Analyse des données

Khaled Larbi

Septembre / Octobre 2018

GIS 2A - Polytech'Lille

Plan du cours

1. Rappel sur la statistique exploratoire bivariée

2. Analyse par composantes principales

3. Analyse factorielle des correspondances

4. Analyse des correspondances multiples

5. Introduction à d'autres méthodes d'analyse multivariée dérivées de l'ACP

Déroulement du cours

Le cours d'analyse des données est constitué de :

- 12 heures de cours (dont des rappels),
- 6 heures de travaux dirigés,
- 10 heures de travaux pratiques.

Les travaux pratiques seront effectués à l'aide du langage R.

L'évaluation aura lieu en deux temps :

- un examen sur table d'une durée de 2 heures,
- un mini-projet sous la forme d'un rapport à rendre.

Rappel sur la statistique exploratoire bivariée

Variables quantitatives et variables qualitatives

Une variable quantitative est une variable dont les valeurs sont numériques. On distingue deux types de variables quantitatives :

- les variables quantitatives discrètes : les valeurs appartiennent à un ensemble au plus dénombrable. Exemple : le nombre d'enfants.
- les variables quantitatives continues : les valeurs appartiennent à un ensemble non dénombrable. Exemple : la taille.

Une variable qualitative est une variable dont les valeurs ne sont pas numériques. Une modalité correspond à une valeur pouvant être prise par une variable qualitative. On distingue deux types de variables qualitatives :

- les variables qualitatives ordinales lors qu'il est possible d'ordonner les modalités. Exemple : "petit", "moyen" et "grand"
- les variables qualitatives nominales lorsqu'elles ne sont pas ordinales. Exemple : "rouge", "vert" et "bleu".

Lien entre deux variables qualitatives

Soient deux variables qualitatives Y_1 et Y_2 ayant respectivement J_1 et J_2 modalités. On considère un $n_{\bullet\bullet}$-échantillon du couple (Y_1, Y_2).

On appelle tableau de contingence associé à Y_1 et Y_2, le tableau suivant :

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1(J_1-1)} & n_{1J_1} \\ n_{21} & n_{22} & \cdots & n_{2(J_1-1)} & n_{2J_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{(J_2-1)1} & n_{(J_2-1)2} & \cdots & \cdots & n_{(J_2-1)(J_1)} \\ n_{J_2 1} & n_{J_2 2} & \cdots & \cdots & n_{J_2 J_1} \end{bmatrix}$$

où n_{ij} décrit le nombre d'individus ayant à la fois la modalité i de la variable Y_2 et la modalité j de la variable Y_1

Le tableau de contingence permet de résumer l'information contenue dans deux variables qualitatives.

Inconvénient : l'information par individu n'est plus disponible.

Notations

On notera :

- $n_{\bullet\bullet} = \sum_{i=1}^{J_2} \sum_{j=1}^{J_1} n_{ij} =$ nombre d'individus total
- $n_{i\bullet} = \sum_{j=1}^{J_1} n_{ij}$ (nombre d'individus ayant la modalité i à la variable Y_2)
- $n_{\bullet j} = \sum_{i=1}^{J_2} n_{ij}$ (nombre d'individus ayant la modalité j à la variable Y_1)
- $f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$ (fréquence d'individus ayant à la fois la modalité i à la variable Y_2 et j à la variable Y_1)
- $f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$ (fréquence d'individus ayant la modalité i à la variable Y_2)
- $f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$ (fréquence d'individus ayant la modalité j à la variable Y_1)

Lien entre deux variables qualitatives

On appelle statistique du Khi-Deux notée $\chi^2(Y_1, Y_2)$, la statistique définie par

$$\chi^2(Y_1, Y_2) = \sum_{i=1}^{J_2} \sum_{j=1}^{J_1} \frac{(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}})^2}{\frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}} = n_{\bullet\bullet} \sum_{i=1}^{J_2} \sum_{j=1}^{J_1} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}.$$

Interprétation : la statistique du $\chi^2(Y_1, Y_2)$ mesure "un écart à l'indépendance".

Considérons deux variables aléatoires discrètes X et Y sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. X et Y sont indépendantes $\iff$
$\forall (x, y) \in X(\Omega) \times Y(\Omega), \quad \mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = i)\mathbb{P}(Y = y).$

On peut s'attendre dans le cas où les deux variables sont indépendantes à ce que $f_{ij} \approx f_{i\bullet} f_{\bullet j}$. Dans ce cas, $\chi^2(Y_1, Y_2)$ sera faible (à effectif total constant).

Sous l'hypothèse d'indépendance des deux variables alors
$\chi^2(Y_1, Y_2) \underset{n_{\bullet\bullet} \to \infty}{\xrightarrow{\mathcal{L}}} \chi^2_{(J_1 - 1)(J_2 - 1)}$.

Lien entre deux variables qualitatives

Problème de la statistique du $\chi^2(Y_1, Y_2)$: pour une structure donnée, la statistique sera d'autant plus grande que l'effectif total $n_{\bullet\bullet}$ l'est.

Remarque : $\chi^2(Y_1, Y_2) < n_{\bullet\bullet} \min(J_1 - 1, J_2 - 1)$

On introduit d'autres indicateurs dérivés de la statistique du χ^2 :

- le coefficient de contingence de Pearson
  $C(Y_1, Y_2) = \sqrt{\dfrac{\chi^2(Y_1, Y_2)}{\chi^2(Y_1, Y_2) + n}}$
- le V de Cramer $V(Y_1, Y_2) = \sqrt{\dfrac{\chi^2(Y_1, Y_2)}{n_{\bullet\bullet} \min(J_1 - 1, J_2 - 1)}}$

$C \in [0; 1]$ et $V \in [0; 1]$

Lien entre une variable qualitative et une variable quantitative

On considère une variable quantitative X, une variable qualitative Y ayant J modalités et un échantillon de taille n du couple (X,Y) qu'on notera $(\{x_{1,1}, y_1\}, ..., \{x_{n_j,J}, y_n\})$.

La variable Y permet de créer une partition des individus en J groupes.

On note $\bar{x}_j$ la moyenne de $\{x_{i,j}\}_{i\in[1,n_j]}$ et $\bar{x}$, la moyenne des $x_{i,j}$.

Décomposition de la variance

$$\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(x_{i,j}-\bar{x})^2}_{=\text{SCT}} = \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(x_{i,j}-\bar{x}_j)^2}_{=\text{SCR}} + \underbrace{\sum_{j=1}^{J} n_j(\bar{x}_j-\bar{x})^2}_{=\text{SCE}}$$

On appelle rapport de corrélation de X et Y noté $\eta^2(X,Y)$, la statistique définie par $\eta^2(X,Y) = \dfrac{\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{x}_j-\bar{x})^2}{\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(x_{i,j}-\bar{x})^2} = \dfrac{\text{SCE}}{\text{SCT}}$

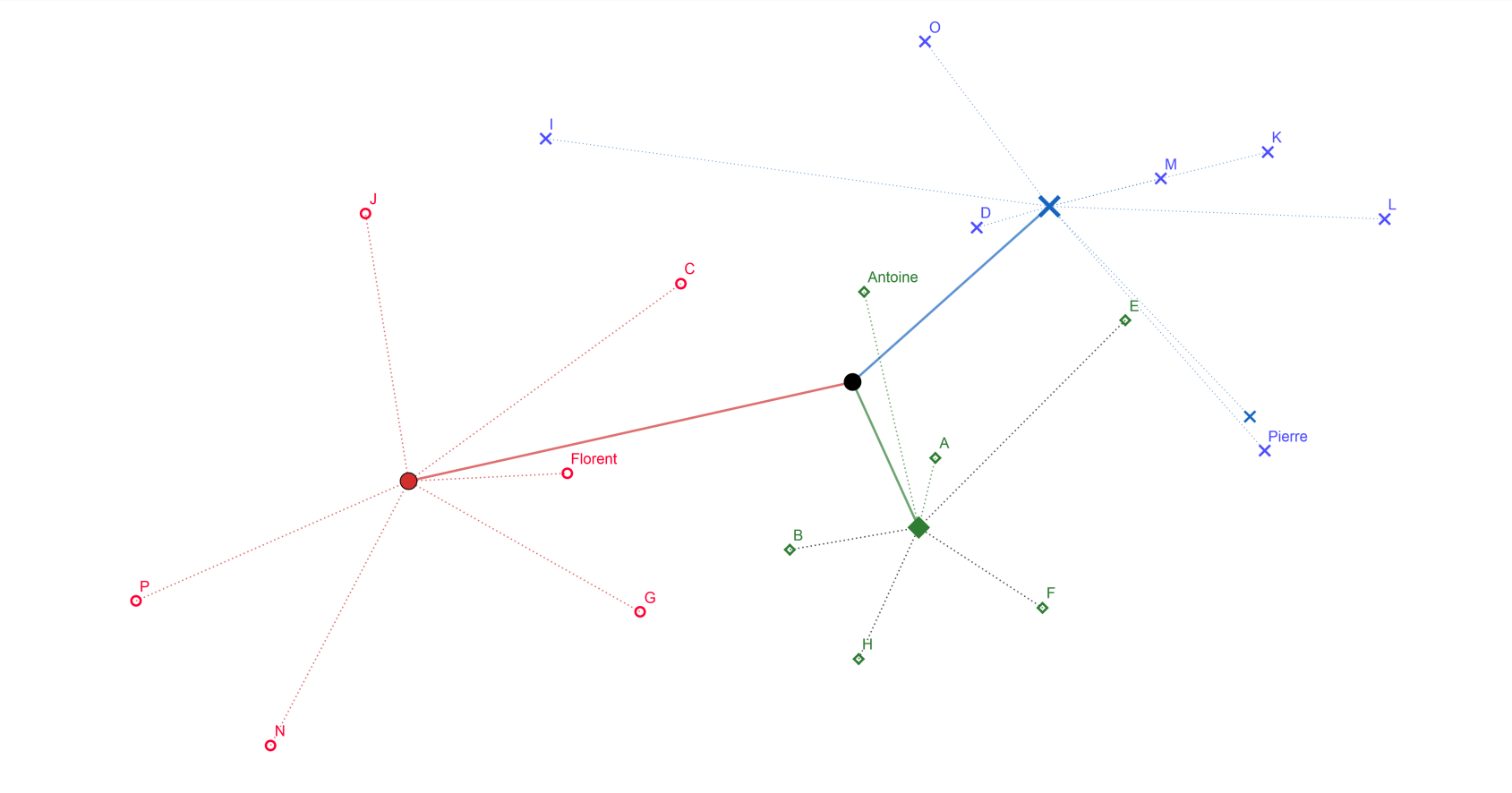


Figure 1: Décomposition de la variance

Lien entre une variable qualitative et une variable quantitative

Lorsque $\eta^2(X, Y) = 0$ alors $\forall j \in [1, J], \bar{x}_j = \bar{x}$: "la valeur de Y ne donne pas d'information sur la valeur de X".

Lorsque $\eta^2(X, Y) = 1$ alors $\forall j \in [1, J], x_{i,j} = \bar{x}_j$: "la valeur de Y détermine entièrement la valeur de X".

Résultat d'ANOVA : On suppose que X suit une loi normale (+ homoscédasticité dans chacun des groupes). On souhaite tester $(H_0) : \bar{x}_1 = \ldots = \bar{x}_J$ vs $\exists j \neq i$ tel que $\bar{x}_i \neq \bar{x}_j : (H_1)$.

Sous (H_0), la statistique $F = \frac{\frac{SCE}{J-1}}{\frac{SCR}{n-J}} \sim \mathcal{F}(J-1, n-J)$

Lien entre deux variables quantitatives

On considère deux variables quantitatives X et Y et un échantillon de taille n du couple (X, Y) qu'on notera $(\{x_1, y_1\}, ..., \{x_n, y_n\})$.

On appelle coefficient de corrélation de Pearson noté $\rho(X, Y)$ la statistique définie par $\rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}$.

- $\rho(X, Y) \in [-1, 1]$
- $\rho(X, Y) = 1 \iff \exists \alpha > 0$ tel que $X = \alpha Y$
- $\rho(X, Y) = -1 \iff \exists \alpha < 0$ tel que $X = \alpha Y$
- $\rho(X, Y) = 0 \implies$ il y a absence de lien linéaire entre X et Y
- $\forall (\lambda, \mu) \in \mathbb{R}^2$ avec $\lambda\mu > 0$, $\rho(\lambda X, \mu Y) = \rho(X, Y)$

Attention : le coefficient de corrélation de Pearson ne permet de détecter que les liens linéaires entre deux variables !

N'hésitez pas à faire des nuages de points de vos variables quantitatives (fonction *pairs* en R) !

