Exercice pour le Mooc Fondamentaux en statistique

Avner Bar-Hen

Vous avez à votre disposition un jeu de données disponible dans un fichier Excel.

Lorsque vous avez terminé l'exercice vous déposez votre/vos documents dans l'espace réservé aux dépôts des productions (un par semaine).

C'est également à partir de cet espace de dépôt que vous pourrez analyser et évaluer les productions de quelques pairs.

La date limite de dépôt est fixé au **jeudi 13 février 2014 à 8h**

Ce fichier va vous suivre pendant l'ensemble de ce cours. À l'issue des cinq semaines, vous aurez réalisé une analyse de données de votre tableau de données

Cette semaine vous avez les l'énoncés et corrigés des semaine 1, 2, 3 ainsi que l'énoncé de la semaine 4.

Bon courage et en avant !

Semaine 1 : Comment résumer l'information d'une variable

Énoncé de l'exercice

En quelques lignes vous décrirez le fichier de données **Geraud.xls** (nombre d'observations, nombre de variables, nombre de modalités). Pour chacune des variables vous calculerez les statistiques descriptives et vous proposerez une représentation graphique.

Merci de ne pas dépasser une page pour le rendu de cet exercice.

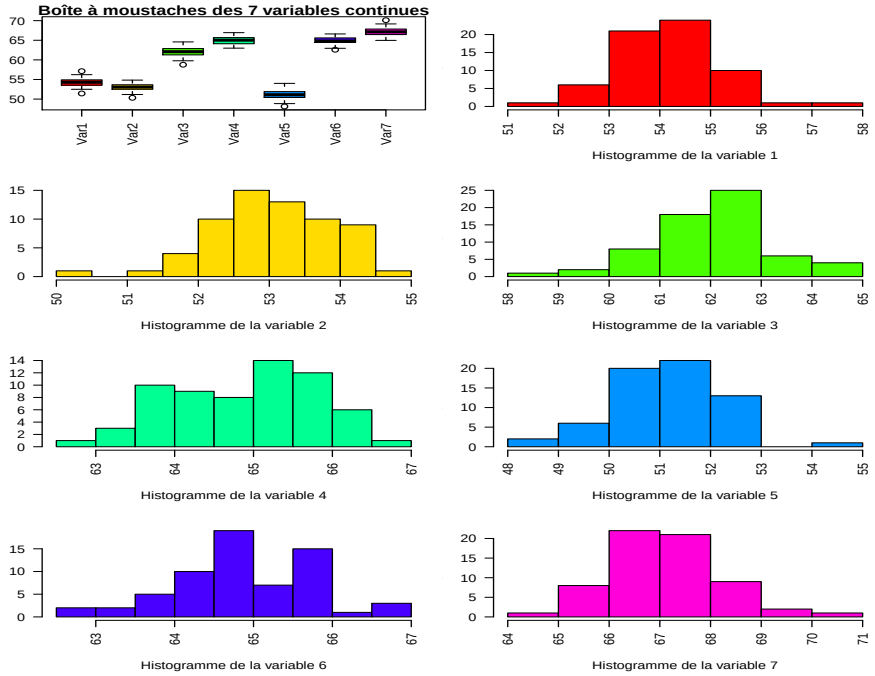
Solution

Le fichier Excel Geraud.xls contient 64 observations et 7 variables continues (dénommées Var1,. . .,Var7).
Le tableau ci-dessous donne les statistiques descriptives pour chacune des 7 variables continues.

	Moyenne	Médiane	Variance	Écart-Type	1er Quartile	3ème quartile	Min	Max
Var1	54.23	54.30	0.99	0.99	53.60	54.91	51.43	57.13
Var2	53.04	53.07	0.73	0.85	52.51	53.65	50.33	54.84
Var3	62.05	62.09	1.34	1.16	61.28	62.87	58.76	64.58
Var4	64.92	65.03	0.87	0.93	64.14	65.67	62.99	66.94
Var5	51.15	51.11	1.18	1.09	50.41	51.90	48.13	54.01
Var6	64.89	64.81	0.73	0.85	64.44	65.60	62.58	66.61
Var7	67.15	67.15	1.08	1.04	66.47	67.85	64.97	70.12

TABLE 1 – Statistiques descriptives des variables continues

On peut donc représenter les histogrammes pour voir la distribution des différentes variables et les boîtes à moustaches pour comparer les caractéristiques de chacune des variables.



