Pràc 3: Components principals sense usar cap funció específica. - 1a. part-

Crea un document `Cognom1Cognom2Nom_pr3.html` que contingui: els enunciats, el codi R, els resultats (taules, gràfiques, etc.) i les interpretacions que es demanin. El document es lliurarà quan acabi la pràctica 3.

Utilitza les dades `decathlon` del paquet `FactoMineR`.
Utilitza les dades `decathlon` del paquet `FactoMineR`.

```
library(fBasics); library(FactoMineR)
data(decathlon, package="FactoMineR")
deca<-decathlon
deca$Competition<-as.factor(deca$Competition)
```

1. Substitueix, en el mateix fitxer `deca`, les variables que són t = temps, de curses, amb la transformació següent: $t \longrightarrow \max(t) - t$. D'aquesta manera, totes les variables tindran el mateix sentin: quan més grans siguin els valors, millor el resultat.

2. Crea els quatre (sub)fitxers de dades (tipus dataframe) següents:

   - El fitxer que denotarem `basic`: conté els casos 1 a 38 (anomenats casos *actius*) i les variables 1 a 10 (dim 38×10). Aquest fitxer bàsic s'utilitzarà per determinar les components principals.
   - El fitxer de variables numèriques suplementàries, que denotarem `n.sup`: conté les columnes 11 i 12 pels casos actius (dim 38×2).
   - El fitxer d'individus suplementaris, que anomenarem `i.sup`: casos 39, 40 i 41 i variables 1 a 10 (dim 3×10). Jugaran el rol de "individus nous" i no determinaran les components, però sí que es projectaran sobre l'espai de les components principals creat a partir dels casos bàsics.
   - Un fitxer compost d'una única variable qualitativa suplementaria pels casos actius, que anomenarem `q.sup`: variable 13 i casos 1 a 38 (dim 38×1). Assegura't que aquesta variable sigui un `factor`. En aquest cas, és un factor amb dos grups o nivells, però podria ser un factor amb k grups.

3. **Tipificació de les dades.** En aquest pas tipificarem els diversos fitxers creats a l'apartat anterior. Fem la tipificació completa (centrar i escalar) i no només el centrament perquè les variables no són homogènies (diferents unitats, rangs, etc.). *Nota:* La funció `scale()` permet tipificar (i també centrar) respecte dels descriptius de les propies dades o respecte d'altres descriptius donats (`(?scale)`).

   Atenció!: la tipificació es fa diferent segons si s'aplica a les dades del fitxer `basic` o al altres fitxers. Fixeu-vos-hi!

   (a) Per què diem que tipificar equival a determinar les components principals de la matriu de correlacions?

   (b) Calcula n com el nombre de files del fitxer `basic` i guarda'l. Aquest és el mateix n en tots els apartats. Calcula p com el nombre de columnes del fitxer `basic` i guarda'l.

   (c) Calcula i guarda els vectors de mitjanes i de desviacions típiques del fitxer de dades bàsiques. Es recomana la funció `apply()`.

   (d) Pel fitxer de dades bàsiques, tipifica les dades respecte dels seus propis descriptius. Anomena `X` el fitxer resultant de tipificar i multiplica tot el fitxer per $1/\sqrt{n-1}$ mantenint el mateix nom `X`.

   (e) Els casos suplementaris es re-escalen utilitzant els descriptius del fitxer `basic`. Anomena `Xisup` el fitxer resultant de tipificar i multiplicar després per $1/\sqrt{n-1}$.

   (f) Les variables suplementàries `n.sup` es tracten igual que a l'apartat (d), utilitzant els seus propis descriptius. Anomena `Xnsup` el fitxer resultant de tipificar i multiplicar després per $1/\sqrt{n-1}$.

