PCA

Melissa Ortega

2022-04-12

Análisis de Componentes Principales

Introducción

El Análisis de componentes principales es un método que sirve para la reducción de la dimensionalidad de las variable originales. Este método permite representar los datos originales (individuos y variables) en un espacio de dimensión inferior del espacio original, mientras limite al máximo la pérdida de información.

Matriz de trabajo

1. Se trabajo con la matriz fiel, extraida del paquete **datos** que se encuentra precargada en R

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

```
library(datos)
```

2. Se selecciona la matriz fiel

```
m<-datos::fiel
```

3. Exploracion de la matriz

4. Dimension de la matriz La matriz cuenta con 272 observaciones y 2 variables

```
dim(m)
```

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

2. Tipo de variables

```
str(m)
```

```
## '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. Nombre de las variables

```
colnames(m)
```

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

4. Busca de datos perdidos

```
anyNA(m)
```

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

Tratamiento de la matriz

Se genera una nueva matriz **X1** filtrada

```
x1<-m[1:2]
```

" ACP Paso a Paso

1. Transformar la matriz en un data frame

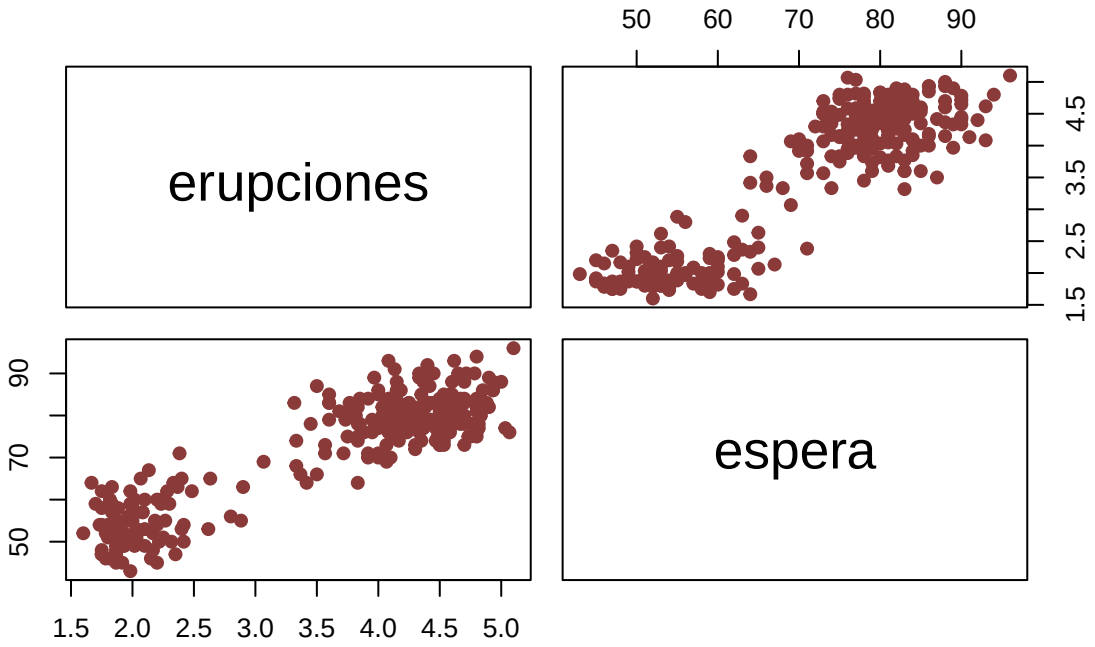
```
x1<-as.data.frame(x1)
```

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

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

3.Generación del Gráfico **scaterplot**

```
pairs(m,col="indianred4", pch=19,
      main="Variables originales")
```



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

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

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

```
s<-cov(x1)
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 la covarianza muestral.

```
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. Separación de la matriz de valores propios

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

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

5.2. Separacion de la matriz de vectores propios

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

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

6. Calcular la proporción de variabilidad

6.1 Para la matriz de valores propios.

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

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

6.2 Acumulada

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

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

7. Obtención de la matriz de correlaciones

```
R<-cor(x1)
R
```

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

8. Obtención de 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. Separación de la matriz de valores propios

9.1 Separación de la matriz de valores propios

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

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

9.2 Separación de la matriz de los vectores propios

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

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

10. Calculo de la proporcion de 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

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

```
## [1] 0.9504056 1.0000000
```

Obtención de coeficientes

12. Centrar los datos respecto a la media 12.1 Construccion de la matriz 1

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

12.2 Construcción de la matriz centrada

