Análisis Estadístico Multivariante

Problemas Propuestos de Análisis de Componentes Principales

Francisco Javier Mercader Martínez

Problema 1

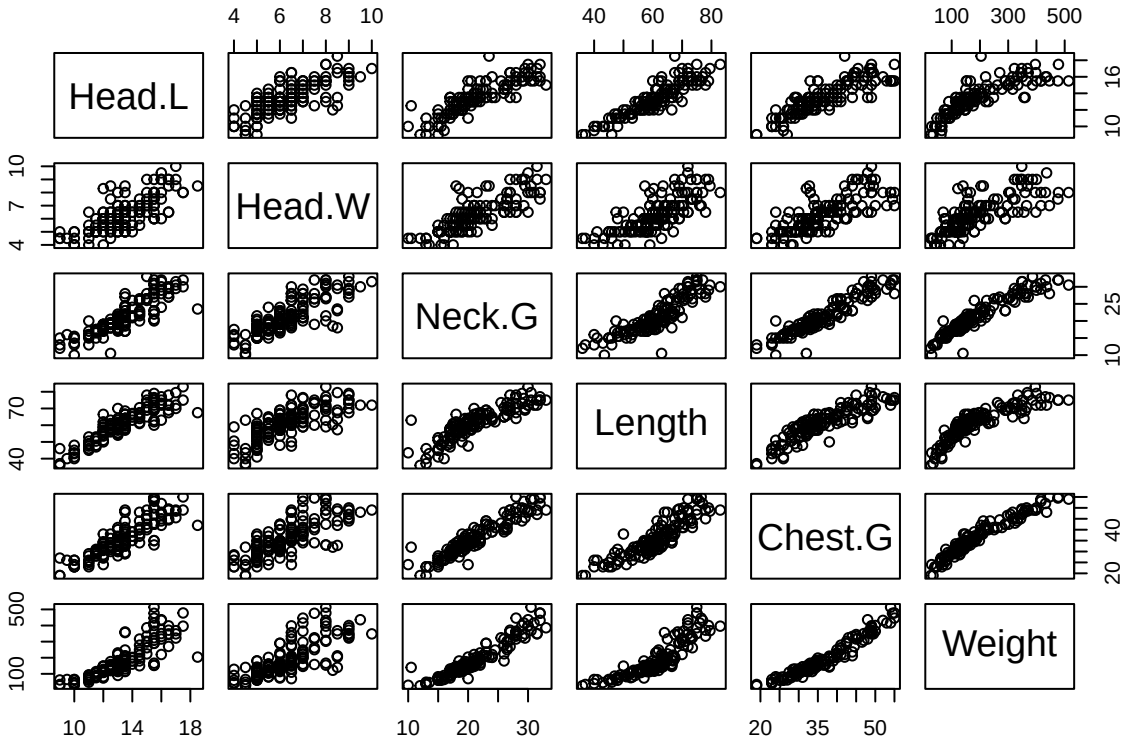
Consideremos los datos del dichero `Bears.rda` (disponible en el aula virtual), que contiene información diversa de 143 osos. En particular las columnas 5 a 10 vienen dadas por: Head.L= longitud de la cabeza (pulgadas), Head.W=anchura de la cabeza (pulgadas), Neck.G=perímetro cuello (pulgadas), Length=altura (pulgadas), Chest.G=perímetro pecho (pulgadas), Weight=peso (libras). Se pretende realizar un Análisis de Componentes Principales usando sólo las 6 variables descritas anteriormente (el resto de variables del fichero sólo se utilizarán para disponer de información complementaria de cada oso, no para el análisis ACP). Se pide:

1) Recuperar los datos usando la función *load()* y realizar un estudio descriptivo previo atendiendo a nuestro objetivo. En particular debes dar respuesta a las cuestiones:

```
load("../data/Bears.rda")
nombres_fila <- make.unique(as.character(d$Name))
row.names(d) <- nombres_fila
d <- d[, 5:10]
```

   a. ¿Tiene sentido plantearse un Análisis de Componentes Principales para estos datos?

```
pairs(d)
```



