Anàlisi de Dades **Examen parcial en línia**
Grau en Ciencia i Enginyeria de Dades **Dilluns 20 d'abril, 2020, 9.00-11.00**

Nom i cognoms: .....

Question:	1	2	3	4	Total
Points:	10	10	10	10	40
Score:					

Indicacions:

- Es poden editar les respostes per l'examen en Word, Latex, R-Markdown, o altre programa informàtic. Assegura't que pots convertir la teva solució a fòrmat PDF. Cal **pujar la teva solució final en fòrmat PDF** dins la tasca corresponent a l'entorn Atenea (atenea.upc.edu).
- Es pot donar fòrmat a les fórmules amb Latex o Word, però no és imprescindible. Pots entrar simplement una fórmula com a text escrit, com por exemple X = U D V', o bé A v = lambda v.
- Procura posar **el teu nom i cognoms** a la primera pàgina del teu examen.
- Procura incloure la teva **declaració de bones pràctiques** a la primera pàgina del teu examen, y complir amb ella: Certifico haver realitzat aquest examen de forma individual, sense cap comunicació amb altres persones.

---

10 points

1. **Àlgebra i anàlisi multivariant.** Sigui $\mathbf{X}$ una matriu $n \times p$ de variables quantitatives. Per contestar les preguntes a continuació, procureu definir adequadament els vectors o les matrius addicionals que es puguin necessitar.

   (a) (1p) Sigui $\mathbf{m}_r$ el vector $n \times 1$ amb les mitjanes de totes les files de $\mathbf{X}$. Dona una expressió matricial per obtenir $\mathbf{m}_r$ a partir de la matriu de dades originals.

   (b) (1p) Es vol centrar $\mathbf{X}$ restant de cada fila la seva mitjana corresponent. Sigui $\mathbf{X}_{rc}$ aquesta matriu centrada per files. Demostra, amb expressió matricial, com obtenir $\mathbf{X}_{rc}$ a partir de $\mathbf{X}$ i $\mathbf{m}_r$.

   (c) (1p) Es vol generar la matriu *doblement centrada* $\mathbf{X}_{dc}$ restant de cada columna de $\mathbf{X}_{rc}$ la seva mitjana corresponent. Indica amb una expressió matricial, com obtenir $\mathbf{X}_{dc}$ a partir de $\mathbf{X}_{rc}$.

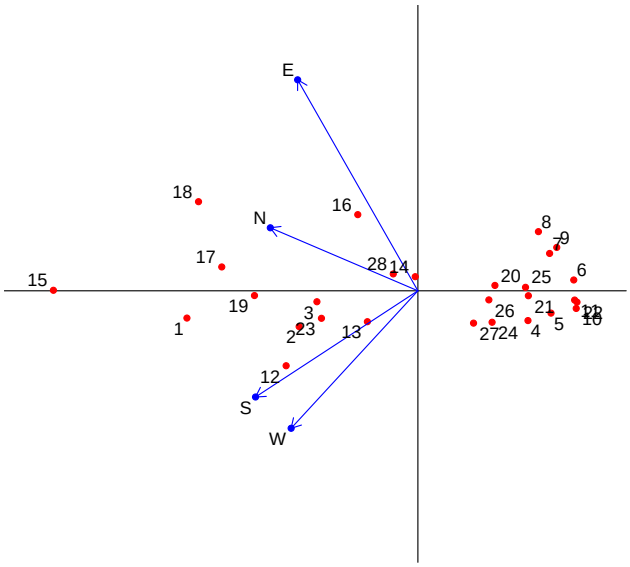
   (d) (2p) Dona una expressió matricial per la matriu de covariancies mostrals de $\mathbf{X}_{dc}$, derivant la seva relació amb la matriu de covariancies de $\mathbf{X}$.

(e) (3p) Descriu la descomposició espectral de la matriu de covariancies de $\mathbf{X}_{dc}$, indicant les propietats de les matrius obtingudes a la descomposició. Que es pot dir sobre els valors propis?

(f) (2p) Indica amb una expressió matricial com es pot obtenir la matriu de correlacions de les columnes de $\mathbf{X}_{dc}$ a partir de la seva matriu de covariancies. Tenen $\mathbf{X}$ i $\mathbf{X}_{dc}$ la mateixa matriu de correlacions?

10 points

2. **Anàlisi de components principals.** S'ha determinant el pes del dipòsit de suro als troncs de 28 arbres a les quatre direccions Nord (N), Est (E), Sud (S) i Oest (W). S'han analitzat les dades amb un anàlisi de components principals. Un biplot, fet amb aspect ratio 1, de la matriu de dades es mostra a la figura mes avall. Els valors propis obtinguts a l'anàlisi són $\lambda_1 = 984.44, \lambda_2 = 59.79, \lambda_3 = 23.91$ and $\lambda_4 = 18.20$.



(a) (1p) Dona la teva interpretació del primer component principal.

(b) (1p) Dona també la teva interpretació del segon component principal.

(c) (1p) Corresponen els resultats donats a un PCA basat en correlacions o en covariancies? Argumenta la teva resposta.

(d) (1p) Representa el biplot components principals estandaritzats o no estandaritzats? Argumenta la teva resposta.

(e) (1p) Segons els resultats de l'anàlisi, quin arbre té els seus dipòsits de suro mes a prop del vector de mitjanes de la mostra?

(f) (1p) Segons els resultats de l'anàlisi, quin arbre té mes dipòsit de suro a la direcció Oest?

(g) (1p) Calcula la bondat d'ajust d'aquesta representació bi-dimensional de la matriu de dades.