```
X.cen<-as.matrix(x1-ones%*%mu)
```

13. Construcción 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. Construcción de la matriz centrada multiplicada por Dxˆ1/2

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

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

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

```
##           [,1]        [,2]
## 1   0.49097422  0.35193210
## 2  -1.92447802  0.16676594
## 3   0.06549949  0.25728314
## 4  -1.20914903  0.28363483
## 5   1.38106424  0.08599050
## 6  -1.20152126 -0.45216609
## 7   1.64056183  0.13856711
## 8   0.80304843  0.66400631
## 9  -1.98758654 -0.08219975
## 10  1.26769136  0.19936337
```

16. Nombramos las columnas PC1...PC2

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

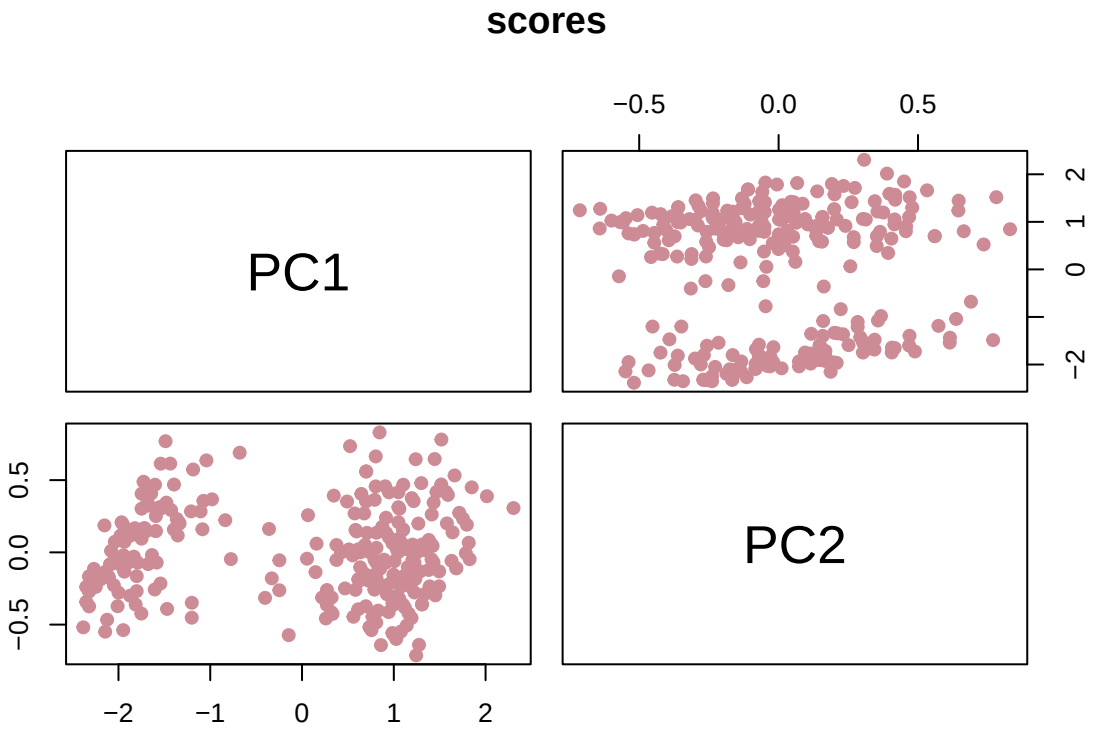
17.Visualizamos

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

```
##            PC1         PC2
## 1   0.49097422  0.35193210
## 2  -1.92447802  0.16676594
## 3   0.06549949  0.25728314
## 4  -1.20914903  0.28363483
## 5   1.38106424  0.08599050
## 6  -1.20152126 -0.45216609
## 7   1.64056183  0.13856711
## 8   0.80304843  0.66400631
## 9  -1.98758654 -0.08219975
## 10  1.26769136  0.19936337
```

18. Generación del Gráfico de los scores

```
pairs(scores, main="scores", col="lightpink3", pch=19)
```



ACP VIA SINTETIZADA

1. Aplicar el calculo de la varianza de las columnas (1=filas, 2=columnas()

```
apply(x1,2,var)
```

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

2. Aplicar la Función **prcomp** para reducir la dimensionalidad y centrado por la media y escalado por la Desviación estandar

```
acp<-prcomp(x1,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. Resumen de la matriz **acp**

```
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 las variables originales

Combinación lineal de las variables originales

z1=-0.707(var1)-0.707(var2)

El primer componente distingue las erupciones

z2=-0.7071(var1)-0.7071(var2)

El segundo componente distingue la espera de la Erupción