Podemos observar que las variables están correlacionadas entre sí, por lo que tiene sentido plantearse un ACP.

b. ¿Todas las variables se miden en magnitudes similares y presentan dispersión similar?

```
summary(d)
```

```
##      Head.L          Head.W          Neck.G          Length     
##  Min.   : 9.00   Min.   : 4.000   Min.   :10.00   Min.   :36.00  
##  1st Qu.:12.00   1st Qu.: 5.500   1st Qu.:18.00   1st Qu.:57.00  
##  Median :13.50   Median : 6.000   Median :20.00   Median :61.00  
##  Mean   :13.42   Mean   : 6.331   Mean   :21.33   Mean   :61.28  
##  3rd Qu.:15.00   3rd Qu.: 7.000   3rd Qu.:25.00   3rd Qu.:67.25  
##  Max.   :18.50   Max.   :10.000   Max.   :33.00   Max.   :83.00  
##     Chest.G          Weight      
##  Min.   :19.00   Min.   : 26.0  
##  1st Qu.:30.75   1st Qu.:118.0  
##  Median :35.00   Median :154.0  
##  Mean   :36.31   Mean   :192.2  
##  3rd Qu.:42.00   3rd Qu.:249.0  
##  Max.   :55.00   Max.   :514.0
```

Las variables no están en la misma escala, por lo que es conveniente estandarizarlas.

c. ¿Conviene usar la matriz de covarianzas, o es preferible la matriz de correlaciones a la hora de extraer las componentes principales?

```
cov(d)
```

```
##             Head.L     Head.W     Neck.G     Length    Chest.G     Weight
## Head.L    3.690092   1.876896   8.398079  16.082678  13.513799   176.9573
## Head.W    1.876896   1.726371   5.364930   9.048429   8.179447   109.7492
## Neck.G    8.398079   5.364930  25.699462  41.388931  39.234424   528.6139
## Length   16.082678   9.048429  41.388931  87.468354  68.442135   904.2627
## Chest.G  13.513799   8.179447  39.234424  68.442135  67.807780   879.3288
## Weight  176.957254 109.749242 528.613892 904.262691 879.328834 12220.0937
```

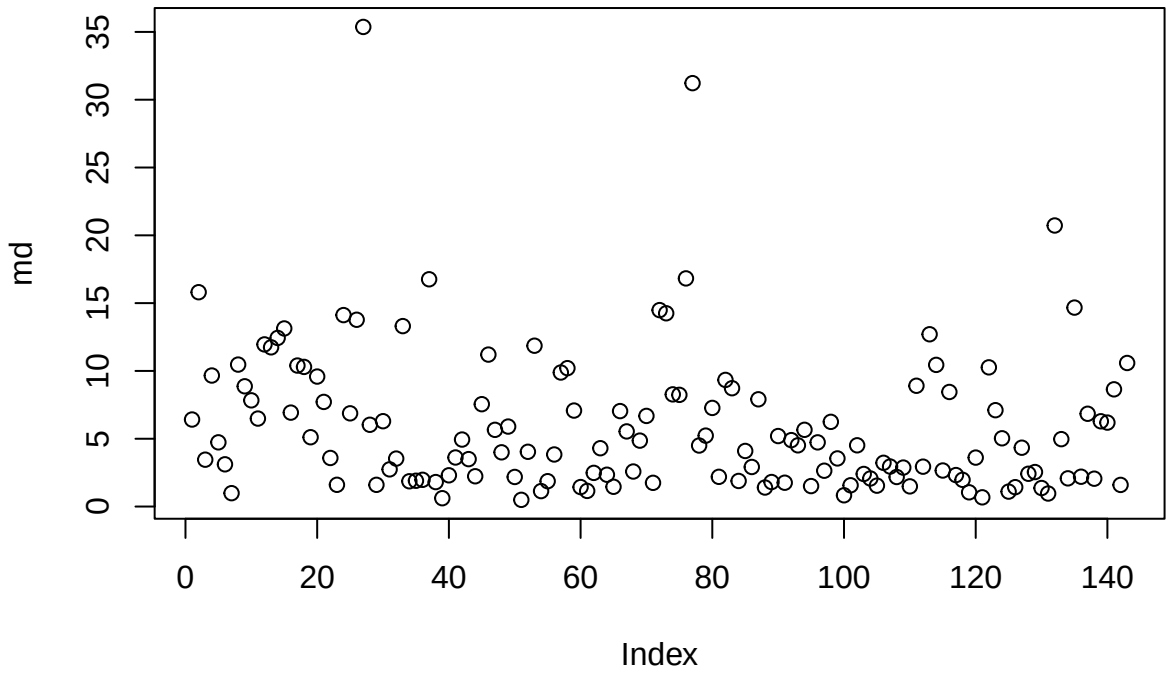
```
cor(d)
```

```
##            Head.L    Head.W    Neck.G    Length   Chest.G    Weight
## Head.L  1.0000000 0.7436261 0.8623814 0.8951881 0.8543171 0.8333214
## Head.W  0.7436261 1.0000000 0.8054434 0.7363439 0.7559920 0.7556092
## Neck.G  0.8623814 0.8054434 1.0000000 0.8729648 0.9398650 0.9432771
## Length  0.8951881 0.7363439 0.8729648 1.0000000 0.8887064 0.8746451
## Chest.G 0.8543171 0.7559920 0.9398650 0.8887064 1.0000000 0.9659937
## Weight  0.8333214 0.7556092 0.9432771 0.8746451 0.9659937 1.0000000
```

