Analyse factorielle des Correspondances Multiples

N. Jégou

Université Rennes 2

Master 1 Géographie

Bibliographie

- Husson *et al.*, Analyse de données avec R
  PUR (2009)

- Pagès J., Analyse Factorielle multiple avec R
  EDP Sciences (2013)

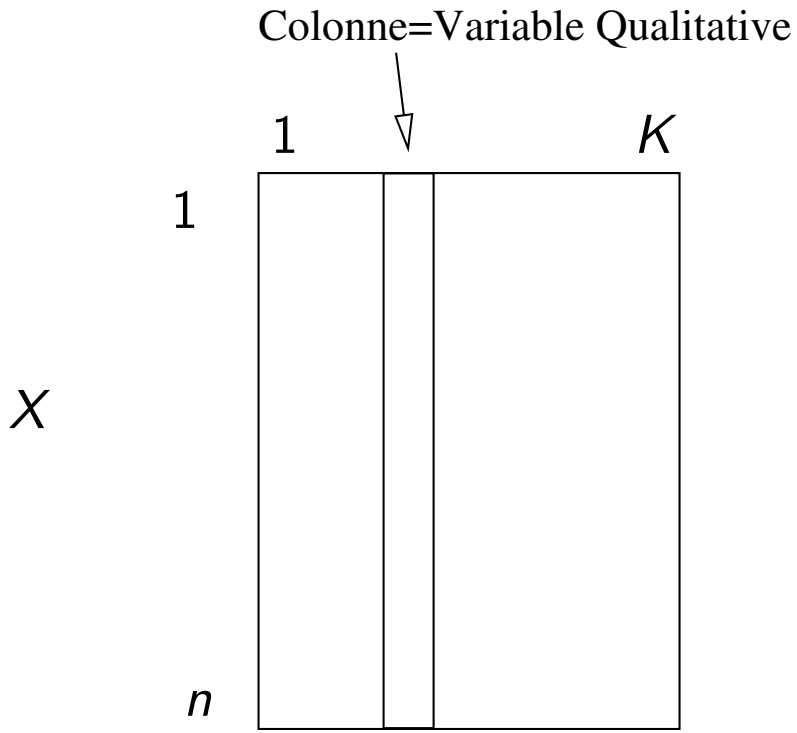
- Cornillon *et al.*, Statistique avec R
  PUR (2012)

Tableau de Données

- K variables qualitatives mesurées sur n individus
- Exemple : enquêtes
  - Quelle est votre catégorie professionnelle ? {cadre, employé, ouvrier, etc}
  - Il faut fermer les centrales nucléaires ; êtes vous {pas d'accord, sans opinion, d'accord,...}
- Les individus sont les enquêtés
- Les variables sont les questions
- Les modalités des variables sont ordonnées ou non

Tableau de Données

- K variables qualitatives mesurées sur n individus



- Chaque colonne est constituée d'observations des modalités

Objectifs

Décrire le jeu de données

- Typologie des individus :
  - Former des groupes d'individus semblables
  - Repérer les individus différents des autres
- Typologie des variables :
  - Faire émerger les liaisons entre variables
  - Faire émerger les associations entre modalités
- Dualité : Quelles variables expliquent le plus la variabilité entre individus ?

Lien entre sexe et salaire

$K = 2$ variables qualitatives

- Sexe : `F` ou `M`
- Classe de salaire :
  - faible [`f`]
  - moyen-inférieur [`m-`]
  - moyen [`m`]
  - moyen-supérieur [`m+`]
  - élevé [e]
- Mesures sur $n = 89$ individus

Extrait des données

```
> quali[1:10,1:2]
   sexe     salaire
1     M        f
2     M        m-
3     M        m-
4     M        m
5     M        m
6     M        m
7     M        m+
8     M        m+
9     M        m+
10    M        m+
```

Tableau de contingence

```
> table(quali[,1:2])
          salaire
sexe     f   m-  m   m+  e | total
   F    15   7   7   2   2 |  33
   M    10  17   11  9   9 |  56
-------------------------------
total  25  24   18  11  11|  89
```

