Facultad de Ciencias.
Licenciatura en Actuaría.
Seminario de Estadística : Aprendizaje estadístico automatizado
Semestre 2020-1 Grupo 9269

Yanely Luna Gutiérrez
Ana Beatriz Ortiz Silva

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Tarea 2
Análisis de Componentes Principales y Análisis de Conglomerados

3 de Noviembre 2020

1. Ejercicio 3

del capítulo 10 Linear Regression, de James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning. With applications in R, Springer, ISL*

En este ejercicio se realizará el método de *k-means* manualmente con $K = 2$ para una muestra con $n = 6$ observaciones y $p = 2$ variables. Las observaciones son:

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

a. Gráfica de las observaciones.

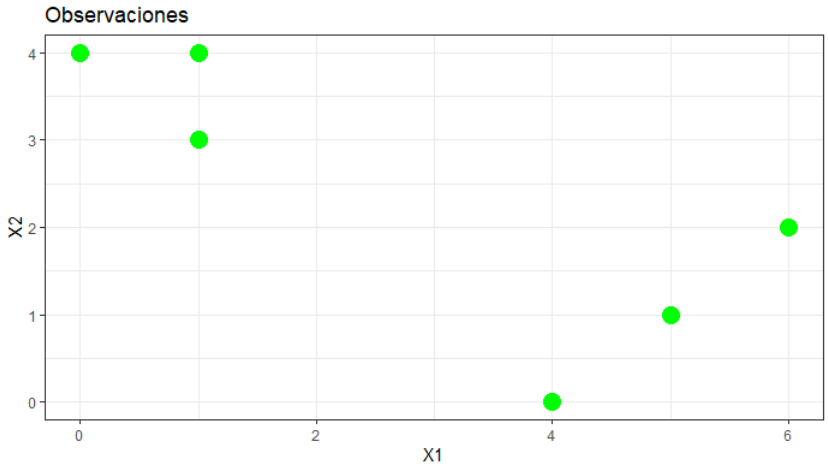

Figura 1: **Gráfica de las observaciones en X_1 y X_2.** *A primera vista se podrían identificar dos grupos que particionan las observaciones: un grupo en la esquina superior izquierda y uno en la esquina inferior derecha.*

b. Asignamos de manera aleatoria un grupo (1 ó 2) a cada observación. El resultado de esta asignación se muestra en el Cuadro 1.

Observación	x_1	x_2	Grupo
1	1	4	2
2	1	3	2
3	0	4	1
4	5	1	1
5	6	2	2
6	4	0	2

Cuadro 1: **Primer agrupación.** *De forma aleatoria.*

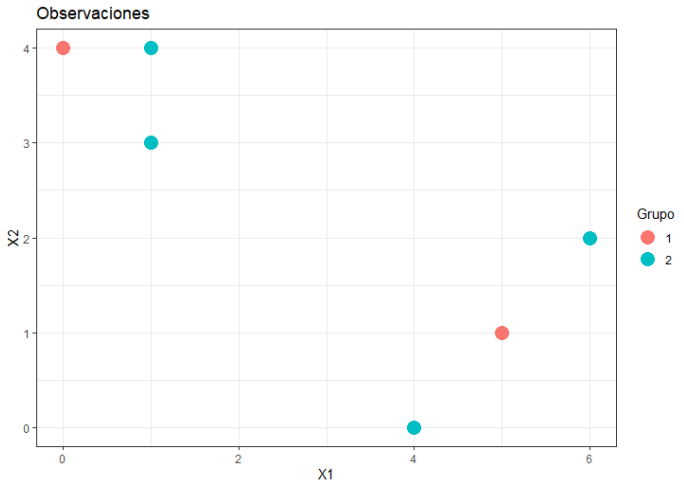

Figura 2: **Primer agrupamiento.**

c. Ahora que tenemos este primer agrupamiento calculamos el centroide de cada uno, el cual es un vector con dos entradas: la media del grupo en X_1 (c_1) y la media del grupo en X_2 (c_2). Los resultados se muestran en el Cuadro 2.

