

## CATALOGUE

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

## QCM

← Codez votre numéro d'étudiant ci-contre en noircissant (ne pas entourer ni cocher) au stylo l'intérieur des cases correspondantes (ligne 1 ⇔ 1 chiffre, etc), et écrivez vos nom et prénom ci-dessous.

Nom et prénom :

*JALABERT Théo*

## M1 ISFA - Analyse de données et Clustering

### Examen du 9/01/2017

Documents de cours et TD autorisés

Durée 3h

Toutes les feuilles sont à rendre en fin d'épreuve.

Les questions faisant apparaître le symbole ☺ peuvent présenter une ou plusieurs bonnes réponses.  
Les autres ont une unique bonne réponse.

Dans tout ce support, le seuil de rejet de l'hypothèse d'indépendance est fixé à 5%.

## Exercice

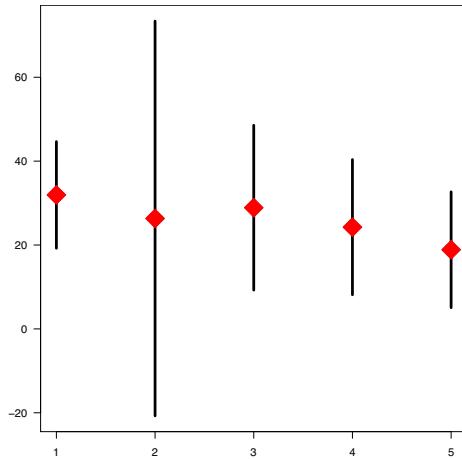
Nous étudions la qualité de l'air à la station permanente de Lyon Centre située rue du Lac dans le troisième arrondissement. Les données ont été prélevées au cours du premier semestre universitaire (du 12 septembre au 17 décembre 2016). Elles sont toutes exprimées en  $\mu\text{g}/\text{m}^3$  :

1. le dioxyde d'azote ( $\text{NO}_2$ ),
2. le monoxyde d'azote ( $\text{NO}$ ),
3. l'ozone ( $\text{O}_3$ ),
4. les particules en suspension dans l'air dont le diamètre est inférieur à 10 micromètres ( $\text{PM}_{10}$ ),
5. les particules fines, particules en suspension dans l'air dont le diamètre est inférieur à 2,5 micromètres ( $\text{PM}_{2.5}$ ).

```
library(ade4)
lyonc <- read.table("Lyon2016c.txt", h=TRUE)
head(lyonc)
  Date  NO2   NO   O3   PM10   PM2.5     Jour     Mois
1 13/09/16   22    1   95    38     20      mardi  septembre
2 14/09/16   38    3   71    39     24     mercredi  septembre
3 15/09/16   25    4   56    11      8      jeudi  septembre
4 16/09/16   35    7   35    10      8     vendredi  septembre
5 17/09/16   13    1   34      8      6     samedi  septembre
6 18/09/16   12    1   28      5      4     dimanche  septembre
dim(lyonc)
[1] 88 8
```

Nous représentons sur la figure 1 les variables par leur moyenne ( $\pm$  deux écarts-types).

## CATALOGUE

FIGURE 1 – Moyennes ( $\pm$  2 écarts-types)

**Question [pol11a]** En vous aidant de la représentation graphique de la figure 1, expliquez en quoi il est plus judicieux de réaliser une analyse en composantes principales normée qu'une analyse en composantes principales centrée.

 F  P  A  B 

L'ACP normée nous permet de mettre toutes les Va sur la m<sup>e</sup> échelle et de les comparer de manière + juste  
Tandis que l'ACP centrée peut conduire à des résultats biaisés lors de la comparaison des Va entre elles car la figure 1 montre qu'elles ne sont pas à la m<sup>e</sup> échelle.

**Question [pol11b]** Proposez un code permettant de réaliser une ACP sur les variables centrées et réduites. Les résultats seront renvoyés dans une variable NR.

 F  P  A  B 

```
library (ade4)
lyonc <- read.table ("Lyon2016C.txt", h= TRUE)
datac <- lyonc [, c ("NO2", "NO", "O3", "PM10", "PM2.5")]
datac_scaled <- scale (datac)
NR <- dudi.pca (datac_scaled, scannf = FALSE, nf = ncol (datac))
```

**Question [pol11c]** Qu'est-ce que l'inertie totale ? Combien vaut-elle dans l'analyse précédente ?

 F  P  A  B 

L'inertie totale est une mesure de variabilité totale des données dans une ACP.  
Elle représente la somme de toutes les variances des Va dans l'ACP et est exprimée en %

Inertie totale = NR\$eig

Les pourcentages d'inertie conservée sur chacun des deux premiers axes valent respectivement 73,27% et 13,57%.

## CATALOGUE

**Question [pol11d]** A partir de quelles informations sont-ils calculés ?

F  P  A  B

*Les % d'inertie conservée sur chaque axe d'une ACP sont calculés à partir de la variance expliquée par chaque axe.*

**Question [pol11e]** Proposez un code permettant d'afficher ces deux pourcentages (selon le même format, i.e. la valeur arrondie à deux chiffres après la virgule).

F  P  A  B

*Faire l'acp NR ci-dessus  
puis NR\$ eig[1:2].*

**Question [pol11f]** Calculez la valeur de l'inertie conservée dans le premier plan factoriel et encodez sa valeur tronquée à la deuxième décimale après la virgule :

0	1	2	3	<input checked="" type="checkbox"/>	5	6	7	8	9
.									
0	1	2	<input checked="" type="checkbox"/>	4	5	6	7	8	9
0	1	2	3	<input checked="" type="checkbox"/>	5	6	7	8	9

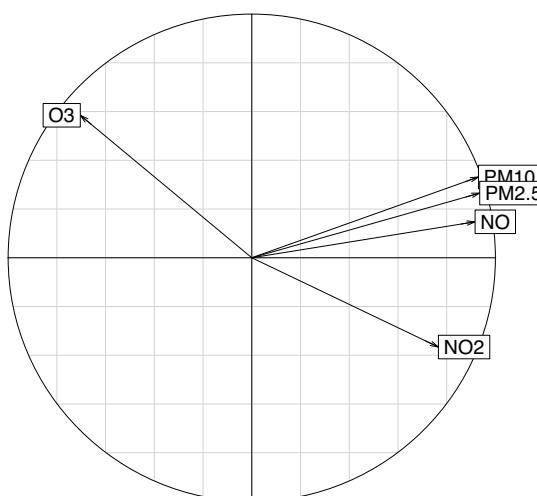


FIGURE 2 – Cercle des corrélations - Axes 1 & 2

## CATALOGUE

**Question [pol11g] ☽** Parmi les propositions suivantes, cochez celles qui sont vraies :

- L'ozone est inversement corrélée aux autres variables sur l'axe 1.
  - L'ozone intervient dans l'interprétation des deux axes.
  - L'axe 1 définit un axe de pollution : de la moins élevée (à gauche) à la plus élevée (à droite).
  - D L'axe 1 est caractérisé par un effet taille.
  - E Le monoxyde d'azote est corrélé à l'axe 2.
  - F L'axe 2 est caractérisé par le monoxyde d'azote et les particules fines.
  - G *Aucune de ces réponses n'est correcte.*

