Analyse des Correspondances Multiples (ACM)

Il est important de souligner que ce polycopié est la composition des documents, ci-dessous, que nous l'avons enrichi par quelques détails théoriques et applications.
Charlotte Baey (2019). Analyse de données. https://baeyc.github.io/teaching/
Alboukadel Kassambara (2019). http://www.sthda.com/french/
François Husson, Sébastien Lê et Jérôme Pagès (2009). Analyse des données avec R, Presses Universitaires de Rennes, 224 p. (ISBN 978-2-7535-0938-2)

1 Introduction

Dans le chapitre précédent, on s'intéressait au lien pouvant exister entre deux variables qualitatives. Dans ce chapitre, on cherche à décrire un nuage de points de p individus décrits par r variables qualitatives. Cette fois-ci, les lignes et les colonnes de la matrice des observations jouent donc des rôles différents. On peut voir l'analyse des correspondances multiples (ACM) comme une version de l'ACP adaptée aux variables discrètes.

L'Analyse des Correspondances Multiples (ACM ou MCA pour multiple correspondence analysis) donc est une extension de l'analyse factorielle des correspondances pour résumer et visualiser un tableau de données contenant plus de deux variables catégorielles.

L'ACM est généralement utilisée pour analyser des données d'enquête ou de sondage.
L'objectif est d'identifier:

1. Un groupe de personnes ayant un profil similaire dans leurs réponses aux questions.
2. Les associations entre les catégories des variables.

Exemple 1.1 *Considérons le cas suivant, où l'on observe deux variables sur $p = 4$ individus. La première variable X_1 codée 1 ou 2 selon l'individu (fumeur on non fumeur), et la deuxième variable X_2 codée $1, 2$ ou 3 selon la classe d'âge d'individus $e_i, i = 1, 2, 3, 4$:*

$$\mathbf{A} := \begin{array}{c|cc} & X_1 & X_2 \\ \hline e_1 & 1 & 3 \\ e_2 & 2 & 2 \\ e_3 & 1 & 1 \\ e_4 & 1 & 1 \end{array} .$$

Si à première vue, ce tableau est proche du type de tableau que l'on obtient en ACP, en réalité le codage des variables qualitatives est arbitraire, et cela n'a donc pas de sens de faire des opérations algébriques directement sur $\mathbf{A}$. *Dans l'exemple précédent, on aurait en effet très bien pu coder le sexe* 0 *ou* 1 *au lieu de* 1 *ou* 2. *Par ailleurs, toutes les variables qualitatives ne sont pas ordinales, et l'ordre induit par le codage peut ne pas correspondre à une réalité physique.*

1.1 Tableau disjonctif complet (TDC)

C'est pourquoi on travaille en ACM avec le tableau disjonctif complet, (TDC), qui contient autant de colonnes qu'il y a de catégories totales. Chaque variable initiale est découpée en autant de sous-variables qu'elle ne peut prendre de valeurs. Par exemple, une variable initiale correspondant au sexe de l'individu et prenant 2 valeurs possibles sera découpée en deux sous-variables. Le TDC ne contient alors que des 0 et des 1, selon que l'individu e_i appartient à la sous-catégorie considérée ou non. On le note N^*. En reprenant l'exemple précédent, on obtient le TDC suivant associé à $\mathbf{A}$:

$$N^* := \begin{pmatrix} & X_{11} & X_{12} & X_{21} & X_{22} & X_{23} \\ e_1 & 1 & 0 & 0 & 0 & 1 \\ e_2 & 0 & 1 & 0 & 1 & 0 \\ e_3 & 1 & 0 & 1 & 0 & 0 \\ e_4 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Le tableau disjonctif complet est de taille $p \times q$, avec $q = \sum_{j=1}^{r} q_j$ le nombre total de modalités, et q_j est le nombre de modalités de la variable X_j. En notant n_{ij} l'élément situé sur la ligne i et la colonne j. On a

$$n_{i\cdot} := \sum_{j=1}^{q} n_{ij} = r \quad \text{et} \quad n_{\cdot j} := \sum_{i=1}^{p} n_{ij},$$

où $n_{\cdot j}$ correspond donc à l'effectif marginal de la modalité j. La somme des élément de la matrice N^* est $n = pr$, qui peut être vue comme "l'effectif total". On note D_c^* la matrice contenant sur sa diagonale les effectifs marginaux des q catégories (i.e. la somme de chaque colonne du TDC):

$$D_c^* := \begin{pmatrix} n_{\cdot 1} & 0 & \cdots & 0 \\ 0 & n_{\cdot 2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{\cdot q} \end{pmatrix} \in \mathcal{M}(q, q).$$

