PARTITION AROUND MEDOIDS (PAM)

Arleth Michell Morales García

2022-05-27

```
library(cluster)
```

Cargar la matriz de datos

```
X<-as.data.frame(state.x77)
```

```
colnames(X)
```

```
## [1] "Population" "Income"     "Illiteracy" "Life Exp"   "Murder"
## [6] "HS Grad"    "Frost"      "Area"
```

Transformación de datos

1. Transformacion de las variables x1, x3 y x8

con la funcion de logaritmo.

```
X[,1]<-log(X[,1])
colnames(X)[1]<-"Log-Population"
```

```
X[,3]<-log(X[,3])
colnames(X)[3]<-"Log-Illiteracy"
```

```
X[,8]<-log(X[,8])
colnames(X)[8]<-"Log-Area"
```

MÉTODO PAM

1. Separacion de filas y columnas.

```
dim(X)
```

```
## [1] 50  8
```

```
n<-dim(X)[1]
p<-dim(X)[2]
```

2. Estandarización univariante

```
X.s<-scale(X)
```

3. Aplicación del algoritmo

```
pam.3<-pam(X.s,3)
```

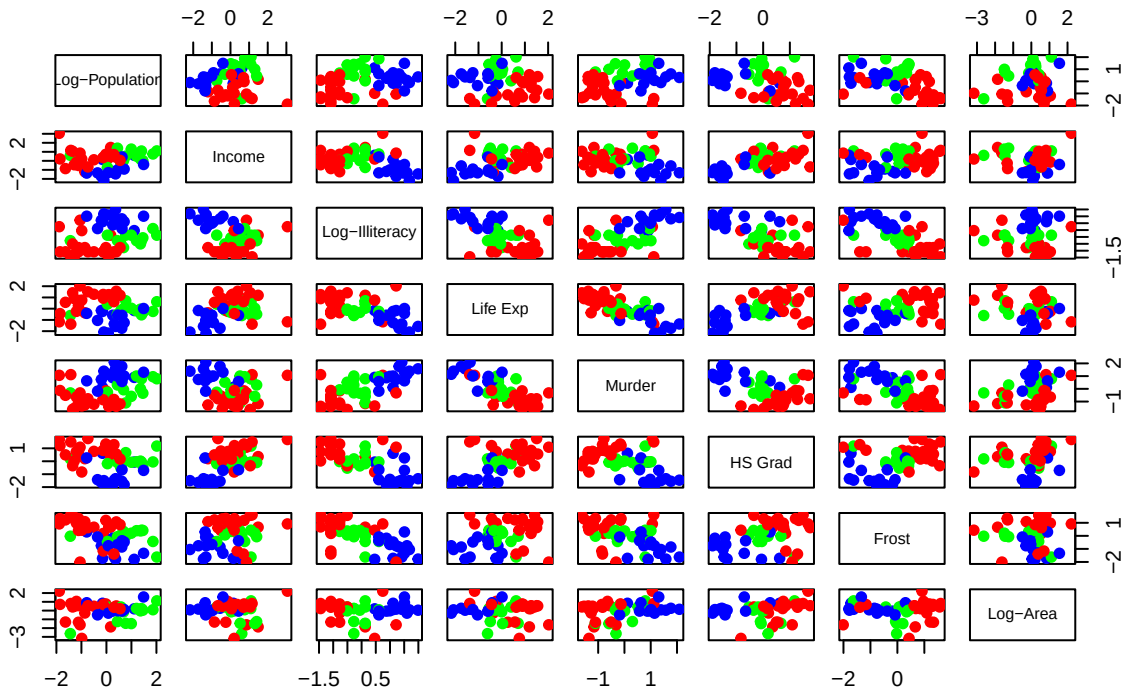
4. Clusters

```
cl.pam<-pam.3$clustering
cl.pam
```

