PUCV - Instituto de Estadística

Analisis por Componentes Principales

Ejercicio 1: Recordatorio del producto vectorial:
Sea $x = (x_1, \ldots, x_I)'$ y $y = (y_1, \ldots, y_I)'$, dos vectores de R^I, entonces:

$$< x, y >= \sum_{i=1}^{I} x_i y_i = \|x\|.\|y\| \cos(\theta),$$

donde $\|.\|$ es la norma Euclidiana y θ es el angulo entre los dos vectores. La proyección ortogonal de y sobre x es dada por

$$\hat{y} = \frac{< x, y >}{\|x\|^2} x,$$

Para poner nos en una situación cercana a la de la ACP, vamos a suponer que $\|x\| = \|y\| = 1$. Responder a las siguientes preguntas:

1- ¿Si $\theta = 0$ a qué es igual $< x, y >$?

2- ¿Si $\theta \simeq 0$ qué podemos decir de $\|\hat{y}\|$?

3- ¿Si $\theta = \pi/2$ a qué es igual $< x, y >$?

4- ¿Si $\theta \simeq 0$ qué podemos decir de $\|\hat{y}\|$?

5- ¿Si x e y corresponden a observaciones centradas y normalizadas de observaciones aleatorias, a qué corresponde $\cos(\theta)$?

6- ¿Sea la recta D definida por el vector unitario x. Mostrar que $\hat{y}$ es la solución del problema

$$\underset{z \in D}{\operatorname{argmin}} \|y - z\|.$$

Deducir que $\hat{y}$ es la representación la más fiel de y en D.

Ejercicio 2:Recordatorio muy breve de la diagonalización de una matriz definida positiva:

a- Si existe un escalar λ y un vector v tales que por una matriz M cuadrada $K \times K$ tenemos que $Mv = \lambda v$, entonces se dice que λ es un valor propio y v un vector propio.

b- Los valores propios se obtienen resolviendo $\det(M - \lambda I_K) = 0$. Los vectores propios cumplen con $Mv_i - \lambda_i v_i = 0$, $i = 1, \ldots, K$.

c- Sea una matriz definida positiva Σ. Se puede mostrar que

$$\Sigma = PDP',$$

donde P contiene los vectores propios de Σ que son ortogonales y tales que $\|P_j\| = 1$, donde P_j, el vector propio j, es una columna de P. Se dice que P es una matriz ortogonal. la matriz $D = diag(\lambda_1, \ldots, \lambda_K)$, $\lambda_1 > \cdots > \lambda_K$ contiene los valores propios de la matriz Σ asociados a los vectores propios P_i.

Sea un vector $X \sim \mathcal{N}(0, \Sigma)$. Es facil de ver que $P'X \sim \mathcal{N}(0, D)$. Queremos buscar ejes definidos por vectores unitarios a (o sea $\|a\| = 1$) tales $\mathrm{Var}(a'X)$ sea maximizada.

1- Mostrar que $\Sigma = PDP' = \sum_{j=1}^{K} \lambda_j P_j P_j'$.

2- Mostrar que $\mathrm{Var}(a'X) = \sum_{j=1}^{K} \lambda_j c_j^2$, donde $c_j = a'P_j$.

En este punto podemos anotart que $|c_j|$ es el módulo de la proyección del vector a sobre P_j. En formula matematica si anotamos $\hat{a}_j$ la proyección del vector sobre P_j, tenemos:

$$\hat{a}_j = \frac{< a, P_j >}{\|P_j\|^2} P_j =< a, P_j > P_j = (a'P_j)P_j = c_j P_j.$$

Dado que los P_j son una base ortonormal, $1 = \|a\| = \sum_{i=1}^{K} c_j^2$.

3- Utilizando los comentarios anteriores mostrar que $\mathrm{Var}(a'X) \leq \lambda_1$.

