

# Análisis Estadístico Multivariante

## Problemas Propuestos de Análisis de Componentes Principales

Francisco Javier Mercader Martínez

### Problema 1

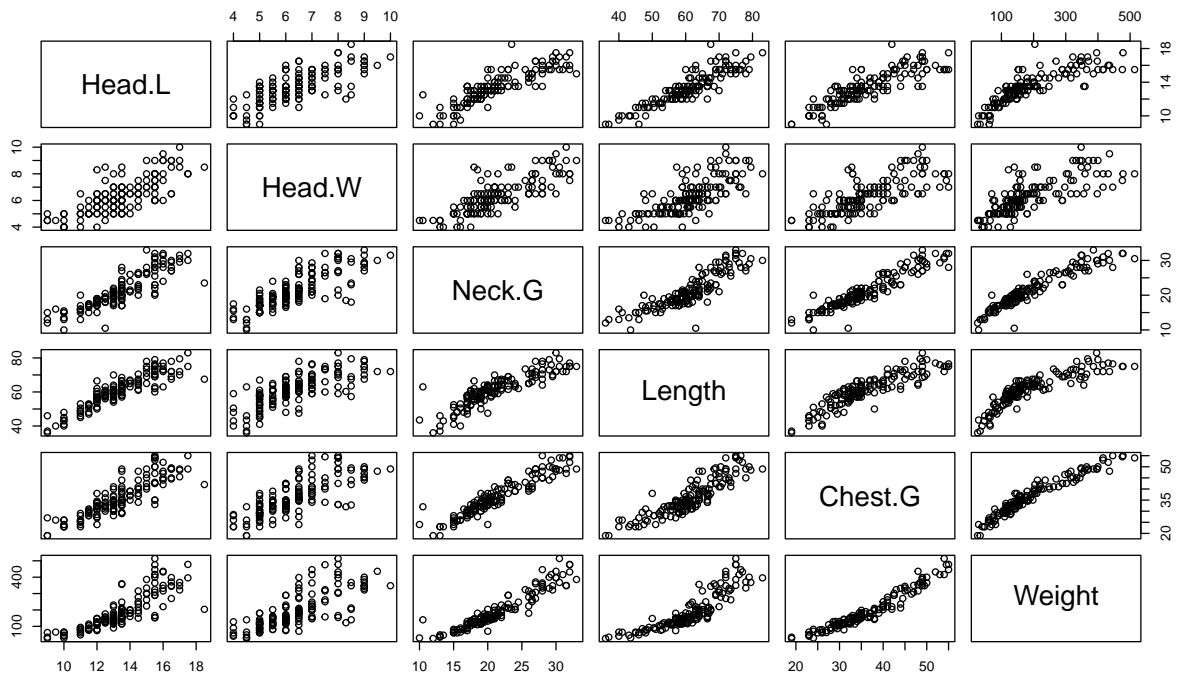
Consideremos los datos del fichero **Bears.rda** (disponible en el aula virtual), que contiene información diversa de 143 osos. En particular, las columnas 5 a 10 vienen dadas por: Head.L= longitud de la cabeza (pulgadas), Head.W=anchura de la cabeza (pulgadas), Neck.G=perímetro cuello (pulgadas), Length=altura (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")  
d <- d[, 5:10]
```

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

```
plot(d)
```



- b. ¿Todas las variables se miden en magnitudes similares y presentan dispersión similar?
  - c. ¿Conviene usar la matriz de covarianzas, o es preferible la matriz de correlaciones a la hora de extraer las componentes principales?
  - d. ¿Existe alguna observación inusual, es decir, alejada del resto atendiendo a la distancia de Mahalanobis?
- 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?
  - 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.
- 4) Repetir el apartado anterior, pero mirando la segunda componente.
  - 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.
  - 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?
  - 7) Calcular las comunidades 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?