Analyse des Correspondances Multiples

ACM

AUTEUR·RICE

Benjamin Charlier, Adrien Mazoyer

1 Objectifs

1.1 Description des données : tableau condensé

On travaille sur des tableaux de données où

- les **lignes** correspondent aux **individus**
- les **colonnes** correspondent aux **variables qualitatives** étudiées (i.e. des modalités discrètes)

	Variable 1	$\cdots$	**Variable p**
Individu 1	x_{11}	$\cdots$	x_{1p}
$\vdots$	$\vdots$		$\vdots$
Individu n	x_{n1}	$\cdots$	x_{np}

x_{ik} : mesure de la variable k pour l'individu i.

Un tel tableau sera nommé tableau **condensé**, et est de taille $n \times p$.

1.2 Description des données : tableau disjonctif complet

Le tableau précédent est transformé avec un codage binaire (*one-hot encoding*)

	Variable 1			$\cdots$	**Variable** p		
	mod. 1	$\cdots$	**mod.** J_1	$\cdots$	**mod.** 1	$\cdots$	**mod.** J_p
Individu 1	$t_{1,(1,1)}$	$\cdots$	$t_{1,(1,J_1)}$	$\cdots$	$t_{1,(p,1)}$	$\cdots$	$t_{1,(p,J_p)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Individu n	$t_{n,(1,1)}$	$\cdots$	$t_{n,(1,J_1)}$	$\cdots$	$t_{n,(p,1)}$	$\cdots$	$t_{n,(p,J_p)}$

Pour chaque variable k ayant J_k modalités, on crée une colonne par modalité (k, j).

$$t_{i,(k,j)} = \begin{cases} 1 & \text{si l'individu } i \text{ possède la modalité } j \text{ de la variable } k, \\ 0 & \text{sinon.} \end{cases}$$

- Ce tableau est nommé **tableau disjonctif complet** (TDC), et est de taille $n \times J$, avec $J = \sum_{k=1}^{p} J_k$.
- Chaque individu possède exactement une modalité par variable qualitative. On a pour tout $i = 1, \ldots, n$

$$\sum_{k,j} t_{i,(k,j)} = p$$

1.3 Exemple

▸ **Tableaux**

Tableau condensé des groupes sanguins

	Groupe sanguin	Rhésus
Ind. 1	A	+
Ind. 2	O	-
Ind. 3	A	+
Ind. 4	A	-
Ind. 5	B	-

Tableau disjonctif complet associé

	Groupe sanguin			Rhésus	
	A	**B**	**O**	**+**	**-**
Ind. 1	1	0	0	1	0
Ind. 2	0	0	1	0	1
Ind. 3	1	0	0	1	0
Ind. 4	1	0	0	0	1
Ind. 5	0	1	0	0	1

1.4 Objectifs de l'ACM

Décrire les liaisons entre p variables qualitatives qui sont simultanément observées sur n individus.

- Étude des individus
  - Identifier des profils d'individus en étudiant les modalités en commun.
  - Étudier la variabilité des profils.
  - Extraire les principales dimensions.
- Étude des variables
  - Identifier les liaisons entre variables.
  - Visualisation d'ensemble d'association entre modalités.
  - Construction de variable synthétique.

> **Astuce**
>
> Effectuer une ACM consiste à appliquer l'AFC au TDC c'est à dire à effectuer une ACP pondérée des nuages des points-individus et des points-modalités (centrés). On reprendra donc largement les résultats du cours d'AFC.

2 Étude du nuage des individus

2.1 Pondération du TDC

Soit $T = (t_{ij})$ la matrice du TDC, de taille $n \times J$, où

$$t_{ij} = \begin{cases} 1 & \text{si l'individu } i \text{ possède la modalité } j, \\ 0 & \text{sinon.} \end{cases}$$

En ACM, on traite le TDC comme un tableau de contingence : la fréquence de la modalité j vaut $f_j = \frac{1}{n}\sum_{i=1}^{n} t_{ij}$.

- **Construction de la matrice X**
  - ***Valorisation de la rareté :*** on divise t_{ij} par f_j pour donner plus de poids aux modalités rares.

- ***Centrage :*** on soustrait 1 pour obtenir des colonnes de moyenne nulle :

$$x_{ij} = \frac{t_{ij}}{f_j} - 1$$

- **Interprétation**
  - **Matrice $X = (x_{ij})$:** c'est le tableau centré-réduit (au sens du χ^2) sur lequel on applique une ACP[1].
  - **Profils individus :** la ligne i représente la position de l'individu i (chacun avec un poids $1/n$).

[1] Comme en AFC cela revient a utiliser la métrique du χ^2 avec poids $D_c^{-1} \in \mathbb{R}^J$

2.2 Exemple (suite)

Pour le groupe sanguin, on a

Modalité	A	B	O	+	-
Fréquence f_j	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$

Les individus de groupes B ou O auront une contribution plus forte, ce qui met en évidence les profils spécifiques.

Le tableau transformé $X = (x_{ij})$ est alors

Individu	A	B	O	+	-
Ind. 1	$\frac{2}{3}$	-1	-1	$\frac{3}{2}$	-1
Ind. 2	-1	-1	4	-1	$\frac{2}{3}$
Ind. 3	$\frac{2}{3}$	-1	-1	$\frac{3}{2}$	-1
Ind. 4	$\frac{2}{3}$	-1	-1	-1	$\frac{2}{3}$
Ind. 5	-1	4	-1	-1	$\frac{2}{3}$

Ainsi, pour l'individu 1 qui est de groupe A,

$$x_{1,A} = \frac{1}{3/5} - 1 = \frac{5}{3} - 1 = \frac{2}{3}.$$

On procède de même pour chaque modalité de toutes les variables, puis on applique l'ACP au tableau X obtenu.