Es preferible usar la matriz de correlaciones, ya que las variables no están en la misma escala.

d. ¿Existe alguna observación inusual, es decir alejada del resto atendiendo a la distancia de Mahalanobis?

```
md <- mahalanobis(d, colMeans(d), cov(d))
plot(md, main = "Distancias de Mahalanobis")
```



No hay observaciones inusuales.

2) Obtener la expresión de todas las componentes principales en función de las variables originales y dar una interpretación de las dos primeras componentes. ¿Para qué podría servir un ACP con estos datos?

```
PCA <- princomp(d, cor = TRUE)
summary(PCA)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4      Comp.5
## Standard deviation     2.2917117 0.5737221 0.47707185 0.31940005 0.238387848
## Proportion of Variance 0.8753238 0.0548595 0.03793292 0.01700273 0.009471461
## Cumulative Proportion  0.8753238 0.9301833 0.96811620 0.98511893 0.994590391
##                            Comp.6
## Standard deviation     0.180160082
## Proportion of Variance 0.005409609
## Cumulative Proportion  1.000000000
```

La primera componente principal está fuertemente asociada con el peso y la longitud de la cabeza, mientras que la segunda componente principal está fuertemente asociada con la anchura de la cabeza y el perímetro del cuello. Un ACP con estos datos podría servir para reducir la dimensionalidad de los datos y estudiar la relación entre las variables.

3) Calcular las puntuaciones (scores) e indicar los nombres de los osos con mayor y menor puntuación en la primera componente. ¿Qué significaría tener una mayor (o menor) puntuación en la primera componente principal?

Realizar un gráfico de las puntuaciones en la primera componente que incluya el nombre de los osos.