On note que les moyennes et les médianes sont proches pour chacune des variables. On rappelle que l'égalité de la moyenne et de la médiane est une indication de la symétrie de la distribution des données de la variable étudiée.

Le tableau ci-dessous donne la liste des observations en dehors des moustaches.

Variable	Individu
Var1	Ind8, Ind19
Var2	Ind30
Var3	Ind40
Var5	Ind55
Var6	Ind6
Var7	Ind40

À première vue il ne semble pas y avoir de données aberrantes.

Les variables (Var4, Var6) ont des moyennes proches.

Les variances des 7 variables sont entre 0.73 et 1.34.

Semaine 2 : Comment résumer l'information d'une variable

Le but de l'exercice de cette semaine est de comprendre les résultats d'une analyse en composantes principales de votre fichier de données. Comme nous n'avons ni le volume horaire, ni les outils techniques pour inclure un enseignement logciel dans ce cours, je vous propose un ensemble de sorties que vous devrez analyser.

Énoncé de l'exercice

Dans un premier temps, vous calculerez les corrélations entre les variables (que vous pourrez compléter par des graphiques).

Si vous utilisez Excel, la fonction COEFFICIENT.CORRELATION devrait vous aider mais vous êtes libre d'utiliser le logiciel que vous souhaitez.

Les résultats devront être commentés.

Cette partie ne devra pas faire plus d'une page.

Ci-dessous vous avez les sorties d'une analyse en composantes principales.

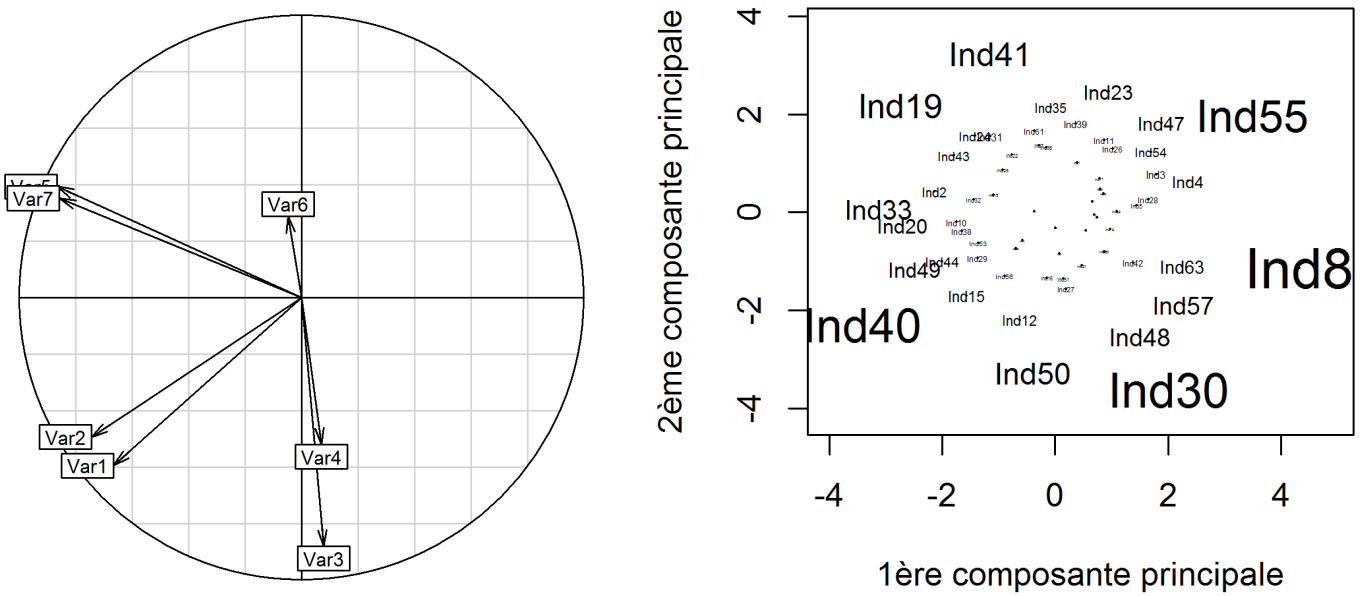
Les coordonnées des individus dans le plan (1,2) de l'ACP est dans la feuille "ACP" du fichier Geraud.xls

```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5    PC6    PC7
Standard deviation     1.582 1.414 1.223 0.7095 0.4661 0.4221 0.3206
Proportion of Variance 0.357 0.286 0.214 0.0720 0.0310 0.0250 0.0150
Cumulative Proportion  0.357 0.643 0.857 0.9290 0.9600 0.9850 1.0000


       PC1   PC2   PC3   PC4   PC5   PC6   PC7
Var1 -0.42  0.42 -0.21  0.33 -0.02  0.66  0.24
Var2 -0.47  0.35 -0.09 -0.43  0.61 -0.29 -0.06
Var3  0.05  0.62 -0.26  0.21 -0.46 -0.50 -0.19
Var4  0.04  0.37  0.60 -0.55 -0.36  0.26 -0.04
Var5 -0.55 -0.28  0.12  0.12 -0.20  0.06 -0.74
Var6 -0.03 -0.20 -0.69 -0.58 -0.34  0.17  0.00
Var7 -0.54 -0.25  0.18  0.01 -0.36 -0.36  0.59
```

