ANALISIS DE COMPONENTES PRINCIPALES ACP

ACP es una tecnica estadistica para reducir la dimensionalidad, transforma los datos con muchas dimensiones a espacios con menos, utilizando TRANSFORMACIONES LINEALES. el objetivo es remplazar un numero grandes de variables que estan correlacionadas entre si con un numero menor de variables que no se encuentren relacionados entre si

```
usarrests <- read.csv("data/tema3/USArrests.csv", stringsAsFactors = F) # para asegurarnos que el nombr
```

para hacer que la primera COLUMNA llamada "X" se convierta en el nombre de las filas utilizamos la FUNCION "rownames" y seleccionamos dicha columna

```
rownames(usarrests) <- usarrests$X
```

Ya que asignamos la columna X como el nombre de las filas entonces ELIMINAMOS LA COLUMNA utilizando la FUNCION "NULL"

```
usarrests$X <- NULL
head(usarrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Una vez que el data frame esta bien definido entonces: vamos a calcular para cada una de las columnas la VARIANZA que existe en dichas variables

```
apply(usarrests, 2, var) # el "2" indica que lo haga por columna, si lo queremos por fila se debe usar
```

```
##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

"prcomp" resta la media (center = TRUE) y divide por la desviacion tipica (escale = TRUE) cada una de las variables con la intencion de perder la variabilidad tan grande desde una de las variables a la otra.

```
acp <- prcomp(usarrests)
print(acp)
```

```
## Standard deviations (1, .., p=4):
## [1] 83.732400 14.212402  6.489426  2.482790
##
```

```
## Rotation (n x k) = (4 x 4):
##                 PC1         PC2         PC3         PC4
## Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
## Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
## UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
## Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

compara las desviaciones que se optienen centrando y escalando la informacion

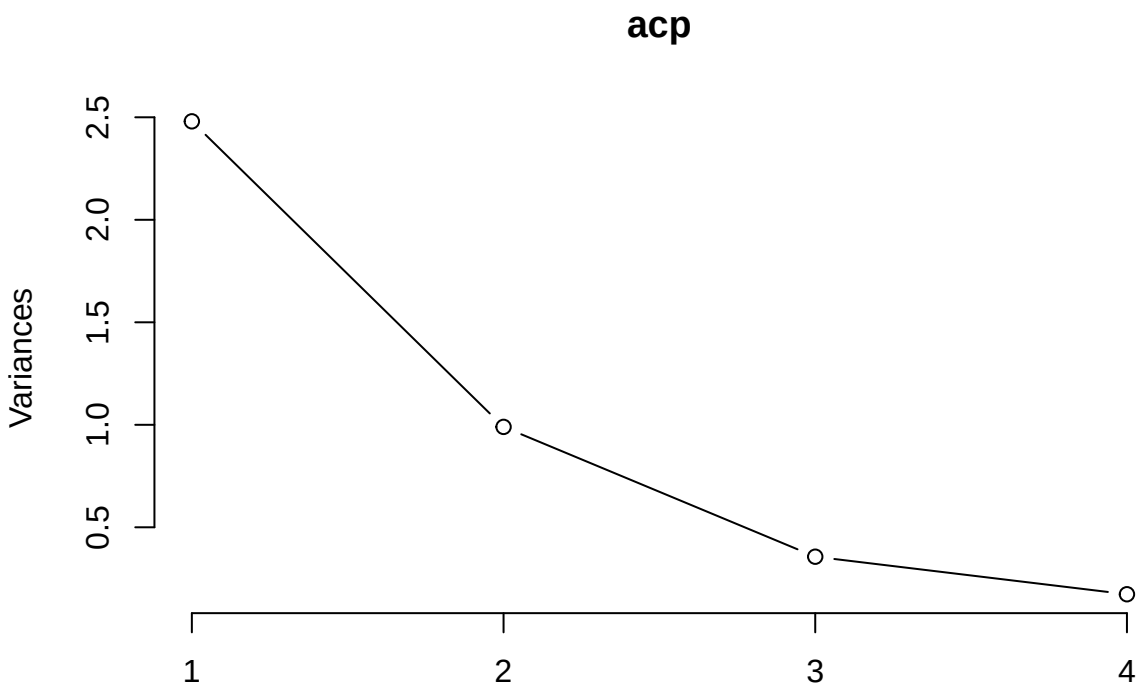
date cuenta que al usar la funcion “prcomp” te arroja como resultado una LISTA no un data frame

```
acp <- prcomp(usarrests,
              center = TRUE, scale = TRUE)
print(acp)
```

```
## Standard deviations (1, .., p=4):
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
##
## Rotation (n x k) = (4 x 4):
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

esto nos da los coeficientes de las combinaciones lineales de las variables continuas, o rotaciones. es decir “LA MATRIZ DE ROTACION” si observamos PC1 podemos observar que “murder” Assault" y ’Rape" tienen numeros parecidos lo que quiere decir que estan muy correlacionadas entre estas 3, sin embargo la segunda componente principal toma valores menores para esos 3 valores que ya han sido explicados en la primera CP y encambio le da un peso mucho valor a la poblacion urbana aun cuando las otras 3 se parecen entre si (por eso si se toman los 2 al mismo tiempo se pueden explicar mejor los datos).

```
plot(acp, type = "l")
```



esta grafica nos es una representacion de los 4 componentes principales y sus varianzas, normalmente los estadistas utilizan la regla del "CODO", por lo que un profesional se quedaria con la primera y la segunda componente principal (punto en la grafica) y rechazaria la tercera y la cuarta.

