ANÁLISIS DE DATOS MULTIVARIANTES.

Curso 2017/18

Departamento de Estadística e Investigación Operativa

GRADO EN MATEMÁTICAS – DOBLE GRADO EN MATEMÁTICAS Y ESTADÍSTICA

LAB. 1 – TEMA 4. Análisis de Componentes Principales

El conjunto de datos "decathlon1989.sav" (formato SPSS) contiene los resultados obtenidos en las pruebas de decatlón ("t100m", "longitud", "peso", "altura", "t400m", "t110val", "disco", "pertiga", "jabalina" y "t1500m") por 34 atletas, además de un código identificativo del atleta ("iden") y la nacionalidad del atleta ("nompais")

1. Análisis descriptivo (en particular, de las correlaciones bivariadas)
2. Análisis de Componentes principales (obtención de componentes, variabilidad explicada y gráfico de sedimentación)
3. Contraste de hipótesis para seleccionar el número de componentes principales (bajo hipótesis de normalidad multivariante)

   $H_0 \ : \ \lambda_{m+1} = \cdots = \lambda_p$

   $$Q = \left(n - \frac{(2+p+11)}{6} \right) \left[(p-m)\, log \left[\frac{1}{p-m} \sum_{k=m+1}^{p} \hat{\lambda}_k \right] - \sum_{k=m+1}^{p} log(\hat{\lambda}_k) \right]$$

   Bajo la hipótesis nula $Q \sim \chi^2_f$ con $f = (p-m+2)(p-m+1)/2$

4. Coeficientes de las componentes principales y correlaciones con las variables originales. Interpretación de las componentes principales.
5. Representación gráfica de los resultados.

Nota. Análisis de componentes principales en R.

Orden: princomp(formula, data = NULL, subset, ...)
princomp(x, cor = FALSE, scores = TRUE, covmat = NULL, subset = rep(TRUE, nrow(as.matrix(x))), ...)

ARGUMENTOS:

formula: formula sin variable respuesta, sólo con variables numéricas. Por ej.: ~ varX1 + varX2 + varX3.
data : un marco de datos opcional que contenga las variables de la fórmula
subset : un vector opcional para seleccionar las filas (observaciones) de la matriz de datos
x : matriz o marco de datos que proporciona los datos que proporciona los datos para el ACP
cor : Valor lógico (T/F) indicando si se usa la matriz de correlación (T) o la matriz de covarianzas (F).
scores : Valor lógico (T/F) indicando si las puntuaciones de cada c.p. deben ser calculadas

RESULTADOS: Crea un objeto "princomp" que recoge la siguiente información
sdev : desviaciones estándar de las comp.principales.
loadings : matriz de cargas (es decir, matriz de autovectores)
center : las medias
scaling : la escala aplicada a cada variable
n.obs : número de observaciones
scores : Si se ha solicitado, las puntuaciones de los datos en las c.p.

LAB. 2 – TEMA 4. Análisis de Componentes Principales

La matriz R dada por:

```
R=matrix(    c(1.0   , 0.584 , 0.615 , 0.601 , 0.570 , 0.600,
               0.584 , 1.0   , 0.576 , 0.530 , 0.526 , 0.555,
               0.615 , 0.576 , 1.0   , 0.940 , 0.875 , 0.878,
               0.601 , 0.530 , 0.940 , 1.0   , 0.877 , 0.886,
               0.570 , 0.526 , 0.875 , 0.877 , 1.0   , 0.924,
               0.600 , 0.555 , 0.878 , 0.886 , 0.924 , 1.0  ),6,6)
```

es la matriz de correlación muestral asociada a p=6 medidas de huesos de una muestra de n=400 aves:

"longcra", "anchcra", "humero", "cubito", "femur", "tibia"

1. El test de Esfericidad de Barlett que contrasta, para una muestra extraída de una población $N_p(\mu, \Sigma)$, la hipótesis
$$H_0 \ : \ \Sigma = \sigma^2 \, \boldsymbol{I}_p \qquad vs. \qquad H_1 \ : \ \Sigma \neq \sigma^2 \, \boldsymbol{I}_p$$
Siendo el estadístico de contraste:
$$G = -2 \log \frac{[\det(\boldsymbol{S})]^{n/2}}{[traza(\frac{1}{p}\,\boldsymbol{S})]^{np/2}}$$
siendo $\boldsymbol{S}$ la matriz suma de cuadrados y productos cruzados asociada a la muestra de tamaño n.
Bajo la hipótesis nula el estadístico G se distribuye según una ley χ^2_g con $g = \left[\frac{p(p+1)}{2}\right] - 1$
Un test más general es el test de independencia, basado en la matriz $\boldsymbol{R}_\rho$ de Correlación poblacional
$$H_0 \ : \ \boldsymbol{R}_\rho = \boldsymbol{I}_p \qquad vs. \qquad H_1 \ : \ \boldsymbol{R}_\rho \neq \boldsymbol{I}_p$$
Siendo el estadístico de contraste:
$$\gamma = -\left[n - 1 - \frac{(2p+5)}{6}\right] \log[\det(\boldsymbol{R})]$$
siendo $\boldsymbol{R}$ la matriz de correlación muestral. Bajo la hipótesis nula el estadístico se distribuye según una ley χ^2_f con $f = \left[\frac{p(p-1)}{2}\right]$. Si se aceptara la independencia, no tendría sentido al ACP.
2. Análisis de Componentes principales (obtención de componentes, variabilidad explicada y gráfico de sedimentación), selección del número de componentes, resumen gráfico.
3. Interpretación de los resultados, aplicando la rotación ortogonal *varimax*.