PCA

Antonio Hernández

2022-03-27

Análisis de Componentes Principales

INTRODUCCION

El análisis de componentes principales (*ACP*) es un método de reducción de la dimensionalidad de las variables originales.

Matriz de trabajo

1.-Se trabajo con la matriz de datos *fiel* que esta en el paquete *datos*.

```
install.packages("datos")
```

```
library(datos)
```

2.- Se selecciona la matriz *fiel*.

```
x<-datos::fiel
```

Exploración de la matriz

1.- Dimensión de la matriz La matriz cuenta con 272 observaciones y 2 variables.

```
dim(x)
```

```
## [1] 272   2
```

2.- Tipo de variables.

```
str(x)
```

```
## 'data.frame':    272 obs. of  2 variables:
##  $ erupciones: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ espera    : num  79 54 74 62 85 55 88 85 51 85 ...
```

3.- Conocer el nombre de las variables.

```
colnames(x)
```

```
## [1] "erupciones" "espera"
```

4.-Buscar los posibles datos perdidos.

```
anyNA(x)
```

```
## [1] FALSE
```

Tratamiento de la matriz

Se genera una nueva matriz filtrada.

```
x1<- x[,]
```

ACP paso a paso

1.- Transformar la matriz en un data frame.

```
x1<- data.frame(x1)
```

2.- Definir n (individuos) y p (variables).

```
n<-dim(x)[1]
n
```

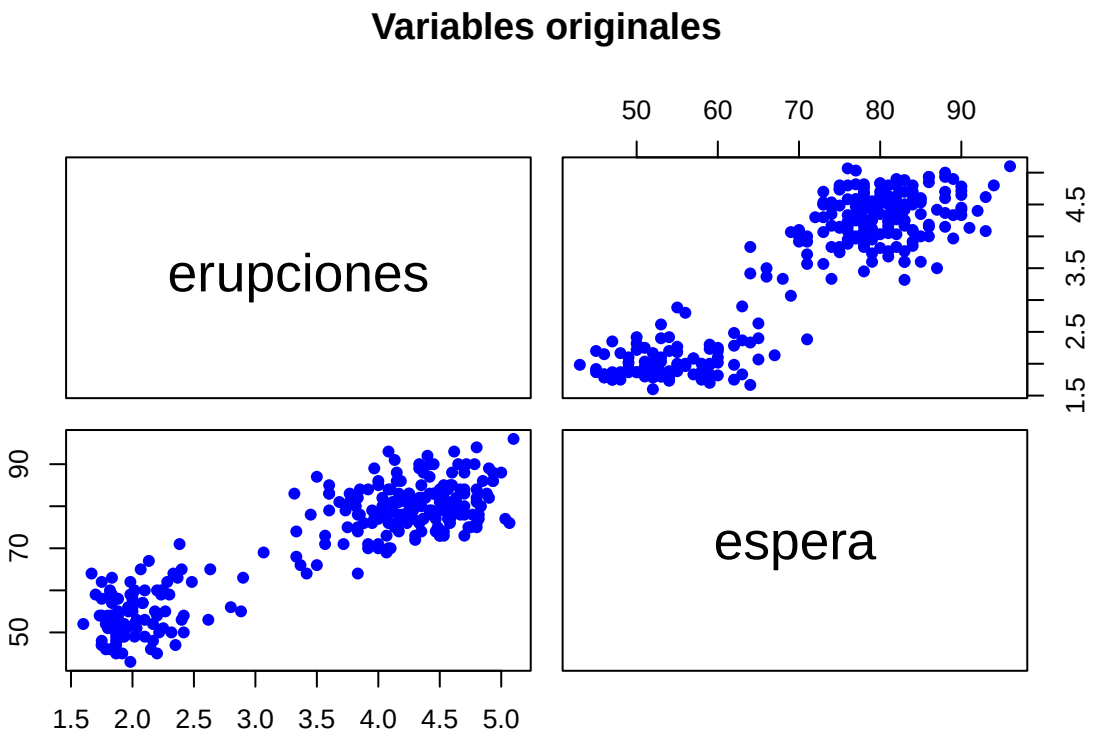
```
## [1] 272
```

```
p<-dim(x)[2]
p
```

```
## [1] 2
```

3.- Generación del gráfico **scaterplot**

```
pairs(x1,col="blue", pch=16,
    main="Variables originales")
```



4.-Obtención de la media por columna y la **la matriz de covarianza muestral**.

```
mu<-colMeans(x1)
mu
```

```
## erupciones     espera
##   3.487783  70.897059
```

```r
s<-cov(x)
s
```

```
##            erupciones    espera
## erupciones   1.302728  13.97781
## espera      13.977808 184.82331
```

5.- Obtención de los **valores** y **vectores** propios desde la matriz de covarianza muestral.

```r
es<-eigen(s)
es
```

```
## eigen() decomposition
## $values
## [1] 185.8818239   0.2442167
##
## $vectors
##           [,1]       [,2]
## [1,] 0.0755118 -0.9971449
## [2,] 0.9971449  0.0755118
```

5.1 .- Obtención de la matriz de autovalores propios.

```r
eigen.val<-es$values
eigen.val
```

```
## [1] 185.8818239   0.2442167
```

5.2 .- Obtención de la matriz de vectores propios.

```r
eigen.vec<-es$vectors
eigen.vec
```

```
##           [,1]       [,2]
## [1,] 0.0755118 -0.9971449
## [2,] 0.9971449  0.0755118
```

6.- Calcular la proporción de la variabilidad.