2.3 Une première ACP sur les profils lignes

On considère un nuage d'individus (ou de profils-lignes) vivant dans $\mathbb{R}^J$.

- **Objectif :** Chercher un sous-espace représenté par des axes factoriels tels que la **forme** du nuage de points soit la moins altérée possible après projection orthogonalisée.
- **Inertie et Distance :** La « forme » du nuage est caractérisée par son **inertie**, dont le calcul nécessite une métrique (notion de distance) appropriée entre deux individus i et i'.

> ⓘ **Métriques usuelles**
>
> - **ACP normée :** Métrique euclidienne usuelle $M = I_J$ (les variables sont réduites, la distance s'écrit $d(i,i')^2 = \sum_j (x_{ij} - x_{i'j})^2$).
> - **AFC / ACM :** Métrique du χ^2 avec $M = \mathrm{diag}(1/f_j)$ (pondération par l'inverse de la fréquence pour sur-pondérer les modalités rares).

2.4 Distance du χ^2 entre deux lignes

Pour tout $i, i' = 1, \dots, n$ on pose:

$$d(i,i')^2 = \sum_{j=1}^{J} \frac{f_j}{p}(x_{ij} - x_{i'j})^2 = \frac{1}{p}\sum_{j=1}^{J} \frac{1}{f_j}(t_{ij} - t_{i'j})^2$$

Donc deux individus sont proches s'ils possèdent les mêmes modalités, sachant que l'on donne plus de "poids" dans cette distance au fait que ces deux individus ont en commun une modalité rare (n_j petit):

- $d(i,i') = 0$ si i et i' prennent les mêmes modalités ;
- $d(i,i')$ est faible si i et i' ont en commun beaucoup de modalités ;
- $d(i,i')$ augmente avec le nombre de modalités différentes entre i et i' ;
- $d(i,i')$ est d'autant plus grande si i ou (exclusif) i' possède une modalité rare.

2.5 Inertie du nuage de points

- Distance d'un individu au barycentre G Comme le tableau est centré, la distance au centre de gravité G s'écrit :

$$d(i,G)^2 = \sum_{j=1}^{J} \frac{f_j}{p} x_{ij}^2 = \frac{1}{p}\sum_{j=1}^{J} f_j \left(\frac{t_{ij}}{f_j} - 1\right)^2 = \frac{1}{p}\sum_{j=1}^{J} \frac{t_{ij}}{f_j} - 1$$

- Inertie totale du nuage d'individus

$$I(L) = \text{Inertie}_{\text{tot}} = \frac{1}{n}\sum_{i=1}^{n} d(i,G)^2 = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{t_{ij}}{f_j} - 1 = \frac{J}{p} - 1$$

Remarque:

- En AFC : L'inertie totale des nuages (profils-lignes ou colonnes) s'interprète directement comme la valeur du $\frac{\chi^2}{n}$, qui mesure l'écart à l'indépendance entre les deux variables qualitatives.
- En ACM : Ce n'est plus le cas. L'inertie dépend uniquement de $\frac{J}{p} - 1$, c'est-à-dire du nombre moyen de modalités par variable moins 1.

2.6 Illustration : jeu de données `hobbies`

▸ **Description**

- 8403 individus (6905 après nettoyage)
- 17 variables qualitatives actives indiquant la pratique ou non du loisir proposé, et une 18ème pour le nombre moyen d'heures journalières à regarder la TV.
- 4 variables qualitatives supplémentaires (sexe, âge, profession, statut matrimonial).
- 1 variable quantitative supplémentaire (nombre d'activités).

```
> library(FactoMineR)
> data(hobbies)
```

	Reading	Listening music	Cinema	Show	Exhibition	Computer	Sport	Walking	Travelling	Playing music	Collecting	Volunteering	Mechanic	Gardening	Knitting	Cooking	Fishing	TV	Sex	Age	Marital status
11000210	1	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0	0	2	F	(55,65]	Married
11000410	1	0	0	0	1	0	1	1	0	0	1	1	1	1	0	0	0	4	M	(45,55]	Married
11000610	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	4	F	(25,35]	Remarried
11000710	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	M	(75,85]	Married
11000810	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3	M	(55,65]	Married
11000910	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	M	(45,55]	Married