Celle-ci peut être réécrite sous la forme suivante

$$D_c^* = \begin{pmatrix} D_1^* & 0 & \cdots & 0 \\ 0 & D_2^* & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & D_r^* \end{pmatrix},$$

où chaque bloc D_j^*, $1 \leq j \leq r$ correspond aux modalités de la variable j. En reprenant l'exemple précédent, on a :

$$D_c^*: = \left(\begin{array}{cc|ccc} 3 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

avec

$$D_1^* = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \text{ et } D_2^* = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Remarque. Pour chaque variable X_j, la somme des q_j colonnes du TDC correspondant à ses modalités vaut 1. Comme il y a r variables, le rang de la matrice N^* est au plus égal à $q - r$ (rang$N^* \leq q - r$).

1.2 Tableau de Burt

Une autre représentation possible du tableau **A** est sous la forme de ce que l'on appelle le *tableau de Burt*. Ce tableau, noté B, s'obtient facilement à partir du TDC: $\mathbf{B} = N^{*t}N^*$. Il est de taille $q \times q$, et contient les croisements deux à deux de toutes les variables, c'est-à-dire tous les tableaux de contingence deux à deux des variables. Toujours en reprenant notre exemple, on trouve:

$$\mathbf{B} := \left(\begin{array}{cc|ccc} 3 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 2 & 0 \\ \hline 2 & 0 & 2 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right).$$

Le tableau de Burt contient moins d'information que le tableau disjonctif complet, car on ne conserve pas l'information sur les individus: tout est agrégé au niveau de la catégorie.

2 Principe général

L'ACM consiste en une extension de l'AFC, en considérant N^* comme un tableau de contingence. Si on s'intéresse à la variabilité des individus et de leurs profils, on travaillera avec les profils-lignes, et si on s'intéresse aux liens entre les modalités, on s'intéressera aux profils-colonnes. Comme on l'a vu au chapitre 2, il y a un lien entre ces deux analyses, et dans la pratique on s'intéressera souvent à la fois aux individus et aux variables.

En pratique, on réalise une ACP du nuage des profils-lignes, en utilisant la métrique du chi-deux.

En reprenant les notations du chapitre 2, on a donc:

- Matrice des profils-lignes $X_r = D_r^{-1}N$, où N^*/n et $D_r = D_r^*/n$ avec $D_r^* := r\mathbf{Id}_p$ est la matrice contenant sur sa diagonale les sommes de chaque ligne de la matrice N^*. Ainsi $D_r^{-1} = r^{-1}\mathbf{Id}_p$.
- Métrique $M_r = D_c^{-1}$, ou $D_c := D_c^*/n$.
- L'inertie totale du nuage de points s'écrit donc

$$I_T = \text{Trace}\,(V_r M_r) \text{ où } V_r := X_r^t D_r X_r - g_r g_r^t,$$

où g_r est le centre de gravité du nuage des profils-lignes. Finalement, on cherche donc les valeurs propres de la matrice $V_r M_r$.

3 Quelques particularités de l'ACM

3.1 Distance entre deux individus

Pour calculer la distance entre deux individus, on se place dans le nuage des profils-lignes et on utilise la métrique du chi-deux D_c^{-1}. On a

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 = \frac{p}{r} \sum_{j=1}^{q} \frac{1}{n_{\cdot j}} \left(n_{ij} - n_{i'j} \right)^2.$$

Autrement dit, deux individus seront proches s'ils possèdent les mêmes modalités, en particulier s'ils ont en commun des modalités rares ($n_{\cdot j}$ petit).