- Forte association des modalités `F` et `f`
- Forte association des modalités `M` et `m+`
- Forte association des modalités `M` et `e`

Test du lien entre sexe et salaire

- Erreur de première espèce: $\alpha = 5\%$
- Modèle $\Pr(\text{Sexe} = i, \text{Sal.} = j) = \mathcal{M}(n, p_1, \ldots, p_{10})$
- Hypothèses : $\mathbf{H}_0$: sexe et salaire sont indépendants
  $\mathbf{H}_1$: sexe et salaire ne sont pas indépendants
- Statistique de test :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n_{ij}^\star)^2}{n_{ij}^\star}$$

où les $n_{ij}^\star$ sont les effectifs théoriques sous $\mathbf{H}_0$

Probabilité critique et conclusion

- Calcul de la probabilité critique

```
> chisq.test(table(quali[,1:2]))
Pearson's Chi-squared test
data:  table(quali[, 1:2])
X-squared = 9.6664, df = 4, p-value = 0.04644
Message d'avis :
In chisq.test(table(quali[, 1:2])) :
  l'approximation du Chi-2 est peut-être incorrecte
```

- Conclusion : $\mathbf{H}_0$ est repoussée (au seuil 5%)
- Conclusion opérationnelle : salaire et sexe sont dépendants

Conclusions

- Problèmes
  - Comment généraliser à plusieurs variables ?
  - Comment connaître les modalités liées (qui contribuent aux écarts à l'indépendance) ?
  - Comment représenter les individus ?

- Idée (vague): se servir du cadre de l'analyse factorielle (ACP)
  - Distances entre individus
  - Distances entre variables (ou modalités ?)

Exemple d'un tableau de variables qualitatives

- Le tableau de départ

$$X = \begin{bmatrix} \texttt{A1} & \texttt{B2} & \texttt{C3} \\ \texttt{A2} & \texttt{B1} & \texttt{C1} \\ \texttt{A2} & \texttt{B2} & \texttt{C2} \\ \texttt{A3} & \texttt{B2} & \texttt{C1} \\ \texttt{A3} & \texttt{B1} & \texttt{C2} \end{bmatrix}$$

- Remplacer `A1` par 1, `A2` par 2, `A3` par 3 ?
- Mauvaise idée...
- A moins (éventuellement) que les variables soient ordinales

Codage disjonctif complet du tableau

U=

$$\left[\begin{array}{ccc|cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right]$$

$$\begin{array}{cccccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ n_1^{(1)} & n_2^{(1)} & n_3^{(1)} & n_1^{(2)} & n_2^{(2)} & n_1^{(3)} & n_2^{(3)} & n_3^{(3)} \\ 1 & 2 & 2 & 2 & 3 & 2 & 2 & 1 \end{array}$$

Exemple

- $n = 15$ individus interrogés dans un même arrondissement

- Deux variables qualitatives
  - Question 1 : Etes vous favorable à l'impôt de solidarité sur la fortune ?
    `oui`, `non`, `peut-etre`

  - Question 2 : Quel est votre statut ?
    `pauvre`, `aisé`, `richissime`

Données

	Question 1	Question 2
1	oui	pauvre
2	oui	pauvre
3	oui	pauvre
4	oui	pauvre
5	peut-etre	pauvre
6	peut-etre	aise
7	peut-etre	aise
8	peut-etre	aise
9	peut-etre	aise
10	peut-etre	aise
11	non	richissime
12	non	richissime
13	non	richissime
14	non	richissime
15	non	richissime

Associations

- Individus identiques :
  - 1-2-3-4
  - 6-7-8-9-10
  - 11-12-13-14-15

- Fortes associations des modalités :
  - `oui` $\leftrightarrow$ `pauvre`
  - `peut-être` $\leftrightarrow$ `aisé`
  - `non` $\leftrightarrow$ `richissime`

