MAITRISE BPE - UV MIAB 2 - JUIN 2002 - STATISTIQUES (2 HEURES)

1. Le vecteur $\mathbf{u} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$ est-il un vecteur propre de la matrice $\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$?

2. Existe-t-il des nombres réels a, b et c pour lesquels $\mathbf{A} = \begin{bmatrix} a & b & c \\ b & a & d \\ c & d & a \end{bmatrix}$ n'est pas diagonalisable ?

3. La somme d'une matrice carrée réelle et de sa transposée possède-t-elle toujours une base de vecteurs propres orthogonaux ?

4. Donner une matrice 2×2 réelle n'ayant aucun vecteur propre.

5. Soit **A** une matrice $p \times p$ réelle diagonalisable de rang r. Quel est le rang de $\mathbf{A}^2$?

Un expérimentateur avisé désire se faire une opinion du comportement de l'analyse des correspondances sur des tableaux de données artificielles. Il considère 3 tableaux comportant n = 16 lignes et p = 16 colonnes. Le premier est appelé **talea** car il a été généré par une procédure de tirage aléatoire. Le second est appelé **tgrad** car il représente une structure simple définie par un gradient.

```
> talea
   a b c d e f g h i j k l m n o p
1  1 0 0 1 1 0 1 0 0 0 0 0 0 0 1 1
2  1 1 1 1 1 0 1 0 0 1 1 0 1 1 1 1
3  1 1 1 1 0 0 0 1 0 0 1 0 0 1 1 0
4  1 1 1 0 0 1 0 1 0 1 0 0 0 1 1 0
5  1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0
6  1 0 0 0 0 0 1 0 1 0 1 1 1 0 1 1
7  1 0 1 1 0 0 0 1 1 0 0 0 1 0 1 1
8  0 1 0 0 1 1 1 1 0 0 0 0 0 1 1 1
9  0 0 1 0 0 0 0 1 1 1 0 1 1 0 0 1
10 0 1 0 0 1 1 0 1 0 1 1 1 1 1 1 0
11 0 0 1 0 1 0 0 0 1 0 1 0 0 0 0 1
12 0 0 1 0 1 0 1 0 0 1 0 1 0 1 1 1
13 1 1 1 0 1 1 0 0 1 0 0 0 1 0 0 0
14 0 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1
15 1 1 1 1 0 1 1 1 0 0 0 0 1 1 0 0
16 0 0 1 1 1 0 1 0 0 0 1 0 1 1 0 1
```

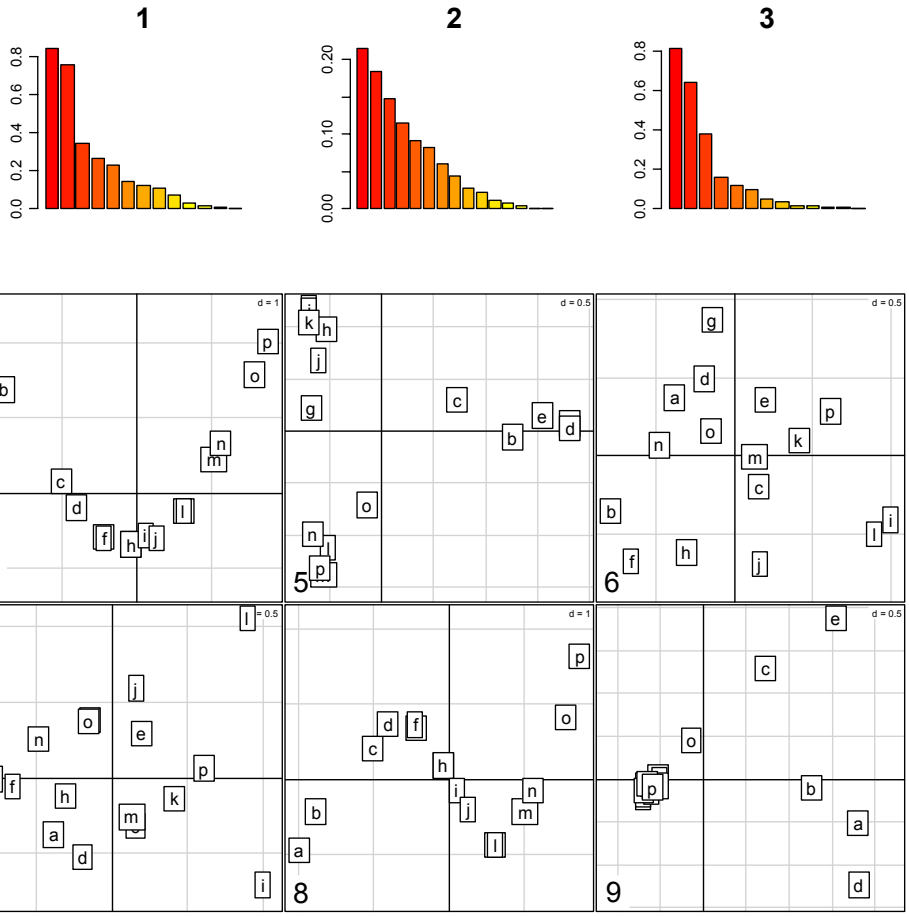
```
> tgrad
   a b c d e f g h i j k l m n o p
1  1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2  1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
3  1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
4  0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
5  0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0
6  0 0 1 1 0 1 1 1 1 0 0 0 0 0 0 0
7  0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0
8  0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
9  0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0
10 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0
11 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 0
12 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0
13 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0
14 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0
15 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
16 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
```