3.2 Distance entre deux modalités

Pour calculer la distance entre deux modalités, on se place dans le nuage des profils-colonnes et on utilise la métrique du chi-deux D_r^{-1}. On a

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 = p \sum_{j=1}^{q} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right)^2.$$

Deux modalités sont proches si elles sont possédées par les mêmes individus.

3.3 Inertie totale

Proposition. L'inertie totale des profils-ligne est

$$I_T = \frac{q}{r} - 1.$$

Preuve. Reprenons les notations de l'AFC, appliquées au TDC N^*. Rappelons que $n_{i\cdot} = r$ (nombre de variables), pour tout les individus i et que $f_{i\cdot} = n_{i\cdot}/n = 1/p$. Ecrivons maintenant

$$\begin{aligned} I_T &= \sum_{i=1}^{p} f_{i\cdot} d^2_{\chi^2}(i, g_r) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{i\cdot}}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{i\cdot}}{n_{\cdot j}} \left(\frac{n_{ij}}{r} - \frac{n_{\cdot j}}{pr}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{r}{n_{\cdot j}} \left(\frac{n_{ij}}{r} - \frac{n_{\cdot j}}{pr}\right)^2 \\ &= r\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{1}{n_{\cdot j}} \left(\frac{n_{ij}}{r} - \frac{n_{\cdot j}}{pr}\right)^2. \end{aligned}$$

Il est claire que cette dernière expression égale à

$$\begin{aligned} & r\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{1}{n_{\cdot j}} \left(\frac{n_{ij}^2}{r^2} - 2\frac{n_{ij}n_{\cdot j}}{r^2 p} + \frac{n_{\cdot j}^2}{p^2 r^2}\right) \\ = & \frac{1}{r}\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{\cdot j}} - \frac{2r}{pr^2}\sum_{i=1}^{p} n_{\cdot j}\sum_{j=1}^{q} n_{ij} + \frac{r}{p^2 r^2}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{\cdot j}^2 \end{aligned}$$

On note que $n_{i\cdot}$ prennent les valeurs 0 ou 1, donc $n_{i\cdot}^2 = n_{i\cdot}$. Donc la dernière quantité devienne

$$\frac{1}{r}\sum_{j=1}^{q} \frac{1}{n_{\cdot j}} \left(\sum_{i=1}^{p} n_{ij}\right) - \frac{2}{pr}\sum_{j=1}^{q} n_{\cdot j} \left(\sum_{i=1}^{p} n_{ij}\right) + \frac{r}{p^2 r^2}\left(\sum_{i=1}^{p} 1\right)\sum_{j=1}^{q} n_{\cdot j}.$$

Rappelons que $\sum_{i=1}^{p} n_{ij} = n_{\cdot j}$ et $\sum_{j=1}^{q} n_{\cdot j} = n = pr$. Après des simplifications on obtient

$$I_T = \frac{q}{r} - 1.$$

Remarque 1. L'inertie totale ne dépend que du nombre de classes et du nombre de variables. Elle n'a pas d'interprétation statistique comme en ACP ou en AFC.

Remarque 2. La distance entre une modalité j et le centre de gravité du nuage des profils-colonnes est :

$$d^2_{\chi^2}(j, g_c) = \frac{p}{n_{\cdot j}} - 1.$$

En effet,

$$\begin{aligned} d^2_{\chi^2}(j, g_c) &= \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{r}{pr} \right)^2 \\ &= p \sum_{i=1}^{p} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{1}{p} \right)^2 = \\ &= p \sum_{i=1}^{p} \left(\frac{n_{ij}^2}{n_{\cdot j}^2} - 2\frac{1}{p}\frac{n_{ij}}{n_{\cdot j}} + \frac{1}{p^2} \right) \\ &= p \sum_{i=1}^{p} \frac{n_{ij}}{n_{\cdot j}^2} - 2 \sum_{i=1}^{p} \frac{n_{ij}}{n_{\cdot j}} + \sum_{i=1}^{p} \frac{1}{p} \\ &= \frac{p}{n_{\cdot j}} - 1. \end{aligned}$$

Plus une modalité est rare, plus elle est loin du centre de gravité du nuage. De même, si on exprime l'inertie totale en utilisant les points du nuage des profils-colonnes, on a

$$I_T = \sum_{j=1}^{q} f_{\cdot j} d^2_{\chi^2}(j, g_c).$$

La part d'inertie due à la modalité j est

$$f_{\cdot j} d^2_{\chi^2}(j, g_c) = \frac{1}{r} \left(1 - \frac{n_{\cdot j}}{p} \right).$$

Plus une modalité est rare, plus la part d'inertie due à cette modalité est importante.