Grupo	c_1	c_2
1	2.5	2.5
2	3	2.25

Cuadro 2: **Centroides de los grupos del primer agrupamiento.**

d. Reasignamos cada observación al grupo con el centroide más cercano. Para esto calculamos la distancia euclidiana entre la observación (x_1, x_2) y el centroide (c_1^i, c_2^i) con $i = 1, 2$. Por ejemplo, si la distancia entre la observación y (c_1^1, c_2^1) es menor que la distancia entre esta misma observación y (c_1^2, c_2^2) asignamos dicha observación al grupo 1. El resultado de este proceso se muestra en el Cuadro 3.

Observación	X_1	X_2	Agrupación 2
1	1	4	1
2	1	3	1
3	0	4	1
4	5	1	2
5	6	2	2
6	4	0	2

Cuadro 3: **Segundo agrupamiento.** *Agrupando cada observación al grupo con centroide más cercano.*

e. Repetimos c. y d. hasta que los nuevos grupos formados no cambien respecto a los anteriores. En este caso los grupos formados en la tercera agrupación son los mismos que se muestran en el Cuadro 3, por lo que terminamos el proceso y nos quedamos con este agrupamiento.

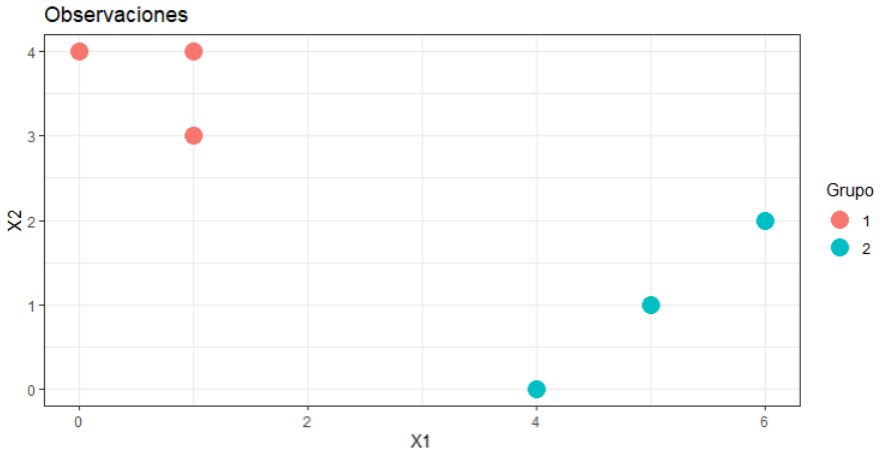

Figura 3: **Agrupamiento final.** *Los dos grupos formados son los que se esperaban intuitivamente al ver la primera gráfica.*

2. Ejercicio 9

Usando la base de datos `USArrests` se realizará un agrupamiento jerárquico para los estados.

a. Agrupamos los estados de acuerdo al método *complete linkage* y usando la distancia euclidiana. Elegimos separar en tres conglomerados a los dendogramas para visualizar mejor los datos.

