Analyse des Correspondances Multiples

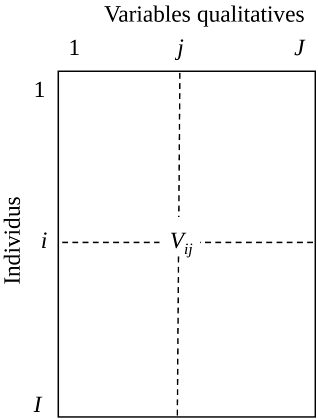
François Husson & Magalie Houée-Bigot

Laboratoire de mathématiques appliquées - AGROCAMPUS OUEST

husson@agrocampus-ouest.fr

Plan

Les données



I individus
J variables qualitatives

v_{ij} : modalité de la variable j possédée par l'individu i

Exemple : enquête où I personnes sont interrogées sur J questions à choix multiples

Les données

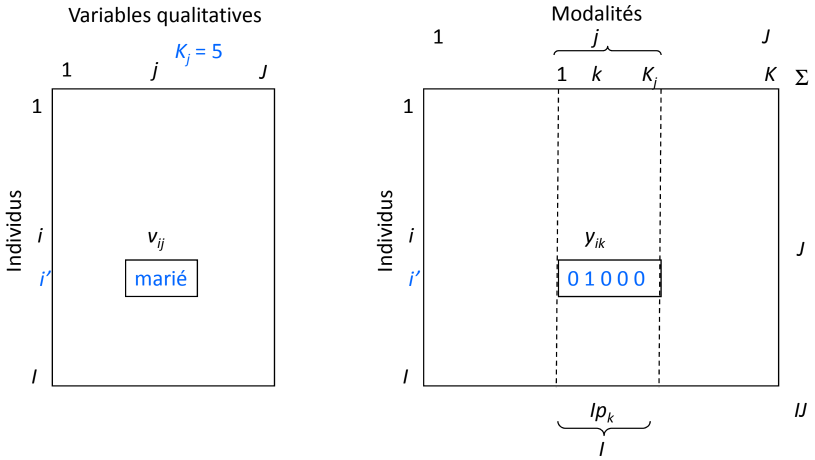


Tableau disjonctif complet (TDC)

Objectifs – problématique

1. Etude des individus

   Un individu = une ligne du TDC = ensemble de ses modalités
   Ressemblance des individus   Variabilité des individus
   Principales dimensions de la variabilité des individus
   (en relation avec les modalités)

2. Etude des variables

   Liaisons entre variables qualitatives
   (en relation avec les modalités)
   Visualisation d'ensemble des associations entre modalités
   Variable synthétique
   (Indicateur quantitatif fondé sur des variables qualitatives)

$\Rightarrow$ Problématique voisine de celle de l'ACP

Les données loisirs

- Extrait d'une enquête de l'Insee de 2003 sur la construction des identités, appelée « Histoire de vie »
- 8403 individus
- 2 sortes de variables :
    - *Parmi les loisirs suivants, indiquez ceux que vous pratiquez régulièrement* : Lecture, Ecouter de la musique, Cinéma, Spectacle, Exposition, Ordinateur, Sport, Marche, Voyage, Jouer de la musique, Collection, Activité bénévole, Bricolage, Jardinage, Tricot, Cuisine, Pêche, nombre d'heures moyen par jour à regarder la TV
    - le signalétique (4 questions) : sexe, âge, profession, statut matrimonial

Les données loisirs

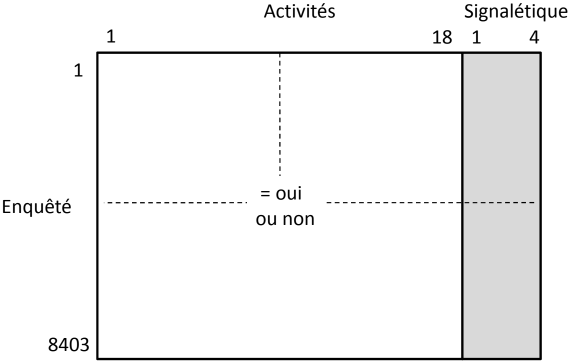
Activités pratiquées

Activité		Effectif
Ecouter de la musique		5947
Lecture		5646
Marche		4175
Cuisine		3686
Bricolage		3539
Voyage		3363
Cinéma		3359
Jardinage		3356
Ordinateur		3158
Sport		3095
Exposition		2595
Spectacle		2425
Jouer de la musique		1460
Tricot		1413
Activité.bénévole		1285
Pêche		945
Collection		862
Nb d'heure à regarder la TV	0	1017
	1	1223
	2	2156
	3	1775
	4	2232