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

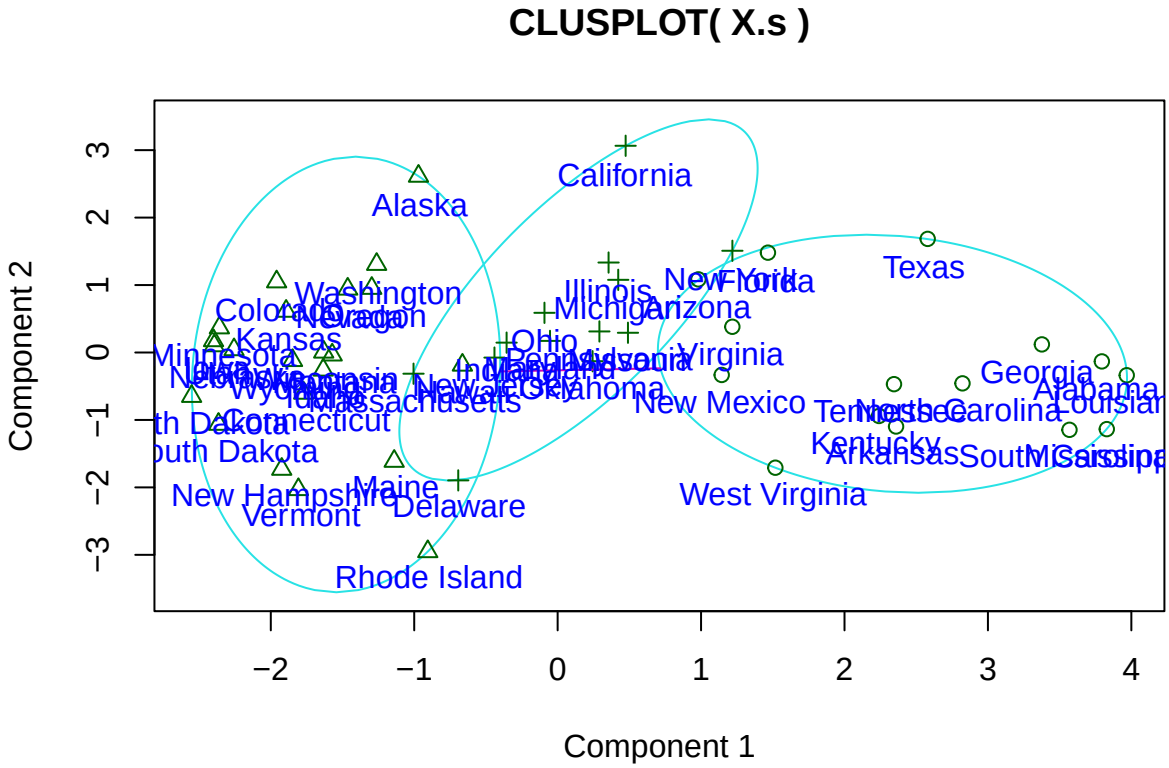
5. Scatter plot de la matriz con los grupos

```
col.cluster<-c("blue","red","green")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=19)
```



