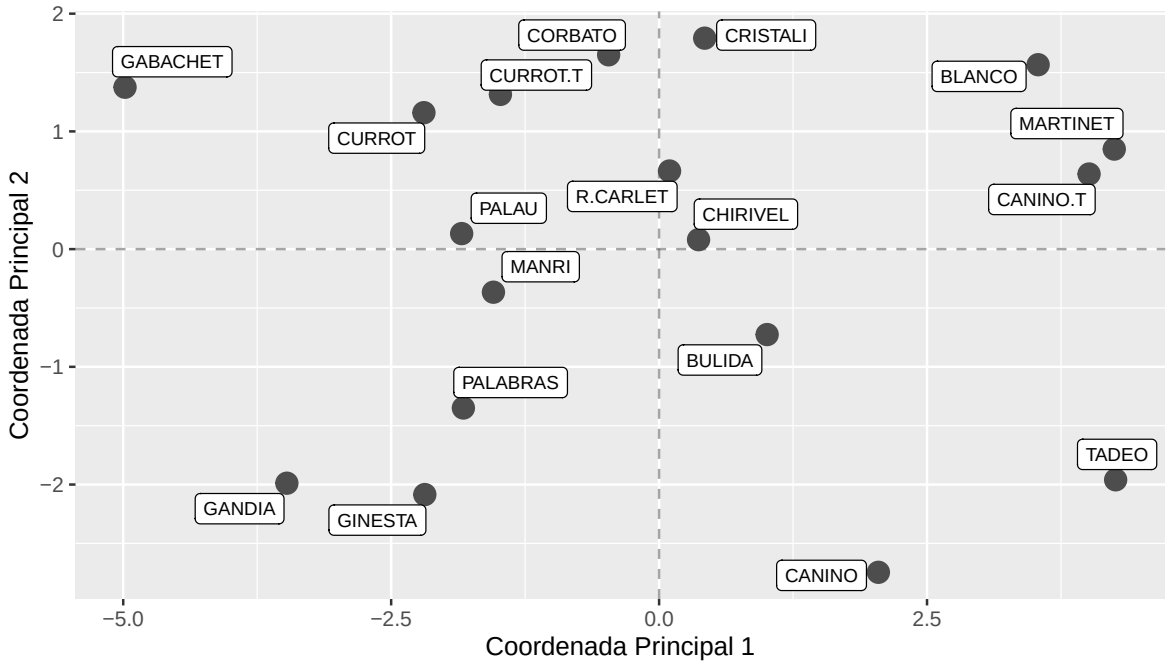


L) Analice los porcentaje de variabilidad explicada por los primeros ejes principales.

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

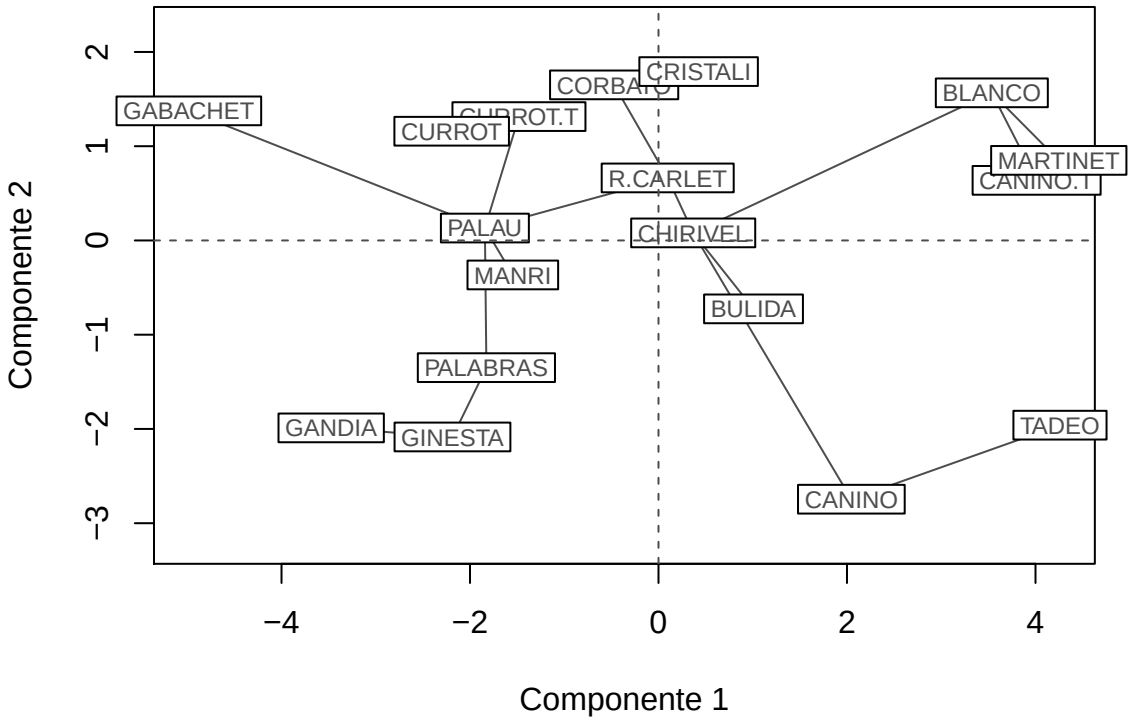
O) Caracterice los grupos determinados en el inciso (N) según los gradientes descriptos en (O).

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

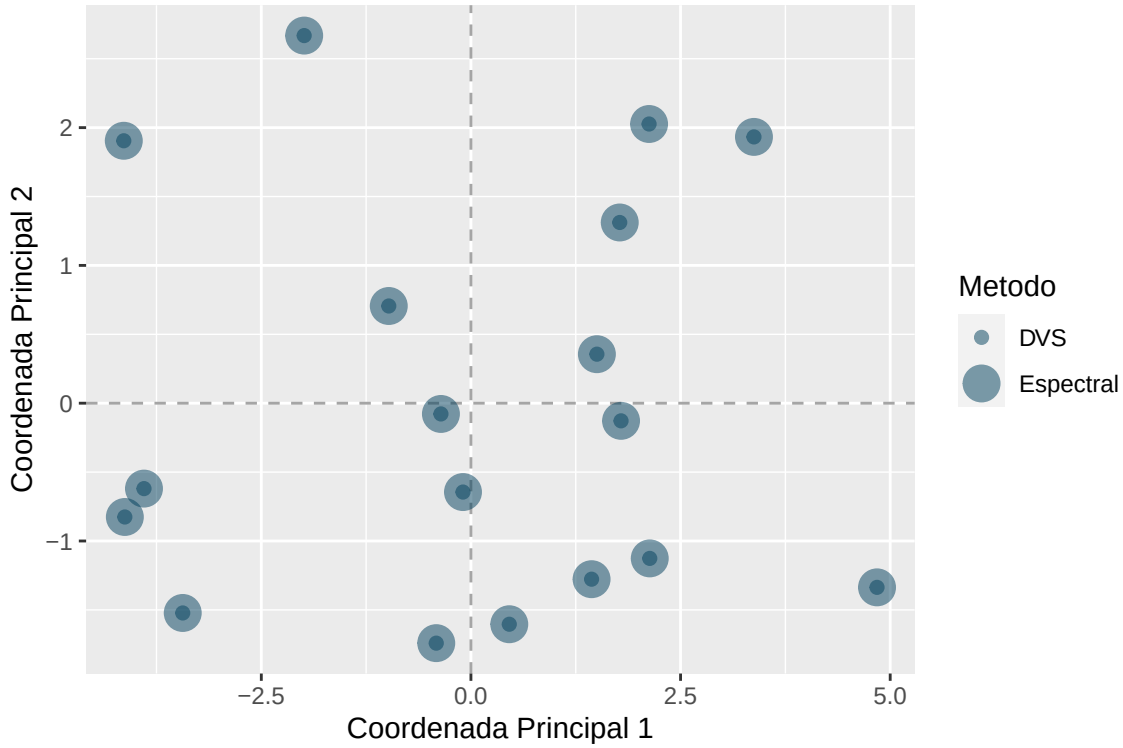
```r
X <- as.matrix(scale(datos))
eig <- eigen(cov(X))
P <- eig$vectors
Y_espectral <- X %*% P
```

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

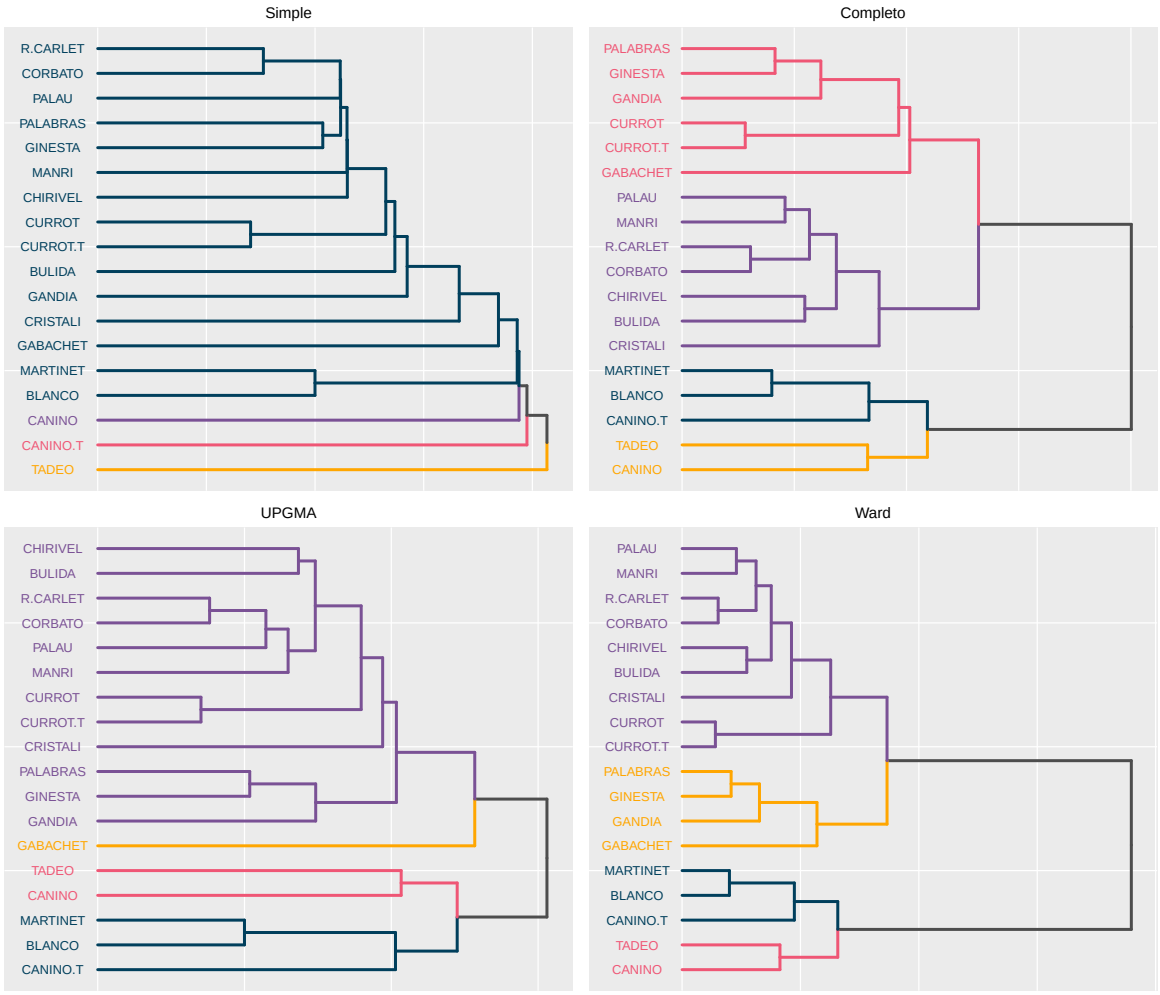
Primero calculamos las coordenadas en el plano principal mediante el enfoque Biplot.

```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).

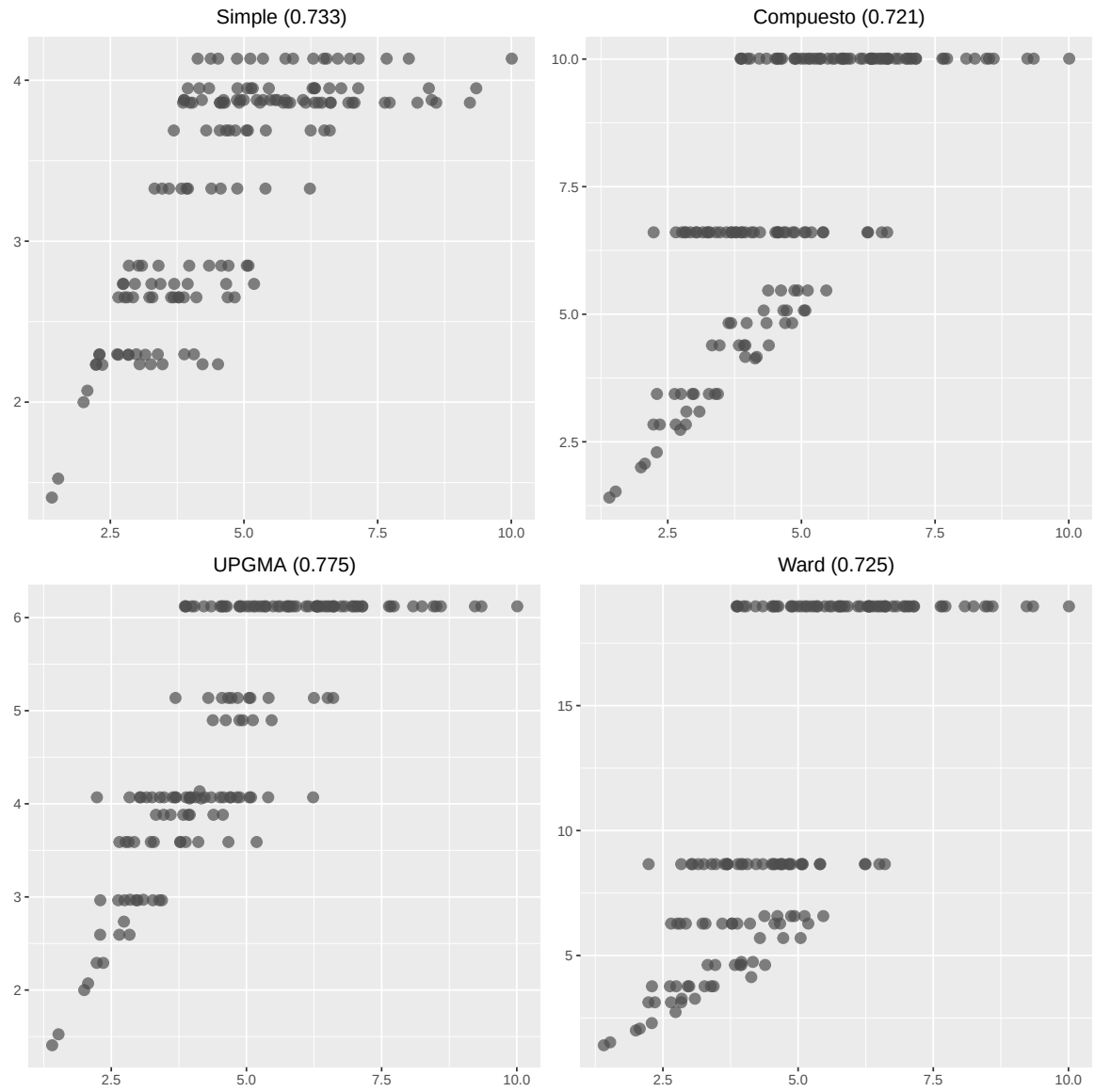


T) Asocie a cada uno de los árboles obtenidos en el inciso anterior la matriz cofenética correspondiente. Que miden los elementos de estas matrices ?