   (g) La variable qualitativa es tractarà com segueix: considerant-la com dos "nous casos suplementaris" representats per les mitjanes de les variables bàsiques en els dos grups, és a dir, representant cada grup com un vector amb p valors mitjans. Anomena `Xqsup` el fitxer resultant de tipificar les mitjanes dels dos grups amb els descritius del fitxer basic i multiplicar després per $1/\sqrt{n-1}$. Es recomanen les funcions `split()`, `lapply()` i `scale()`. Es mostra el codi:

```
## Atenció! les variables qualitat. es tracten com casos suplement. representats per les mitjanes !!
# "split" crea unobjecte de classe llista
llis<- split(data.frame(basic),q.sup) ## una llista de 2 objectes; les matrius de dades bàsiques dels 2 grups
q.sup.m <- lapply(llis,colMeans)  # mitjanes de q.sup. O també: lapply(llis,apply,2,mean)
q.sup.m <- data.frame(q.sup.m)
q.sup.m <- t(q.sup.m)             # transposem per tractar els grups com individus
# tipificar amb els descriptius del fitxer basic:
Xqsup <- scale(q.sup.m, center= ???, scale= ???)
```

4. Per a l'anàlisi de components principals utilitzarem les matrius U, V i D (i $\Lambda = D^2$) de la descomposició en valors singulars (**SVD**) de `X`.

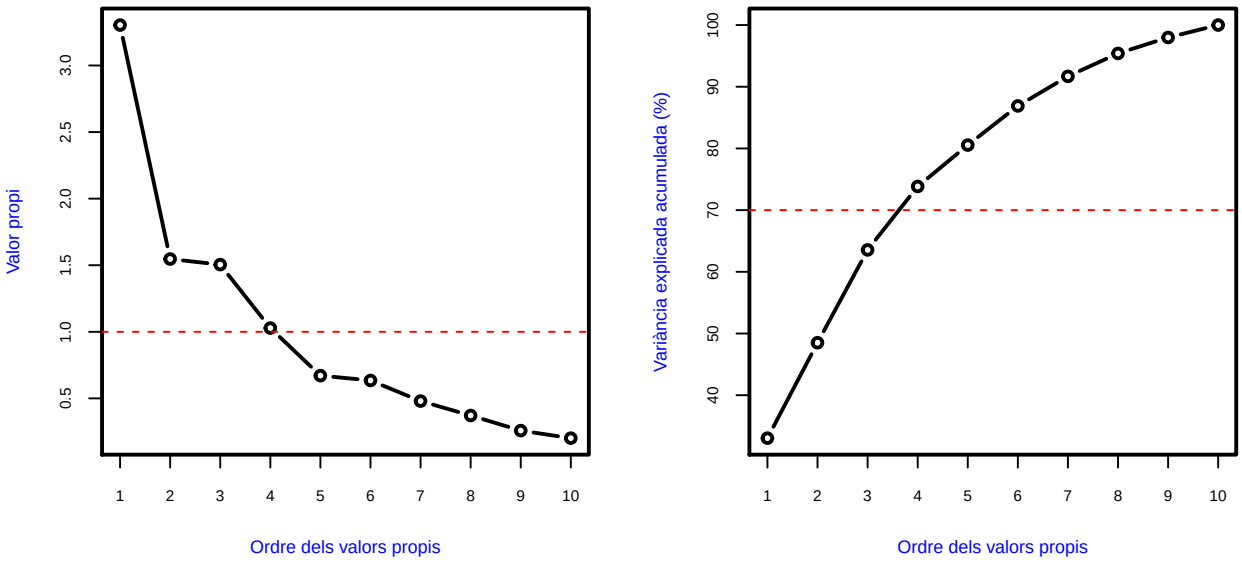
   (a) Recordes la relació entre U, V i D i les matrius de les descomposicions espectrals de X^tX i XX^t ?

   (b) Què és X^tX tal i com has calculat X?

   (c) Calcula i guarda els objectes següents:

```
sing<- .........          # objecte resultant de la svd de X
U<- ............
V<- ............
D<- ............          # classe matriu
Lambda<- .......          # classe matriu
lambda<- .......          # vector diagonal de l'anterior
pvac<- .........          # percentatge de variància acumulada pels components
```

5. Fes una gràfica on a l'eix d'abscises s'hi representin els nombres de 1 a p i a l'eix d'ordendes, els valors propis ordenats de més gran a més petit. Uneix els valors amb una línia. Dibuixa-hi la recta horitzontal $y = 1$ en vermell i menys gruixuda. Idem, representa la variància explicada acumulada i la recta horitzontal al 70%.



Quantes components recomanaria cadascun dels tres criteris (vaps superiors a la mitjana, variància igual o superior al 70%, gràfica de sedimentació) ?

La pràctica 3 continuarà .....