Visualizacion con Componentes Principales

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="blue")
```



Se observa buena variabilidad en los datos.

Silhouette

Representacion grafica de la eficacia de clasificacion

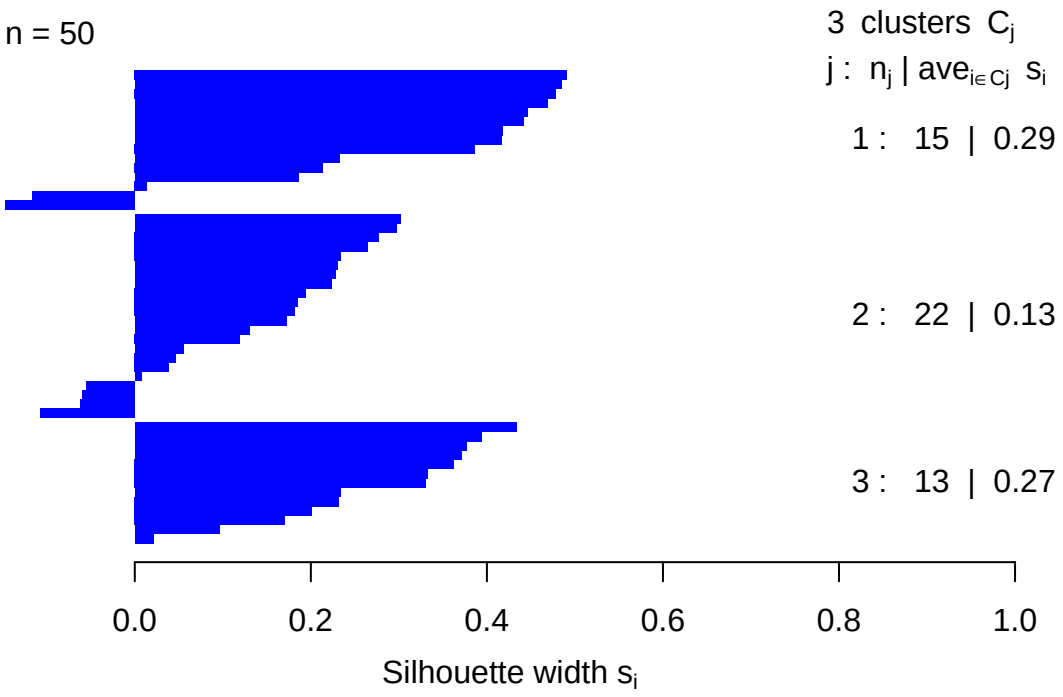
de una observacion dentro de un grupo.

1. Generación de los cálculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

2.Generación del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="blue")
```



Average silhouette width : 0.22

Su clasificación no es buena, tiene promedio de clasificación más baja que en K-MEDIAS, además tiene observaciones negativas

SEGUNDO EJERCICIO

```
library(cluster)
```

Cargar la matriz de datos

```
X<-as.data.frame(state.x77)
```

```
colnames(X)
```

```
## [1] "Population" "Income"     "Illiteracy" "Life Exp"   "Murder"
## [6] "HS Grad"    "Frost"      "Area"
```

Transformación de datos

1. Transformacion de las variables x1, x3 y x8 con la funcion de logaritmo.

```
X[,1]<-log(X[,1])
colnames(X)[1]<-"Log-Population"
```

```
X[,3]<-log(X[,3])
colnames(X)[3]<-"Log-Illiteracy"
```

```
X[,8]<-log(X[,8])
colnames(X)[8]<-"Log-Area"
```