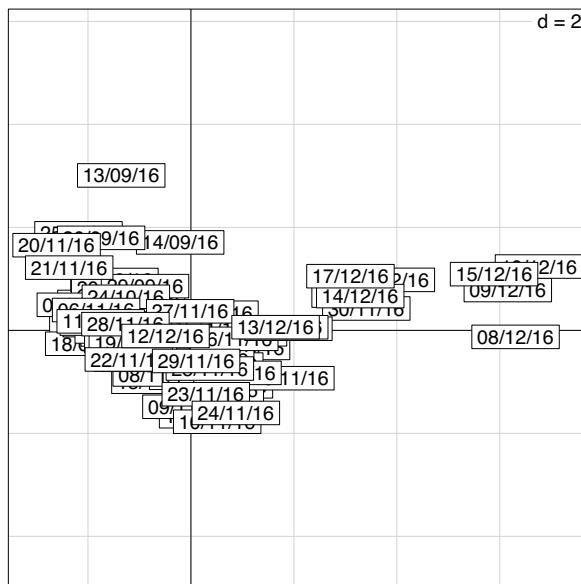


FIGURE 3 – Représentation de la carte factorielle des jours concernés par l'étude

**Question [pol11h]** En se basant sur la figure 3 et les résultats précédents, expliquez ce qu'il s'est passé le 13 septembre 2016 ? F P A B

F P A B

A large, empty rectangular frame with a black border, occupying most of the page.

## CATALOGUE

**Question [pol11i]** Le préfet du Rhône a imposé la circulation alternée le 9 décembre. Si l'analyse le permet, explicitez les raisons de sa décision.

F  P  A  B



Nous souhaitons déterminer s'il existe un lien entre la première composante principale et les jours de la semaine (resp. les mois).

À partir des valeurs de la première composante principale, six classes sont formées, notées C1 (valeurs les plus basses) à C6 (valeurs les plus hautes). En associant chaque individu à la classe à laquelle correspond sa valeur pour la première composante principale, nous formons une variable qualitative notée CP1Qual.

Nous nous intéressons dans un premier temps à l'existence de dépendance entre CP1Qual et les jours de la semaine. Pour cela, nous utilisons la fonction `chisq.test` qui renvoie la valeur 38,95 pour la distance du  $\chi^2$ .

**Question [pol11j]** Pouvons-nous rejeter l'hypothèse d'indépendance ?

non  B oui

**Question [pol11k]** Selon le seuil de rejet fixé et l'annexe fournie, quelle est la plus petite valeur du  $\chi^2$  qui permettrait de rejeter l'hypothèse d'indépendance ?

<input checked="" type="checkbox"/>	1	2	3	4	5	6	7	8	9
0	1	2	3	<input checked="" type="checkbox"/>	5	6	7	8	9
0	1	2	<input checked="" type="checkbox"/>	4	5	6	7	8	9
.									
0	1	2	3	4	5	6	<input checked="" type="checkbox"/>	8	9
0	1	2	3	4	5	6	<input checked="" type="checkbox"/>	8	9

Nous nous intéressons à présent à l'existence de dépendance entre CP1Qual et le mois de la mesure. Cette fois, nous utilisons la fonction `dudi.coa` qui renvoie une inertie totale de 0,7086003.

**Question [pol11l]** Calculez la distance du  $\chi^2$  correspondante et encodez sa valeur tronquée à la deuxième décimale après la virgule :

<input checked="" type="checkbox"/>	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	<input checked="" type="checkbox"/>	7	8	9
0	1	<input checked="" type="checkbox"/>	3	4	5	6	7	8	9
.									
0	1	2	<input checked="" type="checkbox"/>	4	5	6	7	8	9
0	1	2	3	4	<input checked="" type="checkbox"/>	6	7	8	9

## CATALOGUE

**Question [poll1m]** Pouvons-nous rejeter l'hypothèse d'indépendance ?

- non       oui

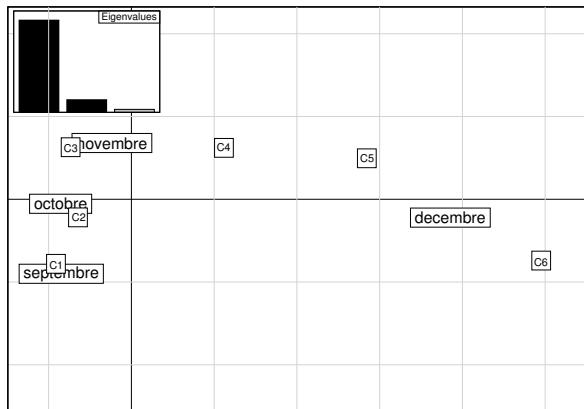


FIGURE 4 – Projection des modalités dans le premier plan factoriel

**Question [poll1n]** Interprétez la figure 4.

- F  P  A  B

## Exercice

Nous étudions la qualité de l'air à la station permanente de Vaulx-en-Velin située rue Maurice Audin à l'ENTPE. Les données ont été prélevées au cours du premier semestre universitaire (du 12 septembre au 17 décembre 2016). Seuls les jours où les données étaient complètes ont été conservées. Elles sont toutes exprimées en  $\mu\text{g}/\text{m}^3$  :

1. le dioxyde d'azote ( $\text{NO}_2$ ),
2. le monoxyde d'azote ( $\text{NO}$ ),
3. l'ozone ( $\text{O}_3$ ),
4. les particules en suspension dans l'air dont le diamètre est inférieur à 10 micromètres ( $\text{PM}_{10}$ ),
5. les particules fines, particules en suspension dans l'air dont le diamètre est inférieur à 2,5 micromètres ( $\text{PM}_{2.5}$ ).

```
library(ade4)
Vaulxc <- read.table("Vaulxc.txt", h=TRUE)
head(Vaulxc)
```

