4to Punto

```
#librerias
library(FactoMineR)#PCA
library(factoextra)
library(gridExtra)#par
library(kableExtra)#tablas
library(dplyr)
```

[PCA, K-medias] En este ejercicio Ud va a generar un conjunto simulado de datos y entonces aplicará PCA y agrupamiento por k-medias sobre dichos datos.

a)

Genere un conjunto de datos simulados con 20 observaciones en cada una de tres clases (es decir, 60 observaciones en total) y 50 variables. Sugerencia: hay una serie de funciones en R que puede utilizar para generar datos. Un ejemplo es la función rnorm(); runif() es otra opción. Asegúrese de agregar un cambio en la media en las observaciones de cada clase a fin de obtener tres clases distintas.

Solución

Primero se fija una semilla que permita la reproductibilidad de los datos obtenidos y luego con ayuda de la función rnorm() y runif() se simula una base de datos con 50 variables, 60 observaciones y 3 clases de población:

```
set.seed(123)
df<- data.frame(matrix(nrow=60,ncol = 50))
for(i in 1:50){
  a=rnorm(n = 20,52,3)
  b=rnorm(n = 20,72,5)
  c=runif(n = 20,min = 30,max = 55)
  df[,i] <- c(a,b,c)
}
df$clase <- c(rep("1",20),rep("2",20),rep("3",20))
```

A continuación, se puede apreciar la estuctura general, con las primeras y últimas filas y también columnas:

	X1	X2	X-	X50	clase
1	50.3185730603434	52.7599555419843	.....	53.2190994144543	1
2	51.3094675315502	51.9143597339539	.....	57.1425955773143	1
3	56.6761249424474	51.871388628126	.....	51.8188403376371	1
4	.....	.....	.....	.....	.....
58	32.339874667814	46.4807581051718	.....	35.7901265332475	3
59	41.669476039242	33.8086654210929	.....	48.4854441497009	3
60	42.7876364975236	44.3216764554381	.....	30.2025894296821	3

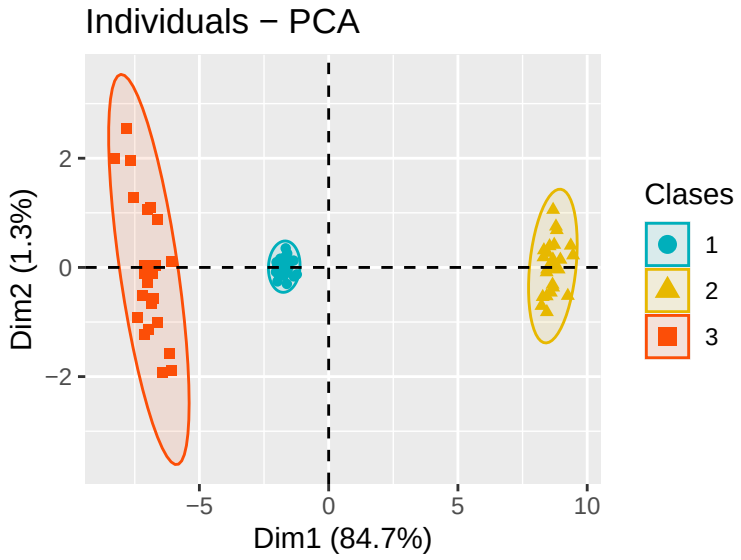
b)

Realice PCA en las 60 observaciones y grafique las observaciones en términos de las 2 primeras variables principales Z1 y Z2. Use un color diferente para indicar las observaciones en cada una de las tres clases. Si las tres clases aparecen separados en esta gráfica, solo entonces continúe con la parte (c). Si no, vuelva al inciso a) y modifique la simulación para que haya una mayor separación entre las tres clases. No continúe con la parte (c) hasta que las tres clases muestren al menos algún grado de separación en los dos primeros vectores de scores de componentes principales.