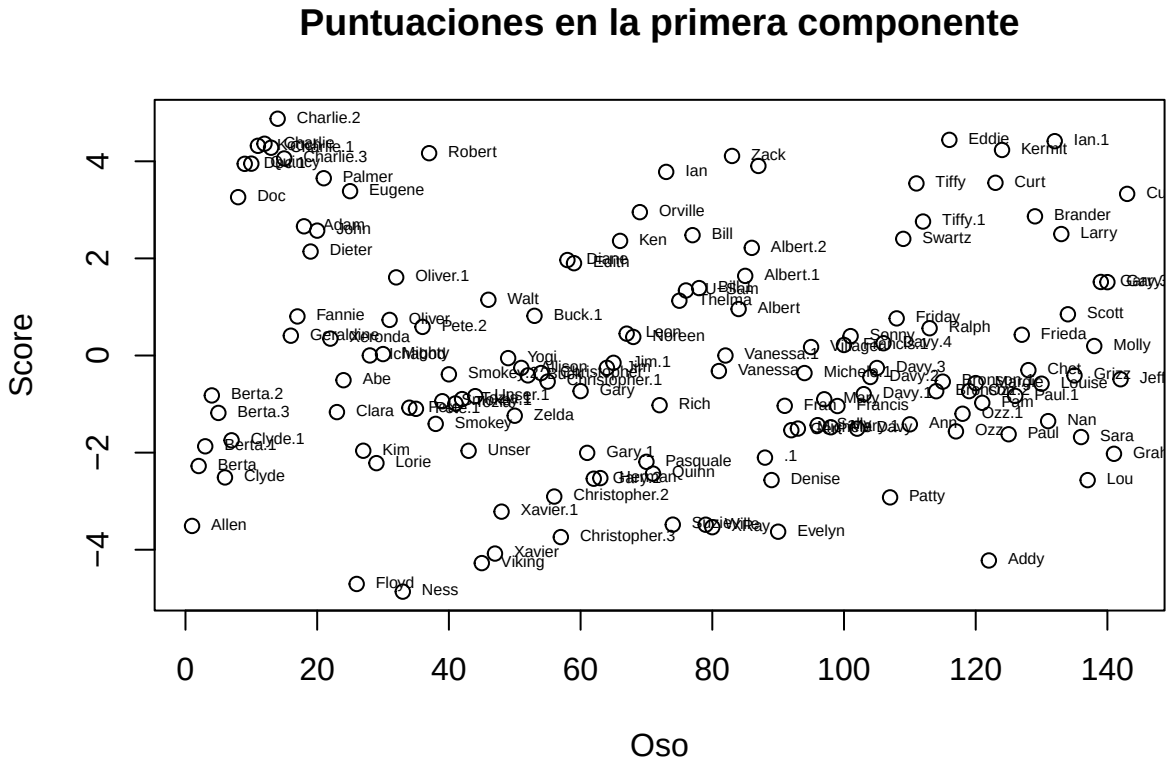
```
scores <- PCA$scores
print(paste("Oso con mayor puntuación en la primera componente:", nombres_fila[which.max(scores[,
 ↪  1])]))
```

```
## [1] "Oso con mayor puntuación en la primera componente: Charlie.2"
```

```
print(paste("Oso con menor puntuación en la primera componente:", nombres_fila[which.min(scores[,
 ↪  1])]))
```

```
## [1] "Oso con menor puntuación en la primera componente: Ness"
```

```
plot(scores[, 1], xlab = "Oso", ylab = "Score", main = "Puntuaciones en la primera componente")
text(1:143, scores[, 1], nombres_fila, cex = 0.5, pos = 4)
```



La puntación en la primera componente principal significa tener un mayor peso y longitud de la cabeza.

4) Repetir el apartado anterior pero mirando la segunda componente.