Le troisième est appelé **tparti** car il représente une structure simple définie par une partition.

```
> tparti
   a b c d e f g h i j k l m n o p
1  0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
2  1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
3  0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 0
4  1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
5  0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0
6  0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
7  0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0
8  0 0 1 0 0 1 1 1 1 1 1 0 0 1 0 0
9  0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 0
10 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0
11 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0
12 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
13 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
14 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1
15 0 1 0 0 0 0 0 0 0 1 0 1 1 0 1 1
```

```
16 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1
```

On pourra penser que les tableaux simulent la présence-absence de 16 espèces (colonnes) sur un échantillon de 16 sites (lignes). Pour les trois analyses des correspondances on garde, dans un ordre arbitraire, le graphe des valeurs propres, la carte f1-f2 des colonnes et la carte f1-f3 des colonnes :



6. Quelles sont les trois figures qui appartiennent à l'analyse de talea ?

7. Quelles sont les trois figures qui appartiennent à l'analyse de tparti ?

8. Quelles sont les trois figures qui appartiennent à l'analyse de tgrad ?

9. Quelle est la corrélation de chacun des nuages de points ?

10. Donner la plus grande valeur propre de l'analyse des correspondances de $\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$.

11. Donner les coordonnées des lignes et des colonnes sur le premier axe de la même analyse.

Une enquête d'opinion (période 1970-1980) portant sur 1000 personnes contient entre autres la question "A quelle famille politique vous rattachez-vous ?" avec 5 modalités de réponse (EG : extrême gauche; G : gauche; C : centre, D : droite et ED : extrême droite) et la question

"Que pensez-vous de l'importance des syndicats ?" avec 4 modalités de réponse (1- importance trop grande; 2- importance convenable; 3- importance insuffisante; 4- sans opinion). La répartition des réponses est consignée dans la table de contingence :

	1	2	3	4
EG	10	31	85	4
G	80	90	118	52
C	55	58	31	36
D	95	38	46	71
ED	40	23	10	27

```
> syndicats
          trop convenable insuffisant nesaispas
extgauche   10          31          85         4
gauche      80          90         118        52
centre      55          58          31        36
droite      95          38          46        71
extdroite   40          23          10        27
> sum(syndicats)
[1] 1000
```

L'analyse des correspondances de ce tableau donne :

```
$tab (le tableau modifié)
              trop convenable insuffisant nesaispas
extgauche -0.72527   -0.00641      1.2546  -0.83806
gauche    -0.15966    0.10294      0.1968  -0.19505
centre     0.09127    0.34259     -0.4061   0.05263
droite     0.35714   -0.36667     -0.3655   0.49474
extdroite  0.42857   -0.04167     -0.6552   0.42105

$cw (les poids des colonnes)
       trop  convenable insuffisant   nesaispas
       0.28        0.24        0.29        0.19

$lw (les poids des lignes)
extgauche    gauche    centre    droite extdroite
     0.13      0.34      0.18      0.25      0.10

$eig (les valeurs propres)
[1] 1.614e-01 1.747e-02 7.105e-05

$nf (le nombre d'axes conservés)
[1] 2

$c1 (les scores normés des colonnes)
                CS1     CS2
trop         0.8873  0.3051
convenable  -0.1711 -1.7549
insuffisant -1.3975  0.7034
nesaispas    1.0416  0.6933

$l1 (les scores normés des lignes)
              RS1     RS2
extgauche -2.1266  0.6529
gauche    -0.4039 -0.3220
centre     0.4571 -1.6072
droite     0.8709  1.3283
extdroite  1.1377 -0.1818

$co (les coordonnées des colonnes)
              Comp1    Comp2
trop         0.35645  0.04033
convenable  -0.06875 -0.23193
insuffisant -0.56140  0.09297
nesaispas    0.41841  0.09163

$li (les coordonnées des lignes)
            Axis1    Axis2
extgauche -0.8543  0.08628
```