Solución

A continuación se realiza el ajuste de componentes principales (PCA), el cual intenta reducir la dimensionalidad de la base de datos, y luego se ilustran las observaciones en las primeras 2 componentes principales, la cual logra segmentar correctamente las clases. La primera componente retiene el 84.7% de la varianza total de los datos originales, mientras la segunda el 1.3%

```
#Ajuste
res.pca <- PCA(df[,-51],graph =F)
#Gráfico
fviz_pca_ind(res.pca,
             geom.ind = "point", #mostrar puntos
             col.ind = df$clase, #clases
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, #elipses
             legend.title = "Clases"
             )+ theme_gray()
```



c)

Desarrolle agrupación de K-medias de las observaciones con K = 3. ¿Qué tan bien funcionan los clústeres que obtuvo con el algoritmo de K-medias comparado con las verdaderas etiquetas de clase?

Sugerencia: puede usar la función table() en R para comparar las verdaderas etiquetas de clase con las etiquetas de clase obtenidas por agrupamiento. Tener cuidado cómo se interpretan los resultados: el agru-

pamiento de K-medias numera los grupos arbitrariamente, por lo que no puede simplemente comprobar si las verdaderas etiquetas de clase y las etiquetas de agrupación son las mismas.

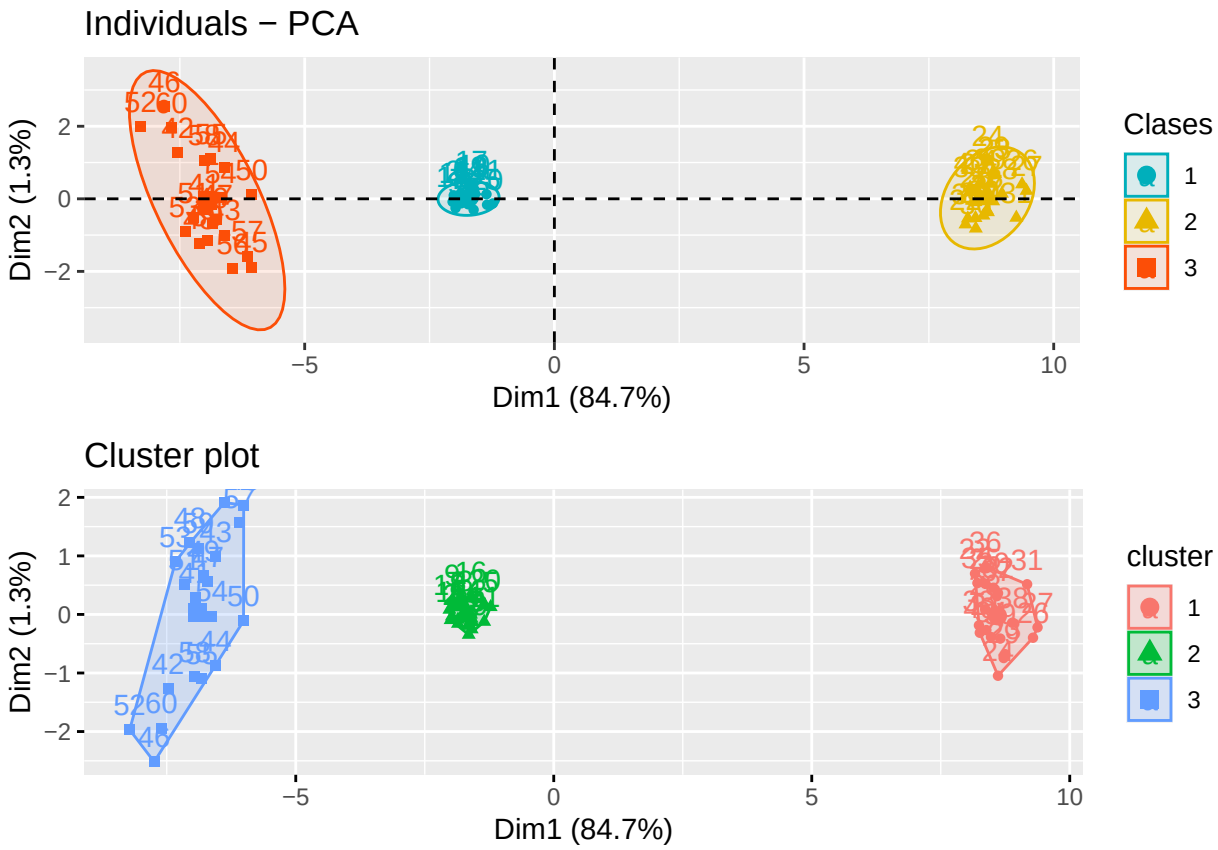
Solución

Se realiza a continuación un análisis visual para saber como se están comportando las etiquetas que resultan de la función kmeans():

```
set.seed(123)
km.out =kmeans (df[,-51],3, nstart =20)
```

```
grid.arrange(
            fviz_pca_ind(res.pca,
             geom.ind = c("point","text"), #mostrar puntos