Signalétique

Sexe	Femme	4616
	Homme	3787
Age	[15,25]	857
	(25,35]	1302
	(35,45]	1646
	(45,55]	1837
	(55,65]	1257
	(65,75]	937
	(75,85]	482
	(85,100]	85
Statut matrimonial	Divorcé	792
	Marié	4333
	remarié	404
	Seul	2140
	Veuf	734
Profession	agent de maîtrise	735
	cadre	1052
	employé	2552
	manoeuvre	792
	ouvrier	1161
	technicien	401
	autre	212
	Non réponse	1498

Les données loisirs



ACM 1 : loisirs en actif, signalétique en supplémentaire

- 1 individu = profil d'activités
- Principales dimensions de variabilité des profils d'activités
- Liaisons entre ces dimensions et le signalétique

ACM 2 : signalétique en actif, loisirs en supplémentaire
ACM 3 : loisirs et signalétique en actif

Transformation du tableau disjonctif complet

Le poids d'un individu est $\frac{1}{I}$

$$
\begin{array}{rcl}
y_{ik} & = & 1 \text{ si } i \text{ possède la modalité } k \text{ de la variable } j \text{ (quel que soit } p_k) \\
 & = & 0 \text{ sinon}
\end{array}
$$

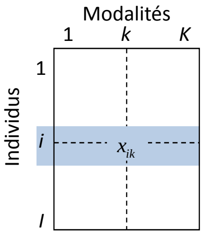
$$\text{Idée :} \quad x_{ik} = y_{ik}/p_k$$

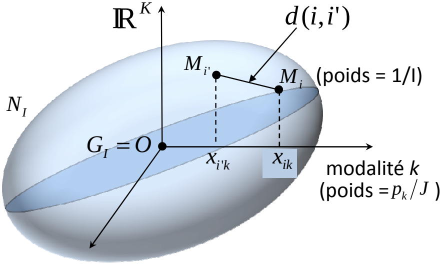
$$\frac{\sum_{i=1}^{I} x_{ik}}{I} = \frac{1}{I} \frac{\sum_{i=1}^{I} y_{ik}}{p_k} = \frac{1}{I} \frac{I \times p_k}{p_k} = 1$$

$$\text{Centrage : } x_{ik} = y_{ik}/p_k - 1$$

Plan

Nuage des individus



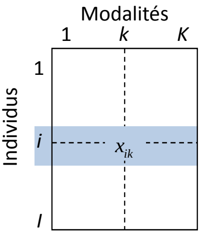


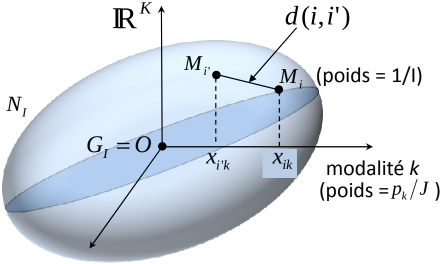
$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - x_{i'k})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$

- 2 individus prennent les mêmes modalités : distance = 0
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus dont l'un des 2 possède une modalité rare : distance grande pour prendre en compte la spécificité d'un des 2
- 2 individus ont en commun une modalité rare : distance petite pour prendre en compte leur spécificité commune

Nuage des individus

Tableau Disjonctif Complet





$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - x_{i'k})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$

$$d(i, G_I)^2 = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - 1 \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{y_{ij}}{p_k} - 1$$

$$\text{Inertie}(N_I) = \sum_{i=1}^{I} \underbrace{\frac{1}{I} d^2(i, O)}_{\text{inertie de } i} = \sum_{i=1}^{I} \left(\frac{1}{IJ} \sum_{k=1}^{K} \frac{y_{ik}}{p_k} - \frac{1}{I} \right) = \frac{K}{J} - 1$$

