# Análisis de Componentes Principales

Ana Karen Martínez Marín

11/3/2022

## Análisis de Componentes Principales de la base "Atmósfera"

### Introducción

El Análisis de Componentes Principales (**ACP**) es un método de reducción de la dimensionalidad de las variables originales, pero procurando que se pierda la menor cantidad posible de información. Este análisis se obtiene a partir de la matriz de correlaciones principalmente.

1.-Se trabajó con la matriz (**Atmósfera**), estraída del paquete **datos** que se encuentra precargado en R.

```
library(datos)
```

2.- Se selecciona la matriz de datos (atmóstera)

```
x<-datos::atmosfera
```

Se trabajará con una parte de la matriz de datos, debido a que son muchos, seleccionando sólo las observaciones del mes de diciembre del año 2000.

```
x<-x[40897:41472,1:11]
```

### Exploración de la matriz

Conocer la dimensión de la matriz

```
dim(x)
```

```
## [1] 576  11
```

Tipos de variables

```
str(x)
```

```
## tibble [576 x 11] (S3: tbl_df/tbl/data.frame)
##  $ latitud        : num [1:576] 36.2 33.7 31.2 28.7 26.2 ...
##  $ longitud       : num [1:576] -114 -114 -114 -114 -114 ...
##  $ anio           : int [1:576] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ mes            : int [1:576] 12 12 12 12 12 12 12 12 12 12 ...
##  $ temp_superficie: num [1:576] 278 284 288 291 293 ...
##  $ temperatura    : num [1:576] 284 289 292 293 294 ...
##  $ presion        : num [1:576] 975 970 990 995 1000 1000 1000 1000 1000 1000 ...
##  $ ozono          : num [1:576] 296 294 290 282 276 274 264 260 254 252 ...
##  $ nube_baja      : num [1:576] 7.5 8 13 15 19.5 21 25 22.5 21 22 ...
##  $ nube_media     : num [1:576] 22.5 18.5 19 14.5 13 14.5 17.5 14 11 15 ...
##  $ nube_alta      : num [1:576] 12 9 8 8 8 10.5 14 14.5 15.5 19.5 ...
```

Nombre de las variables

```
colnames(x)
```

```
##  [1] "latitud"         "longitud"        "anio"            "mes"
##  [5] "temp_superficie" "temperatura"     "presion"         "ozono"
##  [9] "nube_baja"       "nube_media"      "nube_alta"
```

Datos perdidos

```
anyNA(x)
```

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

Como se pudo notar hay datos nulos, que están en la variable nube_baja. Por lo que se quitará de la matriz para que no halla ningún problema.

Para este análisis se requiere solamente variables cuantitativas, por lo que el mes y año las quitaremos de nuestra base de datos. Para reducir un poco más el número de las variables también se quitarán: longitud y latitud.

## Tratamiento de la matriz

Se genera una nueva matriz **x1** filtrada:

```
x$anio <- NULL
x$mes <- NULL
x$latitud <- NULL
x$longitud <- NULL
x$nube_baja <- NULL
x1=x
```

Se comprueba que ya no halla datos nulos:

```
anyNA(x1)
```

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

Salió falso, así que se continúa trabajando sin ningún problema.

## ACP paso a paso

1.- Transformar la matriz en un data.frame(x1).