Tableau disjonctif complet

	oui	peut-etre	non	pauvre	aise	richissime
1	1	0	0	1	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	1	0	0	1	0	0
5	0	1	0	1	0	0
6	0	1	0	0	1	0
7	0	1	0	0	1	0
8	0	1	0	0	1	0
9	0	1	0	0	1	0
10	0	1	0	0	1	0
11	0	0	1	0	0	1
12	0	0	1	0	0	1
13	0	0	1	0	0	1
14	0	0	1	0	0	1
15	0	0	1	0	0	1

Distance entre modalités

- Distance entre modalités `oui` et `pauvre`
- Distance euclidienne classique

$$d^2(\texttt{oui}, \texttt{pauvre}) = \frac{1}{15} \sum_{i=1}^{15} (U_{i1} - U_{i4})^2 = \frac{1}{15}$$

- Interprétation : on cherche à mesurer à quel point les modalités `oui` et `pauvre` ne sont pas associées
  - Un individu est `pauvre` et répond `peut-etre` $\Rightarrow 1$
  - Les modalités `oui` et `pauvre` sont associées 4 fois $\Rightarrow 0$
  - Les autres individus ne sont pas concernés par ces modalités $\Rightarrow 0$
  - $d^2(\texttt{oui}, \texttt{pauvre})$ compte combien de fois (sur 15) les modalités `oui` et `pauvre` ne sont pas prises ensemble

Exemple (suite)

Même étude sur $n = 15$ dans un autre arrondissement

	reponse	statut
1	oui	pauvre
2	peut-etre	pauvre
3	peut-etre	aise
4	peut-etre	aise
5	peut-etre	aise
6	peut-etre	aise
7	peut-etre	aise
8	non	richissime
9	non	richissime
10	non	richissime
11	non	richissime
12	non	richissime
13	non	richissime
14	non	richissime
15	non	richissime

Tableau disjonctif complet

	oui	peut-etre	non	pauvre	aise	richissime
1	1	0	0	1	0	0
2	0	1	0	1	0	0
3	0	1	0	0	1	0
4	0	1	0	0	1	0
5	0	1	0	0	1	0
6	0	1	0	0	1	0
7	0	1	0	0	1	0
8	0	0	1	0	0	1
9	0	0	1	0	0	1
10	0	0	1	0	0	1
11	0	0	1	0	0	1
12	0	0	1	0	0	1
13	0	0	1	0	0	1
14	0	0	1	0	0	1
15	0	0	1	0	0	1

Distance entre modalités

- Les modalités `oui` et `pauvre` ne sont pas prises ensemble 1 fois sur 15
- Distance euclidienne classique

$$d^2(\texttt{oui}, \texttt{pauvre}) = \frac{1}{15} \sum_{i=1}^{15} (U_{i1} - U_{i4})^2 = \frac{1}{15}$$

Problème : même distance dans les 2 cas alors que

- Cas 1 : modalités proches (4 individus `oui` et `pauvre`)
- Cas 2 : modalités moins proches (1 individu `oui` et `pauvre`)
- $\Rightarrow$ modifier la distance pour qu'elle soit plus faible dans le cas 1 que dans le cas 2

Distance modifiée

- Division de chaque colonne $k(j)$ (modalité j de la variable k) de U par la fréquence d'apparition de la modalité $k(j)$ (notée $n_j^{(k)}/n$)
- Cas 1

$$d^2(\texttt{oui}, \texttt{pauvre}) = \frac{1}{15} \sum_{i=1}^{15} \left(\frac{U_{i1}}{4/15} - \frac{U_{i4}}{5/15} \right)^2 = 0.0375$$

