Analyse des données
Master Statistique et Économétrie
Notes de cours

V. Monbet

Master 1 - 2017

Contents

Chapter 1

Introduction

L'analyse statistique multivariée consiste à analyser et comprendre des données de grande dimension. Nous supposons que nous avons un ensemble $\{x_i\}_{i=1,\cdots,n}$ de n observations d'un vecteur de variables X dans $\mathbb{R}^p$. Autrement dit, nous supposons que chaque observation x_i admet p dimensions :

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$$

et que c'est une valeur observée (ou réalisation) d'un vecteur de variables $X \in \mathbb{R}^p$. Le vecteur X est composé de p variables aléatoires :

$$X = (X_1, X_2, ..., X_p)$$

où X_j, pour $j = 1, \cdots, p$, est une variable aléatoire de dimension 1. Comment allons nous analyser ce type de données? Avant de considérer la question de ce qu'on peut inférer à partir de ces données, on doit penser à comment regarder les données. Ceci implique des techniques descriptives. Les questions auxquelles nous pouvons répondre à l'aide d'analyses descriptives sont :

- Y a t'il certaines composantes de X qui sont plus dispersées que d'autres?
- Y a t'il des éléments de X qui indiquent des sous-groupes dans les données?
- Y a t'il des valeurs extrêmes et/ou aberrantes dans des données?
- La distribution des données est-elle "normale"?
- Y a t'il des combinaisons linéaires de faible dimension de X qui montrent des comportements "non-normaux"?

Une difficulté des méthodes descriptives pour les données de grande dimension est le système de perception humain. Les nuages de points en deux dimensions sont faciles à comprendre et à interpréter. Avec les techniques de visualisation interactives modernes on a la possibilité de voir des rotations 3D en temps réel et ainsi percevoir aussi les données à 3 dimensions. Une technique de glissement [1] décrite par Härdle et Scott (1992) permet de matérialiser une 4ème dimension en représentant des contours 3D avec la 4ème dimension en niveau de couleur.

Un saut qualitatif dans les difficultés de représentation apparaît pour des dimensions supérieures à 5, à moins que la structure de grande dimension ne puisse être projetée dans un espace

[1] sliding technic

de dimension plus faible. Certaines caratéristiques telles que des sous-groupes ou des valeurs aberrantes peuvent être détectées par des techniques d'analyses purement graphiques.

Dans le chapitre suivant, nous faisons quelques rappels importants d'algèbre linéaire. Dans le chapitre 3, nous introduisons l'analyse factorielle qui permet de projeter des données de grande dimension dans un espace de dimension plus faible. Nous en déduirons une technique classique : l'analyse en composantes principales. Dans le chapitre 4, nous étudierons un autre type d'analyse factorielle dont l'objectif est davantage un objectif de modélisation que de description : l'analyse en facteurs communs et spécifiques. Dans le chapitre 5, nous considérons un problème dans lequel on cherche des liens entre des variables (explicatives) continues et une variable (à expliquer) catégorielle et nous décrirons l'analyse factorielle discriminante. Puis dans le chapitre 6, nous nous intéresserons aux tableaux de données catégorielles et nous étudierons l'analyse des correspondances et l'analyse des correspondances multiples. En enfin dans le chapitre 7, nous nous tournons vers le problème de la classification non supervisée qui permet de mettre en évidence des sous groupes dans les données. Pour conclure, dans le chapitre 8, nous mettrons en évidence que tous les problèmes évoqués peuvent être formalisés comme des problèmes d'inférence sur une ou plusieurs variables latentes.

Une partie des exemples de ce cours sont empruntés à Härdle et Simar (2007).

Chapter 2

Rappels et compléments d'algèbre linéaire - Décompositions de matrices

2.1 Les projecteurs

La notion de projection est fondamentale en statistique. Par exemple la moyenne est une projection sur la droite des constantes. L'analyse en composante principale est basée sur des projecteurs de même que la régression linéaire.

2.1.1 Sous espaces supplémentaires et projecteurs

Soient F et G deux sous espaces vectoriels de E.

$$F + G = \{x + y | x \in F, y \in G\} \text{ et } F \times G = \{(x, y) | x \in F, y \in G\}.$$

Définition 1. *On dit que F et G sont supplémentaires si $F \cap G = \emptyset$ et $F + G = E$.*
De façon équivalente, tout vecteur x de E s'écrit de manière unique $x = u + v$ avec $u \in F$ et $v \in G$.

Le supplémentaire d'un sous espace vectoriel n'est pas unique.

Proposition 1. *Si F et G sont supplémentaires, les applications p et q de E dans E définies par*

$$\forall x \in E, x = p(x) + q(x) \text{ avec } p(x) \in F \text{ et } q(x) \in G$$

sont linéaires (on dit que ce sont des endomorphismes de E) et vérifient

[P1] $p^2 = p$; $q^2 = q$ (idempotence)

[P2] $poq = qop = 0$

[P3] $p + q = Id_E$

[P4] $Im(p) = F = Ker(q)$ et $Im(q) = F = Ker(p)$

On dit que p est la projection sur F parallèlement à G et que $q = Id_E - p$ est la projection sur G parallèlement à F.
On appelle projecteur dans un espace vectoriel E tout endomorphisme idempotent de E.