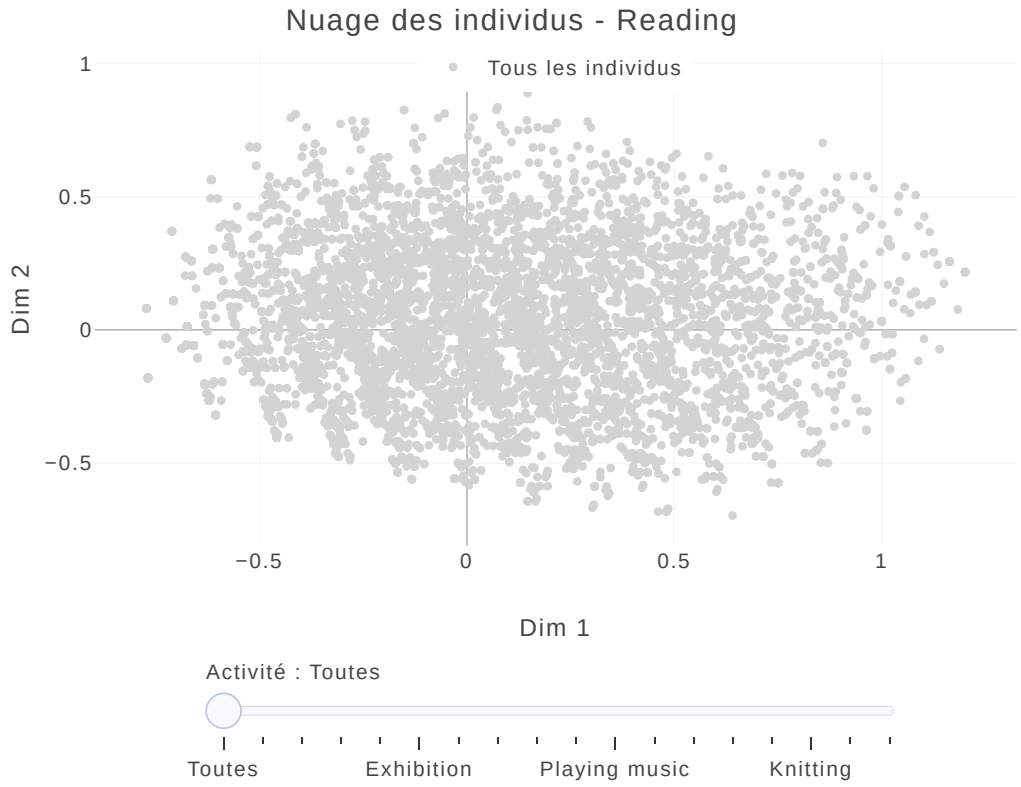
2.6.1 Nettoyage des données

```
> df <- hobbies[, 1:23]
>
> quali_vars <- c(1:22)
> df[quali_vars] <- lapply(df[quali_vars], as.factor)
>
> quanti_vars <- c(18,23)
> df[quanti_vars] <- lapply(df[quanti_vars], as.numeric)
>
> # Supprimer TOUTES les lignes contenant au moins un NA sur ces colonnes
> df_clean <- df[complete.cases(df), ]
>
> cat(dim(df), dim(df_clean))
```

```
8403 23 6905 23
```

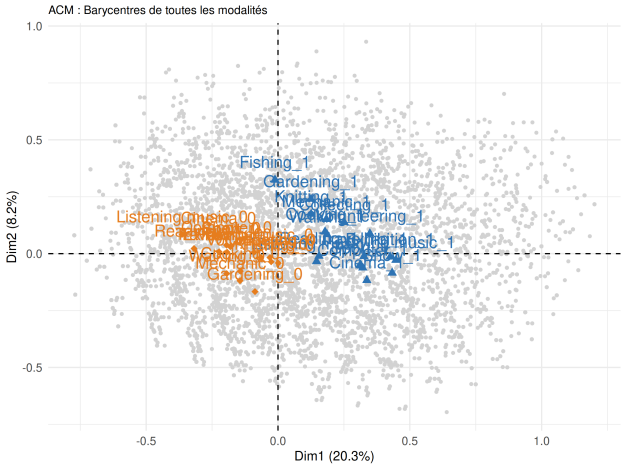
2.7 Représentation graphique du nuage des individus

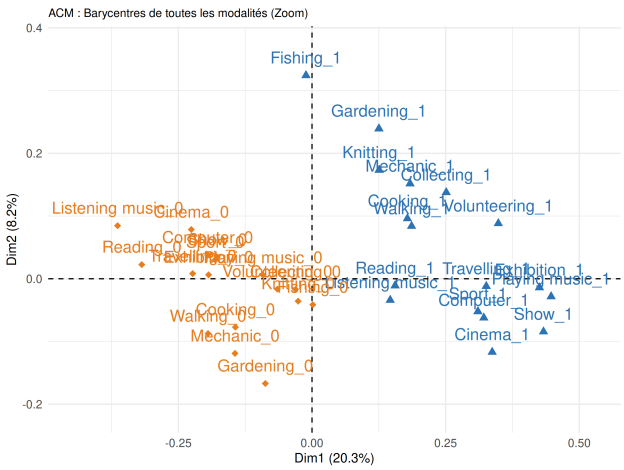
Représentation du nuage des individus sur le 1er plan principal



- Pas d'allure particulière du nuage de points.
- On peut faire apparaître les modalités et les variables pour aider à l'interprétation.

2.8 Représentation graphique des moyennes des profils lignes stratifiés par activités





On peut calculer le barycentre (euclidien) de chaque modalité :

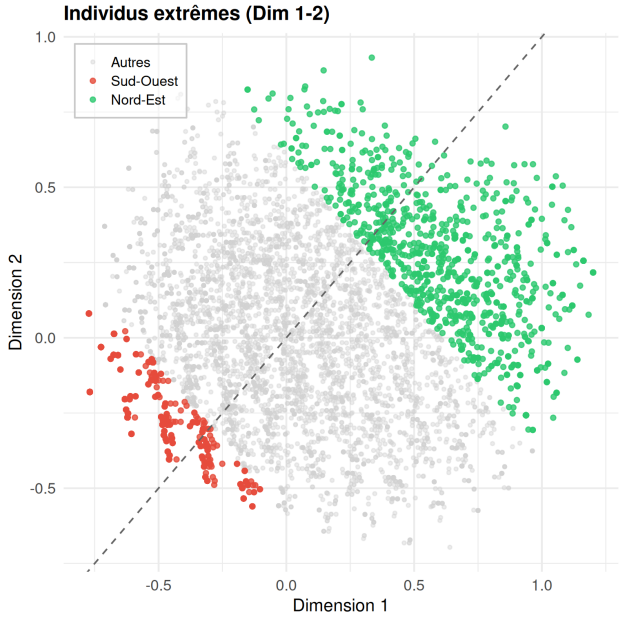
$$G^{(mod)}(j) = \sum_{i=1}^{n} \frac{t_{ij}}{n_j} C_r^{(ind)}(i)$$