Ajustement du nuage des individus

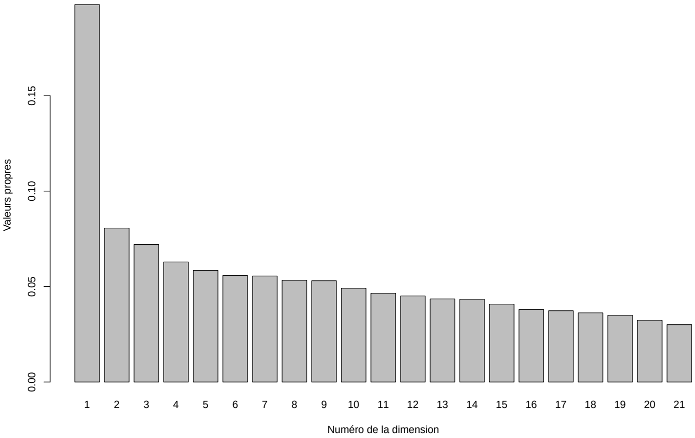
Recherche des dimensions factorielles commme pour toute méthode d'analyse factorielle

Construction séquentielle : recherche d'un axe qui maximise l'inertie et qui est orthogonal aux axes précédemment trouvés

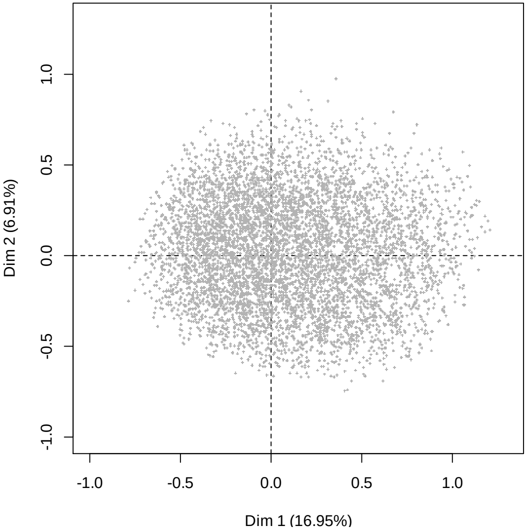
Les données loisirs

- Extrait d'une enquête de 2003 de l'Insee sur la construction des identités, appelée « Histoire de vie »
- 8403 individus
- 2 sortes de variables :
    - *Parmi les loisirs suivants, indiquez ceux que vous pratiquez régulièrement* : Lecture, Ecouter de la musique, Cinéma, Spectacle, Exposition, Ordinateur, Sport, Marche, Voyage, Jouer de la musique, Collection, Activité bénévole, Bricolage, Jardinage, Tricot, Cuisine, Pêche, nombre d'heures moyen par jour à regarder la TV
    - le signalétique (4 questions) : sexe, âge, profession, statut matrimonial

Diagramme des inerties

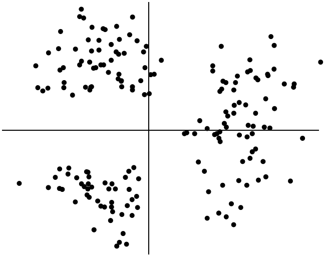


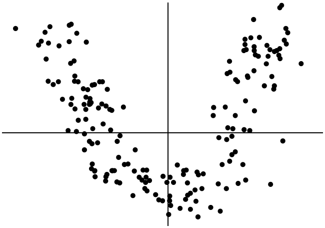
Représentation du nuage des individus



Représentation du nuage des individus

Qu'est-ce qu'une représentation particulière ?