MÉTODO PAM

1. Separacion de filas y columnas.

```
dim(X)
```

```
## [1] 50  8
```

```
n<-dim(X)[1]
p<-dim(X)[2]
```

2. Estandarización univariante

```
X.s<-scale(X)
```

3. Aplicación del algoritmo

```
pam.2<-pam(X.s,2)
```

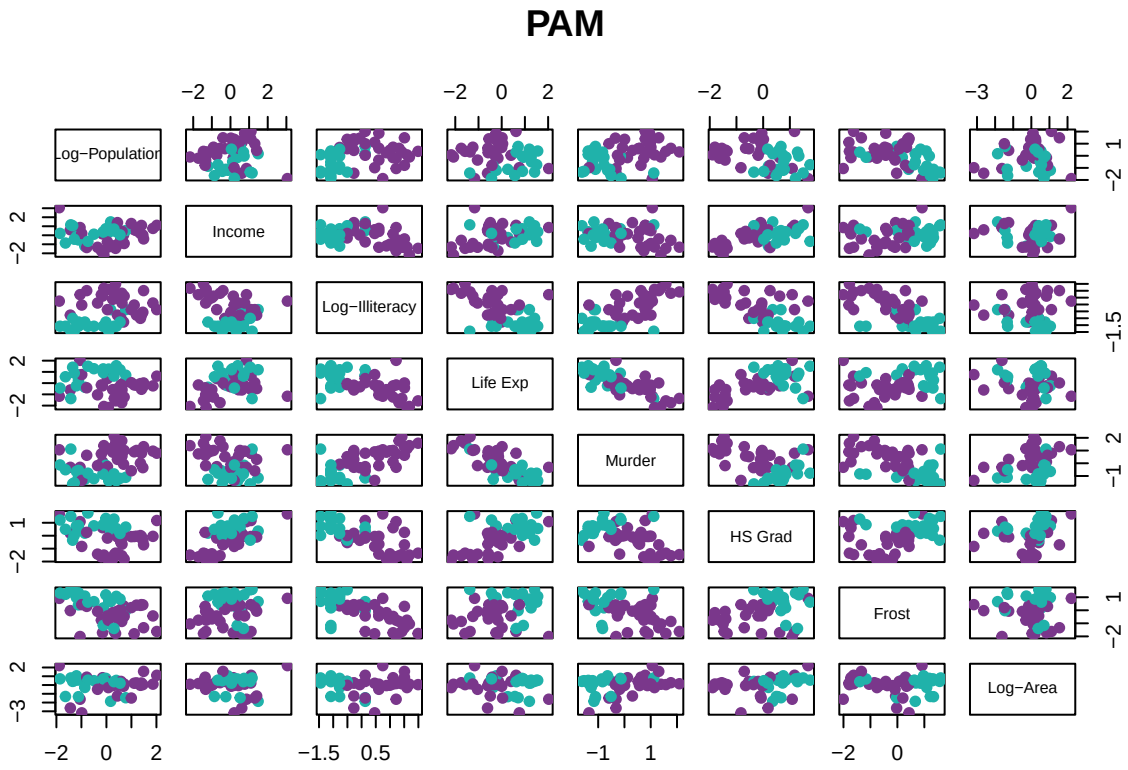
4. Clusters

```
cl.pam<-pam.2$clustering
cl.pam
```

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

5. Scatter plot de la matriz con los grupos

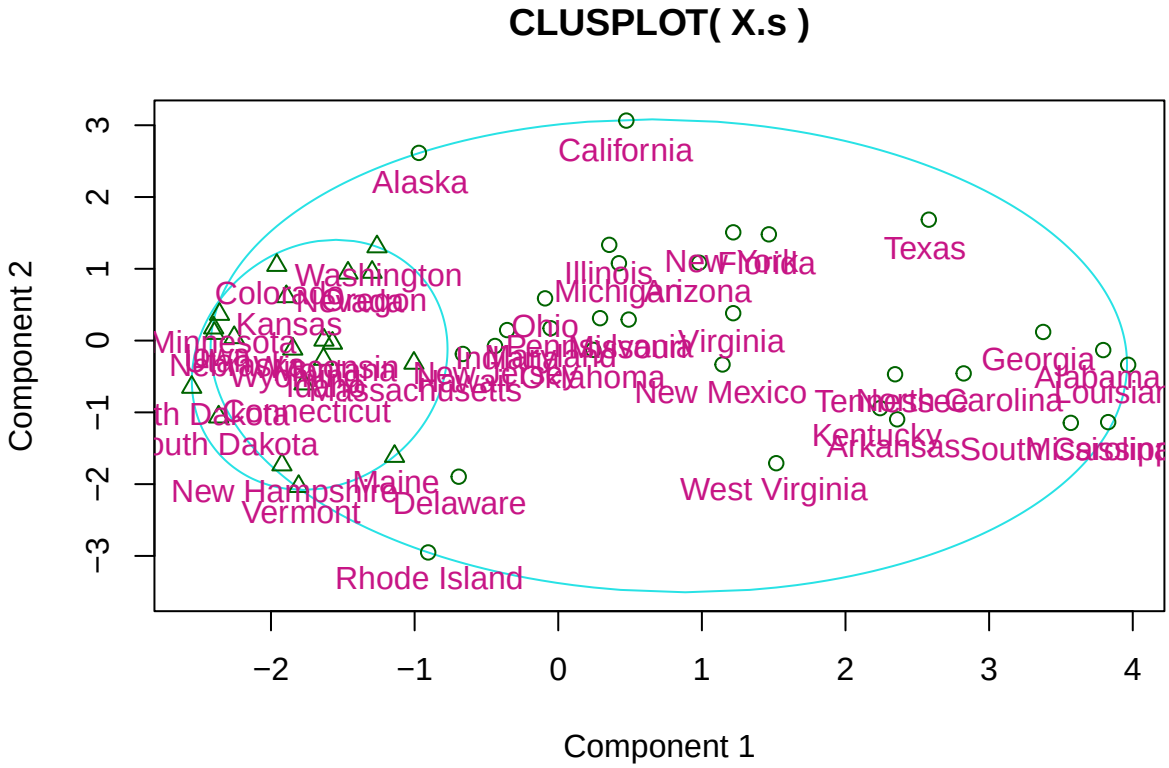
```
col.cluster<-c("mediumorchid4","lightseagreen")[cl.pam]
pairs(X.s, col=col.cluster, main="PAM", pch=19)
```