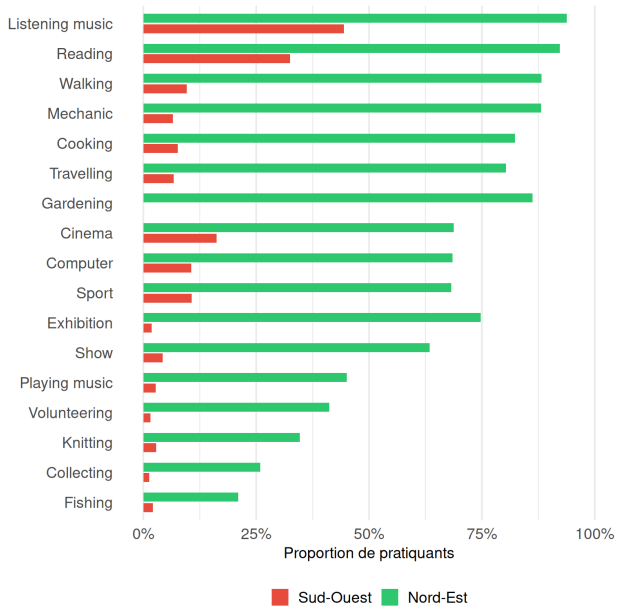
où

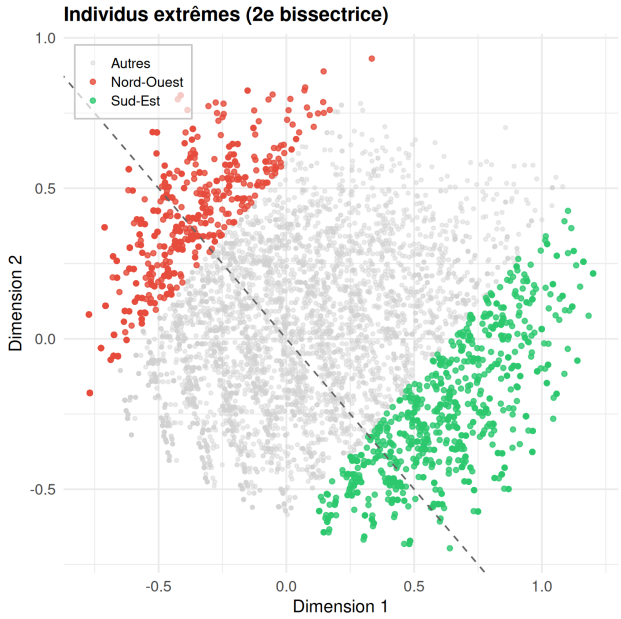
- $C_r^{(ind)}(i)$ désigne la coordonnée de l'individu i sur le r-ième axe factoriel ;
- $t_{ij} \in \{0, 1\}$ l'indicateur d'appartenance à la modalité j ;
- $n_j = \sum_{i=1}^{n} t_{ij}$ l'effectif de cette modalité.

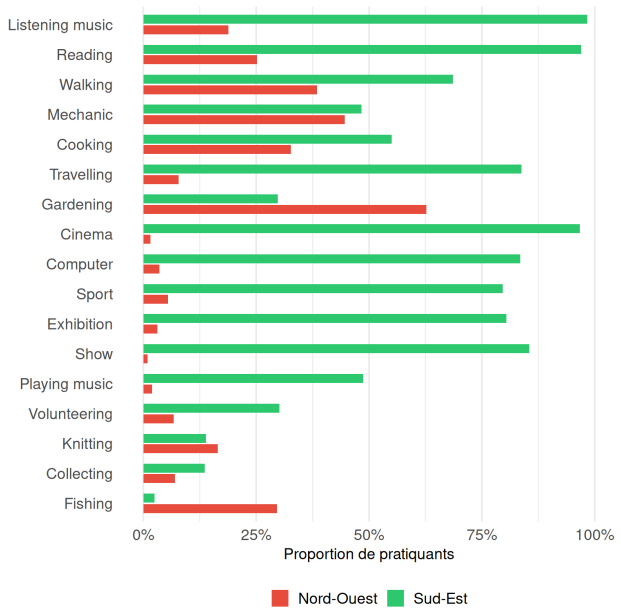
2.9 Validation empirique

- On a vu une **séparation nette des modalités** « Oui » et « Non » le long de la première bissectrice.
- On peut vérifier en observant les 1000 individus situés aux extrémités de certains axes.









2.10 Rapport de corrélation

▸ **Définition**

Pour synthétiser l'information portée par les variables qualitatives k sur chaque axe r, on utilise le **rapport de corrélation** $\eta^2_{r,k}$:

$$\eta^2_{r,k} = \frac{\text{Inertie}_{\text{inter},r,k}}{\lambda_r} \quad \in [0;1]$$

avec $\text{Inertie}_{\text{inter},r,k}$ l'inertie inter-classe de la k-ième variable sur l'axe de rang r :

$$\text{Inertie}_{\text{inter},r,k} = \sum_{j=1}^{J_k} \frac{f_j}{n} \left(\frac{1}{f_j} \sum_{i=1}^{n} t_{ij} C_r^{(\text{ind})}(i) \right)^2$$

- Plus $\eta^2_{r,k}$ est proche de 1:
  - plus la liaison est forte entre la composante r et la variable k
  - plus on observe une séparation des modalités de la variable k sur l'axe r
- Si $\eta^2_{r,k}$ est petit : les barycentres des modalités de la variable k sont tous confondus à l'origine.