             col.ind = df$clase, #clases
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, #elipses
             legend.title = "Clases")+ theme_gray(),

            fviz_cluster(km.out, data = df[,-51]))
```



Así entonces, como el agrupamiento de K-medias numera las clases arbitrariamente, se verifican las clases a las que pertenecen los individuos y se puede notar que las etiquetas 1 y 2 están intercaladas. Luego, se realiza el siguiente procedimiento para intercarbiar dichos valores:

```
km2 <- km.out$cluster
for(i in 1:length(km2)){
  if(km2[i]==1){
    km2[i]<-2
  }else if(km2[i]==2){
    km2[i]<- 1
  }else{
    next
  }}; km2
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [39] 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

De esta manera, logramos obtener las siguiente matriz de confusión, donde todas las observaciones han sido clasificadas correctamente. Es decir, que la tasa de error fue cero, por lo cual se conlcuye que logoritmo de K-medias comparado con las verdaderas etiquetas funciona correctamente.

```
table(km2,df$clase)
```

```
##
## km2  1  2  3
##   1 20  0  0
##   2  0 20  0
##   3  0  0 20
```

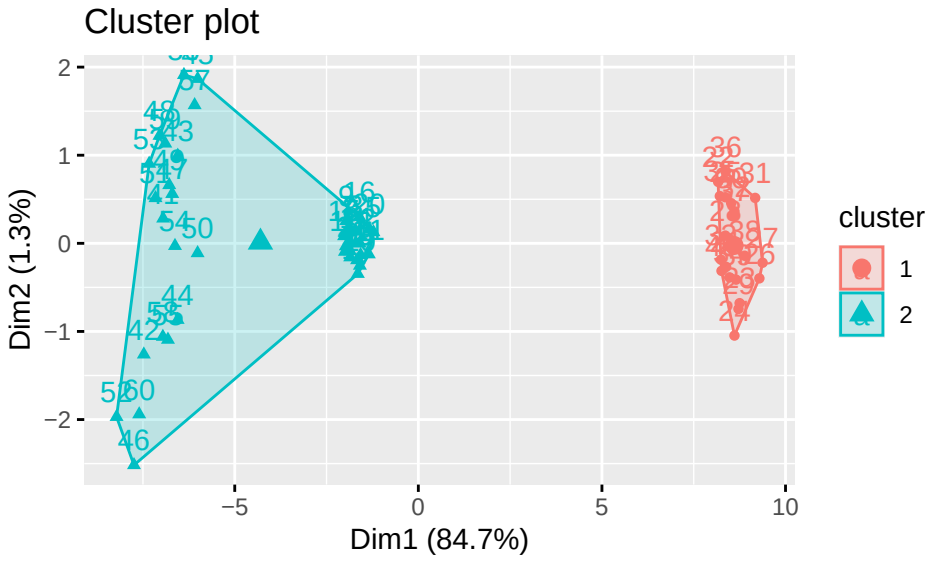
d)

Realice agrupamiento de K-medias con K = 2. Describa sus resultados.

Solución

En el siguiente gráfico podemos observar que realizando el agrupamiento de K-medias con K = 2, el algoritmo clasifica los datos en dos clases, para ello, en este caso unieron las 2 poblaciones más cercanas en una sola (cluster 2, azul), las cuales corresponden a la población original 1 con distribución uniforme y a la población original 3 con distribución normal.

