UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Análisis de componentes principales

Francesc Carmona

5 de abril de 2018

1. Sea $\mathbf{x}$ un vector aleatorio que sigue una distribución normal bivariante de media cero y matriz de covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} 8 & 5 \\ 5 & 4 \end{pmatrix}$$

   a) Escribir la función de densidad $f(x_1, x_2)$ del vector $\mathbf{x}$ y representarla en tres dimensiones[1].

   b) Realizar un análisis de componentes principales de $\mathbf{x}$.

   c) Dibujar un gráfico de curvas de nivel de la función de densidad en el cuadrado $[-6, 6] \times [-6, 6]$ con la función `contour(x,y,z)` de **R**. Añadir a este gráfico los vectores de las componentes principales con la función `arrows()` y explicar el resultado.

   ($*$) Una alternativa gráfica menos exacta para resolver este ejercicio sería simular unos datos aleatorios y utilizar una estimación de la densidad con el siguiente código:

```
# lets first simulate a bivariate normal sample
library(MASS)
bivn <- mvrnorm(1000, mu = c(0, 0), Sigma = matrix(c(8, 5, 5, 4), 2))

# now we do a kernel density estimate
bivn.kde <- kde2d(bivn[,1], bivn[,2], n = 50)

# perspective
persp(bivn.kde, xlab="x", ylab="y", zlab="z", phi = 20, theta = 20)

# fancy contour with image
image(bivn.kde); contour(bivn.kde, add = T)
```

2. Sea la matriz de varianzas-covarianzas poblacionales

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

   correspondiente a un vector aleatorio $\mathbf{x} = (X_1, X_2, X_3)'$ de media cero.

   a) Calcular los valores y vectores propios de $\mathbf{\Sigma}$.

   b) Escribir el vector $\mathbf{y} = (Y_1, Y_2, Y_3)'$ de componentes principales e indicar la proporción de la varianza total que explica cada componente.

   c) Representar la observación $\mathbf{x} = (2, 2, 1)'$ en el plano que definen las dos primeras componentes principales.

[1] El código de la página 2 del documento `http://www.ejwagenmakers.com/misc/Plotting_3d_in_R.pdf` puede ser útil. No confundir la correlación entre dos variables con su covarianza. ¡Alerta! Hay una errata en el signo del `term5`.

3. Con los datos de las $p = 5$ medidas biométricas de los $n = 49$ gorriones hembras que fueron recogidos casi moribundos después de un temporal,

   a) Realizar un análisis de componentes principales[2] y calcular la proporción de varianza explicada por las tres primeras componentes.

   b) Representar los datos sobre el plano de las dos primeras componentes, con dos símbolos distintos en función de si el gorrión sobrevivió o no, e interpretar los ejes como factores de *tamaño* (el primero) y *forma* (el segundo).

   A la vista del gráfico, ¿qué gorriones, según tamaño y forma, tienen menos posibilidades de sobrevivir?

4. Con los datos de los cangrejos de la base de datos crabs del paquete MASS de **R**,

   a) Realizar un análisis de componentes principales de las cinco medidas biométricas y estudiar la bondad de la representación en dos y en tres dimensiones.

   b) Representar los datos con las dos primeras componentes con dos colores, según la especie, y dos símbolos, segun el sexo. Indicar, si existe, alguna interpretación.

   c) Representar los datos con las tres primeras componentes.

   d) El Análisis de componentes principales puede servir para estudiar la capacidad. Supongamos que el caparazón del cangrejo tiene lognitud L, ancho A y alto H. La capacidad sería $C = L^\alpha A^\beta H^\gamma$, donde α, β y γ son parámetros. Aplicando logaritmos, obtenemos

   $$\log C = \log(L^\alpha A^\beta H^\gamma) = \alpha \log L + \beta \log A + \gamma \log H$$

   que podemos interpretar como la primera componente principal de las variables transformadas.

   Obtener las capacidades del caparazón de los cangrejos y un gráfico comparativo según especie y sexo.

5. Florence Nightingale (1820-1910) fué una enfermera británica pionera en la aplicación de la estadística a la epidemiología. Entre otras acciones, ella recogió los datos de soldados muertos por diversas causas en la guerra de Crimea y convenció con sus argumentos estadísticos a las autoridades militares para modificar el sistema hospitalario hacia un modelo con más higiene.