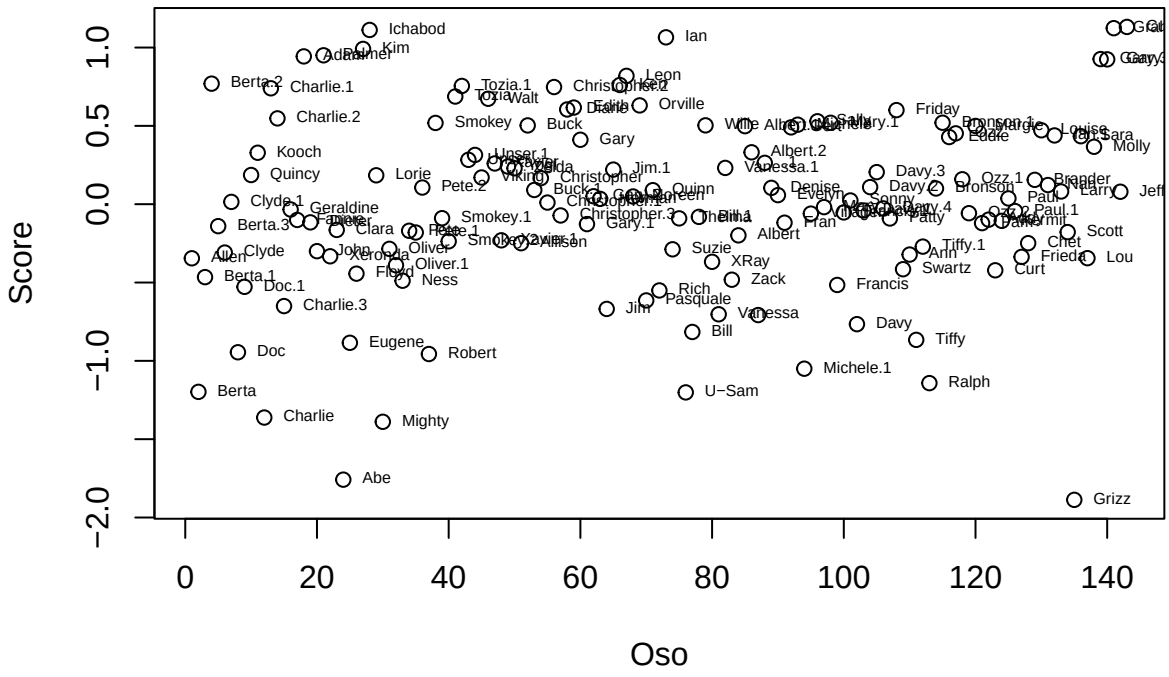
```
print(paste("Oso con mayor puntuación en la segunda componente:", nombres_fila[which.max(scores[, 2])]))
```

```
## [1] "Oso con mayor puntuación en la segunda componente: Curt.1"
```

```
print(paste("Oso con menor puntuación en la segunda componente:", nombres_fila[which.min(scores[, 2])]))
```

```
## [1] "Oso con menor puntuación en la segunda componente: Grizz"
```

```
plot(scores[, 2], xlab = "Oso", ylab = "Score", main = "Puntuaciones en la segunda componente")
text(1:143, scores[, 2], nombres_fila, cex = 0.5, pos = 4)
```



La puntación de la segunda componente significa tener una mayor anchura de la cabeza y perímetro del cuello.

5) Obtener la matriz de saturaciones y utilizarla para revisar la interpretación dada en el apartado (2).