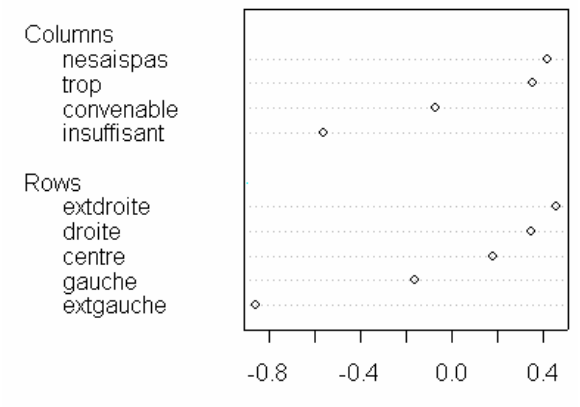
```
gauche    -0.1622 -0.04255
centre     0.1836 -0.21241
droite     0.3498  0.17555
extdroite  0.4570 -0.02403

$N (la somme totale du tableau de départ)
[1] 1000
```

12. Quel est le rang du tableau modifié ?

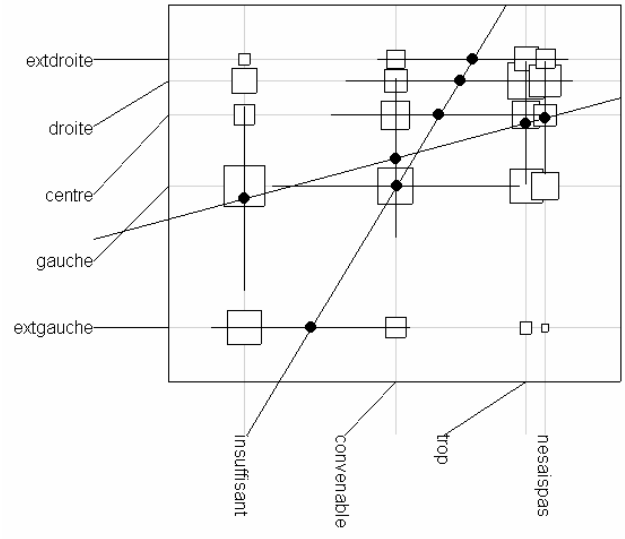
13. Quel est le Khi2 de la table de contingence ?

14. Que vaut la corrélation entre la sensibilité politique et l'opinion sur les syndicats ?

15. Comment est construite cette figure ?



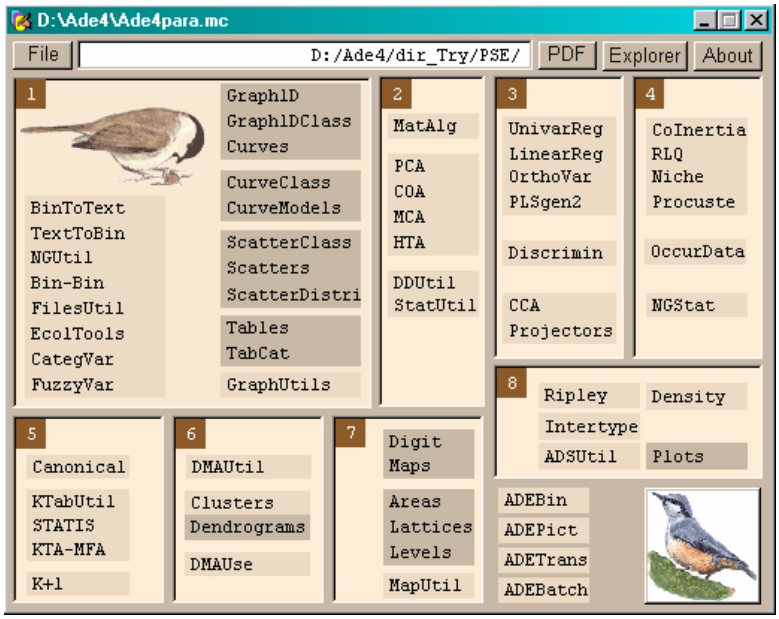
16. Donner une légende pour cette figure :