```
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 ?

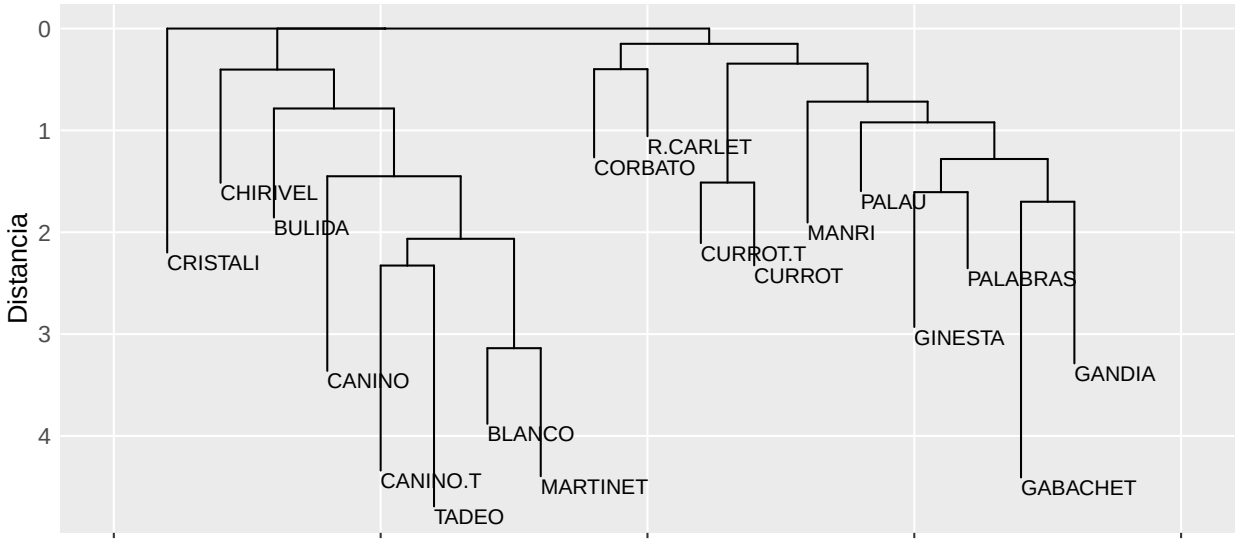


Figure 1: Representacion jerarquizada del arbol aditivo Neighbor-Joining.

Ejercicio 2

A) Cuantifique (en forma manual) la similaridad entre las variedades correspondientes a la primera y segunda fila en función del porcentaje de caracteres comunes respecto al número de caracteres totales. Idem las variedades asociadas a las filas 12 y 13 (incluir la variable TIPO)

Obtenemos dos vectores que representan a cada una de las dos primeras variedades y luego calculamos el la proporcion de variables donde ambas variedades coinciden.

```
variedad1 <- datos[1, ] %>% unlist() %>% unname()
variedad2 <- datos[2, ] %>% unlist() %>% unname()
similaridad_1_2 <- mean(variedad1 == variedad2, na.rm = TRUE) * 100
```

La similaridad entre las variedades de la fila 1 y 2 es del 88.89%. Si tenemos en cuenta que se tienen 9 variables, podemos notar que estas dos variedades coinciden en todas excepto 1. Luego, de manera analoga para el par de variedades 12 y 13

```
variedad12 <- datos[12, ] %>% unlist() %>% unname()
variedad13 <- datos[13, ] %>% unlist() %>% unname()
similaridad_12_13 <- mean(variedad12 == variedad13, na.rm = TRUE) * 100
```

La similaridad entre las variedades de la fila 12 y 13 es del 71.43%, lo que significa que difieren mas que el par de variedades 1 y 2.

B) Halle una matriz de similaridad entre variedades en función del coeficiente SM generalizado.

Para esta tarea utilizamos la funcion `daisy()` del paquete `cluster`, donde especificamos que la metrica a utilizar es `"gower"`, y nos devuelve la matriz de distancia entre las diferentes variedades. En la siguiente linea, convertimos la matriz de distancia a matriz de similaridad.

```
matriz_distancia <- daisy(datos_fct, metric = "gower")
matriz_similaridad <- 1 - matriz_distancia
```

Comprobemos, por ejemplo, si la similaridad entre las variedades de pepino de la primera y segunda fila computada mediante `daisy()` es igual a la que computamos a mano.

```
matriz_similaridad[1] == (similaridad_1_2 / 100)
```

```
## [1] TRUE
```

Por lo que ambos resultados son iguales.

C) Aplique Análisis de Coordenadas principales para representar en un espacio bidimensional la semejanza entre las variedades.

Aplicamos el analisis de coordenadas principales de la siguiente manera:

```
coordenadas_principales <- cmdscale(sqrt(matriz_distancia), k = 2, eig = TRUE)
```

Y graficamos a la caracterizacion cualitativa de las variedades de pepino en el plano principal.

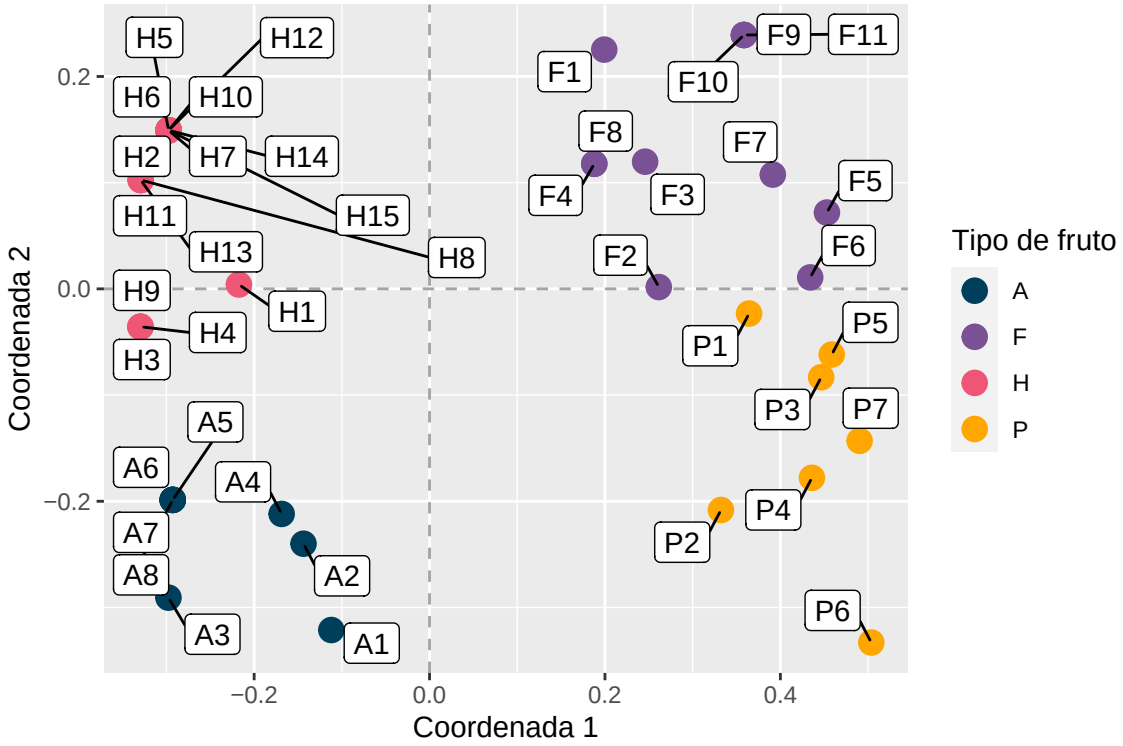


Figure 1: Caracterizacion cualitativa de las variedades de pepino en el plano principal.

En aquellos casos que mas de una etiqueta apunta hacia el mismo punto sucede que las variedades coinciden en terminos de las variables analizadas y en consecuencia los puntos que las representan estan encimados. Por ejemplo, las variedades `A5`, `A6`, `A7`, `A8` tienen idénticos valores para todas las variables.

	TIPO	SEXO	CORN...	TORN...	PEDUNC	VERR...	MOTEADO	CLADOSP	CMV
A5	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A6	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A7	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A8	Alf	Fem	Negra	Espin	Cue	No	No	No	No

D) Conforme grupos de variedades según su homogeneidad en la caracterización agronómica cualitativa.

En la Figura 1 se puede ver que las variedades de los tipos de fruto **A** y **H** se se agrupan de manera que respetan al tipo de fruto y se diferencian del resto. Tambien existe una agrupacion, ya no tan clara, para las variedades de los grupos **F** y **P**. En este caso, si no estuviera el color que diferencie a los tipos de frutos, no podriamos diferenciar a estos dos grupos claramente. Por ejemplo, las variedades **F2**, **F6** y **P1** estan muy cercanas en el plano y las podriamos haber tomado como parte de un mismo grupo.

Tambien podemos ver que la variabilidad de las variedades dentro de cada tipo de fruto difiere. Por ejemplo, para el tipo de fruto **H**, se tiene que casi todas las variedades se corresponden con dos categorizaciones particulares (por eso vemos tantos puntos encimados). Por otro lado, todas las variedades del tipo de fruto **P** se corresponden con una configuracion única de las variables cualitativas.

E) Encuentre el dendograma ultramétrico con ligamiento UPGMA correspondiente

Obtenemos el dendograma utilizando la funcion `hclust()`, a la que le pasamos la matriz de distancia previamente obtenida.

```
cluster_cualitativas <- hclust(matriz_distancia, method = "average")
```

Luego graficamos el dendograma y mostramos con diferentes colores a los clusters que se obtienen al especificar `k = 4`.

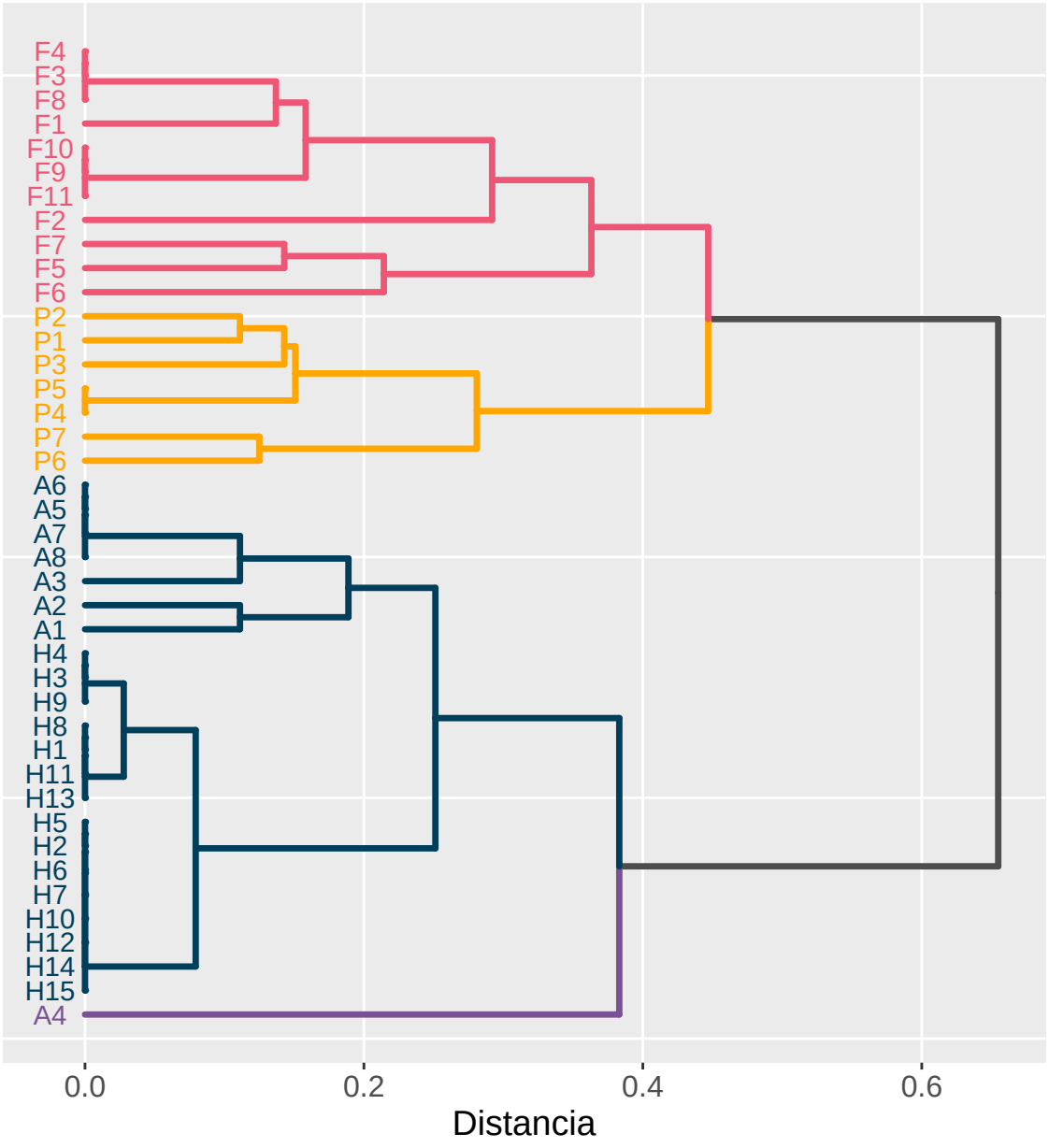


Figure 2: Dendograma Ultrametrico con ligamiento UPGMA.

