Td-TP Ch 3 : Analyse Factorielle des Correspondances

0. Objectifs, données et principes de l'*AFC*

Objectifs et données de l'*AFC*

Cette technique s'applique à des tableaux de contingence croisant deux variables qualitatives avec de nombreuses modalités chacune. Les données sont donc les effectifs des individus croisant deux modalités données. Pour de tels tableaux nous disposons du test d'indépendance du $\chi 2$.

L'**objectif** est ici de faire une synthèse de l'ensemble du tableau afin de répondre aux questions :

- **Pour une variable donnée**, certaines modalités sont-elles proches ou éloignées.
  - La proximité de deux modalités se mesure en comparant leur distribution par rapport à l'autre variable.

  Par exemple, yeux bleus et verts sont proches si les deux groupes ont les mêmes distributions de couleurs de cheveux.
- **Entre les deux variables**, certaines modalités « s'attirent-elles » davantage ou au contraire « se repoussent ».
  - On compare la fréquence observée par rapport à la fréquence attendue sous l'hypothèse d'indépendance, si la fréquence observée est plus forte il y a une plus forte association entre les deux et inversement.

  Par exemple, les yeux bleus et les cheveux blonds « s'attirent », au contraire des yeux noirs et des cheveux blonds.

Remarque : L'*AFC* n'a d'intérêt que si il y a dépendance entre les deux variables, en cas contraire elle n'apporte pas d'information.

Le **tableau** se présente sous la forme :

		Variable qualitative 2				
		Modalité 1	…	Modalité j	…	Modalité J
Variable qualitative 1	Modalité 1	n_{11}		n_{1j}		n_{1J}
	…					
	Modalité i	n_{i1}		n_{ij}		n_{iJ}
	…					
	Modalité I	n_{I1}		n_{Ij}		n_{IJ}

Exemple 1 (trivial)
On examine la répartition des couleurs de cheveux et des yeux.

Tableau de contingence

\cheveux yeux	blond	roux	brun	total
bleu	20	5	5	
vert	0	15	5	
marron	5	5	40	
total				

Tableau théorique sous H_0 (indépendance)

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Test du χ^2 d'indépendance : à la main et sous R : `chisq.test(Ex1)`

Principe de l'*AFC*

Pour mesurer les distances entre modalités, il est nécessaire de calculer au préalable la distribution de chaque modalité d'une variable en fonction de l'autre variable.

On définit ainsi **les profils colonnes et les profils ligne qui sont les distributions respectives des modalités des deux variables.**

Un profil ligne (colonne) est déterminé en divisant la ligne (colonne) par le total de la ligne (colonne).

Profils lignes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Profils des colonnes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Chaque profil est alors assimilé à un point de coordonnées les proportions par rapport aux modalités de l'autre variable. Le nuage des profils ligne est alors projeté sur des axes factoriels en conservant le maximum d'inertie et il en est de même pour le nuage des profils colonne.
Les principales différences avec l'ACP sont :

- **Lignes et colonnes sont transformées au préalable en profil et jouent un rôle symétrique**.
- La distance entre deux profils est calculée à l'aide de la **distance du χ^2** (distance entre deux distributions) :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

- Les profils sont tous dans un hyperplan car la somme de leurs coordonnées est 1. On a donc un axe de moins qu'en ACP et une valeur propre en moins.
- **Les valeurs propres (inertie projetée sur l'axe) sont inférieures à 1**.
- Les deux nuages représentent des profils et il est d'**usage de représenter les deux nuages dans un même plan** (le profil d'une modalité est « quasi » le barycentre des profils des modalités de l'autre variable).
- **La proximité entre modalités des deux variables indique une attirance entre ces modalités** (l'effectif observé est supérieur à celui attendu sous H_0).
- **La proximité entre modalités d'une même variable indique que les distributions sont voisines** pour ces deux modalités au regard de l'autre variable.

Exemple 1 : Les résultats de l'ajustement donnent :

```
> afc=dudi.coa(Ex1) ; afc$eig
[1] 0.4050000 0.3333333
> (afc$eig/sum(afc$eig))
[1] 0.5485327 0.4514673
```

Exemple 2 : CSP
Le tableau décrit la consommation annuelle en francs d'un ménage pour différentes denrées alimentaires en 1972. MA, EM, CA indiquent la catégorie socio-professionnelles et 2, 3, 4, 5 la taille du foyer.

```
> csp=read.table("csp.txt")
pain legu frui vian vola lait vin
MA2 332 428 354 1437 526 247 427
...
```

Statistiques élémentaires

```
> chisq.test(csp)
Pearson's Chi-squared test
X-squared = 1290.386, df = 66, p-value < 2.2e-16
```

Calcul des profils :

Profils-lignes : $\mathbf{L} = \tilde{\boldsymbol{D}}_I^{-1}\mathbf{P}$

```
> round(csp/apply(csp,1,sum),2)
     pain legu frui vian vola lait vin
MA2
EM2 0.08 0.15 0.10 0.40 0.15 0.06 0.07
CA2 0.07 0.15 0.11 0.37 0.18 0.04 0.08
MA3 0.10 0.14 0.08 0.37 0.13 0.08 0.10
EM3 0.09 0.15 0.10 0.36 0.14 0.08 0.09
CA3 0.07 0.14 0.11 0.39 0.19 0.04 0.06
MA4 0.12 0.14 0.08 0.35 0.14 0.09 0.09
EM4 0.09 0.14 0.10 0.37 0.15 0.08 0.08
CA4 0.07 0.13 0.11 0.40 0.19 0.05 0.05
MA5 0.12 0.14 0.08 0.34 0.14 0.09 0.09
EM5 0.10 0.17 0.09 0.35 0.15 0.09 0.05
CA5 0.07 0.15 0.12 0.37 0.16 0.08 0.04
```