Dans le cas particulier où les deux sous espaces supplémentaires sont orthogonaux $E = F \oplus F^{\perp}$ alors les projecteurs p et q associées sont dits projecteurs orthogonaux.

2.1.2 Exemple fondamental

Soient u et v de $\mathbb{R}^n$ muni du produit scalaire usuel, tels que

$$\langle u, v\rangle = v^T u = 1$$

Remarquons que puisque $\langle u, v\rangle = ||u||_2||v||_2 \cos(u, v)$, la condition précédente impose que l'angle vectoriel entre u et v est aigu. Considérons la matrice $n \times n$,

$$P = uv^T.$$

Cette matrice jouit des propriétés suivantes :

$$P^2 = uv^T uv^T = uv^T = P$$

et si $x \in Imu$, c'est à dire si $x = \alpha u$,

$$Px = uv^T(\alpha u) = \alpha uv^T u = \alpha u = x$$

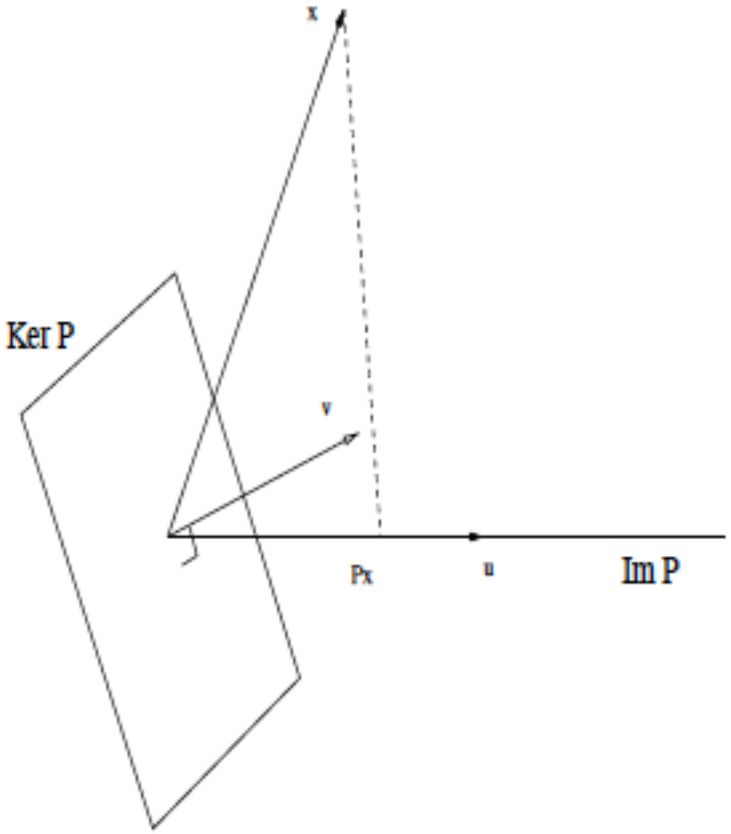
Mais si x est orthogonal à v, alors

$$Px = uv^T x = u(v^T x) = 0$$

L'image de P est donc Imu, le noyau de P est le sous espace vectoriel de dimension $n - 1$ orthogonal à v.

$$\mathbb{R}^n = Imu \oplus (Imv)^{\perp}.$$

P est donc la matrice de l'application linéaire "projection sur u parallèlement à $Imv^{\perp}$".



Si on choisit $v = u$ et $||u||_2 = 1$, le projecteur orthogonal s'écrit

$$P = uu^T.$$

De façon plus générale, soit F donné ainsi qu'une base $\{u_1, \cdots, u_r\}$ orthonormée de F. Soit $U = [u_1, \cdots, u_r]$ alors $U^T U = I_r$. La matrice

$$P = \sum_{i=1}^{r} u_i u_i^T = UU^T$$

est le projecteur orthogonal sur $F = ImU$. Le projecteur ($P^2 = P$) est orthogonal car $P = P^T$. En effet, la projection orthogonale est telle que pour tout vecteur quelconque Y de E, on cherche $\hat{Y} \in F$ tel que

$$(Y - \hat{Y}) \perp F$$

c'est à dire

$$\forall i, \quad u_i^T (Y - \hat{Y}) = 0$$
$$U^T (Y - \hat{Y}) = 0$$

D'oú

$$U^T Y = U^T \hat{Y}$$

Ceci signifie aussi que $\hat{Y}$ s'écrit comme une combinaison linéaire des éléments u_i,

$$\hat{Y} = c_1 u_1 + \cdots + c_k u_k = U \begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_k \end{bmatrix}$$

On a donc

$$U^T Y = U^T U C$$

ce qui s'écrit aussi

$$(U^T U)^{-1} U^T Y = C$$

Et on remarque que $\hat{Y} = UC = U(U^T U)^{-1} U^T Y$ et on obtient que la matrice de projection est $U(U^T U)^{-1} U^T$.