En la Figura 2 se puede ver que el punto de corte `k = 4` no me permite separar a los tipos de pepinos de manera perfecta. El agrupamiento hace un buen trabajo al diferenciar a los pepinos de los tipos **F** y **P**, pero mezcla a las variedades de los tipos **A** y **H**. Estos dos ultimos tipos, son a su vez, los que mayor similaridad entre variedades presentan. Lo podemos ver en la cantidad de uniones que se presentan a una distancia de 0.

Si comparamos la Figura 1 y la Figura 2, puede sorprendernos que la variedad **A4** este tan distante

del resto de las variedades en la Figura 2, ya que se presenta cercano al resto de las variedades de tipo **A** en la Figura 1. Sin embargo, no debemos pasar por alto que la Figura 1 es una proyeccion de posicionamientos en un espacio de mayor dimensionalidad, pudiendo estos puntos estar distantes en ese espacio original.

Tomemos al tipo de fruto **A** y miremos, por ejemplo, a las variedades **A4** y **A2**, que parecen cercanos en la Figura 1 pero estan distantes en la Figura 2, y observemos sus datos crudos.

	TIPO	SEXO	CORN...	TORN...	PEDUNC	VERR...	MOTEADO	CLADOSP	CMV
A1	Alf	Fem	Blanca	Espin	Obt	No	No	No	No
A2	Alf	Fem	Negra	Espin	Obt	No	No	No	No
A3	Alf	Fem	Negra	Pelos	Cue	No	No	No	No
A4	Alf	Fem	Negra	Pelos	Agu	No	No	NA	NA
A5	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A6	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A7	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A8	Alf	Fem	Negra	Espin	Cue	No	No	No	No

En la tabla podemos ver que la variedad **A4** es la unica que presenta extremo pedunculo agudo y es una de las dos unicas que presenta pelos como tipo de ornamentación, lo que alcanza para diferenciarla de las otras variedades, que son mucho mas similares entre si.

F) Mida a través de su matriz cofenética la concordancia con la matriz de distancias que le dio origen

Primero vamos a calcular la concordancia y luego obtenemos un grafico de dispersion donde se muestra la distancia original y la distancia cofenética.

```
distancia_cofenetica <- cophenetic(cluster_cualitativas)
concordancia <- cor(distancia_cofenetica, matriz_distancia)
```

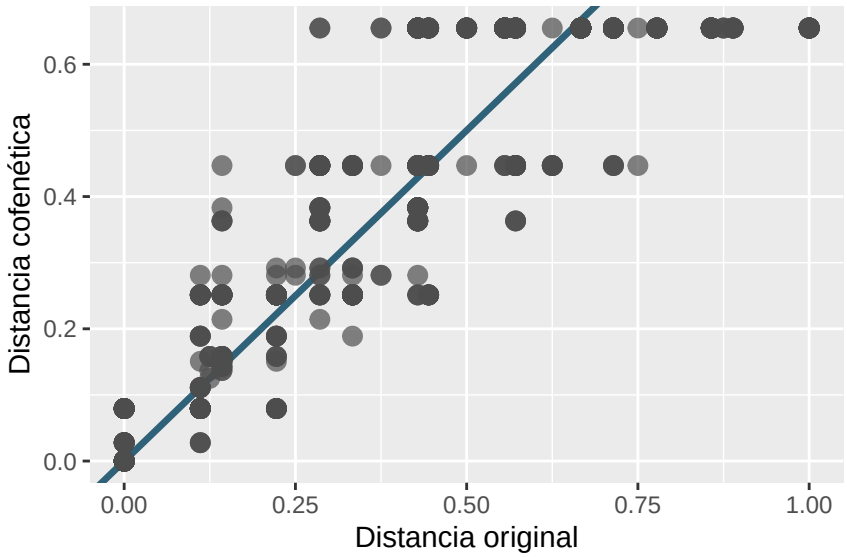


Figure 3: Grafico de dispersion entre distancia original y distancia cofenética a partir de cluster con ligamiento UPGMA. La linea azul representa a la recta identidad.

La concordancia entre la matriz de distancias cofenética y la matriz de distancia original es igual a 0.883, lo que indica una concordancia muy alta entre ambas representaciones.

En la Figura 3 se puede ver que la discrepancia entre estas distancias crece para valores mas altos de la distancia original.

G) Cuantifique concordancia entre plano principal de ACoordP y Cluster

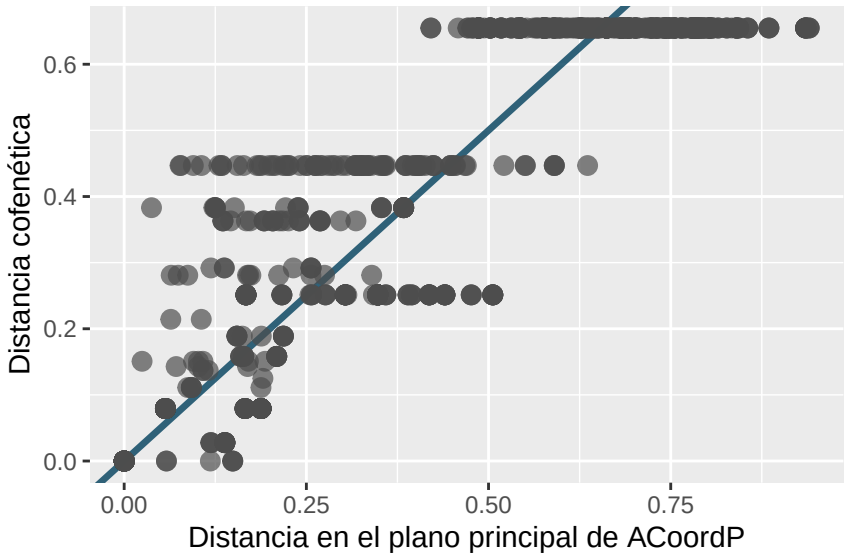


Figure 4: Grafico de dispersion entre distancia original y distancia cofenética a partir de cluster con ligamiento UPGMA. La linea azul representa a la recta identidad.

La concordancia concordancia entre plano principal de ACoordP y Cluster es 0.891, lo que nuevamente nos indica una concordancia muy alta entre ambas representaciones.

Ejercicio 3

Librerias

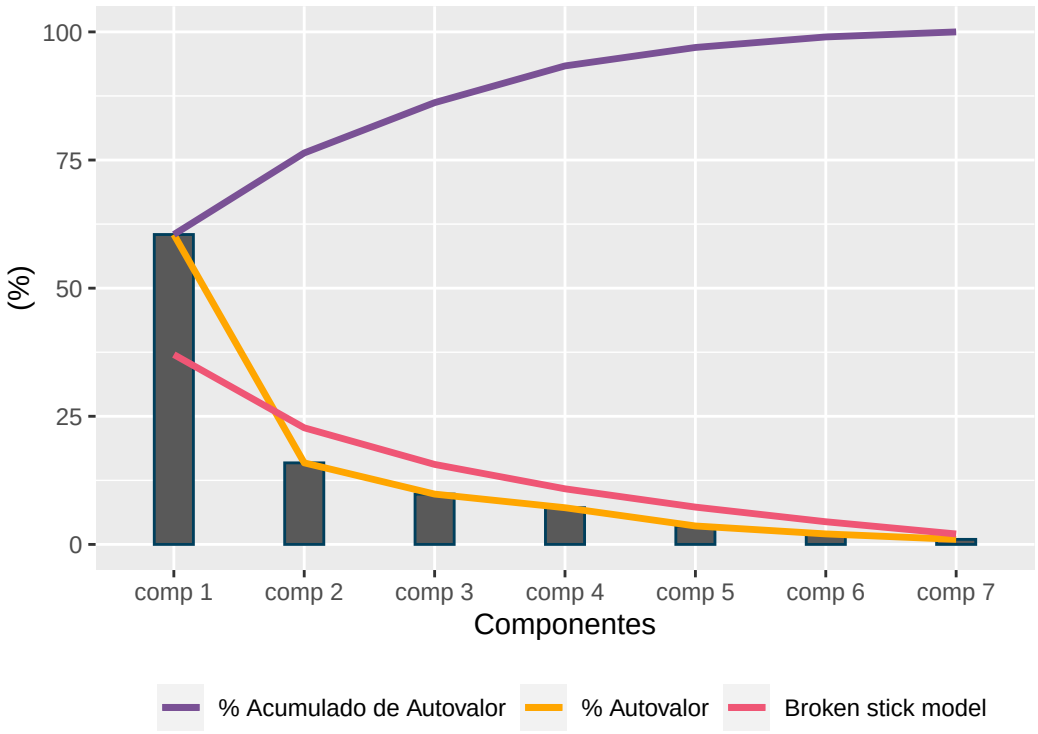
Datos

En el archivo CUANTITATIVAS se presentan los datos correspondientes a la caracterización agronómica cuantitativa de las 41 variedades de pepino tratadas en el ejercicio anterior. Los caracteres cuantitativos analizados fueron: número de flores femeninas por nudo (FLORES), número de espinas en el ovario (ESPINAS), número de aristas (ARISTAS) y estrías (ESTRIAS) en fruto, longitud de fruto (FRUTO), intensidad de cuello (CUELLO) e intensidad de color de cuello (CCUELLO)

A. Aplique sobre estos datos un Análisis de Componentes Principales a partir de matriz de correlaciones.

Para analizar los datos y realizar el Análisis de Componentes Principales, se utilizó la libreria "FactoMineR". Como el objetivo era utilizar la matriz de correlación como base del análisis, se utilizó la opción de escalar los datos dentro del método de la librería. Cabe destacar que el paquete utiliza la noción francesa a la hora de computar las matrices (n, no n-1). A continuación de presentan los resultados:

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.23	60.47	60.47
comp 2	1.11	15.91	76.38
comp 3	0.69	9.82	86.20
comp 4	0.50	7.16	93.36
comp 5	0.25	3.60	96.96
comp 6	0.14	2.05	99.01
comp 7	0.07	0.99	100.00



Como puede observarse en la tabla, las dos primeras componentes acumulan el 76.38% del porcentaje de los autovalores. Tal como denota el gráfico, se observa que el criterio de "broken srick model" apunta a que las dos primeras componentes son suficientes para representar la variabilidad de los datos.

B. Realice la representación de las variedades en el plano principal, encuentre grupos y caracterícelos

A partir de la representación en dos dimensiones de los datos, se podría establecer que existen visualmente 3 grandes grupos de Pepinos. Los P/F, los A y los H. Esta separación coloca a las variedades Holandesas (H), en el cuadrante superior derecho. Este cuadrante está caracterizado por ser la dirección de crecimiento de las variables Cuello, Estrias, Frutos y Flores. En el cuadrante superior izquierdo se alocan las variedades Pepinillo y Francesas (P y F). Tal cuadrante queda caracterizado por ser la dirección de crecimiento de las variables Espinas, Aristas y Ccuello. Finalmente, en el cuadrante inferior, se sitúan las variedades Alpha-Beta (A).

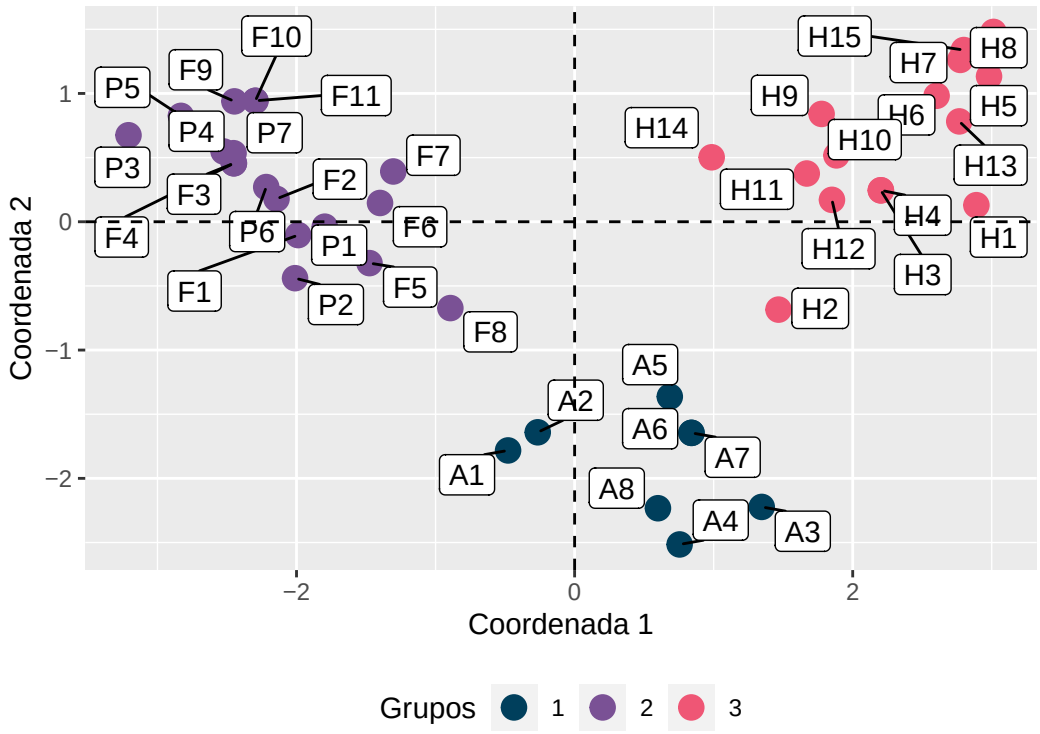


Figure 1: Caracterizacion cuantitativa de las variedades de pepino.