Corrélation entre les variables d'origine et les composantes principales

```
       PC1   PC2   PC3   PC4   PC5   PC6   PC7
Var1 -0.67  0.59 -0.25  0.23 -0.01  0.28  0.08
Var2 -0.74  0.49 -0.11 -0.31  0.28 -0.12 -0.02
Var3  0.08  0.88 -0.32  0.15 -0.22 -0.21 -0.06
Var4  0.07  0.52  0.73 -0.39 -0.17  0.11 -0.01
Var5 -0.87 -0.39  0.14  0.09 -0.10  0.03 -0.24
Var6 -0.05 -0.29 -0.85 -0.41 -0.16  0.07  0.00
Var7 -0.86 -0.35  0.22  0.00 -0.17 -0.15  0.19
```



Dans le graphique de ci-dessus, la taille des individus est proportionnelle à la qualité de la représentation.

Combien d'axes doit-on garder ?

Dans le plan (1,2), les individus (Ind33,Ind20) sont-ils proches ?

Dans le plan (1,2), les individus (Ind64,Ind14) sont-ils proches ?

Que peut-on dire des individus (Ind55,Ind40) dans le plan (1,2) ?

Quelle est la variable la mieux représentée sur l'axe 1 ?

Quelle est la variable la moins bien représentée sur l'axe 2 ?

Quelle est la variable la mieux représentée dans le plan (1,2) ?

À partir du cercle de corrélation, que peut-on dire du lien entre les variables dans le plan (1,2) ?

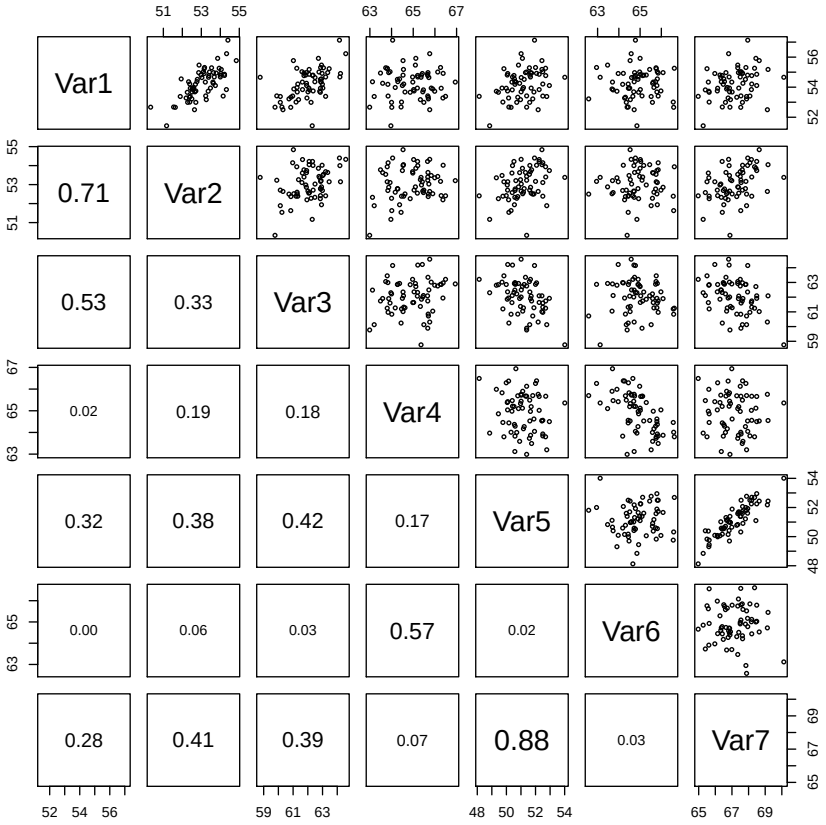
Question bonus : Détailler les calculs permettant d'obtenir les coordonnées de l'individu 8 ?