- Cas 2

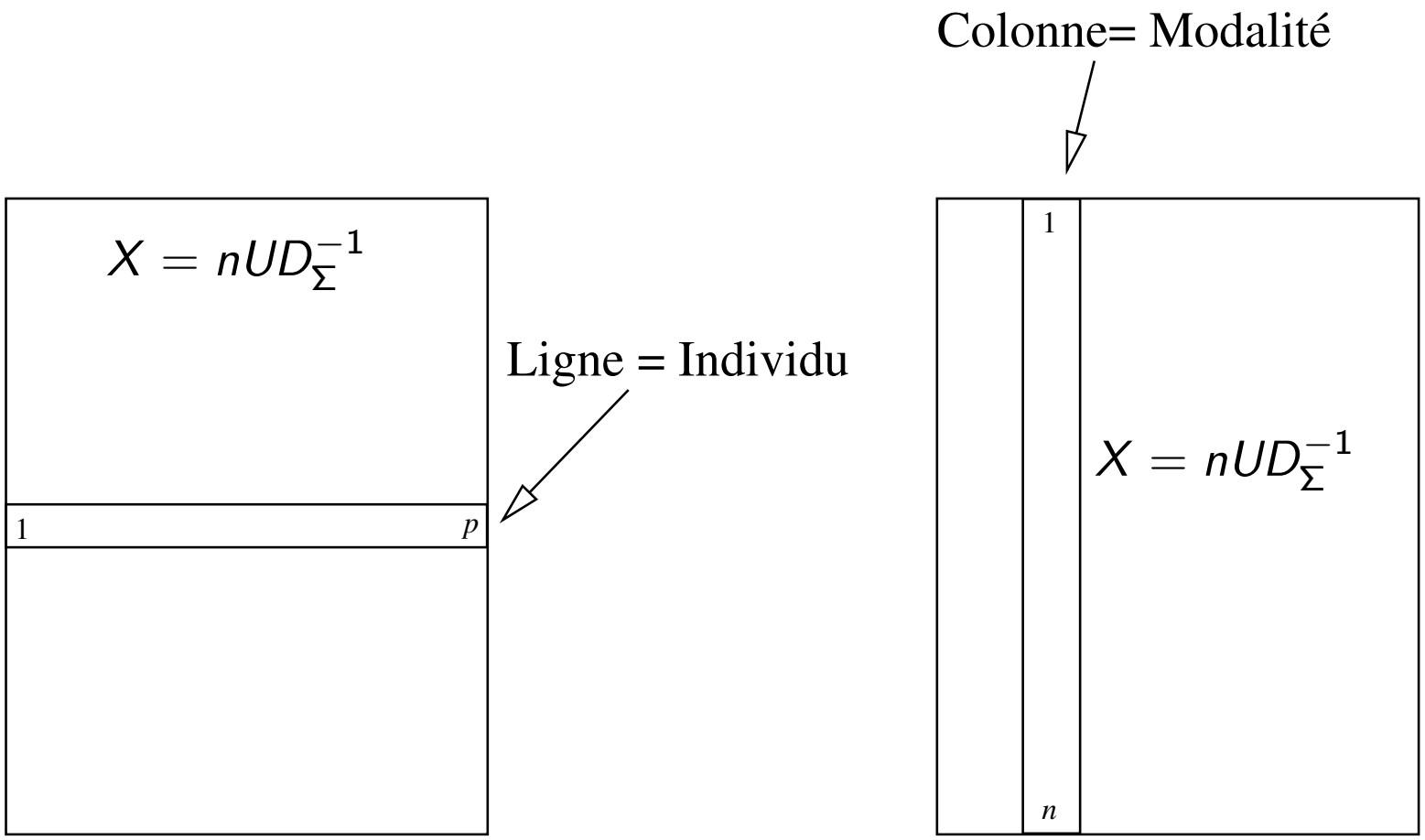
$$d^2(\texttt{oui}, \texttt{pauvre}) = \frac{1}{15} \sum_{i=1}^{15} \left(\frac{U_{i1}}{1/15} - \frac{U_{i4}}{2/15} \right)^2 = 3.75$$

- Ce faisant, on réduit la distance lorsque les associations sont nombreuses
- $\Rightarrow$ cette distance (issue du produit scalaire) permet bien de distinguer les 2 cas

Résumé

- Tableau de variables qualitatives X
- Transformation de X en un tableau disjonctif complet U
- Division des colonnes par la frequence de la modalité (colonne) en question $\rightarrow$ tableau nUD_{Σ}^{-1}
- ACP du tableau nUD_{Σ}^{-1} (pas de centrage/réduction)