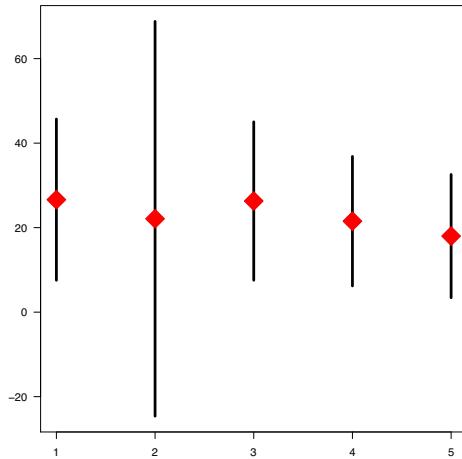
## CATALOGUE

	Date	N02	N0	03	PM10	PM2.5	Jour	Mois
1	12/09/16	17	3	79	22	14	lundi	septembre
2	15/09/16	15	2	54	9	5	jeudi	septembre
3	16/09/16	22	4	37	8	6	vendredi	septembre
4	17/09/16	5	1	31	8	4	samedi	septembre
5	18/09/16	5	1	24	5	3	dimanche	septembre
6	19/09/16	7	1	47	17	13	lundi	septembre

dim(Vaulxc)

[1] 74 8

Nous représentons sur la figure 5 les variables par leur moyenne ( $\pm$  deux écarts-types).

FIGURE 5 – Moyennes ( $\pm$  2 écarts-types)

**Question [pol11a]** En vous aidant de la représentation graphique de la figure 5, expliquez en quoi il est plus judicieux de réaliser une analyse en composantes principales normée qu'une analyse en composantes principales centrée.

 F  P  A  B 

**Question [pol11b]** Proposez un code permettant de réaliser une ACP sur les variables centrées et réduites. Les résultats seront renvoyés dans une variable NR.

 F  P  A  B

## CATALOGUE

**Question [pol11c]** Qu'est-ce que l'inertie totale ? Combien vaut-elle dans l'analyse précédente ?

F  P  A  B

Les pourcentages d'inertie conservée sur chacun des deux premiers axes valent respectivement 79,48% et 13,56%.

**Question [pol11d]** A partir de quelles informations sont-ils calculés ?

F  P  A  B

**Question [pol11e]** Proposez un code permettant d'afficher ces deux pourcentages (selon le même format, i.e. la valeur arrondie à deux chiffres après la virgule).

F  P  A  B

**Question [pol11f]** Calculez la valeur de l'inertie conservée dans le premier plan factoriel et encodez sa valeur tronquée à la deuxième décimale après la virgule :

0	1	2	3	<input checked="" type="checkbox"/>	5	6	7	8	9
.									
0	1	2	3	4	<input type="checkbox"/>	7	8	9	
0	1	2	3	4	<input checked="" type="checkbox"/>	6	7	8	9

**Question [pol11g] ☺** Parmi les propositions suivantes, cochez celles qui sont vraies :

- L'ozone est inversement corrélée aux autres variables sur l'axe 1.
- L'ozone intervient dans l'interprétation des deux axes.
- L'axe 1 définit un axe de pollution : de la moins élevée (à gauche) à la plus élevée (à droite).
- L'axe 1 est caractérisé par un effet taille.
- Le monoxyde d'azote est corrélé à l'axe 2.
- L'axe 2 est caractérisé par le monoxyde d'azote et les particules fines.
- Aucune de ces réponses n'est correcte.

## CATALOGUE

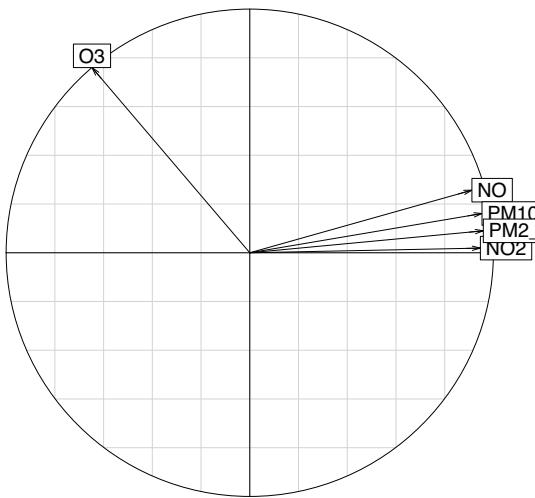


FIGURE 6 – Cercle des corrélations - Axes 1 &amp; 2

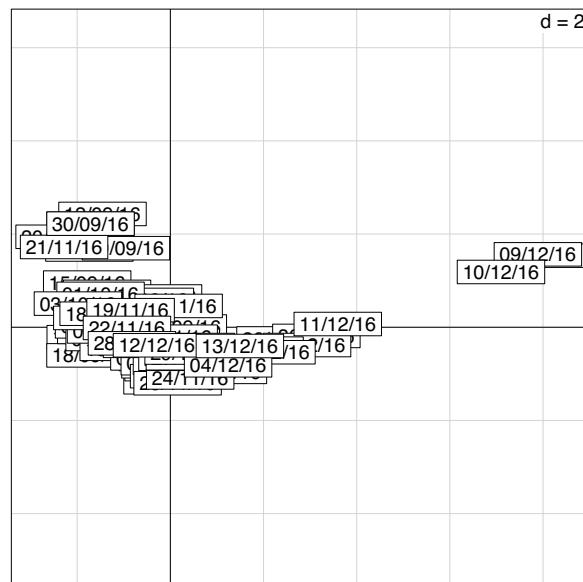


FIGURE 7 – Représentation de la carte factorielle des jours concernés par l'étude

**Question [pol11h]** En se basant sur la figure 7 et les résultats précédents, expliquez ce qu'il s'est passé le 30 septembre 2016 ?

F  P  A  B

## CATALOGUE

**Question [pol11i]** Le préfet du Rhône a imposé la circulation alternée le 9 décembre. Si l'analyse le permet, explicitez les raisons de sa décision.

F  P  A  B



Nous souhaitons déterminer s'il existe un lien entre la première composante principale et les jours de la semaine (resp. les mois).

À partir des valeurs de la première composante principale, six classes sont formées, notées C1 (valeurs les plus basses) à C6 (valeurs les plus hautes). En associant chaque individu à la classe à laquelle correspond sa valeur pour la première composante principale, nous formons une variable qualitative notée CP1Qual.

Nous nous intéressons dans un premier temps à l'existence de dépendance entre CP1Qual et les jours de la semaine. Pour cela, nous utilisons la fonction `chisq.test` qui renvoie la valeur 30,22 pour la distance du  $\chi^2$ .

