IMM con R. Lab5: Análisis de Componentes Principales

Dra. Maite Mascaro y Dr. Edlin Guerra Castro

14/04/2021

Caso real: Gorriones de Bumpus.

Los datos siguientes provienen del famoso estudio del Dr. Herman Bumpus (1899) sobre la morfometría del gorrión *Passer domesicus*. Bumpus analizó 9 medidas corporales de una muestra de gorriones (hembras, machos y juveniles) que fueron encontrados helándose en Providence (USA) durante un invierno particularmente mordaz. Del total de 136 gorriones medidos, 72 sobrevivieron y 64 murieron eventualmente de frío y el Dr. Bumpus aprovechó este experimento natural para explorar si gorriones de ciertas características sucumbían más fácilmente al frío que otros. También exploró las diferencias morfométricas entre los sexos. Los datos se encuentran en el archivo 'bumpus.csv'.

Las medidas que Bumpus registró fueron:

- Total Length (TL) en mm
- Alar Extent (AE) en mm
- Length of Keel of Sternum (SKL) en pulgadas
- Length of Humerus (HL) en pulgadas
- Length of Femur (FL) en pulgadas
- Width of Skull (WS) en pulgadas
- Length of Tibio-Tarsus (LTT) en pulgadas
- Total Weight (WS) en gramos

Parte I: Exploración de datos

1. Considerando la hipotesis de Bumpus, ¿cuáles son las correlaciones que te interesa analizar: aquellas entre variables (columnas) para identificar grupos de observaciones (filas), o entre observaciones para identificar grupos de variables?

2. Explora los datos utilizando herramientas gráficas y numéricas apropiadas. Recuerda que puedes usar subsetting para seleccionar solo las columnas o filas de la tabla que contengan las variables que te interesan. (PISTA: recuerda funciones como `summary`, `plot` y `cor`)

a) ¿Sirven los resúmenes de nivel y dispersión por columnas o filas i.e. boxplots, para describir estos datos? Explica.
b) ¿Cómo es la correlación entre las variables medidas? ¿Hay variables mas fuertemente correlacionadas que otras? ¿Cómo es la dispersión en las correlaciones analizadas?
c) ¿Cuál es la proporción de hembras y machos y sobrevivientes y muertos en el juego de datos?
d) El siguiente código calcula los valores medios de las primeras 6 variables correspondientes a los 4 diferentes grupos de gorriones: hembras y machos que perecieron (hP, mP) y sobrevivieron (hS, mS). ¿Puedes determinar si las aves que sobrevivieron era distintas en su morfología de las que perecieron? ¿Puedes determinar si existen diferencias morfológicas entre machos y hembras?

```r
aggregate(dat[, c(5:10)], by = list(dat$sex, dat$sobreviv), mean)
```

e) Si se tratara de una sola variable morfométrica (BHL, por ejemplo), ¿cómo procederías para explorar la hipótesis de Bumpus? Explora algunas de las variables.

Parte II: Obtención de matriz de VCV y *eigen*-análisis

3. De la exploración aprendiste que las escalas de las variables difieren (i.e., AE está en centenas). Para eliminar el problema de la diferencia en escalas entre variables, copia el siguiente código que obtiene el log natural de las variables AE, BHL, FL, TTL, SW y SKL, y usando el comando `cbind` obtendremos una matriz de datos transformados. El resto del código es para poner nombre a las columnas.

```r
Y <- log(cbind(dat$AE, dat$BHL, dat$FL, dat$TTL, dat$SW, dat$SKL))
colnames(Y, do.NULL = FALSE)
colnames(Y) <- c("AE", "BHL", "FL", "TTL", "SW", "SKL")
```

