Analyse Multivariée des données

Dr Mory Ouattara
Data Science

INPHB

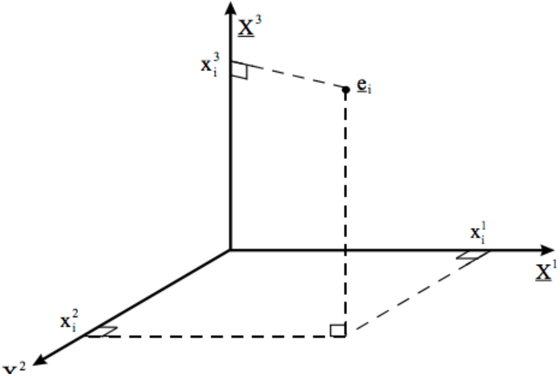
2018

1. Données Multivariées
2. Statistiques
3. Idée de l'ACP
4. Étude des corrélations
5. Lien entre η_j et ξ_i
6. Disque de corrélation
7. Individus et variables supplémentaires
8. Test Khi2
   - Le test du Khi2
9. L'analyse des correspondances simples
   - Notations et présentation
   - ACP du nuage des profils lignes-profils colonnes
   - Lien entre les deux analyses
   - Représentation de l'A.F.C.
   - Aides à l'interprétation : identiques à celles de l'A.C.P.
10. Cas pratique
11. Analyse des correspondances multiples
12. Présentation Formelle

Données Multivariées

A chaque individu noté e_i, on peut associer un point dans R^p = espace des individus.

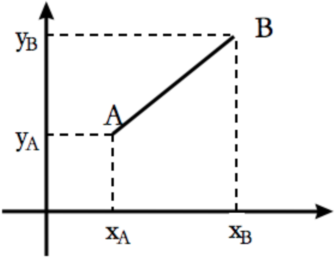
A chaque variable du tableau X est associé un axe de R^p.



Impossible à visualiser dès que p > 3.

Données Multivariées

. LE CHOIX DE LA DISTANCE ENTRE INDIVIDUS



Dans le plan:

$$\boxed{d^2\,(A,B) = (x_B - x_A)^2 + (y_B - y_A)^2}$$

Dans l'espace R^p à p dimensions, on généralise cette notion : la distance euclidienne entre deux individus s'écrit:

$$e_i = \left(x_i^1\ x_i^2\ \ldots\ x_i^p\right) \qquad e_j = \left(x_j^1\ x_j^2\ \ldots\ x_j^p\right)$$

$$d^2\,(e_i, e_j) = \left(x_i^1 - x_j^1\right)^2 + \left(x_i^2 - x_j^2\right)^2 + \ \ldots\ \left(x_i^p - x_j^p\right)^2$$

Données Multivariées

INERTIE TOTALE

$$\boxed{I_g = \sum_{i=1}^{n} \frac{1}{n} d^2(e_i, \underline{g})}$$

ou de façon plus générale

$$\boxed{I_g = \sum_{i=1}^{n} p_i \, d^2(e_i, \underline{g})} \qquad \text{avec} \quad \sum_{i=1}^{n} p_i = 1$$

L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité $\underline{g}$

L'inertie mesure la dispersion totale du nuage de points.

Données Multivariées

L'inertie est donc aussi égale à la somme des variances des variables étudiées.

En notant V la matrice de variances-covariances :

$$V = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ \vdots & s_2^2 & & \vdots \\ \vdots & & & \vdots \\ s_{p1} & & & s_p^2 \end{pmatrix}$$

$$\boxed{I_g = \sum_{i=1}^{p} s_i^2}$$

$$\boxed{I_g = Tr(V)}$$

Remarque

Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1.

L'inertie totale est alors égale à p (nombre de variables).

Données Multivariées

Soit $x \in \mathbb{R}^p$ un vecteur aléatoire :

$$x = (\xi_1, \xi_2, \ldots, \xi_p)^T$$

où v^T désigne la transposée du vecteur v
Un échantillon multidimensionnel est une suite $x_1, \ldots, x_n$ de réalisations aléatoires du vecteur x.

x_{ij} désignera la j ème composante du vecteur x_i

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

Statistiques

1. Les moyennes empiriques

$$\bar{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ki} \;\; k = 1, \ldots, p$$

qui forment le vecteur

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n}\mathbf{X}^T\mathbf{1} \quad \text{avec} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n.$$

2. Les covariances empiriques

$$s_{jk} = \frac{1}{n}\sum_{i}^{n} x_{ij}x_{ik} - \bar{x_j}\bar{x_k} \;\; k, j = 1, \ldots p$$

qui forment la matrice covariance empirique $S = (S_{kj})$

Statistiques

3 Les corrélations empiriques

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \; k, j = 1, \dots p$$

qui forment la matrice de corrélation empirique

$$R = (r_{jk})_{k,j=1,\dots p}$$

Statistiques

Il est facile de voir que

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - \bar{x}\bar{x}^T = \frac{1}{n}XX^T - \frac{1}{n^2}X\mathbf{1}\ \mathbf{1}^T X^T = \frac{1}{n}X^T H X$$

où

$$H = I_n - \frac{1}{n}\mathbf{1}\ \mathbf{1}^T$$

est la matrice de centrage

1. *Montrer que H est un projecteur, i. e. $H = H^2$ et $H^T = H$. Sur quel sous-espace vectoriel de $\mathbb{R}^n$ projette-t-il ?*
2. *Montrer la matrice de covariance empirique S est positive, en effet pour tout vecteur R^p.*

L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :

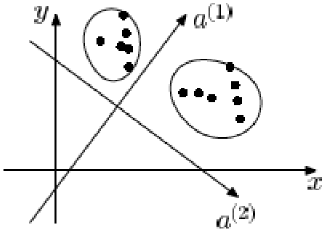
1. visualiser les données (**Notion de distances entre individus**)
2. réduire la dimension effective des données **(en fonction de leurs corrélations)**.