3 Étude du nuage des variables

3.1 Distance du χ^2 entre deux colonnes (modalités)

Le principe est identique à celui appliqué aux individus. On munit l'espace des individus d'une métrique définie par la matrice diagonale des poids $D_r = \text{Diag}(f_i) = \frac{p}{n \cdot p} I_n = \frac{1}{n} I_n$.

- La **distance du χ^2 entre deux modalités** j et j' s'écrit alors :

$$d^2(j,j') = \sum_{i=1}^{n} f_i (x_{ij} - x_{ij'})^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{t_{ij}}{f_j} - \frac{t_{ij'}}{f_{j'}} \right)^2 = \frac{f_j + f_{j'} - 2f_{jj'}}{f_j f_{j'}}$$

où $f_{jj'} = \frac{1}{n} \sum_{i=1}^{n} t_{ij} t_{ij'}$ désigne la fréquence relative des individus possédant à la fois les modalités j et j'.

- Deux modalités j et j' sont donc d'autant plus proches qu'elles sont portées simultanément par les mêmes individus ($f_{jj'} \approx f_j \approx f_{j'}$).
- Deux modalités d'une même variable sont exclusives ($f_{jj'} = 0$), leur distance simplifiée est alors maximale : $d^2(j,j') = \frac{1}{f_j} + \frac{1}{f_{j'}}$.

3.2 Inertie du nuage des modalités

- La distance d'une modalité j au centre de gravité (origine O) s'écrit :

$$d^2(j,O) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{t_{ij}}{f_j} - 1 \right)^2 = \frac{1}{f_j} - 1$$

Cette distance au centre est d'autant plus grande que la modalité j est rare (f_j faible).

- L'**inertie de la modalité** j est donnée par :

$$\text{Inertie}_{\text{mod}}(j) = \frac{f_j}{p} d^2(j,O) = \frac{f_j}{p} \left(\frac{1}{f_j} - 1 \right) = \frac{1 - f_j}{p}$$

Malgré la baisse de son poids, plus une modalité est rare, **plus son inertie propre est grande** (car $1 - f_j$ tend vers 1).

Pré-traitement : On évite en pratique de conserver dans l'analyse des modalités trop rare

3.3 Inertie du nuage des variables

- L'**inertie d'une variable** k possédant J_k modalités est la somme des inerties de ses modalités :

$$\text{Inertie}_{\text{var}}(k) = \sum_{j=1}^{J_k} \text{Inertie}_{\text{mod}}(j) = \sum_{j=1}^{J_k} \frac{1 - f_j}{p} = \frac{J_k - 1}{p}$$

  Propriété fondamentale

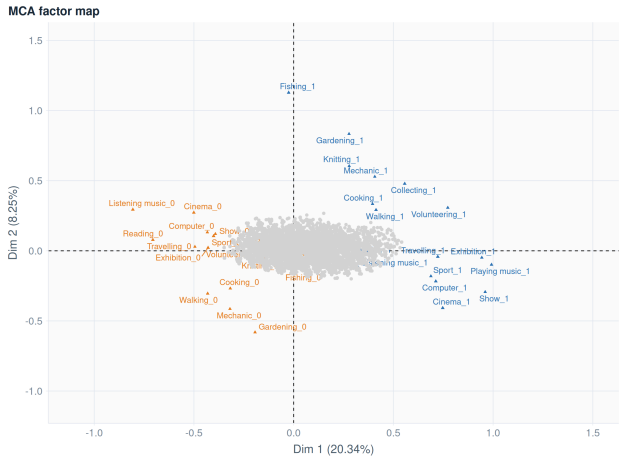
  - La part de l'inertie totale, due à une variable k est d'autant plus grande que le nombre de modalités J_k de cette variable est grand.
  - Si toutes les variables sont binaires ($J_k = 2$), chaque variable apporte exactement la même inertie au nuage : $\frac{2-1}{p} = \frac{1}{p}$.
  - **Pré-traitement :** On évite en pratique de conserver dans l'analyse des variables ayant des nombres de modalités trop différent.

- **Inertie totale** est la somme des p inerties de chaque variable

$$I(C) = \text{Inertie}_{\text{tot}} = \sum_{k=1}^{p} \text{Inertie}_{\text{var}}(k) = \frac{J}{p} - 1.$$

  On retrouve l'inertie des profils lignes $I(L)$.

