Groupe ESIEE **06/11/08**

O.de Cambry **MA-412**

Traitement statistique des données

Examen

Durée :**2 heures** Sujet à traiter **avec** documents

On considère le tableau de données noté $\mathcal{C}$ suivant concernant 12 aliments.

Numéro	Aliment	X_1	X_2	X_3	X_4
1	céréales	13.0	71.0	12.0	2.5
2	pommes de terre	80.0	17.0	2.0	0
3	haricots secs	11.0	61.0	22.0	1.5
4	epinards	90.0	5.5	3.0	0.3
5	chou	92.0	5.5	1.5	0.2
6	orange	86.0	12.0	1.0	0.2
7	poulet	73.5	0	22.0	4.0
8	porc	56.0	0	16.0	27.0
9	cabillaud	81.0	0	17.5	0.5
10	oeuf	74.0	1.0	13.0	11.5
11	yoghourt	88.0	8.0	3.5	2.5
12	camembert	52.0	4.0	17.5	24.5
	Moyenne	66.38	15.42	10.92	6.23
	Ecart-type	27.09	23.25	7.89	9.25

Les variables étudiées sont des quantités de composants fondamentaux dans chacun des aliments

- X_1 :eau
- X_2 :glucides
- X_3 :protides
- X_4 :lipides

- A - Analyse en composantes prindipales

On a effectué une Analyse en Composantes Principales (ACP) sur les données normalisées, dont les résultats sont rassemblés à la suite des questions.

1. *Représenter et commenter la boite à moustaches pour X_1.*
2. *Compléter le cercle des corrélations en ajoutant X_1 et X_2 (Figure 1).*
3. *Commenter les résultats de L'ACP.*

- B - Classification

Dans la suite on travaille sur les données normalisées. Les distances entre les points se trouvent en annexe.

1. Soient A et B, 2 groupements de points disjoints, on définit le critère du lien minimum entre ces 2 groupements par : $D(A,B) = \min_{M\in A, M'\in B} d(M,M')$.

   La hiérarchie $\mathcal{H}(\mathcal{C})$ du saut minimum définie à partir du critère d'aggrégation du lien minimum a donné le dendrogramme représenté sur la Figure 2.

   (a) *Déterminer les indices d'aggrégation et compléter le dendrogramme.*

   (b) *Représenter sur la Figure 3 l'arbre couvrant minimal associé à cette hiérarchie en notant sur les arètes de l'arbre les indices d'aggrégation correspondants.*

2. Dans la suite on notera pour $A \in \mathcal{H}(\mathcal{C})$, $v(A)$ l'indice d'aggrégation associé à la constitution de l'ensemble A dans la hiérarchie. On considère à présent l'application Δ définie pour 2 points M et M' de $\mathcal{C}$ par $\Delta(M,M') = \min_{A\in\mathcal{H}(\mathcal{C})} \{v(A); M \in A, M' \in A\}$

   (a) *Montrer que Δ est une distance ultramétrique.*

   (b) *Déterminer Δ(orange,céréales) et Δ(camembert,yoghourt).*

- C - Discrimination

On considère maintenant la partition en 2 classes des données dans le plan principal définie ci-dessous :

- C_1={cereales, pommes de terre, haricots secs, epinards, chou, orange, yoghourt}
- C_2={poulet, porc, cabillaud, oeuf, yoghourt, camembert}

On notera n_l le nombre d'élément de la classe C_l. Dans la suite on souhaite classer dans C_1 ou C_2 le fromage frais dont les caractéristiques sont :
eau=60, glucides=4, protides=13, lipides=4.

1. *Déterminer les coordonnées dans le plan principal du* fromage frais *et placer le point correspondant sur la figure 4.*

2. On souhaite discriminer les 2 classes C_1 et C_2 en utilisant la méthode des 2 plus proches voisins.
   Déterminer la valeur des fonctions discriminantes des 2-PPV pour le fromage frais *et en déduire la classe dans laquelle il est affecté.*

3. On utilise maintenant la méthode de Bayes dite naive dans laquelle la matrice de covariance est diagonale et identique pour les 2 classes. On dispose donc des 2 probabilité a priori α_1 et α_2 et on suppose que sachant qu'un individu est dans la classe C_i sa position suit une loi $\mathcal{N}(m_i, V)$ où $V = \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix}$.