4- Utilizando la primera pregunta, mostrar que $\mathrm{Var}(P_1'X) = \lambda_1$. Concluir sobre el primer eje que maximiza $\mathrm{Var}(a'X)$.

Ahora vamos con la segunda dirección que maximiza $\mathrm{Var}(a'X)$, tal que a y P_1 sean ortogonales.

5- Mostrar que $\mathrm{Var}(a'X) = \sum_{j=2}^{K} \lambda_j c_j^2$, donde $c_j = a'P_j$ en este caso.

6- Haciendo un calculo similar a las preguntas 3 y 4, deducir que $\text{Var}(a'X) < \lambda_2$. Concluir sobre el segundo eje que maximiza $\text{Var}(a'X)$.

Nota: Podemos seguir así con $P_3, \ldots, P_K, \ldots$

7- Mostrar que $\sum_{j=1}^{K} \lambda_j = K$ si las varianzas del vector aleatorio X son iguales a 1 (caso de una ACP estandardizada).

Ejercicio 3: ACP de datos de autos de los años 80.

En esta parte nos vamos utilizar el paquete "FactoMineR" (existe otros paquetes como por ejemplo "ade4"). Este paquete tiene la posibilidad de hacer un ACP con una interfaz de manera facil.

Por eso: install.packages("RcmdrPlugin.FactoMineR",dep=TRUE)
library(Rcmdr)
Después ir a herramientas y luego cargar "plugins" de Rcmdr.

1- Importar los datos. Agregar la opción `row.names="NOMBRE"` para que los individuos sean identificados (en la parte Rscript). Calcular la matriz de correlaciones para identificar las relaciones lineales entre variables (Estadísticos>Resumenes>matriz de correlaciones). ¿Qué podemos ver?

2- Ver como se presenta el porcentaje de información resumida por los ACP. Comentar las diferentes salidas. ¿Cuanto dimensiones pueden ser suficientes para describir la variabilidad de los datos sin demasiado perdida?

3- Hacer el ACP con todas las variables (FactoMineR>PCA). Comentar la nube de los individuos proyectados a la luz de las variables proyectadas.

4- Comentar la disposición de las variables.

5- Por los 10 primeros individuos, comparar las distancias al centro de la nube y sus contribuciones (en %) a la variabilidad de cada dimensión. ¿Podríamos decir que "RENAULT-30-TS" es un dato extremo que puede tener mucha influencia en fijar Dim1? Hacer el ACP con este

individuo en "supplementary individual", y comparar las salidas. Comentar la salidas. Se puede ver más o menos variables con la opción `"nbelements=x"` o `"nbelements=Inf"` para ver todos las variables. Para los individuos ver la opción `"nbind = 18"`, para ver solo 2 componentes principales `"ncp = 2"`.

7- Comparar la representación de los "LANCIA-BETA-1300" y "CITROEN-GS-CLUB" en las 2 primeras dimensiones con los $\cos^2$. Cual es bien representado por el ACP y cual no lo es?

8- Dar un ejemplo de variable representada a más de 90% y una otra a 70% approximadamente. Interpretar.

9- Incorporar la variable cualitativa "estandar". Comentar el centro de los punto "P", "B" y "MB" al respeto del nube de punto proyectado y a la luz de las variables proyetadas.

10- El v.test corresponde al test que una cierta modalidad de la variable "Lujo" toma valores de manera independiente a una cierta dimensión ("en el mismo sentido–>valor positivo, sentido contrario–>valor negativo). Comentar las salidas del v.test al respeto de la dimensiones 1 y 2. (Un valor mas grande que 2 en valor absoluto sugiere un efecto significativo)

10- Nota que los puntos "B", "MB" y "P" corresponden a los centros de los puntos que son "B", "MB" y "P". Comentar las salidas $\cos^2$.

11- Sacar la variable "precio" de las variables activas. Hacer el ACP y representar la variable precio. ¿La variable precio es bien representada con las otras variables según el $\cos^2$?