   La tabla de los datos en los 24 meses observados puede obtenerse en la página

   `http://understandinguncertainty.org/node/214`

   donde los 12 primeros son antes de aplicar sus nuevos métodos de cuidado en los hospitales militares.

   A partir de las frecuencias de muertes por tres causas: *Zymotic diseases*, *Wounds & injuries* y *All other causes*, Nightingale calculó unos índices de mortalidad anual relativa por cada 1000. Son precisamente esos índices los que vamos a utilizar como variables en este ejercicio.

   a) Observar e interpretar las representaciones descriptivas de la página

   `http://understandinguncertainty.org/node/213`

   en especial el llamado diagrama de *Coxcomb* o rosa de Nightingale.

   b) Realizar un análisis de componentes principales sobre los índices de mortalidad, representar los 24 meses con números correlativos e interpretar el resultado.

6. Con los moluscos gasterópodos conocidos como orejas de mar o abulones[3], muy apreciados por su carne, se estima la edad contando el número de anillos y es un trabajo muy laborioso. Por ello se busca la forma de predecir la edad de un abulón utilizando otras medidas más sencillas de obtener.

   Con el siguiente código se pueden obtener los datos de un conjunto de abulones pertenecientes a un estudio realizado por el Departamento de Industria y Pesca de Tasmania (Australia) en 1994:

[2] En **R** se utiliza la función `prcomp()`, aunque también existe `princomp()` semejante a la función de S-PLUS. Las funciones `print`, `plot` y `biplot` se aplican sobre un objeto de la clase "prcomp".

[3] Ver la página web `http://es.wikipedia.org/wiki/Haliotis`

```
archivo <-
"http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data"

abalone <- read.csv(archivo, header=F)

names(abalone) <-
       c("Sex",             # Sexo: M, F, and I (infant)
         "Length",          # Longitud: mayor medida de la concha (mm)
         "Diameter",        # Diámetro: perpendicular a la longitud (mm)
         "Height",          # Altura: con carne dentro de la concha (mm)
         "Whole weight",    # Peso total: todo el abulón (g)
         "Shucked weight",  # Peso desconchado: peso de la carne (g)
         "Viscera weight",  # Peso de las vísceras: peso de la tripa
                            # (después de sangrar) (g)
         "Shell weight",    # Peso de la concha: después de ser secado (g)
         "Rings")           # Anillos: +1.5 es la edad en años
```

a) Seleccionar un conjunto de 30 abulones hembra al azar[4] y realizar con ellos un análisis de componentes principales con todas las variables numéricas excepto la última (los anillos).

Representar estos abulones con puntos en un diagrama de dispersión formado por los dos primeros ejes principales.

b) Para los primeros 151 abulones, obtener la matriz de correlaciones de las variables numéricas sin la última (los anillos).

¿Qué problema tendremos si pretendemos predecir la edad del abulón con estas variables?

c) Para paliar el problema anterior, realizar una regresión sobre la variable `Rings` con las dos primeras componentes principales obtenidas con la matriz de covarianzas.

Estudiar la bondad de este modelo, frente al modelo de regresión con las variables originales.

7. Para la base de datos `usair` del libro de Everitt(2005)[5]

a) Cargar los datos con la instrucción:

```
usair <- source("/(path)/chap3usair.dat")$value
```

y estudiar la base de datos con las funciones `str()` y `summary()`.

Realizar un gráfico descriptivo de los datos.

b) Representar las ciudades en el plano de las dos primeras componentes principales calculadas con la matriz de correlaciones de todas las variables excepto el `SO2`.

Estudiar las posibles ciudades atípicas (outliers).

c) Utilizar una matriz de correlaciones robusta[6] y calcular de nuevo las componentes principales.

Comparar los resultados con los del apartado anterior.

d) (∗) Investigar la utilización de la regresión múltiple con las variables del clima y poblacionales para predecir la polución en dióxido de sulfuro, teniendo en cuenta el problema de alta correlación entre algunas variables predictoras.

[4] Se puede utilizar la función `sample()` de **R**.

[5] Ver la página web `http://biostatistics.iop.kcl.ac.uk/publications/everitt/`.

[6] Como coeficiente de correlación robusto se puede calcular la rho de Spearman o la tau de Kendall. (∗) También se puede utilizar una estimación robusta de la matriz de correlaciones por el método del elipsoide de mínimo volumen.