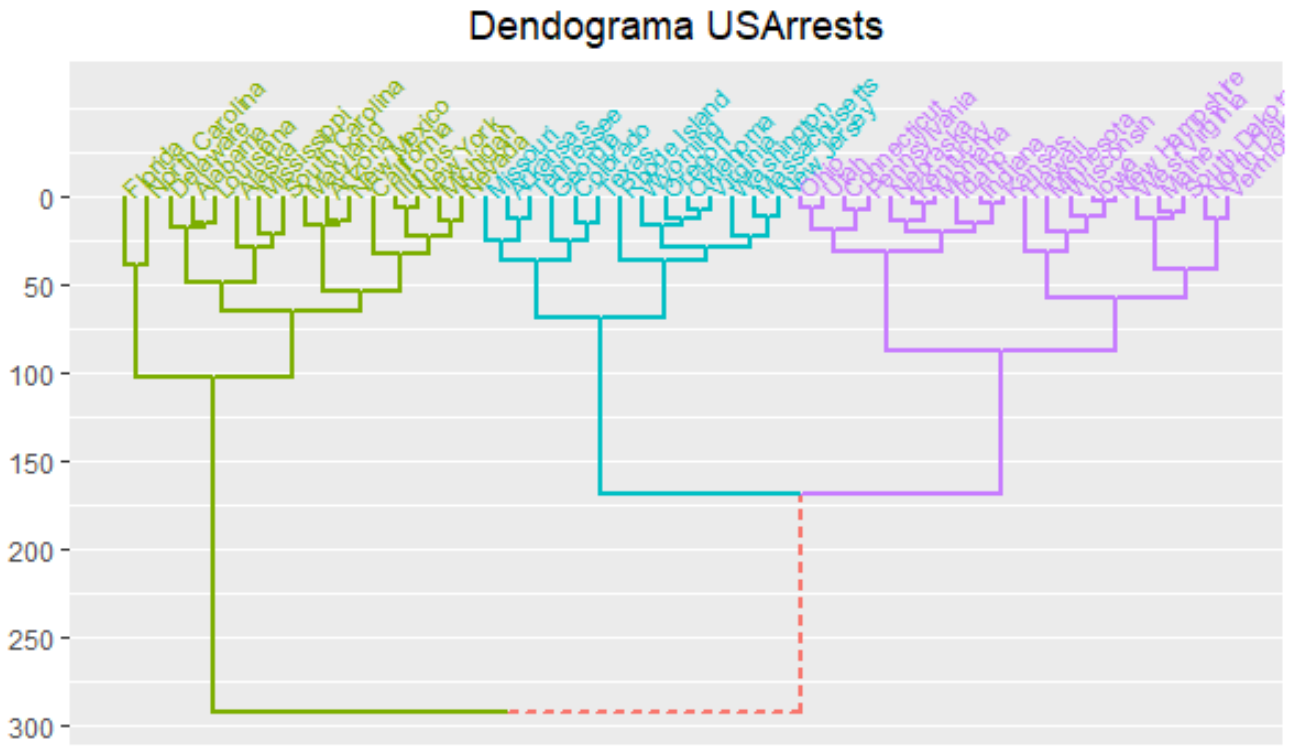

Figura 4: **Dendograma USArrest.** *Agrupación jerárquica de USArrest usando método complete, podemos observar una partición intuitiva de los estados.*

b. Cortamos el dendograma en k=3 y notamos los siguientes agrupamientos en el cuadro 4.

Agrupación 1	Agrupación 2	Agrupación 3
Alabama	Arkansas	Connecticut
Alaska	Colorado	Hawaii
Arizona	Georgia	Idaho
California	Missouri	Indiana
Delaware	New Jersey	Iowa
Florida	Oklahoma	Kansas
Illinois	Oregon	Kentucky
Louisiana	Rhode Island	Maine
Maryland	Tennessee	Minnesota
Michigan	Texas	Montana
Mississippi	Virginia	Nebraska
Nevada	Washington	New Hampshire
New Mexico	Wyoming	North Dakota
New York	*	Ohio
North Carolina	*	Pennsylvania
South Carolina	*	South Dakota
*	*	Utah
*	*	Vermont
*	*	West Virginia
*	*	Wisconsin

Cuadro 4: **Agrupamiento con corte en k=3** *De acuerdo a la distancia euclidiana, la separación de los datos es semejante en los tres conglomerados.*

c. Escalamos las variables para tener desviación estándar uno.