```
set.seed(123) #semilla
km.out =kmeans (df[,-51],2, nstart =20)
fviz_cluster(km.out, data = df[,-51])
```



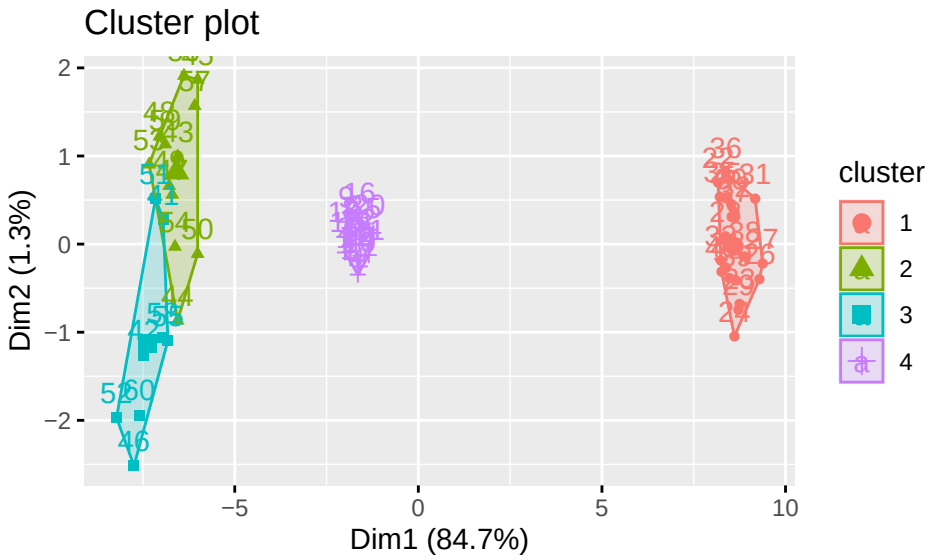
e)

Ahora realice agrupamiento de K-medias con K = 4 y describa sus resultados.

Solución

En el siguiente gráfico podemos observar que realizando el agrupamiento de K-medias con K = 4, el algoritmo clasifica los datos en cuatro clases, para ello, en este caso separó en dos la población con mayor dispersión (cluster 2 y 3) que corresponden el las verdaderas etiquetas a las poblaciones 1 y 3 respectivamente.

```
set.seed(123) #semilla
km.out =kmeans(df[,-51],4, nstart =20)
fviz_cluster(km.out, data = df[,-51])
```



f)

Ahora realice agrupamiento de K-medias con K = 3 en los dos primeros vectores de scores de componentes principales, en lugar de los datos en las variables originales. Es decir, realice la agrupación de K-medias en la matriz de 60 O 2, cuya primera columna es la coordenada zi1 en la primera componente principal Z1 y la segunda columna es la coordenada zi2 en la segunda componente principal Z2. Comente los resultados.

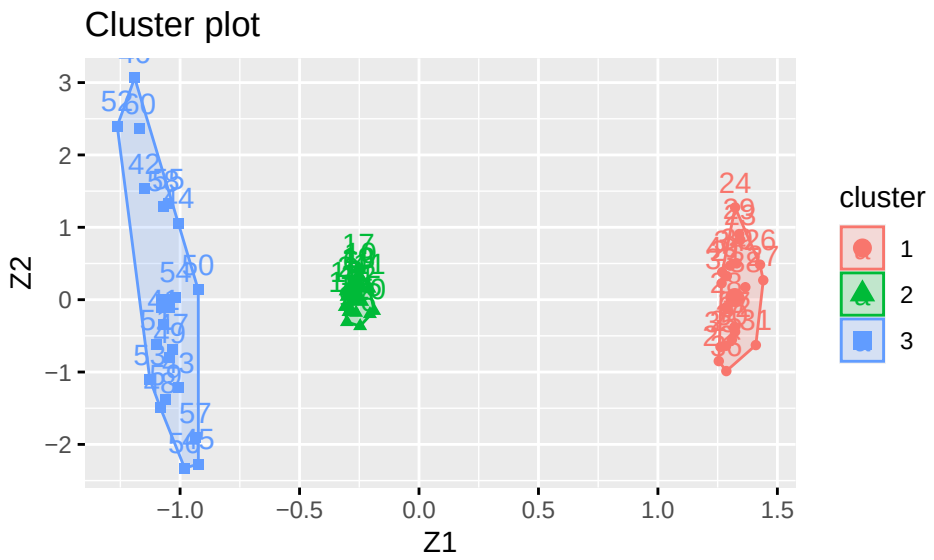
Solución

Primero se obtienen las coordenadas de las dos primeras componentes principales, de la siguiente manera:

```
Z1 <- res.pca$ind$coord[,1]
Z2 <- res.pca$ind$coord[,2]
Z<- data.frame(Z1,Z2)
```

Luego, se realiza la agrupación de K-medias con K = 3 con los datos de Z

```
set.seed(123)
km.out =kmeans(Z,3, nstart =20)
fviz_cluster(km.out, data = Z)
```



Como se puede observar, realizar agrupación de K medias con las coordenadas obtenidas en las dos primeras componenetes principales se logra también una correcta clasificación de los datos originales. Cabe mencionar que tener conocimiento acerca del número de clases, lo cual favorece los resultados del análisis y que se manifestó un cambio en escala con respercto a la original.

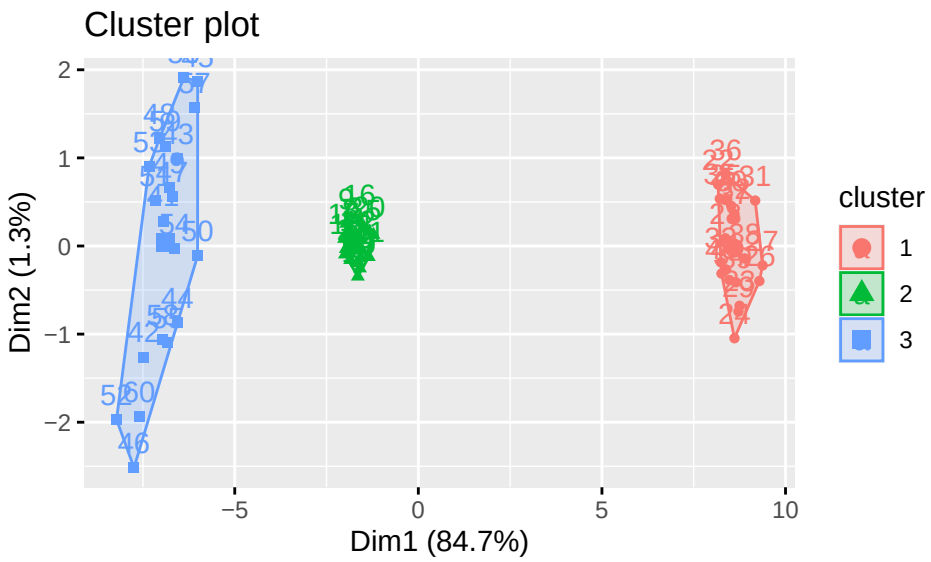
g)

Con la función scale(), realice agrupamiento de K-medias con K = 3 en los datos después de escalar cada variable para tener una desviación estándar de uno. ¿Cómo se comparan estos resultados con los obtenidos? en (b)? Explique.

Solución

```
set.seed(123)
sd.data=scale(df[,-51])
km.out =kmeans (sd.data,3, nstart =20)
```

```
fviz_cluster(km.out, data = sd.data)
```



En este caso, el agrupamiento de K-medias de los datos escalados, es decir con desviación estándar de uno, también logra clasificar correctamente las tres diferentes poblaciones. Es decir, con respecto a (b) donde se hizo uso de análisis de componentes principales, ambos logran clasificar correctamente los individuos de las tres diferentes distribuciones.