L'idée de l'Analyse en composantes principales (ACP)

si $a = (a_1, \ldots, a_p)^T \in R^p$ est une direction de projection.

- les données projetées $(a^T x_1, ..., a^T x_n)$ forment un échantillon de dimension 1.
- que l'on peut visualiser et qui est donc plus facile à interpréter que l'échantillon de départ $(x_1, \ldots, x_n)$.

L'idée de l'Analyse en composantes principales (ACP)

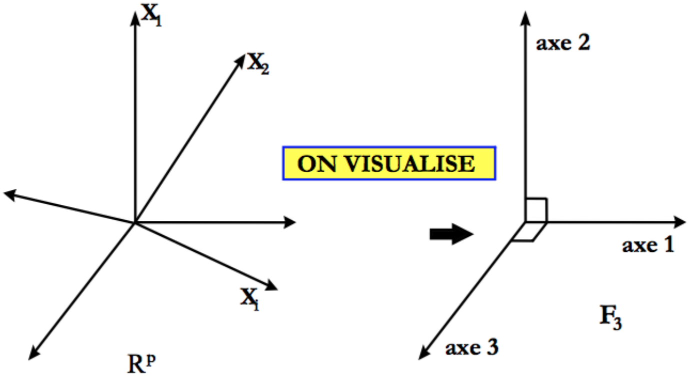


Bonne et mauvaise directions de projection.

L'ACP a pour objectif de trouver un sous-espace linéaire de $\mathbb{R}^p$ de dimension $p* << p$ tel que la projection sur ce sous-espace "capte" presque toute la structure des données.

L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :



L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :

① F_k devra être « ajusté » le mieux possible au nuage des individus: la somme des carrés des distances des individus à F_k doit être minimale.

⇕

② F_k est le sous-espace tel que le nuage projeté ait une **inertie** (dispersion) maximale.

L'idée de base de l'ACP est de chercher la direction $\boldsymbol{a} \in \mathbb{R}^p$ qui maximise en $\boldsymbol{a}$ la variance empirique de l'échantillon unidimensionnel $(a^T x_1, \ldots, a^T x_n)$

$$s_a^2 \overset{\text{déf}}{=} \frac{1}{n}\sum_{i=1}^{n}(a^T \mathbf{x}_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}(a^T \mathbf{x}_i)\right)^2$$

$$= \frac{1}{n}\, a^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\right) a - \frac{1}{n^2}\, a^T \left(\sum_{i=1}^{n} \mathbf{x}_i \sum_{i=1}^{n} \mathbf{x}_i^T\right) a = a^T S a,$$

la direction la plus intéressante $\hat{a}$ est une solution de

$$\max_{a \in \mathbb{R}^p : \|a\| = 1} a^T S a = \hat{a}^T S \hat{a}$$

ou

$$\hat{a} = \arg\max_{a \in \mathbb{R}^p : \|a\| = 1} a^T S a$$

ACP

Nous nous intéressons à la solution du problème suivant :

$$a^* = \underset{a \in \mathbb{R}^p : \|a\|=1}{arg \ \ max} \, Var(a^T X) \ avec \ E(\| X \|^2) < \infty$$

Soit

$$\Sigma = \Gamma \Lambda \Gamma^T$$

une décomposition spectrale de covariance

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \qquad \Gamma = (\gamma_{(1)}, \dots, \gamma_{(p)})$$

Où les λ_i sont les valeurs propres de Σ rangées dans l'ordre croissant et

$$\| \gamma_{(i)} \| = 1 \ avec \ \gamma_{(i)}^T \gamma_{(k)} = 0 \ pour \ i \neq k$$

ACP : Composante principale $\eta_{(j)}$

Composante principale $\eta_{(j)}$

La variable aléatoire $\eta_{(j)} = \gamma_{(i)}^T(x - \mu)$ est dite j ème composante principale du vecteur aléatoire $x \in R^p$.

Les $\gamma_{(j)}$ sont les vecteurs propres de la matrice de covariance Σ du vecteur aléatoire x, on obtient :

$$\mathrm{Var}[\eta_j] = E[\gamma_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\gamma_{(j)}] = \gamma_{(j)}^T\Sigma\gamma_{(j)} = \gamma_{(j)}^T\lambda_j\gamma_{(j)} = \lambda_j,$$

$$\mathrm{Cov}(\eta_j, \eta_k) = E[\gamma_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\gamma_{(k)}] = \gamma_{(j)}^T\Sigma\gamma_{(k)} = \gamma_{(j)}^T\lambda_k\gamma_{(k)} = 0,$$

ACP

Soit $\mathbf{x}$ un vecteur aléatoire de $\mathbb{R}^2$ de moyenne nulle et de matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad 0 \leq \rho \leq 1.$$

ACP

Exemple 7.1. Soit $\mathbf{x}$ un vecteur aléatoire de $\mathbb{R}^2$ de moyenne nulle et de matrice de covarian

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad 0 \leq \rho \leq 1.$$

Considérons les vecteurs propres orthonormés de cette matrice

$$\gamma_{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \gamma_{(2)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Donc si les coordonnées de $\mathbf{x}$ sont ξ_1 et ξ_2, les composantes principales de x valent

$$\eta_1 = \frac{\xi_1 + \xi_2}{\sqrt{2}}, \qquad \eta_2 = \frac{\xi_1 - \xi_2}{\sqrt{2}}.$$