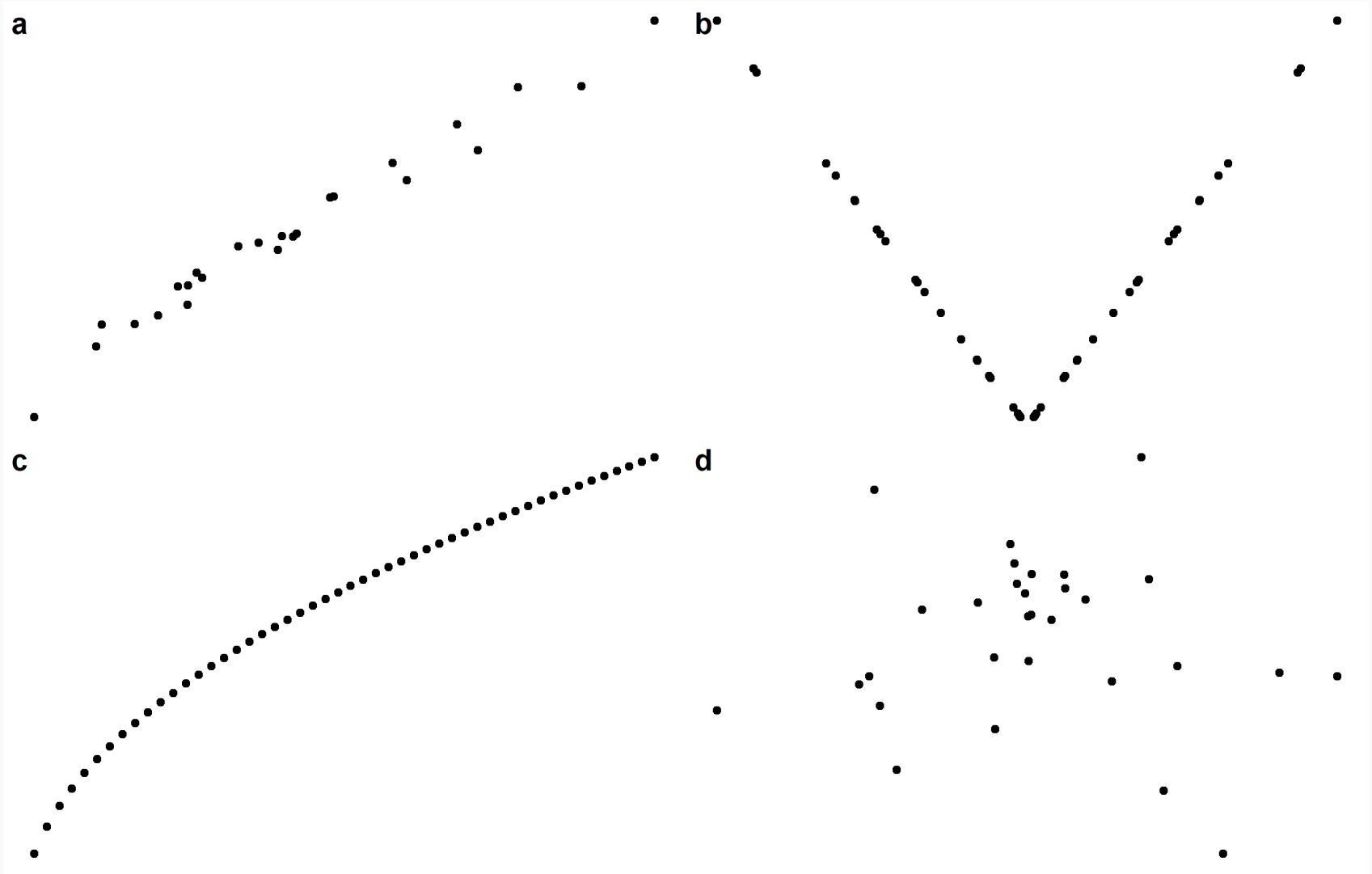


Figure 2: Coefficients de corrélation de Pearson et nuage de points

Les coefficients de corrélation de Pearson obtenus : (a) 0.988 (b) 0 (c) 0.982 (d) 0.03.

Attention : une corrélation n'est pas une relation de causalité.

Limites

Dans cet exemple, les valeurs de trois variables Y, X_1 et X_2 vont être générées de la manière suivante :

- $X_1 \sim \mathcal{N}(0, 1)$
- $X_2 = X_1 + \nu$ avec $\nu \sim \mathcal{N}(0, 0.25)$
- $Y = -2X_1 + 3X_2 + 4 + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, 0.5)$

	Y	X_1	X_2
Y	1	0.624	0.862
X_1	0.624	1	0.917
X_2	0.862	0.917	1

Table 1: Matrice des corrélations

Le coefficient de corrélation de Pearson ne permet pas de prendre en compte le lien linéaire de plus de variables. Solution : utiliser des méthodes de statistique exploratoire multivariée.

Analyse par composantes principales

Problématique

L'ACP permet d'analyser des tableaux de données de type n individus x p variables quantitatives.[1]

L'ACP met en lumière :

- des liens linéaires entre les p variables (variables très corrélées positivement, négativement ou décorrélées)
  De plus, ces liens pourront être synthétisées à l'aide d'un petit nombre de variables synthétiques appelées *composantes principales*.
  On dit que l'ACP est une méthode de réduction de la dimension.
- des ressemblances entre les n individus à l'aune des p variables.

L'ACP ne nécessite pas d'hypothèses probabilistes sur les données.

[1] Il est possible d'ajouter des variables qualitatives dans une ACP si elles sont placées en variables supplémentaires (voir)

Données brutes et notations

On considère un tableau de données n individus x p variables quantitatives. On associe à ce tableau une matrice $Y \in M_{n,p}(\mathbb{R})$ où le terme général $y_{i,j}$ correspond à la valeur prise par l'individu i pour la variable j.

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p-1} & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p-1} & y_{2,p} \\ \ldots & \ldots & y_{i,j} & \ldots & \ldots \\ y_{n-1,1} & y_{n-1,2} & \cdots & y_{n-1,p-1} & y_{n-1,p} \\ y_{n,1} & y_{n,2} & \cdots & y_{n,p-1} & y_{n,p} \end{bmatrix}$$

- Chaque individu i est décrit par un vecteur de $\mathbb{R}^p$: $\mathbf{y_i} = {}^t(y_{i,1}, ..., y_{i,p})$
- Chaque variable j est décrite par un vecteur de $\mathbb{R}^n$: $\mathbf{y^j} = {}^t(y_{1,j}, ..., y_{n,j})$

Dans la suite, on supposera $p < n$.

Matrice de poids et métrique

Matrice de poids

Les données peuvent avoir été obtenues à l'issue d'un échantillonage ou avoir fait l'objet de traitement de correction de la non-réponse par repondération.

Chaque individu dispose donc d'un poids. L'ensemble des poids est décrit à l'aide d'une matrice de poids P telle que

$$P = \text{diag}(p_1, ..., p_n) \text{ avec } \sum_{i=1}^{n} p_i = 1$$

où p_i désigne le poids associé à l'individu i.

Matrice de poids et métrique

Métrique

On considère $\mathbb{R}^p$ et $\mathbb{R}^n$ comme des espaces affines euclidiens (de directions respectives $\mathbb{R}^p$ et $\mathbb{R}^n$).

On note $<,>_M$ le produit scalaire associé à $\mathbb{R}^p$, d la distance euclidienne associée et M la matrice associée à la forme bilinéaire $<,>_M$ dans la base canonique de $\mathbb{R}^p$. M est appelée la *métrique*.

Projection orthogonale d'un individu sur une droite : soient un individu i et une droite Δ dont u est un vecteur unitaire (i.e $< \mathrm{u}, \mathrm{u} >_M = 1$). Alors la projection orthogonale de l'individu i est définie par

$$P_\Delta(y_i) = \underbrace{< y_i, u >_M}_{= \text{coordonnée}} \mathrm{u} = ({}^t y_i M u) u$$

Distance entre individus : soient deux individus i_1 et i_2, la distance (au carré) entre ces deux individus est définie par

$$d^2(y_{i_1}, y_{i_2}) =< y_{i_1} - y_{i_2}, y_{i_1} - y_{i_2} >_M$$

Nuages des individus et des variables

L'idée de l'ACP est de chercher des espaces de plus faible dimension sur lesquels projeter orthogonalement le nuage des individus en déformant le moins possible le nuage.

Nuages des individus

L'ensemble des points $\{y_i\}_{i\in[1,n]}$ est appelé nuage des individus et est noté $\mathcal{N}_I$. Chaque point appartient à $\mathbb{R}^p$.

La recherche d'espaces de plus faible dimension représentant le nuage des individus s'appelle l'analyse directe.

Nuages des variables

L'ensemble des points $\{y^j\}_{j\in[1,p]}$ est appelé nuage des variables et est noté $\mathcal{N}_V$. Chaque point appartient à $\mathbb{R}^n$

La recherche d'espaces de plus faible dimension représentant le nuage des individus s'appelle l'analyse duale.

Exemple de nuages des individus

On considère le tableau de données dont les quatre premières lignes sont les suivantes :

Nom	Taille (en cm)	Âge (en année)
Pierre	187	23
Antoine	160	30
David	186	24.5
Florent	140	22

La matrice associée est la matrice

$$Y = \begin{pmatrix} 187 & 23 \\ 160 & 30 \\ 186 & 24.5 \\ 140 & 22 \\ \vdots & \vdots \end{pmatrix}$$