**Question [pol11j]** Pouvons-nous rejeter l'hypothèse d'indépendance ?

non  B oui

**Question [pol11k]** Selon le seuil de rejet fixé et l'annexe fournie, quelle est la plus petite valeur du  $\chi^2$  qui permettrait de rejeter l'hypothèse d'indépendance ?

<input checked="" type="checkbox"/>	1	2	3	4	5	6	7	8	9
0	1	2	3	<input checked="" type="checkbox"/>	5	6	7	8	9
0	1	2	<input checked="" type="checkbox"/>	4	5	6	7	8	9
.									
0	1	2	3	4	5	6	<input checked="" type="checkbox"/>	8	9
0	1	2	3	4	5	6	<input checked="" type="checkbox"/>	8	9

Nous nous intéressons à présent à l'existence de dépendance entre CP1Qual et le mois de la mesure. Cette fois, nous utilisons la fonction `dudi.coa` qui renvoie une inertie totale de 0,8233881.

**Question [pol11l]** Calculez la distance du  $\chi^2$  correspondante et encodez sa valeur tronquée à la deuxième décimale après la virgule :

<input checked="" type="checkbox"/>	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	<input checked="" type="checkbox"/>	7	8	9
<input checked="" type="checkbox"/>	1	2	3	4	5	6	7	8	9
.									
0	1	2	3	4	5	6	7	8	<input checked="" type="checkbox"/>
0	1	2	<input checked="" type="checkbox"/>	4	5	6	7	8	9

## CATALOGUE

**Question [poll1m]** Pouvons-nous rejeter l'hypothèse d'indépendance ?

non       oui

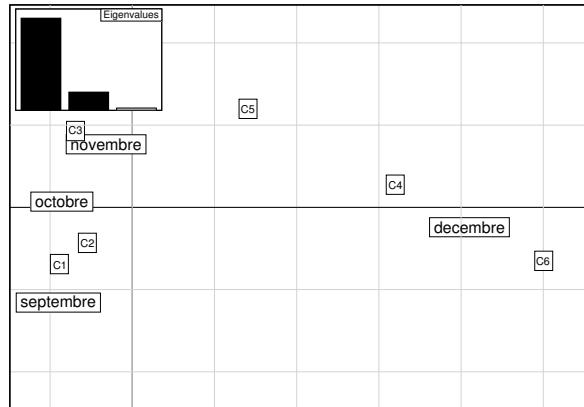


FIGURE 8 – Projection des modalités dans le premier plan factoriel

**Question [poll1n]** Interprétez la figure 8.

F  P  A  B

---

### Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 1. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut minimal.

	$V_1$	$V_2$	$V_3$
$e_1$	7	4.5	3
$e_2$	9	1.5	3.5
$e_3$	4.5	9	8
$e_4$	5.5	6	3.5
$e_5$	9.5	3	2.5
$e_6$	5.5	9.5	8.5
$e_7$	10	0.5	3.5
$e_8$	6.5	3.5	3
$e_9$	9.5	1.5	4

TABLE 1 – Valeurs des individus

Distance de Tchebychev :  $d_{\text{Cheb}}(A, B) = \max(|x_{1A} - x_{1B}|, |x_{2A} - x_{2B}|, \dots, |x_{mA} - x_{mB}|)$

Distance de Manhattan :  $d_{\text{Man}}(A, B) = |x_{1A} - x_{1B}| + |x_{2A} - x_{2B}| + \dots + |x_{mA} - x_{mB}|$

Distance euclidienne standard :  $d = \sqrt{(x_{1A} - x_{1B})^2 + (x_{2A} - x_{2B})^2 + \dots + (x_{mA} - x_{mB})^2}$

## CATALOGUE

**Question [cha-1-1]** Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- A  $d_\infty(e_2, e_3) = 7,5$      B  $d_\infty(e_2, e_4) = 8$      C  $d_{\text{Man}}(e_7, e_8) = 7$   
 D  $d_{\text{Man}}(e_6, e_3) = 2$      E  $d^2(e_2, e_7) = 2$      F  $d^2(e_8, e_9) = 14$   
 G Aucune de ces réponses n'est correcte.

**Question [cha-1-2]** Reportez dans la table 2 les distances de  $e_8$  à tous les autres points. Faites de même pour  $e_9$ . *Il s'agit de la distance euclidienne standard.* F P A M N L

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$	$\sqrt{1.25 + \frac{15}{2}}$	$\sqrt{10.5}$	$\sqrt{59.25}$	$\sqrt{7.5}$	$\sqrt{95}$	$\sqrt{67.25}$	$\sqrt{21.25}$	0
$e_9$	$\sqrt{16.25}$	$\sqrt{0.5}$	$\sqrt{59.25}$	$\sqrt{35.25}$	$\sqrt{45}$	$\sqrt{100.25}$	$\sqrt{1.25}$	$\sqrt{14}$

TABLE 2 – Tableau des distances

**Question [cha-1-3]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
. V1 V2 V3
1 7 4.5 3.0
2 9 1.5 3.5
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permettant de représenter la CAH à partir de D selon la stratégie indiquée ci-dessus : F P A B

```
D <- dist(C, method = "chebychev")
hc <- hclust(D, method = "average")
plot(hc)
```

**Question [cha-1-4]** Reportez dans la table 3 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 9. F P A M N L

---

Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 4. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut minimal.

## CATALOGUE

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 3 – Ultramétrique associée

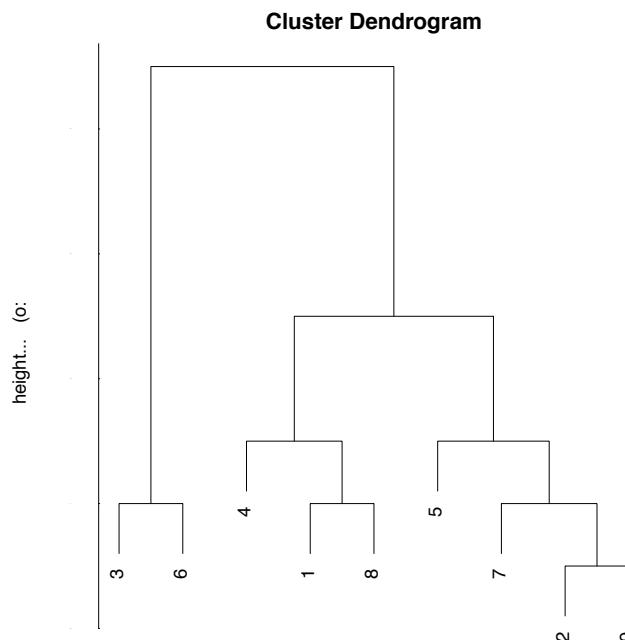


FIGURE 9 – Dendrogramme

	$V_1$	$V_2$	$V_3$
$e_1$	9	2.5	11.5
$e_2$	1.5	9	5.5
$e_3$	0.5	2	4.5
$e_4$	1.5	0	3.5
$e_5$	11	5	12.5
$e_6$	-1.5	1.5	3.5
$e_7$	2	9	4.5
$e_8$	11	4.5	10.5
$e_9$	9	5	11.5