ACP

Théorème

Soit $x \in R^p$ un vecteur aléatoire tel que $E(\| x \|) < \infty$. Alors $\hat{a} = \gamma_{(1)}$ est vérifie :

$$Var(\hat{a}^T X) = \max_{a \in \mathbb{R}^p : \|a\| = 1} (a^T X) = \max_{a \in \mathbb{R}^p : \|a\| = 1} (a^T (X - \mu))$$

ACP

Preuve. La décomposition spectrale de la matrice Σ est de la forme

$$\Sigma = \Gamma\Lambda\Gamma^T = \sum_{j=1}^{p} \lambda_j \gamma_{(j)}\gamma_{(j)}^T.$$

On a donc

$$\mathrm{Var}[a^T \mathrm{x}] = \sum_{j=1}^{p} \lambda_j (a^T \gamma_{(j)})(\gamma_{(j)}^T a) = \sum_{j=1}^{p} \lambda_j c_j^2,$$

où $c_j = a^T\gamma_{(j)}$ est la projection du vecteur a sur la direction $\gamma_{(j)}$. Puisque les vecteurs $\gamma_{(j)}$ forment une base orthonormée de $\mathbb{R}^p$, on a $c_1^2 + \cdots + c_p^2 = \|a\|^2$. Comme $\lambda_j \leq \lambda_1$, on en déduit que

$$\mathrm{Var}[a^T \mathrm{x}] = \sum_{j=1}^{p} \lambda_j c_j^2 \leq \lambda_1 \sum_{j=1}^{p} c_j^2 = \lambda_1 \|a\|^2 = \lambda_1.$$

Par ailleurs, si $a = \hat{a} = \gamma_{(1)}$, les coefficients c_j sont tous nuls sauf le premier $c_1 = 1$. On a donc $\mathrm{Var}[\hat{a}^T \mathrm{x}] = \lambda_1$. Par conséquent, $\hat{a}$ est une solution du problème de maximisation (7.2) et $\mathrm{Var}[\hat{a}^T \mathrm{x}] = \lambda_1 = \mathrm{Var}[\eta_1]$. ∎

Étude des corrélations

On définit la variance totale de x par

$$E(\| X - \mu \|^2) = E(X - \mu)^T (X - \mu) = E(X - \mu)^T \Gamma \Gamma^T (X - \mu)$$

avec

$$\Gamma^T(\mathbf{x} - \mu) = \begin{pmatrix} \gamma_{(1)}^T(\mathbf{x} - \mu) \\ \vdots \\ \gamma_{(p)}^T(\mathbf{x} - \mu) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_p \end{pmatrix} \overset{\text{déf}}{=} \mathbf{y}.$$

Compte tenu de ces notations et de l'égalité $E(\eta_i^2) = \lambda_i$, on obtient

$$E(\| X - \mu \|^2) = E(\eta_1^2 + \ldots + \eta_p^2)$$

ACP

Donc : Comment détermine t-on le meilleur sous espace de projection ?

Étude des corrélations

Part de variance expliqué

On appelle part de la variance totale de x expliquée par les k premières composantes principales $(\eta_1, \ldots, \eta_k)$ la quantité

$$\frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

Étude des corrélations

Combien d'axes ?

Différentes procédures sont complémentaires:

① **Pourcentage d'inertie souhaité : a priori**

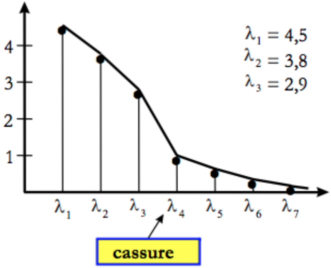
② Diviser l'inertie totale par le nombre de variables initiales

⇨ inertie moyenne par variable : I.M.

Conserver tous les axes apportant une inertie supérieure à cette valeur (inertie > 1 si variables centrées réduites).

③ **Histogramme**

Conserver les axes associés aux valeurs propres situées avant la cassure.



Lien entre η_j et ξ_i

Calculons d'abord la matrice de covariance des vecteurs aléatoires x et y.

$$C(\mathbf{x},\mathbf{y}) = E[(\mathbf{x}-\mu)\mathbf{y}^T] = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Gamma] = \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda.$$

Comme $\mathrm{Cov}(\xi_i,\eta_j)$ est le $(i,j)^{\text{ème}}$ élément de cette matrice, on obtient

$$\mathrm{Cov}(\xi_i,\eta_j) = \gamma_{ij}\lambda_j.$$

La corrélation $\tilde{\rho}_{ij} = \mathrm{Corr}(\xi_i,\eta_j)$ entre ξ_i et η_j vaut

$$\tilde{\rho}_{ij} = \frac{\mathrm{Cov}(\xi_i,\eta_j)}{\sqrt{\mathrm{Var}(\xi_i)\mathrm{Var}(\eta_j)}} = \gamma_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}}.$$

Lien entre η_j et ξ_i

Proposition

Soit $x \in R^p$ un vecteur aléatoire, tel que $E(\| x \|^2) < \infty$ et $\sigma_{ii} > 0$ pour tout $i = 1, \ldots, p$.
Alors,

$$\sum_{j=1}^{p} \tilde{\rho}_{ij}^2 = 1$$

On appelle $\tilde{\rho}_{ij}^2$ part de variance de la variable ξ_i expliquée par la j ème composante principale η .
Pour tout sous-ensemble J de 1, ..., p,

$$\sum_{j \in J} \lambda_j = \sum_{j=1}^{p} \sigma_{ii} \tilde{\rho}_{iJ}^2 \; avec \; \tilde{\rho}_{iJ}^2 = \sum_{j \in J} \rho_{ij}^2$$

Lien entre η_j et ξ_i

Preuve.

