Lliurament: **Divendres 13/03/2020**.

# 1 Introducció

A la primera part d'aquesta pràctica aplicarem l'anàlisi de components principals (ACP) i aprendrem a interpretar els seus resultats. A la segona part aplicarem escalament multdimensional (MDS).
Cada grup d'estudiants ha de realitzar els exercicis marcats pel professor de la llista a continuació, realitzant els càlculs i gràfics necessaris en l'entorn R. Els resultats obtinguts i les respostes a les preguntes s'han de recollir en un document amb la vostre solució, fet per exemple en R markdown, Microsoft Word o Latex. Procureu posar els vostres noms i cognoms i número de grup a l'inici del document. Empleneu la mateixa numeració dels items de l'enunciat al document amb la vostre solució. S'ha de **lliurar la solució del qüestionari en format .pdf**, pujant-la a l'entorn Atenea (atenea.upc.edu), a l'apartat de la tasca corresponent, no mes tard que la data límit de lliurament.

# 2 Exercicis

1. (15p) **ACP dels resultats de la competició del FIS**.

   Les dades que analitzarem dades són els resultats de la $7^a$ competició FIS Cup, un esdeveniment de salt d'esquí que es fa a Einsiedeln (Suïssa) anualment. Les dades són extruïdes de la web oficial de la FIS (http://www.fis-ski.com/), disponibles dins el fitxer `Einsiedeln.dat`. El fitxer conté les variables definides a continuació:

   - *Surname, Name, Nationality*: Nom, cognom i nacionalitat de l'atleta.
   - *Day, Month, Year*: Data de naixement de l'atleta.
   - *Bib*: Variable desconeguda.
   - *Speed*: Velocitat de l'atleta durant el salt.
   - *Dist*: Distància realitzada amb el salt.
   - *DP*: Una variable índex (resum) pel salt.
   - *A, B, C, D, E*: Puntuacions de 5 jutges pel salt realitzat.
   - *JP*: Variable índex de les puntuacions dels jutges.
   - *Total*: Puntuació final de l'atleta dins la classificació.

   (a) (1p) Carregueu les dades dins l'entorn de R amb la instrucció *read.table*. Calculeu l'edat (*Age*) dels atletes en anys amb *Age = 2019 - Year of birth*. Cal crear una nova matriu (o dataframe) contenint nomes les 10 variables a continuació que utilitzarem per un anàlisi de components principals: *Age, Speed, Dist, DP, A, B, C, D, E* i *JP*. Convé etiquetar les files de la nova matriu amb els noms dels atletes, i etiquetar les columnes de la matriu amb els noms de les variables. La variable *Total* es guarda per separat i no s'utilitza com a variable d'entrada en el ACP. Quantes observacions conté la base de dades? Feu boxplots de les variables amb la instrucció *boxplot*. Feu una matriu de diagrames bivariants (scatterplot matrix) de les variables seleccionades. Comenteu els vostres resultats. Calculeu la matriu de correlacions.

(b) (2p) Fem un ACP utilitzant la matriu de correlacions. Calculeu la taula amb la descomposició de la variabilitat, posant les variàncies dels components i els percentatges de variància explicada i els percentatges acumulats. Pots utilitzar la funció `princomp` per tal propòsit. Quants components creeu que són necessaris per descriure aquestes dades? Feu un scree-plot dels valors propis com ajut per prendre aquesta decisió.

(c) (1p) Podeu explicar perquè apareixen valors propis que són zeros a l'anàlisi?

(d) (2p) Feu un biplot corresponent a l'anàlisi anterior, utilitzant la funció `biplot`. Interpreteu els primers dos components. Comenteu possibles anomalies. Si existeixen anomalies, cal identificar-les, i investigar en quin sentit són inusuals.

(e) (1p) Repetiu l'anàlisi de components principals sense la anomalia mes extrema. Repetiu la descomposició de la variabilitat i refeu el biplot. A partir d'aquí, feu tots els anàlisis sense la anomalia.

(f) (1p) Extrèieu els components principals calculats per `princomp` accedint a l'objecte "scores" dins la llista produïda per `princomp`. Calculeu les variàncies dels components extrets. Compareu aquestes variàncies amb la taula amb la descomposició que heu fet abans. Comenteu els resultats.

(g) (1p) Calculeu la matriu de correlacions entre els components principals. Comenteu els resultats.

(h) (1p) Fem també l'anàlisi de components principals "a mà". Calculeu els valors i vectors propis de la matriu de correlacions fent servir la funció `eigen`. Calculeu els components principals mitjançant la postmultiplicació de les dades estandarditzades per la matriu de vectors propis. Compareu els resultats amb els que ha trobat al funció `princomp`. Són idèntics? Expliqueu possibles diferències. Cal posar les 10 primeres files de la matriu de components calculada a mà i les 10 primeres files dels components obtinguts amb la matriu obtinguda per `princomp` dins l'informe.

(i) (1p) Calculeu la matriu de correlacions entre les variables originals i els primers dos components principals obtinguts amb `princomp`. Quines variables tenen correlació alta amb quins components?