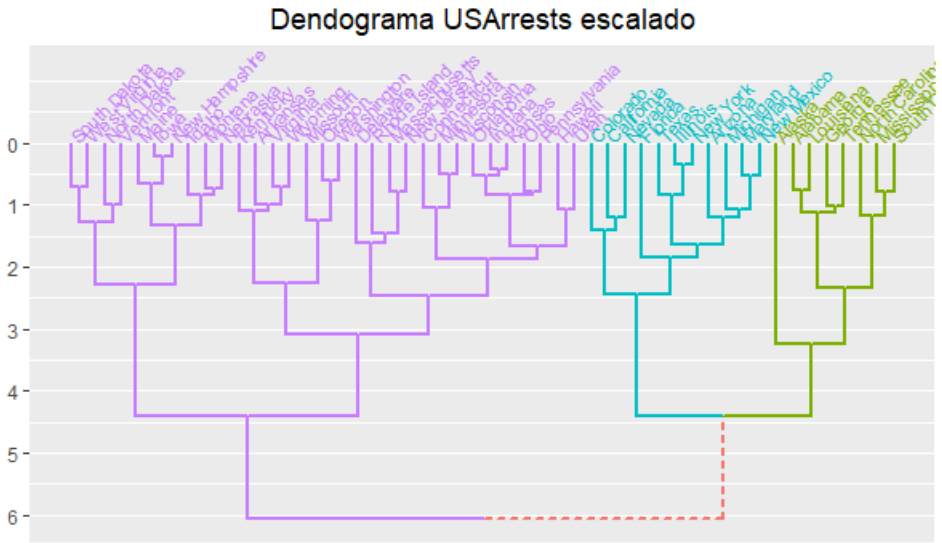

Figura 5: **Dendograma USArrest escalado** *Agrupamiento jerárquico de los datos escalados. Podemos notar una distribución distinta de las conglomeraciones en relación con la figura anterior.*

Hacemos un corte en k=3 y notamos como los estados se agrupan en el cuadro 7.

Agrupación 1	Agrupación 2	Agrupación 3
Alabama	Arizona	Arkansas
Alaska	California	Connecticut
Georgia	Colorado	Delaware
Louisiana	Florida	Hawaii
Mississippi	Florida	Hawaii
North Carolina	Illinois	Idaho
South Carolina	Maryland	Indiana
Tennessee	Michigan	Iowa
*	Nevada	Kansas
*	New Mexico	Kentucky
*	New York	Maine
*	Texas	Massachusetts
*	*	Missouri
*	*	Montana
*	*	Nebraska
*	*	New Hampshire
*	*	New Jersey
*	*	North Dakota
*	*	Ohio
*	*	Oklahoma
*	*	Pennsylvania
*	*	Rhode Island
*	*	South Dakota
*	*	Utah
*	*	Vermont
*	*	Virginia
*	*	Washington
*	*	West Virginia
*	*	Wisconsin
*	*	Wyoming

Cuadro 5: **Agrupamiento escalado de los estados.** *Las proporciones entre las conglomeraciones ya no es semejante y cambia radicalmente respecto al cuadro anterior.*

d. ¿Escalar antes o después de computar las observaciones?

Observar los datos antes y después de escalar nos puede dar una perspectiva de los cambios de agrupación de los estados. Sin embargo, al final, son con las variables escaladas con las que vamos a trabajar y nos darán una mejor descripción de la base.

Podemos observar esa diferencia significativa en los cuadros 6 y 7 que explican las medias de los tipos de arrestos de cada estado antes y después de escalar. También lo podemos ver en los cuadros 4 y 5 , ya que el cuadro 5 es más estricto al determinar el acercamiento de los estados según la distancia euclidiana

Agrupación	Murder	Assault	UrbanPop	Rape
1	11.81	272.56	68.31	28.37
2	8.21	173.28	70.64	22.84
3	4.27	87.55	59.75	14.39

Cuadro 6: **Medias según tipo de arresto.** *La agrupación con medias mayores es la 1 y la agrupaión con medias menores es la 3 aunque solo se difieran de 4 estados.*

Agrupación	Murder	Assault	UrbanPop	Rape
1	14.09	252.75	53.50	24.53
2	11.05	264.09	79.09	32.61
3	5.00	116.48	63.84	16.33

Cuadro 7: **Medias según tipo de arresto escalado.** *Ahora la agrupación con medias mayores es la 2 y la agrupaión con medias menores es la 3 exceptuando en Urban pop.*

Podemos explicar el aumento de las medias en grupo 3 por el escalonamiento de las variables, ya que gracias a eso el grupo 3 tiene muchos más estados que antes y por lo tanto su media aumenta.

3. Ejercicio 10

En este ejercicio se simularán datos para posteriormente realizar análisis de componentes principales y agrupamiento con el método de *k-means*.

a. Generar un conjunto de datos con 20 observaciones en cada una de tres clases sobre 50 variables.