$$\sum_{i=1}^{p} \sigma_{ii}\tilde{\rho}_{iJ}^2 = \sum_{i=1}^{p} \sigma_{ii} \sum_{j\in J} \gamma_{ij}^2 \frac{\lambda_j}{\sigma_{ii}} = \sum_{j\in J} \lambda_j \sum_{i=1}^{p} \gamma_{ij}^2.$$

Le résultat de la proposition découle du fait que la dernière somme vaut 1, car $\|\gamma_{(j)}\|^2 = \sum_{i=1}^{p} \gamma_{ij}^2 = 1$. ∎

Disque des corrélations

Proposition

Soient ξ_i et ξ_k deux variables entièrement expliquées par les deux premières composantes principales, i.e.

$$\tilde{\rho_{i1}}^2 + \tilde{\rho_{i2}}^2 = 1 \ et \ \tilde{\rho_{k1}}^2 + \tilde{\rho_{k2}}^2 = 1$$

Alors, la corrélation de ξ_i et ξ_k est donnée par la formule

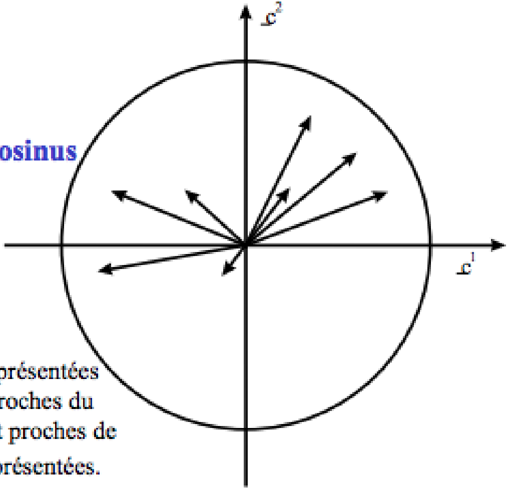
$$\rho_{ik} = \tilde{\rho_{i1}}\tilde{\rho_{k1}} + \tilde{\rho_{i2}}\tilde{\rho_{k2}} = cos(\varphi),$$

où φ est l'angle formé par les vecteurs $(\tilde{\rho_{i1}}, \tilde{\rho_{i2}})$ et $(\tilde{\rho_{k1}}, \tilde{\rho_{k1}})$.

Disque des corrélations

Le cercle des corrélations est la projection du nuage des variables sur le plan des composantes principales.

corrélation = cosinus



Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

Variables

- On calcule le coefficient de corrélation entre la variable supplémentaire et les composantes principales.
- Ceci permet sa représentation sur le cercle des corrélations.

Individus

- Individu de poids nul ne participant pas à l'analyse
- Appliquer aux coordonnées de l'individu les expressions définissant les composantes principales.

L'Analyse des Correspondances Simples
A F C

Structure de base des données

Objective : mesurer des liaisons entre deux variables qualitatives :Khi-deux
exemple : Il s'agit de tester l'indépendance de deux variables qualitatives. Y a-t-il indépendance entre :

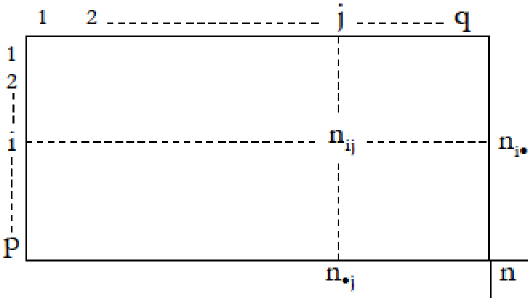
- la catégorie socioprofessionnelle et le vote à l'élection présidentielle ?
- le niveau d'études et les journaux lus ?
- Que pensez vous de cette affirmation : *On en a assez de ceux qui bloquent la vie du pays par leurs revendications par les armes.*
    1. pas du tout d'accord
    2. pas tellement d'accord
    3. bien d'accord
    4. entièrement d'accord

Existe-t- il un lien entre les réponses et la tendance politique ?

Tendance Politique	1	2	3	4	5	Total
PIT	714	71	0	143	71	1000
FPI	284	216	199	174	127	1000
MFA	87	106	228	335	244	1000
RDR	16	86	156	271	471	1000
PDCI	71	71	0	214	643	1000
Indifférent	82	120	244	301	263	1000
Non Reponse	88	124	269	285	233	1000

Le tableau de contingence

Croisement de deux variables qualitatives I et J à p et q modalités.



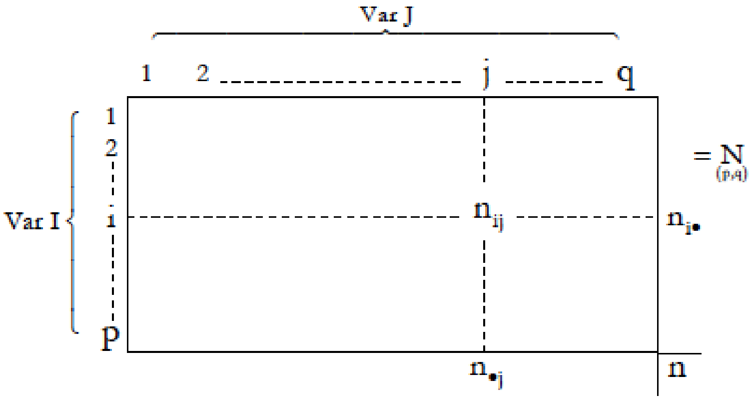
$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \quad \text{(total ligne)}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij} \quad \text{(total colonne)}$$