Se observa buena variabilidad en los datos.

Visualizacion con Componentes Principales

```
clusplot(X.s,cl.pam)
text(princomp(X.s)$scores[,1:2],
     labels=rownames(X.s),pos=1, col="mediumvioletred")
```



These two components explain 62.5 % of the point variability.

El cluster está solapado, algunos estados del cluster 1 comparten similitudes con el cluster 2

Silhouette

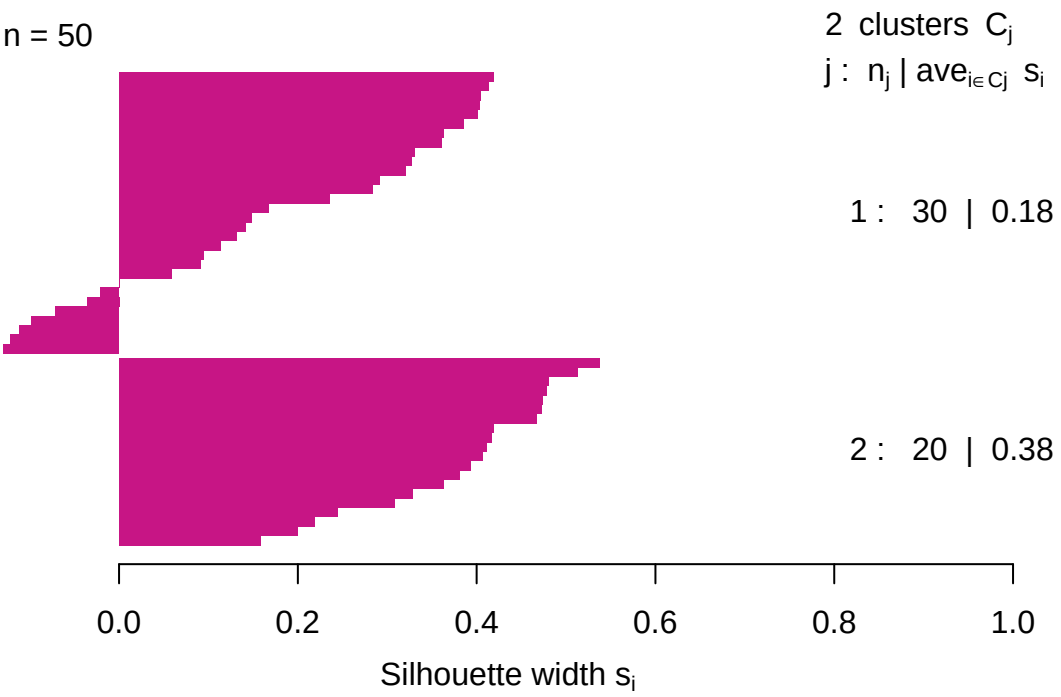
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1. Generación de los cálculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.pam<-silhouette(cl.pam, dist.Euc)
```

2.Generación del gráfico

```
plot(Sil.pam, main="Silhouette for PAM",
     col="mediumvioletred")
```



Average silhouette width : 0.26

Con 2 clusters se obtiene una mejor clasificación, ya que su clasificación promedio es de 0.26 en cambio con 3 clusters es de 0.22 En este caso, es mejor utilizar solo 2 clusters.