(j) (1p) Extrèieu l'objecte "loadings" dels resultats del `princomp`. Que representen els loadings?

(k) (1p) Feu un altre biplot, utilitzant els components estandarditzats. Pinteu un cercle unitari dins el biplot. Quines variables estan mal representades?

(l) (1p) Inspeccionant el biplot, quines característiques dels atletes aprecien els jutges sobre tot?

(m) (1p) Estudieu la relació entre la variable amb la puntuació final de l'esdeveniment (*Total*) i les dues primeres components principals. Comenteu els resultats.

2. (20p) **ACP dels llavors de blat**.

Considerem dades morfològiques de grans de blat. Les variables recollides en l'estudi són:

- *area*: àrea
- *perimeter*: perímetre
- *compactness*: compacitat

- *length*: longitud
- *width*: amplada
- *asymmetry*: coeficient d'asimetria
- *groove*: longitud ranura

Les dades es troben al fitxer `kernels.dat` (podeu descarregarles clicant en el nom del fitxer, o bé llegir el fitxer directament dins l'entorn R). Feu els càlculs i els gràfics que es demanen a continuació. Comuniqueu resultats numèrics amb una precisió de tres decimals.

(a) Llegiu el fitxer de dades amb la funció `read.table`.

(b) (1p) Feu un scatterplot matrix per explorar les relacions entre les variables. Calculeu la matriu de correlacions. Quina parella de variables sembla tenir la relació lineal mes forta entre elles?

(c) (2p) Fem un ACP utilitzant la matriu de correlacions. Calculeu la taula amb la descomposició de la variabilitat, posant les variàncies dels components i els percentatges de variància explicada i els percentatges acumulats. Pots utilitzar la funció `princomp` per tal propòsit. Quants components creeu que són necessaris per descriure aquestes dades? Feu un scree-plot dels valors propis com ajut per prendre aquesta decisió.

(d) (2p) Feu un biplot corresponent a l'anàlisi anterior, utilitzant la funció `biplot`. Interpreteu els primers dos components.

(e) (1p) Quina parella de variables té, segons aquest biplot, la correlació mes alta? Coincideix amb la parella identificada a l'apartat (b)?

(f) (1p) Extrèieu els components principals calculats per `princomp` accedint a l'objecte "scores" dins la llista produïda per `princomp`. Calculeu les variàncies dels components extrets. Compareu aquestes variàncies amb la taula amb la descomposició que heu fet abans. Comenteu els resultats.

(g) (1p) Extrèieu també l'objecte "loadings" dels resultats del `princomp`. Que representen els loadings?

(h) (2p) Grafiqueu els primers dos components principals extrets en un diagrama bivariant amb `plot`. Etiqueteu els punts amb el número de l'observació fent servir la funció `text`. Compareu la gráfica amb el biplot i comenteu possibles diferències.

(i) (1p) Calculeu la matriu de correlacions entre els components principals. Comenteu els resultats.

(j) (1p) Calculeu la matriu de dades estandaritzades, i feu la descomposició en valors singulars d'aquesta matriu. Doneu els valors singulars.

(k) (2p) Calculeu a partir dels resultats de la descomposició els components principals estandaritzades. Mostreu els valors per les primeres sis observacions amb `head`.

(l) (2p) Calculeu les coordenades per les variables (columnes) al biplot. Feu el biplot amb els components estandaritzades, i representeu les variables mitjançant fletxes, utilitzant la funció `arrows`. Traçeu un cercle unitari al biplot amb `draw.circle` del paquet `plotrix`.

(m) (1p) Quina variable té la pitjor qualitat de representació al biplot?

(n) (1p) Quina és la variància total en aquest análisis que estem fent amb la SVD?

(o) (1p) Quin tan percent d'aquesta variància queda explicada per tres components principals?

(p) (1p) Quina és la teva interpretació del tercer component principal?

3. (20p) **MDS de breakfast cereals**.

Les dades que fem servir en aquesta pràctica provenen d'un estudi de cereals. Les observacions corresponen a 43 cereals enregistrats al fitxer `Cereals.dat`. Per cada cereal, es disposa de les variables a continuació:

*Brand*: marca del cereal
*Manufacturer*: fabricant (G (General Mills), K (Kellogg) o Q (Quaker))
*Calories*: Quantitat de calories
*Protein*: Contingut en proteïnes
*Fat*: Contingut en grassa
*Sodium*: Contingut en sodi
*Fiber*: Contingut en fibra
*Carbohydrates*: Contingut d'hidrats de carbó
*Sugar*: Contingut en sucre
*Potassium*: Contingut en potassi

(a) Carregueu el fitxer Cereals.dat a l'entorn R.

(b) (1p) Calculeu la matriu de distàncies Euclidianes entre els cereals, fent servir nomes les variables *calories* i les 7 components dels cereals, i estandarditzant les variables abans del càlcul de la matriu de distàncies. Podeu utilitzar les funcions de R `scale` i `dist`. Procureu posar les distàncies entre els primers 5 objectes al vostre informe.

(c) (1p) Extrèieu les distàncies entre cereals per sota la diagonal de la matriu de distàncies, utilitzant la funció `lower.tri` i representeu-les en un histograma. Existeixen anomalies?