TABLE 4 – Valeurs des individus

**Question [cha-2-1] ☺** Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- [D]  $d_\infty(e_2, e_3) = 7$       [B]  $d_\infty(e_2, e_4) = 8$       [F]  $d_{\text{Man}}(e_7, e_5) = 21$   
 [D]  $d_{\text{Man}}(e_6, e_3) = 3$       [E]  $d^2(e_2, e_3) = 51$       [F]  $d^2(e_7, e_9) = 113$   
 [G] Aucune de ces réponses n'est correcte.

## CATALOGUE

**Question [cha-2-2]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
. V1 V2 V3
1 9 2.5 11.5
2 1.5 9 5.5
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permettant de représenter la CAH à partir de D selon la stratégie indiquée ci-dessus : F P A B C

**Question [cha-2-3]** Reportez dans la table 5 les distances de  $e_8$  à tous les autres points. Faites de même pour  $e_9$ .

F P A M N L C

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$								
$e_9$								

TABLE 5 – Tableau des distances

**Question [cha-2-4]** Reportez dans la table 6 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 10.

F P A M N L C


---

## Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 7. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut minimal.

**Question [cha-3-1] ☺** Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- |   |   |   |
|---|---|---|
| <span style="border: 1px solid black; padding: 2px;">D</span> | <span style="border: 1px solid black; padding: 2px;">B</span> | <span style="border: 1px solid black; padding: 2px;">F</span> |
| $d_\infty(e_2, e_3) = 5$                                      | $d_\infty(e_2, e_4) = 8$                                      | $d_{\text{Man}}(e_7, e_5) = 9$                                |
| $d_{\text{Man}}(e_6, e_3) = 9.5$                              | <span style="border: 1px solid black; padding: 2px;">E</span> | <span style="border: 1px solid black; padding: 2px;">G</span> |
|   | $d^2(e_2, e_4) = 32$  | $d^2(e_8, e_9) = 71$  |
| <i>Aucune de ces réponses n'est correcte.</i>                 |   |   |

## CATALOGUE

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 6 – Ultramétrique associée

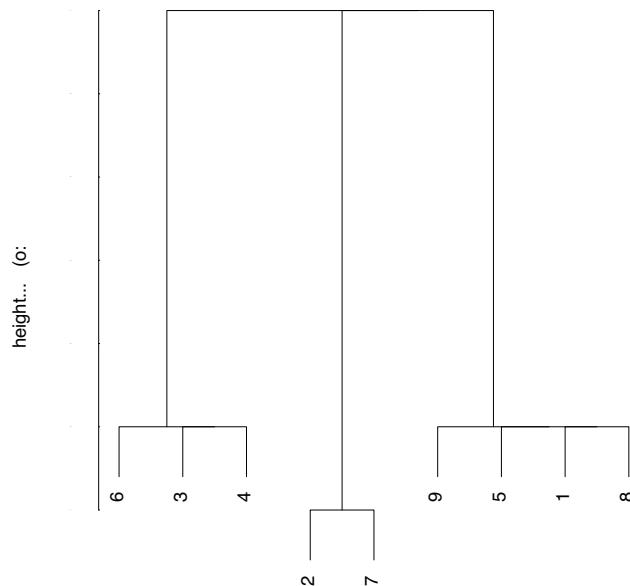
**Cluster Dendrogram**

FIGURE 10 – Dendrogramme

	$V_1$	$V_2$	$V_3$
$e_1$	1	5.5	3.5
$e_2$	1	7.5	2.5
$e_3$	6	6	6.5
$e_4$	5	6.5	6.5
$e_5$	8	4.5	3
$e_6$	9	3	3.5
$e_7$	0.5	5.5	2.5
$e_8$	0.5	6	1
$e_9$	8.5	4	3

TABLE 7 – Valeurs des individus

**Question [cha-3-2]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
```

```
. V1 V2 V3
```

```
1 1 5.5 3.5
```

```
2 1 7.5 2.5
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permet-

## CATALOGUE

**Question [cha-3-3]** Reportez dans la table 8 les distances de  $e_8$  à tous les autres points. Faites de même pour  $e_9$ .

[F] [P] [A] [M] [N] [L]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$								
$e_9$								

TABLE 8 – Tableau des distances

**Question [cha-3-4]** Reportez dans la table 9 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 11.

[F] [P] [A] [M] [N] [L]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 9 – Ultramétrique associée

---

## Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 10. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut maximal.

**Question [cha-4-1]** ☺ Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- []  $d_\infty(e_2, e_3) = 4$     []  $d_\infty(e_2, e_4) = 4$     []  $d_{\text{Man}}(e_7, e_5) = 12$   
 []  $d_{\text{Man}}(e_6, e_3) = 3.5$     []  $d^2(e_1, e_4) = 42$     []  $d^2(e_8, e_9) = 3$   
 [] *Aucune de ces réponses n'est correcte.*

## CATALOGUE

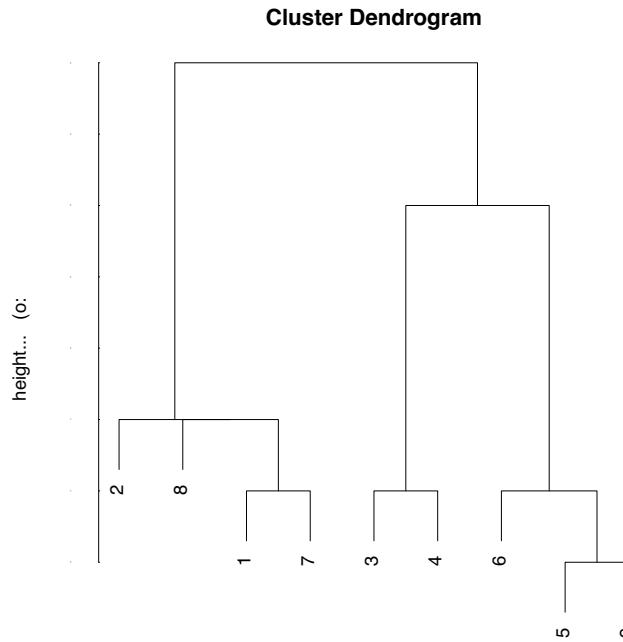


FIGURE 11 – Dendrogramme

	$V_1$	$V_2$	$V_3$
$e_1$	7.5	3.5	5.5
$e_2$	9	4	5
$e_3$	5.5	8	4
$e_4$	7.5	8.5	1.5
$e_5$	4.5	9.5	2.5
$e_6$	6.5	9	3
$e_7$	7.5	2.5	4.5
$e_8$	8	2.5	5.5
$e_9$	7.5	3.5	4