Profils-colonnes : $\mathbf{C} = \mathbf{P}\tilde{\boldsymbol{D}}_J^{-1}$

```
> round(t(t(csp)/apply(t(csp),1,sum)),2)
    pain legu frui vian vola lait vin
MA2 0.06 0.05 0.06 0.06 0.05 0.06
EM2 0.05 0.06 0.06 0.07 0.06 0.06
CA2 0.07 0.09 0.09 0.09 0.10 0.05
MA3 0.08 0.06 0.06 0.07 0.06 0.08
EM3 0.07 0.07 0.07 0.07 0.06 0.07
CA3 0.08 0.10 0.11 0.10 0.12 0.06
MA4 0.10 0.08 0.06 0.07 0.07 0.10 0.09
EM4 0.09 0.08 0.08 0.08 0.08 0.09 0.09
CA4 0.07 0.09 0.10 0.10 0.12 0.07 0.06
MA5 0.12 0.09 0.07 0.08 0.08 0.12 0.11
EM5 0.11 0.11 0.09 0.09 0.09 0.12 0.07
CA5 0.10 0.12 0.15 0.12 0.12 0.13 0.06
```

Poids des lignes et colonnes

Profil-colonne moyen $\bar{x}_C = \left|f_{1+}, \cdots, f_{I+}\right|^{\mathbf{T}} \in \mathbb{R}^I$

```
> round(apply(csp,1,sum)/sum(csp),2)
MA2 EM2 CA2 MA3 EM3 CA3 MA4 EM4 CA4 MA5 EM5 CA5
0. 0.06 0.09 0.07 0.07 0.10 0.08 0.08 0.10 0.09 0.10 0.1
```

Profil-ligne moyen $\bar{x}_L = \left|f_{+1}, \cdots, f_{+J}\right| \in \mathbb{R}^J$

```
> round(apply(csp,2,sum)/sum(csp),2)
pain legu frui vian vola lait vin
0. 0.14 0.10 0.37 0.16 0.07 0.07
```

Comment décrire ce tableau ? Quelles sont les relations entre variables et modalités de chaque variable ?
Comment présenter ce tableau à un public non averti ?

Guide pratique de l'analyse AFC

Aides à l'interprétation

1. VALEURS PROPRES λ ET CHOIX DES AXES

Pour définir le nombre d'axes étudiés, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projeté sur un axe donné.

Remarques importantes:

- La somme des valeurs propres est toujours égale à l'inertie totale du nuage. On caractérise ainsi chaque axe par le % d'inertie qu'il permet d'expliquer.
- En AFC, les **valeurs propres sont toutes inférieures à 1**.

On ne retient donc que les axes avec les plus fortes valeurs propres. Le choix des axes retenus est un peu délicat. On peut donner quelques règles :

- **Règle du coude :** On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme. On retient les axes avant le décrochage.
- **Règle de l'inertie minimale** : On sélectionne les premiers axes afin d'atteindre un % donné d'inertie expliquée (70% par exemple).
- **Règle du bon sens :** On analyse les plans et axes et on ne retient que ceux interprétables.

Exemple 2 CSP

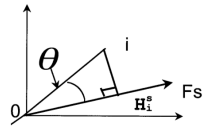
Réaliser une AFC avec **ade4**:

```
> afc=dudi.coa(csp)
> afc$eig  # Extraire les valeurs propres :
> afc$eig/sum(afc$eig)  # Calcul des % :

> round(afc$eig,3)
[1] 0.014 0.005 0.001 0.001 0.000 0.000
> round(afc$eig/sum(afc$eig)*100)
[1] 66 25 5 2 1 1
```


2. QUALITE DE REPRESENTATION *QLT*

Les profils projetés dans un plan factoriel ne sont pas forcément correctement représentés. Dans ce cas, les interprétations sont erronées. Il est indispensable de vérifier la bonne représentation au préalable.



Qualité de représentation *qlt*:

Si l'angle ɵ est grand, le point initial est éloigné de sa projection. On utilise le paramètre $\cos^2$ɵ pour caractériser la **qualité de représentation** (*qlt*) sur un axe.

- Plus *qlti* est proche de 1 plus il est bien représenté.
- Plus *qlti* est proche de 0 plus il est mal représenté.
- Dans un plan, on calcule la somme des deux *qlt*, par exemple $qlt_{F1} + qlt_{F2}$ pour le plan F^1 F^2 .
- *qlt* correspond en fait au rapport de l'inertie du projeté sur l'inertie du point initial.

Qualité globale : Dans un plan donné, on définit également la qualité globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la *qlt* d'un profil.

Bilan : On commencera donc toujours l'analyse d'un plan factoriel en précisant l'existence (ou non) de profils mal représentés et en justifiant par les *qlt*.

3. CONTRIBUTION *ctr*

Lors de la construction d'un axe factoriel, certains profils ont des rôles plus importants. On calcule un paramètre appelé **contribution**, *ctr*, qui permet de calculer cette influence.

Définition: La contribution *ctr* est définie comme la proportion de l'inertie de l'axe expliquée par le profil pour un axe donné.

Règles d'interprétation :

- L'analyse se fait axe par axe, en parallèle sur les profils ligne et colonne.
- Plus *ctr* est grande, plus l'influence du profil est grande. On ne retient donc que les plus fortes valeurs
  (il y a souvent un décrochage après quelques valeurs).
- *ctr* est considéré comme positif si le profil est dans la partie positive de l'axe.
- *ctr* est considéré comme négatif si le profil est dans la partie négative de l'axe.
- Le bilan des *ctr* peut être présenté pour un axe donné sous forme d'un **tableau** avec les principales *ctr* + et – des profils, en précisant la valeur de *ctr* :

Ctr axe F_1	-	+
Profils colonnes		
Profils lignes		

On réalise ensuite une interprétation.