Remarque - On reconnait les formules du modèle de régression linéaire pour des variables centrées (en prenant $U = X$, on a bien $C = var(X)^{-1} cov(X, Y)$).

Exercice Calculer la matrice de projection Q sur le sous espace de $\mathbb{R}^4$ engendré par les vecteurs $(1, 1, 0, 2)$ et $(-1, 0, 0, 1)$. Donner la projection de $x = (0, 2, 5, -1)$ sur le sous espace.

Exercice Calculer la projection de $v = (1, 1, 0)$ sur le plan $x + y - z = 0$.

2.2 Matrices carrées diagonalisables

Définition 2. *Une matrice carrée A d'ordre n est diagonalisable si elle est semblable à une matrice diagonale $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ ie qu'il existe une matrice inversible S telle que*

$$\Lambda = S^{-1} A S$$

La $i$ème colonne de S est le vecteur propre de A associé à la valeur propre λ_i.

Condition nécessaire et suffisante - Une condition nécessaire et suffisante pour que A carrée d'ordre n, soit diagonalisable est que ses n vecteurs propres soient linéairement indépendants.
Condition suffisante : Les vecteurs propres associés à des valeurs propres distinctes sont linéairement indépendants. Si toutes les valeurs propres de A sont distinctes, alors A est diagonalisable.

Décomposition spectrale de A diagonalisable
Soit A diagonalisable telle que $A = S\Lambda S^{-1}$. Notons u_j la $j$ème colonne de S et v_j^T la $j$ème ligne de S^{-1}, associés à λ_j. La décomposition spectrale de A s'écrit

$$A = \sum_{j=1}^{n} \lambda_j u_j v_j^T$$

Le vecteur v_j est le vecteur propre de A^T associé à $\bar{\lambda}_j$ et $v_j^T u_i = 0$ si $j \neq i$. Ceci signifie que les vecteurs propres distincts de A et A^T sont orthogonaux.

- Une matrice symétrique et réelle est diagonalisable et on a $A = S\Lambda S^T$.
- Une matrice symétrique et réelle est (semi) définie positive si et seulement si toutes ses valeurs propres sont positives (non négatives).

2.3 Décomposition en valeurs singulières

Pour une matrice rectangulaire, la notion de valeur propre n'a pas de sens. Néanmoins, les matrices carrées A^TA et AA^T sont symétriques semi définies positives. De plus,

$$rang(A) = rang(AA^T) = rang(A^TA) = r$$

et les r valeurs propres non nulles (positives) de A^TA et AA^T sont identiques.

Définition 3. *On appelle valeurs singulières de A les racines carrées des valeurs propres non nulles et A^TA ou de AA^T.*

$$\mu_i = \sqrt{\lambda_i(A^TA)} = \sqrt{\lambda_i(AA^T)}$$

Soit A $m \times n$ telle que $rang(A) = r$. Alors

$$A = U\Lambda_r^{1/2}V^T = \sum_{j=1}^{r} \mu_j u_j v_j^T$$

avec

- $U = [u_1, \cdots, u_r]$ unitaire $m \times r$ est telle que u_j est le vecteur propre de AA^T associé à la valeur propre non nulle λ_i.
- $V = [v_1, \cdots, v_r]$ unitaire $n \times r$ est telle que v_j est le vecteur propre de A^TA associé à la valeur propre non nulle λ_i.
- $\Lambda_r = diag(\lambda_1, \cdots, \lambda_r)$ et $\Lambda_r^{1/2} = diag(\mu_1, \cdots, \mu_r)$ où $\mu_j = \lambda_j^{1/2}$ est la $j$ème valeur singulière de A.

Remarques/résultats importants -

- Dans la pratique, le nombre de valeurs singulières non nulles fourni le rang de la matrice.
- Dans le calcul de U et de V , on ne calcule les vecteurs propres de AA^T ou de A^TA que pour celle de ces matrices de plus petite dimension, les vecteurs propres de l'autre se déduisent par des "formules de transition" (2.1) et (2.2).

$$U = AV\Lambda_r^{-1/2} \tag{2.1}$$

avec $\Lambda_r^{-1/2} = (\Lambda_r^{1/2})^{-1} = diag(1/\mu_1, \cdots, 1/\mu_r)$

$$V = A^TU\Lambda_r^{-1/2} \tag{2.2}$$

- La décomposition en valeurs singulières donne

$$A^TA = V\Lambda_rV^T = \sum_{i=1}^{r} \mu_i^2 v_i v_i^T$$

$$AA^T = U\Lambda_rU^T = \sum_{i=1}^{r} \mu_i^2 u_i u_i^T$$

- Il y a d'importantes projections orthogonales associées à la décomposition en valeurs singulières. Soit A supposée de rang r et $A = U\Lambda_r^{1/2}V^T = P\Lambda Q^T$, la SVD de A. Rappelons que les partitions des colonnes de P et Q

$$P = [U|\tilde{U}], \quad Q = [V|\tilde{V}]$$

avec U les r premières colonnes de P et $\tilde{U}$ les suivantes (idem pour V et Q)

  - VV^T = projection orthogonale sur $\{KerA\}^\perp = ImA^T$
  - $\tilde{V}\tilde{V}^T$ = projection orthogonale sur $KerA$
  - UU^T = projection orthogonale sur ImA
  - $\tilde{U}\tilde{U}^T$ = projection orthogonale sur $\{ImA\}^\perp = KerA^T$

- On peut montrer que l'approximation d'une matrice A de rang p par une matrice de B de rang $q < p$ est donnée par la décomposition en valeurs singulière $B = U\tilde{\Lambda}V^T$ avec $\tilde{\Lambda}$ une matrice diagonale qui contient les q plus grandes valeurs singulières de A.