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} \quad \text{(total)}$$

Notations : tableau de contingence N

Croisement de deux variables qualitatives à p et q modalités



Profils lignes - profils-colonnes - profils marginaux

- p **Profils des lignes** $\dfrac{n_{ij}}{n_{i\bullet}} = \dfrac{f_{ij}}{f_{i\bullet}}$

  profil de la ligne i noté ℓ_i

$$\left(\frac{n_{i1}}{n_{i\bullet}} \quad \frac{n_{i2}}{n_{i\bullet}} \quad \cdots\cdots \quad \frac{n_{iq}}{n_{i\bullet}}\right) \Leftrightarrow \left(\frac{f_{i1}}{f_{i\bullet}} \quad \frac{f_{i2}}{f_{i\bullet}} \quad \cdots\cdots \quad \frac{f_{iq}}{f_{i\bullet}}\right)$$

- q **Profils des colonnes**

  profil de la colonne j

  noté c_j

$$\begin{pmatrix} \dfrac{n_{1j}}{n_{\bullet j}} \\ \dfrac{n_{2j}}{n_{\bullet j}} \\ \dfrac{n_{pj}}{n_{\bullet j}} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \dfrac{f_{1j}}{f_{\bullet j}} \\ \dfrac{f_{2j}}{f_{\bullet j}} \\ \dfrac{f_{pj}}{f_{\bullet j}} \end{pmatrix}$$

Profils lignes - profils-colonnes - profils marginaux

Si les deux variables qualitatives I et J étaient indépendantes, les profils lignes seraient tous identiques, et donc identiques au profil marginal correspondant.

$$Independance \Rightarrow \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \Rightarrow n_{ij} = \frac{n_{\bullet j} * n_{i\bullet}}{n}$$

Profils lignes - profils-colonnes - profils marginaux

- On pouvait établir la relation précédente en raisonnant sur les profils colonnes.

$$\boxed{f_{ij} = f_{i\bullet}\, f_{\bullet j}}$$

$$\text{avec } f_{ij} = \frac{n_{ij}}{n} \quad \text{et} \quad f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Elle exprime clairement que dans le cas de l'indépendance le tableau de contingence est entièrement déterminé par ses marges

Définition du Khi-deux

- Pour chaque case, on peut donc calculer le nombre de cas attendus (sous hypothèse d'indépendance) $n_{ij} = \frac{n_{\bullet j} * n_{i\bullet}}{n}$
- On peut comparer les nombres de cas attendus E_{ij} aux nombres observés.

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i\bullet}\, n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\, n_{\bullet j}}{n}}$$

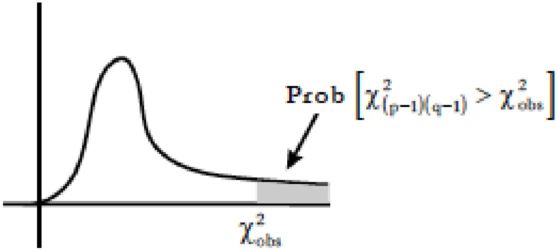
$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Si les deux variables sont réellement indépendantes, cette expression suit une distribution du Khi-deux avec un nombre de degrés de liberté égal à : $(p-1)(q-1)$
Dans une table on lit $\chi^2_{\alpha,k}$ valeur ayant une probabilité α d'être dépassée
pour une distribution du khi-deux avec $k=(p-1)(q-1)$ degrés de liberté.

1. Si $\chi^2 \leq \chi^2_{\alpha,k}$ On accepte H_o : independance
2. Si $\chi^2 > \chi^2_{\alpha,k}$ On rejette H_o : independance

Pratique sous logiciel statistique

- Calcul du χ^2 associé au tableau de contingence noté χ^2_{obs}.
- Probabilité pour une v.a. suivant une loi du khi-deux à $(p-1)(q-1)$ d.d.l. de dépasser χ^2_{obs}.



Si cette probabilité est faible (en général < 5 %), on rejette l'hypothèse d'indépendance entre les deux variables qualitatives.

Représentation des profils lignes

- Les profils lignes sont considérés comme des individus.
- Les p profils-lignes forment un nuage de p points dans R^q
- A chaque profil-ligne est associé un poids égal à sa fréquence marginale profil ligne poids $f_{i\bullet}$.

On note $N(I)$ le nuage de points formé des profils-lignes pondérés : $(l_i; f_{i\bullet})$
Le centre de gravité g est défini par :

$$g_I = \sum_{i=1}^{p} f_{i\bullet} l_i$$

La jième coordonnée de g_I vaut $f_{\bullet j}$
Donc g_I = profil marginal de la variable J (à q modalités) $g_I = f_J$

Représentation des profils colonne

$N(J) =$ nuage de points formé des q profils - colonnes pondérés $\left(c_j, f_{\bullet j}\right)$

$$\boxed{\text{Le centre de gravité } g_c \text{ est le profil marginal de la variable I à p modalités.}}$$

$$g_c = f_I$$

Le problème qui se pose est **l'étude de la dépendance** entre les deux variables qualitatives.

Dans le cas où les deux variables sont **indépendantes**, on a identité des profils :

(1) $\dfrac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$ profil - ligne

(2) $\dfrac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$ profil - colonne

$$\boxed{f_{ij} = f_{i\bullet}\, f_{\bullet j}}$$

Dans le cas de l'indépendance, le nuage des profils-lignes se réduit à un point g_ℓ

De même, le nuage des profils-colonnes se réduit à un point g_c.

$\Longrightarrow$ L'étude de la dépendance consiste à étudier la **forme des nuages**.

$\Longrightarrow$ Problème d'analyse en composantes principales.

Quelle métrique ?

Métrique du $\chi 2$

- Pour les profils lignes :

$$d^2_{\chi^2}(l_i, l'_i) = \sum_{j=1}^{q} \frac{n}{n_{\bullet j}} \left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{i'j}}{n_{i'\bullet}}\right)^2$$

  - Donne un poids important aux différences portant sur les petits pourcentages.
  - Vérifie le principe d'équivalence distributionnelle : si deux colonnes ont le même profil, on les réunit en une seule d'effectif somme sans modifier les distances entre profils lignes.