Tableau pour l'ACP



Les données

- Enquête menée sur en 2008 135 personnes
- Sujet : prise de position sur les OGM
- 2 groupes de questions :
  - Lien aux OGM des personnes interrogées : 16 questions (variables actives)
  - Variables de signalétique : 5 questions (variables supplémentaires)
- Objectifs :
  - Typologie des individus selon leur rapport aux OGM
  - Voir le lien avec les variables de siganlétique

Questions sur les OGM

- Vous sentez-vous concerné(e) par la polémique sur les OGM ?
  `beaucoup`, `moyennement`, `un peu`, `pas du tout`
- Quelle est votre position quant à la culture d'OGM en France ?
  `favorable`, `plutôt défavorable`, `pas favorable du tout`
- Quelle est votre position quant à l'incorporation de matière première OGM dans les produits alimentaires destinés à l'alimentation humaine ?
  `favorable`, `plutôt défavorable`, `pas favorable du tout`
- Quelle est votre position quant à l'incorporation de matière première OGM dans les produits alimentaires destinés à l'alimentation animales ?
  `très favorable`, `favorable`, `plutôt défavorable`, `pas favorable du tout`

Questions sur les OGM

- Avez-vous déjà participé à une manifestation contre les OGM ? `oui`, `non`
- Faites-vous vous même la démarche de vous informer sur le sujet ? `oui`, `non`
- Pensez-vous que l'utilisation d'OGM puisse permettre la réduction d'usage des fongicides ? `oui`, `non`
- Pensez-vous que l'utilisation d'OGM puisse permettre la réduction des problèmes de famine dans le monde ? `oui`, `non`
- Pensez-vous que l'utilisation d'OGM puisse permettre l'amélioration des conditions de vie des agriculteurs ? `oui`, `non`

Questions sur les OGM

- Pensez-vous que l'utilisation d'OGM puisse permettre de futurs progrès scientifiques ? `oui`, `non`
- Pensez-vous que les OGM représentent un éventuel danger pour notre santé ? `oui`, `non`
- Pensez-vous que les OGM représentent une menace pour l'environnement ? `oui`, `non`
- Pensez-vous que les OGM représentent un risque économique pour les agriculteurs ? `oui`, `non`
- Pensez-vous que les OGM représentent un procédé scientifique inutile ? `oui`, `non`
- Pensez-vous que nos grand-parents avaient une alimentation plus saine ? `oui`, `non`