Exemple de nuages des individus

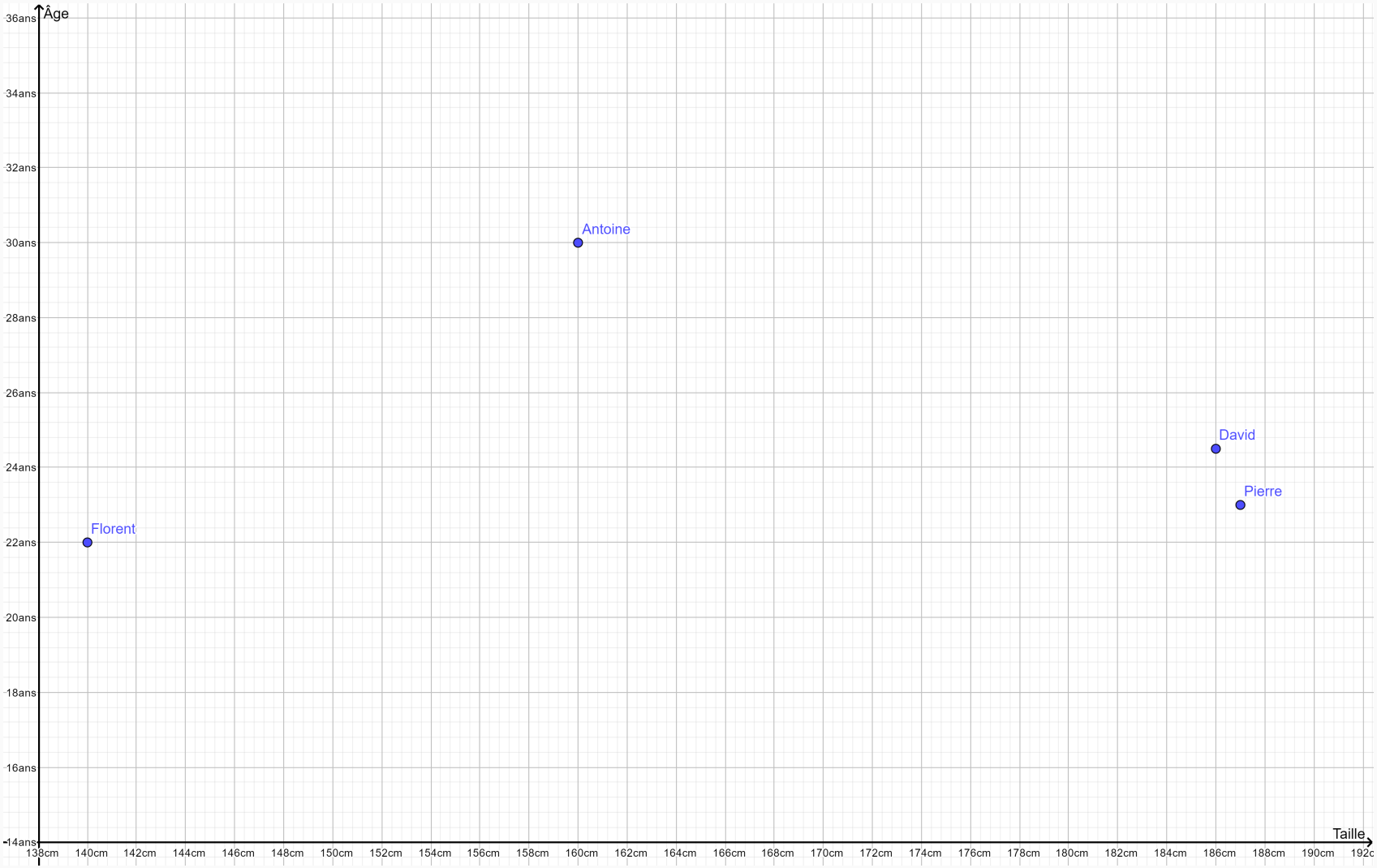


Figure 3: Nuages des individus

Exemple de nuage des variables

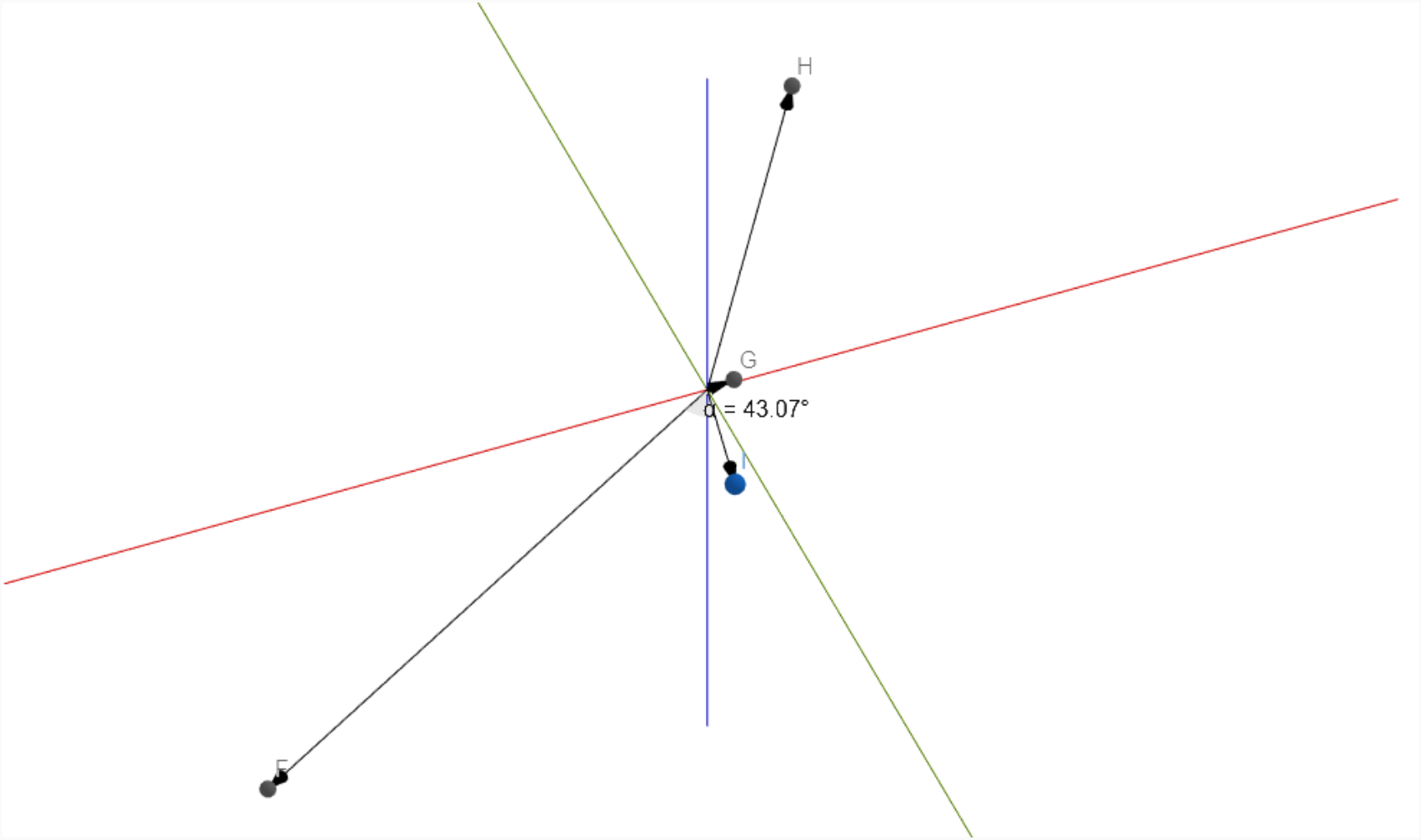


Figure 4: Nuage des variables avec $n = 3$ et $p = 4$

Données centrées-réduites et données brutes

Dans cet exemple, on considère $\mathbb{R}^p$ muni de sa structure euclidienne canonique.

	Pierre	Antoine	David
Antoine	27.8		
David	1.8	26.6	
Florent	47.0	21.5	46.0

Table 2: Distances entre individus quand la taille est en cm.

Pierre	Antoine	David
7.0		
1.5	5.5	
1.1	8.0	2.54

Table 3: Distances entre individus quand la taille est en m.

Les distances entre individus sont sensibles à l'unité des variables. Plus une variable prend de grandes valeurs, plus son poids dans la distance sera grand. Solution : réduire les données.

Le tableau de données est :

- centré si $\bar{Y} = {}^t(\bar{y^1}, ..., \bar{y^p}) = {}^t 0_{\mathbb{R}^p}$
- réduit si $\forall j \in [1, p], (\sigma^j)^2 = \sum_{i=1}^{n} p_i (y_{i,j} - \bar{y^j})^2 = 1$

Centrer

Soit Y la matrice des données alors la matrice centrée de Y notée X_c est la matrice de terme général $x^c_{i,j} = (y_{i,j} - \bar{y^j})$.

Dans le nuage des individus, *centrer* engendre une translation du nuage de manière à ce que le centre de gravité se confonde avec l'origine.

Dans le nuage des variables, *centrer* engendre une projection du nuage sur l'orthogonal (au sens de la métrique P) de la première bissectrice de $\mathbb{R}^n$. En effet, $X_c = (I_n - \mathbf{1}^t\mathbf{1}P)Y$ avec $\mathbf{1} = {}^t(1, ..., 1)$.

En pratique, l'ACP est toujours réalisée sur des données centrées. Les données sont automatiquement centrées par R.

Exemple de nuages des individus

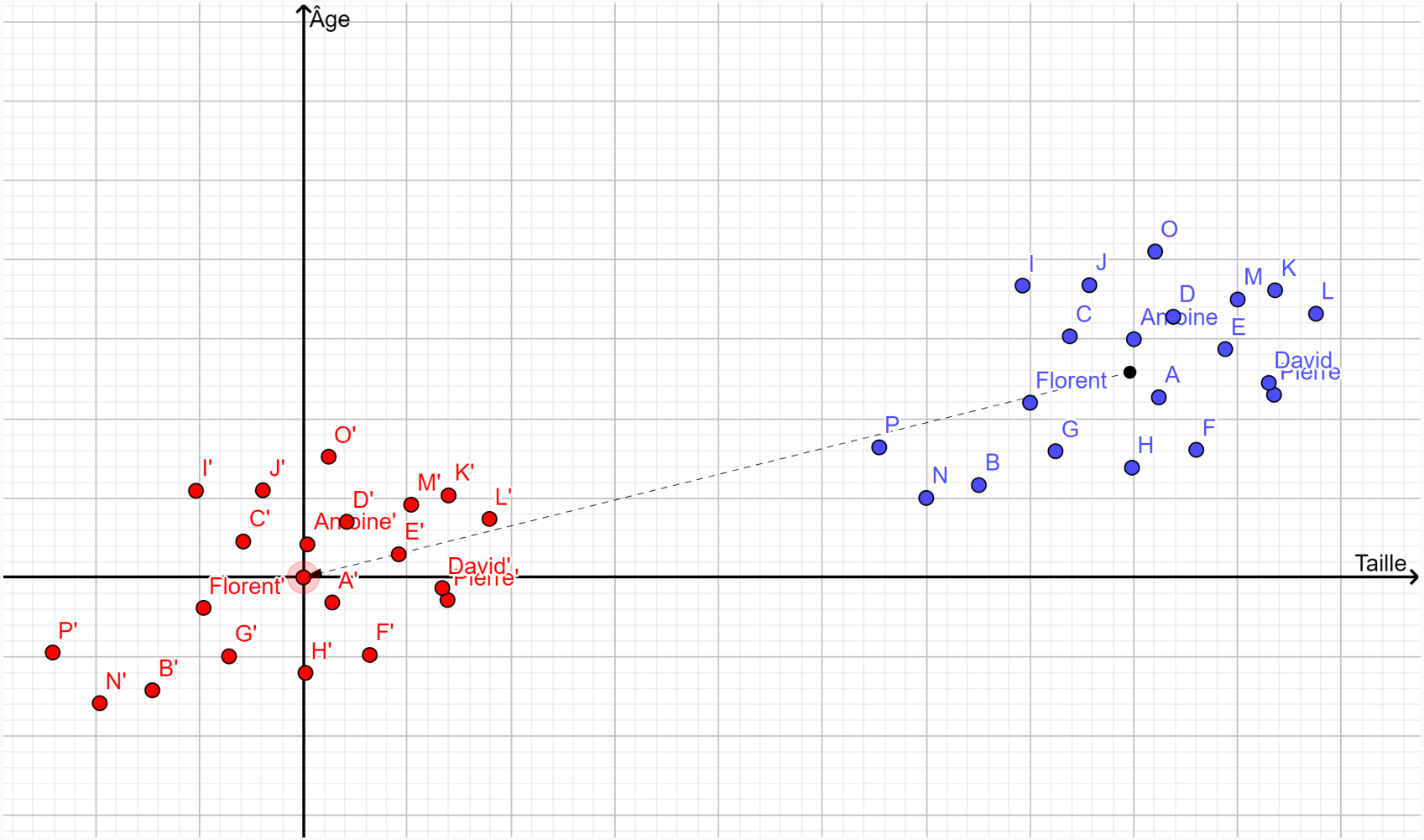


Figure 5: Nuage des individus brut et centré

Données centrées-réduites et données brutes

Réduire

Soit Y la matrice des données et X_c la matrice centrée de Y.

La matrice réduite de Y est la matrice $Y\Sigma^{-1}$ avec $\Sigma = \text{diag}(\sigma^1, ..., \sigma^p)$.

Dans le nuage des variables, *réduire* revient à normaliser chaque variable. Autrement dit, $\forall j \in [1, p], ||x^j||_P^2 = 1$: les variables sont donc sur la sphère unité $\mathbb{S}^{p-1}$.

La matrice centrée réduite de Y notée X est telle que
$X = (I_n - \mathbf{1}^t\mathbf{1}P)Y\Sigma^{-1} = X_c\Sigma^{-1}$ d'où $(x_{i,j} = \frac{y_{i,j} - \bar{y^j}}{\sigma^j})$

Quand réduit-on les données ?

- Quand les variables ne sont pas homogènes (ex : vitesse vs poids);
- Quand les données ne sont pas de la même échelle.