TABLE 10 – Valeurs des individus

**Question [cha-4-2]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
. V1 V2 V3
1 7.5 3.5 5.5
2 9.0 4.0 5.0
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permettant de représenter la CAH à partir de D selon la stratégie indiquée ci-dessus : F P A B C

## CATALOGUE

**Question [cha-4-3]** Reportez dans la table 11 les distances de  $e_8$  à tous les autres points.

Faites de même pour  $e_9$ .

[F] [P] [A] [M] [N] [L]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$								
$e_9$								

TABLE 11 – Tableau des distances

**Question [cha-4-4]** Reportez dans la table 12 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 12.

[F] [P] [A] [M] [N] [L]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 12 – Ultramétrique associée

---

## Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 13. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut minimal.

**Question [cha-5-1]** ☺ Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- [ ]  $d_\infty(e_4, e_5) = 4$     [ ]  $d_\infty(e_4, e_8) = 4$     [ ]  $d_{\text{Man}}(e_7, e_5) = 10$   
 [ ]  $d_{\text{Man}}(e_6, e_3) = 5.5$     [ ]  $d^2(e_1, e_9) = 18$     [F]  $d^2(e_8, e_5) = 2$   
 [G] Aucune de ces réponses n'est correcte.

## CATALOGUE

Cluster Dendrogram

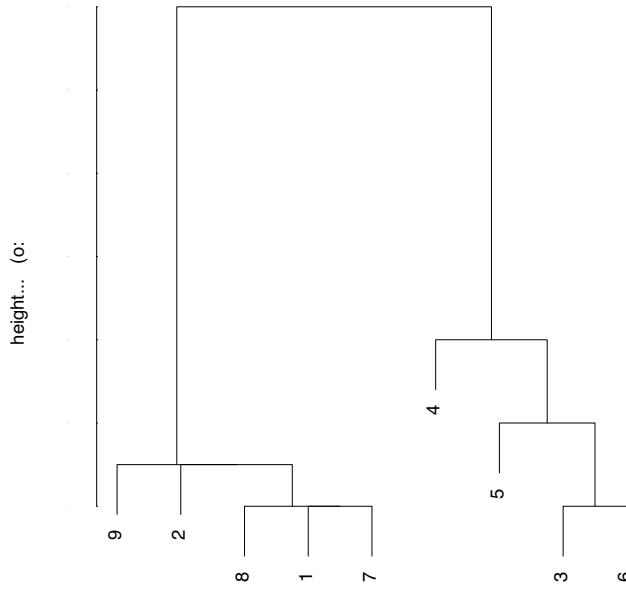


FIGURE 12 – Dendrogramme

	$V_1$	$V_2$	$V_3$
$e_1$	11	6	7
$e_2$	10.5	4.5	4.5
$e_3$	9.5	7.5	5
$e_4$	7	6.5	3
$e_5$	8.5	10.5	2.5
$e_6$	9.5	11	3
$e_7$	10	4.5	5
$e_8$	9.5	10.5	0.5
$e_9$	8	6	4

TABLE 13 – Valeurs des individus

**Question [cha-5-2]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
. V1 V2 V3
1 11.0 6.0 7.0
2 10.5 4.5 4.5
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permettant de représenter la CAH à partir de D selon la stratégie indiquée ci-dessus : F P A B C

## CATALOGUE

**Question [cha-5-3]** Reportez dans la table 14 les distances de  $e_8$  à tous les autres points.

Faites de même pour  $e_9$ .

[F] [P] [A] [M] [N] [L] 

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$								
$e_9$								

TABLE 14 – Tableau des distances

**Question [cha-5-4]** Reportez dans la table 15 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 13.

[F] [P] [A] [M] [N] [L] 

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 15 – Ultramétrique associée

---

## Exercice

---

Nous considérons un groupe de neuf individus, notés  $e_1$  à  $e_9$ , dont les données sont fournies par la table 16. Nous souhaitons réaliser une classification ascendante hiérarchique basée sur la distance de Tchebychev et le saut minimal.

**Question [cha-6-1]** ☺ Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- [A]  $d_\infty(e_4, e_5) = 1.5$     [B]  $d_\infty(e_4, e_8) = 1.5$     [C]  $d_{\text{Man}}(e_7, e_5) = 4.5$   
 [D]  $d_{\text{Man}}(e_6, e_3) = 4.5$     [E]  $d^2(e_1, e_9) = 11$     [F]  $d^2(e_8, e_5) = 6$   
 *Aucune de ces réponses n'est correcte.*

## CATALOGUE

Cluster Dendrogram

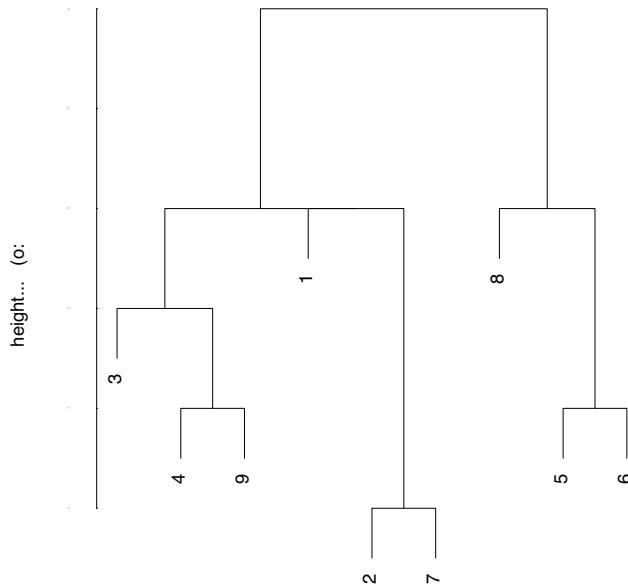


FIGURE 13 – Dendrogramme

	$V_1$	$V_2$	$V_3$
$e_1$	3	9	10.5
$e_2$	5	10	8
$e_3$	11	2	8
$e_4$	3	10	8.5
$e_5$	3	9	9.5
$e_6$	11.5	3	6
$e_7$	4	7.5	8
$e_8$	3	10.5	7.5
$e_9$	4.5	9.5	7.5

TABLE 16 – Valeurs des individus

**Question [cha-6-2]** Nous nous plaçons dans le contexte du logiciel R. Supposons que les données soient contenues dans le dataframe C :

```
> head(C, n=2L)
. V1 V2 V3
1 3 9 10.5
2 5 10 8
```

Proposez un ensemble de commandes permettant de placer dans une variable D la matrice des distances de Tchebychev entre les individus de C, de construire et d'afficher le dendrogramme permettant de représenter la CAH à partir de D selon la stratégie indiquée ci-dessus : F P A B C

## CATALOGUE

**Question [cha-6-3]** Reportez dans la table 17 les distances de  $e_8$  à tous les autres points.

Faites de même pour  $e_9$ .