3.4 Représentation des modalités

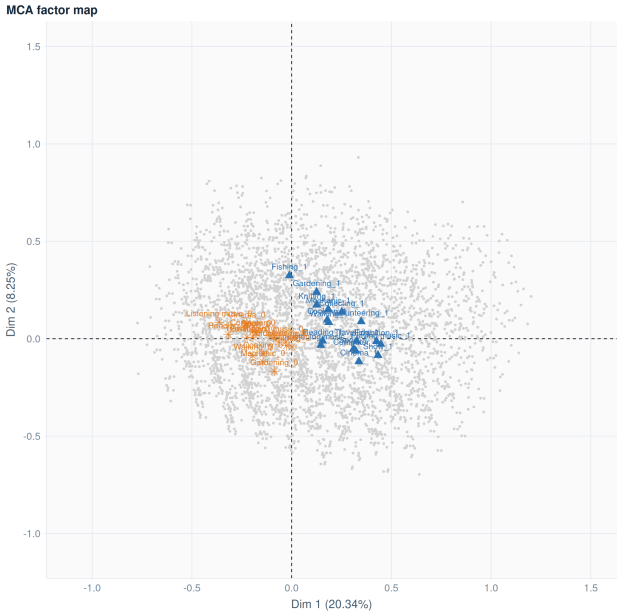


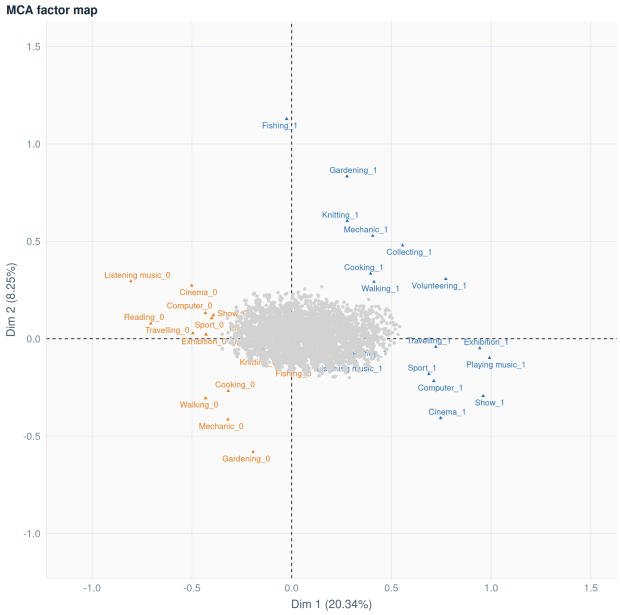
On peut tracer le nuage des points des individus comme le barycentre des modalités qu'il possède

$$G_r^{(ind)}(i) = \sum_{j=1}^{J} \frac{t_{ij}}{p} C_r^{(mod)}(j)$$

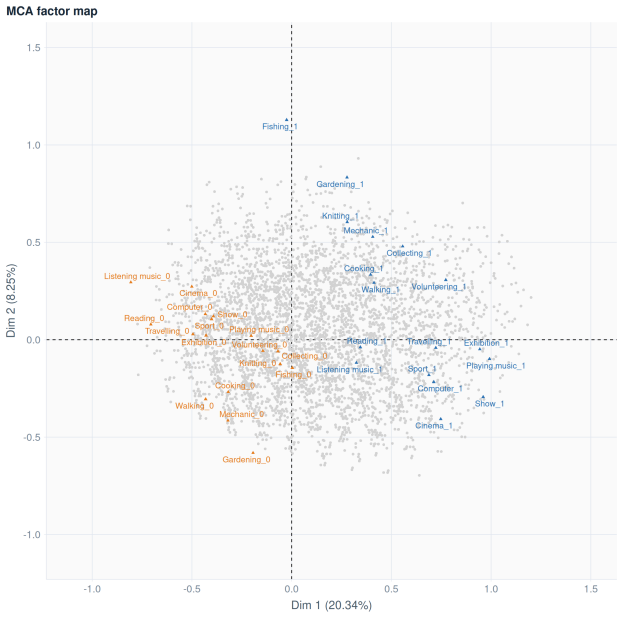
3.5 Relation de transition

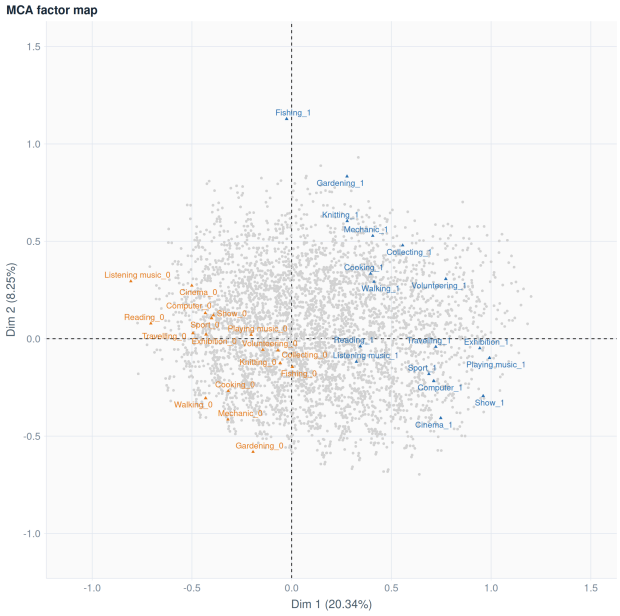
$$G_r^{(mod)}(j) = \sum_{i=1}^{n} \frac{t_{ij}}{n_j} C_r^{(ind)}(i) \qquad G_r^{(ind)}(i) = \sum_{j=1}^{J} \frac{t_{ij}}{p} C_r^{(mod)}(j)$$





$$G_r^{(mod)}(j) = \frac{1}{\sqrt{\lambda_r}} \sum_{i=1}^{n} \frac{t_{ij}}{n_j} C_r^{(ind)}(i) \qquad G_r^{(ind)}(i) = \frac{1}{\sqrt{\lambda_r}} \sum_{j=1}^{J} \frac{t_{ij}}{p} C_r^{(mod)}(j)$$





3.6 Relation de transition en R

▶ 1. Individus → \rightarrow → Modalités (Barycentre des Individus)

On calcule la position d'une modalité comme le centre de gravité des individus qui la possèdent.

Approche 1 : Propriété de transition. Utilisation des coordonnées de l'ACM (`FactoMineR`)

```
> # Coordonnées des modalités (normalisées)
> bary1_mod <- hobbies.mca$var$coord[, c("Dim 1",
      "Dim 2")]
>
> # Reprojection barycentrique (multiplication
      par √λ)
> bary1_mod <- sweep(
+   bary1_mod,
+   MARGIN = 2,
+   STATS = sqrt(hobbies.mca$eig[1:2,
      "eigenvalue"]),
+   FUN = "*"
+ )
> print_table(bary1_mod, scroll = 5)
```

	Dim 1	Dim 2
Reading_0	-0.318	0.023
Reading_1	0.155	-0.011
Listening music_0	-0.364	0.085

- **Formule :** $G_{js} \cdot \sqrt{\lambda_s} = \frac{1}{n_j} \sum_{i \in I_j} F_{is}$
- Multiplication par $\sqrt{\lambda_s}$ pour annuler la sur-dilatation des modalités et retrouver le centre de gravité exact du nuage d'individus F_{is}.