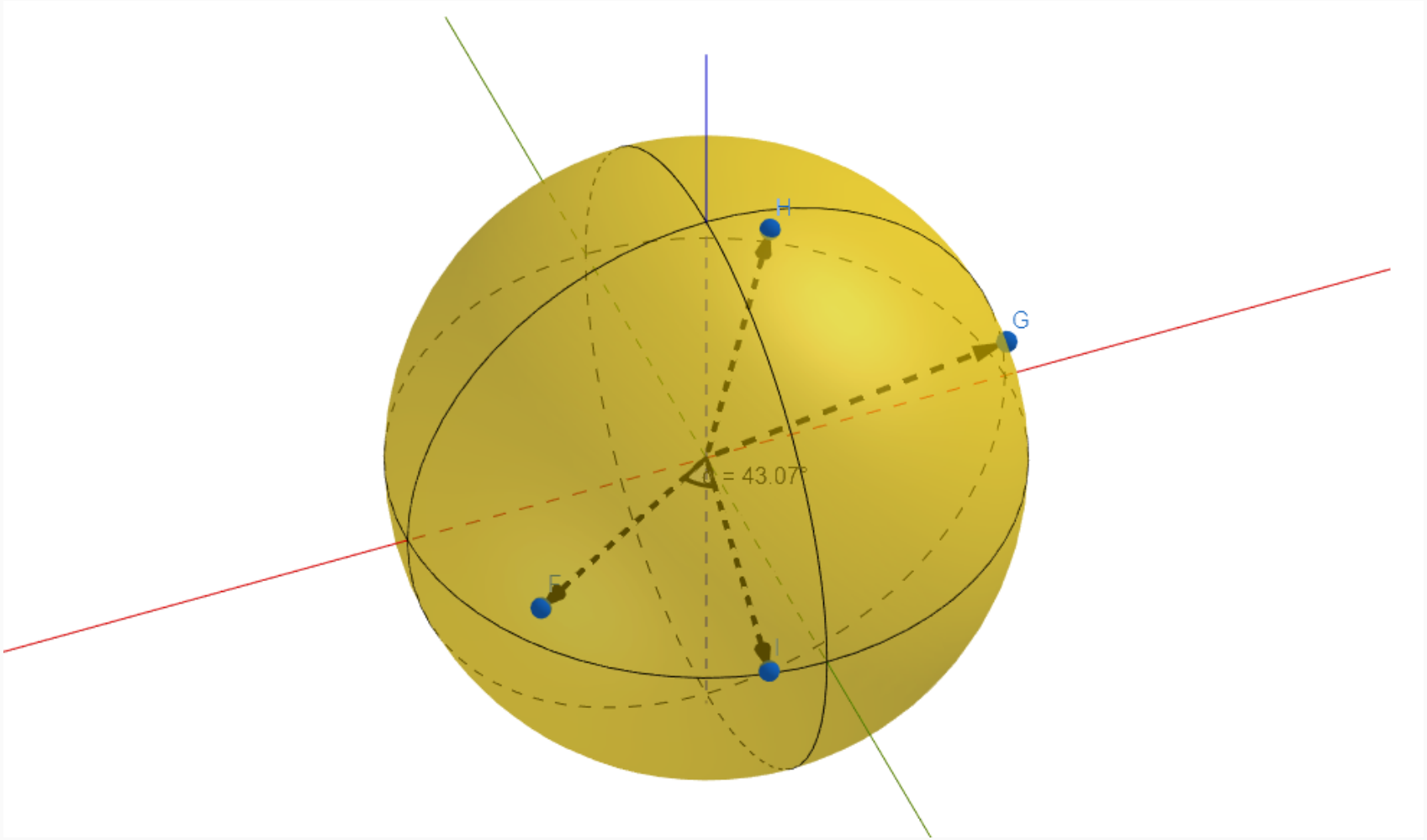


Figure 6: Nuage des variables réduit

Remarque : le cosinus de l'angle formé par deux variables correspond à leur coefficient de corrélation de Pearson.

Analyse directe et analyse duale

Une analyse est définie à l'aide d'un triplet de matrice :

- une matrice contenant les données (centrée au moins),
- une matrice de poids,
- une métrique.

Pour l'analyse directe, le triplet est (X_c, P, M) et pour l'analyse duale, le triplet est $({}^tX_c, M, P)$.

On dira que :

- l'analyse est une ACP lorsque la matrice des données correspond à une table de type individus x variables,
- l'analyse est une ACP normée lors l'ACP est effectuée sur des données centrées et réduites. Le triplet de l'analyse directe est donc (X, P, M) et celui de l'analyse duale est $({}^tX, M, P)$.

Dans la suite, on prendra $M = Id_p$.

Inertie d'un nuage de points

L'inertie du nuage des individus par rapport au barycentre est définie par

$$I_0(G) = \sum_{i=1}^{n} p_i d^2(i, G) = \underbrace{\sum_{i=1}^{n} p_i d^2(i, O)}_{\text{si centré}}.$$

L'inertie permet de quantifier la variabilité d'un nuage de points.

Remarque : dans le cas d'une ACP normée, $I_0(G) = p$.

Analyse directe

L'idée de l'ACP est de chercher une droite Δ_1 telle que l'inertie de la projection orthogonale du nuage des individus sur cette droite $I_1(G)$ soit maximale (= minimise la perte d'information engendrée par la projection (application 1-lipschitzienne)).

$$I_1(G) = \sum_{i=1}^{n} p_i d^2(P_{\Delta_1}(x_i), 0) \leq \sum_{i=1}^{n} p_i d^2(x_i, 0) = I_0(G)$$

Cette démarche est équivalente à chercher une droite qui maximise

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d^2(P_{\Delta 1}(x_i), P_{\Delta 1}(x_j)) \ (= 2nI_1(G))$$

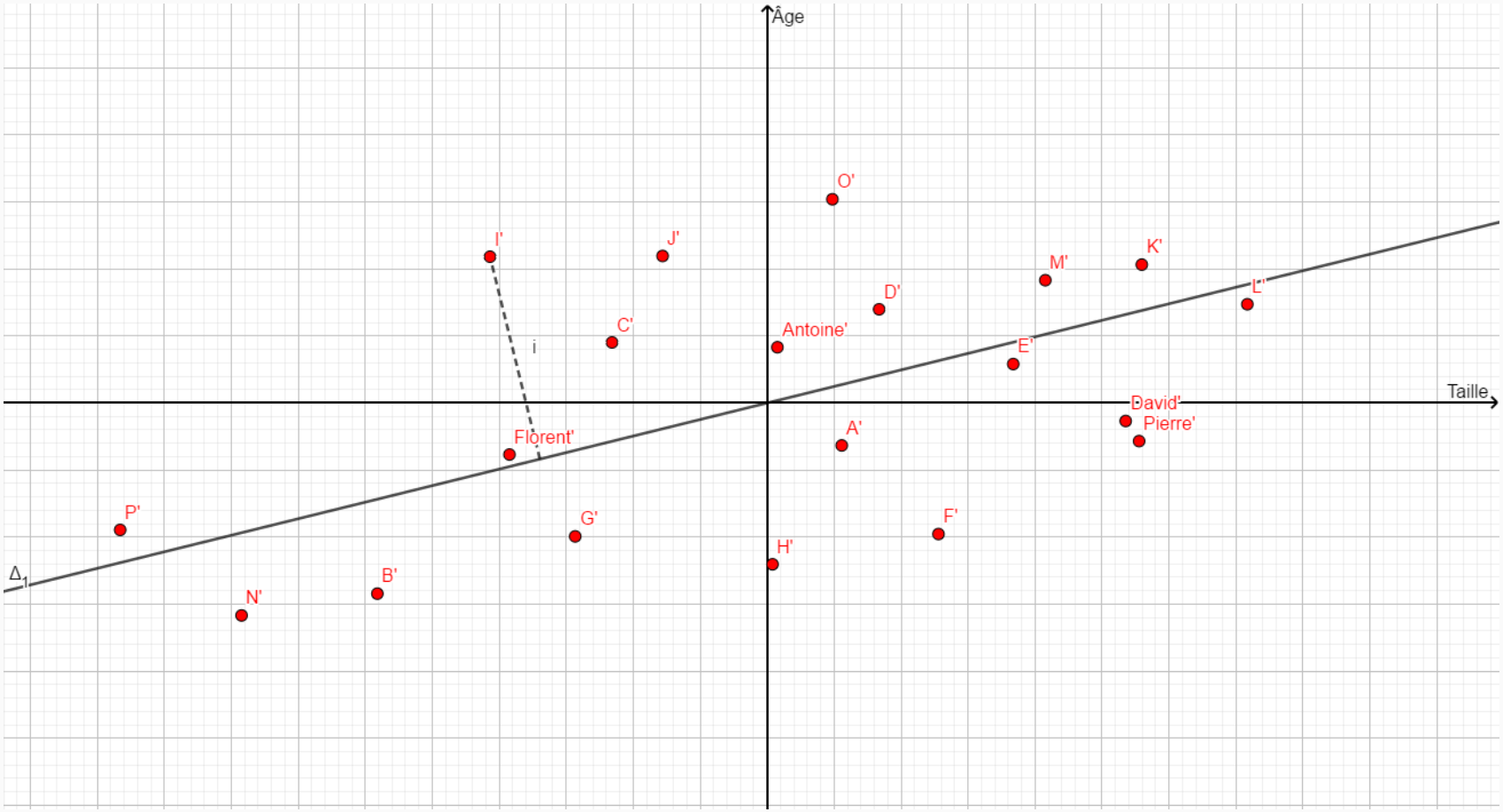


Figure 7: Recherche du premier axe factoriel

Analyse directe

Puis une fois la droite Δ_1 obtenue (à l'aide d'un vecteur unitaire), on cherche une deuxième droite Δ_2 qui maximise l'inertie de la projection du nuage des individus sur cette droite et qui est orthogonale à Δ_1.

On réitère en cherchant une droite Δ_k qui maximise l'inertie projetée sur cette droite et qui est orthogonale à $\Delta_1, ..., \Delta_{k-1}$

On a ainsi p droites chacune définie à l'aide d'un vecteur unitaire. Ces droites sont appelées *axes factoriels*.

On notera u_k, un vecteur unitaire de Δ_k.

Pour obtenir une représentation en plus faible dimension, le nuage des individus sera projeté sur l'espace $\text{Vect}(u_1, ..., u_k)$. Mais comment choisir k ?

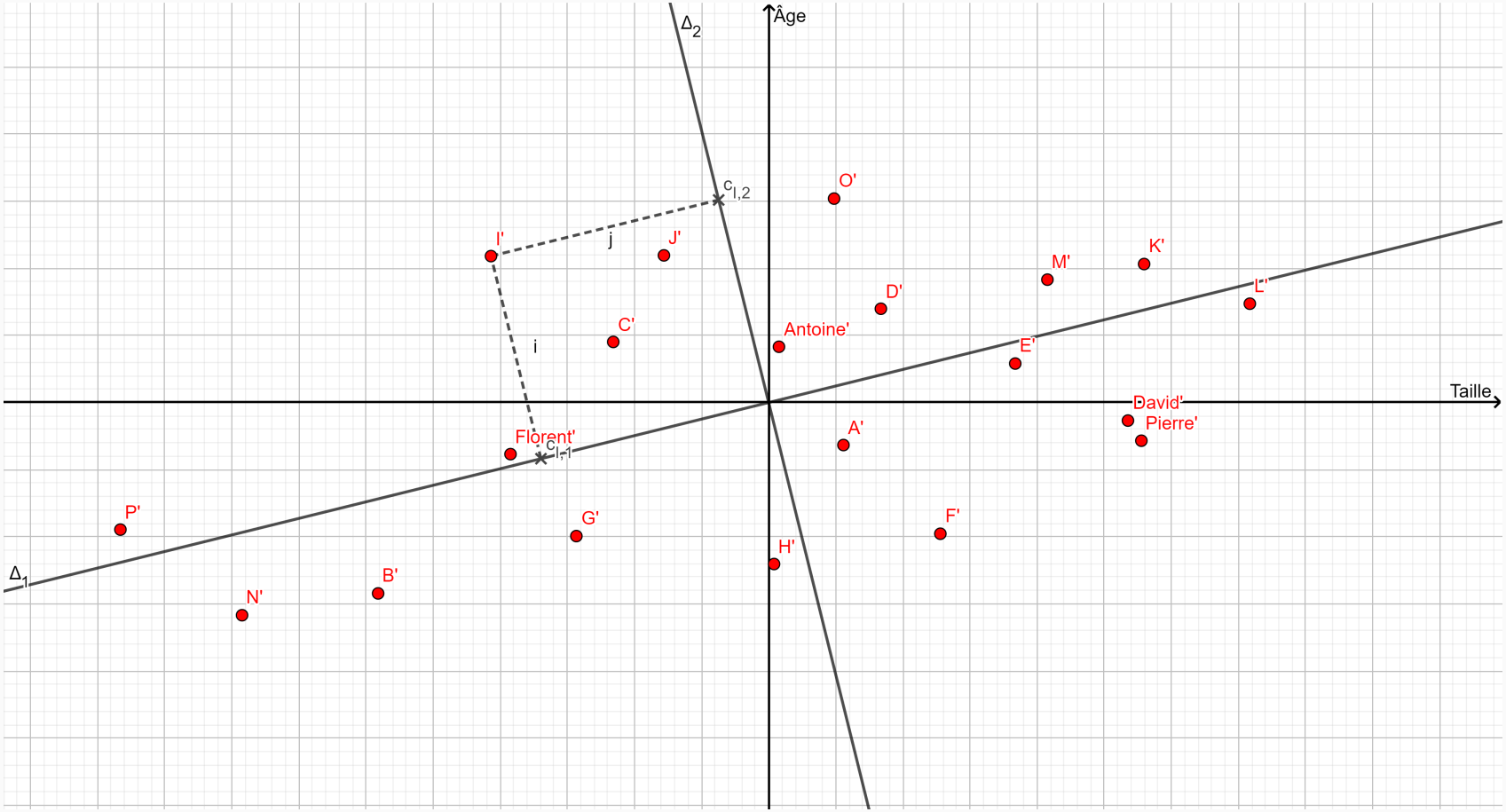


Figure 8: Recherche du deuxième axe factoriel

Formalisation de l'analyse directe

On souhaite déterminer une expression des axes factoriels de l'analyse directe (X, P, M).

Soit u_1 un vecteur unitaire de la droite maximisant I_1.