[F] [P] [A] [M] [N] [L] [■]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_8$								
$e_9$								

TABLE 17 – Tableau des distances

**Question [cha-6-4]** Reportez dans la table 18 les valeurs de l'ultramétrique correspondant à la CAH construite. Pour accélérer vos calculs, vous pourrez vous aider du dendrogramme de la figure 14.

[F] [P] [A] [M] [N] [L] [■]

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_2$								
$e_3$								
$e_4$								
$e_5$								
$e_6$								
$e_7$								
$e_8$								
$e_9$								

TABLE 18 – Ultramétrique associée

---

### Quelques questions

---

## CATALOGUE

Cluster Dendrogram

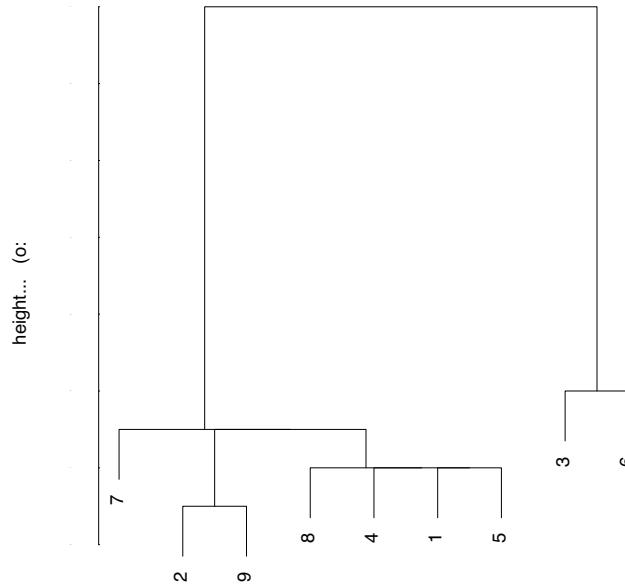


FIGURE 14 – Dendrogramme

**Question [theo1]** Explicitez deux critères pouvant être mobilisés afin de justifier le choix d'un nombre de classes lors d'une classification ascendante hiérarchique.

F  P  A  B

## CATALOGUE

**Question [theo2]** Explicitez deux critères pouvant être employés pour justifier le choix du nombre d'axes retenus lors d'une analyse factorielle.

F  P  A  B

**Question [theo3]** De quelle limite souffre la classification ascendante hiérarchique ? Décrivez une contre-mesure à cette limitation ?

F  P  A  B

**Question [theo4]** Quelles réserves doivent accompagner les résultats d'une classification obtenue à l'issue d'une procédure de type k-means ?

F  P  A  B

## CATALOGUE

**Question [theo5]** Décrivez une procédure permettant d'exploiter à la fois des variables quantitatives et des variables qualitatives sur une population, et cela dans un objectif de classification.

F	P	A	B	■
---	---	---	---	---

**Question [code1]** Considérons deux vecteurs de valeurs quantitatives A et B et de même taille dans R. Proposez un ensemble de commandes permettant de construire un vecteur de couleurs de taille identique et pour lequel la ième couleur est `green` si la ième valeur de A est supérieure à la ième valeur de B et `red` sinon.

F	P	A	B	■
---	---	---	---	---

## CATALOGUE

**Question [code2]** On souhaite réaliser une classification ascendante hiérarchique (avec la distance euclidienne et la stratégie de Ward) sur un dataframe D comportant 15 variables quantitatives, à l'issue de laquelle on souhaite en plus construire une représentation graphique de cette classification. Proposez un ensemble de commandes pour cela :

F  P  A  B

**Question [code3]** On dispose d'un dataframe T1 correspondant aux données du premier tour d'une élection, et comportant les variables `Abs` (le nombre d'abstentions), `B1N1` (le nombre de bulletins blancs ou nuls), `X` (le nombre de voix pour le candidat X) et `Y` (idem pour le candidat Y). On dispose d'un dataframe T2 contenant les données analogues pour le second tour de l'élection (`X` et `Y` sont bien présents aux deux tours). On souhaite construire un dataframe T contenant :

- la progression du nombre d'abstentions entre les deux tours ;
- les variables `X` et `Y` pour les deux tours ;
- un facteur contenant les valeurs `X_gagnant` et `Y_gagnant` selon que X est vainqueur au second tour ou non.

Proposez un ensemble de commande permettant de construire T :

F  P  A  B

## CATALOGUE

**Question [code4]** Nous disposons d'un objet `acp` contenant les résultats renvoyés par la fonction `dudi.pca` appliquée à des variables centrées et réduites. L'analyse a été limitée aux résultats des 4 premiers axes principaux. Proposez un ensemble de commandes permettant de rassembler la contribution à l'inertie totale et la contribution à l'inertie de chacune des quatre composantes principales pour chaque individu.