Approche 2 : Moyenne empirique directe. Calcul explicite sur les individus du tableau nettoyé

```
> # Coordonnées des individus sur les axes 1 et 2
> coords_ind <- hobbies.mca$ind$coord[, 1:2]
>
> # Calcul manuel du barycentre pour chaque
      modalité
> bary2_mod <-
      t(sapply(rownames(hobbies.mca$var$coord),
      function(mod) {
+   parts <- unlist(strsplit(mod, "_"))
+   var_name <- parts[1]
+   mod_val <- paste(parts[-1], collapse = "_")
+
+   idx <-
      which(as.character(df_clean[[var_name]])
      == mod_val)
+   colMeans(coords_ind[idx, , drop = FALSE])
+ }))
>
> print_table(bary2_mod, scroll = 5)
```

	Dim 1	Dim 2
Reading_0	-0.318	0.023
Reading_1	0.155	-0.011
Listening music_0	-0.364	0.085

- Applique directement la définition statistique du barycentre (moyenne arithmétique des coordonnées des individus $i \in I_j$).

4 Aides à l'interprétation

4.1 Inertie et valeurs propres dans l'ACM

- Il y a (au plus) $\min(n-1, J-p)$ valeurs propres non nulles[2].
- La valeur propre λ_r associée à l'axe de rang r s'écrit :

$$\lambda_r = \frac{1}{p} \sum_{k=1}^{p} \eta^2_{r,k} \in [0,1]$$

[2] Chaque individu i, la somme des modalités choisies vaut exactement p

avec $\eta^2_{r,k}$ les **rapports de corrélation**.

- Les inerties sont beaucoup plus faibles qu'en ACP/AFC ($J-p$ est très grand). Mais la moyenne des valeurs propres vaut :

$$\bar{\lambda} = \frac{1}{J-p} \sum_{r=1}^{J-p} \lambda_r = \frac{1}{p}$$

$\Rightarrow$ On interprétera en priorité les directions associées à une inertie $\lambda_r \geq \frac{1}{p}$.

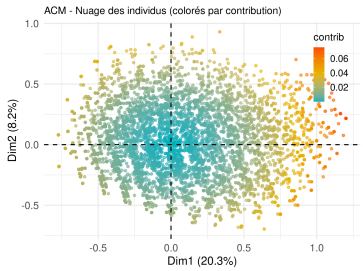
4.2 Contribution des individus

▸ **Définition**

- **Contribution de l'individu i à l'axe de rang r :**

$$\mathrm{CTR}_r^{(\mathrm{ind})}(i) = \frac{\left(C_r^{(\mathrm{ind})}(i)\right)^2}{n\,\lambda_r}$$

où $C_r^{(\mathrm{ind})}(i)$ est la coordonnée de l'individu i sur l'axe r, n le nombre total d'individus et λ_r la valeur propre de l'axe r.



Interprétation : Les individus les plus excentrés sur les axes factoriels sont ceux qui contribuent le plus.

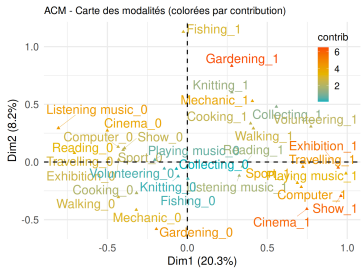
4.3 Contribution des modalités

▸ **Définition**

- **Contribution de la modalité j à l'axe de rang r :**

$$\mathrm{CTR}_r^{(\mathrm{mod})}(j) = \frac{f_j\left(C_r^{(\mathrm{mod})}(j)\right)^2}{p\,\lambda_r}$$

où $f_j = \frac{n_j}{n}$ est la fréquence relative de la modalité j, $C_r^{(\mathrm{mod})}(j)$ sa coordonnée sur l'axe r et p le nombre de variables actives.



Attention : Comme en AFC, la position sur le graphe seul ne suffit pas pour interpréter le rôle d'une modalité. Les modalités les plus excentrées ne sont pas nécessairement celles qui contribuent le plus, car leur contribution dépend aussi de leur fréquence f_j.

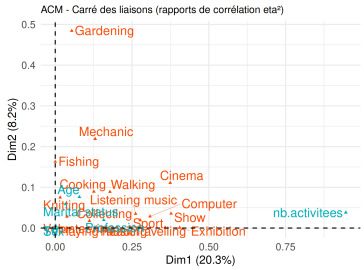
4.4 Contribution des variables

▸ **Définition**

- **Contribution de la variable k à l'axe de rang r :**

$$\mathrm{CTR}_r^{(\mathrm{var})}(k) = \sum_{j \in J_k} \mathrm{CTR}_r^{(\mathrm{mod})}(j) = \frac{\eta^2(k, F_r)}{p\,\lambda_r}$$

où J_k désigne l'ensemble des modalités de la variable k, et $\eta^2(k, F_r)$ est le rapport de corrélation entre la variable qualitative k et la composante factorielle F_r.



Interprétation : La contribution d'une variable correspond à la somme des contributions de ses modalités. Sur le carré des liaisons, les variables proches du coin supérieur droit $(1, 1)$ sont celles qui contribuent le plus à la construction des deux premiers axes.

4.5 Qualité de représentation : $\cos^2$

▸ **Description**

On utilise les mêmes formules de $\cos^2$ pour mesurer la qualité de projection des **individus** et des **modalités** sur les axes factoriels.