Pour ces raisons, il peut être intéressant, dans certains cas, de regrouper entre elles les catégories d'une même variable.

Remarque 3. La part d'inertie due à une variable j la somme des inerties dues à chacune de ses q_j modalités. Cette part d'inertie est égale à

$$(q_j - 1)/r.$$

Plus une variable a de modalités, plus la part d'inertie liée à cette variable est importante. Dans la pratique, on évitera d'avoir des déséquilibres trop importants dans le nombre de modalités des variables (par exemple, des variables avec peu de modalités et d'autres avec un grand nombre de modalités).

3.4 Propriétés des valeurs propres

3.4.1 Premier critère pour le choix du nombre d'axes

De par la nature des données du TDC (variables binaires, colonnées démultipliées par le nombre de modalités, redondance de l'information, ...), il est difficile de concentrer l'inertie sur les premiers

facteurs de l'ACM. Cependant, on peut proposer un premier seuil pour éliminer certaines valeurs propres. En effet, en supposant que la matrice N^* est de plein rang on a $\tau := q - r$ valeurs propres non nulles. Dans ce cas

$$\sum_{k=1}^{\tau} \lambda_k = I_T = \frac{q}{r} - 1 \Leftrightarrow \frac{1}{q} \sum_{k=1}^{\tau} \lambda_k = \frac{1}{p}.$$

Autrement dit, une valeur propre vaut en moyenne $1/r$. On peut donc sélectionner dans un premier temps les valeurs propres supérieures à ce seuil, c'est-à-dire celles qui contribuent plus que la moyenne. En revanche, comme les pourcentage d'inertie expliqués par chaque axe sont petits par construction de l'ACM, on n'utilise pas en pratique le critère du pourcentage d'inertie pour choisir le nombre d'axes à retenir.

3.4.2 Correction des valeurs propres

On a vu dans la section précédente qu'il n'est pas possible en ACM d'interpréter l'inertie totale de façon aussi intéressante qu'en ACP ou AFC. Par contre, on peut relier la somme des carrés des valeurs propres à des indices statistiques. En effet, si on note λ_k, $k = 1, ..., \tau$ les valeurs propres de $V_r M_r$ avec τ est le rang de la matrice $V_r M_r$, alors λ_k^2 est valeurs propres de $(V_r M_r)^t V_r M_r$. On peut alors montrer que:

$$\sum_{k=1}^{\tau} \lambda_k^2 = \frac{1}{r^2} \sum_{j=1}^{r} (q_j - 1) + \frac{1}{r^2} \sum_{j=1}^{r} \sum_{j'=1, j \neq j'}^{r} \varphi_{jj'}^2,$$

où

$$\varphi_{jj'}^2 := \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2$$

est le coefficient de Pearson pour le croisement des variables X_j et $X_{j'}$.

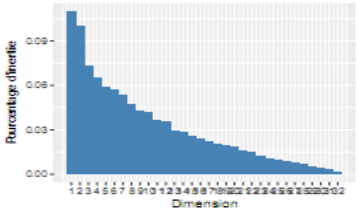
On peut définir de la même manière la distance du chi-deux entre deux profils-colonnes.

Avec cette distance, si on observe de grands écarts sur des modalités peu représentées, ceux-ci ont plus de poids dans le calcul de la distance. Et inversement, on donne moins de poids à des écarts importants qui pourraient être dus au fait que l'on a seulement observé plus de points sur cette modalité.