2.4 Les projecteurs M-orthogonaux

En statistique on est souvent amené à définir des produits scalaires différents du produit scalaire usuel et basés sur des métriques M, où M est une matrice symétrique définie positive, différentes de l'identité.

$$\langle x, y\rangle_M = y^TMx \text{ et } ||x||_M = x^TMx.$$

Définition 4. *Soit l'espace vectoriel Euclidien $\mathbb{E} = \mathbb{R}^m$ mini d'un M produit scalaire et soit $\mathbb{E}_1$ un sous espace vectoriel de $\mathbb{E}$ tel que $\mathbb{E} = \mathbb{E}_1 \oplus \mathbb{E}_1^\perp$ où $\mathbb{E}_1^\perp = \{y \in \mathbb{E} | \langle y, x\rangle_M = 0, x \in \mathbb{E}_1\}$. Pour tout x de $\mathbb{E}$ la décomposition*

$$x = x_1 + y_1, \ x \in \mathbb{E}_1, \ y \in \mathbb{E}_1^\perp$$

est unique. P est un projecteur M-orthogonal sur $\mathbb{E}_1$ si et seulement si

$$Px = x_1 (I - P)x = y_1$$

La notion de M-orthogonalité est liée à une notion de symétrie particulière, la M-symétrie. La symétrie usuelle correspond au cas où M est l'identité.

Définition 5. *Une matrice $A \in \mathbb{R}^{m\times m}$ est M-symétrique si*

$$MA = A^T M$$

c'est à dire que MA est symétrique.

Proposition 2. *Un projecteur P est un projecteur M-orthogonal si et seulement si P est M-symétrique.*

Preuve : Soit P un projecteur ($P^2 = P$) sur $\mathbb{E}_1$ tel que

$$\forall x, y \in \mathbb{E},\ Px \in \mathbb{E}_1,\ (I-P)y \in \mathbb{E}_1^{\perp} \text{ au sens de } M.$$

c'est à dire que $x^T P^T M(I-P)y = 0 \equiv P^T M(I-P) = 0 \equiv P^T M = (P^2)^T M = P^T MP$.
Puisque M est symétrique, $P^T M$ est aussi symétrique, $P^T M = MP$. $\diamond$

Chapter 3

Analyse en Composantes Principales

3.1 Introduction

L'objectif de ce chapitre est d'étudier les méthodes classiquement utilisées pour décrire et visualiser des données multivariées issues de variables continues : l'analyse en composantes principales et le positionnement multidimensionnel. Ces techniques d'analyse descriptive seront utilisées, notamment, pour visualiser les données dans un sous espace représentatif, pour détecter des groupes d'individus et/ou de variables, des valeurs extrêmes ou abbérantes ou pour aider au choix de variables. Ces méthodes permettent aussi de répondre à des questions du type : quels individus se ressemblent du point de vue de l'ensemble des variables? ou réciproquement quelles variables sont semblables du point de vue de l'ensemble des individus?

L'analyse en composantes principales est un outil de réduction de dimension qui permet de retirer la redondance ou la duplicité dans un ensemble de variables corrélées. L'ensemble initial est alors représenté par un ensemble réduit de variables dérivées des variables observées. Ces facteurs sont, en théorie, indépendants les uns des autres et on peut les classer par ordre d'importance.

Soit $\{X_1, \cdots, X_p\}$ un ensemble de p variables observées sur n individus indépendants. On notera