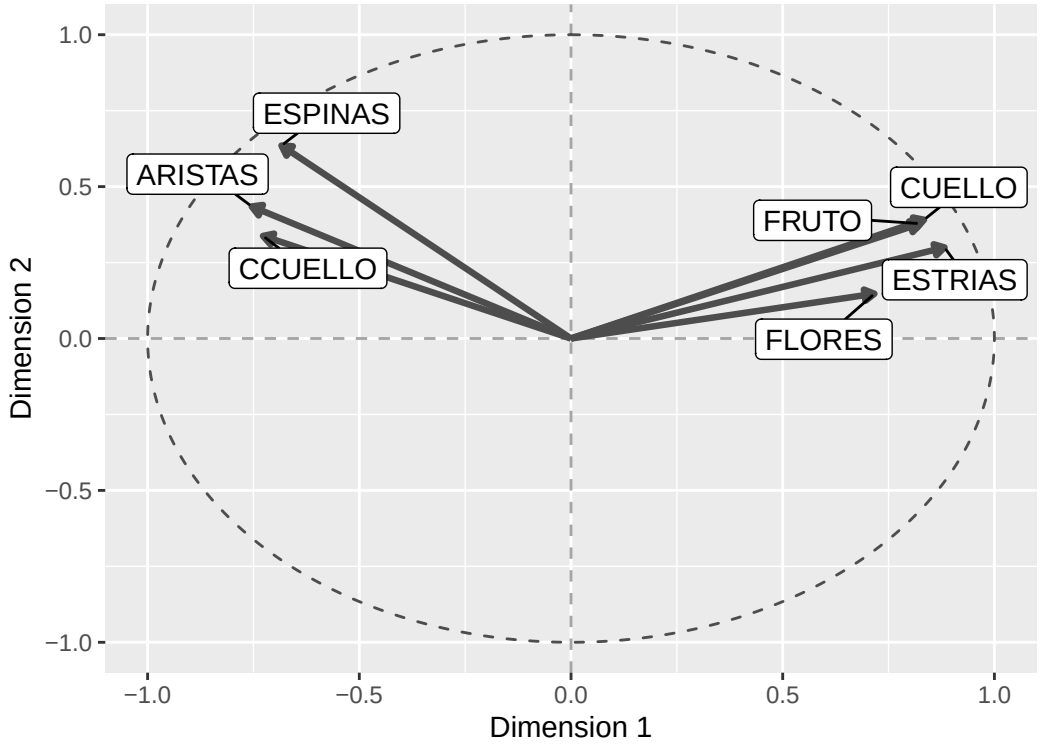


Figure 2: Gradiente de las Variables

C. Evalúe visualmente y a través de la correlación entre matrices la concordancia general entre esta configuración y la hallada en función de los caracteres cualitativos

Visualmente, se observa que la clusterización en base a componentes principales no logra separar las variadades P y F en dos dimensiones. En contra posición, se observa que con dos dimensiones, el análisis basado en las variables cualitativas lográ discriminar con notable exactitud las variedades y hasta colocarlas en cuadrantes opuestos.

Desde un punto de vista cuantitativo, si se calculan las matrices de distancias a partir de la representación en dos dimensiones para tipo de análisis y luego se toma la correlación entre los elementos de la diagonal subinferior, se observa que hay una concordancia del 0.7797612. Esto mismo puede corroborarse en el diagrama de dispersión de puntos.

```
## [1] 0.7797612
```

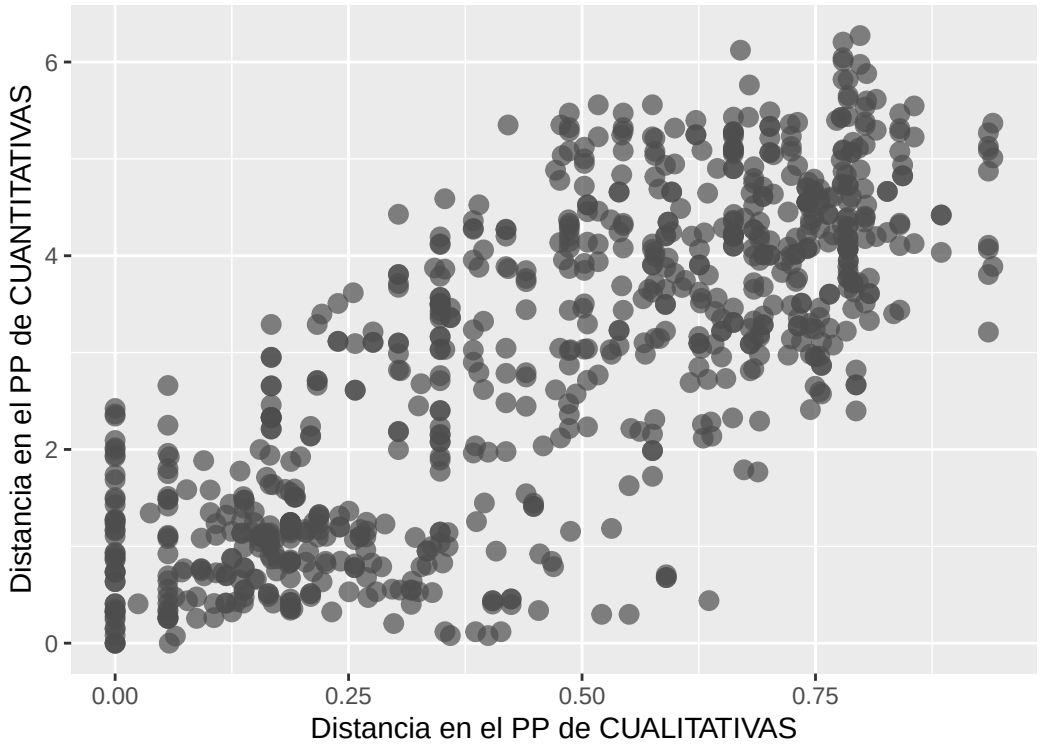


Figure 3: Distancia entre Matrices

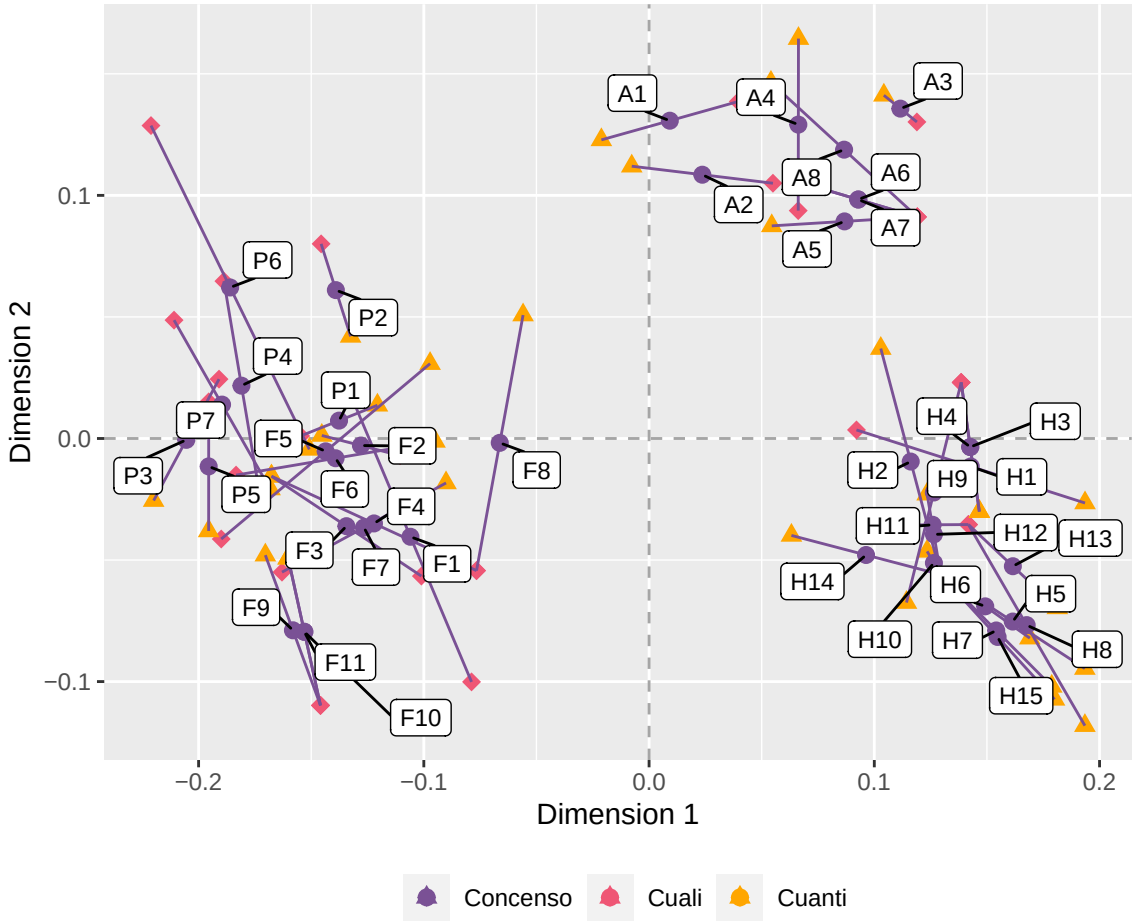
Ejercicio 4

Datos

	c1_cuali	c2_cuali	c1_cuanti	c2_cuanti
A1	-0.11	-0.32	-0.48	-1.78
A2	-0.14	-0.24	-0.27	-1.64
A3	-0.30	-0.29	1.35	-2.22
A4	-0.17	-0.21	0.76	-2.52
A5	-0.29	-0.20	0.68	-1.36
A6	-0.29	-0.20	0.84	-1.65

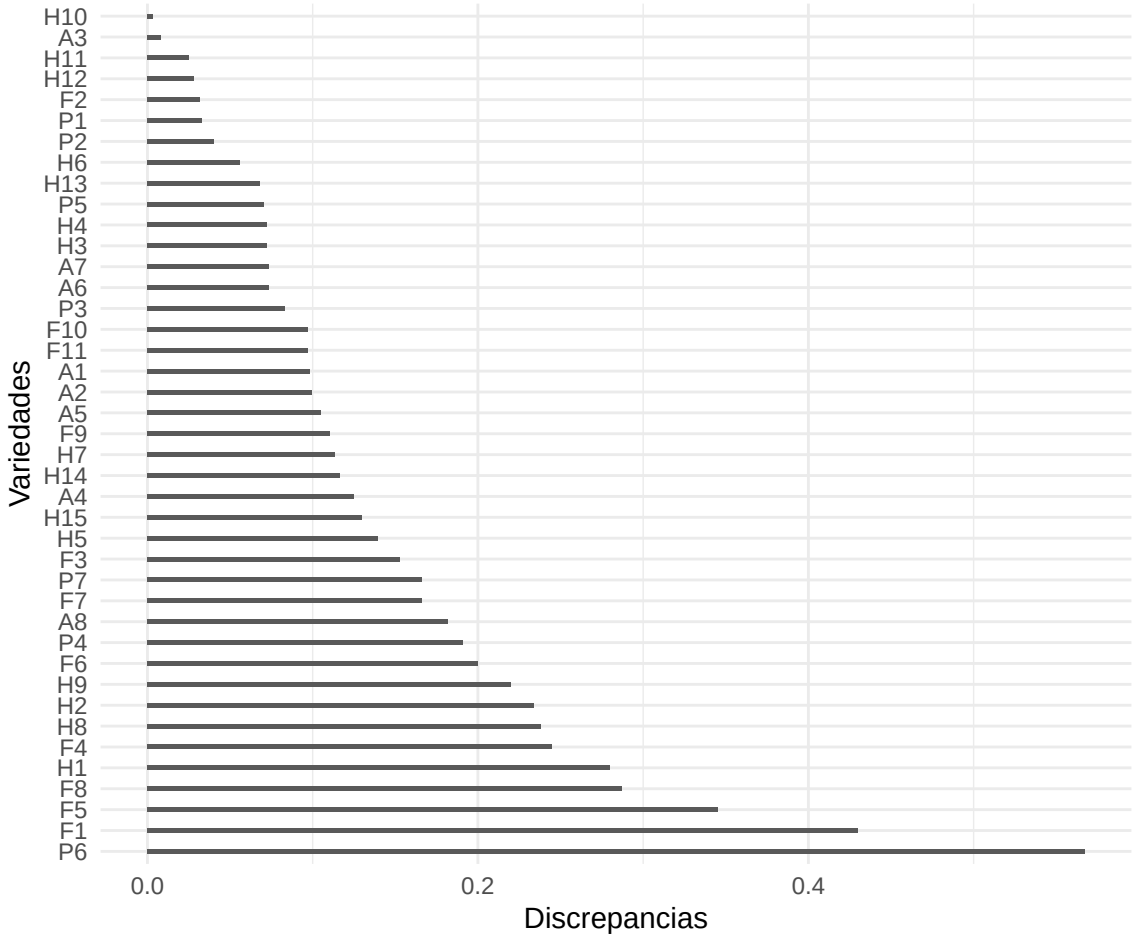
A. Halle la configuración de consenso recurriendo a Análisis de Procrustes Generalizados

Se utilizó para el análisis la librería FactorAnalysis y el método GPA con la semilla "1234" para poder hacer reproducible el análisis. A continuación se muestra su resultádo gráfico:



B. Podría decir para cual o cuales variedades existe mayor concordancia entre la caracterización cuantitativa y cualitativa ?

Se puede observar la discrepancia y concordancia resultante del análisis de Procrustes en el siguiente gráfico:



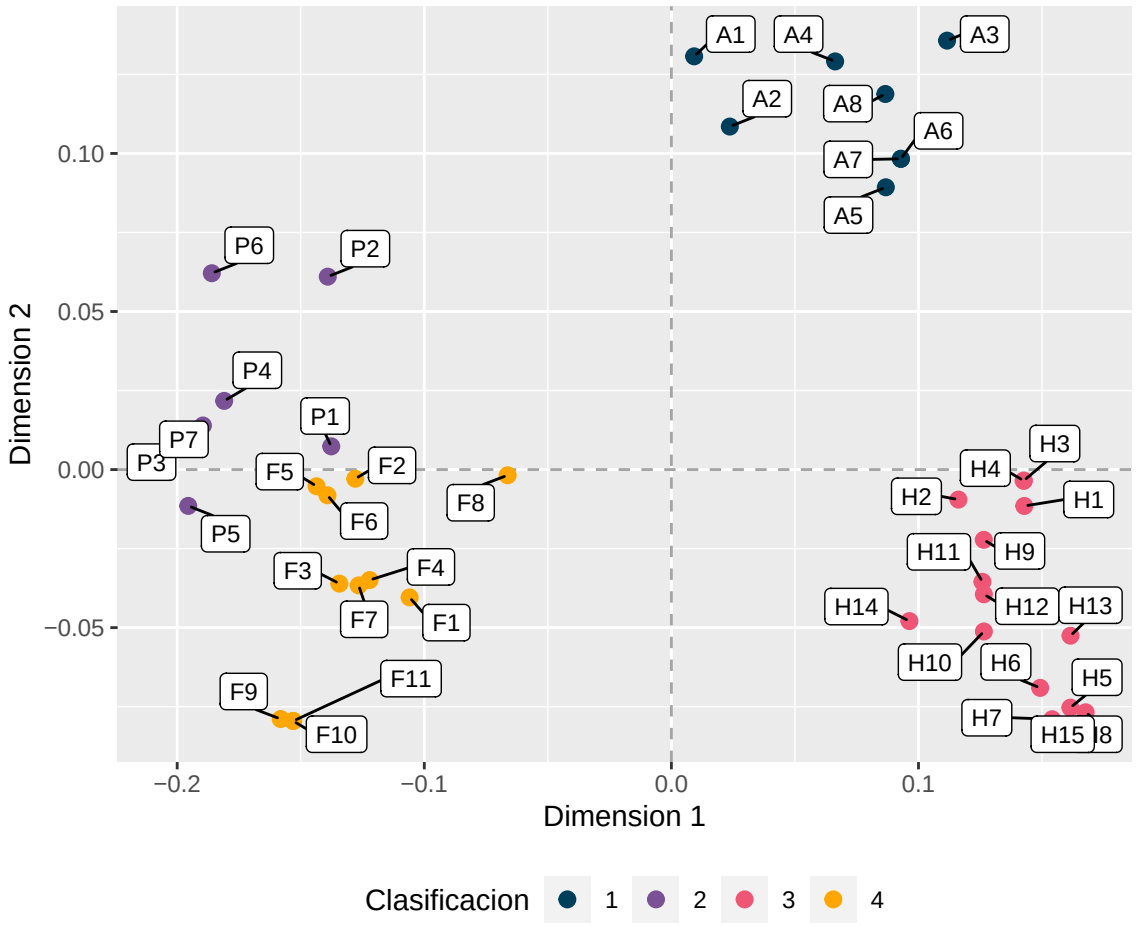
Aquellas cinco que presentan mayor concordancia son H10, A3, H11, H12, F2. En términos agregados, se podría decir que en promedio la Alpha-Beta poseen una mayor concordancia, seguidas por los Pepinillos.