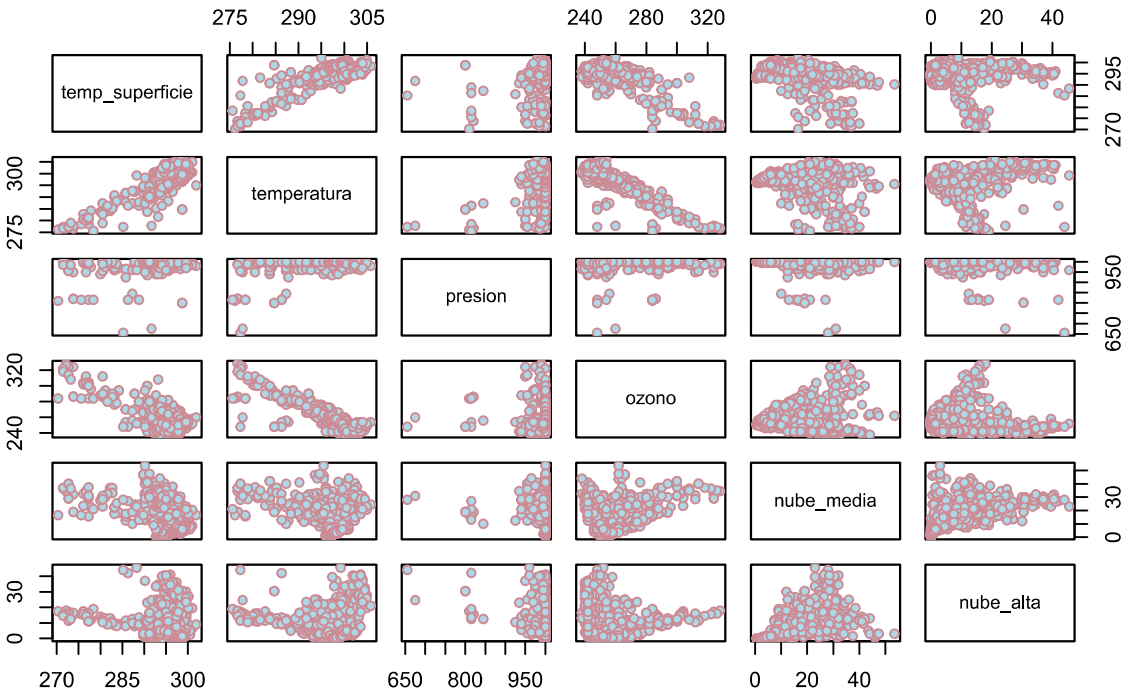
```
x1<-as.data.frame(x1)
```

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

```
n<-dim(x1)[1]
p<-dim(x1)[2]
```

3.- Generación de un scatterplot de las variables originales de la matriz ya filtrada.

```
pairs(x1,col="lightpink3", pch=21,bg = "lightblue", cex = 0.9, lwd=0.9, main="Variables originales")
```




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

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

```
## temp_superficie     temperatura         presion           ozono      nube_media
##        294.09167        296.38125        990.16493        258.88889         17.51823
##       nube_alta
##         10.81424
```

```
S<-cov(x1)
S
```

```
##                 temp_superficie temperatura    presion     ozono nube_media
## temp_superficie       31.830678   28.097252   69.48746 -70.20545  -25.44576
## temperatura           28.097252   34.401978   86.21875 -81.67826  -15.46583
## presion               69.487464   86.218750 1115.32058 -56.94686  -53.85084
## ozono               -70.205449  -81.678261  -56.94686 259.91111   42.11072
## nube_media          -25.445761  -15.465832  -53.85084  42.11072  110.20619
## nube_alta            -1.362159    7.563207  -84.63453 -20.02763   55.90731
##                  nube_alta
## temp_superficie  -1.362159
## temperatura       7.563207
## presion         -84.634526
## ozono           -20.027633
## nube_media       55.907305
## nube_alta       104.744562
```

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

```
es<-eigen(S)
es
```

```
## eigen() decomposition
## $values
## [1] 1142.583485  307.131223  153.398246   42.180763    8.339653    2.781724
##
## $vectors
##             [,1]        [,2]        [,3]        [,4]        [,5]         [,6]
## [1,] -0.07024988  0.24464054  0.04239502 -0.11174021  0.89442567  0.347757768
## [2,] -0.08452296  0.27119837 -0.06356428 -0.04213744  0.27305275 -0.915933148
## [3,] -0.98554758 -0.11591882 -0.09324648 -0.02475154 -0.06214501  0.045708190
## [4,]  0.07809815 -0.89964620  0.07837047 -0.24999312  0.28789976 -0.181693763
## [5,]  0.06201475 -0.18679566 -0.70649979  0.65850163  0.16861690  0.007971697
## [6,]  0.08168005  0.09438945 -0.69295271 -0.69928891 -0.10131624  0.070466968
```

5.1.- Separación de la matriz de valores propios.

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

```
## [1] 1142.583485  307.131223  153.398246   42.180763    8.339653    2.781724
```

5.2.- Separación de la matriz de vectores propios.

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

```
##             [,1]        [,2]        [,3]        [,4]        [,5]         [,6]
## [1,] -0.07024988  0.24464054  0.04239502 -0.11174021  0.89442567  0.347757768
## [2,] -0.08452296  0.27119837 -0.06356428 -0.04213744  0.27305275 -0.915933148
## [3,] -0.98554758 -0.11591882 -0.09324648 -0.02475154 -0.06214501  0.045708190
## [4,]  0.07809815 -0.89964620  0.07837047 -0.24999312  0.28789976 -0.181693763
## [5,]  0.06201475 -0.18679566 -0.70649979  0.65850163  0.16861690  0.007971697
## [6,]  0.08168005  0.09438945 -0.69295271 -0.69928891 -0.10131624  0.070466968
```

6.- Calcular la proporción de variabilidad para cada valor.