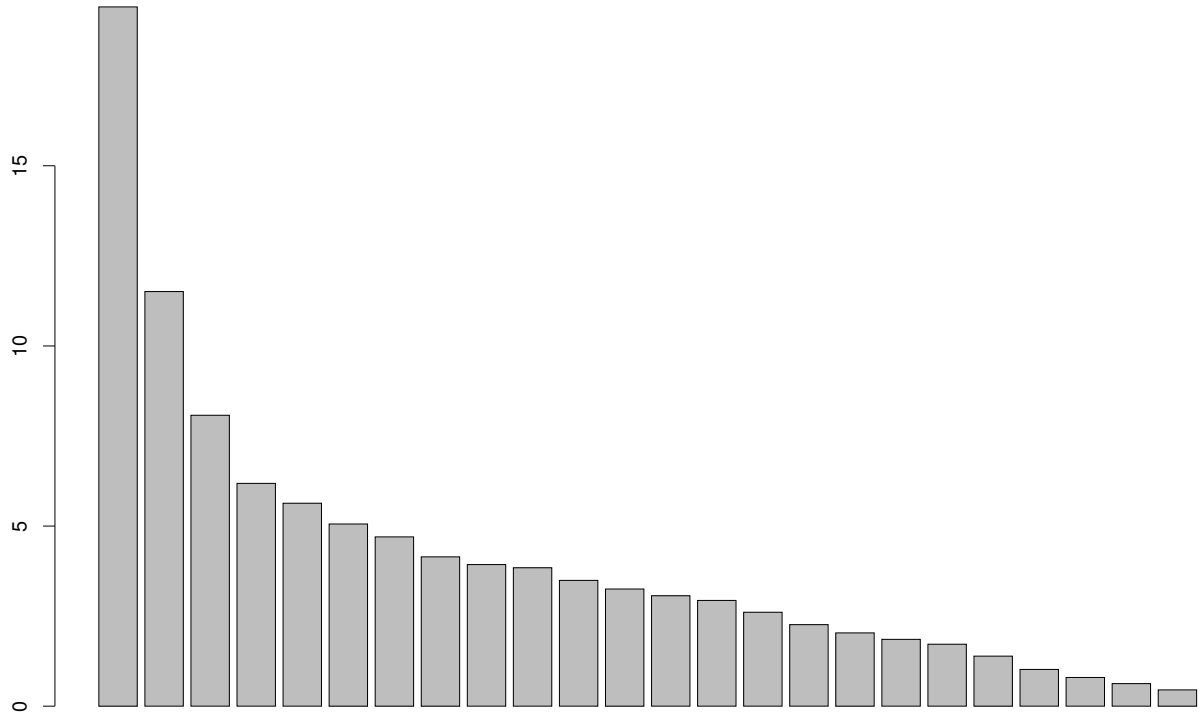
Questions de signalétique

- Sexe `masculin`, `féminin`
- Catégorie socio-professionnelle `agriculteur`, `étudiant`, `ouvrier`, `carde`, `fonction publique`, `libéral`, `technicien`, `commerçant`, `autre actif`, `non actif`, `retraité`
- Age `-25 ans`, `25-40 ans`, `40-60 ans`, `+60 ans`
- Exercez-vous des études, un métier en rapport avec l'agriculture ou la pharmaceutique ? `oui`, `non`
- A quel parti politique vous identifiez-vous le plus ?
  `extrême gauche`, `verts`, `PS`, `centre`, `UMP`, `FN`

FactoMineR

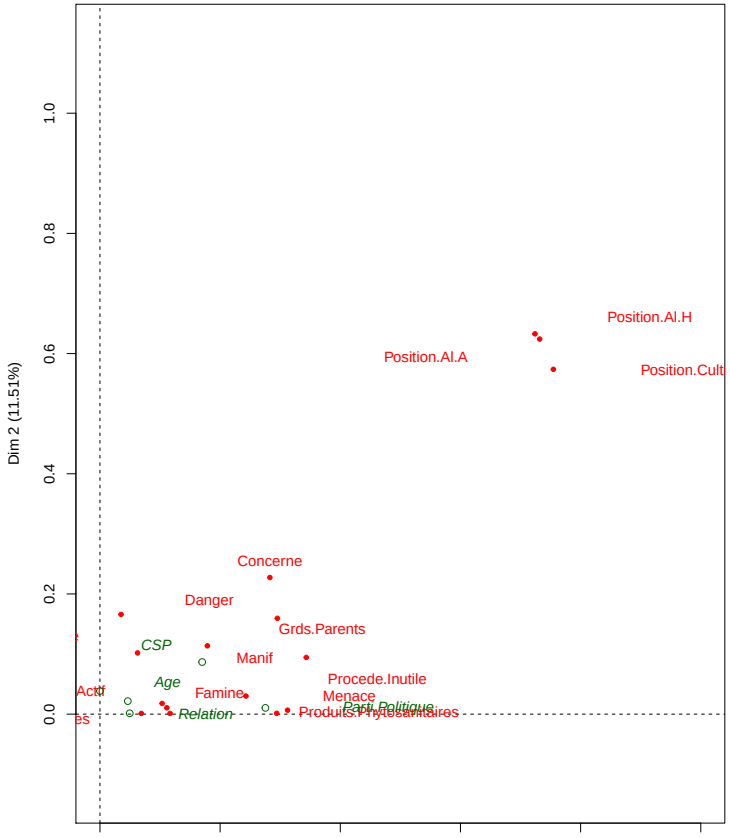
```
library(FactoMineR)
res <- MCA(ogm, ncp=5, quali.sup=17:21, graph= FALSE)
```

Choix du nombre d'axes : 2



Graphe des variables

- Quelles variables contribuent à séparer les points sur les axes ?
- $\Rightarrow$ Rapports de corrélation entre variables et coordonnées des individus



Graphe des variables

- Trois variables participent beaucoup de la typologie des individus sur l'axe 1 et l'axe 2
- Quelles modalités sont associées ?