Sous R, ces paramètres sont obtenus avec les commandes suivantes :

Pour les lignes

```
> inertieL<-inertia.dudi(afc, row.inertia=TRUE)
> inertieL$row.abs/100  # ctr des lignes en % :
> inertieL$row.rel/100  # qlt des lignes en % :
```

Pour les colonnes

```
> inertieC<-inertia.dudi(afc, col.inertia=TRUE)
> inertieC$col.abs/100  # ctr des colonnes en % :
> inertieC$col.rel/100  qlt des colonnes en % :
```

Représentation graphique des profils

Les profils ligne (respectivement colonne) sont associés à des points de l'espace dont les coordonnées sont les distributions conditionnelles. On peut mesurer la **distance entre ces profils** en utilisant la distance du χ^2 entre ces deux points.

La construction des composantes principales conduit à **rendre minimale la déformation des distances entre profils lorsque l'on projette les profils dans le plan factoriel** F^1 F^2. Ainsi les distances que l'on observe entre les profils dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces profils.

L'analyse des plans factoriels permet ainsi d'observer les profils proches entre eux (distribution similaire) ou au contraire éloignés. Il est ainsi possible de construire des groupes, d'observer des tendances ...

Les règles de lecture des plans factoriels sont :

- **Les profils ligne et colonne sont projetés simultanément** afin notamment d'observer les modalités des deux variables présentant de fortes (attraction) ou faibles (répulsion) associations.
- **Seuls les profils bien représentés sont pris en compte** dans l'interprétation.
  - On calcule la somme des qtl dans le plan et on vérifie que cette somme n'est pas trop faible par rapport à la qualité moyenne du plan.
- On réalise le **bilan en positif et en négatif des profils qui ont la plus forte contribution** pour un axe donné.
  - On donne ainsi en parallèle sur les lignes et colonnes une signification concrète à ces axes en termes d'attraction, de similarité, entre modalités ou en tendance particulière.
- **On réalise des groupes**, à l'aide éventuelle de la fonction `s.class`, dans le cas de groupes préexistants (homme-femme par exemple) ou on construit arbitrairement ces groupes en raison des proximités entre profils.
- L'utilisation de **profils supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

En résumé Guide pratique de l'analyse *AFC*

- **Etape 1 :** Sélection des axes et des plans retenus principalement par rapport aux valeurs propres.
- **Etape 2 :** Projection des profils ligne et colonne dans un plan donné (F^1 F^2 en premier)
  - Examen des *qlt* dans le plan pour éliminer les profils mal représentés
  - Bilan des *ctr* pour un axe afin de donner un sens à cet axe (opposition, tendance ...)
  - Topographie des profils afin d'identifier des groupes, des oppositions, des tendances notamment à l'aide de la fonction `s.class`
  - Utiliser ses connaissances sur le sujet pour proposer des explications sur les résultats de l'analyse
  - Utiliser des profils supplémentaires ou des profils type (moyenne des H et des F par exemple)

I Exercice « à la main »

Tableau de contingence N

1. Construire le tableau.

cheveux yeux	blond	roux	brun	total
bleu	10	10	10	
vert	7	6	7	
marron	13	4	33	
total				

2. Construire la matrice **P** des fréquences relatives, les vecteurs colonnes (resp lignes) P_I (resp. P_J) des fréquences marginales lignes (resp colonnes), les matrices $\tilde{D}_I = n^{-1}D_I$, $\tilde{D}_J$ et leur inverse $\tilde{D}_I^{-1} et\ \tilde{D}_J^{-1}$.
3. Déterminer les matrices **L** et **C** des profils lignes et colonnes puis $L_0 = \tilde{D}_I^{-1}P - 1_I\bar{L}$ (où $(f_{+1},...,f_{+J}) = \bar{L}$) des profils lignes centrés et $C_0 = P\tilde{D}_J^{-1} - \bar{C}1_J^T$ ((où $(f_{I+},...,f_{I+})' = \bar{C}$) des profils colonnes centrés.
4. Etudier la liaison entre les deux variables.
5. Réaliser la DVS sur :
   a. $(L, \tilde{D}_J^{-1}, \tilde{D}_I)$ et $(L_0, \tilde{D}_J^{-1}, \tilde{D}_I)$
   b. $(C^T, \tilde{D}_I^{-1}, \tilde{D}_J)$ et $(C_0^T, \tilde{D}_I^{-1}, \tilde{D}_J)$
   c. $(X = \tilde{D}_I^{-1}P\tilde{D}_J^{-1}, Q = \tilde{D}_J, D = \tilde{D}_I)$

   Comparer les résultats et effectuer les projections des profils.

II Etude d'un tableau à l'aide d'une *AFC*

Partie A : calculs « A la main »

On considère le tableau de contingence suivant :

	X	Y	Z
A	1	0	0
B	1	1	0
C	1	0	1
D	0	0	1

1. **Calcul des fréquences et profils**
   Calculer le tableau des fréquences relatives, **F**, les fréquences marginales f_{i+} et f_{+j}et les profils lignes **L** et colonnes **C**
2. **Calcul des distances**
   Calculer les distances entre les modalités A, B, C et D de la première variable. Les résultats seront représentés sous forme d'un tableau.

$$Rappel : d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

3. Construire la matrice **X**

$$\mathbf{X} = \left[\frac{f_{ij}}{\sqrt{f_{i+}}\sqrt{f_{+j}}}\right] = \left[\frac{n_{ij}}{\sqrt{n_{i+}}\sqrt{n_{+j}}}\right]$$

4. Valeurs propres et inertie

Les valeurs propres de $\mathbf{X}^T\mathbf{X}$ sont $\lambda_0 = 1; \lambda_1 = \frac{2}{3}; \lambda_2 = \frac{1}{4}$.

a. Quel est l'inertie totale du nuage de points?
b. En déduire les % d'inertie projetée sur les différents axes.