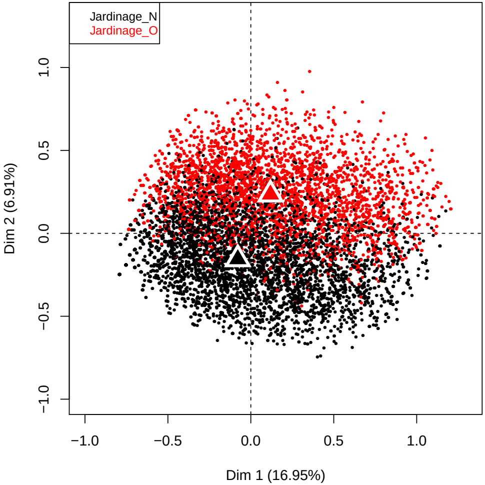


Effet Guttman

Représentation des individus en fonction du jardinage

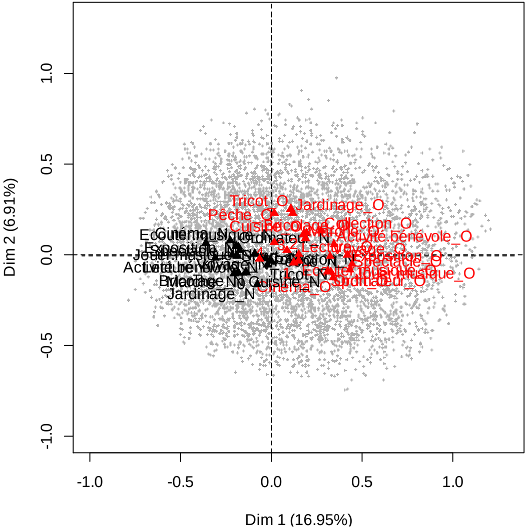
Idée : utiliser les modalités et les variables pour interpréter le graphe des individus

Une modalité au barycentre des individus qui possèdent cette modalité



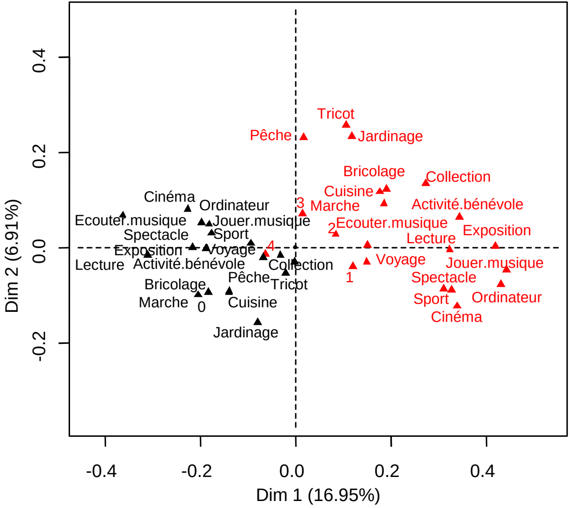
Représentation des modalités dans le nuage des individus

Chaque modalité est au barycentre des individus qui la prennent



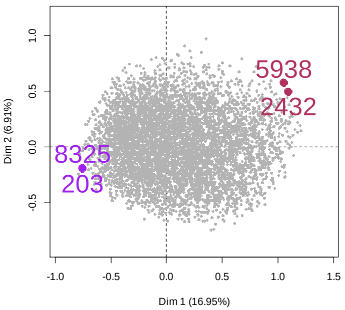
Activité non choisie – activité choisie

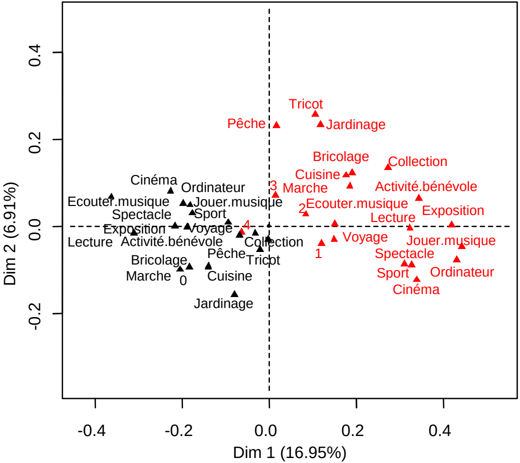
Représentation des modalités dans le nuage des individus