17. Placer sur la figure les étiquettes manquantes.

18. Interpréter les résultats.

19. La carte d'entrée du logiciel ADE-4 affiche la liste des modules. Indiquer très simplement ce qu'on peut faire avec 5 d'entre eux.



20. A votre avis, quelles sont les limites des méthodes multivariées pour l'analyse des grands tableaux (au delà de 10 000 lignes et/ou 1000 colonnes)

MAITRISE BPE - UV MIAB 2 -STATISTIQUES (2 HEURES)

NOM : **PRENOM :**

Question 1 Le vecteur **u** …

Question 2 Existe-t-il des nombres *a, b* et *c* …

Question 3 La somme d'une matrice …

Question 4 Donner une matrice 2×2 …

Question 5 Quel est le rang de …

Question 6 Les trois figures de l'analyse de talea

Question 7 Les trois figures de l'analyse de tparti

Question 8 Les trois figures de l'analyse de tgrad

Question 9 Quelle est la corrélation de chacun des nuages …

Question 10 Donner la plus grande valeur propre …

Question 11 Donner les coordonnées …

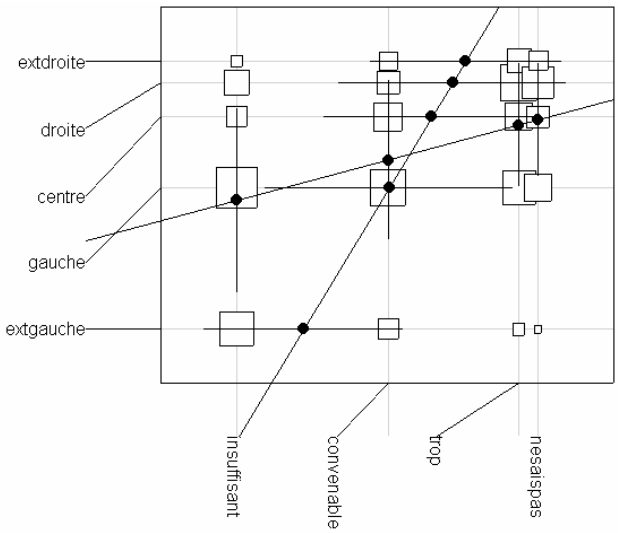
Question 12 Quel est le rang du tableau …

Question 13 Quel est le Khi2 de la table de contingence

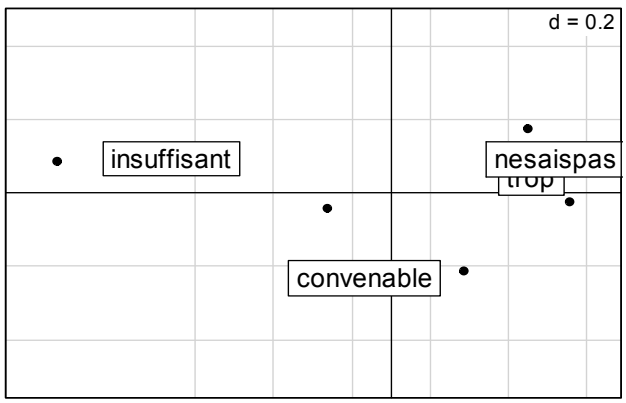
Question 14 Que vaut la corrélation …

Question 15 Comment est construite la figure ?

Question 16 Donner une légende



Question 17 Placer les étiquettes :



Question 18 Interpréter

Question 19 Ce qu'on peut faire avec les modules d'ADE-4

Question 20 Quelles sont les limites ?

SOLUTION

Question 1 Le vecteur **u** …

OUI car $\mathbf{Du} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ -2 \\ 2 \end{bmatrix} = 2\mathbf{u}$. 2 est valeur propre.