5. Vecteurs propres et projection des profils ligne

a. Calculer les vecteurs propres normés $\boldsymbol{u}_0^*$ $\boldsymbol{u}_1^*$$\boldsymbol{u}_2^*$ (pour la norme usuelle) de $\mathbf{X}^T\mathbf{X}$.
b. En déduire les composantes principales $\boldsymbol{F}^1$ et $\boldsymbol{F}^2$.

Rappel : $F^1 = \mathbf{L}\tilde{\mathbf{D}}_J^{-1/2}u_1^*$ avec $\tilde{\mathbf{D}}_J$ la matrice diagonale (f_{+j}).

c. Représenter les points ligne dans le plan ($\boldsymbol{F}^1$,$\boldsymbol{F}^2$)

6. Projection des profils colonne

a. Calculer les coordonnées $\boldsymbol{G}^1$ et $\boldsymbol{G}^2$ pour les profils colonne à l'aide des formules de transition.

Rappel : $G^1 = \sqrt{\lambda_1}\tilde{D}_J^{-1/2}u_1^*$.

b. Représenter les points colonnes dans le même plan ($\boldsymbol{G}^1$,$\boldsymbol{G}^2$)=($\boldsymbol{F}^1$,$\boldsymbol{F}^2$).

Partie B : Calculs à l'aide du logiciel R

1. Traiter les différentes questions de la partie A à l'aide du logiciel R.

La suite des commandes, ainsi que les résultats, seront notés dans un fichier Word. Toutes les commandes nécessaires ont été utilisées au TP1.

2. Construire une fonction dans R qui pour tout tableau A donné en argument donne en sortie le tableau des fréquences relatives, les fréquences marginales, les profils lignes et colonnes, les valeurs propres non triviales, les axes de projection, les coordonnées F et G.

Partie C : Etude d'un second exemple

Reprendre les étapes du **II** parties A et B (manuel + vérification sous R) avec le tableau de données :

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array}\begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix} \end{array}$$

III Analyse de tableaux de données

Exemple I : Catégories socio-professionnelles et alimentation

Le tableau décrit la consommation annuelle en francs d'un ménage pour différentes denrées alimentaires en 1972. MA, EM, CA indiquent la catégorie socio-professionnelle et 2,3,4,5 la taille du foyer.

```
> csp=read.table("csp.txt")
      pain  legu  Frui vian  vola  Lait vin
MA2   332   428   354 1437    526   247 427
EM2   293   559   388 1527    567   239 258
CA2   372   767   562 1948    927   235 433
MA3   406   563   341 1507    544   324 407
EM3   386   608   396 1501    558   319 363
CA3   438   843   689 2345   1148   243 341
MA4   534   660   367 1620    638   414 407
EM4   460   699   484 1856    762   400 416
CA4   385   789   621 2366   1149   304 282
MA5   655   776   423 1848    759   495 486
EM5   584   995   548 2056    893   518 319
CA5   515  1097   887 2630   1167   561 284
```

1. a. A quelles questions permet de répondre ce tableau ?
b. Quelles méthodes d'analyse de ce tableau peut-on envisager ?

2. a. Calculer le profil-ligne de `MA2` et le profi- colonne de `vin`.
b. Que représentent ces profils ?
c. Interpréter les poids des lignes et colonnes.

```
    pain legu frui vian vola lait vin
MA2 ????????????????????????????????????
EM2 0.08 0.15 0.10 0.40 0.15 0.06 0.07
CA2 0.07 0.15 0.11 0.37 0.18 0.04 0.08
MA3 0.10 0.14 0.08 0.37 0.13 0.08 0.10
EM3 0.09 0.15 0.10 0.36 0.14 0.08 0.09
CA3 0.07 0.14 0.11 0.39 0.19 0.04 0.06
MA4 0.12 0.14 0.08 0.35 0.14 0.09 0.09
EM4 0.09 0.14 0.10 0.37 0.15 0.08 0.08
CA4 0.07 0.13 0.11 0.40 0.19 0.05 0.05
MA5 0.12 0.14 0.08 0.34 0.14 0.09 0.09
EM5 0.10 0.17 0.09 0.35 0.15 0.09 0.05
CA5 0.07 0.15 0.12 0.37 0.16 0.08 0.04
```

Profils-lignes

```
> round(csp/apply(csp,1,sum),2)
```

```
     pain legu frui vian vola lait  vin
MA2  0.06 0.05 0.06 0.06 0.05 0.06    ?
EM2  0.05 0.06 0.06 0.07 0.06 0.06    ?
CA2  0.07 0.09 0.09 0.09 0.10 0.05    ?
MA3  0.08 0.06 0.06 0.07 0.06 0.08    ?
EM3  0.07 0.07 0.07 0.07 0.06 0.07    ?
CA3  0.08 0.10 0.11 0.10 0.12 0.06    ?
MA4  0.10 0.08 0.06 0.07 0.07 0.10    ?
EM4  0.09 0.08 0.08 0.08 0.08 0.09    ?
CA4  0.07 0.09 0.10 0.10 0.12 0.07    ?
MA5  0.12 0.09 0.07 0.08 0.08 0.12    ?
EM5  0.11 0.11 0.09 0.09 0.09 0.12    ?
CA5  0.10 0.12 0.15 0.12 0.12 0.13    ?
```