Activité non choisie – activité choisie

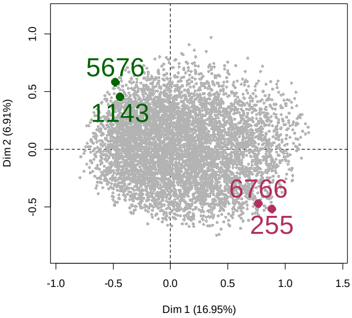
Représentation des modalités dans le nuage des individus

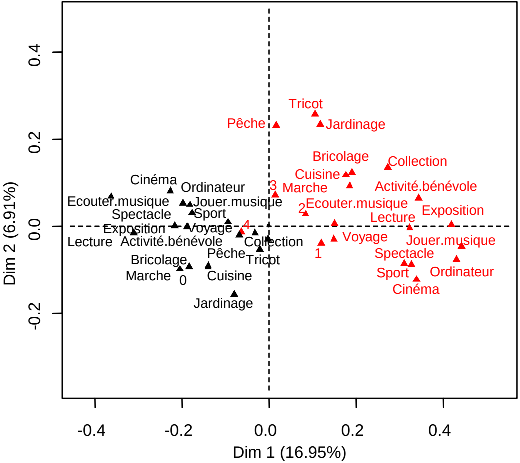




	Lecture	Ecouter musique	Ciné	Spectacle	Expo	Ordi	Sport	Marche	Voyage	Jouer musique	Collec	Activité bénévole	Bricol	Jardin	Tricot	Cuisine	Pêche	TV
5938	O	O	N	O	O	O	O	O	O	O	O	O	O	O	O	O	N	3
2432	O	O	O	O	O	O	N	O	O	O	O	O	O	O	O	O	N	2
8325	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	4
203	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	4

Représentation des modalités dans le nuage des individus





	Lecture	Ecouter musique	Ciné	Spectacle	Expo	Ordi	Sport	Marche	Voyage	Jouer musique	Collec	Activité bénévole	Bricol	Jardin	Tricot	Cuisine	Pêche	TV
255	O	O	O	O	O	O	O	O	O	O	N	O	N	N	N	N	N	1
6766	O	O	O	O	O	O	O	O	O	O	N	N	N	N	N	O	N	0
5676	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	O	N	4
1143	O	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	N	4

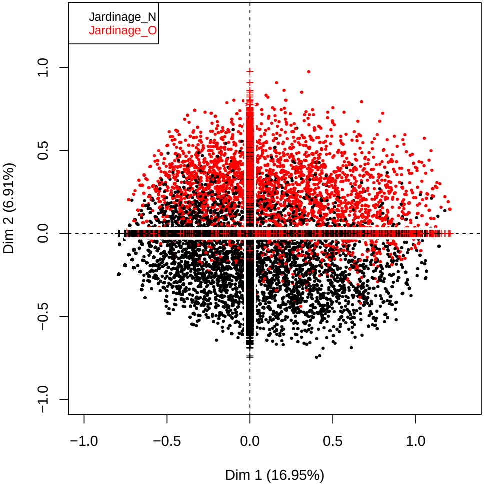
Représentation des variables pour interpréter les dimensions

Idée : considérer les coordonnées des projetés des individus sur un axe et calculer un indicateur de liaison entre ces coordonnées et chaque variable qualitative

Rapport de corrélation entre la variable j et la composante s : $\eta(v_{.j}, F_s)$

$$\eta^2(F_2, Jardinage) = 0.453$$

$$\eta^2(F_1, Jardinage) = 0.047$$

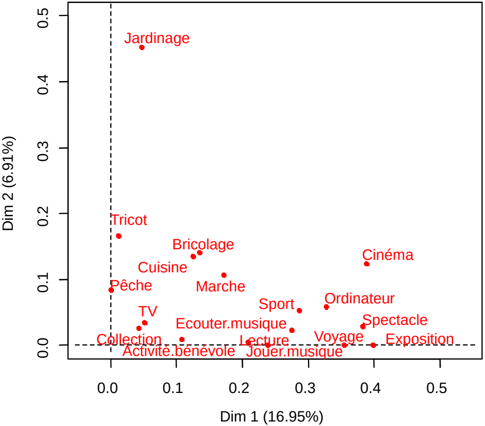


Représentation des variables pour interpréter les dimensions

Utilisation des rapports de corrélation au carré

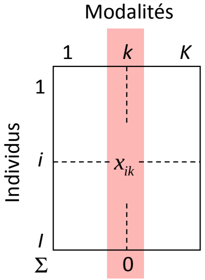
L'axe s est orthogonal à tout axe t ($t < s$) et est le plus lié aux variables qualitatives au sens du η^2 :