6.1.- Para la matriz de valores propios.

```r
pro.var<-eigen.val/sum(eigen.val)
pro.var
```

```
## [1] 0.998687896 0.001312104
```

6.2.- Acumulada.

```r
pro.var.acum<-cumsum(eigen.val)/sum(eigen.val)
pro.var.acum
```

```
## [1] 0.9986879 1.0000000
```

7.- Obtener la matriz de correlacion

```r
R<-cor(x)
R
```

```
##            erupciones    espera
## erupciones  1.0000000 0.9008112
## espera      0.9008112 1.0000000
```

8.- Obtener los valores y vectores propios a partir de la matriz de correlaciones.

```
eR<-eigen(R)
eR
```

```
## eigen() decomposition
## $values
## [1] 1.90081117 0.09918883
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

9.- Obtención de la matriz de autovalores propios.

9.1- Separación de matriz de valores propios.

```
eigen.val.R<-eR$values
eigen.val.R
```

```
## [1] 1.90081117 0.09918883
```

9.2 .- Obtención de la matriz de vectores propios.

```
eigen.vec.R<-eR$vectors
eigen.vec.R
```

```
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

10.-Cálculo de la proporción de la variabilidad. 10.1.- Para la matriz de valores propios.

```
pro.var.R<-eigen.val.R/sum(eigen.val.R)
pro.var.R
```

```
## [1] 0.95040558 0.04959442
```

10.2.- Acumulada. En este punto se selecciónа en número de componentes, siguiendo el criterio del 95% de varianza explicada.

Para este ejemplo se va a seleccionar un factor (0.95% de varianza explicada)

```
pro.var.acum.R<-cumsum(eigen.val)/sum(eigen.val)
pro.var.acum.R
```

```
## [1] 0.9986879 1.0000000
```

11.- Obtener la media de los autovalores.

```
mean(eigen.val)
```

```
## [1] 93.06302
```

##Obtención de coeficientes.

12.-Centrar los datos con respecto a la media.

12.1.-Construcción de la matriz de unos.

```
ones<-matrix(rep(1,n),nrow=n, ncol=1)
```

12.1.- Construccion de la matriz centrada.

```
X.cen<-as.matrix(x)-ones%*%mu
```

13.- Construccion de la matriz diagonal de las varianzas.

```
Dx<-diag(diag(s))
Dx
```

```
##          [,1]     [,2]
## [1,] 1.302728   0.0000
## [2,] 0.000000 184.8233
```

14.-Construccion de la matriz centrada multiplicada. # por Dxˆ1/2

```
Y<-X.cen%*%solve(Dx)^(1/2)
```

15.- Construccion de los coeficientes o scores eigen.vec.R matriz de autovectores.

```
scores<-Y%*%eigen.vec
scores[1:10,]
```

```
##          [,1]        [,2]
## 1   0.6017472 -0.05303002
## 2  -1.3510033  1.38065799
## 3   0.2173499  0.15245928
## 4  -0.7322759  1.00312613
## 5   1.1035529 -0.83480761
## 6  -1.2060067  0.44006280
## 7   1.3346412 -0.96404192
## 8   1.0418267 -0.01970367
## 9  -1.5611193  1.23294917
## 10  1.0914459 -0.67493191
```