6.1.- La matriz de valores propios.

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

```
## [1] 0.689792968 0.185419237 0.092608578 0.025465092 0.005034761 0.001679364
```

6.2.- Acumulada.

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

```
## [1] 0.6897930 0.8752122 0.9678208 0.9932859 0.9983206 1.0000000
```

7.- Obtención de la matriz de correlaciones.

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

```
##                 temp_superficie temperatura    presion      ozono nube_media
## temp_superficie      1.00000000   0.8490811  0.3687938 -0.7718545 -0.4296250
## temperatura          0.84908114   1.0000000  0.4401596 -0.8637784 -0.2511763
## presion              0.36879384   0.4401596  1.0000000 -0.1057688 -0.1535994
## ozono               -0.77185452  -0.8637784 -0.1057688  1.0000000  0.2488154
## nube_media          -0.42962496  -0.2511763 -0.1535994  0.2488154  1.0000000
## nube_alta           -0.02359061   0.1259936 -0.2476179 -0.1213812  0.5203551
```

```
##                   nube_alta
## temp_superficie -0.02359061
## temperatura      0.12599361
## presion         -0.24761786
## ozono           -0.12138121
## nube_media       0.52035511
## nube_alta        1.00000000
```

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

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

```
## eigen() decomposition
## $values
## [1] 2.97716910 1.53743542 0.90203016 0.35776748 0.16749582 0.05810202
##
## $vectors
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## [1,] -0.54263200  0.04598222  0.06049457 -0.08956966  0.82113269  0.13226839
## [2,] -0.54515372  0.19869019 -0.09895998  0.07857706 -0.23139229 -0.77059705
## [3,] -0.27769208 -0.24887053 -0.85229807 -0.17306404 -0.16997688  0.27512772
## [4,]  0.49756370 -0.25402695 -0.26600510 -0.39116438  0.40786118 -0.54569374
## [5,]  0.28475212  0.53743694 -0.43327282  0.61155516  0.26042348 -0.02307216
## [6,]  0.05090814  0.73695532 -0.04063586 -0.65486066 -0.09557951  0.12114463
```

9.- Separación de la matriz de los valores propios a partir de la matriz de correlaciones.

9.1.- Separación de la matriz de los valores propios.

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

```
## [1] 2.97716910 1.53743542 0.90203016 0.35776748 0.16749582 0.05810202
```

9.2.- Separación de la matriz de los vectores propios.

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

```
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## [1,] -0.54263200  0.04598222  0.06049457 -0.08956966  0.82113269  0.13226839
## [2,] -0.54515372  0.19869019 -0.09895998  0.07857706 -0.23139229 -0.77059705
## [3,] -0.27769208 -0.24887053 -0.85229807 -0.17306404 -0.16997688  0.27512772
## [4,]  0.49756370 -0.25402695 -0.26600510 -0.39116438  0.40786118 -0.54569374
## [5,]  0.28475212  0.53743694 -0.43327282  0.61155516  0.26042348 -0.02307216
## [6,]  0.05090814  0.73695532 -0.04063586 -0.65486066 -0.09557951  0.12114463
```

10.- Cálculo de la proporción de variabilidad.

10.1.- Para la matriz de valores propios.

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

```
## [1] 0.49619485 0.25623924 0.15033836 0.05962791 0.02791597 0.00968367
```

10.2.- Acumulada.

En este punto se seleciiona en número de componentes, siguiendo el criterio del _____ de la varianza aplicada.

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

```
## [1] 0.4961949 0.7524341 0.9027724 0.9624004 0.9903163 1.0000000
```

11.- Calcular la 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.- Construcción de la matriz de 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 covarianzas.

```
Dx<-diag(diag(S))
Dx
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## [1,] 31.83068  0.00000    0.000   0.0000   0.0000   0.0000
## [2,]  0.00000 34.40198    0.000   0.0000   0.0000   0.0000
## [3,]  0.00000  0.00000 1115.321   0.0000   0.0000   0.0000
## [4,]  0.00000  0.00000    0.000 259.9111   0.0000   0.0000
## [5,]  0.00000  0.00000    0.000   0.0000 110.2062   0.0000
## [6,]  0.00000  0.00000    0.000   0.0000   0.0000 104.7446
```

14.- Construcción 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 matriz de autovectores.