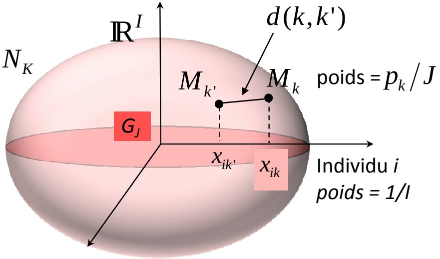
$$F_s = \max_F \sum_{j=1}^{J} \eta^2(F, v_{.j})$$



Plan

Nuage des modalités





$$Var(k) = d^2(k,O) = \sum_{i=1}^{I} \frac{1}{I} x_{ik}^2 = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - 1 \right)^2 = \frac{1}{p_k} - 1$$

	p_k	$1/2$	$1/5$	$1/10$	$1/101$
	$d(k,O)$	1	2	3	10
(si $J = 10$)	$Inertie(k)$	0.05	0.08	0.09	0.099

$$Inertie(k) = \frac{p_k}{J} d^2(k,O) = \frac{1-p_k}{J}$$

$$d^2(k,k') = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

Inertie d'une modalité et d'une variable

$$Inertie(k) = \frac{1 - p_k}{J}$$

$$Inertie(j) = \frac{1}{J} \sum_{k=1}^{K_j} (1 - p_k) = \frac{K_j - 1}{J}$$

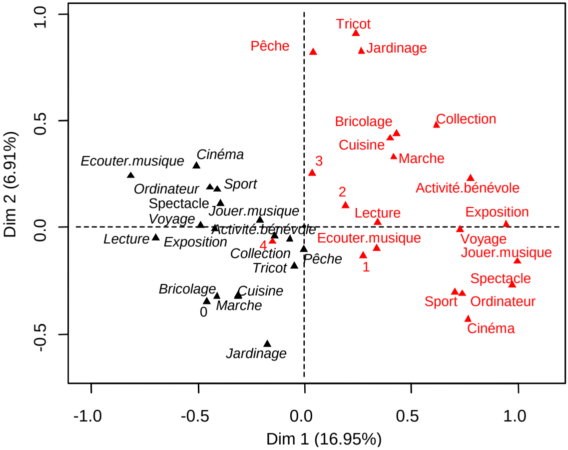
Variable	Nb de modalité	Inertie	nb dim. du sous-espace
sexe	2	$1/J$	1
région	21	$20/J$	20
département	96	$95/J$	95

MAIS : l'inertie $\frac{K_j - 1}{J}$ se répartit dans un ss-espace à $K_j - 1$ dim.

$$Inertie\ totale = \sum_{j=1}^{J} \frac{K_j - 1}{J} = \frac{K}{J} - 1$$

Ajustement du nuage des modalités

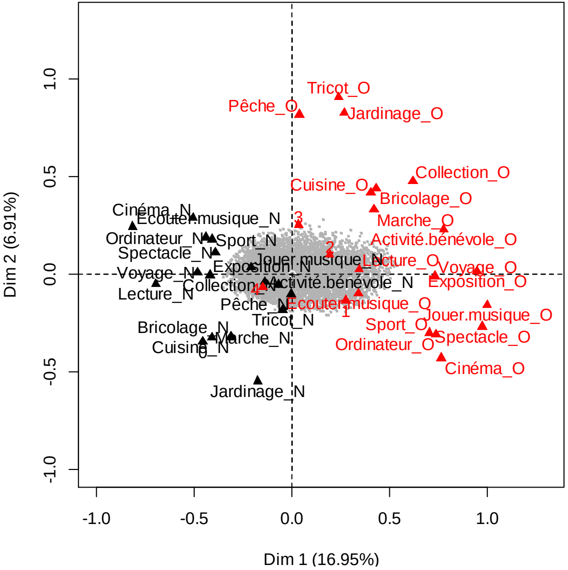
Recherche séquentielle des dimensions commme pour toute méthode d'analyse factorielle : un axe doit maximiser l'inertie et être orthogonal aux axes précédents