- Pour les profils-colonnes :

$$d^2_{\chi^2}(c_j, c'_j) = \sum_{j=1}^{q} \frac{n}{n_{i\bullet}} \left(\frac{n_{ij}}{n_{\bullet j}} - \frac{n_{ij'}}{n_{\bullet j'}}\right)^2$$

Inertie du nuage N(I)

$I_{N(I)}$ l'inertie du nuage$N(I)$ calculée par rapport au centre de gravité f_J vaut

$$\frac{\chi^2}{n}$$

où $\chi^2 =$ Khi-deux associé au tableau de contingence étudié.
On obtient le même résultat pour l'inertie du nuage N(J).

l'A.C.P. du nuage des profils-lignes :

- Les profils-lignes jouent le rôle d'individus ; ils sont affectés des poids $f_{i\bullet}$
- La métrique utilisée pour le calcul des distances entre individus est la métrique du khi-deux
- Le premier axe principal du nuage des profils-lignes est la droite passant le plus près possible de l'ensemble des points de N(I).

l'A.C.P. du nuage des profils-lignes

Notons $\underline{a}^1$ la première composante principale

$$\underline{a}^1 = \begin{pmatrix} \vdots \\ \vdots \\ \vdots \end{pmatrix} \leftarrow \text{coordonnées des p profils-lignes sur l'axe 1}$$

Notons λ_1 la variance de $\underline{a}^1$ (égale à l'inertie portée par l'axe qui lui est associé).

$\underline{a}^2$ = deuxième composante principale de variance λ_2

$\underline{a}^3$ = troisième composante principale de variance λ_3

'A.C.P. du nuage des profils-colonnes

Notons $\underline{b}^1$ la première composante principale

$\underline{b}^1 = \begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \leftarrow$ coordonnées des q profils-colonnes sur l'axe 1

$\underline{b}^2 =$ deuxième composante principale

Les composantes principales de l'A.C.P. des profils-colonnes sont associées aux mêmes valeurs propres que les composantes principales de l'A.C.P. des profils-lignes.

$\underline{b}^1$ a pour variance λ_1

$\underline{b}^2$ a pour variance λ_2

Formules de transition

En notant b_j et a_i les j^{me} et i^{me} coordonnées des composantes principales b et a associées à la même valeur propre λ :

$$\sqrt{\lambda} b_i = \sum_{i=1}^{p} \frac{n_{ij}}{n_{\bullet j}}$$

$$\sqrt{\lambda} a_i = \sum_{j=1}^{q} \frac{n_{ij}}{n_{i\bullet}}$$

À λ près, la coordonnée d'une modalité i d'une variable est la moyenne des coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de i.

Les modalités de la variable I sont représentées en tant qu'individus (profils-lignes) de l'A.C.P. des profils-lignes.

La modalité i de la variable I a pour coordonnées dans un espace de dimension k :

$\left(a_i^1 \,,\, a_i^2 \,,\, \ldots \,,\, a_i^k\right)$

avec a_i^1 = $i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^1$

a_i^2 = $i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^2$

.....

Les modalités de la variable I sont représentées en tant qu'individus (profils-lignes) de l'A.C.P. des profils-lignes.

Pour les modalités de la variable J, la modalité j a pour coordonnées :

$$\left(\sqrt{\lambda_1}\, b_j^1 \,,\, \sqrt{\lambda_2}\, b_j^2 \,,\, \ldots\ldots\, \sqrt{\lambda_k}\, b_j^k\right)$$

$$b_j^1 = j^{\text{ème}} \text{ coordonnée du vecteur } \underline{b}^1$$

$$b_j^2 = j^{\text{ème}} \text{ coordonnée du vecteur } \underline{b}^2$$

Les modalités du deuxième groupe (J) sont les barycentres des modalités du premier groupe (variable I).

(voir formules de transition)

Abandon du principe barycentrique

Les modalités de chaque ensemble sont représentées par les :

$$a_i^k, i = 1, \dots, p$$

$$b_j^k, j = 1, \dots, q$$

Cette représentation permet de déterminer les proximités entre certains éléments de I et certains éléments de J (compte tenu de la qualité de la représentation).

Contributions

de la ligne i à l'axe k

$$\frac{f_{i\bullet}\left(a_i^k\right)^2}{\lambda_k} \qquad \text{avec } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

de la colonne j à l'axe k

$$\frac{f_{\bullet j}\left(a_j^k\right)^2}{\lambda_k} \qquad \text{avec } f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Cosinus carrés

Modalité i représentée sur l'axe k

$$\frac{\left(a_i^k\right)^2}{d^2(i,G)}$$

Modalité j représentée sur l'axe k

$$\frac{\left(b_j^k\right)^2}{d^2(j,G)}$$

Aspects pratiques de l'interprétation

- L'interprétation peut se faire à partir des représentations graphiques (en s'assurant de la qualité de représentation de chaque modalité à l'aide des cos2).
- Quand le nombre de modalités est élevé, il est conseillé d'éditer d'abord le graphique des profils-lignes, puis celui des profilscolonnes, enfin la représentation simultanée.
- Les profils ayant des poids différents la lecture de leurs contributions à l'inertie de chaque axe s'avère très utile.
- On peut repérer les profils dont la contribution est supérieure au poids

But

- Étendre l'AFC au cas de $p \geq 2$ variables $\xi_1, \xi_2, \ldots, \xi_p$ à $m_1, \ldots, m_p$ modalités

$$\begin{array}{cccc} \xi_1 & \ldots & \xi_p & \text{variables} \\ m_1 & \ldots & m_p & \text{modalités} \end{array}$$

  Utile pour l'exploration d'enquêtes où les questions sont à réponses multiples.
- L'analyse des correspondances utilise une table de contingence qui est difficilement généralisable au cas $p \geq 2$

Trouver un moyen différent d'analyser $p > 2$ variables et vérifier que les résultats sont comparables à l'AFC pour p = 2.