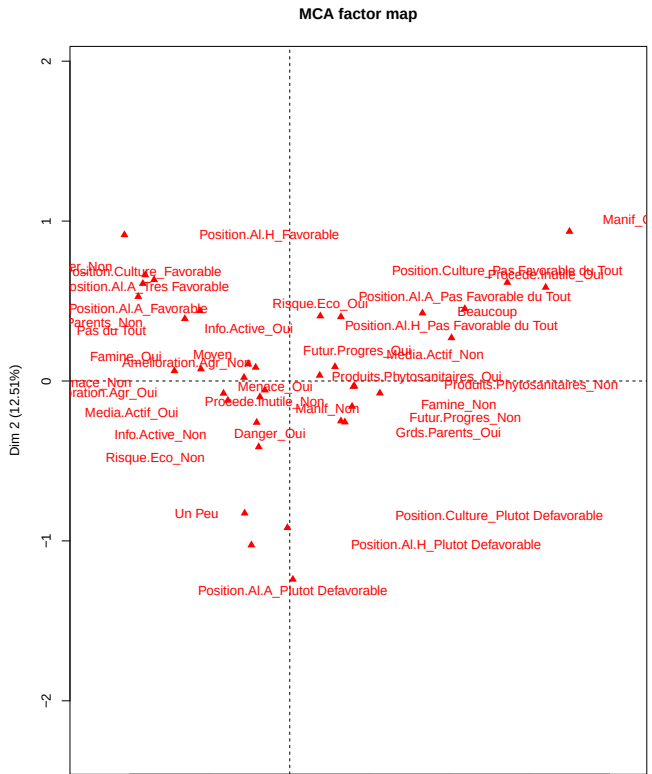
Exemple : Position culture $\times$ OGM dans l'alimentation H

	Favorable	Pas favorable	Plutôt défavorable
Favorable	34	1	13
Pas favorable	0	32	1
Plutôt defavorable	4	17	33