C. Idem para las variedades con mayor discrepancia entre ambas configuraciones.

Aquellas cinco que presentan mayor discrepancia son H1, F8, F5, F1, P6. En términos agregados, se podría decir que en promedio las Francesas poseen una mayor discrepancia, seguidas por las Holandesas.

D. Grafique sólo los puntos de consenso y diga si pudo conformar grupos asociados a los tipos de frutos. Compare con los resultados de los ejercicios 2 y 3.

Al graficar los puntos de consenso, se nota como la combinación de la información cualitativa junto con la cuantitativa permite separar casi a la perfección los diferentes grupos. A diferencia del PCA sobre las variables cuantitativas, la representación logra separar en los cuatro cuadrantes a los diferentes tipos de frutos e intensifica la separación entre la variedad Holandesa y la Alpha-Beta. Si en cambio, comparamos la representación de las variables cualitativas, se observa que la separación en cuadrantes mejora, aunque no así con la separación entre los Pepinillos y las Francesas.


A1
A4
A3
A2
A8
A6
A7
A5
P6
P2
P4
P1
P7
P3
F5
F2
F8
F6
P5
F3
F4
F7
F1
F9
F11
F10
H3
H4
H2
H1
H11
H9
H14
H12
H13
H10
H6
H5
H7
H15
0.10
0.05
0.00
−0.05
Dimension 2
−0.2
−0.1
0.0
0.1
Dimension 1
Clasificacion
1
2
3
4

Ejercicio 5

A) Halle la distancia genética de Prevosti entre variedades

Para calcular esta distancia usamos la siguiente sentencia

```
matriz_distancia <- dist(datos, method = "manhattan", diag = TRUE) / 33
```

Mostramos la matriz resultante de manera parcial, ya que es muy grande como para mostrarla de manera completa en una pagina.

	A1	A2	A3	A4	A5	A6	A7	A8	F1	F2
A1	0.00	0.27	0.30	0.33	0.30	0.39	0.39	0.33	0.55	0.47
A2	0.27	0.00	0.33	0.30	0.33	0.30	0.30	0.24	0.45	0.33
A3	0.30	0.33	0.00	0.03	0.18	0.15	0.15	0.09	0.42	0.41
A4	0.33	0.30	0.03	0.00	0.15	0.12	0.12	0.06	0.39	0.38
A5	0.30	0.33	0.18	0.15	0.00	0.09	0.09	0.15	0.42	0.41
A6	0.39	0.30	0.15	0.12	0.09	0.00	0.00	0.06	0.39	0.38
A7	0.39	0.30	0.15	0.12	0.09	0.00	0.00	0.06	0.39	0.38
A8	0.33	0.24	0.09	0.06	0.15	0.06	0.06	0.00	0.45	0.38
F1	0.55	0.45	0.42	0.39	0.42	0.39	0.39	0.45	0.00	0.18
F2	0.47	0.33	0.41	0.38	0.41	0.38	0.38	0.38	0.18	0.00

Pero lo que si podemos observar son los pares mas similares y los mas distintos.

Variedad 1	Variedad 2	Distancia
A6	A7	0
F3	F4	0
H3	H4	0
H5	H6	0
H7	H8	0
H7	H15	0
H8	H15	0

Variedad 1	Variedad 2	Distancia
F5	H7	0.59
F5	H8	0.59
F5	H11	0.59
F5	H15	0.59
A1	F5	0.60

Donde vemos que hay 7 pares de variedades que presentan valores identicos para sus variables moleculares, y que la distancia maxima entre pares es 0.6.

B) Podría aplicar el coeficiente de similaridad SM ? Porque ?

Si, pero no lo hacemos porque perderiamos informacion, ya que las bandas presentan mas de 2 valores posibles. Categorizar los valores observados en solamente dos categorias implicaria una perdida de informacion.

C) Realice un Análisis de Coordenadas Principales para encontrar la configuración de las variedades de pepino en función de esta caracterización molecular. Encuentra asociaciones en función del tipo de pepino?

```
coordenadas_principales <- cmdscale(matriz_distancia, eig = TRUE)
```

En la Figura 1 podemos ver que los pepinos del tipo **F** suelen encontrarse en el primer cuadrante, los del tipo **H** en el segundo cuadrdante, los de tipo **A** en el tercero, y los de tipo **P** en el cuarto. Sin embargo esta ordenacion es un tanto imprecisa, ya que por ejemplo, hay pepinos de los tipo **F** y **A** en el cuarto cuadrante, asi como pepinos del tipo **H** en el tercero. Sino estuvieran los colores que indican los tipos de pepinos, probablemente obtendriamos agrupamientos que estuvieran compuestos en su mayoria un unico tipo de pepino, pero que tambien incluirian pepinos de otros tipos.

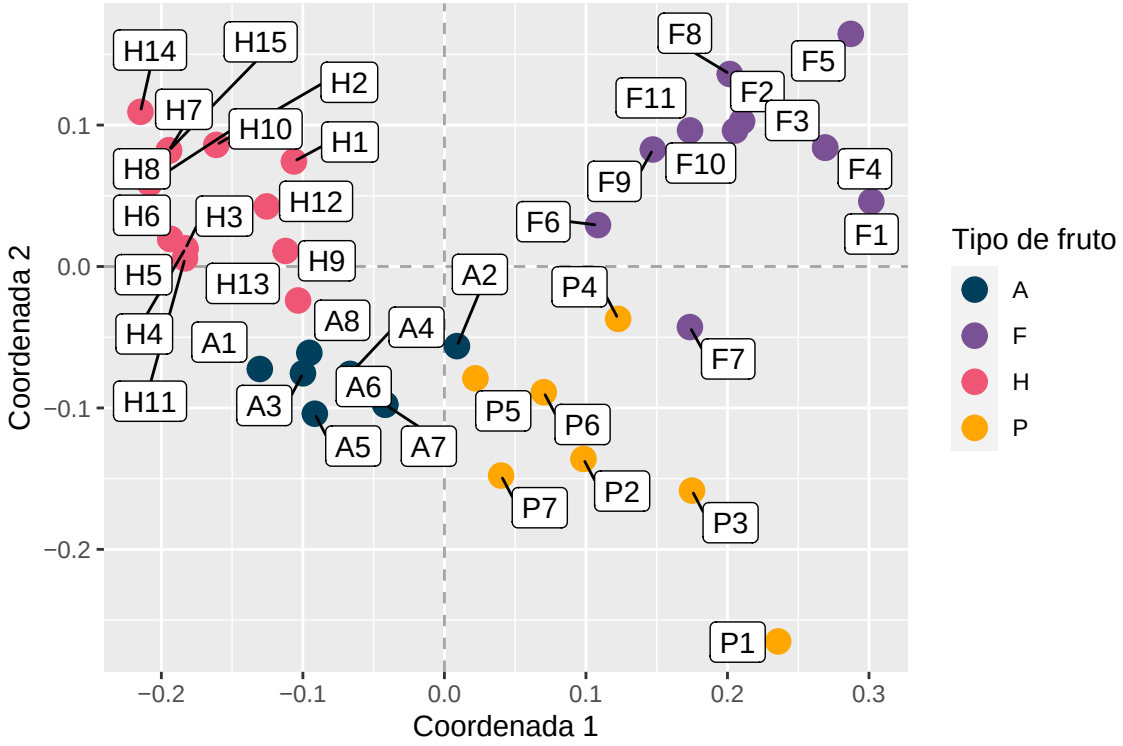


Figure 1: Caracterizacion molecular de las variedades de pepino en el plano principal.

D) Encuentre el dendograma ultramétrico con ligamiento UPGMA

```
cluster_molecular <- hclust(matriz_distancia, method = "average")
```

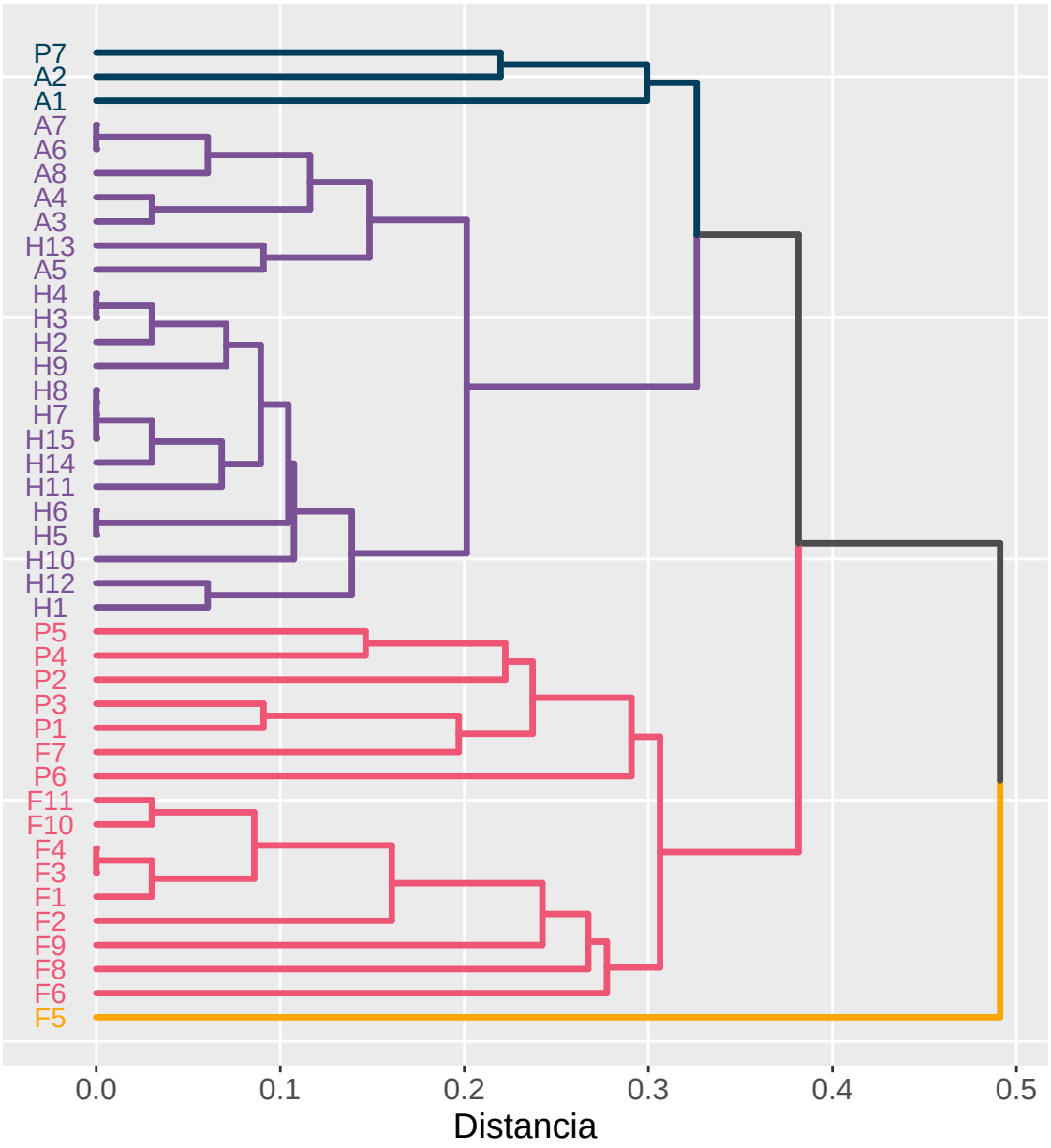


Figure 2: Dendograma Ultrametrico con ligamiento UPGMA para las variedades de pepino en base a caracteristicas moleculares.

Resulta interesante comparar los agrupamientos representados en la Figura 2 con los representados en la Figura 2 del Ejercicio 2. En ambos casos sucede que hay una mezcla de variedades pertenecientes a los tipos alfa-beta y holandes. Sin embargo, en el agrupamiento de este ejercicio se observa que las variedades **A1** y **A2** no pertenecen al mismo cluster que el resto de las alfa-betas, sino que se unen con una variedad del tipo pepinillo, **P7**. A priori, hubiera sido dificil imaginarse este agrupamiento si uno solo hubiera tenido en cuenta a la configuracion de los puntos en el plano principal tal como se ve en la Figura 1. En este dendograma, asi como sucede con el dendograma de la Figura 2 del Ejercicio 2, tambien se observa una variedad que se diferencia sustancialmente del resto, la variedad **F5**. Si bien en la Figura 1 se ve que esta variedad esta en el extremo del grafico, no queda tan claro que se diferencia sustancialmente del resto.

E) Mida a través de su matriz cofenética la concordancia con la matriz de distancias que le dio origen

La concordancia entre la matriz de distancias cofenética y la matriz de distancia original es igual a 0.872, lo que habla de una alta similaridad entre las mismas. En la Figura 3 se puede ver la asociacion positiva entre las dos medidas de distancia. La dispersion en la nube de puntos aumenta a medida que la distancia es mayor, lo que significa que las dos medidas de distancia tienden a diferir mas cuando la distancia entre los tipos de pepinos es mayor.