Profils-colonnes

```
> round(t(t(csp)/apply(t(csp),1,sum)),2)
```

Poids des lignes et colonnes

```
> round(apply(csp,1,sum)/sum(csp),2)
 MA2  EM2  CA2  MA3  EM3  CA3  MA4  EM4  CA4  MA5  EM5  CA5
0.06 0.06 0.09 0.07 0.07 0.10 0.08 0.08 0.10 0.09 0.10 0.12
```

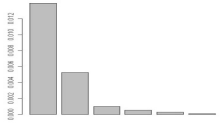
```
> round(apply(csp,2,sum)/sum(csp),2)
pain legu frui vian vola lait  vin
0.09 0.14 0.10 0.37 0.16 0.07 0.07
```

3. Valeurs propres et axes

a. Justifier que 2 axes aient été retenus.

b. **Expliquer le calcul de 6 valeurs propres**

```
> library(ade4)
> afc <-dudi.coa(csp)
> round(afc$eig,3)
[1] 0.014 0.005 0.001 0.001 0.000 0.000
> round(afc$eig/sum(afc$eig)*100)
[1] 66 25 5 2 1 1
```



4. Profils-lignes et colonnes

a. **Etudier la qualité de représentation des profils-lignes dans le plan F^1-F^2.**

b. Il manque le profil `vin` sur le plan F^1-F^2. Replacer le.

c. Etudier la qualité de représentation des profils-colonnes dans le plan F^1-F^2.

d. Interpréter chacun des axes.

e. Proposer une synthèse de ces résultats.

```
> inertie <-inertia.dudi(afc, row.inertia=TRUE)
```

Coordonnées profils ligne

```
> round(afc$li,2)
```

	Axis1	Axis2
MA2	-0.10	-0.14
EM2	0.04	-0.02
CA2	0.08	-0.09
MA3	-0.13	-0.05
EM3	-0.08	-0.01
CA3	0.16	-0.05
MA4	-0.15	0.02
EM4	-0.05	-0.01
CA4	0.17	-0.02
MA5	-0.17	0.03
EM5	-0.03	0.12
CA5	0.11	0.10

[CTR en %]

```
> round(inertie$row.abs/100)
```

	Axis1	Axis2
MA2	4	24
EM2	1	0
CA2	4	14
MA3	8	3
EM3	3	0
CA3	18	4
MA4	13	1
EM4	2	0
CA4	20	1
MA5	18	2
EM5	1	27
CA5	10	24

[QLT en %]

```
> round(inertie$row.re/100)
```

	Axis1	Axis2	con.tra
MA2	-29	-64	10
EM2	31	-5	2
CA2	34	-49	7
MA3	-86	-11	6
EM3	-80	-2	3
CA3	87	-7	14
MA4	-94	2	9
EM4	-81	-2	1
CA4	90	-1	14
MA5	-91	3	13
EM5	-5	89	8
CA5	46	43	14

```
> inertie <-inertia.dudi(afc, col.inertia=TRUE)
```

Coordonnées profils colonne

```
> round(afc$co,2)
```

	Comp1	Comp2
pain	-0.19	0.04
legu	0.01	0.07
frui	0.13	0.01
vian	0.04	-0.03
vola	0.12	-0.02
lait	-0.19	0.15
vin	-0.23	-0.19

[CTR en %]

```
> round(inertie$col.abs/100)
```

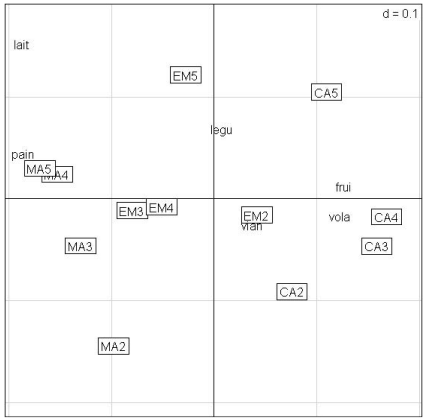
	Comp1	Comp2
pain	22	3
legu	0	13
frui	11	0
vian	4	5
vola	17	1
lait	18	31
vin	29	48

[QLT en %]

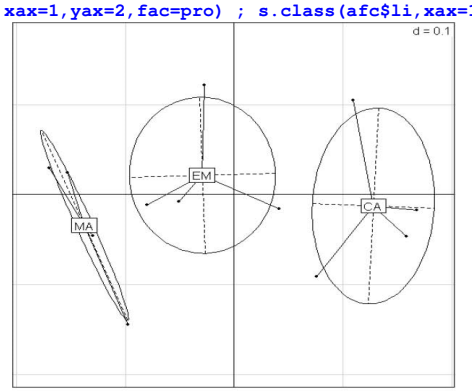
```
> round(inertie$col.re/100)
```

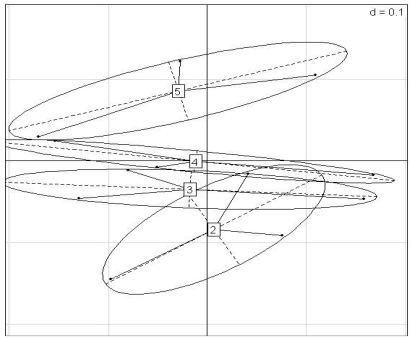
	Comp1	Comp2	con.tra
pain	-87	5	17
legu	1	66	5
frui	79	1	10
vian	51	-25	5
vola	84	-2	13
lait	-58	38	20
vin	-61	-38	31


```
> s.label(afc$li,xax=1,yax=2)
> s.label(afc$co,xax=1,yax=2,add.plot=T,boxes=F)
```




```
> pro <-as.factor(c("MA","EM","CA","MA","EM","CA","MA","EM","CA","MA","EM","CA"))
> nb <-as.factor(c(2,2,2,3,3,3,4,4,4,5,5,5))
> s.class(afc$li,xax=1,yax=2,fac=pro) ; s.class(afc$li,xax=1,yax=2,fac=nb)
```





Exemple II : Elections présidentielles de 2002

Dans un sondage, on a interrogé les lecteurs de 12 périodiques sur leur vote au premier tour des présidentielle 2002 (100 lecteurs par périodiques). Les résultats sont dans le tableau ci-dessous.