- Forte structuration des individus selon ces variables

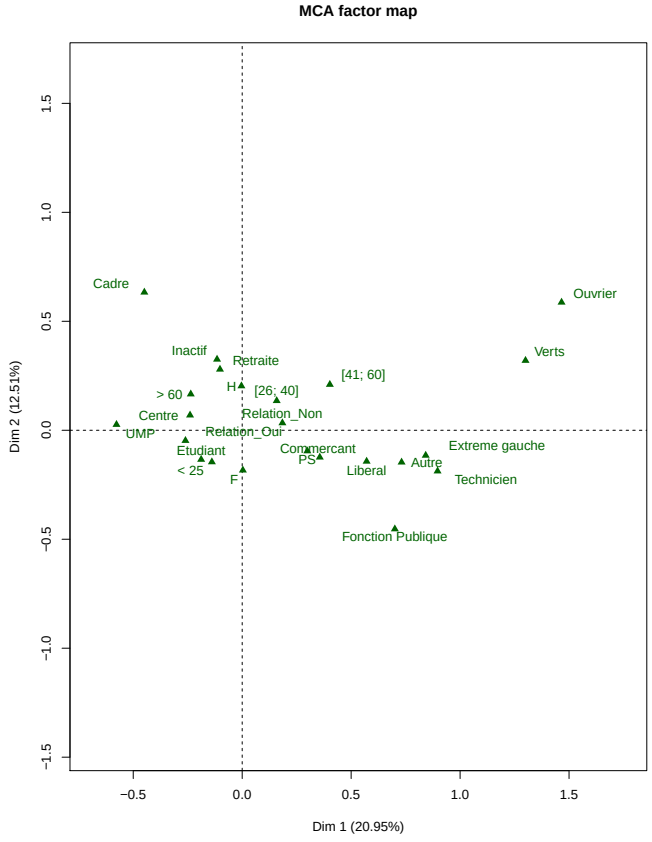
Graphe des modalités

- Axe 1 : Opposition
  - individus concernés par la question des OGM, défavorables à leur utilisation
  - individus peu concernés et plutôt favorables
- Axe 2 : axe des avis moins tranchés

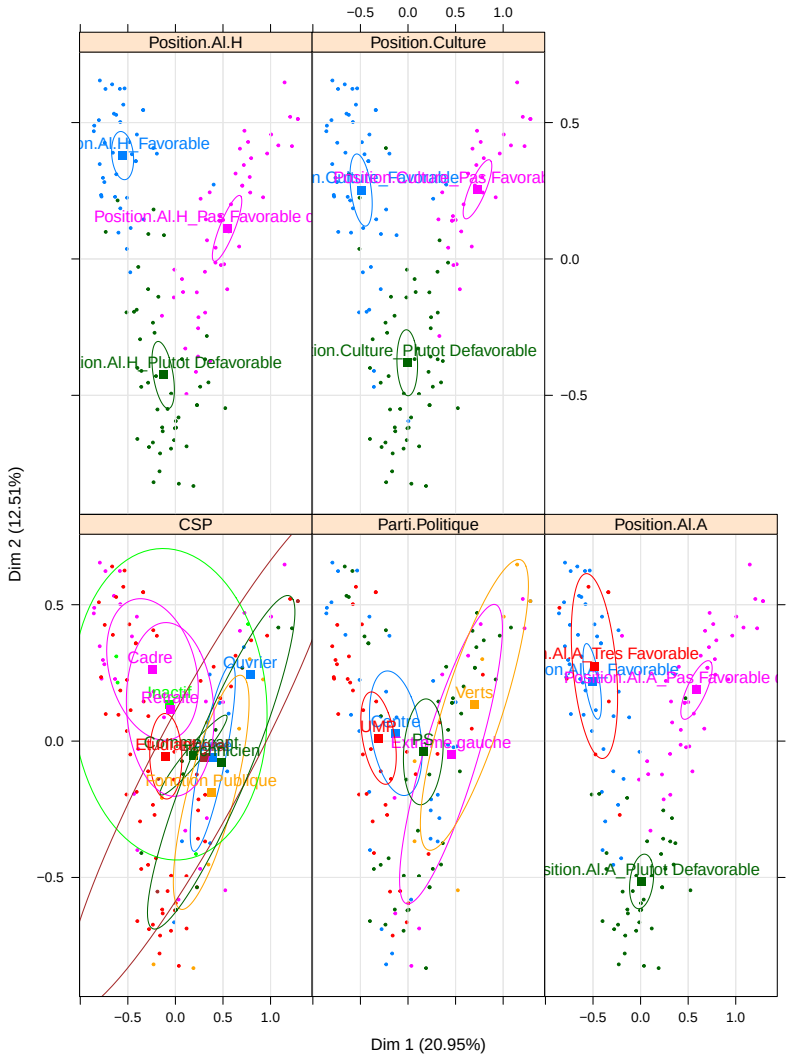


Variables de signalétique

- Structuration forte pour les variables CSP et parti politique



Ellipses


Position.Al.H
Position.Culture
CSP
Parti.Politique
Position.Al.A
Position.Al.H_Favorable
Position.Al.H_Ras Favorable
ion.Al.H_Plutot Defavorable
ion.Culture_Plutot Defavorable
Cadre
Ouvrier
Inactif
Retraite
Profession Publique
Verts
UMP
Centre
PS
Extreme.gauche
Al.A_Tres Favorable
sition.Al.A_Plutot Defavorable
Dim 1 (20.95%)
Dim 2 (12.51%)