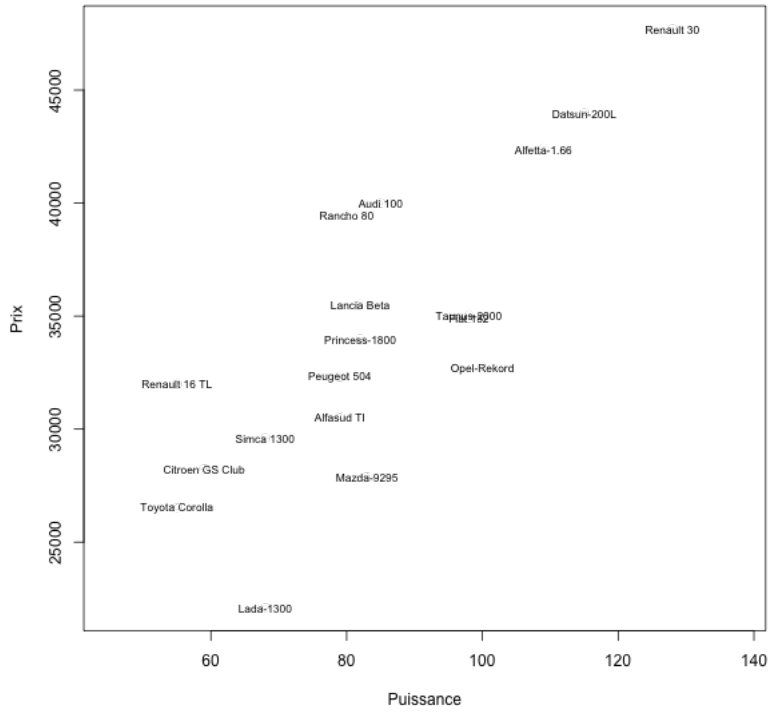
$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & & \cdots \\ x_{np} & \cdots & x_{np} \end{pmatrix}$$

le tableau des n observations des p variables. Typiquement, n est grand devant p. Pour tout $j \in \{1, \cdots, p\}$, $x_j \in \mathbb{R}^n$. Chaque ligne du tableau représente un individu et chaque colonne une variable. Chaque individu est un point de l'espace $\mathbb{R}^p$. On dira que $\mathbb{R}^p$ est l'espace des variables et $\mathbb{R}^n$ l'espace des individus.

Par exemple, dans le tableau ci-dessous la première colonne (Modèle) est l'identifiant et on observe deux variables, la puissance de la voiture et son prix.

Modèle	Puissance	Prix
Alfasud TI	79	30570
Audi 100	85	39990
Simca 1300	68	29600
Citroen GS Club	59	28250
Fiat 132	98	34900
Lancia Beta	82	35480
Peugeot 504	79	32300
Renault 16 TL	55	32000
Renault 30	128	47700
Toyota Corolla	55	26540
Alfetta-1.66	109	42395
Princess-1800	82	33990
Datsun-200L	115	43980
Taunus-2000	98	35010
Rancho	80	39450
Mazda-9295	83	27900
Opel-Rekord	100	32700
Lada-1300	68	22100

Quand on n'observe que deux variables, la représentation des individus est directe : on représente les individus dans le plan de $\mathbb{R}^2$, chaque axe représentant une variable (voir Figure **??**).



L'objectif de l'analyse en composantes principales (ACP) est de représenter les individus quand $p > 2$. L'idée est la suivante. Supposons dans un premier temps que les individus soient en fait concentrés dans un plan de $\mathbb{R}^p$. La solution la plus simple consiste à faire un changement

de base où les deux premiers axes sont dans le plan et les autres leurs sont orthogonaux (et les coordonnées des individus sur les axes 3 à p axes seront nulles). Considérons maintenant un nuage de points qui est presque concentré sur un plan. En pratique, on cherche un sous espace vectoriel dans lequel la dispersion entre les observations est la mieux représentée. On cherche aussi à préserver au mieux les distances entre les individus. On peut faire l'analogie avec une photographie. Si on photographie un objet en 3 dimensions (par exemple un poisson), on va chercher un plan tel qu'on reconnaisse aisément que c'est un poisson, c'est à dire un plan dans lequel les informations importantes sont restituées au mieux. Dans la figure 3.1 l'image représentant le poisson de profil (plus grande dispersion) restitue davantage d'information que celle du poisson de face (moins de dispersion).

Figure 3.1: Le poisson clown.

D'un point de vue plus "mathématique", l'ACP correspond à l'approximation d'une matrice $n \times p$ par une matrice de même dimension mais de rang $q < p$, q étant souvent de petite valeur 2, 3 pour la construction de graphiques facilement compréhensibles. Plus précisément, les objectifs poursuivis par l'ACP sont

- la représentation graphique "optimale" des individus en minimisant les déformations du nuage des points, dans un sous espace de dimension $q < p$ (autrement dit on cherche à préserver les distances entre individus) ;
- la représentation graphique des variables dans un sous espace E_q en explicitant "au mieux" les liaisons entre ces variables ;
- la réduction de la dimension (compression), ou approximation du tableau de données X par une matrice de rang $q < p$.

On peut construire l'ACP de plusieurs façons. L'approche la plus classique (en France) est l'approche géométrique.

3.2 ACP par projection : approche géométrique

En ACP, on travaille toujours sur les données centrées. Notons $\tilde{x}_i$ et $\tilde{\mathbf{x}}$ les individus centrés :

$$\tilde{x}_i = x_i - \bar{x}_i, \tilde{\mathbf{x}} = \begin{pmatrix} x_1 - \bar{x}_1 \\ \dots \\ x_n - \bar{x}_n \end{pmatrix}$$

La moyenne empirique $\bar{x}$ est parfois appelée centre de gravité.

La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage par rapport à son barycentre se mesure par l'inertie.