   Organizamos estos datos en una tabla con 60 observaciones (20 de cada clase) y 51 variables (1 representa la clase a la que corresponden: A, B o C). Para las observaciones simulamos variables aleatorias normales con la diferencia que en la clase A estas tienen media 0, en la clase B tienen media 1 y en la clase C tienen media 2. Todas tienen varianza igual a 1.

b. Realizamos el análisis de componentes principales sobre los datos.

	PC1	PC2	PC3	PC4	PC5
Desviación estándar	5.71	1.82	1.74	1.64	1.60
Proporción de la varianza	0.42	0.04	0.04	0.03	0.03
Proporción acumulada	0.42	0.46	0.50	0.53	0.56

Cuadro 8: **Primeros cinco componentes principales.** *A partir del cuarto componente principal tenemos más del cincuenta por ciento de la varianza explicada.*

En el Cuadro 8 notamos que el primer componente explica un gran porcentaje de la varianza y a partir del segundo solo aumenta menos del cinco por ciento para cada componente que sigue. En la Figura 6 se muestran los valores de las observaciones en los primeros dos componentes principales y se nota una clara separación de las tres clases de observaciones.

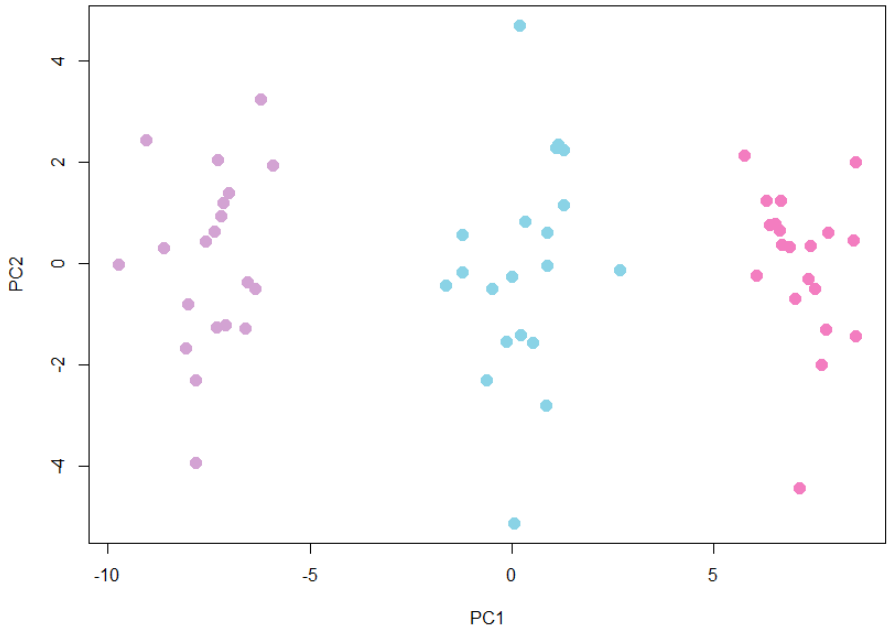

Figura 6: **Gráfica de los primeros dos componentes principales.** *Las observaciones están coloreadas según la categoría a la que pertenecen.*

c. Formamos grupos usando el método de *k-means* con $K = 3$ sobre el conjunto de observaciones. Para ver que tan parecido es el agrupamiento resultante respecto al original usamos el comando `table` y notamos primero que cada grupo formado tiene 20 observaciones, lo que coincide con los grupos originales. Sin embargo, esto no nos indica que cada grupo contenga las mismas observaciones que los grupos originales.

   Para ver si cada grupo formado tiene las 20 observaciones correspondientas a un mismo grupo original seleccionamos las observaciones del conjunto que recibieron la misma etiqueta como resultado de *k-means* y revisamos si estas pertenecen a un mismo grupo original. Para esto usamos el comando `distinct` de la paquetería `dplyr` y vimos cuantas observaciones de grupos distintos formaron un mismo grupo.