Merci de ne pas dépasser deux pages pour l'ensemble du travail de la semaine.

Solution

Le graphique ci-dessous donne les nuages de points deux à deux (dessus de la diagonale) et les corrélations deux à deux (dessous de la diagonale).

Afin de rendre les résultats plus lisibles, les fontes pour écrire les corrélations sont proportionnelles à leur valeur (plus la corrélation est grande et plus le chiffre écrit est grand)



Si l'on ne considère que les corrélations plus grandes (en valeur absolue) que 0.7, on note que l'on a 5 groupes de variables. Les groupes sont (Var1,Var2) (Var3) (Var4) (Var5,Var7) (Var6)

Passons maintenant à l'ACP. Comme nous n'avons pas d'indication sur les unités de mesure, il est préférable d'effectuer l'ACP sur la matrice de corrélation. Ceci revient donc à standardiser les variables.
La première question concerne en général le nombre d'axes à garder. Si nous appliquons la règle de Kaiser, nous conservons 3 axes car 3 valeurs propres sont supérieures à 1.

Pour ce corrigé, nous gardons les deux premiers axes.

Attention : ici ce sont les écarts-types qui sont présentés. Pour retrouver le pourcentage de variance il faut utiliser les carrés des écarts-types. Par exemple le pourcentage de variance de l'axe 1 est donné par la formule

$$\frac{1.58^2}{1.58^2 + 1.41^2 + 1.22^2 + 0.71^2 + 0.47^2 + 0.42^2 + 0.32^2} = 0.36$$

De la même manière le pourcentage de variance expliquée par les deux premiers axes vaut :

$$\frac{1.58^2 + 1.41^2}{1.58^2 + 1.41^2 + 1.22^2 + 0.71^2 + 0.47^2 + 0.42^2 + 0.32^2} = 0.64$$

Les 3 premiers axes représentent 85.69% de la variance du jeu de données d'origine.

Regardons maintenant les observations. Les coordonnées de l'observation 8 (ligne 8 du fichier de données) sont

```
 Var1  Var2  Var3  Var4  Var5  Var6  Var7
51.43 51.17 62.30 63.97 48.85 64.85 65.27
```

On voit aussi sur le graphique que les données sont centrées (afin d'avoir le point (0,0) au centre) et réduites (c'est à dire que la variance de chaque variable vaut 1) car l'ACP est réalisée avec la matrice de corrélation.

Concrètement pour chaque observation x de la variable V_i, on calcule $\frac{x-\text{moy}(V_i)}{\sqrt{\text{Var}(V_i)}}$
Les coordonnées centrées réduites de l'observation 8 sont

```
 Var1  Var2  Var3  Var4  Var5  Var6  Var7
-2.82 -2.20  0.22 -1.02 -2.11 -0.04 -1.81
```

Pour obtenir les coordonnés de cette observation dans le nouveau repère on utilise la première valeur propre (PC1)
(-0.42)*(-2.82)+(-0.47)*(-2.2)+(0.05)*(0.22)+(0.04)*(-1.02)+(-0.55)*(-2.11)+(-0.03)*(-0.04)+(-0.54)*(-1.81)=4.33