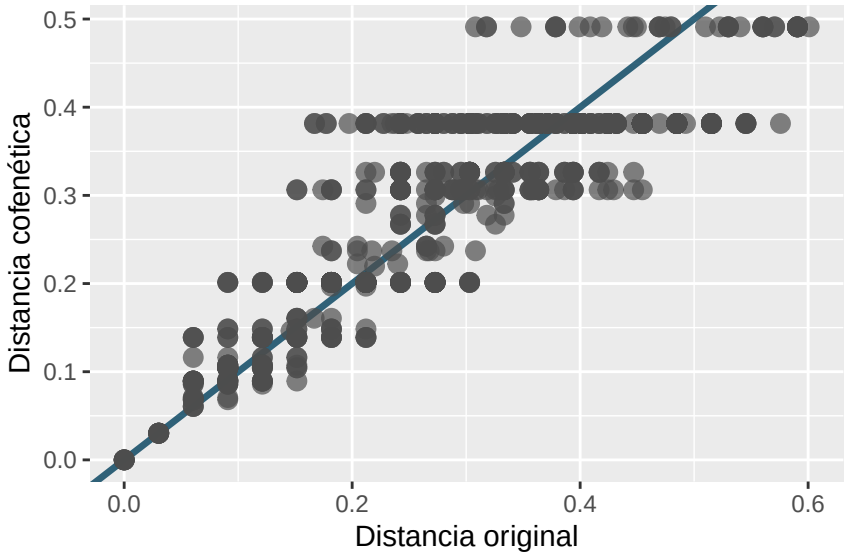


Figure 3: Grafico de dispersion entre distancia original y distancia cofenética a partir de d endograma Ultrametico con ligamiento UPGMA. La linea azul representa a la recta identidad.

F) Halle el dendograma aditivo Neighbor Joining

```
rapds_nj <- nj(matriz_distancia)
```

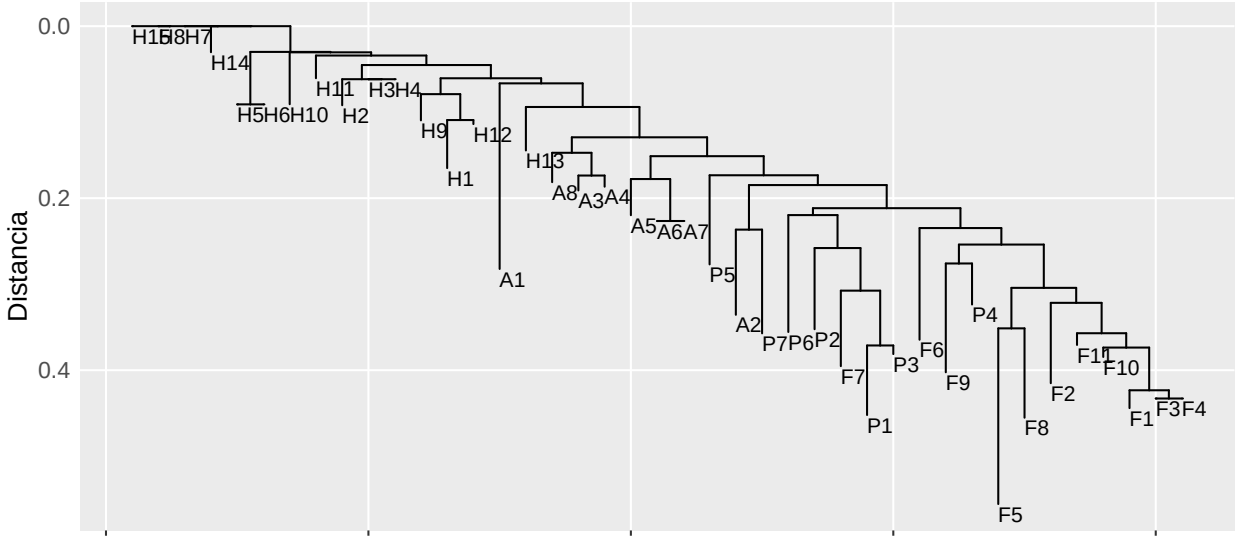


Figure 4: Representacion jerarquizada del arbol aditivo Neighbor-Joining

G) Mida su concordancia con matriz de distancia original

```
distancia_cofenetica_nj <- as.dist(cophenetic(rapds_nj), diag = TRUE, upper = FALSE)
concordancia <- cor(distancia_cofenetica_nj, matriz_distancia)
```

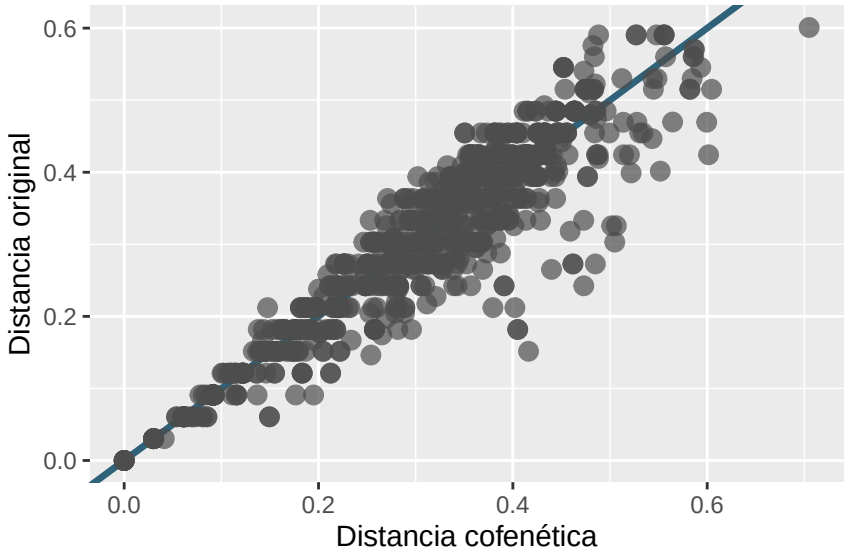


Figure 5: Grafico de dispersion entre distancia original y distancia cofenética a partir del dendograma aditivo. La linea azul representa a la recta identidad.

La concordancia entre la matriz de distancias cofenética construida a partir del dendograma aditivo Neighbor Joining y la matriz de distancia original es igual a 0.926, por lo que este arbol aditivo es el que mejor representa las distancias originales entre las variedades, y se corresponde con lo mencionado en clase de que en general estos arboles aditivos dan mejor.

H) Relacione ambos dendogramas y saque conclusiones

```
concordancia <- cor(distancia_cofenetica_nj, distancia_cofenetica_upgma)
```

La concordancia entre ambos dendogramas es 0.872. Es decir, en ambos casos se preserva altamente el ordenamiento entre variedades pero no de manera perfecta. Por ejemplo, si miramos la Figura 4, vemos que la variedad **P7** se uniria primero a la variedad **A2**, pero no asi a la variedad **A1**, como si sucede en la Figura 2.

Ejercicio 6

Librerias

Para cada una de las configuraciones que se midieron en los pepinos, cualitativa, cuantitativa y molecular, obtenemos la matriz de distancia entre los puntos en el plano principal y calculamos la correlacion entre pares de matrices.

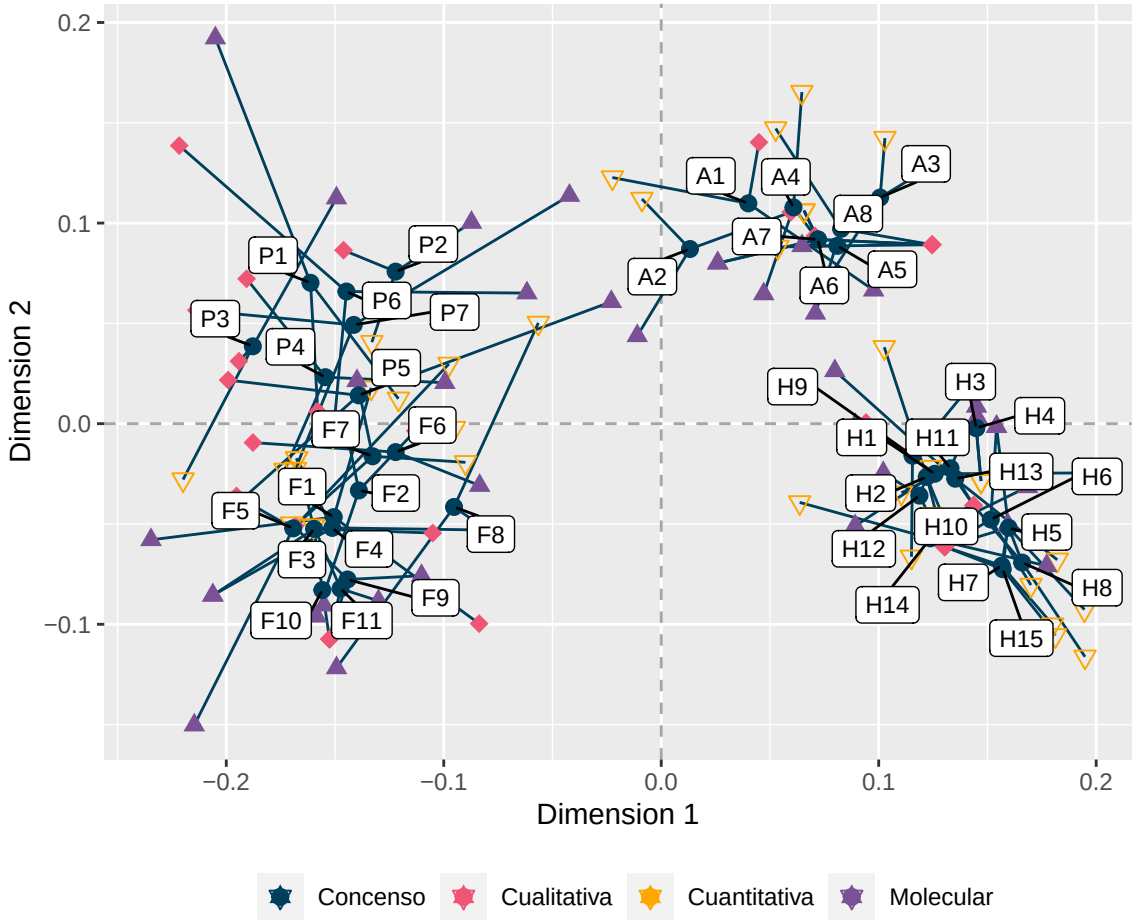
Al obtener las correlacion entre las matrices de distancias de las diferentes caracterizaciones sobre el plano principal, observamos que la correlacion mas baja se obtiene al comparar la caracterizacion cuantitativa con la molecular, siendo de 0.6. Este resultado no es para nada desalentador, por el contrario, nos indica que al incorporar una nueva dimension (caracteristicas moleculares) sobre la informacion cuantitativa que ya conociamos estamos adquiriendo nueva informacion, permitiendo conocer y explicar de una mejor forma el comportamiento y caracterizacion de los pepinos. Luego, la correlacion entre las matrices de distancias de las caracterizaciones cualitativa y molecular es de 0.67, indicando una concordancia media-alta. La correlacion entre las matrices de distancias de las caracterizaciones cualitativa y cuantitativa en el plano principal es de 0.78, lo cual nos indica una concordancia alta entre estas caracterizaciones.

A) Mida la concordancia entre la caracterización agronómica (cualitativa + cuantitativa) y molecular (planos principales)

Al medir la conconrdancia entre la caracterizacion agronomica y molecular observamos que la misma es de 0.66. Al igual que se planteo en el punto anterior, este resultado es alentador ya que nos confirma que al incorporar nueva informacion sobre las caracteristicas agronomicas de los pepinos se puede enriquecer aun mas el analisis.

A) Con APG halle el consenso entre las configuraciones (plano principal) obtenidas en base a datos cualitativos, cuantitativos y moleculares.

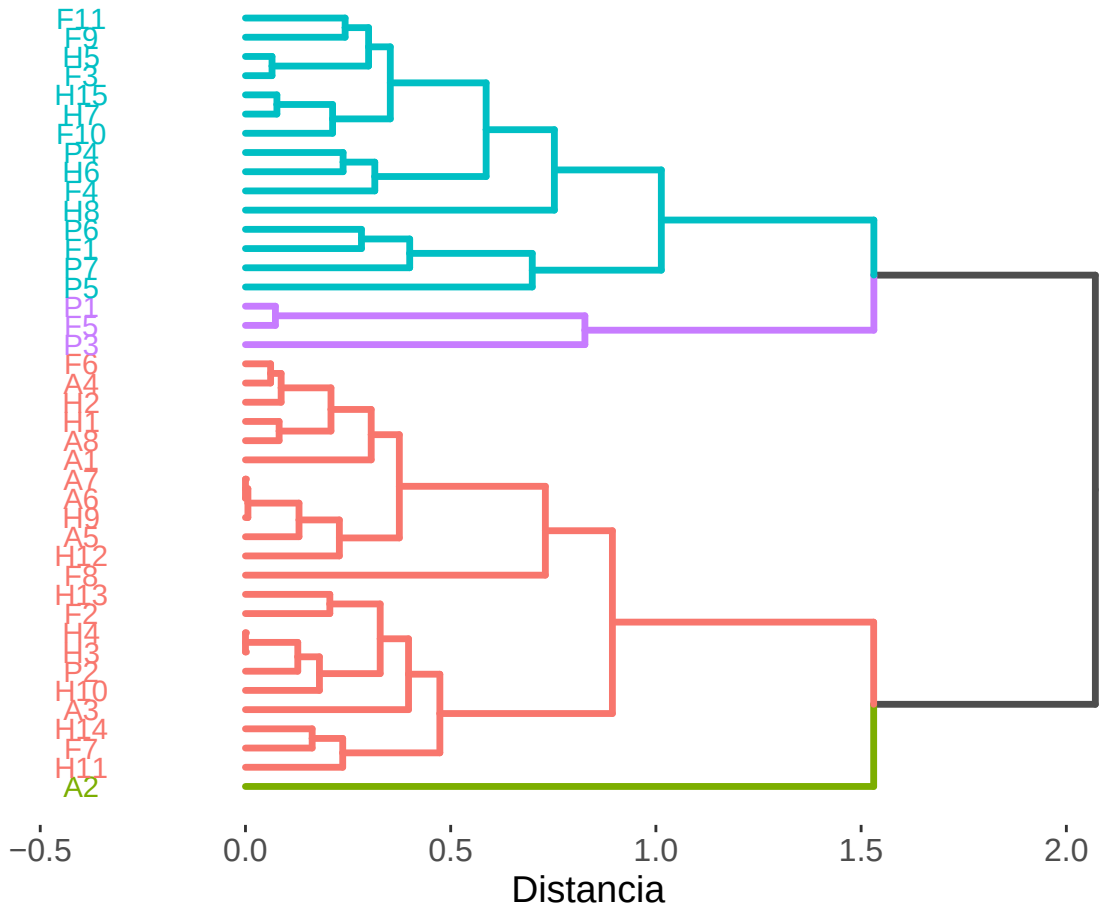
tipo	media
P	0.60
F	0.33
A	0.20
H	0.15



A) Identifique si hay algún tipo de pepino para el cual hay más discrepancia entre estas tres caracterizaciones

Para identificar si hay algun tipo de pepino que tiene mas discrepancia entre las caracteristicas cualitativas, cuantitativas y moleculares se calculan la sumas de cuadrados residuales promedio para cada una de las variedades. Como podemos observar, los pepinillos son los que presentan mayor discrepancia entre las caracterizaciones. Los Peninos Holandeses, tal como pudo observarse en la configuracion cualitativa son los mas parecidos **VER!**

A) Finalice el análisis con un cluster UPGMA obtenido a partir de la configuración de consenso


F11
F9
H5
F3
H15
H7
F10
P4
H6
F4
H8
P6
F1
P7
P5
P1
F5
P3
F6
A4
H2
H1
A8
A1
A7
A6
H9
A5
H12
F8
H13
F2
H4
H3
P2
H10
A3
H14
F7
H11
A2
−0.5
0.0
0.5
1.0
1.5
2.0
Distancia

Ejercicio 7

A) Cuantifique la concordancia de la configuración de poblaciones nativas de maíz en ambos ambientes en el espacio original mediante el coeficiente de correlación de Pearson entre matrices de distancias euclídeas estandarizadas entre individuos, y utilizando el coeficiente Rv.