Question 2 Existe-t-il des nombres *a, b* et *c* …

Non, la matrice est toujours diagonalisable parce qu'elle est symétrique.

Question 3 La somme d'une matrice …

Oui, car elle est symétrique puisque $\left(\mathbf{A}+\mathbf{A}^t\right)^t = \mathbf{A}^t + \mathbf{A}^{tt} = \mathbf{A}^t + \mathbf{A} = \mathbf{A} + \mathbf{A}^t$. Les matrices symétriques sont diagonalisables et ont au moins une base de vecteurs propres orthogonaux.

Question 4 Donner une matrice 2×2 …

La matrice d'une rotation convient : $\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$

Question 5 Quel est le rang de …

Les deux matrices ont même sous-espace propre ($\mathbf{Au} = \lambda\mathbf{u} \Rightarrow \mathbf{A}^2\mathbf{u} = \lambda\mathbf{Au} = \lambda^2\mathbf{u}$) donc même noyau donc même rang.

Question 6 Les trois figures de l'analyse de talea

2/6/7 Le graphe des valeurs propres, caractéristique d'une absence de structure est 2 (noter aussi la faible valeur 0.2 contre 0.8 pour les autres). Les deux cartes factorielles qui ne font ni partition ni ordination sont 6 et 7.

Question 7 Les trois figures de l'analyse de tparti

1/5/9 La partition explicite avec 3 groupes visibles sur le tableau impose 5. Le lien par l'axe des x renvoie à 9. L'axe des y de la figure 9 n'indique aucune cohérence. La structure est exprimée sur deux axes auxquels correspondent deux valeurs propres. On y associe donc la figure 1.

Question 8 Les trois figures de l'analyse de tgrad

3/4/8 Par élimination en confirmant les autres. L'analyse prend 3 scores pour exprimer un gradient (3 valeurs propres sur la figure 3) et forme deux polynômes (de degré 2 sur 4 et de degré 3 sur 9). On a un effet Guttman.

Question 9 Quelle est la corrélation de chacun des nuages …

Ces corrélations sont toutes nulles *à condition seulement de les calculer avec les pondérations marginales du tableau* (cw et lw dans le listing).

Question 10 Donner la plus grande valeur propre de l'analyse des correspondances de $\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$

Observer que les pondérations marginales sont :

$$\left(\frac{1}{2},\frac{1}{2}\right) \text{ et } \left(\frac{1}{4},\frac{1}{2},\frac{1}{4}\right)$$

L'AFC donne des scores normés maximisant la variance des moyennes conditionnelles. Il n'y a avec 2 lignes qu'une seule façon de fabriquer un score centré réduit des lignes. Pour le score normé $(-1,1)$ des lignes, les moyennes conditionnelles sont $(-1,0,1)$. Elles sont centrées pour la pondération des lignes et de variance $\frac{1}{4}+\frac{1}{4}=0.5=\lambda_1$. L'autre valeur propre est toujours nulle.

Ou encore, on peut diagonaliser (schéma 5 du cours) :

$$\mathbf{B}=\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{bmatrix}=\begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}$$

La trace est 1.5 à laquelle on enlève la valeur propre parasite 1 et il reste $\lambda_1=0.5$.

Question 11 Donner les coordonnées

On prend le score $(-1,1)$. Il est normé. On le multiplie par $\sqrt{\lambda_1}$. On a $c=\left(-1/\sqrt{2},1/\sqrt{2}\right)$ (variance 0.5). On prend les moyennes par ligne, on multiplie par $\sqrt{\lambda_1}$ et on a $l=(-1/2,0,1/2)$ (variance 0.5).

Vérification :

```
> corresp(matrix(c(1,1,0,0,1,1),2,3,byrow=T))
First canonical correlation(s): 0.7071

 Row scores:
R 1 R 2
 -1   1

 Col scores:
   C 1    C 2    C 3
-1.414  0.000  1.414
corresp donne les scores normés (variance 1)

> unclass(dudi.coa(data.frame(matrix(c(1,1,0,0,1,1),2,3,byrow=T)),scan=F))
$tab
  X1 X2 X3
1  1  0 -1
2 -1  0  1

$cw
  X1   X2   X3
0.25 0.50 0.25

$lw
  1   2
0.5 0.5

$eig
[1] 0.5

$rank
[1] 1

$nf
[1] 1

$co
   Comp1
X1    -1
X2     0
X3     1
```

```
$li
    Axis1
1 -0.7071
2  0.7071
```

Question 12 Quel est le rang du tableau …

3 (4 colonnes - 1 valeur propre nulle). La troisième valeur propre est petite mais non nulle.