Données

- **Données brutes** : chaque individu est décrit par les numéros des modalités qu'il possède pour chacune des p variables ξ_j. Impossible de faire des calculs sur ce tableau : valeurs arbitraires.
- **Tableau disjonctif** : Remplacer la j-ième colonne par m_j colonnes d'indicatrices : mettre un zéro dans chaque colonne, sauf celle correspondant à la modalité de l'individu i qui reçoit 1.

Technique de description de données qualitatives

n individus décrits par p variables qualitatives

$$\begin{array}{cccc} \mathcal{X}_1 & \dots & \mathcal{X}_p & \text{variables} \\ m_1 & \dots & m_p & \text{modalités} \end{array}$$

- L'A.C.M. décrit les relations deux à deux entre p variables qualitatives à travers une représentation des groupes d'individus correspondant aux diverses modalités.
- Cette méthode est particulièrement bien adaptée à l'exploration d'enquêtes.

Données

- Chaque individu est décrit par les numéros des catégories ou il est classé pour les p variables. Les données brutes se présentent sous forme d'un tableau à n lignes et p colonnes.
- Les éléments de ce tableau sont des codes arbitraires sur lesquels aucune opération arithmétique n'est licite.
- La forme mathématique utile pour les calculs est alors le tableau disjonctif des indicatrices des p variables obtenu en juxtaposant les p tableaux d'indicatrices de chaque variable $\mathcal{X}_i$

Données

Exemple :
On interroge 6 personnes sur la couleur de leurs cheveux (CB, CC et CR pour blond, châtain et roux), la couleur de leurs yeux (YB, YV et YM pour bleu, vert et marron) et leur sexe (H/F). Les tableaux brut (ci-dessous à gauche) sont équivalents aux tableaux disjonctifs (à droite).

$$\begin{pmatrix} CB \\ CB \\ CC \\ CC \\ CR \\ CB \end{pmatrix} \begin{pmatrix} YB \\ YV \\ YB \\ YM \\ YV \\ YB \end{pmatrix} \begin{pmatrix} H \\ H \\ F \\ H \\ F \\ F \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Tableau disjonctif et tableau de contingence

A chaque variable ξ_j est associée un tableau disjonctif $X_j(n \times m_j)$.
Pour 2 variables ξ_j et ξ_l le tableau de contingence est donné par :

$\mathcal{X} = X_j \vert X_l$	$N_{j'l} = X_j' X_l$	$X_j' X_j$	$X_l' X_l$
Disjonctifs	Contingence	Marge ξ_j	Marge ξ_l

$$\mathbf{N}_{12} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{D}_1 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Tableau disjonctif joint

$$\mathcal{X}(n \times m) = X_1 | X_2 \ldots | X_p$$
$$m = m_1 + \ldots + m_p$$

Exemple Pour les variables précédentes, on a le tableau disjonctif joint suivant

$$\mathbf{X} = \left(\begin{array}{ccc|ccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

Chaque somme de lignes vaut 3. Les sommes de colonnes valent

$$(3 \quad 2 \quad 1 | 3 \quad 2 \quad 1 | 3 \quad 3)$$

Le tableau de Burt

C'est un super-tableau de contingence des variables $X_1, ..., X_p$, formé de tableaux de contingence et de matrices d'effectifs marginaux. :

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \cdots & \mathbf{X}_1'\mathbf{X}_p \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{X}_p'\mathbf{X}_1 & \cdots & & \mathbf{X}_p'\mathbf{X}_p \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{D}_1 & \mathbf{N}_{12} & \cdots & \mathbf{N}_{1p} \\ \mathbf{N}_{21} & \mathbf{D}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{N}_{p1} & \cdots & & \mathbf{D}_p \end{bmatrix}$$

Exemple Toujours pour les mêmes variables

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

L'ACM : une AFC sur tableau disjonctif

Chercher une représentation des $m_1 + \ldots + m_p$ catégories comme points d'un espace de faible dimension.
Méthode Faire une AFC sur le tableau disjonctif joint

$$\mathcal{X}(n \times m) = X_1 | X_2 | \ldots | X_p$$

L'ACM : une AFC sur tableau disjonctif

- **Les lignes** : La somme des éléments de chaque ligne de $\mathcal{X}$ est égale à p. Le **tableau des profils-lignes** est donc $\frac{1}{p}\mathcal{X}$
- **Les colonnes** : la somme des éléments de chaque colonne de $\mathcal{X}$ est égale à l'effectif marginal de la catégorie correspondante.

Le tableau des profils colonnes est donc $\mathcal{X}D^{-1}$ où D est la matrice diagonale par blocs.

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}_p \end{pmatrix}$$

Les coordonnées factorielles des catégories

On note $a_k = (a_{1k}, \ldots, a_{pk})$ vecteur à $m_1 + \ldots + m_p$ composantes des coordonnées factorielles des catégories sur l'axe k.

Calcul de l'AFC sur $\mathcal{X}$

La matrice des profils lignes est

$$\frac{1}{p}\mathcal{X}$$

et celle des profils colonnes

$$\mathcal{X}D^{-1}$$

.