Lo primero que hacemos es calcular las matrices de distancia entre las poblaciones de maiz a partir de las variables estandarizadas, uilizando la distancia euclidea.

```
dist_m1 <- dist(scale(datos_m1), method = "euclidean")
dist_m2 <- dist(scale(datos_m2), method = "euclidean")
correlacion <- cor(dist_m1, dist_m2)
```

La correlacion entre las matrices de distancia es igual a 0.676. Esto indica que existe una concordancia media-alta entre la configuracion de las poblaciones de maiz en ambos ambientes.

Por otro lado, tambien calculamos el coeficiente `RV`.

```
coef_rv <- coeffRV(scale(datos_m1), scale(datos_m2))
```

que resulta 0.567.

Mientras que la correlacion entre las matrices de distancia mide la similaridad entre las posiciones relativas de las poblaciones de maiz, en terminos de las variables medidas, el coeficiente RV mide directamente la correlacion entre los valores de estas variables para ambos ambientes.

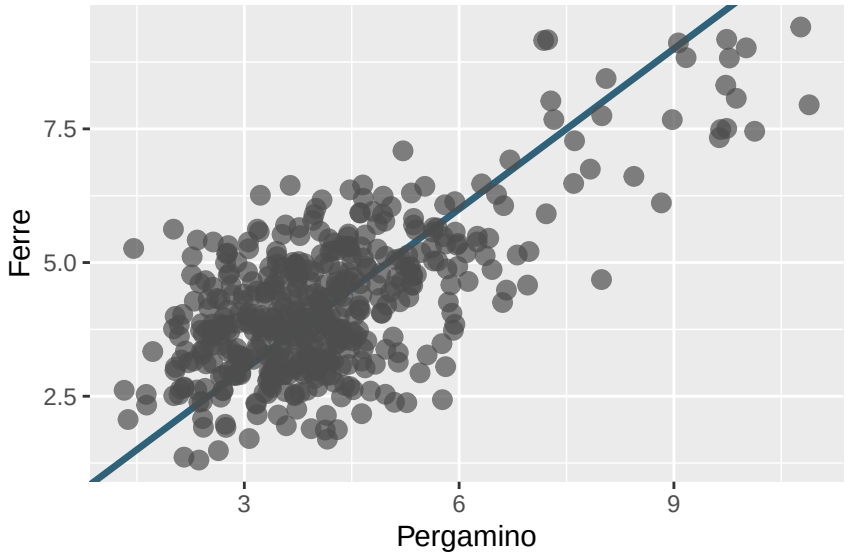


Figure 1: Distancia entre poblaciones para los ambientes Pergamino y Ferre. La linea azul representa a la recta identidad.

B) Realice un ACP para cada ambiente, compare semejanzas y diferencias entre ambas caracterizaciones tanto para individuos como para variables.

Utilizamos la sentencia `PCA()` de la libreria `FactoMineR`.

```
acp_m1 <- PCA(datos_m1, ncp = 2, graph = FALSE)
acp_m2 <- PCA(datos_m2, ncp = 2, graph = FALSE)
```

Y luego obtenemos los graficos para los individuos y para las variables en el plano principal.

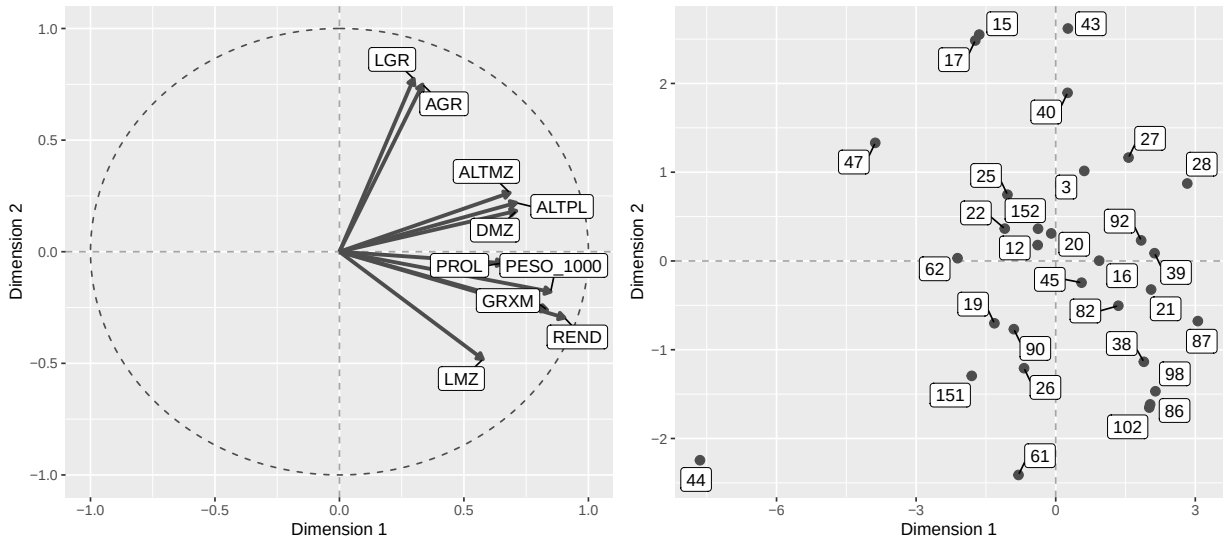


Figure 2: Caracterizacion de los individuos y las variables en el plano principal del ACP para el ambiente Pergamino. 64% variabilidad explicada.

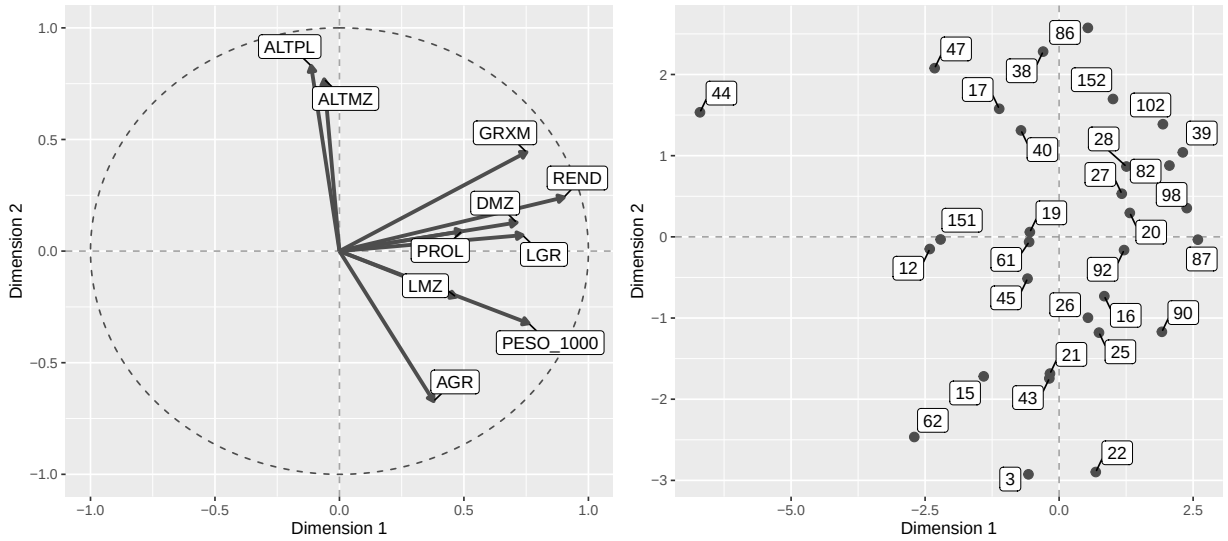


Figure 3: Caracterizacion de los individuos y las variables en el plano principal del ACP para el ambiente Ferre. 57% variabilidad explicada.

Respecto al grafico de las variables, vemos que en ambos ambientes se da que **rendimiento**, **peso cada 1000 granos**, y **granos por metro** son las que mas contribuyen a la primera dimension. En cambio, las variables que mas contribuyen al segundo eje dependen del ambiente. En el ambiente Pergamino se trata de las variables asociadas al grano, **largo** y **ancho**, mientras que en el ambiente Ferre se trata

de la **altura de la planta** y la **altura de la mazorca**, donde ademas el ancho del grano tiene una correlacion negativa cercana a -0.7.

El angulo entre los vectores asociados a las cargas nos brinda informacion sobre la relacion entre las variables, y podemos ver si estas relaciones varian segun el ambiente. Por ejemplo, en Pergamino se da que la altura de la planta y de la mazorca se relacionan positivamente con el rinde, indicando que plantas con mayor altura y mazorcas mas largas se asocian a rindes mayores. Sin embargo, esta asociacion no sucede en el ambiente Ferre, donde vemos por el angulo entre los vectores, que el rendimiento de la planta no se asocia a estas variables de altura. Otro ejemplo similar ocurre con la asociacion entre largo y ancho de grano, que en el ambiente Pergamino resultan altamente dependientes, mientras que en el ambiente Ferre su correlacion es casi nula.

En cuanto al grafico de los individuos, lo primero que se observa es que la poblacion **44** presenta en ambos ambientes un comportamiento muy diferente al resto. En ambos ambientes tiene valores muy bajos en la primer componente, lo que indica que se trata de una poblacion con un rendimiento y un peso de grano muy inferior al resto de las poblaciones. Sin embargo, en el ambiente Pergamino se corresponde con plantas de poca altura con mazorcas cortas, mientras que en el ambiente Ferre se presenta plantas altas con mazorcas altas.

C) Cuantifique la relación de las dos configuraciones en el plano principal originado por los ACP

Para cada ambiente obtenemos la matriz de distancia entre los puntos en el plano principal y calculamos la correlacion entre ambas matrices.

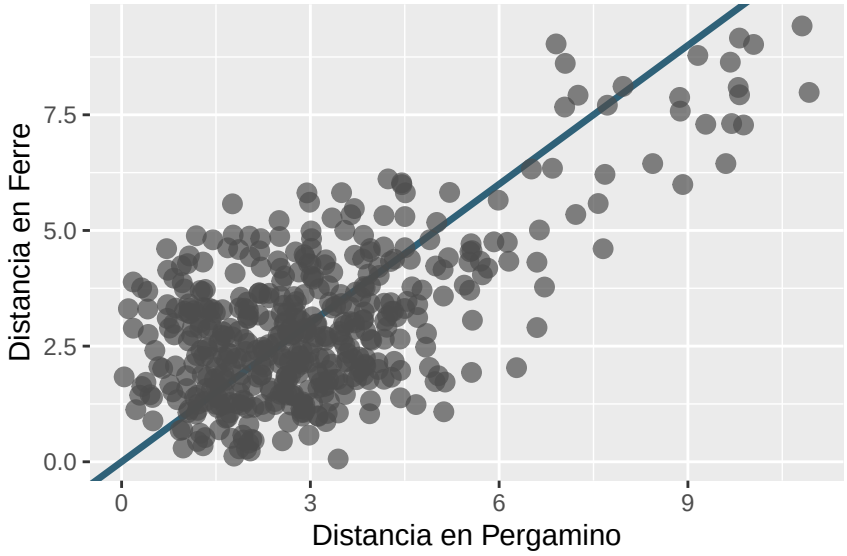


Figure 4: Distancia en Pergamino vs. distancia en Ferre para cada poblacion de maiz.

La correlacion entre ambas matrices de distancia en el plano principal es de 0.642. A partir de este valor podemos decir que hay una concordancia media o media-alta entre las configuraciones de las poblaciones en ambos ambientes. Esto significa que el comportamiento de las poblaciones de maiz es en general similar en ambos ambientes, pero tambien presenta caracteristicas distintas en cada uno de ellos.

D) Encuentra indicios de interacción tanto genotipo-ambiente como variable-ambiente ?

Si, en ambos casos, por todo lo explicado en el inciso **b**. Como ejemplo de la interaccion genotipo-ambiente tenemos a la poblacion 44, que en el ambiente Pergamino se corresponde con plantas bajas y mazorcas cortas, mientras que en Ferre son plantas altas y mazorcas largas.

Y como ejemplo de la interaccion variable-ambiente tenemos que en Pergamino se da que la altura de la planta y de la mazorca se relacionan positivamente con el rinde, indicando que plantas con mayor altura y mazorcas mas largas se asocian a rindes mayores. Sin embargo, esta asociacion no sucede en Ferre, donde vemos que el rendimiento de la planta no se asocia a estas variables de altura.

E) Como se quiere encontrar una caracterización 'media' o 'promedio' para las 31 poblaciones en función de la información dada en ambos ambientes proceda a realizar una ACP sobre el promedio de las variables para ambos ambientes

Obtenemos un nuevo data frame que representa al promedio entre ambos ambientes, y realizamos el ACP como lo hicimos anteriormente.

```
datos_media <- (datos_m1 + datos_m2) / 2
acp_media <- PCA(datos_media, ncp = 2, graph = FALSE)
```

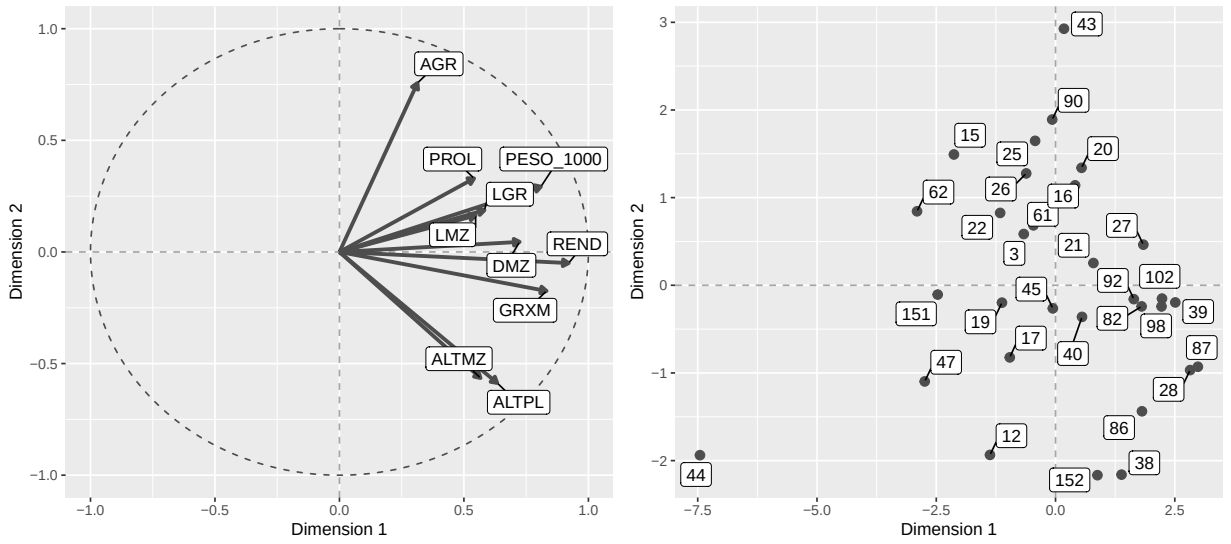


Figure 5: Caracterizacion de los individuos y las variables en el plano principal del ACP para el promedio de ambientes. 60% variabilidad explicada.