Or, on a vu au début de ce chapitre que la tableau de Burt pouvait s'obtenir à partir du TDC: $\mathbf{B} = N^{*t} N^*$. En faisant une AFC du tableau de Burt, on va alors récupérer les valeurs propres λ_k^2, $1 \leq k \leq \tau$. Cette pratique est courante dans le monde anglo-saxon, et présente l'avantage de fournir une meilleure interprétation des valeurs propres, mais rend l'interprétation en terme d'individus plus difficile car non immédiate.

Deux corrections ont été proposées pour améliorer les pourcentages d'inertie expliqués par chaque axe, en partant des résultats obtenus dans le cas d'une AFC du tableau de Burt. La figure 3.1

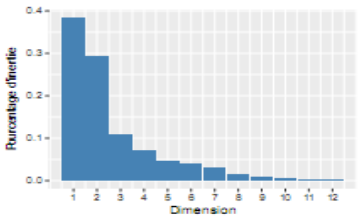
présente un exemple d'application de ces corrections.



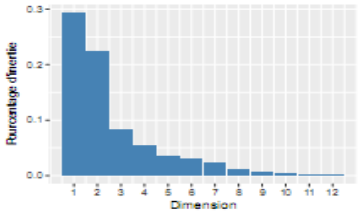
(a) *À partir des valeurs propres initiales*

Axe	Sans correction	Benzécri	Greenacre
1	10.98	38.37	29.32
2	9.99	29.30	22.39
3	7.27	10.66	8.15
4	6.48	6.99	5.34
5	5.85	4.60	3.52
6	5.66	4.00	3.06

(b) *Pourcentage d'inertie cumulée*



(c) *Après correction de Benzécri*



(d) *Après correction de Greenacre*

Figure 3.1 – Pourcentages d'inertie obtenus à partir des valeurs propres initiales obtenues par l'ACP des profils-lignes du TDC, ou après correction de Benzécri et de Greenacre.

3.4.3 Correction de Benzécri

En utilisant les propriétés du tableau de Burt et l'interprétation des valeurs propres de l'AFC de ce tableau, Benzecri (1979) a proposé la correction suivante :

1. sélectionner les ℓ valeurs propres supérieures au seuil $1/p$.
2. calculer les valeurs propres corrigées:
$$\widetilde{\lambda}_k := \left[\left(\frac{p}{p-1}\right)\left(\lambda_k - \frac{1}{p}\right)\right]^2.$$
3. calculer la somme des valeurs propres corrigées.
4. tracer l'éboulis des valeurs propres corrigées en traçant le pourcentage d'inertie cumulée corrigée.

3.4.4 Correction de Greenacre

La correction de Benzécri permet d'augmenter la part d'inertie expliquée par les premiers axes, mais elle a tendance à légèrement surestimer cette part. C'est pourquoi, Greenacre (1993) a proposé une correction supplémentaire pour le calcul de l'inertie expliquée par chaque axe. Quand la correction de Benzécri propose de calculer le pourcentage d'inertie d'un axe k par $\widetilde{\lambda}_k / \sum_{k=1}^{\ell} \widetilde{\lambda}_k$ Greenacre

propose d'utiliser :

$$\frac{\widetilde{\lambda}_k}{I_G} \text{ où } I_G := \left[\left(\frac{p}{p-1}\right)\left(\sum_{k=1}^{\ell}\widetilde{\lambda}_k^2 - \frac{q-p}{p^2}\right)\right]^2.$$

4 Interprétation des résultats

4.1 Contributions des individus et des modalités

Définition. *La contribution d'un individu i à l'axe factoriel l est donnée par :*

$$ctr_l(i) = f_{i\cdot}\frac{a_{il}^2}{\lambda_l} = \frac{a_{il}^2}{p\lambda_l},$$

où a_{il} est la coordonnée de l'individu i sur l'axe factoriel l. La contribution d'une modalité j à l'axe factoriel l est données par:

$$ctr_l(j) = f_{\cdot j}\frac{a_{jl}^2}{\lambda_l},$$

où a_{jl} est la coordonnées de la modalité j sur l'axe factoriel l.

La contribution d'une variable est égale à la somme des contributions de chacune de ses modalités. On remarque que la contribution d'une modalité à un axe dépend de sa fréquence: on peut alors retenir les modalités qui contribuent plus que leurs poids, i.e. celles pour les quelles $|a_{jl}| > \sqrt{\lambda_j}$.

