**Groupe ESIEE** **04/11/09**

**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 $\mathcal{C}$, se trouvant en annexe, concernant 16 années sur lesquelles on s'est intéressé à la qualité du vin de Bordeaux.

On considère les variables $X_1$, somme des températures moyennes journalières de l'année et $X_3$, nombre de jours de grande chaleur. Les variables étudiées sont des mesures météorologique dans la région.

- $X_1$ :somme des températures moyennes journalières de l'année
- $X_2$ :ensoleillement en heures
- $X_3$ :nombre de jours de grande chaleur
- $X_4$ :hauteur de pluies en mm

## - A - Analyse des données

1. (a) *Déterminer les coefficients de la régression linéaires de $X_3$ par rapport à $X_1$.*
   (b) *Représenter la droite de régression sur la figure 1.*
   (c) *Peut-on considérer que le nombre de jours de grande chaleur dépend linéairement de températures journalières ?* .
2. On a effectué une Analyse en Composantes Principales (ACP) sur les données normalisées, dont les résultats sont rassemblés à la suite des questions.
   *Commenter les résultats de L'ACP.*

## - B - Classification

Dans la suite on travaille sur les données normalisées dans le plan principal.Les coordonnées des points se trouvent en annexe.

1. On considère une classification en 3 classes par la méthode des kmeans en 2 phases :
   **phase1 :** algorithme des centres mobiles
   **phase2 :** algorithme de transfert
   A l'issue de l'étape $m$ l'algorithme est toujours dans la phase 1 et la classification suivante est obtenue :
   $L_m = [2 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 2 \quad 3 \quad 2 \quad 1 \quad 3 \quad 2 \quad 3 \quad 2 \quad 2 \quad 2]$
   Les éléments concernant cette partition sont donnés en annexe.
   A l'issue de l'algorithme on obtient la classification suivante :
   $L_{final} = [2 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 2 \quad 3 \quad 2 \quad 2 \quad 2]$

(a) *Décrire la fin de l'algorithme.*

(b) *Calculer la valeur finale du critère de la somme des inerties.*

2. Soient $A$ et $B$, 2 groupements de points disjoints, on définit la distance du lien maximal entre ces 2 groupements par : $D(A,B) = \max_{x\in A, y\in B} d(A,B)$.

   La hiérarchie obtenue sur les données à partir de ce critère a donné le dendrogramme incomplet représenté sur la Figure 2.

   (a) *Déterminer les autres éléments de la hiérarchie en utilisant l'algorithme des plus proches voisins réciproques sur les ensembles obtenus $(A_i)_{1\leq i\leq 6}$, en l'initialisant sur $A_6$ et en détaillant les étapes qui suivent.*

   (b) *Représenter sur la Figure 2 le dendrogramme complet.*

   (c) *En déduire la classification en 3 classes obtenue.*

   (d) La valeur du critère de la somme des inerties pour cette classe est $W = 18.92$. *Que peut-on penser de la classification obtenue à la question - B - 1 ?*

## - C - Discrimination

On considère maintenant la partition en 3 classes des données suivantes :
$L_q = [3 \quad 3 \quad 3 \quad 2 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1 \quad 2 \quad 3 \quad 2 \quad 2 \quad 3 \quad 2]$.
Cette classification a été obtenue par dégustation en notant les vins :

**1** : bonne année

**2** : année moyenne

**3** : année médiocre

Elle est représentée sur la Figure 3 et les caractéristiques des classes se trouvent en annexe.

1. (a) *Calculer la valeur finale du critère de la somme des inerties pour cette classification.*

   (b) *Au vu du résultat précédent et de la figure 3 pensez-vous que les variables choisies sont discriminantes pour la qualité d'une année.*

   (c) *Déterminer la limite des classes pour la discrimination par la distance au centre et représenter les sur la figure 3. Quel est le taux de points mal classés par cette méthode.*

2. On souhaite déterminer la classe de l'année 39 dont les coordonnées sont :
   $An_{39} = [3022 \quad 1285 \quad 9 \quad 303]$.