   On rappelle qu'une fonction discriminante est linéaire si pour le point $M(x_1,x_2)$ la décision est de la forme $\delta(x) = \left| \begin{array}{lll} 1 & \text{si} & L(M) \leq 0 \\ 2 & \text{si} & L(M) \geq 0 \end{array} \right.$ avec $L(M) = a_1x_1 + a_2x_2 + b$.

   (a) *Vérifier que la fonction discriminante est linéaire.*

   (b) *Déterminer les estimations de α_1, α_2, m_1 et m_2.*

(c) Si on note $\sigma_i^2(l)$ la variance sur l'axe u_i pour les éléments de la classe C_l, on estime les coefficient de la matrice V par $s_i^2 = \dfrac{(n_1-1)\sigma_i^2(1)+(n_2-1)\sigma_i^2(2)}{n_1+n_2-2}$.
En déduire V.

(d) On a obtenu $a_1 = 0.55$, $a_2 = 5.5$, $b = -1.1$.
Tracer la droite séparatrice correspondante sur la Figure 4 et déterminer l'affectation de fromage frais

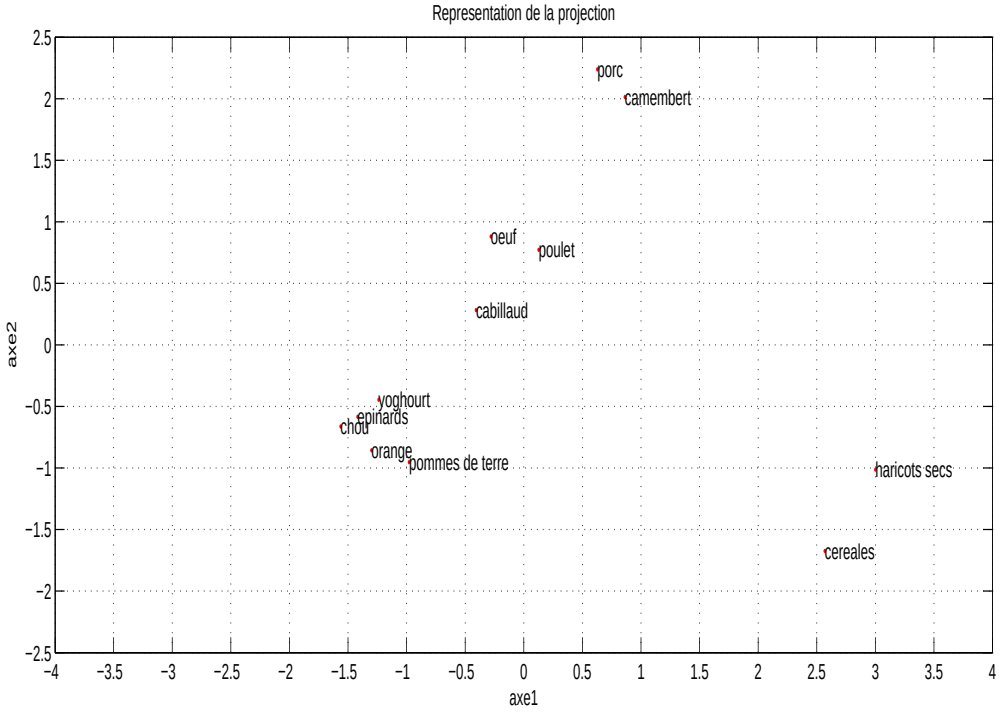
Barème indicatif :

- A - : 5pts

- B - : 7pts

- C - : 8pts

Annexe



- A -

Matrice de corrélation : $R = \begin{pmatrix} 1.0000 & -0.8346 & -0.5972 & -0.2178 \\ -0.8346 & 1.0000 & 0.1617 & -0.2898 \\ -0.5972 & 0.1617 & 1.0000 & 0.4160 \\ -0.2178 & -0.2898 & 0.4160 & 1.0000 \end{pmatrix}$

Valeurs propres :

$$\begin{aligned}\lambda_1 &= 2.137\\ \lambda_2 &= 1.414\\ \lambda_3 &= 0.448\\ \lambda_4 &= 0.001\end{aligned}$$

Vecteurs propres :

$$U = \begin{pmatrix} -0.6746 & 0.0859 & -0.1931 & 0.7073 \\ 0.5201 & -0.5290 & 0.2406 & 0.6259 \\ 0.4930 & 0.4107 & -0.7348 & 0.2198 \\ 0.1770 & 0.7377 & 0.6041 & 0.2442 \end{pmatrix}$$