On peut évidemment fait la même chose sur le deuxième axe en utilisant la deuxième valeur propre (PC2)
(0.42)*(-2.82)+(0.35)*(-2.2)+(0.62)*(0.22)+(0.37)*(-1.02)+(-0.28)*(-2.11)+(-0.2)*(-0.04)+(-0.25)*(-1.81)=-1.14

Les coordonnées de l'individu 8 sont donc $(4.33, -1.14)$

Dans le plan (1,2), les individus (Ind33,Ind20) sont proches et bien représentés. On peut donc conclure qu'ils sont proches. Par contre dans le plan (1,2), les individus (Ind64,Ind14) sont proches mais mal représentés. On ne peut donc rien conclure sur leur proximité.

Dans le plan (1,2), les individus (Ind55,Ind40) sont loins et nous pouvons conclure qu'ils sont très différents (pour les variables mesurées). Ici ces individus sont bien représentés (et donc plus simples à voir) mais nous n'avons pas besoin de leur qualité de représentation pour conclure à leur éloignement.

Le cercle de corrélation permet de tirer l'étude des variables.

La variable la mieux représentée sur l'axe 1 (axe horizontal) est celle avec la plus grande coordonnée (en valeur absolue) sur l'axe 1. C'est-à-dire la variable 5.

De même la variable la moins bien représentée sur l'axe 2 (axe vertical) est celle avec la plus petite coordonnées (en valeur absolue) sur l'axe 2. C'est-à-dire la variable 6.

La variable la mieux représentée est la variable qui est la plus proche du cercle unité. C'est-à-dire la variable 5. Dans le plan (1,2), $(-0.87)^2 + (-0.39)^2 = 90.6\%$ de la variabilité de cette variable est représentée.

Les variables Var1, Var2, Var3, Var5, Var7 sont bien représentées dans le plan (1,2)

Les couples de variables (Var1, Var2) et (Var5, Var7) ont un angle presque nul (donc corrélation proche de 1)

Semaine 3 : Apprentissage/Classification

Énoncé de l'exercice.

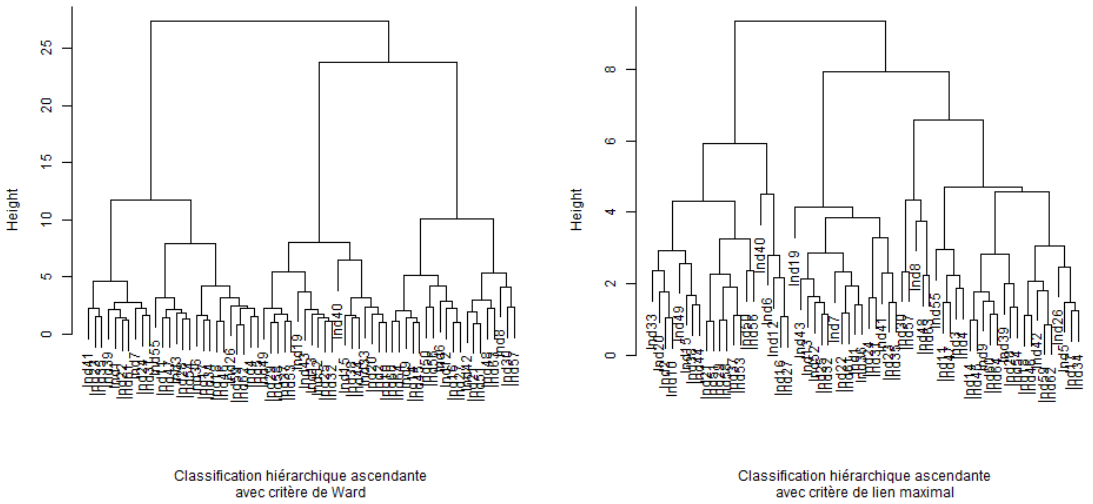
Première partie : La matrice ci-dessous est une matrice de distance Z obtenue en utilisant la distance euclidienne sur 5 observations, notées A, B, C,D, E.

Réaliser une classification hiérarchique ascendante en utilisant d'abord le critère du lien maximal puis en utilisant le critère du saut minimal. Vous contruirez d'abord la matrice de distance pour chacune des étapes de la classification puis vous dessinerez les deux dendrogrammes correspondants. Enfin vous commenterez les résultats obtenus.