   (a) *Déterminer les coordonnées de $An_{39}$ dans le plan principal et ajouter ce point à la figure 3.*

   (b) *Dans quelle classe doit-on le classer par la méthode précédente ?*

## Barème indicatif :

**- A - :** 6pts

**- B - :** 8pts

**- C - :** 6pts

# Annexes

## Données

| | Année | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| 1 | 40 | 3094 | 1329 | 11 | 339 |
| 2 | 41 | 3009 | 1210 | 15 | 536 |
| 3 | 42 | 3227 | 1331 | 21 | 414 |
| 4 | 43 | 3308 | 1368 | 24 | 282 |
| 5 | 44 | 3212 | 1289 | 17 | 302 |
| 6 | 45 | 3381 | 1444 | 25 | 253 |
| 7 | 46 | 3061 | 1175 | 12 | 261 |
| 8 | 47 | 3478 | 1317 | 42 | 259 |
| 9 | 48 | 3126 | 1248 | 11 | 315 |
| 10 | 49 | 3458 | 1508 | 43 | 286 |
| 11 | 50 | 3252 | 1361 | 26 | 346 |
| 12 | 51 | 3052 | 1186 | 14 | 443 |
| 13 | 52 | 3270 | 1399 | 24 | 306 |
| 14 | 53 | 3198 | 1299 | 20 | 367 |
| 15 | 54 | 2904 | 1164 | 6 | 311 |
| 16 | 55 | 3247 | 1277 | 19 | 375 |
| | Moyenne | 3205 | 1307 | 20.6 | 337 |
| | Ecart-type | 154.2 | 94.25 | 9.98 | 73.78 |

- $X_1$ :somme des températures moyennes journalières de l'année
- $X_2$ :ensoleillement en heures
- $X_3$ :nombre de jours de grande chaleur
- $X_4$ :hauteur de pluies en mm

## - A - Analyse des données

**Matrice de corrélation** :

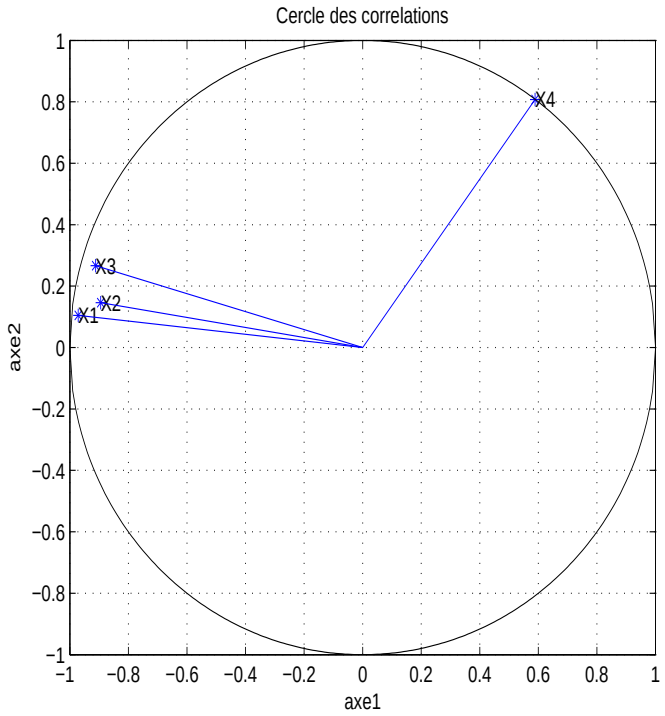
$$R = \begin{pmatrix} 1.0000 & 0.8319 & 0.9173 & -0.4852 \\ 0.8319 & 1.0000 & 0.7427 & -0.3973 \\ 0.9173 & 0.7427 & 1.0000 & -0.3341 \\ -0.4852 & -0.3973 & -0.3341 & 1.0000 \end{pmatrix}$$

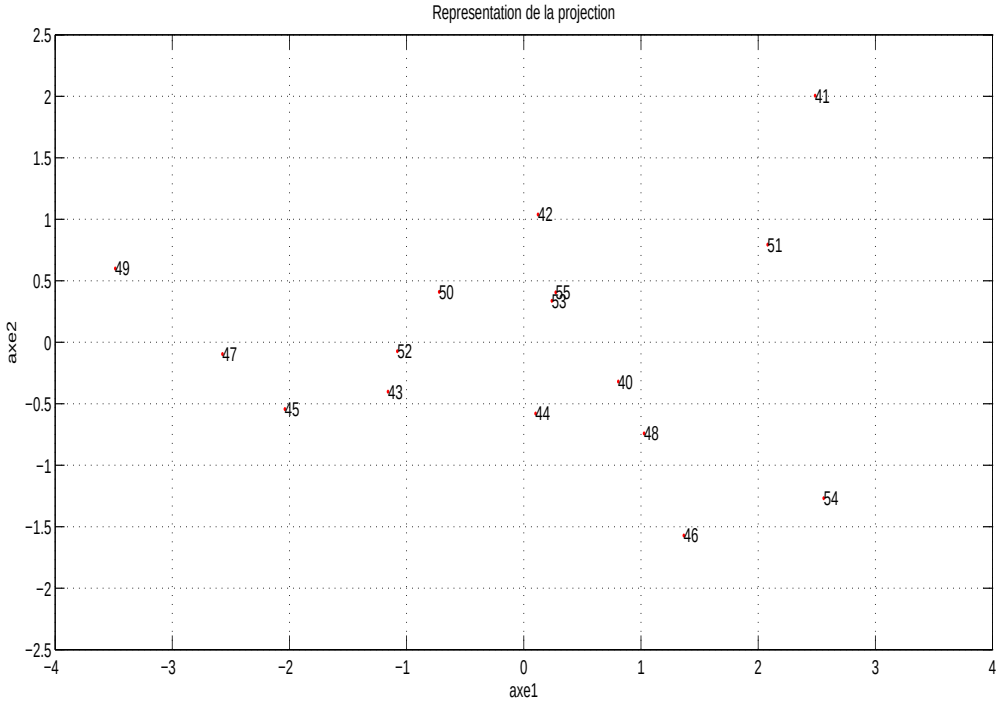
**Valeurs propres** :