Activité non choisie – activité choisie

Projection des individus

Chaque individu au barycentre des modalités qu'il possède



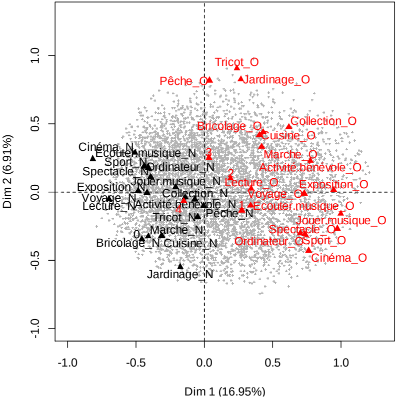
Représentations barycentriques – représentation simultanée

Représentation optimale des individus
Modalités au pseudo-barycentre :

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{y_{ik}}{I_k} F_s(i)$$

Représentation optimale des modalités
Individus au pseudo-barycentre :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{y_{ik}}{J} G_s(k)$$



Plan

Inertie et pourcentage d'inertie en ACM

$$\lambda_s = \frac{1}{J} \sum_{j=1}^{J} \eta^2(F_s, v_{.j})$$

$\Rightarrow \lambda_s$ est la moyenne des carrés des rapports de corrélation

- Individus vivent dans $\mathbb{R}^{K-J} \Rightarrow$ pourcentages d'inertie faibles
- Pourcentage maximum pour une dimension s :

$$\begin{aligned} \frac{\lambda_s}{\sum_{t=1}^{K-J} \lambda_t} \times 100 \quad &\leq \quad \frac{1}{\frac{K-J}{J}} \times 100 \\ &\leq \quad \frac{J}{K-J} \times 100 \end{aligned}$$

Avec $K = 100$, $J = 10$: $\lambda_s \leq 11.1$ %

- Moyenne des valeurs propres non nulles : $\frac{1}{K-J} \times \sum_t \lambda_t = \frac{1}{K-J} \times \left(\frac{K}{J} - 1\right) = \frac{1}{J}$

  $\Rightarrow$ interpréter les dimensions d'inertie supérieure à $1/J$

Contribution et qualité de représentation

- Contribution et $\cos^2$ pour les individus et les modalités

  $\Rightarrow$ Modalités extrêmes ne contribuent pas nécessairement beaucoup (cela dépend de leur fréquence)
  $\Rightarrow$ $\cos^2$ petits ... ce qui est attendu car bcp de dimensions

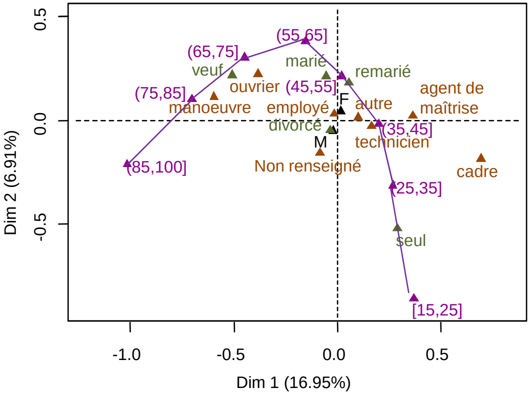
- Contribution absolue d'une variable :

$$CTR(j) = \sum_{k=1}^{K_j} CTR(k) = \frac{\eta^2(F_s, v_{.j})}{J}$$

- Contribution relative : $CTR(j) = \frac{\eta^2(F_s, v_{.j})}{J\lambda_s}$

Représentation des modalités supplémentaires

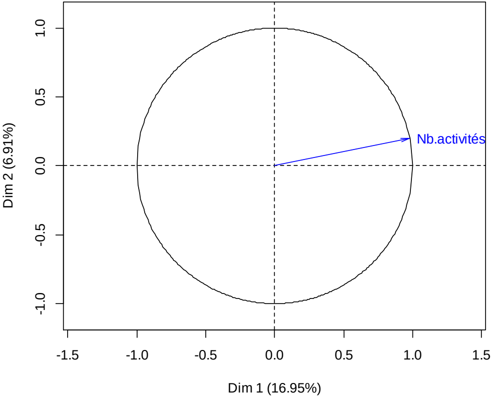
Utilisation des relations de transition pour les éléments (individus, modalités) supplémentaires



Variable quantitative supplémentaire

$\Rightarrow$ Comment faire avec les variables quantitatives ?