Les coordonnées des projections sur Δ_1 du nuage des individus sont données par $C_1 = XMu_1 = Xu_1$.

Or $I_1(G) = {}^tC_1PC_1 = {}^tu_1{}^tM^tXPXMu_1$.

u_1 est donc solution du problème d'optimisation suivant :

$$\max_{u \in \mathbb{R}^p} \quad {}^tu^tM^tXPXMu \quad \text{s.c } {}^tuMu = 1$$

Formalisation de l'analyse directe

La solution de ce problème est le vecteur unitaire (au signe près) de l'espace propre associé à la plus grande valeur propre de la matrice ${}^tXPXM$ (si l'espace propre est de dimension supérieure à une, on prendra un vecteur d'une base orthonormée).

On montre que le vecteur unitaire (au sens près) maximisant I_k et orthogonal aux vecteurs $u_1, ..., u_{k-1}$ est le vecteur unitaire associé à la k-ième plus grande valeur propre de la matrice ${}^tXPXM$

La matrice $V = {}^tXPXM$ est appelé matrice d'inertie.

Lorsque les données sont centrées et réduites alors la matrice d'inertie se confond avec la matrice de corrélation.

Quelques résultats

Décomposition de l'inertie sur les axes factoriels

Soit I_k, l'inertie de la projection orthogonale du nuage des individus sur la droite Δ_k (i.e $I_k = \sum_{i=1}^{n} p_i d^2(P_{\Delta_k}(x_i), 0)$). On a $I_0 = \sum_{k=1}^{p} I_k$

De plus, $I_k = \lambda_k$.

En ACP normée, $I_0 = p$.

Composantes principales

Composantes principales

On appelle k-ième composante principale, noté C_k, le vecteur $C_k = {}^tXMu_k = {}^tXu_k$.

- une composante principale est une combinaison linéaire des $\{x^j\}_{j\in[1,p]}$,
- une composante principale est centrée
- la k-ième composante principale a une variance empirique égale à λ_k
- deux composantes principales différentes sont décorrélées (empiriquement).

Analyse duale

De manière similaire, on cherche une représentation de plus faible dimension du nuage des variables. Cela revient à effectuer l'analyse du triplet $({}^tX, M, P)$.

Comme précédemment, on cherche une droite Δ_1' de vecteur unitaire v_1 telle que l'inertie de la projection du nuage des variables sur la droite soit maximale.

v_1 est solution du problème d'optimisation suivant :

$$\max_{v \in \mathbb{R}^n} \quad {}^tv\,{}^tPXM\,{}^tXPv \quad \text{s.c } {}^tvPv = 1$$

De même on montre que v_1 est le vecteur unitaire (au signe près) de l'espace propre associée à la plus grande valeur propre de la matrice $XM\,{}^tXP$

Puis on cherche séquentiellement d'autres vecteurs unitaires orthogonaux à ceux déjà trouvés et maximisant l'inertie projetée.

Relations de transition

Si $p \leq n$ alors $\mathrm{Sp}(XM{}^tXP) = \mathrm{Sp}({}^tXPXM) \backslash \{0\}$.

De plus,

$$\forall k \in [1, p], u_k = \frac{1}{\sqrt{\lambda_k}} {}^tXPv_k$$

$$\forall k \in [1, n], v_k = \frac{1}{\sqrt{\lambda_k}} XMu_k$$

$$\forall k \in [1, p], C_k = \frac{1}{\sqrt{\lambda_k}} XMD_k$$

$$\forall k \in [1, n], D_k = \frac{1}{\sqrt{\lambda_k}} {}^tXPC_k$$

Les résultats de l'analyse directe permettent de retrouver les résultats de l'analyse duale.

Choix du nombre d'axes

Pour obtenir une représentation en plus faible dimension, le nuage des individus sera projeté sur l'espace Vect$(u_1, ..., u_k)$. Mais comment choisir k ?

Plusieurs critères :

- critère de l'inertie moyenne : on garde tous les axes tels que $I_k > \frac{I_0}{p}$. (lorsque l'ACP est normée ce critère devient le critère de Kaiser : on considère les axes dont l'inertie projetée est supérieure à 1)
- critère du coude : on observe l'apparition de coudes dans le diagramme en bar formé par les valeurs propres de V,
- scree-test de Catell : on garde tous les axes avant que la différence seconde des valeurs propres de V devienne négative.

Si différents nombres d'axes à retenir sont obtenus en utilisant les différents critères, on considérera l'interprétabilité des axes.

Il existe d'autres critères : laplacien d'Anderson, bootstrap, ...

Choix du nombre d'axes

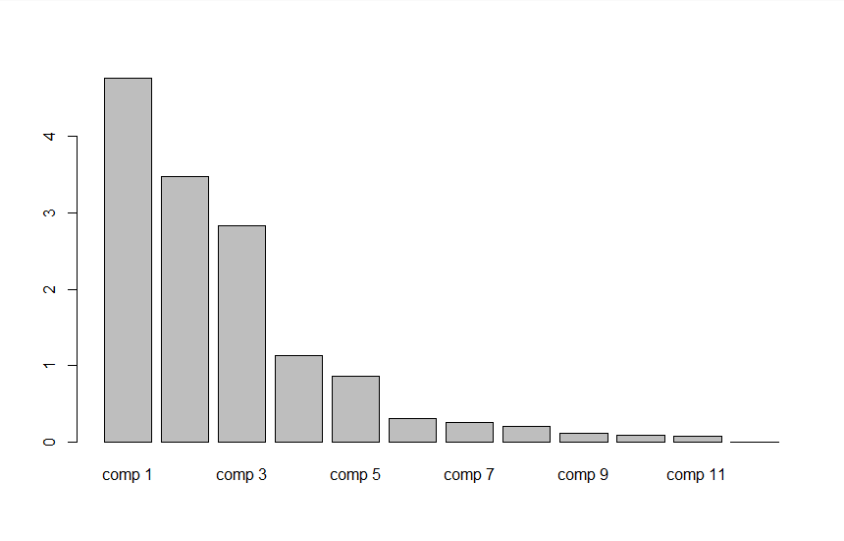


Figure 9: Histogramme des valeurs propres

	λ_k	δ_k	ε_k
Axe 1	4.76		
Axe 2	3.48	1.28	
Axe 3	2.83	0.65	0.63
Axe 4	1.13	1.70	-1.05
Axe 5	0.86	0.27	1.43
Axe 6	0.30	0.56	-0.29
Axe 7	0.26	0.05	0.51
Axe 8	0.21	0.05	-0.00
Axe 9	0.12	0.09	-0.04
Axe 10	0.09	0.02	0.06
Axe 11	0.07	0.02	0.00
Axe 12	0.00	0.07	-0.05

Table 4: Différences des valeurs propres : $\delta_k = \lambda_{k-1} - \lambda_k$ et $\varepsilon_k = \delta_{k-1} - \delta_k$

Aide à l'interprétation pour les individus

Certains indivudus peuvent être mal représentés sur un axe ou ils peuvent contribuer très fortement à la conception de l'axe : les aides à l'interprétation vont nous aider à les identifier.

- La coordonnée de l'individu i sur l'axe k est notée c_{ik}. Il s'agit de la i-ème composante de la k-ème composante principale.
- La contribution (absolue) de l'individu i à l'inertie projetée sur l'axe k notée $\text{CTRB}_k(i)$ est telle que $\text{CTRB}_k(i) = \dfrac{p_i c_{ik}^2}{\lambda_k}$.

  Remarque : $\displaystyle\sum_{i=1}^{n} \text{CTRB}_k(i) = 1$
- La qualité de représentation d'un individu i sur l'axe k noté $\text{QLT}_k(i)$ est telle que $\text{QLT}_k(i) = \dfrac{< x_i, u_k >_M^2}{||x_i||_M^2 \underbrace{||u_i||_M^2}_{=1}} = \dfrac{c_{ik}^2}{||x_i||_M^2} = \cos^2(\theta_{x_i,u_i})$

  Remarque : $\displaystyle\sum_{k=1}^{p} \text{QLT}_k(i) = 1$

	Forte qualité de représentation	Faible qualité de représentation
Forte contribution	Individu ayant participé fortement à la formation de l'axe et sur lequel il est bien représenté. Cet individu risque d'être moins bien représenté sur les autres axes.	Individu ayant participé fortement à la formation de l'axe mais mal représenté. Cet individu est bien représenté sur d'autres axes.
Faible contribution	Individu ayant peu contribué à la formation de l'axe mais bien représenté dessus. Cet individu *illustre* bien l'axe	

Table 5: Cas possibles en fonction de la contribution et de la qualité de représentation des individus

Aide à l'interprétation pour les variables

De même, il est nécessaire de disposer d'aide à l'interprétation lors de l'analyse du nuage des variables.

- La coordonnée d'une variable j sur l'axe k est notée $d_j k$. Il s'agit de la j-ième composante de D_k.
- La contribution (absolue) de la variable j à l'inertie projetée sur l'axe k noté $\text{CTRB}_k(j)$ est telle que $\text{CTRB}_k(j) = \frac{d_{jk}^2}{\lambda_k}$.

  Remarque : $\sum_{j=1}^{p} \text{CTRB}_k(j) = 1$
- La qualité de représentation d'une variable j sur l'axe k notée $\text{QLT}_k(j)$ est telle que $\text{QLT}_k(j) = \frac{< x^j, v_k >_P^2}{\underbrace{||x^j||_P^2}_{=1 \text{ en ACP normée}} \underbrace{||v_k||_P^2}_{=1}} = d_{jk}^2$.

  Remarque : $\sum_{k=1}^{n} \text{QLT}_k(j) = 1$

$$\text{CTRB}_k(j) = \frac{\text{QLT}_k(j)}{\lambda_k} = \frac{d_{jk}^2}{\lambda_k} \text{ en ACP normée}$$

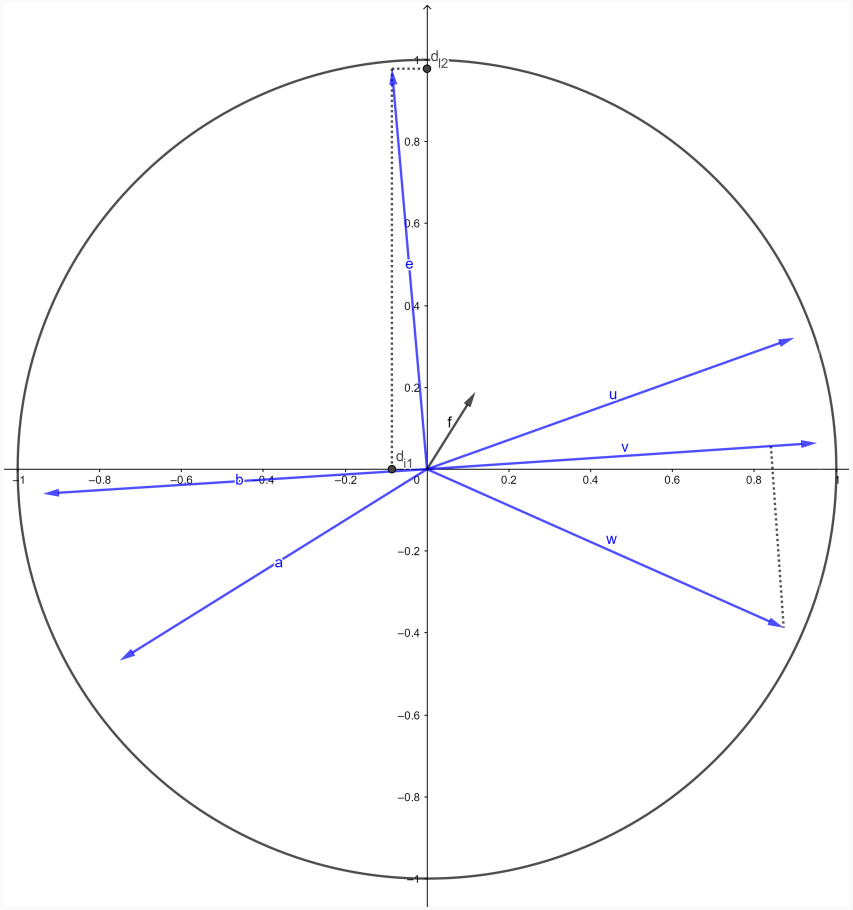


Figure 10: Premier plan factoriel (nuage des variables)

- en ACP normée, les projections des variables dans un plan factoriel sont à l'intérieur du cercle unité.
- en ACP normée, la corrélation entre une composante C_k et une variable x_j correspond à la coordonnée de cette variable sur l'axe factoriel correspondant.
- en ACP normée et sur un plan factoriel, le coefficient de corrélation entre deux variables x^{j_1} et x^{j_2} est approximativement la coordonnée de la projection de l'une des variables sur l'autre si les projections sur le plan factoriel sont près du bord du cercle.