Grupo formado \ Grupo original	A	B	C
1	0	20	0
2	20	0	0
3	0	0	20

Cuadro 9: **Grupos formados con** $K = 3$. *Cada grupo contiene observaciones de un mismo grupo original.*

d. Realizamos ahora un agrupamiento con $K = 2$ usando *k-means*.

   Usamos un proceso similar al ejercicio anterior para ver la estructura de los grupos resultantes.

Grupo formado \ Grupo original	A	B	C	Total de observaciones
1	20	0	0	20
2	0	20	20	40

Cuadro 10: **Grupos formados con** $K = 2$.

En este caso, el primero grupo contiene solo observaciones del grupo A, mientras que el segundo tiene todas las observaciones de los grupos B y C. El método hizo un buen agrupamiento al dejar un grupo igual al grupo original, pues el método pudo distingir claramente a un grupo del resto.

e. Con $K = 4$. En este caso también fue posible distinguir dos grupos originales (A y B), mientras que el las observaciones del grupo C fueron separadas para formar dos grupos. Esto tal vez sea debido a que las observaciones del grupo C resultaron tener más diferencia entre sí que las de los otros grupos.

Grupo formado \ Grupo original	A	B	C	Total de observaciones
1	0	20	0	20
2	0	0	6	6
3	0	0	14	14
4	20	0	0	20

Cuadro 11: **Grupos formados con** $K = 4$.

f. Realizamos *k-means* con $K = 3$ pero ahora sobre los primeros dos componentes principales en lugar de hacerlo sobre las observaciones originales (como en los incisos anteriores).

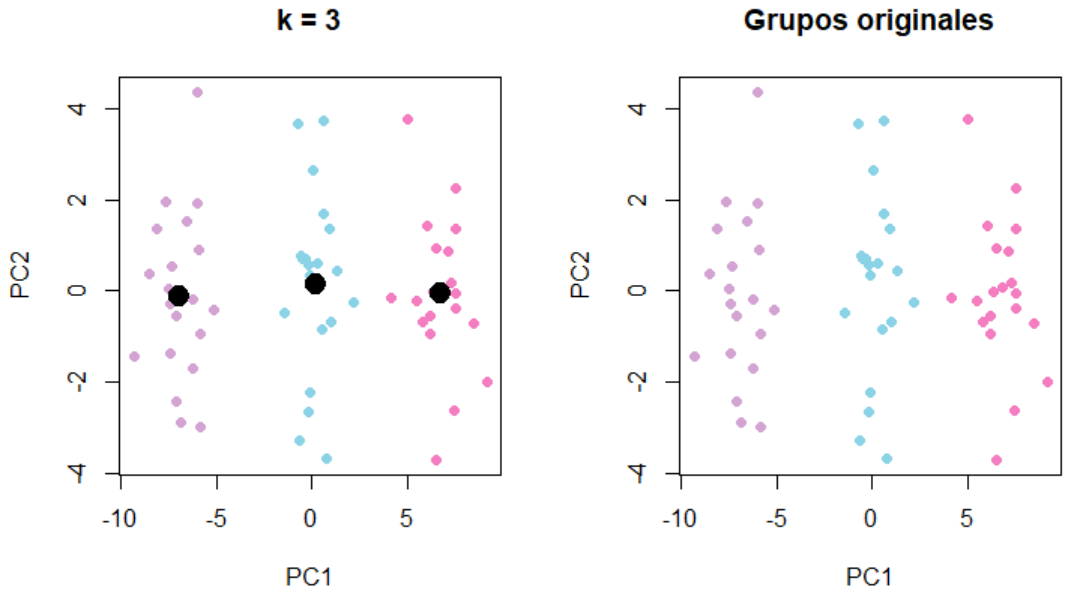

Figura 7: **Grupos formados vs Grupos originales**. *Los puntos negros indican la ubicación del centroide de cada grupo.*

En la Figura 7 podemos observar que los grupos identificados con *k-means* coinciden con los grupos originales, al igual que pasó al aplicar el método sobre las observaciones, sin embargo al aplicarlo ahora sobre los primeros dos componentes principales nos es posible visualizar el agrupamiento en vez de solo analizarlo con una tabla.

g. Aplicamos *k-means* con $K = 3$ pero ahora sobre las variables escaladas con desviación estándar igual a 1 usando el comando `scale`.