	La Croix	Le Figaro	Libération	Le Monde	Le Parisien	Le Canard	L'Express	Marianne	Nouvel Obs	Paris Match	Télérama	Le Point	Total	%
Laguiller	0	2	6	5	4	8	4	6	4	4	4	2	**49**	**4**
Besancenot	2	2	8	6	3	7	2	6	7	2	8	2	**55**	**5**
Hue	2	0	3	2	4	5	1	2	2	1	5	1	**28**	**2**
Jospin	3	7	41	26	12	29	15	19	35	15	28	7	**237**	**20**
Taubira	2	1	5	3	2	2	2	3	3	0	4	3	**30**	**3**
Chevènement	2	2	5	5	4	7	4	15	5	2	7	2	**60**	**5**
Mamère	4	1	10	7	6	9	5	4	8	2	13	1	**70**	**6**
Lepage	5	3	0	2	1	2	2	2	1	2	2	2	**24**	**2**
Saint Joss	3	1	1	1	1	4	3	3	1	3	0	1	**22**	**2**
Bayrou	20	8	2	5	6	4	8	10	6	7	10	8	**94**	**8**
Madelin	2	9	2	4	2	2	9	5	3	4	3	9	**54**	**5**
Chirac	29	35	9	18	23	8	22	9	14	29	9	41	**246**	**21**
Boutin	8	3	0	1	2	0	1	1	1	2	2	2	**23**	**2**
Megret	2	3	0	2	2	2	5	3	0	2	0	1	**22**	**2**
Le Pen	14	22	7	12	27	10	16	10	9	22	3	16	**168**	**14**
Blanc	2	1	1	1	1	1	1	2	1	3	2	2	**18**	**2**
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

a. Indiquer, en justifiant vos propos, la (es) méthode(s) factorielle(s) que l'on pourrait utiliser pour traiter les données.

b. Analyser avec soins les résultats obtenus.

Exemple III : Reprenez ces exemples en utilisant la library FactoMineR

Exemple IV : AFCM préférences de consommateurs (`PrefConsom`)

Cet exemple est tiré de l'ouvrage de Bellanger et Tomassone (2014), il contient les résultats d'une enquête sur la préférence de consommateurs dans laquelle on a trois variables : `sexe`, `classe d'âge` et `produit` (C=3) ; le total des dénombrements des personnes dans les 24 combinaisons vaut $n = 1000$. La matrice $\mathbf{Z}$ a donc 1000 lignes et 9 colonnes ; mais on peut traiter directement le tableau de Burt que l'on peut construire à partir des données (Tab.1).

TABLEAU 1 - `PrefConsom` : données pour 4 produits en fonction de l'âge et du sexe.

Classe d'âge	exe	Produit			
A1 < 20 ans		8			4
		2	0		6
A2 20-60 ans		20	0	0	0
		0	0	0	0
A3 > 60 ans		0	2	0	
		0	8	0	2

Le premier facteur met bien en évidence le goût pour le produit `D` préféré par la classe d'âge des moins de 20 ans, sans aucune différence entre hommes et femmes. Le second oppose le produit `A`, préféré par la classe d'âge de plus de 60 ans et par les hommes, aux deux autres `B` et `C` préférés par la classe d'âge entre 20 et 60 ans et les femmes. Le troisième introduit une nuance à ce qu'a décelé l'axe précédent : il oppose le produit `B` préféré par les femmes et la classe d'âge de plus de 60 ans au produit `C` préféré par les hommes et la classe d'âge entre 20 et 60 ans (Tab.2 et Fig.1).