Définition 6. *L'inertie des individus est donnée par la quantité*

$$I = \frac{1}{n}\sum_{i=1}^{n} ||\tilde{x}_i||^2$$

On remarque que l'*inertie* est définie comme la somme des distances au carré des points à leur centre de gravité. Dans le cas où les variables dont quantitatives, c'est aussi la somme des variances empiriques de chacune des variables, c'est à dire la trace de la matrice de variance-covariance empirique $\hat{\Sigma}$. En effet,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^T \tilde{x}_i = \frac{1}{n}\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}, \hat{\Sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ij}\tilde{x}_{ik}$$

L'inertie est une quantité réelle qui mesure la dispersion des individus dans l'espace à p dimensions.

Soit P un projecteur de $\mathbb{R}^p$. Par abus, on notera également P la matrice associée à P dans la base canonique. La projection d'un vecteur x_i sera

$$P(x_i) = x_i P^T, \quad XP^T = \begin{pmatrix} x_1 P^T \\ \dots \\ x_n P^T \end{pmatrix}$$

Soit E un sous-espace de $\mathbb{R}^p$ et P_E le projecteur orthogonal sur E, on note I_E l'inertie des individus projetés :

$$I_E = \frac{1}{n}\sum_{i=1}^{n} ||P_E(\tilde{x}_i)||^2 = \frac{1}{n}\sum_{i=1}^{n} ||\tilde{x}_i||^2 - \frac{1}{n}\sum_{i=1}^{n} ||\tilde{x}_i - P_E(\tilde{x}_i)||^2$$

par Pythagore. L'inertie I_E est donc également une mesure de la dispersion des individus après projection sur E. Il est facile de vérifier que

$$I_E = Tr(P_E \hat{\Sigma} P_E)$$

Soit $u_1, \cdots, u_q$ une base orthogonale de E. Alors $P_E = UU^T$ où U est la matrice rectangulaire formée des vecteurs U_i en colonne : $U = [u_1, \cdots, u_q]$. Donc, la trace étant invariante pas changement de base,

$$I_E = Tr(P_E \hat{\Sigma} P_E) = Tr(U^T \hat{\Sigma} U) = \sum_{i=1}^{q} u_i^T \hat{\Sigma} u_i$$

.

Raisonnons dans un premier temps avec un seul axe de projection u_1, ie $q = 1$. La projection d'un individu observé $\tilde{x}_i \in \mathbb{R}^p$ sur l'axe u est définie par

$$P_u(x_i) = x_i^T \frac{u}{||u_1||}$$

Et on cherche l'axe u^* qui conduit à la projection qui conserve au mieux les distances entre individus :

$$u^* = \min_{u\in\mathbb{R}^p, ||u||=1} \sum_{i=1}^{n} ||\tilde{x}_i - P_u(\tilde{x}_i)||^2 \tag{3.1}$$

avec $\tilde{\mathbf{x}}$ le nuage de points centré (et éventuellement réduit) et $\tilde{x}_i$ le $i$ème individu correspondant. Par le théorème de Pythagore, on sait que $||\tilde{x}_i - P_u(\tilde{x}_i)||^2 = ||\tilde{x}_i||^2 - ||P_u(\tilde{x}_i)||^2$, ainsi le problème de l'équation (3.1) est équivalent à

$$u^* = \max_{u\in\mathbb{R}^p, ||u||=1} \sum_{i=1}^{n} ||P_u(\tilde{x}_i)||^2 \tag{3.2}$$

soit encore en utilisant la définition de l'opérateur de projection :

$$u^* = \max_{u\in\mathbb{R}^p, ||u||=1} u^T \tilde{\mathbf{x}}^T \tilde{\mathbf{x}} u$$

On remarque que la variance empirique de $P_u(\tilde{\mathbf{x}})$ vaut

$$\frac{1}{n} P_u(\tilde{\mathbf{x}})^T P_u(\tilde{\mathbf{x}}) = u^T \cdot \underbrace{\frac{1}{n} \tilde{\mathbf{x}}^T \tilde{\mathbf{x}}}_{\hat{\Sigma}} \cdot u$$

où $\hat{\Sigma}$ est la matrice de covariance empirique de $\tilde{\mathbf{x}}$. Ainsi, pour le premier vecteur propre, on cherche un vecteur unitaire u^* tel que

$$u^* = arg \max_{\{u\in\mathbb{R}^n, u^T u=1\}} u^T \hat{\Sigma} u \tag{3.3}$$

Nous cherchons donc le vecteur u^* tel que la projection du nuage sur u^* ait une inertie (ou une variance) maximale. En introduisant les multiplicateurs de Lagrange pour s'affranchir de la contrainte dans le problème de maximisation, (3.3) est équivalent à

$$(u^*, \lambda) = arg \max_{\{u\in\mathbb{R}^n, \lambda\in\mathbb{R}\}} u^T \hat{\Sigma} u - \lambda(u^T u - 1)$$

La solution est la racine de la dérivée de l'expression ci-dessous.

$$\frac{\partial \left(u^T \hat{\Sigma} u - \lambda(u^T u - 1) \right)}{\partial u} = 2(\Sigma u - \lambda u)$$

$$\frac{\partial \left(u^T \hat{\Sigma} u - \lambda(u^T u - 1) \right)}{\partial \lambda} = u^T u - 1$$

Si on remarque maintenant que

$$\max_{\{u\in\mathbb{R}^n, u^T u=1\}} u^T \hat{\Sigma} u = \max_{\{u\in\mathbb{R}^n, u^T u=1\}} u^T \lambda u = \max_{\{u\in\mathbb{R}^n, u^T u=1\}} \lambda$$

on a que le premier axe factoriel u^* est associé à la plus grande valeur propre de $\hat{\Sigma} u$.