Question 13 Quel est le Khi2 de la table de contingence

$$\chi^2 = N\sum \lambda_k = 1000\left(0.1614 + 0.0174 + 0.00007\right) = 178.9$$

Question 14 Que vaut la corrélation …

La corrélation canonique est $\sqrt{\lambda_1} = \sqrt{0.1614} = 0.40$. C'est le maximum possible de la corrélation entre deux scores marginaux de la table de contingence.

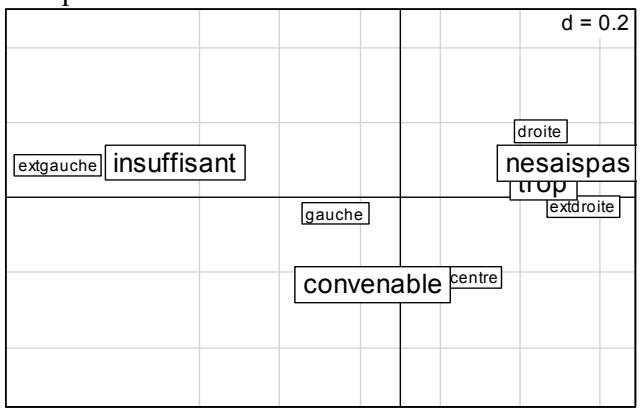
Question 15 Comment est construite la figure ?

Elle place les lignes et les colonnes avec la première scoordonnée ($li et $co de la liste). Elle exprime le "dual scaling" dit aussi "canonical scoring" et résume l'AFC comme analyse canonique. On peut aussi faire la figure avec les scores normés ($l1 et $c1).

Question 16 Donner une légende …

La table de contingence est représentée par des carrés proportionnels (en surface) aux effectifs. Les lignes et les colonnes sont positionnées par la première coordonnée ($li et $co) ou par le premier score ($l1 et $c1) ce qui est la même chose à une constante près $\sqrt{\lambda_1}$. Les points sont les moyennes conditionnelles par classe et les droites sont les deux droites de régression. Les scores de l'AFC donnent une régression doublement linéaire et λ_1 est le carré de corrélation et le rapport de corrélation.

Question 17 Placer les étiquettes :



Question 18 Interpréter

L'analyse montre la concordance du gradient droite - gauche avec celui d'une bienveillance croissante envers les syndicats. Elle souligne que ce lien n'est pas caricatural et que l'amplitude de variation d'une opinion dans la classe de l'autre est forte. Elle distingue bien "extrême gauche" et "gauche" mais confond "droite" et "extrême droite". Elle positionne enfin la catégorie "Ne sais pas" franchement "à droite" ce qui ne peut être le fruit du hasard. Elle permet d'éditer le résultat sous la forme :

```
                nesaispas trop convenable insuffisant
```

extgauche	0.03	0.08	0.24	**0.65**
gauche	0.15	0.24	**0.26**	**0.35**
centre	0.20	**0.31**	**0.32**	0.17
droite	**0.28**	**0.38**	0.15	0.18
extdroite	**0.27**	**0.40**	0.23	0.10

Question 19 Ce qu'on peut faire avec les modules d'ADE-4

Entre autre , BinToText ou ADEBin, lectures des fichiers binaires ; TextToBin et ADETrans : lecture et édition des fichiers binaires, Scatters et ScatterClass : dessin des nuages de points, PCA : analyses en composantes principales, COA : analyse des correspondances et variantes, DDUtil : utilitaires sur les schémas de dualité, …

Question 20 Quelles sont les limites ?

On peut invoquer les limites techniques (temps de calcul et place mémoire), les difficultés de dépouillement (saturation des graphiques et énormité des listings d'inertie). On peut aussi se demander le pourquoi d'une analyse d'un tableau énorme, l'absence possible d'objectif précis. On peut aussi dire qu'il n'y a pratiquement pas de limite si on cherche essentiellement des images de structures sans s'intéresser aux positions particulières. On peut également sous-échantillonner les grands tableaux, vérifier la stabilité des analyses partielles donc parler de redondance inutile ou regrouper par une classification aveugle et projeter en individus supplémentaires.