F  P  A  B

## CATALOGUE

**Annexe - Table de valeurs du  $\chi^2$** 

La table ci-dessous donne, en fonction de la valeur du nombre de degré de liberté (abrégé par ddl dans la suite) et en fonction de  $P$ , la valeur du  $\chi^2$  telle que la probabilité pour une variable aléatoire suivant une loi du  $\chi^2$  de dépasser cette valeur est  $P$ .

ddl	1 - $P$													
	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
1	0.000	0.000	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.51
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.603	3.053	3.816	4.575	5.578	7.584	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	3.074	3.571	4.404	5.226	6.304	8.438	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.565	4.107	5.009	5.892	7.041	9.299	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.075	4.660	5.629	6.571	7.790	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.601	5.229	6.262	7.261	8.547	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.142	5.812	6.908	7.962	9.312	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.697	6.408	7.564	8.672	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.265	7.015	8.231	9.390	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.844	7.633	8.907	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.434	8.260	9.591	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	8.034	8.897	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.643	9.542	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.260	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.886	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.65	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
31	14.46	15.66	17.54	19.28	21.43	25.39	30.34	35.89	41.42	44.99	48.23	52.19	55.00	61.10
32	15.13	16.36	18.29	20.07	22.27	26.30	31.34	36.97	42.58	46.19	49.48	53.49	56.33	62.49
33	15.82	17.07	19.05	20.87	23.11	27.22	32.34	38.06	43.75	47.40	50.73	54.78	57.65	63.87
34	16.50	17.79	19.81	21.66	23.95	28.14	33.34	39.14	44.90	48.60	51.97	56.06	58.96	65.25
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27	66.62
36	17.89	19.23	21.34	23.27	25.64	29.97	35.34	41.30	47.21	51.00	54.44	58.62	61.58	67.98
37	18.59	19.96	22.11	24.07	26.49	30.89	36.34	42.38	48.36	52.19	55.67	59.89	62.88	69.35
38	19.29	20.69	22.88	24.88	27.34	31.81	37.34	43.46	49.51	53.38	56.90	61.16	64.18	70.70
39	20.00	21.43	23.65	25.70	28.20	32.74	38.34	44.54	50.66	54.57	58.12	62.43	65.48	72.06
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.1	118.5	124.3	129.6	135.8	140.2	149.4
120	83.85	86.92	91.57	95.70	100.6	109.2	119.3	130.1	140.2	146.6	152.2	159.0	163.6	173.6
140	100.7	104.0	109.1	113.7	119.0	128.4	139.3	150.9	161.8	168.6	174.6	181.8	186.8	197.4
160	117.7	121.3	126.9	131.8	137.5	147.6	159.3	171.7	183.3	190.5	196.9	204.5	209.8	221.0
180	134.9	138.8	144.7	150.0	156.2	166.9	179.3	192.4	204.7	212.3	219.0	227.1	232.6	244.4
200	152.2	156.4	162.7	168.3	174.8	186.2	199.3	213.1	226.0	234.0	241.1	249.4	255.3	267.5
240	187.3	192.0	199.0	205.1	212.4	224.9	239.3	254.4	268.5	277.1	284.8	293.9	300.2	313.4
300	240.7	246.0	253.9	260.9	269.1	283.1	299.3	316.1	331.8	341.4	349.9	359.9	366.8	381.4
400	330.9	337.2	346.5	354.6	364.2	380.6	399.3	418.7	436.6	447.6	457.3	468.7	476.6	493.1

## CATALOGUE

---

## Annexe - Aide R - fonction dist

---

```

dist {stats} R Documentation

Distance Matrix Computation

Description
This function computes and returns the distance matrix
computed by using the specified distance measure to
compute the distances between the rows of a data
matrix.

Usage
dist(x, method = "euclidean", diag = FALSE, upper =
FALSE, p = 2)

as.dist(m, diag = FALSE, upper = FALSE)
## Default S3 method:
as.dist(m, diag = FALSE, upper = FALSE)

## S3 method for class 'dist'
print(x, diag = NULL, upper = NULL,
      digits = getOption("digits"), justify = "none",
      right = TRUE, ...)

## S3 method for class 'dist'
as.matrix(x, ...)

Arguments
x           a numeric matrix, data frame or "dist"
object.
method       the distance measure to be used. This
must be one of "euclidean", "maximum",
"manhattan", "canberra", "binary" or
"minkowski". Any unambiguous substring
can be given.
diag         logical value indicating whether the
diagonal of the distance matrix should
be printed by print.dist.
upper        logical value indicating whether the
upper triangle of the distance matrix
should be printed by print.dist.
p            The power of the Minkowski distance.
m            An object with distance information to
be converted to a "dist" object. For
the default method, a "dist" object, or
a matrix (of distances) or an object
which can be coerced to such a matrix
using as.matrix(). (Only the lower
triangle of the matrix is used, the
rest is ignored).
digits, justify passed to format inside of print().
right, ...    further arguments, passed to other
methods.

Details
Available distance measures are (written for two
vectors x and y):

euclidean:
  Usual distance between the two vectors (2 norm
  aka L_2), sqrt(sum((x_i - y_i)^2)).

maximum:
  Maximum distance between two components of x and
  y (supremum norm)

manhattan:
  Absolute distance between the two vectors (1 norm
  aka L_1).

canberra:
  sum(|x_i - y_i| / |x_i + y_i|). Terms with zero
  numerator and denominator are omitted from the
  sum and treated as if the values were missing.
  This is intended for non-negative values (e.g.,
  counts): taking the absolute value of the
  denominator is a 1998 R modification to avoid
  negative distances.

binary:
  (aka asymmetric binary): The vectors are regarded
  as binary bits, so non-zero elements are 'on' and
  zero elements are 'off'. The distance is the
  proportion of bits in which only one is on
  amongst those in which at least one is on.

minkowski:
  The p norm, the pth root of the sum of the pth
  powers of the differences of the components.

Missing values are allowed, and are excluded from all
computations involving the rows within which they
occur. Further, when Inf values are involved, all pairs
of values are excluded when their contribution to the
distance gave NaN or NA. If some columns are excluded
in calculating a Euclidean, Manhattan, Canberra or
Minkowski distance, the sum is scaled up proportionally
to the number of columns used. If all pairs are
excluded when calculating a particular distance, the
value is NA.

The "dist" method of as.matrix() and as.dist() can be
used for conversion between objects of class "dist" and
conventional distance matrices.

as.dist() is a generic function. Its default method
handles objects inheriting from class "dist", or
coercible to matrices using as.matrix(). Support for
classes representing distances (also known as
dissimilarities) can be added by providing an as.matrix
() or, more directly, an as.dist method for such a
class.

Value
dist returns an object of class "dist".
The lower triangle of the distance matrix stored by
columns in a vector, say do. If n is the number of
observations, i.e., n <- attr(do, "Size"), then for i <
j <= n, the dissimilarity between (row) i and j is do
[n*(i-1) - i*(i-1)/2 + j-i]. The length of the vector
is n*(n-1)/2, i.e., of order n^2.
The object has the following attributes (besides
"class" equal to "dist"):

Size      integer, the number of observations in the
dataset.
Labels    optionally, contains the labels, if any, of
the observations of the dataset.
Diag, Upper logicals corresponding to the arguments
diag and upper above, specifying how the
object should be printed.
call      optionally, the call used to create the
object.
method    optionally, the distance method used;
resulting from dist(), the (match.arg()ed)
method argument.

References
[snip...snip]

-----
[Package stats version 3.4.0 Index]

```