16.- Nombramos las columnas PC1...PC8.

```
colnames(scores)<-c("PC1","PC2")
```

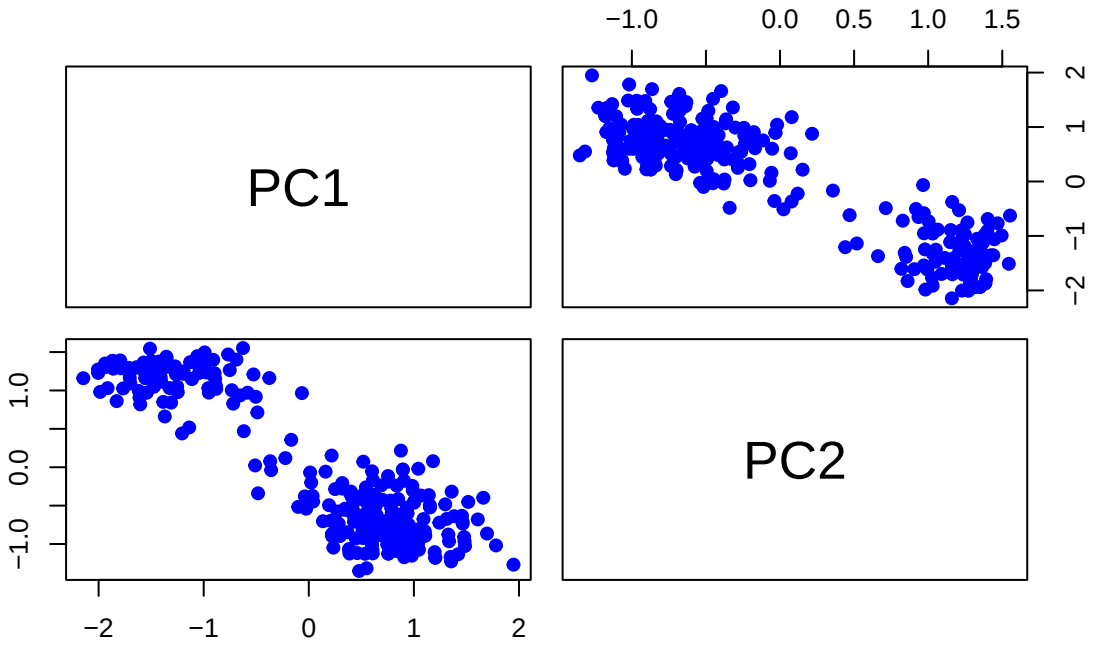
17.- Visualización de los scores.

```
scores[1:10,]
```

```
##           PC1         PC2
## 1   0.6017472 -0.05303002
## 2  -1.3510033  1.38065799
## 3   0.2173499  0.15245928
## 4  -0.7322759  1.00312613
## 5   1.1035529 -0.83480761
## 6  -1.2060067  0.44006280
## 7   1.3346412 -0.96404192
## 8   1.0418267 -0.01970367
## 9  -1.5611193  1.23294917
## 10  1.0914459 -0.67493191
```

18.- Generacion del grafico de los scores

```
pairs(scores, main="scores", col="blue", pch=19)
```



##ACP SINTETIZADO

1.-# Aplicar el cálculo de la varianza a las columnas 1=filas, 2=columnas.

```
apply(x, 2, var)
```

```
## erupciones     espera
##   1.302728 184.823312
```

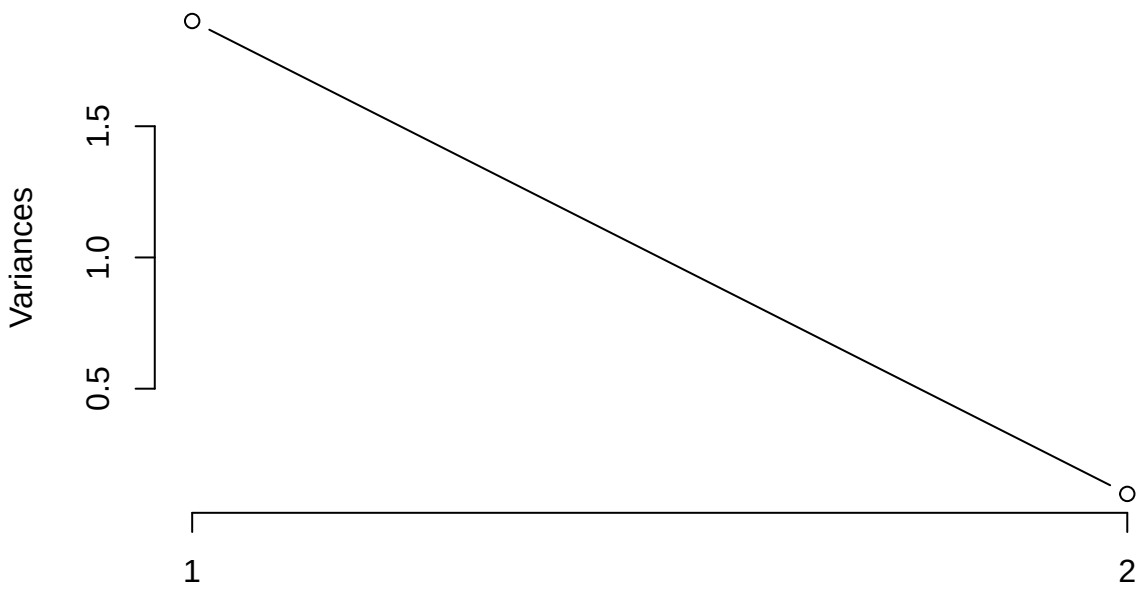
2.- Aplicar la función **acp** para reducir la dimensionalidad y centrado por la media y escalada por la desviacion standar (dividir entre sd).

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

```
## Standard deviations (1, .., p=2):
## [1] 1.3786991 0.3149426
##
## Rotation (n x k) = (2 x 2):
##                   PC1        PC2
## erupciones -0.7071068  0.7071068
## espera     -0.7071068 -0.7071068
```

3.-Generación del gráfico screeplot.

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



4.- Resumen.

```
summary(acp)
```

```
## Importance of components:
##                           PC1     PC2
## Standard deviation     1.3787 0.31494
## Proportion of Variance 0.9504 0.04959
## Cumulative Proportion  0.9504 1.00000
```

CONSTRUCCIÓN DE LOS CP CON VARIABLES PRINCIPALES

Convinacion lineal de las variables originales.

z= -0.707(var1)- (-707(var2))

Tiempo de espera entre erupciones.

z2= 0.707(var1)- 0.707(var2)