PCA

Saraí Campos Varela

2022-03-25

Análisis de Componentes Principales

Introducción

El análisis de componentes principales (**ACP**) es una herramienta utilizada para describir un conjunto de datos en términos de nuevas variables (componentes) no correlacionadas. Esta técnica es útil para reducir la dimensionalidad de un conjunto de datos.

El ACP busca la proyección en la que los datos queden mejor representados en términos de mínimos cuadrados. Convierte un conjunto de observaciones de variablesmcorrelacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales. El ACP se emplea generalmente en análisis exploratorio de datos y construcción de modelos predictivos.

Matriz de trabajo

1.- Se trabajo con la matriz Flores, extraída del paquete *datos* que se encuentra precargado en R.

```
library (datos)
```

2.- Se selecciona la matriz **Flores**.

```
x<- datos::flores
```

Exploración de la matriz.

1.- Dimensión de la matriz. La matriz cuenta con 150 observaciones y 5 variables.

```
dim (x)
```

```
## [1] 150   5
```

2.-Tipo de variables.

```
str (x)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Largo.Sepalo: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Ancho.Sepalo: num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Largo.Petalo: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Ancho.Petalo: num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Especie     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

3.- Nombre de las variables.

```
colnames(x)
```

```
## [1] "Largo.Sepalo" "Ancho.Sepalo" "Largo.Petalo" "Ancho.Petalo" "Especie"
```

4.-Buscamos datos perdidos en la matriz.

```
anyNA(x)
```

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

En este caso no se encuentran valores nulos.

Tratamiento de la matriz.

Generación de la nueva matriz

1.- Selección de variables cuantitativas unicamente de la especie versicolor.