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

```
##  [1]  4.1115927  2.9249240  1.9815001  1.2455680  0.6075567  0.3341787
##  [7] -0.1404411 -0.6582945 -0.9583432 -1.0234141
```

16.- Nombramos las columnas PC1...PCN

```
colnames(scores)<-c("PC1","PC2","PC3","PC4","PC5","PC6")
```

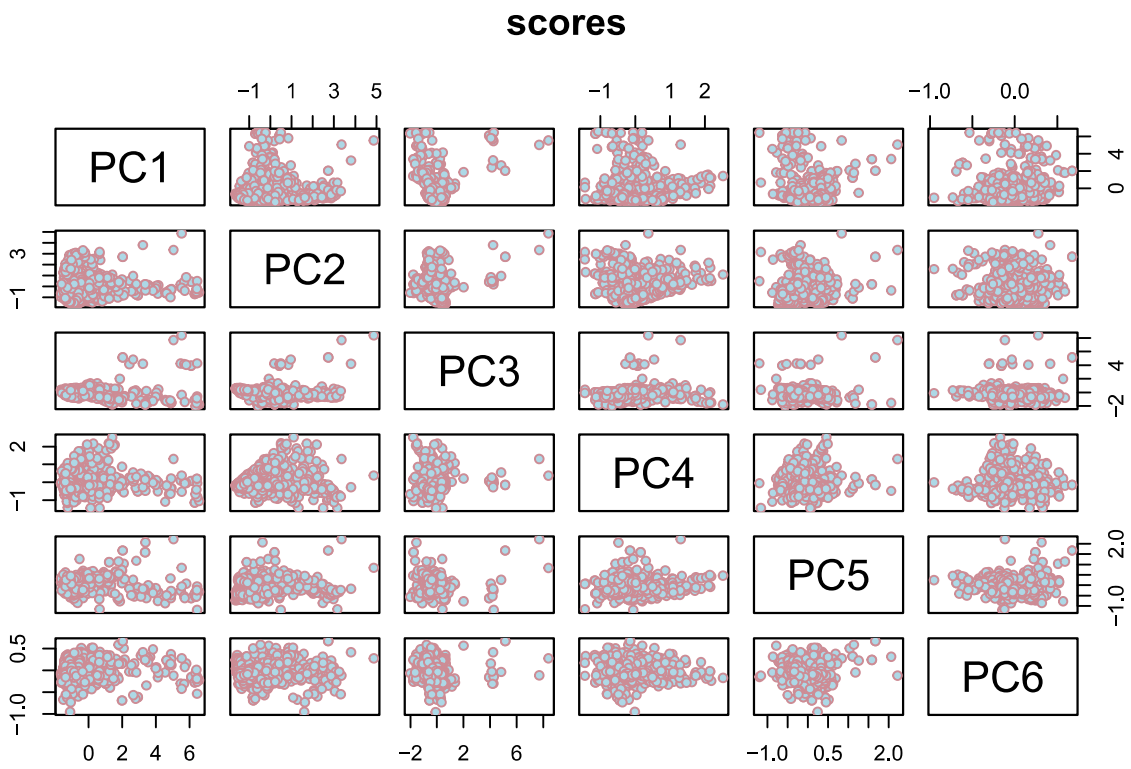
17.- Visualización de los datos.

```
scores[1:5,]
```

```
##            PC1        PC2         PC3        PC4        PC5        PC6
## [1,] 4.1115927 -0.6767437 -0.40478628 -0.5120509 -0.7620104 -0.1595395
## [2,] 2.9249240 -0.8207820 -0.07617081 -0.5186566 -0.1083725 -0.6140685
## [3,] 1.9815001 -0.8086879 -0.55601454 -0.4437369  0.1002020 -0.6891463
## [4,] 1.2455680 -0.9096964 -0.34323541 -0.5770312  0.1630793 -0.3649820
## [5,] 0.6075567 -0.8614067 -0.30674646 -0.5656794  0.2572528 -0.2427290
```

18.- Genaración del gráfico de los scores

```
pairs(scores,col="lightpink3", pch=21,bg = "lightblue", cex = 0.9, lwd=0.9,
      main="scores")
```




## ACP VÍA SINTETIZADA DE ACP

```
head(x1)
```

```
##   temp_superficie temperatura presion ozono nube_media nube_alta
## 1           277.8       284.2     975   296       22.5       12.0
## 2           284.2       288.8     970   294       18.5        9.0
## 3           287.8       292.2     990   290       19.0        8.0
## 4           290.7       292.7     995   282       14.5        8.0
## 5           293.2       294.1    1000   276       13.0        8.0
## 6           294.6       295.5    1000   274       14.5       10.5
```

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

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

```
## temp_superficie     temperatura         presion           ozono      nube_media
##        31.83068        34.40198      1115.32058       259.91111       110.20619
##       nube_alta
##       104.74456
```

2.- Aplicar la función **prcomp** para reducir la dimensionalidad y centrado por la media y escalada por la desviación estándar (dividir entre sd).