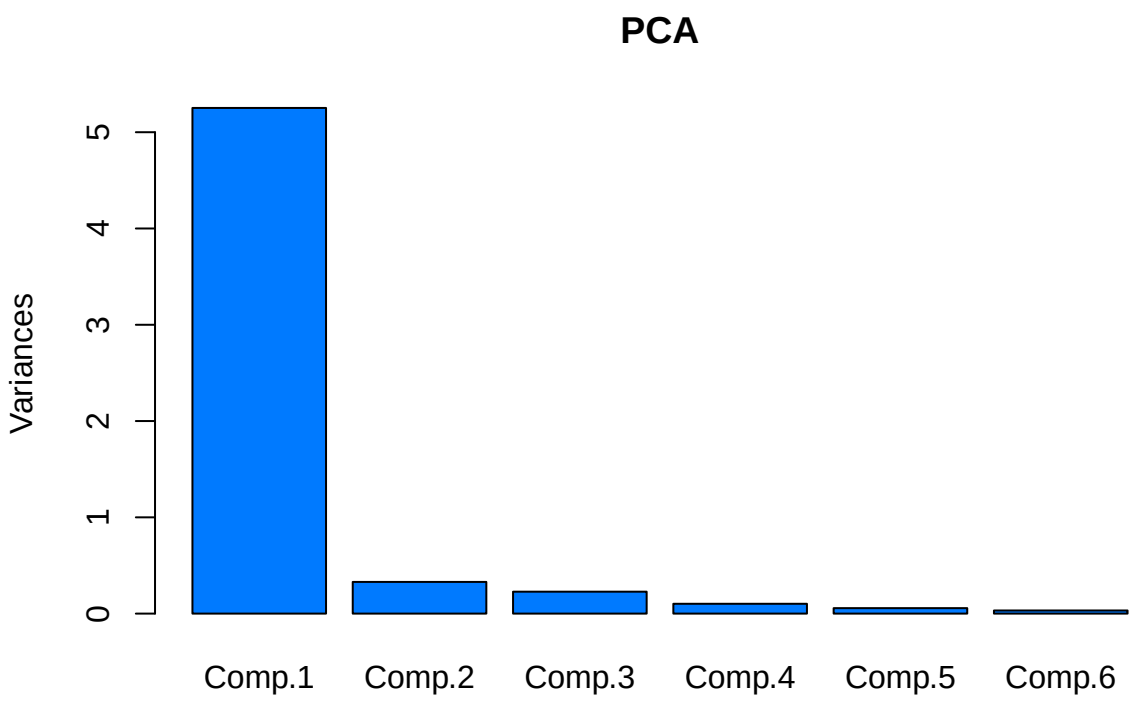
   Atendiendo a la matriz de saturaciones, identificar la variable mejor y peor representada (explicada) por cada componente principal.

```
matriz_saturaciones <- cor(d, PCA$scores)

mejor_variable <- apply(matriz_saturaciones, 2, function(x) which.max(abs(x)))

peor_variable <- apply(matriz_saturaciones, 2, function(x) which.min(abs(x)))

for (i in 1:6) {
  print(paste("Variable mejor representada por la componente", i, ":", colnames(d)[mejor_variable[i]]))
  print(paste("Variable peor representada por la componente", i, ":", colnames(d)[peor_variable[i]]))
}
```

```
## [1] "Variable mejor representada por la componente 1 : Neck.G"
## [1] "Variable peor representada por la componente 1 : Head.W"
## [1] "Variable mejor representada por la componente 2 : Head.W"
## [1] "Variable peor representada por la componente 2 : Neck.G"
## [1] "Variable mejor representada por la componente 3 : Head.L"
## [1] "Variable peor representada por la componente 3 : Head.W"
## [1] "Variable mejor representada por la componente 4 : Length"
## [1] "Variable peor representada por la componente 4 : Weight"
## [1] "Variable mejor representada por la componente 5 : Neck.G"
## [1] "Variable peor representada por la componente 5 : Head.W"
## [1] "Variable mejor representada por la componente 6 : Weight"
## [1] "Variable peor representada por la componente 6 : Head.W"
```

6) Determinar el número de componentes a retener usando diferentes criterios (porcentaje de variabilidad explicada, regla de Rao, regla de Kaiser y gráfico de sedimentación). ¿Sería razonable considerar sólo las 2 primeras componentes?

```r
# Regla de Kaiser
screeplot(PCA, col="#007AFF")
```




```r
# Regla de Rao
eigen(cov(d))$values
```

```
## [1] 1.237690e+04 2.189934e+01 4.190140e+00 2.336676e+00 6.252904e-01
## [6] 5.294126e-01
```

7) Calcular las comunalidades de cada variable en el caso de retener sólo las 2 primeras componentes. Identificar la variable mejor y peor representada (explicada) al retener sólo 2 componentes.

8) Representar a los individuos de la muestra (los osos) en el nuevo sistema de referencia dado por las 2 primeras componentes principales. Es decir representar la nube de puntos de las puntuaciones sin estandarizar correspondientes a las 2 primeras componentes.

9) Repetir el apartado anterior pero considerando puntuaciones estandarizadas e incluyendo las saturaciones. ¿Qué variable queda mejor representada en la segunda componente principal?

10) Para los gráficos de los dos apartados anteriores, etiquetar cada observación con el "sexo" del oso en lugar de su nombre. ¿Qué podemos destacar de dichos gráficos?