Cette partie ne doit pas faire plus d'une page

$$Z = \begin{pmatrix} 0 & 57 & 18 & 36 & 33 \\ 57 & 0 & 51 & 31 & 40 \\ 18 & 51 & 0 & 50 & 43 \\ 36 & 31 & 50 & 0 & 29 \\ 33 & 40 & 43 & 29 & 0 \end{pmatrix}$$

Deuxième partie : Vous avez ci-dessous les résultats de la classification hiérarchique ascendante réalisée sur votre jeu de données avec les critères du lien maximal et du saut minimal



Combien de groupes avez vous envie de choisir ?

L'onglet classif du fichier Geraud.xls regroupe les affectations des individus. Chaque colonne correspond au résultat d'une classification hiérarchique ascendante pour un critère (Max : lien maximal ; Ward : critère de Ward) et un nombre de groupes allant de 2 à 6. Par exemple la colonne Max4 correspond donc à une classification hiérarchique ascendante pour le critère du lien maximal et un niveau de coupure donnant 4 groupes

Combien d'individus par classe avez vous (pour le nombre de groupes choisi précédemment) ?

À l'aide de statistiques descriptives vous décrirez vos classes.

Cette partie ne doit pas faire plus de deux pages.

Solution

Première partie :
Le principe général de la méthode est de regrouper itérativement les deux lignes et colonnes correspondant aux individus ou groupes les plus proches. La nouvelle ligne (respectivement colonne) est obtenue par fusion des deux lignes respectivement colonne) en utilisant le critère.

Pour le critère du saut minimal on remplace par la plus petite valeur

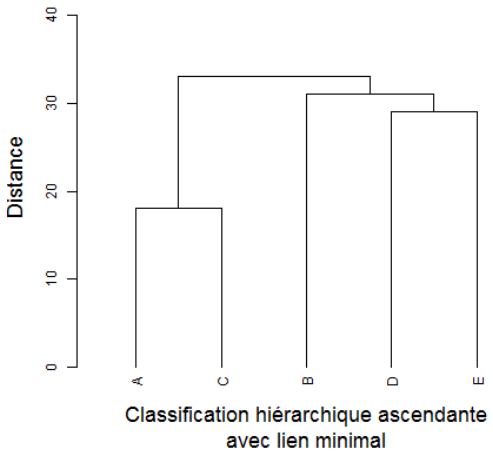
Étape 0	Étape 1	Étape 2	Étape 3
$\begin{pmatrix} 0 & 57 & 18 & 36 & 33 \\ 57 & 0 & 51 & 31 & 40 \\ 18 & 51 & 0 & 50 & 43 \\ 36 & 31 & 50 & 0 & 29 \\ 33 & 40 & 43 & 29 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 51 & 36 & 33 \\ 51 & 0 & 31 & 40 \\ 36 & 31 & 0 & 29 \\ 33 & 40 & 29 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 51 & 33 \\ 51 & 0 & 31 \\ 33 & 31 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 33 \\ 33 & 0 \end{pmatrix}$
Variable A,B,C,D,E	Variable {A,C},B,D,E	Variable {A,C},B,{D,E}	Variable {A,C},{B,{D,E}}

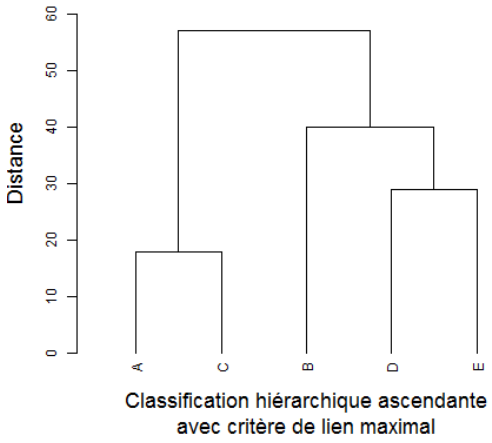
Pour le lien maximal on remplace par la plus grande valeur