- B -

Distance entre les points

céréales	1	0											
pommes de terre	2	3.63	0										
haricots secs	3	1.34	4.06	0									
epinards	4	4.17	0.63	4.47	0								
chou	5	4.28	0.67	4.63	0.20	0							
orange	6	3.96	0.33	4.38	0.41	0.36	0						
poulet	7	3.99	2.68	3.50	2.53	2.73	2.78	0					
porc	8	4.37	3.60	4.22	3.56	3.69	3.67	2.68	0				
cabillaud	9	4.02	2.10	3.73	1.88	2.08	2.16	0.74	3.01	0			
oeuf	10	3.88	2.00	3.81	1.86	2.02	2.05	1.40	1.84	1.34	0		
yoghourt	11	4.02	0.59	4.33	0.28	0.40	0.44	2.43	3.32	1.84	1.66	0	
camembert	12	4.06	3.50	3.85	3.49	3.63	3.60	2.43	0.40	2.81	1.72	3.25	0
		1	2	3	4	5	6	7	8	9	10	11	12

- C -

Coordonnées dans le plan principal et caractéristiques des classes

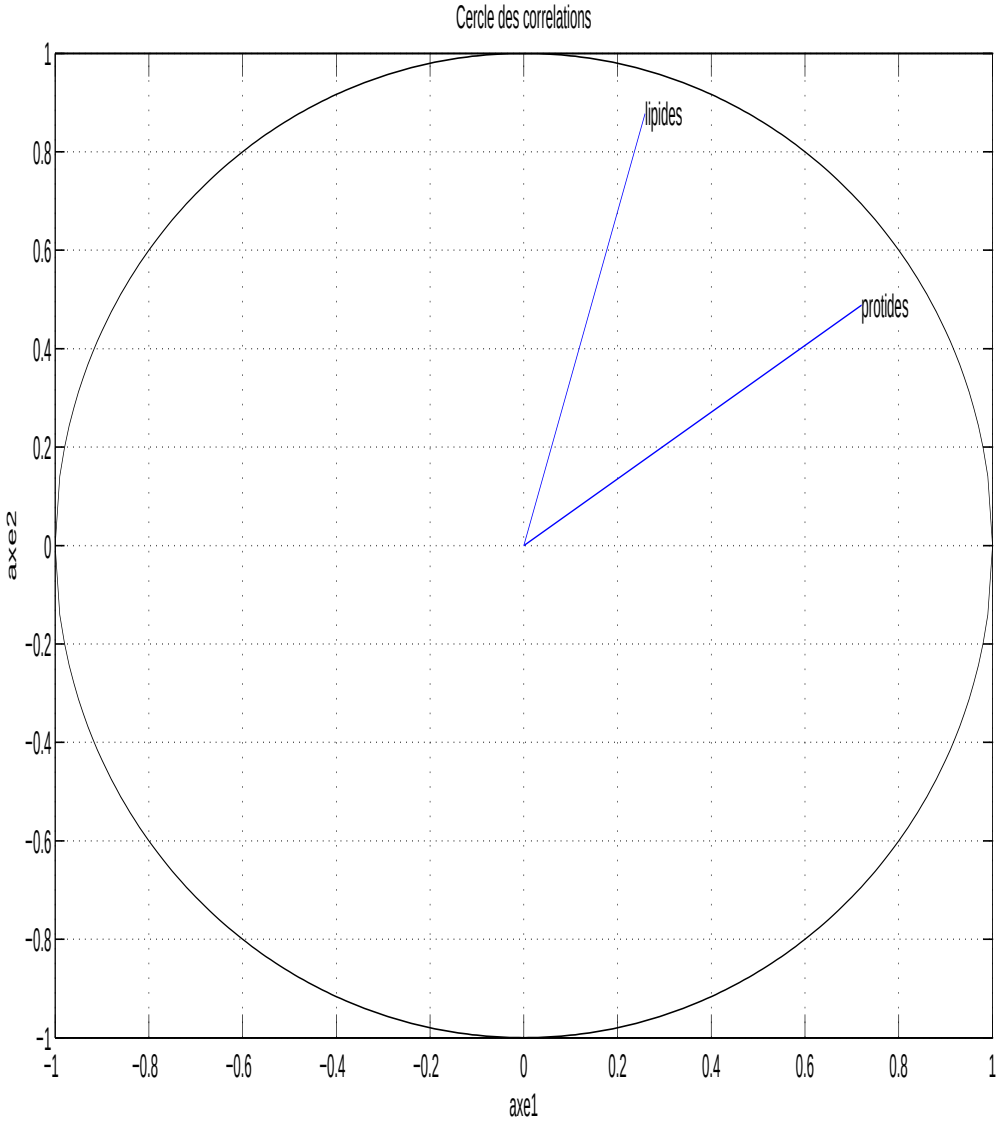
d : distance entre l'aliment et fromage frais dans le plan principal.
Xp_1, Xp_2 : coordonnées du point dans le plan principal