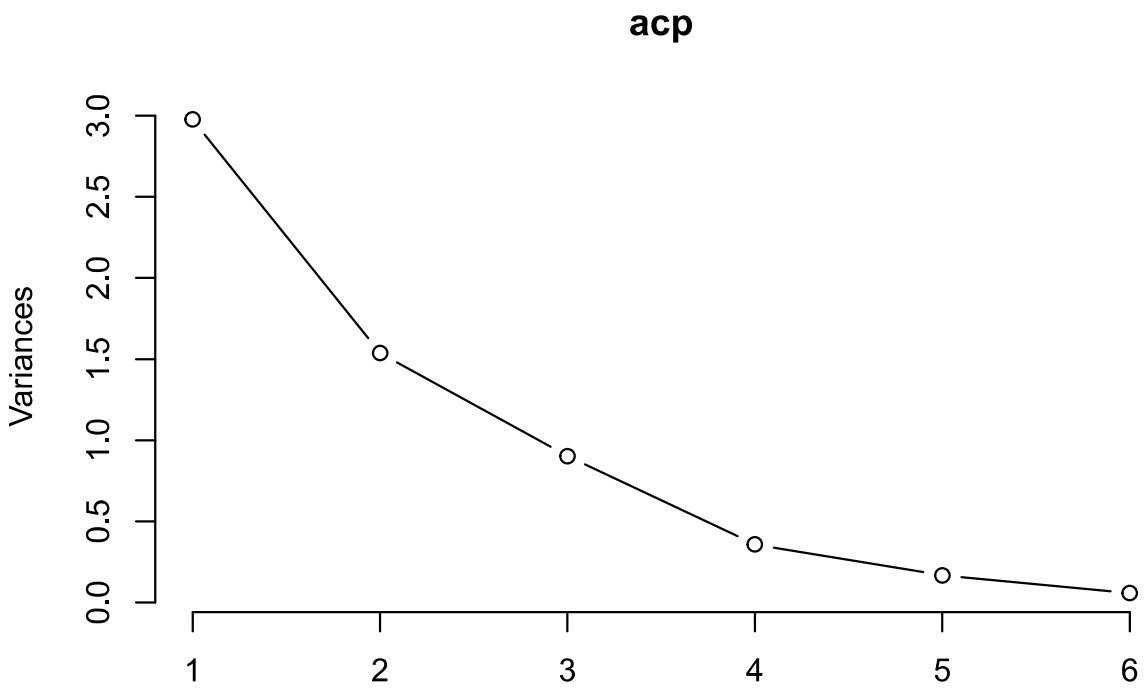
```
acp<-prcomp(x1, center=TRUE, scale=TRUE)
acp
```

```
## Standard deviations (1, .., p=6):
```

```
## [1] 1.7254475 1.2399336 0.9497527 0.5981367 0.4092625 0.2410436
##
## Rotation (n x k) = (6 x 6):
##                          PC1         PC2         PC3         PC4         PC5
## temp_superficie  0.54263200  0.04598222 -0.06049457 -0.08956966  0.82113269
## temperatura      0.54515372  0.19869019  0.09895998  0.07857706 -0.23139229
## presion          0.27769208 -0.24887053  0.85229807 -0.17306404 -0.16997688
## ozono           -0.49756370 -0.25402695  0.26600510 -0.39116438  0.40786118
## nube_media      -0.28475212  0.53743694  0.43327282  0.61155516  0.26042348
## nube_alta       -0.05090814  0.73695532  0.04063586 -0.65486066 -0.09557951
##                          PC6
## temp_superficie -0.13226839
## temperatura      0.77059705
## presion         -0.27512772
## ozono            0.54569374
## nube_media       0.02307216
## nube_alta       -0.12114463
```

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

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




4.- Resumen de la matriz **acp**.

```
summary(acp)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7254 1.2399 0.9498 0.59814 0.40926 0.24104
## Proportion of Variance 0.4962 0.2562 0.1503 0.05963 0.02792 0.00968
## Cumulative Proportion  0.4962 0.7524 0.9028 0.96240 0.99032 1.00000
```

## Construcción de los CP con las variables originales.

Combinación lineal de las variables originales.

z1= 0.542(temp_superficie)+0.545(temperatura)+0.277(presion)-0.497(ozono)-0.284(nube_media)-0.0509(nube_alta)

Se tomaron los primeros 3 componentes: - El primer componente distingue entre la temperatura y la temperatura de la superficie.

- El segundo componente distingue entre las nubes altas y medias.
- Y el tercer componente se distingue presión y nube media.