a) Confirma que sea una matriz. ¿Qué dimensiones tiene?
b) Obtén la matriz de varianza-covarianza del objeto `Y` mediante la función `cov` y llámala `C`. ¿Qué dimensiones tiene la matriz C?
c) ¿Qué representan los elementos de la diagonal? ¿Qué representan los elementos fuera de la diagonal? Confirma esto usando tu conocimiento en estadística básica.
d) ¿Qué representa la traza (suma de elementos de la diagonal) de esta matriz en el juego de datos? Obtén la traza de la matriz. Usando tu conocimiento sobre ANOVA, piensa cuál sería otra forma de calcular dicho valor.
e) El siguiente códgio es para obtener la matriz varianza-covarianza a mano. A un lado están comentarios para ayudarte a seguir las instrucciones del código que están en la presentación. Córre el código paso por paso, y compara el resultado con el obtenido con la función `cov`.

```r
#Obtención de un vector vertical unitario
ones <- matrix(1, nrow = nrow(Y), ncol = 1)
dim(ones)

#Cálculo del número de filas (número de observaciones)
numFilas <- t(ones) %*% ones
numFilas

#Obtención del inverso de esa matriz de un único elemento
invnumFilas <- solve(t(ones) %*% ones)
invnumFilas

#Obtención de la sumatoria de elementos por columnas
sumaCol <- t(ones) %*% Y
sumaCol

#Mutiplica la sumaCol por el inverso del numFilas (divide la sumatoria entre n filas).
invnumFilas %*% sumaCol

#El vector de medias también se puede obtener así.
colMeans(Y)

#Multiplica el vector horizontal de medias por el vector vertical unitario para obtener
#una matriz de medias.
Y.bar <- ones %*% invnumFilas %*% sumaCol
dim(Y.bar)
```

```r
#Sustraer la matriz de datos de la matriz de medias.
Z <- Y - Y.bar

#Esto también se logra obteniendo los residuales de una regresion lineal contra solo un
#intercepto de valor 1.
Z <- resid(lm(Y ~ 1))

#Obtención de la matriz de suma de cuadrados y cross-products (SSCP matrix) (elevar al
# cuadrado la matriz Z)
S <- t(Z) %*% Z
S

#SSCP es una matriz triangular con la SS en la diagonal y los CP en los elementos fuera
#de la diagonal.
dim(S)

#Obtener la matriz de varianza covarianza: Dividir SSCP (suma de cuadrados y productos
#cruzados) entre n-1 (g.l.)
Cmano <- 1 / (nrow(Y) - 1) * S
Cmano
C
```

4. Aplica la función `cor` sobre la matriz `Y`, llama `R` al objeto generado e identifica cuál es el resultado de esta operación.

a) ¿Qué representan los valores en la diagonal? ¿Qué representan los valores fuera de la diagonal? ¿Por qué se repiten en la diagonal opuesta?
b) ¿Cómo representarías la matriz R de forma gráfica?
c) El siguiente códgio es para obtener la matriz de correlación a mano. Córrelo y compara el resultado.

```r
#Sustraer la matriz de datos de la matriz de medias para calcular la variación
#total: Y-Y.bar
Z <- resid(lm(Y ~ 1))

#Obtiene la matriz diagonal W, cuyos elementos (en la diagonal) son el inverso de
#la desviación estándar de cada una de las variables medidas
W <- diag(sqrt(1 / apply(Y, 2, var)))

#Obtiene una medida ponderada de Z usando el inverso de la desviación
#estandar W: (Y-Y.bar)* W
Zp <- Z %*% W

#Eleva al cuadrado y suma las distancias ponderadas Zp
Sr <- t(Zp) %*% Zp

#Divide los valores de la matriz Sr entre g.l.=n-1
Rmano <- Sr / (nrow(Y) - 1)
Rmano
```

Una forma de obtener un vector con los elementos diagonales de una matriz es la función `apply`. El valor `2` es para indicar que `var` debe aplicarse a las columnas de la matriz Y (si es `1` se aplicaría a las filas).

```r
apply(Y, 2, var)
```