Étape 0	Étape 1	Étape 2	Étape 3
$\begin{pmatrix} 0 & 57 & 18 & 36 & 33 \\ 57 & 0 & 51 & 31 & 40 \\ 18 & 51 & 0 & 50 & 43 \\ 36 & 31 & 50 & 0 & 29 \\ 33 & 40 & 43 & 29 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 57 & 50 & 43 \\ 57 & 0 & 31 & 40 \\ 50 & 31 & 0 & 29 \\ 43 & 40 & 29 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 57 & 50 \\ 57 & 0 & 40 \\ 50 & 40 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 57 \\ 57 & 0 \end{pmatrix}$
Variable A,B,C,D,E	Variable {A,C},B,D,E	Variable {A,C},B,{D,E}	Variable {A,C},{B,{D,E}}

Quelque soit le critère, la première fusion est réalisée sur la même paire d'observations.

Et il suffit donc de dessiner les arbres



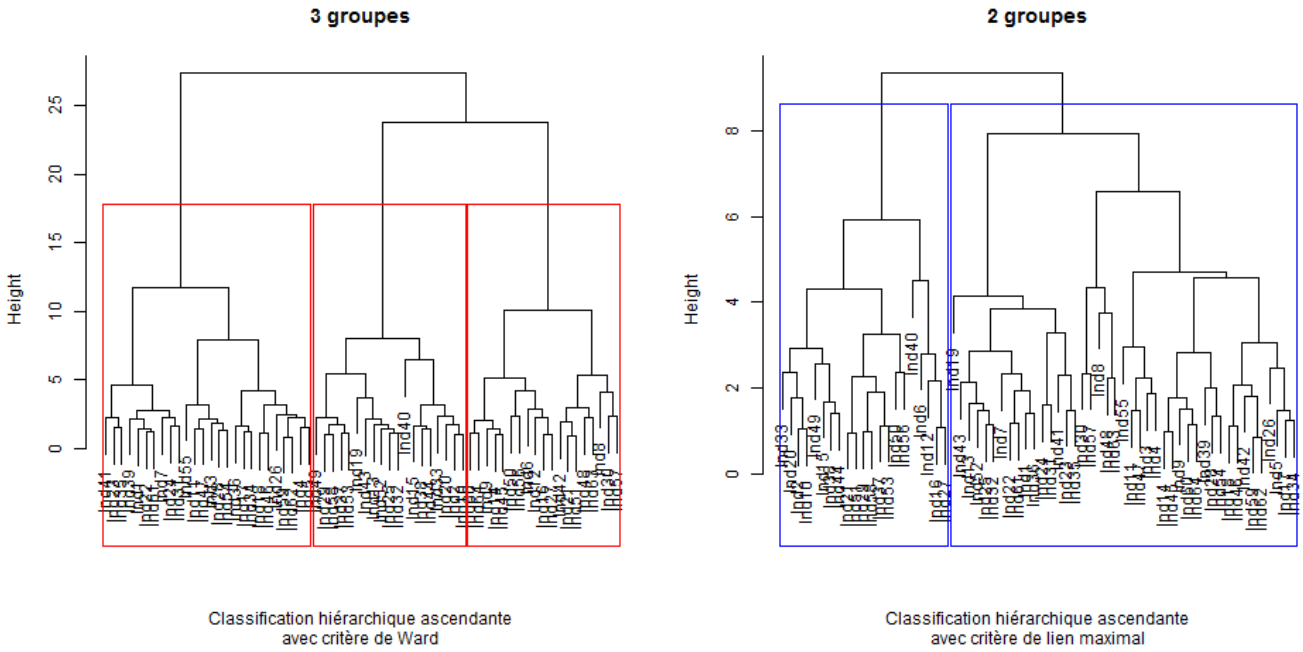


Deuxième partie :
Si l'on regarde les résultats de la classification hiérarchique ascendante, il est tentant de couper à l'endroit de plus grand saut.

Nous avons donc :
– Pour un critère du lien maximal : 2 groupes
– Pour un critère de Ward : 3 groupes

On voit que le nombre groupe est différent pour les deux critères.