Effet taille

- Supposons que la matrice de corrélation soit à terme strictement positif
- D'après le théorème de Perron-Frobenius, il existe un vecteur propre unitaire dont les coordonnées sont strictement positives (associé au premier axe).

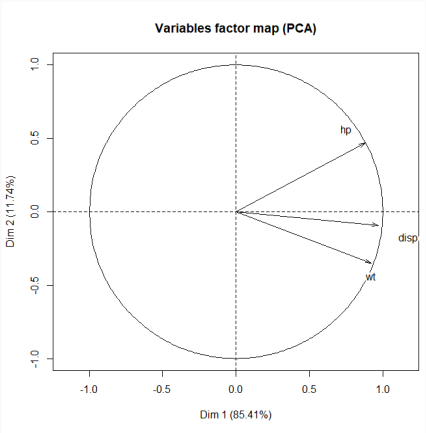


Figure 11: Visualisation de l'effet taille sur le premier plan factoriel

- Or $d_{jk} = \sqrt{\lambda_k} u_{jk}$.

Conséquences :

- Toutes les variables seront du même côté sur le premier axe factoriel
- Le premier axe va opposer les individus prenant de fortes valeurs sur toutes les variables vs les individus prenant de faibles valeurs (apparition d'un gradient).

Éléments supplémentaires

Il peut être intéressant d'inclure des individus ou des variables dans une analyse sans pour autant qu'ils contribuent à la conception des axes (individus atypiques, variables signalétiques ...).

On procède en deux étapes : les axes factoriels sont calculés sans ces individus et ces variables, puis ils y sont projetés.

Ces éléments sont appelés *individus et variables supplémentaires*. Les éléments non supplémentaires sont les *éléments actifs*.

Soient $x_{i,sup}$ et $x^{j,sup}$ respectivement un individu et une variable quantitative supplémentaires. Avec les notations précédentes, on a pour le kème axe :

$$c_{(i,sup),k} =< x_{(i,sup)}, u_k >_M \text{ et } d_{(j,sup),k} =< x^{(j,sup)}, v_k >_P$$

Question : que vaut la contribution d'un élément supplémentaire ?

Il est également possible d'ajouter des variables supplémentaires qualitatives.

Considérons $x^{j,sup}$ une variable qualitative à Q_j modalités.

Pour chaque modalité q, on note $\bar{x}_q = (\sum_{i \in \mathcal{N}_q} p_i)^{-1} \sum_{i \in \mathcal{N}_q} p_i x_i$ l'individu moyen ayant cette modalité.

La variable qualitative supplémentaire sera décrite à l'aide du sous-nuage $\{\bar{x}_m\}_{m \in [1, Q_j]}$ de $\mathbb{R}^p$ alors qu'une variable quantitative supplémentaire est décrite par un vecteur de $\mathbb{R}^n$.

Les aides à l'interprétation disponibles sont sur chaque axe factoriel :

- les coordonnées (individus, variables et individus moyen par modalité),
- la qualité de représentation (individus, variables et individus moyen par modalité),
- la valeur-test par modalité (seulement pour les variables supplémentaires qualitatives).

La valeur-test d'une modalité q d'une variable qualitative $x^{j,sup}$ pour un axe factoriel k est la statistique d'un test de nullité de la moyenne des coordonnées des projections (sur cet axe) des individus ayant cette modalité.

On teste $(H_0) : \bar{c}_{q,k} = (\sum_{i \in \mathcal{N}_q} p_i)^{-1} \sum_{i \in \mathcal{N}_q} p_i c_{ik} = 0$ vs non$(H_0) : (H_1)$

La statistique de test (ici donc la valeur-test) vaut $VT = \dfrac{\bar{c}_{q,k}}{\sqrt{\frac{n-n_q}{n_q(n-1)}}}$ avec

$n_q = \text{Card}(\mathcal{N}_q)$.

Sous (H_0), $VT \xrightarrow[n_q \to \infty]{\mathcal{L}} \mathcal{N}(0,1)$.

Méthode pour réaliser une ACP

1. Vérifier les données ("qualité", type, format ...), choisir entre ACP et ACP normée et choisir les individus et variables supplémentaires.
2. Choix du nombre d'axes à l'aide des critères de sélection (coude, scree-test de Catell, critère de l'inertie moyenne ...).
3. Pour chaque axe :
   - analyse des variables : sélection des variables ayant contribué le plus (et les mieux représentés si ACP non normée), étude des coordonnées puis synthèse,
   - analyse des individus : sélection des individus ayant contribué le plus et les mieux représentés, étude des coordonnées puis synthèse,
   - analyse des éléments supplémentaires,
   - analyse conjointe.
4. Refaire l'analyse en modifiant les individus / variables supplémentaires.

Analyse factorielle des correspondances

Problématique

L'analyse factorielle des correspondances est une méthode de réduction de la dimension.

L'AFC permet d'analyser des tableaux de contingences : elle permet de mettre en avant des lignes (resp. des colonnes) qui se ressemblent (ou se différencient) à l'aune de "leurs répartitions" sur les modalités en colonne (resp. en ligne).

L'AFC permet de réaliser des typologies des lignes et des colonnes.

Contrairement à l'ACP, en AFC les lignes et les colonnes jouent des rôles symétriques : ils représentent les modalités de chaque variable.

L'AFC permet d'obtenir une représentation graphique

Définition

Soient deux variables qualitatives Y_1 et Y_2 ayant respectivement J_1 et J_2 modalités. On considère un échantillon de taille $n_{\bullet\bullet}$ du couple (Y_1, Y_2).

On considère le tableau de contingence associé à cet échantillon :

$$C = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1(J_2-1)} & n_{1J_2} \\ n_{21} & n_{22} & \cdots & n_{2(J_2-1)} & n_{2J_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{(J_1-1)1} & n_{(J_1-1)2} & \cdots & \cdots & n_{(J_1-1)(J_2)} \\ n_{J_1 1} & n_{J_1 2} & \cdots & \cdots & n_{J_1 J_2} \end{bmatrix}$$

On note $F = \frac{1}{n_{\bullet\bullet}} C$.

Cheveux	Yeux				
	Bleu	Clair	Marron	Noir	Marge
Blond	326	688	343	98	1455
Roux	38	116	84	48	286
Châtain clair	241	584	909	403	2137
Châtain foncé	110	188	412	681	1391
Brun	3	4	26	85	118
Marge	718	1580	1774	1315	5387

Définition profils-lignes

On appelle matrice des profils-lignes, la matrice suivante :

$$\begin{bmatrix} \frac{f_{11}}{f_{1\bullet}} & \frac{f_{12}}{f_{1\bullet}} & \dots & \frac{f_{1(J_2-1)}}{f_{1\bullet}} & \frac{f_{1J_2}}{f_{1\bullet}} \\ \frac{f_{21}}{f_{2\bullet}} & \frac{f_{22}}{f_{2\bullet}} & \dots & \frac{f_{2(J_2-1)}}{f_{2\bullet}} & \frac{f_{2J_2}}{f_{2\bullet}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{f_{(J_1-1)1}}{f_{(J_1-1)\bullet}} & \frac{f_{(J_1-1)2}}{f_{(J_1-1)\bullet}} & \dots & \dots & \frac{f_{(J_1-1)(J_2)}}{f_{(J_1-1)\bullet}} \\ \frac{f_{J_1 1}}{f_{J_1\bullet}} & \frac{f_{J_1 2}}{f_{J_1\bullet}} & \dots & \dots & \frac{f_{J_1 J_2}}{f_{J_1\bullet}} \end{bmatrix}$$

Les lignes de la matrice des profils lignes sont les profils lignes.

La i-ième ligne de la matrice des profils lignes correspond à la distribution empirique de la variable Y_2 conditionnellement à $Y_1 = y_i$.

À chaque profil ligne, on associe un point de $\mathbb{R}^{J_2}$ dont les coordonnées dans la base canonique correspondent à son profil ligne.

On appelle nuage des profils lignes, le nuage de points formé par l'ensemble des profils lignes.

Cheveux	Yeux: Bleu	Clair	Marron	Noir	Marge
Blond	22.4	47.3	23.6	6.7	100
Roux	13.3	40.6	29.4	16.8	100
Châtain clair	11.3	27.3	42.6	18.9	100
Châtain foncé	7.9	13.4	29.9	48.7	100
Brun	0.2	3.4	22.0	72.0	100
Profil-ligne moyen	13.3	29.3	32.9	24.4	

Pondération et distance entre profils-lignes

Dans le cas de l'AFC :

- le poids d'un profil-ligne p_i correspond à la fréquence d'apparition de la modalité correspondante :

$$p_i = f_{i\bullet}$$

et

$$D_{J_1} = \text{diag}(f_{1\bullet}, ..., f_{J_1\bullet})$$

Il en vient que le barycentre des profils-lignes correspond au profil ligne moyen (observé sur toute la population) $f_{\bullet \text{moyen}} = (f_{\bullet 1}, ..., f_{\bullet J_2})$

- la distance entre profils-lignes est la distance du χ^2 :

$$\forall\ (i_1, i_2) \in [1, J_1]^2, \quad d^2_{\chi^2}(i_1, i_2) = \sum_{j=1}^{J_2} \frac{1}{f_{\bullet j}} \left(\frac{f_{i_1 j}}{f_{i_1 \bullet}} - \frac{f_{i_2 j}}{f_{i_2 \bullet}} \right)^2$$

Le terme $\frac{1}{f_{\bullet j}}$ permet d'équilibrer l'influence des modalités trop fréquentes.

Propriété d'équivalence distributionnelle de la distance du χ^2 : agréger deux profils-lignes (resp. profils-colonnes) similaires ne change pas les distances entre profils-colonnes (resp. profils-lignes).

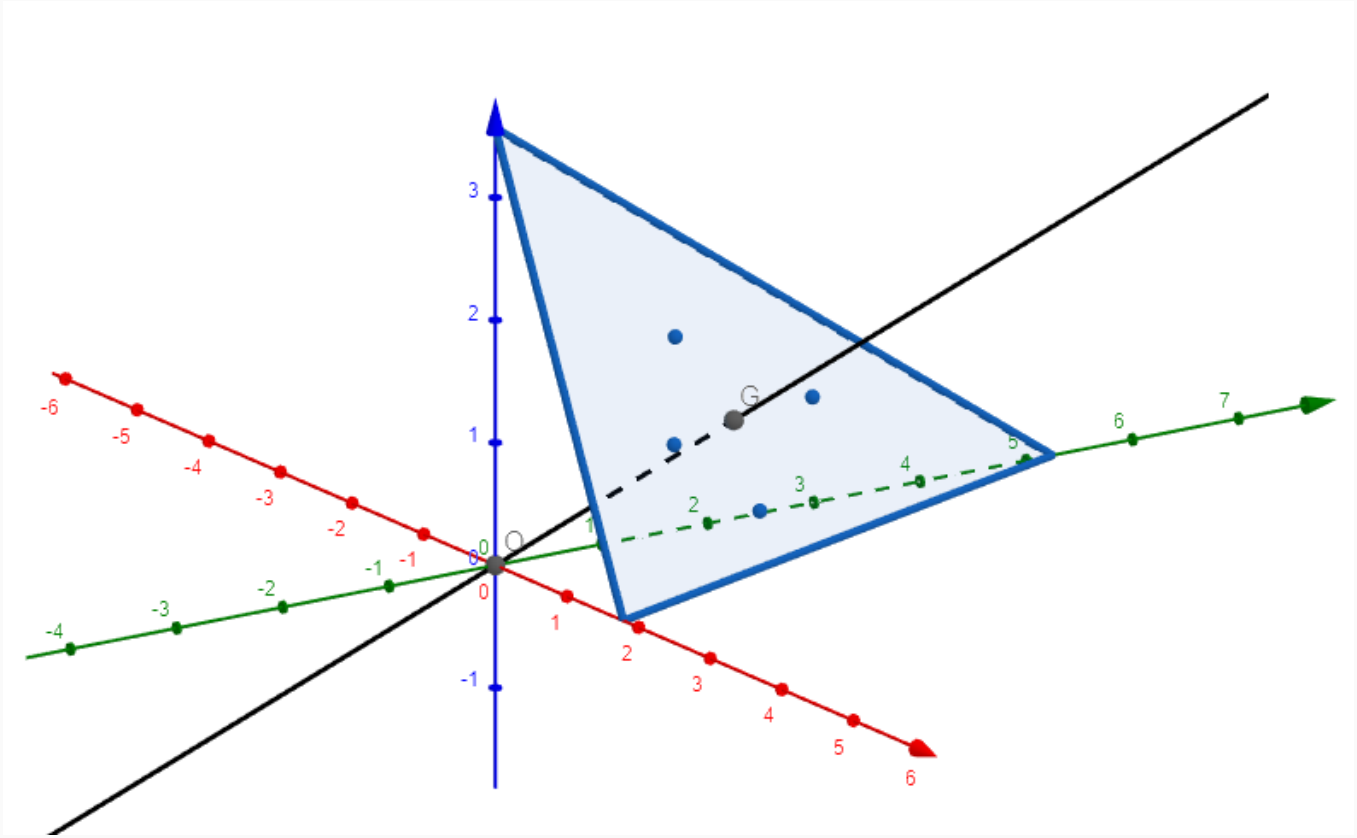
Exemple d'application : CSP, départements.