```
X<-x[51:100,1:4]
```

PCA paso a paso

1.-Transformar la matriz en un data frame.

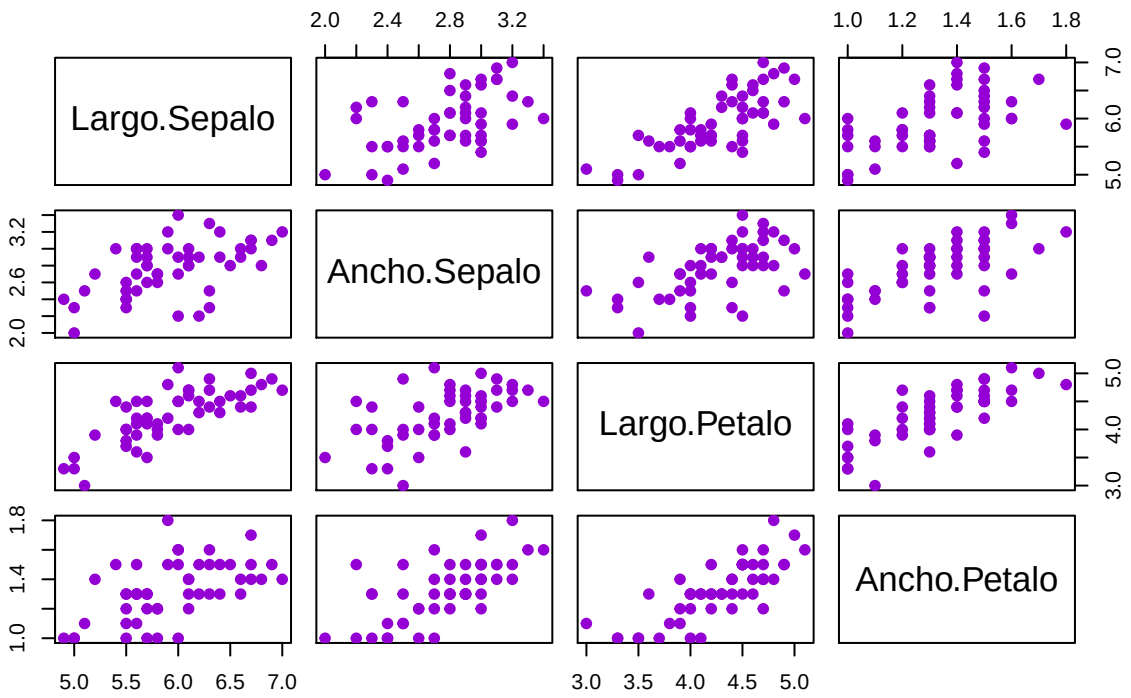
```
X<-as.data.frame(X)
```

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

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

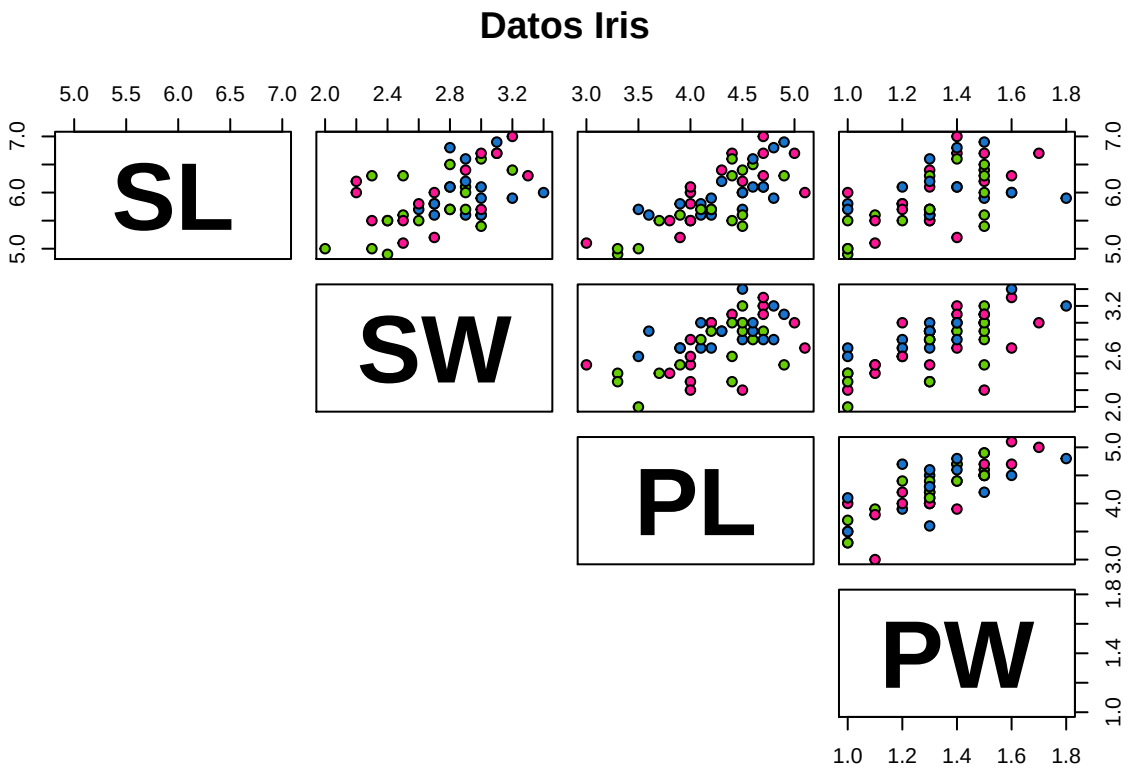
3.- Generación de un **scatterplot** de las variables originales.

```
pairs(X,col="darkviolet", pch=19,
      main="Variables Originales")
```



Gráfico

```
pairs(X, main = "Datos Iris", pch = 21, bg = c("deeppink1", "chartreuse3", "dodgerblue3"),
      lower.panel=NULL, labels=c("SL","SW","PL","PW"), font.labels=2, cex.labels=4.5)
```



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

```
mu<-colMeans(X)
mu
```

```
## Largo.Sepalo Ancho.Sepalo Largo.Petalo Ancho.Petalo
##        5.936        2.770        4.260        1.326
```

Covarianza.

```
s<-cov(X)
s
```

```
##              Largo.Sepalo Ancho.Sepalo Largo.Petalo Ancho.Petalo
## Largo.Sepalo   0.26643265   0.08518367   0.18289796   0.05577959
## Ancho.Sepalo   0.08518367   0.09846939   0.08265306   0.04120408
## Largo.Petalo   0.18289796   0.08265306   0.22081633   0.07310204
## Ancho.Petalo   0.05577959   0.04120408   0.07310204   0.03910612
```

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

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

```
## eigen() decomposition
## $values
## [1] 0.487873944 0.072384096 0.054776085 0.009790365
##
## $vectors
##            [,1]       [,2]        [,3]       [,4]
## [1,] -0.6867238  0.6690891 -0.26508336  0.1022796
## [2,] -0.3053470 -0.5674653 -0.72961786 -0.2289194
## [3,] -0.6236631 -0.3433270  0.62716496 -0.3159668
## [4,] -0.2149837 -0.3353051  0.06366081  0.9150409
```

5.1 Matriz de valores propios.

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