Plus généralement, la maximisation de l'inertie I_E sur toutes les familles de q vecteurs orthogonaux est réalisée en choisissant les q vecteurs associés aux q plus grandes valeurs propres de $\hat{\Sigma}$ et on a les théorèmes suivants.

Théorème 1. *L'espace de dimension q d'inertie maximale est engendré par les q vecteurs propres associés aux q plus grandes valeurs propres de la matrice de variance-covariance des données (si des valeurs propres sont égales il n'y a pas unicité).*

Théorème 2. *Les composantes principales sont données par la transformation linéaire* $Y = U^T(X - E(X))$ *où* $\hat{\Sigma} = Var(X) = U\Lambda U^T$. *De plus on a :*

$$E(Y_j) = 0,\ \forall j = 1, \cdots, p$$

$$Var(Y_j) = \lambda_j,\ \forall j = 1, \cdots, p$$

$$Cov(Y_j, Y_k) = 0,\ \forall j, k = 1, \cdots, p$$

3.3 Représentations graphiques et aide à l'interprétation

L'analyse en composantes principales est principalement utilisée pour donner une représentation graphique des individus et des variables.

3.3.1 Les individus

En pratique, on projette orthogonalement les observations $\tilde{\mathbf{x}}$ sur les plans factoriels. Les coordonnées de $\mathbf{x}_i - \bar{\mathbf{x}}$ sur le sous espace de dimension q sont les q premiers éléments de la matrice $C = U\Lambda^{1/2}$. Voir l'exemple ci-dessous. Les graphiques obtenus permettent de représenter au mieux les distances euclidiennes inter-individus.

La qualité globale des représentations est mesurée par la *part de dispersion expliquée* ou la *portion d'inertie expliquée* :

$$r_Q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

Tandis que la qualité de la représentation de chaque point est donnée par

$$cs_i^2 = \frac{\sum_{k=1}^{q} d(O, y_i)_k^2}{\sum_{j=1}^{p} d(O, y_i)_k^2}$$

où $d(O, y_i)_k = c_{iK}$ et O représente le centre de gravité du nuage de point.

La contribution de chaque individu à l'inertie du nuage permet de détecter les observations les plus influentes et éventuellement aberrantes.

$$\gamma_i = \frac{\sum_{j=1}^{p} c_{ij}^2}{\sum_{j=1}^{p} \lambda_j}$$

Si la contribution d'un individu à un ou plusieurs axes est beaucoup plus importante que celle des autres il faut vérifier si cet individu n'est pas abberrant.

On peut projeter des individus supplémentaires $\mathbf{s}$ sur un sous espace factoriel en calculant ses coordonnées :

$$U^T(\mathbf{s} - \bar{\mathbf{x}})$$

Ici U joue le rôle d'une matrice de changement de base.

3.3.2 Les variables

La projection des variables sur les plans factoriels peuvent aider à l'interprétation des composantes. Cette représentation des variables peut s'interpréter comme le positionnement, pour chaque variable, d'un individu type, pour lequel les autres variables auraient leur valeur moyenne et la variable considérée serait amplifiée. Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables et, si celles ci ne sont pas réduites, leurs variances. On obtient le cercle des corrélations par projection orthogonale sur le sous espace factoriel E_q. La coordonnée de la variable x_j sur u_k est donnée par

$$\sqrt{\lambda} u_{jk}$$

La qualité de la représentation de chaque x_j est mesurée par

$$\frac{\sum_{j=1}^{q} \lambda_j v_{jk}^2}{\sum_{j=1}^{p} \lambda_j v_{jk}^2}$$

3.4 Exemple

A titre d'exemple, on considère un jeu de données établissant la composition du lait de 25 espèces de mammifères. On mesure 5 variables : la teneur en protéines, en lactose, en graisse, en eau et en minéraux. On obtient pour les matrices U et Λ suivantes.

$$U = \begin{pmatrix} 0.76 & -0.16 & -0.57 & -0.25 & -0.01 \\ -0.16 & 0.85 & -0.27 & -0.39 & -0.14 \\ -0.62 & -0.44 & -0.55 & -0.34 & 0.01 \\ 0.09 & -0.18 & 0.54 & -0.82 & 0.04 \\ -0.01 & 0.13 & -0.06 & -0.02 & 0.99 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 282.1 & 0 & 0 & 0 & 0 \\ 0 & 8.1 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{pmatrix}$$

En première approximation, on peut dire que les composantes principales correspondent dans l'ordre à