(h) (1p) Quin tan percent de la variància total quedaria explicada per una gràfica del primer versus el tercer component principal?

(i) (1p) Segons els resultats de l'anàlisi, quina parella de variables té coeficient de correlació mes elevada?

(j) (1p) El biplot representat dona una approximació a la matriu de correlacions de les variables. És possible millorar l'approximació de les correlacions? Argumenta la resposta.

10 points

3. **Normal multivariant.** Un estadístic analitza una matriu de variables quantitatives $\mathbf{X}$ amb $n = 100$ observacions. Les primeres 50 observacions corresponen al grup dels homes, i la segona meitat de la mostra representen les dones. Es fan càlculs en R, obtenint els resultats a continuació.

```
> HotellingsT2(X,mu=c(0,0),test="f")

Hotelling's one sample T2-test

data:  X
T.2 = 0.026962, df1 = 2, df2 = 98, p-value = 0.9734
> X1 <- X[1:50,]
> X2 <- X[51:100,]
> m1 <- colMeans(X1)
> m2 <- colMeans(X2)
> S1 <- cov(X1)
> S2 <- cov(X2)
```

```
> T2 <- (m1-m2)%*%solve(S1/50+S2/50)%*%(m1-m2)
> T2
         [,1]
[1,] 2.738152
> pchisq(T2,ncol(X),lower.tail = FALSE)
           [,1]
[1,] 0.2543419
```

(a) (1p) Quantes variables té la matriu $\mathbf{X}$?

(b) (1p) Descriu formalment la hipòtesi nul.la i alternativa del primer contrast realitzat en els càlculs.

(c) (1p) Quina és la conclusió del primer contrast realitzat?

(d) (1p) Descriu formalment la hipòtesi nul.la i alternativa del segon contrast realitzat en els càlculs.

(e) (1p) Quin és el valor p del segon contrast realitzat en els càlculs, i quina és la conclusió del test?

(f) (2p) Quina és la distribució de referència de l'estadístic de prova al segon contrast realitzat en els càlculs.

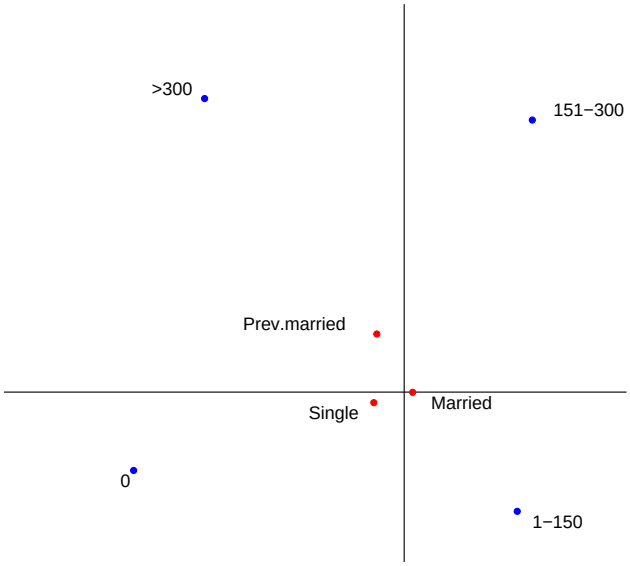
(g) (1p) En quins supòsits es base el segon contrast realitzat?

(h) (2p) Quines ventatges té l'ús de la prova amb el T^2 de Hotelling en comparació de fer proves amb una t de Student per comparar les mitjanes dels dos grups?

10 points

4. **Anàlisi de correspondències.** En un estudi d'una mostra de 3888 dones es va obtenir una taula creuada del seu estat civil ("single", "married" o "previously married") i el seu consum de cafeïna (0, 1-150, 151-300 or >300). La taula de contingencia obtinguda de l'estat civil (en files) per consum de cafeïna (en columnes) s'analitza amb un anàlisi de correspondències. El biplot de les perfils fila amb els resultats numèrics de l'anàlisi es mostren a continuació.

```
> out <- ca(caff); summary(out)
```




```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.008492  63.9  63.9  ***************
 2      0.004793  36.1 100.0  *********
        -------- -----
 Total: 0.013286 100.0


Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | Mrrd |  779 1000  141 |   49 999 221 |   -1   1   0 |
2 | Prvm |   36 1000  378 | -159 182 107 |  336 818 856 |
3 | Sngl |  185 1000  481 | -176 893 672 |  -61 107 143 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |    0 |  233 1000  383 | -144 955 571 |  -31  45  48 |
2 | 1150 |  491 1000  219 |   60 614 210 |  -48 386 234 |
3 | 1513 |  191 1000  238 |   68 282 105 |  109 718 473 |
4 |  300 |   85 1000  161 | -106 450 113 |  118 550 245 |
```

(a) (1p) Existeix associació significativa entre consum de cafeïna i l'estat civil? Argumenta la resposta.

(b) (1p) Quina és la inercia total de la taula de contingencia?

(c) (1p) Quin tan percent de la inerica total queda explicada pel biplot?

(d) (1p) Quin és el perfil fila marginal de la taula de contingencia?

(e) (1p) Quantes dimensions es necessiten per representar la matriu de perfiles fila sense error?

(f) (1p) Quina categoria de consum de cafeïna és la mes comuna?

(g) (1p) Quin estat civil s'assembla mes al perfil marginal de les files?

(h) (1p) Quin tipus de dona es troba mes entre les no-consumidores de cafeïna?

(i) (1p) Quina és la inercia màxima possible per una taula creuada d'aquestes dimensions?

(j) (1p) Quina categoria de les dades queda millor explicada per la segona dimensió de l'anàlisi?