- Information supplémentaire : projetée sur les dimensions, coefficient de corrélation calculé avec chaque dimension
- Information active : découper la variable en classes



Description des dimensions

Par les variables qualitatives (test de Fisher), les modalités (test de Student) et les variables quantitatives (corrélation)

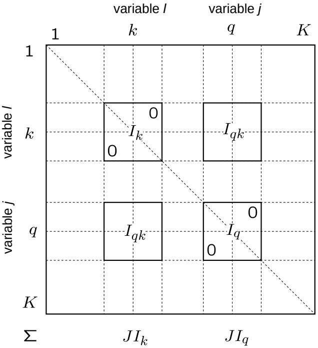
```
Variables quantitatives
              correlation p.value
Nb.activités    0.9753459       0

           Variables qualitatives           Modalités
                       R2   p.value                      Estimate p.value
Lecture             0.239  0.00e+00  Jouer.musique_O     0.268        0
Ecouter.musique     0.275  0.00e+00          Voyage_O     0.270        0
Cinéma              0.389  0.00e+00          Marche_O     0.184        0
Spectacle           0.383  0.00e+00           Sport_O     0.247        0
Exposition          0.399  0.00e+00      Ordinateur_O     0.263        0
Ordinateur          0.327  0.00e+00      Exposition_O     0.304        0
Sport               0.287  0.00e+00       Spectacle_O     0.304        0
Marche              0.172  0.00e+00           Sport_N    -0.247        0
Voyage              0.355  0.00e+00      Ordinateur_N    -0.263        0
Jouer.musique       0.209  0.00e+00      Exposition_N    -0.304        0
Bricolage           0.135 8.82e-267       Spectacle_N    -0.304        0
Cuisine             0.125 9.42e-247          Cinéma_N    -0.283        0
Profession          0.128 7.20e-245 Ecouter.musique_N    -0.257        0
Activité.bénévole 0.109 2.25e-212           Lecture_N    -0.231        0
```

Autre présentation de l'ACM : tableau de Burt

Tableau de Burt :

- Ensemble des liaisons entre variables prises 2 à 2 (tableau analogue à la matrice des corrélations entre variables quantitatives)
- Analyse des correspondances sur le tableau de Burt
- Résultats uniquement sur les modalités : même représentation mais avec des valeurs propres différentes $\lambda_s^{Burt} = (\lambda_s^{TDC})^2$
- λ_s^{TDC} moyenne des carrés des rapports de corrélation

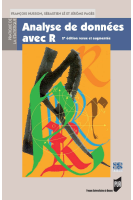


$\Rightarrow$ L'ACM ne dépend que des liaisons entre les variables prises 2 à 2 (comme l'ACP ne dépend que de la matrice des corrélations)

Conclusion

- L'ACM est la méthode factorielle adaptée aux tableaux individus $\times$ variables qualitatives
- Les valeurs propres s'interprètent comme des moyennes de rapports de corrélation au carré
- Le carré des liaisons est précieux en particulier lorsqu'il y a beaucoup de variables
- Revenir aux données en analysant des tableaux de contingence par AFC
- La convergence entre l'analyse du TDC et celle du tableau de Burt est un argument en faveur de l'intérêt de la méthode
- L'ACM comme pré-traitement d'une classification

Suppléments



Analyse de données avec R (2016), 2e édition.
Husson, Lê, Pagès.
Presses Universitaires de Rennes.

Package `FactoMineR` pour faire des ACP :
`http://factominer.free.fr/index_fr.html`

Vidéos sur Youtube :

- une chaîne Youtube : `youtube.com/HussonFrancois`
- une playlist de vidéos en français
- une playlist de vidéos en anglais