5. Aplica un eigenanálisis a la matriz `C` para obtener los eigenvectores y eigenvalores de un PCA (de covarianza) de Y. Usa la función `eigen`.

a) ¿Qué porcentaje de la variación total en la morfometría de los gorriones está explicada por el primer componente? ¿Cuánta por el segundo?
b) ¿Cuál es la variable con la carga positiva más alta en el segundo componente? ¿Qué significa eso?
c) ¿Con qué componentes principales te parece que queda explicada suficiente variación en este experimento?
d) Aplica la función `prcomp` sobre la matriz Y, usando el siguiente código. Compara con el PCA a mano.

```
pca.cov <- prcomp(Y,
                  retx = T,
                  center = TRUE,
                  scale. = FALSE)
```

e) La función `names` aplicada a un objeto que es producto de ciertos procedimiento (p.e. clase anova, clase prcomp), proporciona el nombre de las listas con los resultados del procedimiento. Usa la función `names` sobre pca.cov para explorar los distintos resultados que ofrece `prcomp`. Intenta identificar que contiene cada uno de ellos.

6. Aplica un eigenanálisis a la matriz R que obtuviste a mano.

a) ¿Cómo difieren los eigenvalores de este PCA con respecto al realizado sobre la matriz C?
b) Aplica la función `prcomp` sobre la matriz Y, pero adiciona el argumento `scale.=T` y guarda el objeto como pca.cor. Compara la salida con la eigenanálisis hecho a mano sobre la matriz R.

```
pca.cor <- prcomp(Y,
                  retx = T,
                  center = TRUE,
                  scale. = TRUE)
```

c) ¿Por qué crees que aplicar un PCA sobre una matriz de correlación equivale a aplicar un PCA escalado sobre una matriz de covarianza?
d) ¿Cuál de los dos es un análisis más útil para los propósitos de un PCA?
e) ¿Qué porcentaje de la variación total en la morfometría de los gorriones está explicada ahora por el primer componente?¿Cuánta por el segundo? ¿Puedes explicar las diferencias?

Parte III: Visualización

7. La función enlatada `prcomp` calcula automáticamente las proyecciones de los eigenvectores sobre los ejes principales para poder visualizarlas. Estas proyecciones (o scores) las encuentras en la lista del objeto bajo el nombre `$x` del objeto clase `prcomp`. Para obtener las proyecciones a mano se requiere del siguiente código:

```
#Ajudica el nombre E.cov a una matriz con los eigenvectores de un PCA de covarianza
E.cov <- eigen(C)$vectors

#Obtiene las proyecciones como el producto matricial de Z (matriz de residuales Y-Y.barra)
#y la matriz E.cov
P.cov <- Z %*% E.cov

#Dimensiones igual a la matriz original Y.
dim(P.cov)

#Ajudica el nombre E.cor a una matriz con los eigenvectores de un PCA de correlación
E.cor <- eigen(Rmano)$vectors

#Obtiene las proyecciones como el producto matricial de Z
#(matriz de residuales Y-Y.barra)* W, y la matriz E.cor
P.cor <- Z %*% W %*% E.cor
```

```
#Recuerda que W es una matriz diagonal con el inverso de la desviación estándar de
#las variables de Y.

#Dimensiones igual a la matriz original Y.
dim(P.cor)
```

a) Obtén las proyecciones del PCA de correlación usando la versión enlatada y compara con las obtenidas a mano. Las proyecciones están en `$x` de la lista. (PISTA: para comparar matrices usa la función `head` te permite ver sólo las primeras 6 líneas de una matriz.