```
## [1] 0.487873944 0.072384096 0.054776085 0.009790365
```

5.2 Matriz de vectores propios.

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

```
##            [,1]       [,2]        [,3]       [,4]
## [1,] -0.6867238  0.6690891 -0.26508336  0.1022796
## [2,] -0.3053470 -0.5674653 -0.72961786 -0.2289194
## [3,] -0.6236631 -0.3433270  0.62716496 -0.3159668
## [4,] -0.2149837 -0.3353051  0.06366081  0.9150409
```

6.- Proporción de variabilidad para cada vector.

6.1.- Proporción de variabilidad de valores propios.

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

```
## [1] 0.78081758 0.11584709 0.08766635 0.01566898
```

6.2.- Proporción de variabilidad acumulada.

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

```
## [1] 0.7808176 0.8966647 0.9843310 1.0000000
```

7.- Obtención de la matriz de correlaciones.

```
R<-cor(X)
R
```

```
##              Largo.Sepalo Ancho.Sepalo Largo.Petalo Ancho.Petalo
## Largo.Sepalo    1.0000000    0.5259107    0.7540490    0.5464611
## Ancho.Sepalo    0.5259107    1.0000000    0.5605221    0.6639987
## Largo.Petalo    0.7540490    0.5605221    1.0000000    0.7866681
## Ancho.Petalo    0.5464611    0.6639987    0.7866681    1.0000000
```

8.- Obtencion de valores/vectores propios de la **matriz de correlaciones**.

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

```
## eigen() decomposition
## $values
## [1] 2.9263407 0.5462747 0.3949976 0.1323871
##
## $vectors
##            [,1]       [,2]       [,3]       [,4]
## [1,] -0.4823284  0.6107980 -0.4906296  0.3918772
## [2,] -0.4648460 -0.6727830 -0.5399025 -0.1994658
## [3,] -0.5345136  0.3068495  0.3402185 -0.7102042
## [4,] -0.5153375 -0.2830765  0.5933290  0.5497778
```

9.- Separación de valores propios desde la matriz de correlaciones.

9.1.- Obtención de valores propios.

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

```
## [1] 2.9263407 0.5462747 0.3949976 0.1323871
```

9.2.- Obtención de vectores propios.

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

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -0.4823284  0.6107980 -0.4906296  0.3918772
## [2,] -0.4648460 -0.6727830 -0.5399025 -0.1994658
## [3,] -0.5345136  0.3068495  0.3402185 -0.7102042
## [4,] -0.5153375 -0.2830765  0.5933290  0.5497778
```

10.-Calculo de la Proporcion variabilidad.

10.1.- Variabilidad para valores propios.

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

```
## [1] 0.121968486 0.018096024 0.013694021 0.002447591
```

10.2.- Proporcion de variabilidad acumulada de valores propios. Aquí se seleccionan el número de componentes, siguiendo el criterio del 80% de varianza explicada. En el ejemplo se seleccionarán 2 factores (0.868% de varianza explicada).

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

```
## [1] 0.7315852 0.8681538 0.9669032 1.0000000
```

11.- Media de los valores propios.

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

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

Obtención de coeficientes.

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

12.1 Matriz 1

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

12.2.- Construccion de la matriz centrada.

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

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

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

```
##           [,1]       [,2]      [,3]       [,4]
## [1,] 0.2664327 0.00000000 0.0000000 0.00000000
## [2,] 0.0000000 0.09846939 0.0000000 0.00000000
## [3,] 0.0000000 0.00000000 0.2208163 0.00000000
## [4,] 0.0000000 0.00000000 0.0000000 0.03910612
```

14.- Construccion de la matriz centrada multiplicada por Dx^1/2.