On retient que ce sont les modalités et les individus les plus excentrés sur les axes qui contribuent le plus.

4.2 Rapport de corrélation

En ACP, on mesure la corrélation entre les variables initiales et les composantes principales, toutes les variables étant quantitatives. En ACM, les variables initiales étant qualitatives et les composantes principales quantitatives, on a besoin d'un autre critère statistique pour étudier la liaison entre les variables et les composantes principales. Le rapport de corrélation est un critère reposant sur la décomposition de l'inertie totale d'une variable quantitative y selon les modalités d'une variable qualitative X, en une somme de l'inertie interclasses et des inerties interclasses, (décomposition due au théorème de Huygens). L'inertie interclasses correspond à l'inertie des centres de gravité des classes définies par les modalités de N^*, et les inerties interclasses correspondent à l'inertie des individus d'une classe par rapport au centre de gravité de la classe à laquelle ils appartiennent. Le rapport de corrélation est défini comme le ratio entre l'inertie interclasses et l'inertie totale de la variable Y, et varie entre 0 et 1.

Lorsque ce ratio est proche de 1, cela signifie que l'inertie interclasses est élevée et que les inerties interclasses sont faibles: autrement dit, les individus d'une même classe sont regroupés autour du

centre de gravité de la classe, et les classes sont bien séparées les unes des autres. Dans ce cas, il y a une liaison forte entre la variable qualitative X et la variable quantitative Y. En revanche, lorsque ce ratio est proche de 0, cela signife que l'inertie interclasses et faible et qu'au moins une des inerties interclasses est élevée: les classes sont alors proches les unes des autres, et les individus dispersés au sein des classes. C'est une situation où les variables X et Y ne sont pas liées. On trouvera une illustration de ces deux situations sur la figure 3.2. Le rapport de corrélation entre la variable j et l'axe factoriel l est donné par :

$$R_l(j) := \frac{1}{\lambda_l} \sum_{k=1}^{q_j} \frac{n_{\cdot k}}{p}$$

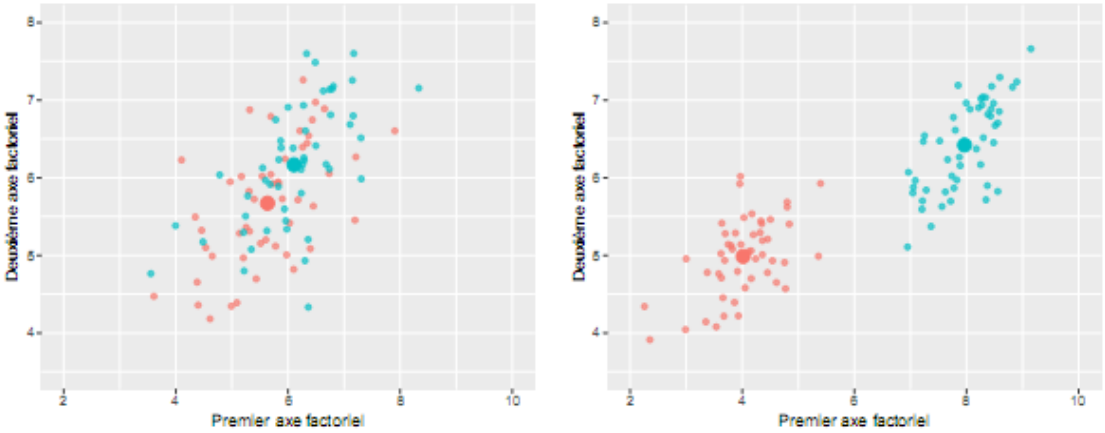


FIGURE 3.2 – Illustration de deux situations correspondant à un rapport de corrélation proche de 0 (à gauche) et proche de 1 (à droite).

4.3 Relations barycentriques

Il est possible de représenter les individus et les modalités sur le même graphe. On a alors l'interprétation suivante:

1. les individus sont au barycentre des modalités qu'ils possèdent.

2. les modalités sont au barycentre des individus qui les possèdent.