En pratique, si deux points sont bien projetés ($\cos^2$ élevé), leur proximité sur le plan reflète une réelle proximité dans l'espace d'origine.

ⓘ **Proximité entre deux individus**

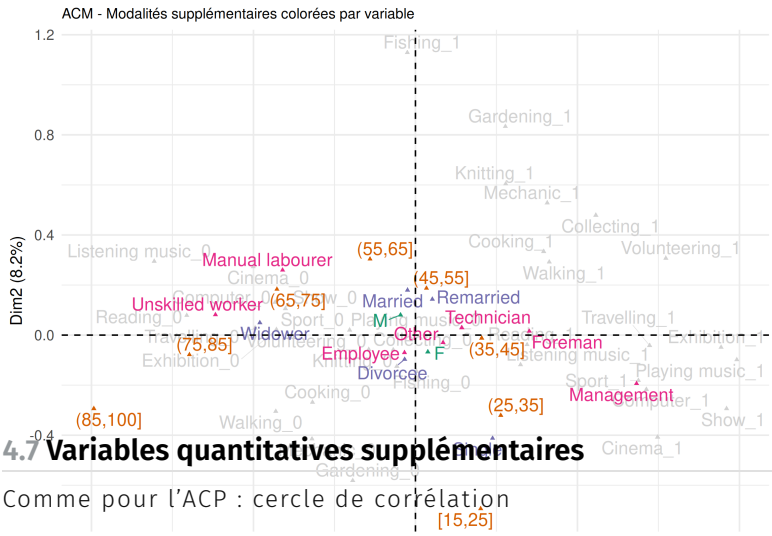
- S'interprète en termes de **distance du χ^2** : deux individus se ressemblent s'ils ont choisi les mêmes modalités.
- C'est cohérent avec la relation barycentrique : un individu se situe au centre de gravité des modalités qu'il possède.

ⓘ **Proximité entre deux modalités (de variables différentes)**

- S'interprète en termes de **distance du χ^2** : deux modalités se ressemblent si elles sont choisies par les mêmes individus.
- C'est cohérent avec la relation barycentrique : une modalité se situe au centre de gravité des individus qui la possèdent.

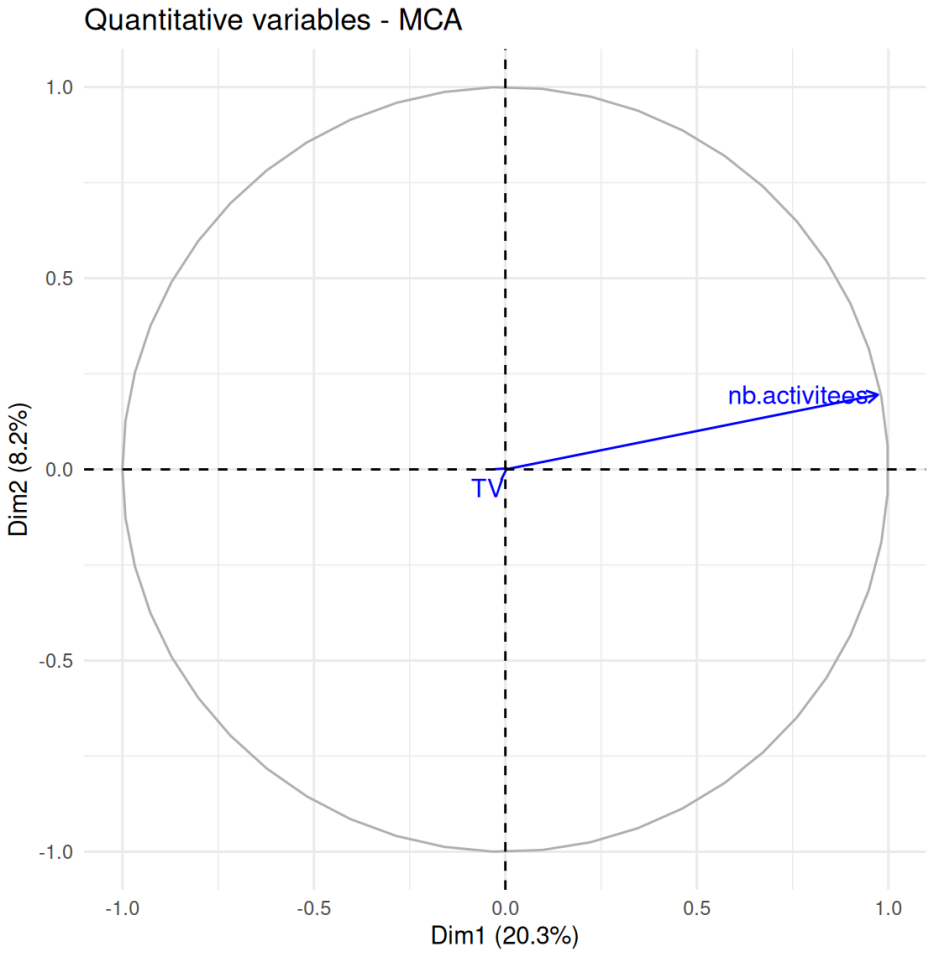
4.6 Variables qualitatives supplémentaires

Comme pour l'AFC : on se sert des relations de transitions pour calculer les coordonnées des variables



4.7 Variables quantitatives supplémentaires

Comme pour l'ACP : cercle de corrélation



4.8 Références

- (Saporta 2006)
- François Husson
- Marie Chavent
- Vivien Roussez et Pascal Irz

Saporta, Gilbert. 2006. *Probabilités, analyse des données et statistique / Gilbert Saporta,...* 2e édition révisée et augmentée. Éditions Technip.