Les coordonnées factorielles des catégories

On note $a_k = (a_{1k}, \ldots, a_{pk})$ vecteur à $m_1 + \ldots + m_p$ composantes des coordonnées factorielles des catégories sur l'axe k.

a_k **est vecteur propre** de

$$(\mathcal{X} D^{-1})' \frac{1}{p} \mathcal{X} = \frac{1}{p} D^{-1} \mathcal{X}' \mathcal{X} = \frac{1}{p} D^{-1} B$$

l'équation des coordonnées des catégories est donc

$$\frac{1}{p} D^{-1} B a_k = \mu_k a_k$$

Avec la convention de normalisation suivantes

$$\frac{1}{np} {a_k}' D a_k = \mu_k$$

Resolution cas p=2

On note $a = (a_1, a_2)$ vecteur à $m_1 + m_2$ composantes factorielles des catégorie et μ_k la valeur propre correspondante.

Calcul de l'AFC sur $\mathcal{X}$

$$\frac{1}{2}\mathbf{D}^{-1}\mathbf{B}\begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{D}_1^{-1}\mathbf{N} \\ \mathbf{D}_2^{-1}\mathbf{N}' & \mathbf{I}_{m_2} \end{bmatrix}\begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} = \mu_k \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix}$$

Resolution cas p=2

On note $a = (a_1, a_2)$ vecteur à $m_1 + m_2$ composantes factorielles et μ_k la valeur propre correspondante.

On obtient les équations

$$\left\{ \begin{array}{l} \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}_k = (2\mu_k - 1)\mathbf{a}_k \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = (2\mu_k - 1)\mathbf{b}_k \end{array} \right.$$

et donc on retrouve les coordonnées des modalités de lignes et de colonnes dans l'AFC classique (avec $\mu_k = (2\lambda_k - 1)^2$)

$$\left\{ \begin{array}{l} \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}_k = (2\mu_k - 1)^2\mathbf{b}_k \\ \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = (2\mu_k - 1)^2\mathbf{a}_k \end{array} \right.$$

Différences ACM/AFC pour p = 2

- **Nombre de valeurs propres** : on a a priori $m_1 + m_2 - 2$ valeurs propres non nulles, En particulier pour chaque λ_k, on a deux μ_k possibles

$$\begin{cases} \mu_k = \frac{1+\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} \\ \mu'_k = \frac{1-\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix} \mathbf{a}_k \\ -\mathbf{b}_k \end{bmatrix} \end{cases}$$

  On ne garde donc que les valeurs $\mu_k > 0.5$

- **Inertie** : l'interprétation de la part d'inertie expliquée par les valeurs propres est maintenant très différente. En particulier les valeurs propres qui étaient très séparées dans l'AFC de N le sont beaucoup moins dans celle de X.

Formules barycentriques

- Les coordonnées des individus

$$\mathbf{c}_k = \frac{1}{\sqrt{\mu_k}} \frac{1}{p} \mathbf{X} \mathbf{a}_k \quad \text{et donc} \quad c_{ik} = \frac{1}{\sqrt{\mu_k}} \frac{1}{p} \sum_{j \text{ catégorie de } i} a_{jk}$$

Avec variance

$$\operatorname{var} \mathbf{c}_k = \frac{1}{n} \mathbf{c}_k' \mathbf{c}_k = \mu_k$$

- Les coordonnées des catégories

$$\mathbf{a}_k = \frac{1}{\sqrt{\mu_k}} \mathbf{D}^{-1} \mathbf{X}' \mathbf{c}_k \quad \text{c-à-d} \quad a_{jk} = \frac{1}{\sqrt{\mu_k}} \frac{1}{n_j} \sum_{i \text{ de catégorie } j} c_{ik}$$

Barycentres et représentation

- Les points représentatifs des catégories sont barycentres des groupes d'individus.
- Moyennes comme c_k est une variable de moyenne nulle, la formule de barycentre indique que pour chaque variable X_i les coordonnées de ses catégories sont de moyenne nulle.
- Pour que les catégories se trouvent visuellement au barycentre des individus qui les représentent on peut remplacer a_k

$$\boldsymbol{\alpha}_k = \mathbf{D}^{-1}\mathbf{X}'\mathbf{c}_k = \sqrt{\mu_k}\mathbf{a}_k$$

Sélection des axes

- règle courante : garder les axes tels que $\mu_k > \frac{1}{p}$ (la moyenne des valeurs propres est $\frac{1}{p}$).
- les axes intéressants sont ceux que l'on peut interpréter, en regardant les contributions des variables actives et les valeurs-tests associées aux variables supplémentaires.
- En pratique on se contente souvent d'interpréter le premier plan principal.

Sélection des axes

- Si n_j est l'effectif de la catégorie j et a_{jk} sa coordonnée sur l'axe factoriel k, alors

$$\text{var } \mathbf{a}_k = \sum_{j \in \text{catégories}} \frac{n_j}{np}(a_{jk})^2 = \mu_k$$

- **Catégorie** La contribution de la catégorie j à l'axe factoriel est :

$$\frac{n_j}{np}\frac{(a_{jk})^2}{\mu_k},$$

- **Variable** : la contribution totale de la variable ξ_v à l'axe factoriel est

$$\frac{1}{\mu_k}\frac{1}{np}\sum_{j \text{ modalité de } \mathcal{X}_v} n_j(a_{jk})^2$$

Contribution d'un individu

- Elle est égale pour l'individu i à
$$\frac{1}{n}\frac{(c_{ik})^2}{\mu_k}$$
- Qualité de la représentation pour le sous-espace formé par les premier axes, la qualité de la représentation de l'individu i est le cosinus carré habituel
$$\frac{\sum_{k=1}^{\ell}(c_{ik})^2}{\sum_{k=1}^{q}(c_{ik})^2}$$

Les variables supplémentaires

-