C_1	Nom	Xp_1	Xp_2	d	C_2	Nom	Xp_1	Xp_2	d
1	céréale	2.57	-1.67	3.17	7	poulet	0.13	0.77	0.62
2	pommes de terre	-0.98	-0.96	1.48	8	porc	0.63	2.24	2.16
3	haricots secs	3.00	-1.01	3.23	9	cabillaud	0.41	0.28	0.41
4	epinards	-1.42	-0.58	1.60	10	oeuf	0.28	0.88	0.76
5	chou	-1.56	-0.66	1.76	12	camembert	0.86	2.01	2.04
6	orange	-1.30	-0.86	1.65					
11	yoghourt	-1.24	-0.45	1.38					
	Moyenne	-0.13	0.88			Moyenne	0.19	1.24	
	Ecart-type	1.86	0.37			Ecart-type	0.50	0.76	

Ne pas oublier de rendre cette page et les suivantes avec la copie

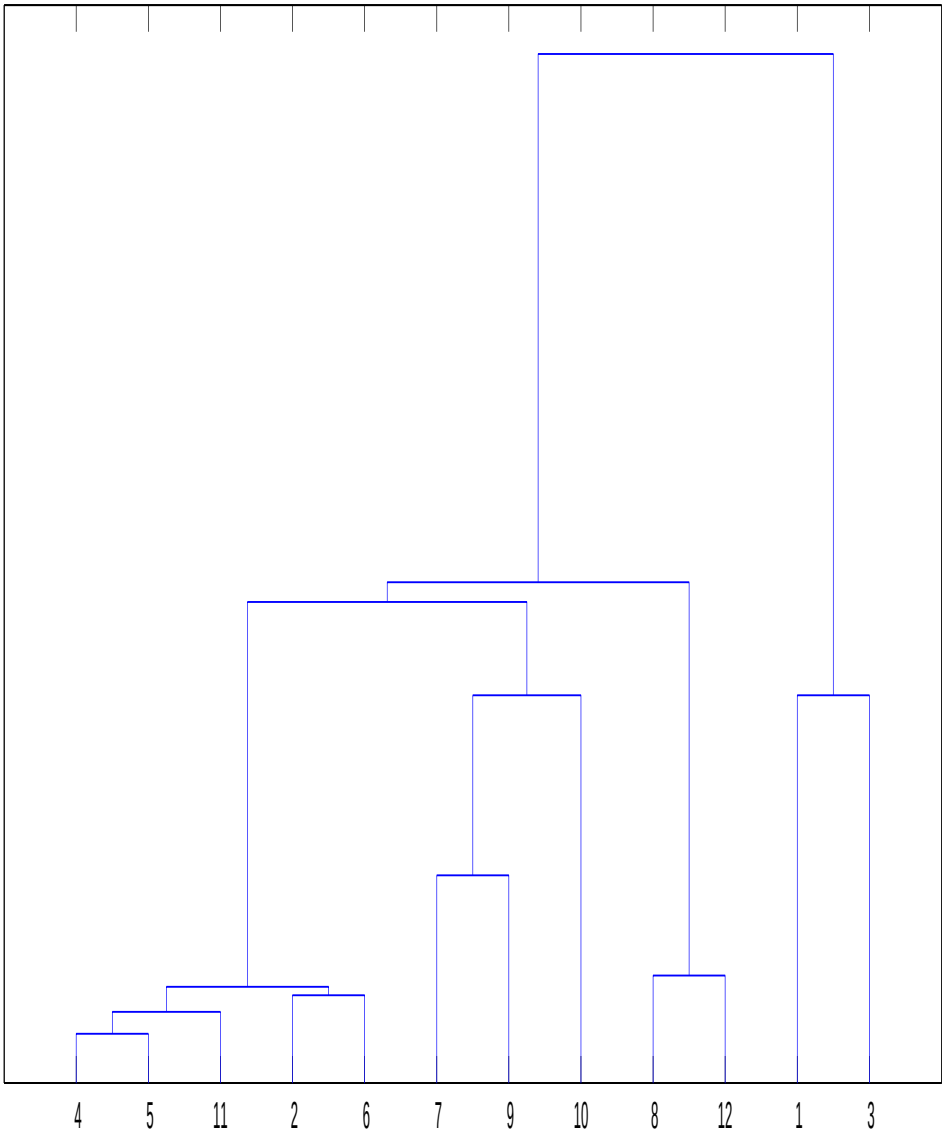
NOM :