Remarque :

$$\forall i \in [1, J_1],\ d^2_{\chi^2}(i, G) = \sum_{j=1}^{J_2} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 = \sum_{j=1}^{J_2} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij} - f_{i\bullet} f}{f_{i\bullet}} \right)^2$$

On pose $D_{J_1} = \text{diag}(f_{1\bullet}, ..., f_{J_1\bullet})$ et $D_{J_2} = \text{diag}(f_{\bullet 1}, ..., f_{\bullet J_2})$.

On utilise toujours l'inertie comme critère $\Longrightarrow$ Analyse du triplet $(D_{J_1}^{-1} F, P = D_{J_1}, M = D_{J_2}^{-1})$



- Remarquons que la matrice $D_{J_1}^{-1}F$ n'est pas centrée. Pourquoi ?
- Les points du nuage des points appartiennent à un hyperplan affine.
- La matrice d'inertie est donnée par $V = {}^tFD_{J_1}^{-1}FD_{J_2}^{-1}$. On montre que $Sp(V) \subset [0, 1]$.
- 1 est toujours valeur propre, l'axe factoriel associé est appelé "axe trivial".
- Lorsqu'on centre $D_{J_1}^{-1}F$, cela revient à ne considérer que les $J_2 - 1$ valeurs propres non triviales.

Lien entre l'inertie par rapport à l'origine $I_0 = \sum_{i=1}^{J_1} f_{i\bullet} d^2_{\chi^2}(i, 0)$ et l'inertie par rapport au centre de gravité $I_G = I_0 = \sum_{i=1}^{J_1} f_{i\bullet} d^2_{\chi^2}(i, G)$: $I_0 = I_G + 1$.

En diagonalisant V, on obtient $J_2 - 1$ axes factoriels (en omettant l'axe trivial) auxquels on associe des valeurs propres inférieures à 1 : la projection sur les espaces engendrés par les axes factoriels associés aux plus grandes valeurs propres permet d'avoir une représentation en plus faible dimension du nuage des profils-lignes.

Comme en ACP, on utilise des critères de sélection d'axes :

- critère du coude,
- scree-test de Catell,
- critère de l'inertie moyenne

Les coordonnées des profils-lignes sur le k-ième axe factoriel sont données par le vecteur $C_k = D_{J_1}^{-1} F D_{J_2}^{-1} u_k$

Aide à l'interprétation pour les profil-lignes

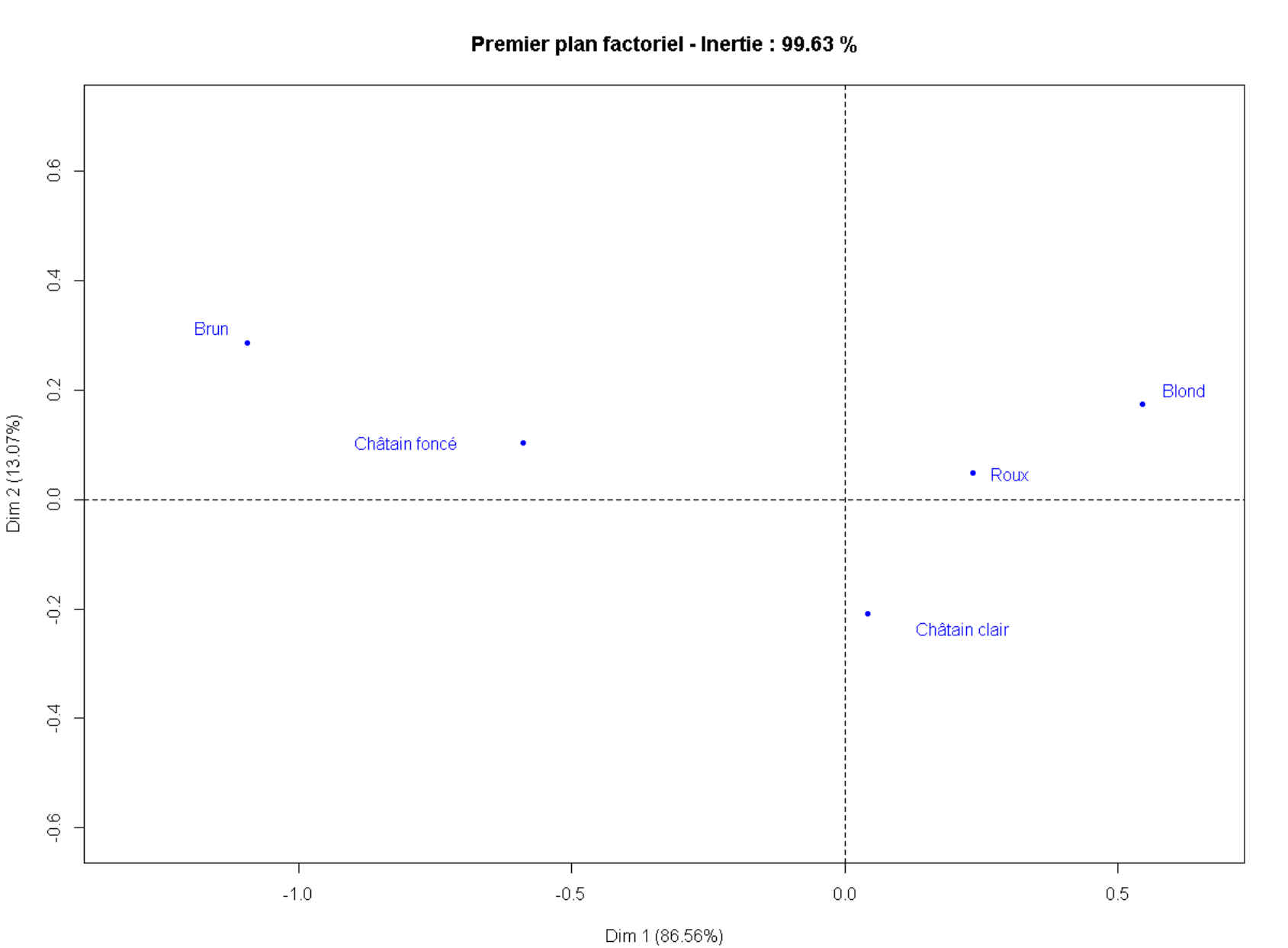
Certains profil-ligne peuvent être mal représentés sur un axe ou ils peuvent contribuer très fortement à la conception de l'axe : les aides à l'interprétation vont nous aider à les identifier.

- La coordonnée du profil-ligne i sur l'axe k est notée c_{ik}. Il s'agit de la i-ème composante du vecteur C_k.
- La contribution du profil-ligne i à l'inertie projetée sur l'axe k notée $\text{CTRB}_k(i)$ est telle que $\text{CTRB}_k(i) = \dfrac{f_{i\bullet} c_{ik}^2}{\lambda_k}$.

  Remarque : $\displaystyle\sum_{i=1}^{J_1} \text{CTRB}_k(i) = 1$
- La qualité de représentation du profil-ligne i sur l'axe k noté $\text{QLT}_k(i)$ est telle que $\text{QLT}_k(i) = \dfrac{c_{ik}^2}{||f_i||_M^2} = \cos^2(\theta_{f_i,u_i})$

  Remarque : $\displaystyle\sum_{k=1}^{J_2} \text{QLT}_k(i) = 1$

Premier plan factoriel



Aide à l'interprétation

	Axe 1	Axe 2	Axe 3
Blond	0.54	0.17	0.01
Roux	0.23	0.05	-0.12
Châtain clair	0.04	-0.21	0.00
Châtain foncé	-0.59	0.10	0.01
Brun	-1.09	0.29	-0.05

Table 6: Coordonnées

	Axe 1	Axe 2	Axe 3
Blond	40.12	27.13	4.93
Roux	1.45	0.41	86.09
Châtain clair	0.35	57.21	0.48
Châtain foncé	44.92	9.27	3.07
Brun	13.17	5.97	5.42

Table 7: Qualités de représentation

	Dim 1	Dim 2	Dim 3
Blond	0.91	0.09	0.00
Roux	0.77	0.03	0.20
Châtain clair	0.04	0.96	0.00
Châtain foncé	0.97	0.03	0.00
Brun	0.93	0.06	0.00

Table 8: Contributions

Définition profils-colonnes

On appelle matrice des profils-colonnes, la matrice suivante :

$$\begin{bmatrix} \frac{f_{11}}{f_{\bullet 1}} & \frac{f_{21}}{f_{\bullet 1}} & \dots & \frac{f_{(J_2-1)1}}{f_{\bullet 1}} & \frac{f_{J_2 1}}{f_{\bullet 1}} \\ \frac{f_{12}}{f_{\bullet 2}} & \frac{f_{22}}{f_{\bullet 2}} & \dots & \frac{f_{(J_2-1)2}}{f_{\bullet 2}} & \frac{f_{J_2 2}}{f_{\bullet 2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{f_{1(J_1-1)}}{f_{\bullet(J_1-1)}} & \frac{f_{2(J_1-1)}}{f_{\bullet(J_1-1)}} & \dots & \dots & \frac{f_{J_2(J_1-1)}}{f_{\bullet(J_1-1)}} \\ \frac{f_{1J_1}}{f_1} & \frac{f_{2J_1}}{f_1} & \dots & \dots & \frac{f_{J_2 J_1}}{f_{\bullet J_1}} \end{bmatrix}$$

Les lignes de la matrice des profils-colonnes sont les profils-colonnes.

La i-ième ligne de la matrice des profils lignes correspond à la distribution empirique de la variable Y_1 conditionnellement à $Y_2 = y_i$.

À chaque profil-colonne, on associe un point de $\mathbb{R}^{J_1}$ dont les coordonnées dans la base canonique correspondent à son profil-colonne.

On appelle nuage des profils-colonnes, le nuage de points formé par l'ensemble des profils-colonnes.

Pondération et distance entre profils-colonnes

Dans le cas de l'AFC :

- le poids d'un profil-colonnes p_j correspond à la fréquence d'apparition de la modalité correspondante :

$$p_j = f_{\bullet j}$$

Il en vient que le barycentre des profils-colonnes correspond au profil-colonne moyen (observé sur toute la population)
$f_{\text{moyen}\bullet} = (f_{1\bullet}, ..., f_{J_1\bullet})$

- la distance entre profils-colonnes est la distance du χ^2 :

$$\forall\ (j_1, j_2) \in [1, J_2]^2, \quad d^2_{\chi^2}(j_1, j_2) = \sum_{i=1}^{J_1} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij_1}}{f_{\bullet j_1}} - \frac{f_{ij_2}}{f_{\bullet j_2}} \right)^2$$

Le terme $\frac{1}{f_{i\bullet}}$ permet d'équilibrer l'influence des modalités trop fréquentes.

Analyse du nuage des profils-colonnes

L'analyse du nuage des profils-colonnes est obtenue en effectuant l'analyse du triplet $(D_{J_2}^{-1\,t}F, P = D_{J_2}, M = D_{J_1}^{-1})$.

Les valeurs propres λ_k sont les mêmes dans les deux analyses (au nombre de valeurs propres nulles près).

Les coordonnées des profils-colonnes sur le k-ième axe factoriel sont données par le vecteur $D_k = D_{J_2}^{-1\,t}FD_{J_1}^{-1}v_k$.