Nous pouvons visualiser ces résultats :



Regardons donc les effectifs par groupes :

(a) Critère de Ward

Max	Groupe = 1	Groupe = 2
nombre d'observations	43	21

(b) Lien maximal

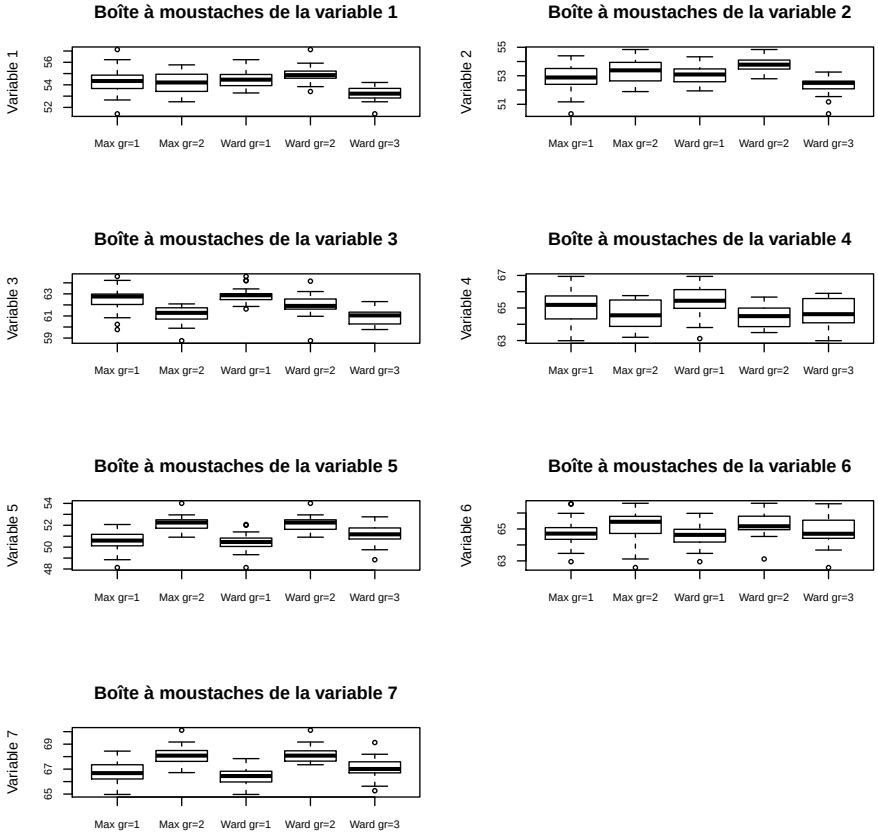
Ward	Groupe = 1	Groupe = 2	Groupe = 3
nombre d'observations	26	19	19

Pour étudier la cohérence entre les classifications, nous pouvons construire le tableau croisé entre les classifications obtenues avec les deux critères

	Max1	Max2
Ward1	26	0
Ward2	6	13
Ward3	11	8

Les observations du groupe Ward1 sont dans le groupe Max1

Regardons les boîtes à moustaches pour les variables au sein chaque groupe.



Il est possible aussi d'effectuer les k-means sur notre jeu de données. Au vu des résultats précédents il est tentant de regarder les résultats pour 2, 3, 4 groupes.

Pour filtrer le bruit et visualiser les résultats, nous allons effectuer cette classification sur les coordonnées des deux premiers axes de l'analyse en composantes principales réalisée la semaine dernière.

Les résultats sont présentés sous forme d'animation. N'hésitez pas à jouer avec les boutons de contrôle.

2 groupes 3 groupes 4 groupes

Semaine 4 : Tests non paramétriques

Énoncé de l'exercice

Première partie
La table ci-dessous donne les valeurs pour deux échantillons A et B.

	1	2	3	4	5	6
A	8	9	4	8	8	
B	7	2	5	7	5	3

Calculer le test de Wilcoxon sur ces données. Au seuil 5% peut-on conclure à la différence entre la distribution des deux échantillons ?

Cette partie de doit pas faire plus d'une demi-page.

Deuxième partie
La distribution des variables 4 et 6 est-elle significativement différente au seuil 5% ?

La distribution des variables 3 et 5 est-elle significativement différente au seuil 5%.

Préciser si vous considérez les données comme appariées ou non. Expliquer (rapidement) votre choix.

Commenter vos résultats.

Cette partie de doit pas faire plus d'une demi-page