Question - A - 2 : Figure 1



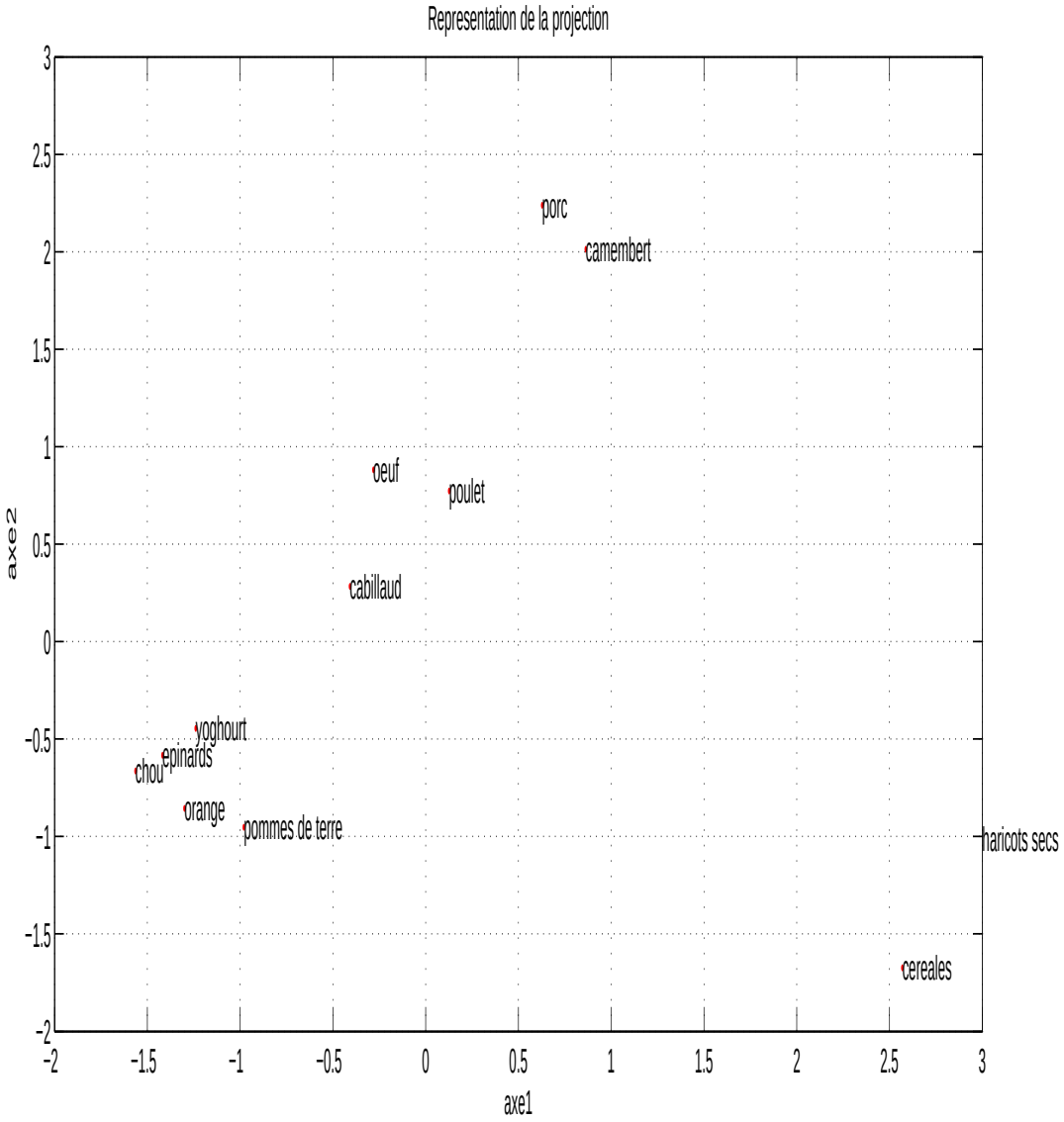
NOM :

Question - B - 1 - a : Figure 2



NOM :

Question - B - 1 - b : Figure 3



NOM :

Question - C - : Figure 4