$$\begin{aligned} \lambda_1 &= 2.92 \\ \lambda_2 &= 0.75 \\ \lambda_3 &= 0.27 \\ \lambda_4 &= 0.06 \end{aligned}$$

**Vecteurs propres** :

$$U = \begin{pmatrix} -0.5682 & 0.1202 & 0.2007 & -0.7889 \\ -0.5234 & 0.1682 & -0.8120 & 0.1960 \\ -0.5334 & 0.3068 & 0.5449 & 0.5696 \\ 0.3444 & 0.9291 & -0.0589 & -0.1215 \end{pmatrix}$$


Cercle des correlations
X1
X2
X3
X4
axe1
axe2


Representation de la projection
axe1
axe2

## - B - Classification

**Coordonnées des points dans le plan principal** :

| | Année | $C_1$ | $C_2$ | | Année | $C_1$ | $C_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 40 | 0.81 | -0.32 | 9 | 48 | 1.03 | -0.74 |
| 2 | 41 | 2.49 | 2.01 | 10 | 49 | -3.49 | 0.60 |
| 3 | 42 | 0.12 | 1.04 | 11 | 50 | -0.72 | 0.41 |
| 4 | 43 | -1.16 | -0.40 | 12 | 51 | 2.08 | 0.79 |
| 5 | 44 | 0.10 | -0.58 | 13 | 52 | -1.08 | -0.07 |
| 6 | 45 | -2.04 | -0.54 | 14 | 53 | 0.24 | 0.34 |
| 7 | 46 | 1.37 | -1.57 | 15 | 54 | 2.56 | -1.27 |
| 8 | 47 | -2.57 | -0.10 | 16 | 55 | 0.27 | 0.41 |

### B - 1 : Kmeans

**Partition à l'étape m** $L_m = [2 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 2 \quad 3 \quad 2 \quad 1 \quad 3 \quad 2 \quad 3 \quad 2 \quad 2 \quad 2]$

**Centres de gravités à l'étape m** :

$$\begin{aligned} G_1^m &= (-3.49; 0.60) \\ G_2^m &= (1.11; 0.01) \\ G_3^m &= (-1.51; -0.14) \end{aligned}$$

**Distances aux centres des classes** :

| | Année | $d(M, G_1^m)$ | $d(M, G_2^m)$ | $d(M, G_3^m)$ |
|---|---|---|---|---|
| 1 | 40 | 4.39 | 0.45 | 2.33 |
| 2 | 41 | 6.13 | 2.43 | 4.54 |
| 3 | 42 | 3.63 | 1.42 | 2.02 |
| 4 | 43 | 2.53 | 2.30 | 0.44 |
| 5 | 44 | 3.78 | 1.17 | 1.67 |
| 6 | 45 | 1.84 | 3.19 | 0.66 |
| 7 | 46 | 5.32 | 1.60 | 3.22 |
| 8 | 47 | 1.15 | 3.68 | 1.06 |
| 9 | 48 | 4.71 | 0.76 | 2.61 |
| 10 | 49 | 0 | 4.63 | 2.11 |
| 11 | 50 | 2.77 | 1.87 | 0.96 |
| 12 | 51 | 5.57 | 1.25 | 3.72 |
| 13 | 52 | 2.50 | 2.18 | 0.44 |
| 14 | 53 | 3.74 | 0.93 | 1.82 |
| 15 | 54 | 6.33 | 1.94 | 4.23 |
| 16 | 55 | 3.76 | 0.92 | 1.87 |