si hacemos un resumen de los componentes con "summary" nos da las desviaciones estandar de cada uno de los componentes, se utilizan componentes principales hasta poder explicar cierto porcentaje de los casos, por ejemplo si quicieramos explicar el ~80% de casos podriamos utilizar los primeros dos componentes principales como nos lo dise la proporcion acumulativa, o los primeros 3 componentes si queremos explicar el 95% de los casos. Sin embargo, si solo nos quedamos con una solo podremos explicar el 62% de los casos.

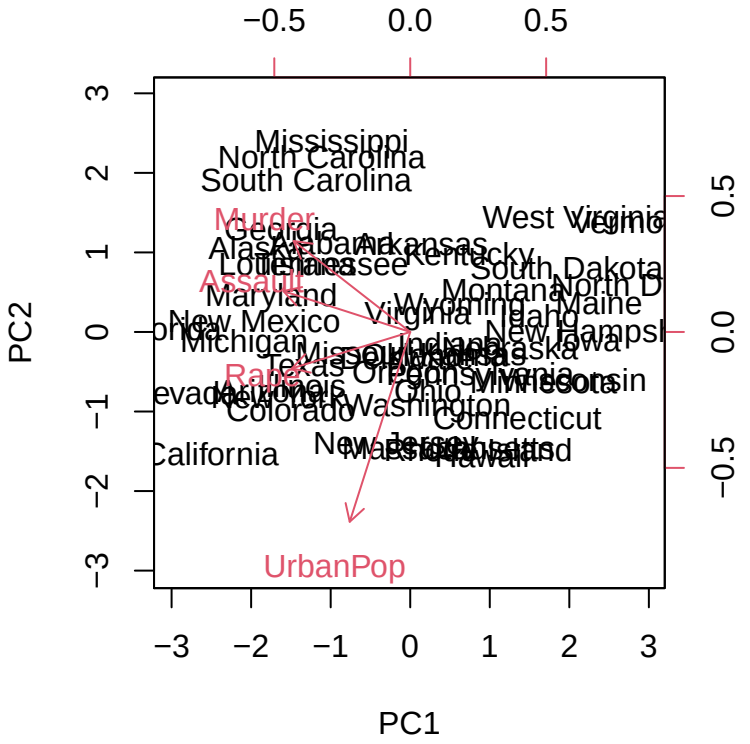
```
summary(acp)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

una representacion grafica de lo que nos quiere decir esta tecnica se logra graficando con la FUNCION "biplot"

Al ver las flrchas se puede decir que las variables murder assault y rape tienden hacia PC2 (se mueven en el eje horizontal), mientras que la poblacion urbana tiende mas hacia PC1.

```
biplot(acp, scale = 0)
```



si quieres traducir del data frame original para tener explicitamente el primer y el segundo componente principal al data frame original puedo computarle la primera componente principal haciendo lo siguiente, “sumando el producto de la primera columnade la matriz de rotacion y multiplicarlo por el data set original por filas (”1") y lo mismo para la PC2, por eso nos quedamos con la segunda columna de la matriz de rotacion (acp$rotation[,2])

```
pc1 <- apply(acp$rotation[,1]*usarrests, 1, sum)
pc2 <- apply(acp$rotation[,2]*usarrests, 1, sum)
```

y ahora agrego esta informacion en el data frame original generando dos nuevas columnas llamadas PC1 y PC2

```
usarrests$pc1 <- pc1
usarrests$pc2 <- pc2
head(usarrests)
```

```
##            Murder Assault UrbanPop Rape       pc1        pc2
## Alabama      13.2     236       58 21.2 -109.7067 -194.71128
## Alaska       10.0     263       48 44.5 -200.9300  -40.54732
## Arizona       8.1     294       80 31.0 -198.6759   59.01456
## Arkansas      8.8     190       50 19.5 -154.1308   29.54465
## California    9.0     276       91 40.6 -141.6652 -234.51235
## Colorado      7.9     204       78 38.7 -181.9864  -24.46027
```

y ya que se tienen los datos de los componentes principales 1 y 2 que explican el ‘~80% de los casos se podria finalmente eliminar toda la informacion de la tabla original y quedarte solo con esos datos de la siguiente manera

```r
usarrests[,1:4] <- NULL
head(usarrests)
```

```
##                  pc1        pc2
## Alabama    -109.7067 -194.71128
## Alaska     -200.9300  -40.54732
## Arizona    -198.6759   59.01456
## Arkansas   -154.1308   29.54465
## California -141.6652 -234.51235
## Colorado   -181.9864  -24.46027
```