Il n'est pas nécessaire de faire les deux analyses : il existe des relations de transition :

- $u_k = \frac{1}{\sqrt{\lambda_k}}{}^tFD_{J_1}^{-1}v_k$
- $v_k = \frac{1}{\sqrt{\lambda_k}}FD_{J_2}^{-1}u_k$
- $C_k = \frac{1}{\sqrt{\lambda_k}}D_{J_1}^{-1}FD_k$
- $D_k = \frac{1}{\sqrt{\lambda_k}}D_{J_2}^{-1\,t}FC_k$

Aide à l'interprétation pour les profils-colonnes

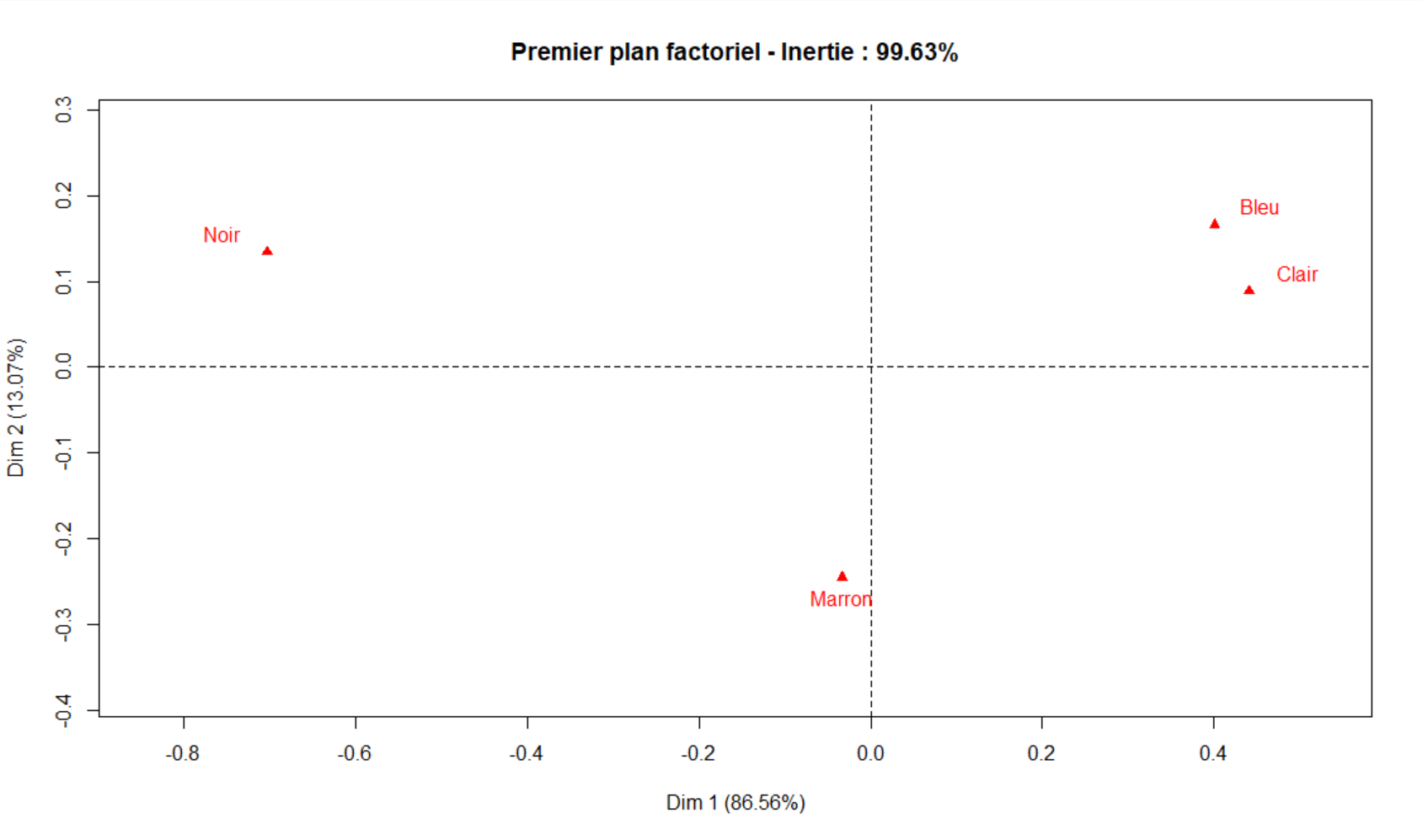
Certains profils-colonnes peuvent être mal représentés sur un axe ou ils peuvent contribuer très fortement à la conception de l'axe : les aides à l'interprétation vont nous aider à les identifier.

- La coordonnée du profil-colonne j sur l'axe k est notée d_{jk}. Il s'agit de la j-ème composante du vecteur D_k.
- La contribution du profil-colonne j à l'inertie projetée sur l'axe k notée $\text{CTRB}_k(j)$ est telle que $\text{CTRB}_k(j) = \dfrac{f_{\bullet j} d_{jk}^2}{\lambda_k}$.

  Remarque : $\displaystyle\sum_{j=1}^{J_2} \text{CTRB}_k(j) = 1$
- La qualité de représentation d'un profil-colonne j sur l'axe k noté $\text{QLT}_k(j)$ est telle que $\text{QLT}_k(j) = \dfrac{d_{ij}^2}{||f_j||_P^2} = \cos^2(\theta_{f_j, v_i})$

  Remarque : $\displaystyle\sum_{k=1}^{J_1} \text{QLT}_k(j) = 1$

Premier plan factoriel



Aide à l'interprétation

	Axe 1	Axe 2	Axe 3
Bleu	0.40	0.17	0.06
Clair	0.44	0.09	-0.03
Marron	-0.03	-0.25	0.01
Noir	-0.70	0.13	-0.00

Table 9: Coordonnées

	Axe 1	Axe 2	Axe 3
Bleu	0.84	0.14	0.02
Clair	0.96	0.04	0.00
Marron	0.02	0.98	0.00
Noir	0.96	0.04	0.00

Table 10: Qualités de représentation

	Axe 1	Axe 2	Axe 3
Bleu	10.72	12.12	63.83
Clair	28.59	7.63	34.45
Marron	0.19	65.70	1.18
Noir	60.50	14.55	0.54

Table 11: Contributions

On déduit des relations de transition deux relations appelées *relations barycentriques* :

$$c_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J_2} \frac{f_{ij}}{f_{i\bullet}} d_{jk}$$

$$d_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{J_1} \frac{f_{ij}}{f_{\bullet j}} c_{ik}$$

À un facteur multiplicatif près, chaque projection de profil-ligne est au barycentre des projections des profils-colonnes (pondérés par leurs distributions conditionnelles).

Exemple : la projection de la modalité "Roux" se trouve au barycentre des projections des profils-colonnes (couleurs des yeux) pondérés par la fréquence des couleurs des yeux sachant qu'on est roux.

Contraitement à l'ACP, les relations barycentriques permettent la représentation des projections des profils-lignes et des profils-colonnes sur les mêmes plans factoriels.

Représentation simultanée

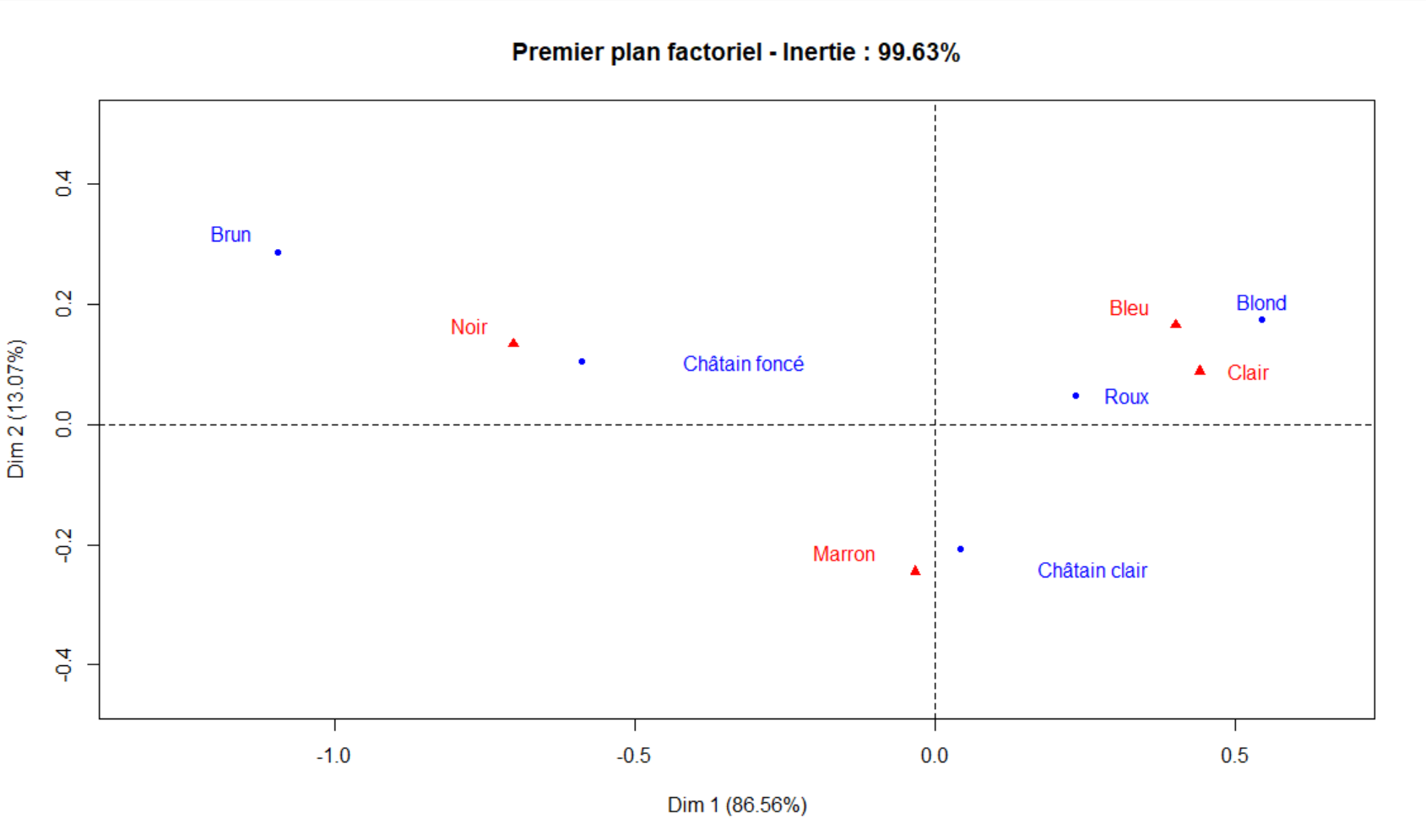


Figure 13: Représentation simultanée

Analyse des correspondances multiples

Problématique

L'ACM permet d'analyser des tableaux de données de type n individus x Q variables qualitatives.

L'ACM s'apparente à :

- l'ACP dans le type de données à analyser,
- l'AFC dans la méthode.

L'ACM et l'ACP sont des cas particuliers d'AFDM (cf chapitre 5).

Comme en ACP, l'ACM permet de mettre en lumière des liens :

- entre variables à l'aune des valeurs prises sur les variables qualitatives,
- entre individus.

Données brutes et notations

On considère un tableau de données n individus x Q variables qualitatives. On associe à ce tableau une matrice $Y \in M_{n,p}(\mathbb{N})$ où le terme général $y_{i,j}$ correspond à la valeur prise par l'individu i pour la variable j.

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p-1} & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p-1} & y_{2,p} \\ \cdots & \cdots & y_{i,j} & \cdots & \cdots \\ y_{n-1,1} & y_{n-1,2} & \cdots & y_{n-1,p-1} & y_{n-1,p} \\ y_{n,1} & y_{n,2} & \cdots & y_{n,p-1} & y_{n,p} \end{bmatrix}$$

Notations : si $y_{\bullet,j}$ est la j-ème variable et qu'elle est décrite par J_j modalités alors on associera un entier entre 1 et J_j à chaque modalité.

Tableau disjonctif complet ou TDC

On appelle tableau disjonctif complet associé à X qu'on notera TDC(X), la matrice suivante.

$$\text{TDC(X)} = \begin{pmatrix} x_{1,1,1} & \dots & x_{1,1,J_1} & x_{1,2,1} & \dots & x_{1,2,J_2} & \dots & x_{1,Q,1} & \dots & x_{1,Q,J_Q} \\ x_{2,1,1} & \dots & x_{2,1,J_1} & x_{2,2,1} & \dots & x_{2,2,J_2} & \dots & x_{2,Q,1} & \dots & x_{2,Q,J_Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & x_{i,j,k} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-1,1,1} & \dots & x_{n-1,1,J_1} & x_{n-1,2,1} & \dots & x_{n-1,2,J_2} & \dots & x_{n-1,Q,1} & \dots & x_{n-1,Q,J_Q} \\ x_{n,1,1} & \dots & x_{n,1,J_1} & x_{n,2,1} & \dots & x_{n,2,J_2} & \dots & x_{n,Q,1} & \dots & x_{n,Q,J_Q} \end{pmatrix}$$

avec $x_{i,j,k} = 1$ si l'individu i a la modalité j pour la k-ième variable.
$x_{i,j,k} = 0$ sinon.

Principe de l'ACM

L'analyse des correspondances multiples correspond à une analyse factorielle des correspondances du tableau disjonctif complet.

Introduction à d'autres méthodes d'analyse multivariée dérivées de l'ACP

References i