**Somme des inerties** $W = 22.51$

**B - 2 : Classification hiérarchique**

**Partition obtenue**

- $A_1 = \{42 \quad 53 \quad 55\}$
- $A_2 = \{43 \quad 50 \quad 52\}$
- $A_3 = \{40 \quad 44 \quad 48\}$
- $A_4 = \{41 \quad 51\}$
- $A_5 = \{46 \quad 54\}$
- $A_6 = \{45 \quad 47 \quad 49\}$

**Tableau des distances pour le lien maximal**

| | | | | | |
|---|---|---|---|---|---|
| $A_1$ | 3.76 | | | | |
| $A_2$ | 2.77 | 1.92 | | | |
| $A_3$ | 4.70 | 1.99 | 2.21 | | |
| $A_4$ | 6.13 | 2.79 | 4.36 | 3.51 | |
| $A_5$ | 6.32 | 3.35 | 3.83 | 2.55 | 3.74 |
| | $A_6$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ |

## - C - Discrimination

**Caractéristiques des classes**

- $C_1 = \{45 \quad 47 \quad 49\}$
- $C_2 = \{43 \quad 44 \quad 46 \quad 50 \quad 52 \quad 53 \quad 55\}$
- $C_3 = \{40 \quad 41 \quad 42 \quad 48 \quad 51 \quad 54\}$

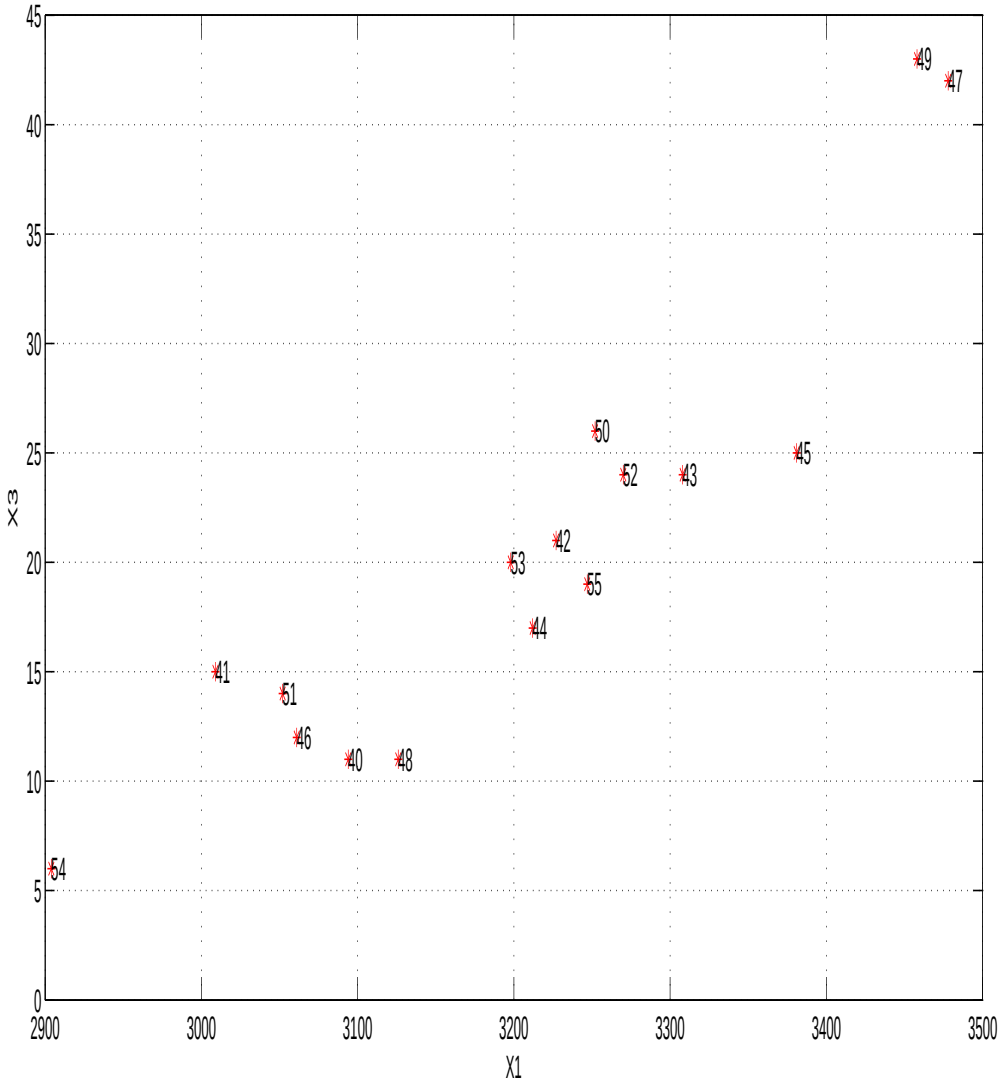
**Caractéristiques**

| | $C_1$ | | $C_2$ | | $C_3$ | |
|---|---|---|---|---|---|---|
| Centre | -2.70 | 0.01 | -0.14 | -0.21 | 1.51 | 0.25 |
| Inertie | 0.578 | | 1.144 | | 2.109 | |