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

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

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

```
##           [,1]       [,2]        [,3]         [,4]
## 51 -2.32455278  0.5185273 -1.21059316  0.075191200
## 52 -1.79699308 -0.4652131 -0.48504815  0.199955742
## 53 -2.57106666  0.6020469 -0.49865033  0.038577169
## 54  1.46714905  0.3591890  0.95682822  0.288414020
## 55 -1.41164332  0.5760181  0.18051660  0.378999671
## 56 -0.02915352 -0.1496476  0.26845808 -0.633250224
## 57 -2.33977751 -0.8104931 -0.11721324  0.036211804
## 58  3.45770058 -0.5928617 -0.05182738 -0.006758222
## 59 -1.13202813  0.7662442 -0.68666085 -0.164670936
## 60  1.00808930 -1.0618537  0.78140281  0.235542894
```

16.- Nombramos las columnas PC1, PC2, PC3, PC4.

```
colnames(scores)<-c("PC1","PC2","PC3","PC4")
```

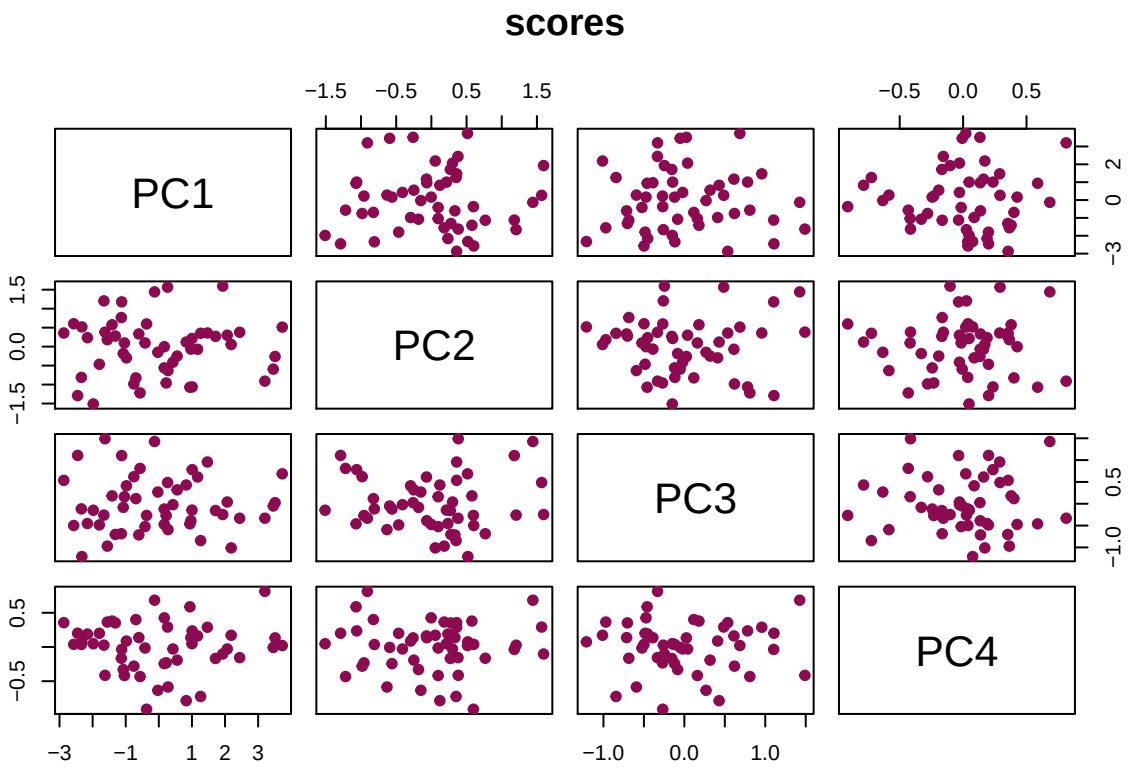
17.- Visualizamos los primeros 10 datos.

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

```
##            PC1        PC2         PC3          PC4
## 51 -2.32455278  0.5185273 -1.21059316  0.075191200
## 52 -1.79699308 -0.4652131 -0.48504815  0.199955742
## 53 -2.57106666  0.6020469 -0.49865033  0.038577169
## 54  1.46714905  0.3591890  0.95682822  0.288414020
## 55 -1.41164332  0.5760181  0.18051660  0.378999671
## 56 -0.02915352 -0.1496476  0.26845808 -0.633250224
## 57 -2.33977751 -0.8104931 -0.11721324  0.036211804
## 58  3.45770058 -0.5928617 -0.05182738 -0.006758222
## 59 -1.13202813  0.7662442 -0.68666085 -0.164670936
## 60  1.00808930 -1.0618537  0.78140281  0.235542894
```

18.- Elaboración del gráfico de score.

```
pairs(scores, main="scores", col="deeppink4", pch=19)
```