(d) (1p) Realitzeu un MDS mètric de les dades utilitzant la funció `cmdscale`. Feu una gràfica de la solució en dos dimensions, etiquetant els cereals amb els seu nom abreviat. Utilitzeu colors o símbols diferents per identificar els fabricants de cadascun dels cereals.

(e) (1p) Quina parella de cereals és, segons la solució en dues dimensions, la mes similar?

(f) (1p) Quina parella de cereals és, segons la solució en dues dimensions, la que és mes diferent?

(g) (2p) És possible obtenir una representació dels 43 cereals en $k$ dimensions que aproximi la matriu de distàncies originals sense error? Argumenteu la resposta. Si la representació exacta fos possible, quantes dimensions ens caldrien per aconseguir-la?

(h) (1p) Doneu els valors propis obtinguts pel mètode, i calculeu la bondat de l'ajust de la solució en dues dimensions.

(i) (1p) Existeixen valors propis que són zero? Podeu explicar perquè apareixen?

(j) (2p) Calculeu les distàncies ajustades segons la solució en dues dimensions. Feu una gràfica les distàncies ajustades versus les observades. Que observeu?

(k) (1p) Realitzeu un MDS no-mètric amb la funció `isoMDS`. Feu una gràfica de la solució en dues dimensions, etiquetant els cereals amb el nom (abreviat) de la marca i utilitzant diferents símbols o colors pels diferents fabricants. Quin és el valor del stress de la solució?

(l) (1p) Quina parella de cereals és, segons la solució en dues dimensions, la mes similar?

(m) (1p) Calculeu les distàncies ajustades segons el mapa no-mètric de la solució en dues dimensions. Feu una gràfica de les distàncies ajustades versus les distàncies observades. Valoreu la bondat de l'ajust.

(n) (2p) Calculeu el stress per les solucions amb 1, 2, 3, 4 i 5 dimensions. Feu un plot del stress en funció del número de dimensions. Quantes dimensions són necessàries per un bon ajust?

(o) (2p) Feu un scatterplot matrix amb les primeres dues dimensions de la solució mètrica i les dues primeres solucions de la solució no mètrica (utilitzeu $k = 2$). Calculeu la matriu de correlaciones entre les quatre variables i comenteu els resultats.

(p) (2p) Hem utilitzat la distància Euclidià amb variables estandarditzades. Hagués-sim obtingut els mateixos resultats en el MDS mètric transformant les distàncies prèviament prenent l'arrel quadrada de les distàncies? I en l'anàlisi no mètric? Argumenteu la resposta.

4. (10p) **MDS amb distàncies genètiques**.

S'han calculat les distàncies genètiques entre els 800 individus d'un país. Els individus són de tres raçes auto-declarades, Africa-Americà, Caucasià o bé Asiàtic. La matriu de distàncies conté les distàncies de 342, 361 i 97 individus d'aquestes raçes, en aquest ordre respectivament.

(a) Genereu la variable categòrica raça pels 800 individus utilitzant al funció `rep`. Llegiu la matriu de distàncies que es troba dins el fitxer de GeneticDist.txt a l'entorn R amb `read.table`.

(b) (1p) Realitzeu un MDS mètric de les dades utilitzant la funció `cmdscale`. Feu una gràfica de la solució en dos dimensions, utilitzant colors o símbols diferents per identificar la raça dels invidus.

(c) (1p) És possible obtenir una representació dels 800 individus en $k$ dimensions que aproximi la matriu de distàncies originals sense error? Argumenteu la resposta. Si fos el cas, per quin valor de $k$?

(d) (1p) Existeixen valors propis que són zero? Si n'hi han, podeu explicar perquè apareixen?

(e) (1p) Calculeu la bondat de la representació en dues dimensions, utilitzant l'ajust de prendre valor absolut dels valors propis.

(f) (1p) Feu un diagrama bivariant de la primera versus la tercera dimensió. Quina és la bondat d'ajust d'aquesta representació? Ajuda la tercera dimensión a distingir les raçes?

(g) (2p) Calculeu les distàncies ajustades segons la solució en dues dimensions. Feu una gràfica les distàncies ajustades versus les observades. Que observeu?

(h) (1p) Realitzeu un MDS no-mètric amb la funció `isoMDS`, utilitzant una configuració aleatòria inicial, feta amb les instruccions:

```
set.seed(1234)
init <- scale(matrix(runif(n*2),ncol=2),scale=FALSE)
```

Feu una gràfica de la solució en dues dimensions, utilitzant diferents símbols o colors per les diferentes raçes. Quin és el valor del stress de la solució? Es distingueixen els tres grups?

(i) (1p) Repetiu l'anàlisi no-metric, cridant isoMDS sense cap configuració inicial. Que observeu?

(j) (1p) Feu 100 execucions del isoMDS, cada vegada amb una configuració inicial aleatòria diferent, i guardeu el stress obtingut en cada execució. Heu trobat alguna configuració amb un stress menor que trobat a (i)? Si és el cas, grafiqueu la solució corresponent i doneu el seu stress.