TABLEAU 2 - `PrefConsom` : résultats de l'*AFCM*.

```
> Pref<-read.table("PrefConsom.txt",h=T)
> Pref
       Nb Produit Sexe Age
    1  28       A    M  A1
    2   8       B    M  A1
   …
   24  12       D    F  A3
> summary(Pref)
       Nb          Produit Sexe   Age
 Min.   :  6.00   A:6     F:12   A1:8
 1st Qu.: 11.50   B:6     M:12   A2:8
 Median : 34.00   C:6            A3:8
 Mean   : 41.67   D:6
 3rd Qu.: 65.50
 Max.   :120.00
# Tableau disjonctif complet d'un data frame ne contenant que les facteurs
# (acm.disjonctif)
> disj<-acm.disjonctif (Pref[,-1]) #création du tableau disjonctif
> disj
    Produit.A Produit.B Produit.C Produit.D Sexe.F Sexe.M Age.A1 Age.A2 Age.A3
1           1         0         0         0      0      1      1      0      0

23          0         0         1         0      1      0      0      0      1
24          0         0         0         1      1      0      0      0      1
# AFCM du tableau des facteurs ; « row.w » fournit les pondérations,
# ici les effectifs Nb
> Pref.acm<-dudi.acm(df = as.data.frame(Pref[,-1]), row.w=as.vector(Pref$Nb), scannf = FALSE,
  nf =6) #l'analyse factorielle
```

```
#  Eboulis des valeurs propres (Fig.5A)
> inertie<-Pref.acm$eig/sum(Pref.acm$eig)*100
> barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))
> title("Eboulis des valeurs propres en %")
#  Valeurs propres
> round(Pref.acm$eig,4)
     [1] 0.4585 0.4188 0.3598 0.3167 0.2414 0.2048
> round(Pref.acm$eig/sum(Pref.acm$eig)*100,2) #les valeurs propres en %
     [1] 22.92 20.94 17.99 15.84 12.07 10.24
# Plans factoriels (Nuages par modalité des facteurs ; Fig.5B)
> scatter(Pref.acm)
> par(mfrow=c(1,2))  # plan 1-2 et plan 1-3
> s.value(Pref.acm$li, Pref.acm$li[,2])
> s.value(Pref.acm$li, Pref.acm$li[,3])
# Aide à l'interprétation : axe 1 (Fig.5C : Ch4B-Pref-AideAxe1-a.jpg)
> modal<-as.data.frame(Pref.acm$co)
> modal<-modal[sort.list(modal$Comp1),]
> dotchart(modal[,1],labels = row.names(modal),cex=0.8)
> title(sub="Répartition des modalités sur l'axe 1") ; abline(v=0)
# Aide à l'interprétation : axe 2 (Fig.5C : Ch4B-Pref-AideAxe2-a.jpg)
> modal<-as.data.frame(Pref.acm$co)
> modal<-modal[sort.list(modal$Comp2),]
> dotchart(modal[,2],labels = row.names(modal),cex=0.8)
> title(sub="Répartition des modalités sur l'axe 2") ; abline(v=0)
# Autre représentation : les variables (Fig.5D : Pref-Variables-1&2)
> plot(Pref.acm$co[,1],Pref.acm$co[,2],type="n",xlab="Axe 1",ylab="Axe 2", xlim=c(-1.4,1.4))
> text(Pref.acm$co[,1], Pref.acm$co[,2], label= colnames(disj))
> title("Préférences consommateurs - plan des variables")
> abline(h=0,v=0)
# Autre représentation : les individus (Fig.5D : Pref-Obs-1&2)
> plot(Pref.acm$li[,1],Pref.acm$li[,2],type="n",xlab="Axe 1",ylab="Axe 2", xlim=c(-1.4,1.4))
> text(Pref.acm$li[,1], Pref.acm$li[,2], label=row.names(disj))
> title("Préférences consommateurs - Plan des individus") ;  abline(h=0,v=0)
# Contributions (absolues) des modalités des variables
# à la construction de chaque axe
> inertia.dudi(Pref.acm,col.inertia = T)$col.abs
              Comp1 Comp2 Comp3 Comp4 Comp5 Comp6
Produit.A       968  2632   114     0   604  2282
Produit.B       260  1088  1858  3436  1274     4
Produit.C       227  1666  1971  2671  1942     2
Produit.D      3280     4    50   131  1220  2315
Sexe.F          158  1029  1406   887   936   344
Sexe.M          174  1133  1547   976  1030   379
Age.A1         3105   210   672   317  1536  2161
Age.A2           63   879  1208   733  1113     4
Age.A3         1765  1360  1175   847   345  2507
# Contributions (absolues) des individus à la construction de chaque axe
> inertia.dudi(Pref.acm,row.inertia = T)$row.abs
    Axis1 Axis2 Axis3 Axis4 Axis5 Axis6
1     166   638     8     1   327  1133
2      68     0    68   151   498    66
...
23    164    35     2   600   174   132
24      0    17   222   204    69   593
```

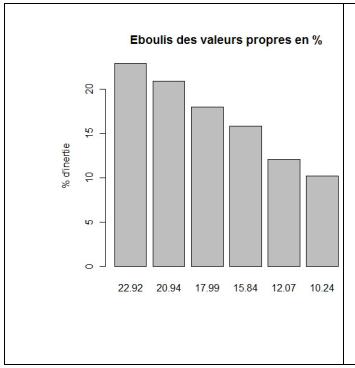



FIGURE 1A. **PrefConsom** : *éboulis des valeurs propres.*

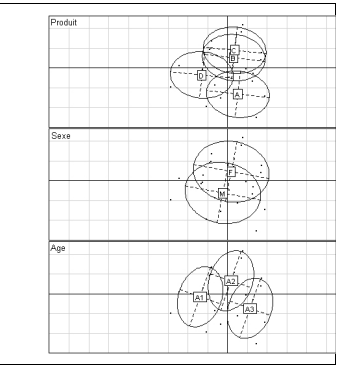


FIGURE 1B. **PrefConsom** : *nuages par modalité des facteurs.*

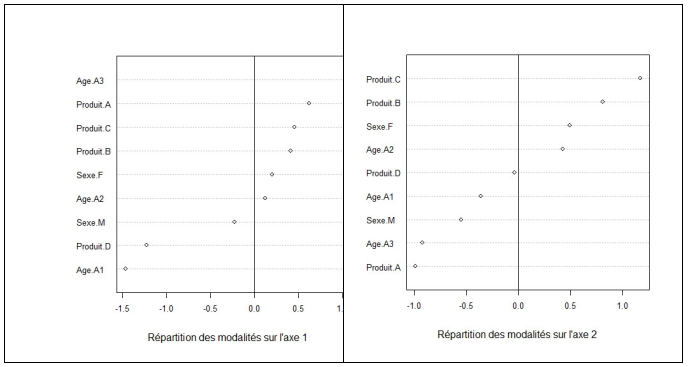


FIGURE 1C. **PrefConsom** : *répartition des modalités des variables sur l'axe 1 et 2.*