ACP VÍA SINTETIZADA

1.- Cálculo de la varianza a las columnas (1=filas, 2=columnas).

```
apply(X, 2, var)
```

```
## Largo.Sepalo Ancho.Sepalo Largo.Petalo Ancho.Petalo
##   0.26643265   0.09846939   0.22081633   0.03910612
```

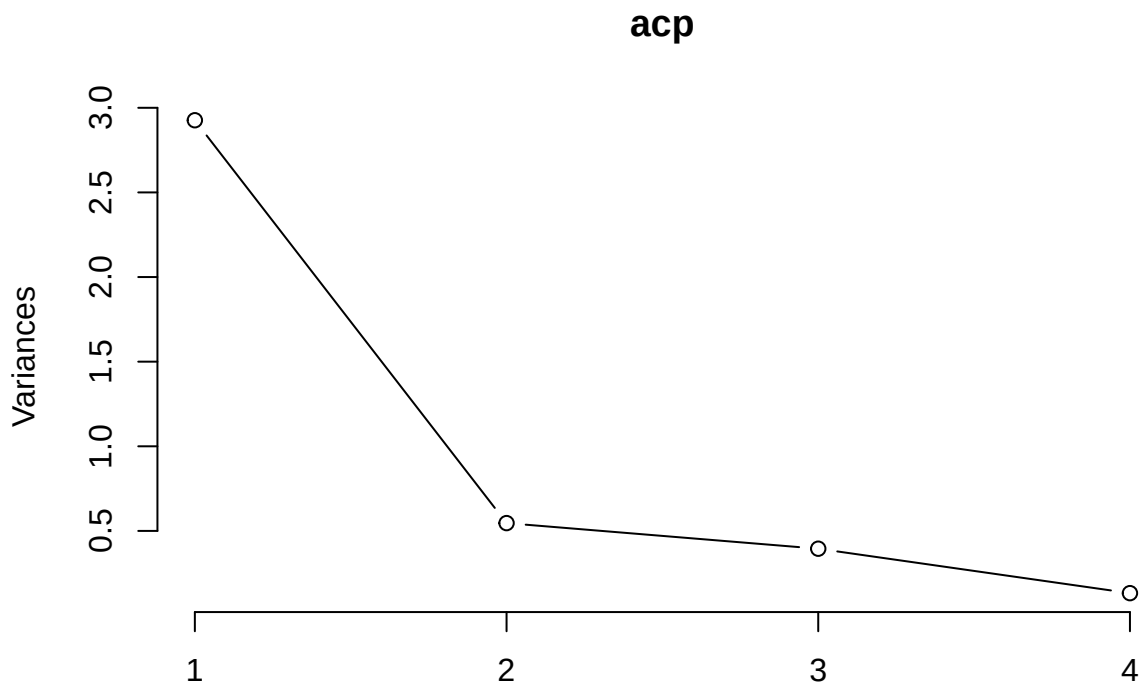
2.- Aplicar la función **prcomp** para la reducción de dimensión y centrado porla media y escalada por ladesviación estandar (división entre sd).

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

```
## Standard deviations (1, .., p=4):
## [1] 1.7106550 0.7391040 0.6284883 0.3638504
##
## Rotation (n x k) = (4 x 4):
##                      PC1        PC2        PC3        PC4
## Largo.Sepalo -0.4823284 -0.6107980  0.4906296  0.3918772
## Ancho.Sepalo -0.4648460  0.6727830  0.5399025 -0.1994658
## Largo.Petalo -0.5345136 -0.3068495 -0.3402185 -0.7102042
## Ancho.Petalo -0.5153375  0.2830765 -0.5933290  0.5497778
```

3.- Generación del gráfico **screeplot**.

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



4.- Visualizar el resumen de la matriz **acp**.

```
summary(acp)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4
## Standard deviation     1.7107 0.7391 0.62849 0.3639
## Proportion of Variance 0.7316 0.1366 0.09875 0.0331
## Cumulative Proportion  0.7316 0.8681 0.96690 1.0000
```

Construccion de los componentes principales con las variables originales .

Combinación lineal de las variables originales.

Z1= -0.482(var1)-0.464(var2)-0.534(var3)-0.515(var4)

El primer componente distingue entre flores grandes y pequeñas.

Sépalo corto Sépalo angosto *Pétalo corto* Pétalo angosto

Z2= -0.610(var1)+0.672(var2)-0.306(var3)+0.283(var4)

El segundo componente distingue entre flores por especie.

Sépalo corto Sépalo angosto *Pétalo corto* Pétalo angosto