   El resultado es el mismo que obtuvimos en el inciso b. Los tres grupos formados coinciden con los grupos originales, como se observa en el Cuadro 12.

Grupo formado \ Grupo original	A	B	C
1	20	0	0
2	0	20	0
3	0	0	20

Cuadro 12: **Grupos formados con** $K = 3$. *Cada grupo contiene observaciones de un mismo grupo original.*

Este resultado no es inesperado pues las variables fueron simuladas como normales con varianza 1.

4. Ejercicio 11

Trabajaremos con la base de datos `Ch10Ex11.csv`

a. La base de datos es un conjunto de datos de expresión génica. Consta de 40 muestras de tejido con mediciones de 1.000 genes. Las primeras 20 muestras son de pacientes sanos, mientras que las segundas 20 son de un grupo de enfermos.

b. Realizamos agrupamientos jerárquicos a nuestros datos utilizando la distancia basada en la correlación. Elegimos separar en dos conglomerados a los dendogramas para visualizar mejor los datos.

   - Método single

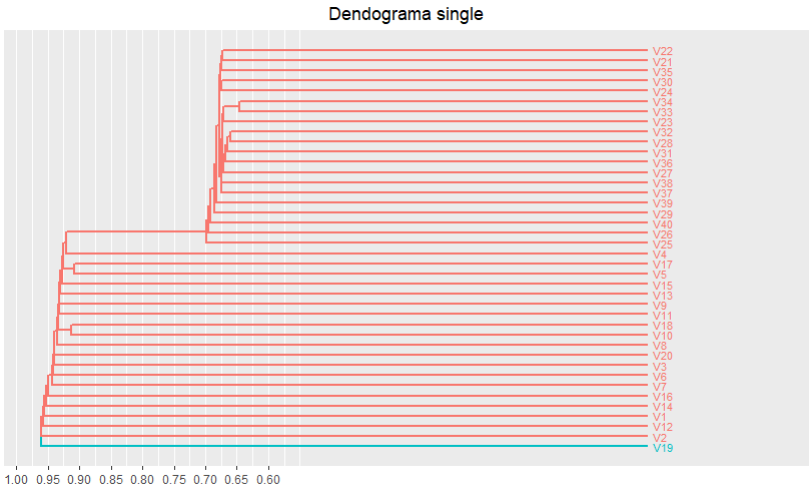

Figura 8: **Dendograma single** *No tenemos una buena agrupación. Para poder observar una mejor agrupación de los datos deberíamos hacer un corte en k=2.El dendograma es inestable*

- Método complete

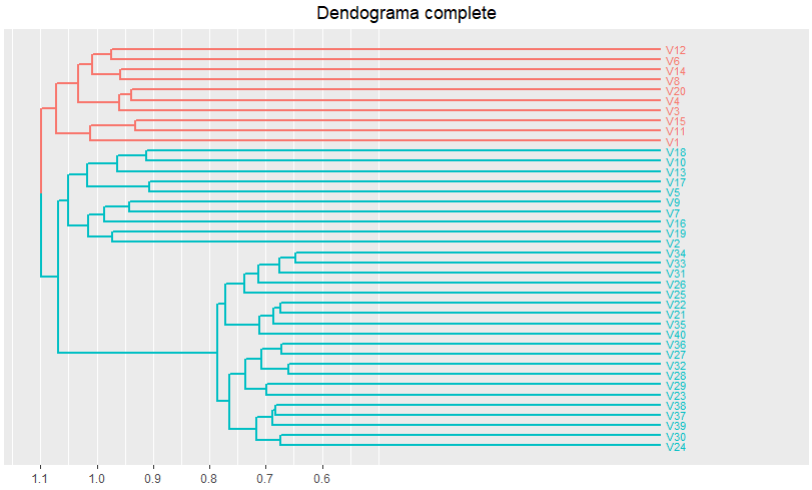

Figura 9: **Dendograma complete.** *Aquí se observa una partición mas clara de las muestra y más proporcionada.*