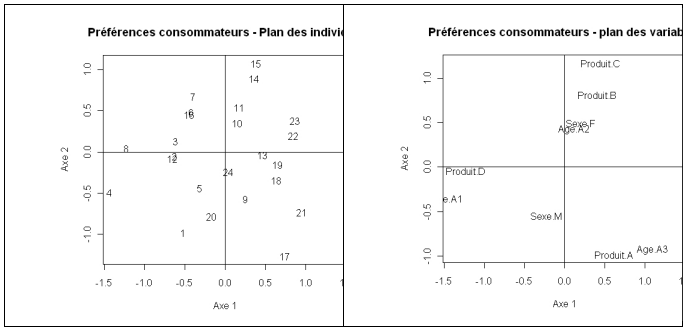


FIGURE 1D. **PrefConsom** : *représentation des observations et des variables.*

L'*AFCM* est une méthode d'analyse bien adaptée au traitement de données faisant intervenir plusieurs variables qualitatives. C'est précisément ce type de variables que l'on rencontre dans un **dépouillement d'enquêtes**, qui constitue un domaine d'application important de *l'AFCM* : elle permet d'éviter l'examen fastidieux, et souvent inutile, de tous les tableaux de contingence entre tous les couples de variables qualitatives. Elle permet de mettre en évidence des associations et des interactions entre les facteurs. Dans notre exemple, il y a association entre les trois facteurs `produit`, `sexe` et `classe d'âge`. Les produits `B` et `C` sont préférés par les femmes et par la classe d'âge médiane ; le produit `A` est préféré par les plus de 60 ans, alors que le produit *D* est préféré par les moins de 20 ans.

Comme dans les autres analyses factorielles, il est possible d'introduire des variables supplémentaires ; dans une enquête, on peut travailler sur une base de variables actives et étudier comment d'autres variables leur sont liées. Les premières peuvent être des variables comme le sexe, la classe d'âge, la catégorie socio-professionnelle ; les secondes sont des variables comme des intentions de vote, des positions sur des problèmes sociologiques. L'explication des secondes par les premières s'apparente aux méthodes de régression qui sortent du cadre du cours d'Analyse de Données.

D'autres fonctions permettent sous *R* d'effectuer une *AFCM* : **`MCA(FactoMineR)`**, **`mca(MASS)`** sont les principales alternatives.

Autres données pour l'analyse factorielle des correspondances

Exercice 1 : "heberg"
Ce fichier est un tableau de contingence croisant les catégories socio-professionnelles avec des modes de résidence en vacances. (source: M. Goguel 1965). Les catégories professionnelles sont :
AGRI : agriculteurs
SALAG : salariés agricoles
PATR : patrons
CADSUP : cadres supérieurs & prof libérales
CADMOY : cadres moyens
EMPL : employés
OUVR : ouvriers
SERV : personnels de service
AUTRE : autres actifs
NONACT : non actif
Les modes de résidence sont :
HOTEL : Hôtel, pension de famille
LOCAT : Maison louée chez l'habitant
MAISON : Maison en propriété
PARENT : chez des parents
AMIS : chez des amis
CAMPING : Tente, caravane
VVF : Village de vacances
DIVERS : autres modes

Exercice 2 : Proximité entre couleur des yeux et couleur des cheveux. Vous pouvez directement saisir les données dans votre logiciel.
Couleur des yeux Couleur des cheveux

Y/X	CBlond	CBrun	CNoir	CRoux
YBleu	1768	807	189	47
YGrisVert	946	1387	746	53
YBrun	115	438	288	16

Exercice 3 : "mariage"
Ce fichier est un tableau de contingence décrivant les mariages entre catégories professionnelles. Il croise pour les hommes et les femmes les catégories professionnelles suivantes :
agri : agriculture
ouva : ouvrier agricole
pat : patron
sup : cadre supérieur
moy : cadre moyen
emp : employé
ouv : ouvrier
serv : services
aut : autre

Exercice 4 : "psysoc"
Pour 19 pays, ce fichier présente le nombre de morts en fonction des causes. Les variables sont :
Pays : nom du pays
Suici : nombre de suicides
Omic : nombre d'homicides
Arout : nombre d'accidents de la route
Aindu : nombre d'accidents industriels
Aautr : autres type de mort
Cirfo : nombre de ciroses du foi

Exercice 5 : "science"
Sur six années consécutives, ce fichier présente les choix de types d'études suivis par les étudiants de grandes écoles.

Exercice 6 : "house"
This data frame contains four columns : wife, altern, husband and jointly. Each column is a numeric vector.
Les tâches ménagères à assumer sont :
• Laundry Main_meal Dinner Breakfeast
• Tidying Dishes Shopping Official Driving
• Finances Insurance Repairs Holidays

AFC multiple

exercice 1 : "bledur"
Vous trouverez dans le fichier "**bledur**" les variables suivantes (dans l'ordre) :
RDT='Rendement en Qt/ha'
PLM='Nb de pieds levés par m2'
ZON='Numero de zone'
ARG='Argile du sol'
LIM='Limon du sol'
SAB='Sable du sol'
VRT='Variete cultivee'
PGM='Poids de 1000 grains'
MST='Matière sèche totale'
AZP='Azote dans la plante'
Comme le découpage en classe des données est souvent une activité fastidieuse, on utilisera le fichier "**bledurCD**" dans lesquels les données ont déjà été codées en classes.
Effectuez une AFC multiple sur ces données, caractérisez les axes 1 et 2 du plan factoriel.

exercice 2 : "sencd"
Le fichier **sencd** est issu d'une enquête sur la traction animale. Les variables sont :
EX : numéro d'exploitation
QU : numéro de quartier (village)
AC : nombre d'actifs dans la famille
SP : surface possédée
SU : surface agricole utile
AT : nombre d'ânes de traction
CT : nombre de chevaux de trait
BT : nombre de paires de boeux de trait
VT : nombre de paires de vaches de trait
BV : nombre de bovins hors exploitation
OV : nombre d'ovins
CP : nombre de caprins
Les variables codées en classes sont : ACcd, SPcd, ..., OVcd, CPcd