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

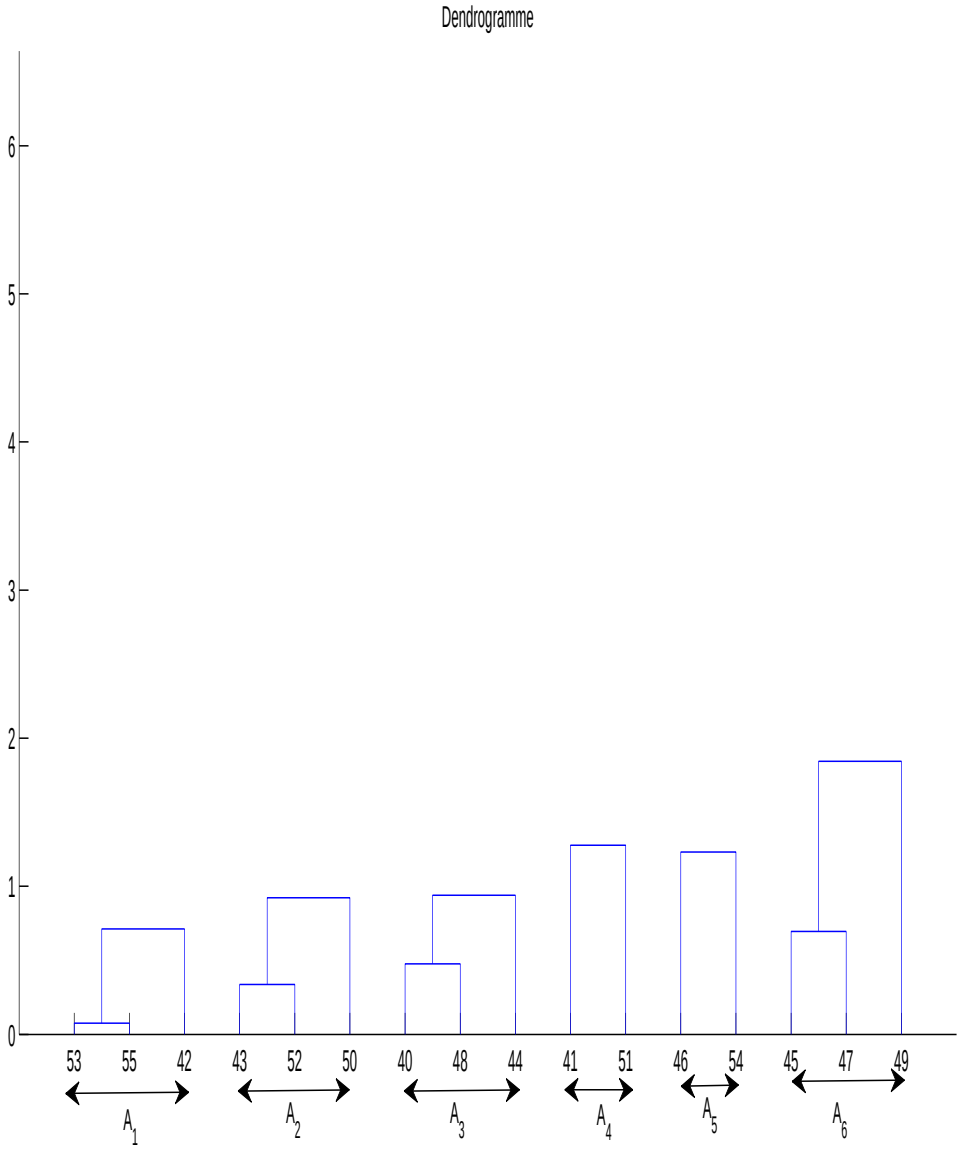
**NOM :**

**Question - A - 2 : Figure 1**



**NOM :**

**Question - B - 2 : Figure 2**



**NOM :**

**Question - C - : Figure 3**