En el panel izquierdo de la Figura 5 vemos que el ACP realizado sobre el promedio de las variables en ambos ambientes presenta una primera coordenada que esta muy correlacionada positivamente con el rendimiento, el peso cada 1000 granos, la cantidad de granos por metro, la altura de la planta, y la altura de la mazorca. Por otro lado, la segunda dimension esta relacionada, principalmente, con el ancho del grano.

Cuando el comportamiento de la poblacion es consistente en ambos ambientes, vemos que esa informacion tambien se ve reflejado aqui. Por ejemplo, la variedad 43 presenta rendimientos medios y granos anchos y largos tanto en Ferre como Pergamino, que se corresponde con lo que se observa en este ACP basado en el promedio.

Sin embargo, observando el angulo que se forma entre los vectores que representan a las variables comenzamos a notar ciertos problemas con esta representacion. Por ejemplo, el angulo entre ancho de grano (`AGR`) y altura de mazorca (`ALTMZ`) y altura de planta (`ALTPL`) es de un poco mas de 90 grados, indicando casi una total independencia entre `AGR` y el resto. Sin embargo, este comportamiento *promedio*, no representa lo que sucede en ninguno de los dos ambientes. En el ambiente Pergamino `AGR` tiene una correlacion media con `ALTMZ` y `ALTPL` (angulo aproximado 45 grados) y en el ambiente Ferre tiene una correlacion negativa muy fuerte (angulo aproximado 180 grados).

F) Ahora concatene ambos archivos por filas y columnas y realice un ACP para ambas situaciones, encuentra respuesta para las interacciones planteadas en el inciso b ?

```
datos_long <- rbind(datos_m1, datos_m2)
acp_long <- PCA(datos_long, ncp = 2, graph = FALSE)
```

```
datos_wide <- cbind(datos_m1, datos_m2)
acp_wide <- PCA(datos_wide, ncp = 2, graph = FALSE)
```

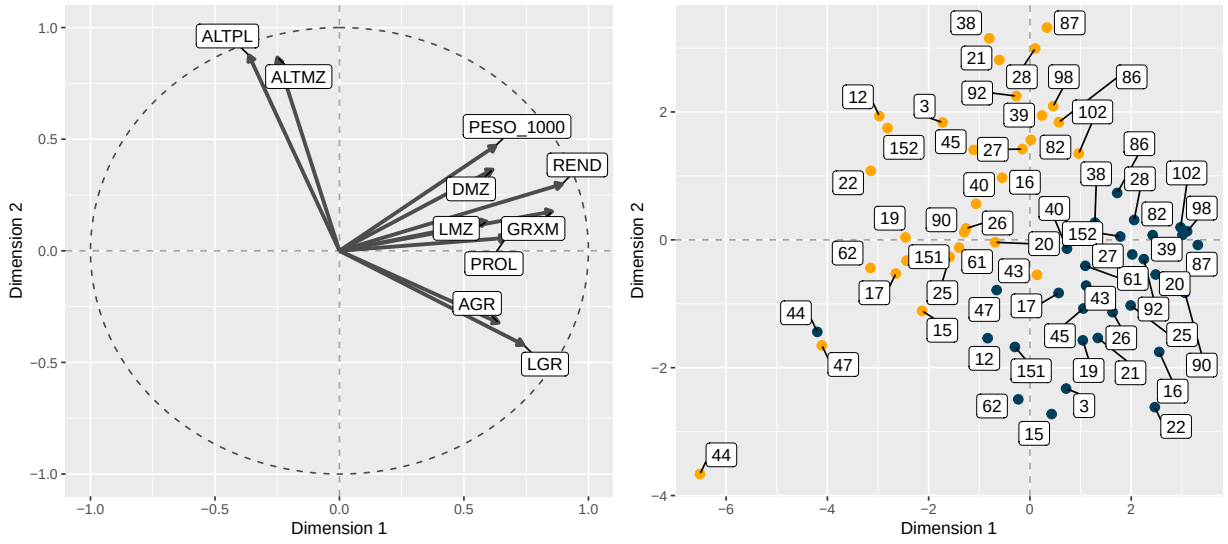


Figure 6: Caracterizacion de los individuos y las variables en el plano principal del ACP al concatenar filas . 65% variabilidad explicada. Ferre en azul, Pergamino en amarillo.

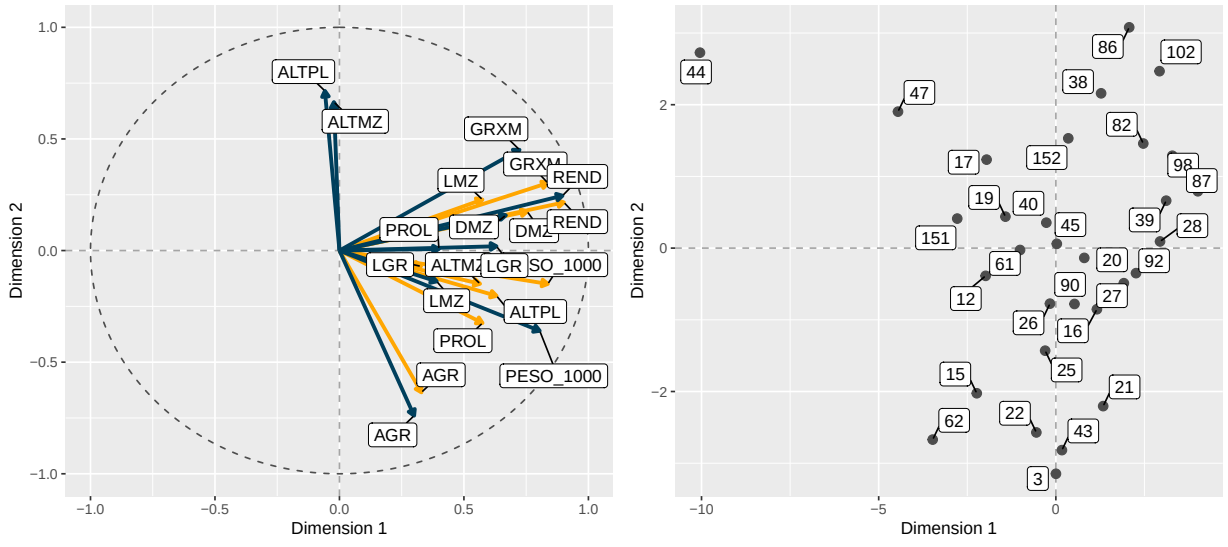


Figure 7: Caracterizacion de los individuos y las variables en el plano principal del ACP para al concatenar columnas. 51% variabilidad explicada. Ferre en azul, Pergamino en amarillo.

De la Figura 6 podemos apreciar que la primer dimension esta muy relacionada con el rinde de la

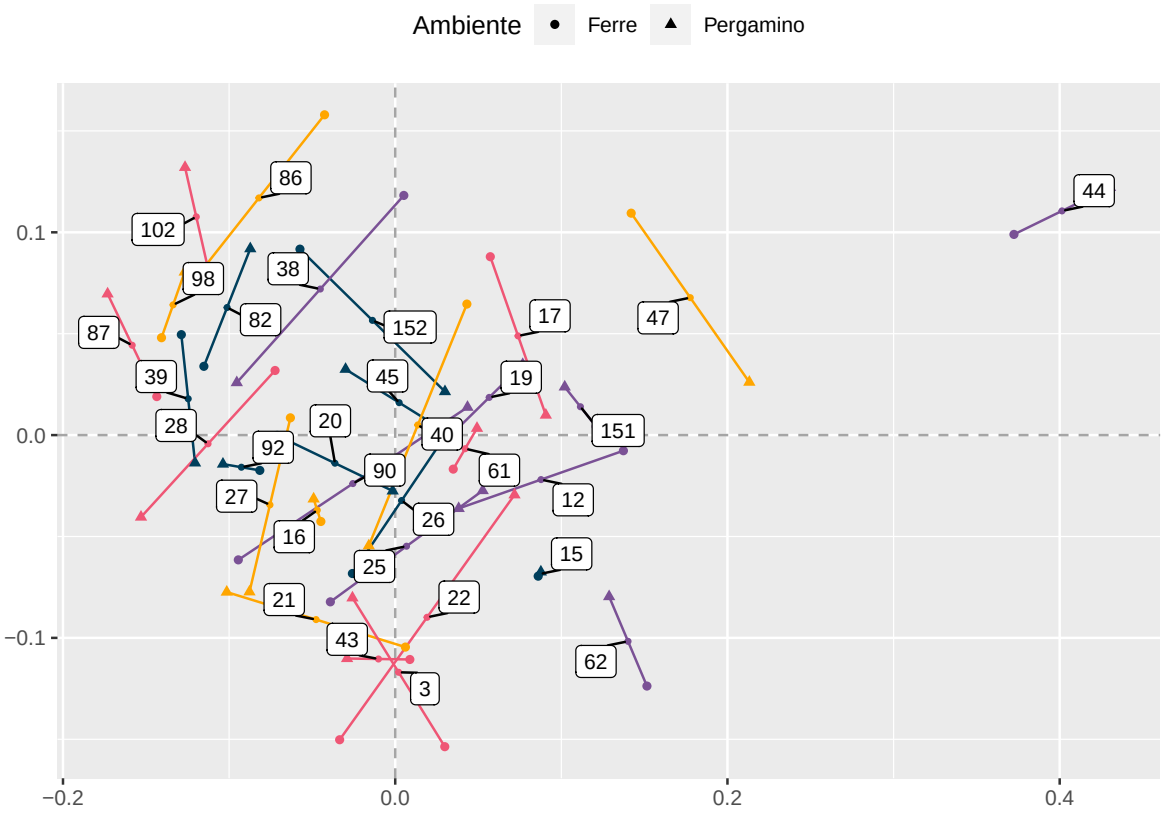
cosecha, la cantidad de granos por metro y el ancho del grano. La segunda dimension, en cambio, representa variables de altura, como son la altura de la planta y de la mazorca. Al observar el panel derecho de esta figura, podemos ver que los cultivos en Ferre presentaron en general mejores rindes, pero menores tamanos de planta, mientras que las plantas de Pergamino fueron mas grandes, pero con menor rinde. En otras palabras, en Pergamino se presentaron mejores caracteristicas en terminos de la planta, pero en Ferre se presentaron mejores caracteristicas en cuanto al grano y el rendimiento de la cosecha.

En el panel izquiero de la Figura 7 podemos ver, por ejemplo, que el primer eje esta asociado al rendimiento, en ambos ambientes. Por lo tanto, las poblaciones que se encuentren hacia la derecha en el panel derecho presentaron rindes altos tanto en Pergamino como en Ferre. En cambio, el segundo eje representa a la altura de planta y mazorca solo para el ambiente Ferre. Alli podemos ver que la variedad 44 en el ambiente Ferre resulto ser de plantas altas, con mazorcas altas, pero rinde muy por debajo del promedio. Por otro lado, la variedad 102 tambien se caracterizo por su altura, pero tambien presento rindes elevados. Finalmente, notamos que las poblaciones 44 y 47 se destacan por haber tenido rindes muy pobres en ambos ambientes.

G) Aplicar APG para tener otra visualización de la interacción genotipo-ambiente (retenga todas las dimensiones de ambas configuraciones).

```
acp_m1 <- PCA(datos_m1, ncp = 10, graph = FALSE)
acp_m2 <- PCA(datos_m2, ncp = 10, graph = FALSE)
df <- data.frame(cbind(acp_m1$ind$coord, acp_m2$ind$coord))
gpa <- GPA(df, group = c(10, 10), name.group = c("Pergamino", "Ferrre"),
           graph = FALSE)
```

H) Visualizar del gráfico correspondiente las 3 poblaciones con mayor efecto ambiente y las 3 con menor efecto



A la hora de observar el grafico hay que tener en cuenta que la escala de los ejes no es la misma. El eje horizontal va desde -0.2 a 0.45 y el eje vertical va desde -0.175 a 0.175 aproximadamente.

Las tres poblaciones con mayor efecto ambiente son 22, 90 y 38. Por otro lado, las tres poblaciones con menor efecto ambiente son la 15, 16 y 98.

Tambien podemos obtener la longitud de los segmentos y obtener a los que presentan mayor y menor efecto ambiente de manera numerica.

	Longitud del segmento		Longitud del segmento
15	0.00001	38	0.01863
16	0.00014	90	0.02470
92	0.00050	22	0.02567

I) Comparar lo encontrado en h) con el ANOVA correspondiente. Realice comentarios

Imprimimos las 3 poblaciones con mayor y menor efecto ambiente segun la suma de cuadrados residual.

	SSresidual		SSresidual
61	0.14821	17	1.10177
102	0.20504	90	1.13240
92	0.21264	22	1.30115

Alli vemos que la informacion que obtenemos en el plano de dos dimensiones no es exactamente la misma que la que se obtiene al calcular los residuos utilizando las ubicaciones de los puntos en el espacio original.

Aca podemos observar que los ambientes 22 y 90 siguen siendo los que mas efecto ambiente presentan, pero aparece el ambiente 17 en tercer lugar. Por otro lado, los ambientes 61 y 102 no aparecieron cuando buscamos los ambientes con menor efecto, y ahora figuran como los dos con menor residuo.

Estas diferencias se dan porque en el grafico estamos analizando una proyeccion de los datos en un plano de dos dimensiones, cuando en realidad se encuentran en un espacio de 10 dimensiones.