b) Elabora una gráfica 2D del PC1 y PC2 mediante el siguiente código. ¿Qué representa cada punto en la gráfica?

```
plot(proy[,2]~proy[,1],asp=1,cex=1, xlab = "PC1", ylab = "PC2" )
```

c) Para facilitar la interpretación del gráfico, pinta de colores distintos los puntos de acuerdo al factor sobreviv. Identifica cuál es cuál según su color.

```
plot(
  proy[, 2] ~ proy[, 1],
  asp = 1,
  cex = 1,
  xlab = "PC1",
  ylab = "PC2"
)
points(proy[dat$sobreviv == 'T', ],
       cex = 1.5,
       pch = 21,
       bg = 'green')
points(proy[dat$sobreviv == 'F', ],
       cex = 1.5,
       pch = 21,
       bg = 'black')
text(proy[, 1],
     proy[, 2],
     as.character(dat$sex),
     pos = 1,
     cex = 1)
```

d) ¿Puedes distinguir evidencias gráficas de que los gorriones sobreviviente tenían características distintas a las de los que perecieron?

e) Haz un gráfico similar pero ahora para los gorriones hembra y macho, usaando otros colores. ¿Hay diferencias sexuales en estas variables morfométricas?

f) Con base en esta última gráfica, y tu respuesta en el inciso b del numeral 5, responde ¿cuál de los sexos tiene un mayor SKL?

g) Es conocimiento común que en un PCA sobre variables morfológicas, el PC1 representa cambios en el tamaño general de los individuos. Con esto en mente, ¿podrías explorar la posibilidad de que los gorriones juveniles posean tamaños menores en lo general que los adultos? ¿Donde deben estar localizados los puntos correspondientes a los gorriones juveniles en la ordenación?

8. Es frecuente representar los vectores de las variables junto con el mapa de ordenación en las mismas escalas, es decir, un biplot. Aplica la función `biplot` de correlación (o sea, la escala para elevar lambda es 0) de acuerdo con el código siguiente:

```
biplot(
  princomp(Y, cor = T),
  choices = 1:2,
```

```
  scale = 0,
  var.axes = T,
  arrow.len = 0.1,
  col = c("black", "red"),
  cex = 0.7,
  asp = 1,
  main = "Biplot (alfa=0)",
  xlab = "PC1",
  ylab = "PC2"
)
```

a) Interpreta el resultado siguiendo las instrucciones vistas en clase. ¿Cuáles son las variables más correlacionadas entre si? ¿Coincide tu respuesta con lo que concluiste de la exploración de los datos inicial? ¿En qué variables difieren mas el gorrión 103 y 110? ¿En qué variables difieren mas los gorriones 129 y 91?

b) Obtén el biplot de distancia (scale=1) y compara con el anterior. ¿Cuáles gorriones son más parecidos entre si: el 24 y 65, o el 85 y el 1?¿¿En qué variables difieren uno y otro par?

9. Con el siguiente código, elabora un screeplot para ayudarte a decidir cuántos componentes principales bastan para explicar la variación contenida en los datos de Bumpus. Después de ver la gráfica, ¿continuas con la misma respuesta que diste en el inciso c del numeral 5?

```
screeplot(pca.cor, type = "lines", main = "Scree-plot")
L.cor <- eigen(R)$values
scree.porcent <- L.cor / sum(L.cor) * 100
plot(
  scree.porcent ~ c(1:6),
  type = "l",
  main = "Scree-plot",
  xlab = "PC",
  ylab = "Variación (%)"
)
```

10. Otras visualizaciones: gráfico en 3D y con centroides.

```
install.packages("rgl")
library(rgl)
plot3d(
  proy[, 1],
  proy[, 2],
  proy[, 3],
  xlab = "PC1",
  ylab = "PC2",
  zlab = "PC3",
  asp = 1,
  col = "red"
)


install.packages("pca3d")
library(pca3d)
pca.cor <- prcomp(Y, scale. = TRUE)
pca2d(pca.cor,
      group = dat$sex,
      fancy = T,
      biplot = T)
```

```
pca3d(pca.cor,
      group = dat$sex,
      fancy = T,
      biplot = T)
```

11. Explore el paquete `ggfortify` y genere ordenaciones basados en PCA visualmente amenos.