- Método centroid

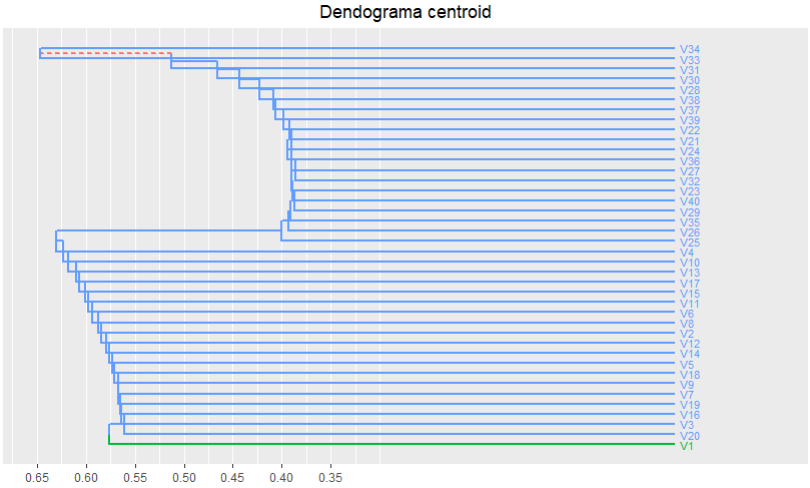

Figura 10: **Dendograma centroid** *Tenemos los mismos problemas que en la agrupación de single pero con una distribución más curveada.*

- Método average

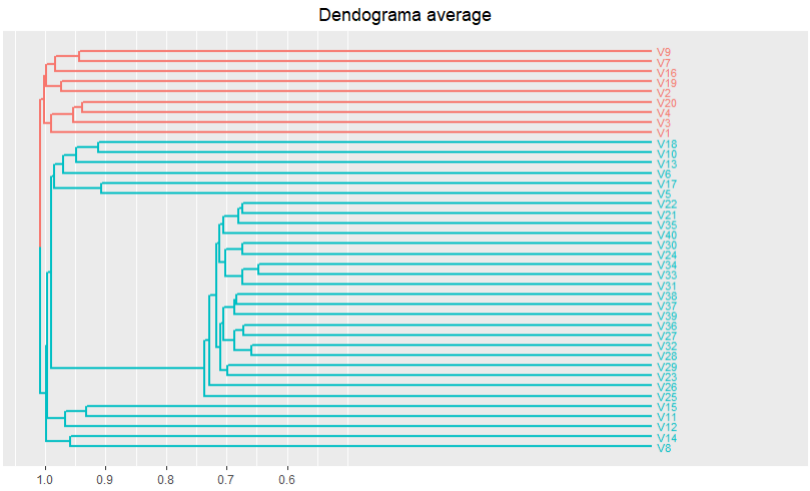

Figura 11: **Dendograma average** *Aquí ya tenemos al menos tres agrupaciones importantes por lo que average se aleja mucho del objetivo.*

Con esto podemos afirmar que el tipo de *linkage* usado determinará el grado de significancia de la separación deseada.

En los métodos *average* y *complete* los genes computan una agrupación inicial más marcada. Aunque sólo en *complete* es inicialmente binaria. Sin embargo una vez hecho un corte en el dendograma, tanto el método single como el centroid, tendrán una separación más pronunciada.

Con el método *complete* tenemos una partición que deja en un grupo parte del grupo original 1, (los que están sanos) pero en el otro grupo tenemos todos los del grupo original 2 (enfermos) y la mitad del grupo original 1, sin embargo, es el método que más se acerca a la partición original de los datos (sanos y enfermos).

c. ¿Qué genes difieren más entre los dos grupos?

Realizamos un análisis de componentes principales. Sacamos los valores absolutos de la cargas totales de cada gen para obtener su peso característico y después creamos un indíce para encontrar los primeros i genes que más difieren.

Posición	Gen
1	865
2	68
3	911
4	428
5	624
6	803
7	524
8	980
9	822
10	25

Cuadro 13: **Los 10 genes que más difieren.**