- la proportion d'eau sur la proportion de graisse
- la teneur en protéines
- la proportion de lactose sur celle d'eau et de graisse
- la teneur en lactose
- la teneur en sel minéraux

Le fait que la première valeur propre soit grande devant les autres signifie que les individus se démarquent surtout par la proportion d'eau par rapport à la graisse dans leur lait. La figure 3.2 montre, dans le premier plan factoriel, les graphes des variables et des individus pour l'ACP non réduite. Ce plan explique 99.5% de la variance. On observe sur le graphe de projection des variables que l'eau est le composant le plus important suivi par la matière grasse dans le composant du lait. Associé au graphe des individus, on peut voir, par exemple, que le dauphin et le phoque ont des laits plus gras que les autres mammifères. Le graphe permet de visualiser que les variables qui contribuent fortement au premier axe factoriel sont la matière grasse et l'eau. Le deuxième axe factoriel apporte peu d'information supplémentaire ; ce sont essentiellement les protéines qui contribuent à cet axe.

Le plus souvent, il est préférable d'interpréter une ACP réduite dans laquelle chaque variable va avoir la même contribution. Le résultat est alors indépendant des unités utilisées. Dans le cas de l'exemple les différents composants du lait sont mesurées dans les même unités. On peut alors préférer ne pas normaliser car les grandeurs relatives des variables sont importantes.

Si on normalise les données, on obtient pour des matrices U et D analogues à celles du cas non normalisé,

$$U = \begin{pmatrix} 0.47 & 0.35 & 0.37 & 0.11 & 0.71 \\ -0.47 & 0.32 & 0.15 & -0.79 & 0.19 \\ -0.45 & -0.48 & -0.31 & 0.18 & 0.67 \\ 0.48 & 0.06 & -0.78 & -0.38 & 0.11 \\ -0.35 & 0.74 & -0.38 & 0.43 & -0.00 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 3.88 & 0 & 0 & 0 & 0 \\ 0 & 0.89 & 0 & 0 & 0 \\ 0 & 0 & 0.13 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 \\ 0 & 0 & 0 & 0 & 0.01 \end{pmatrix}$$

La figure 3.3 montre, dans le premier plan factoriel, les graphes des variables et des individus pour l'ACP réduite. Le graphe des variables est aussi appelé cercle des corrélations. Le premier plan factoriel restitue 95.3% de la variance. Le cercle des corrélations permet de dire que les laits à forte teneur en matière grasse ou protéines sont généralement à faible teneur en lactose et eau car ces variables sont opposées sur le graphe. Ce sont ces variables qui contribuent au premier axe factoriel. Le second axe oppose les laits riches en protéines et minéraux aux laits riches en matières grasses. On remarque à l'aide du graphe des individus que les animaux qui ont un lait riche en eau sont surtout des animaux de régions chaudes.

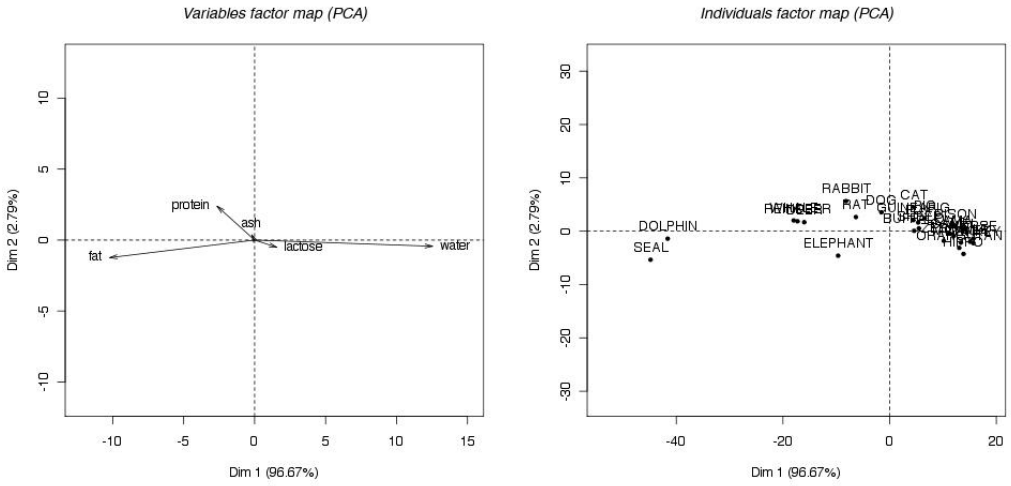


Figure 3.2: Composition du lait - Cercle des corrélations (à gauche) et graphe des individus (à droite) sur le premier plan principal de l'ACP non réduite

3.5 Propriétés asymptotiques des estimateurs de composantes principales

En pratique l'ACP est réalisée à partir de données. On manipule donc des estimateurs. Il est utile de connaitre leurs propriétés.

Théorème 3. *Soit $\Sigma > 0$ ayant des valeurs propres distinctes et soit $\hat{\Sigma} \sim n^{-1} W_p(\Sigma, n)$ tels que $\Sigma = \Gamma \Lambda \Gamma^T$ et $\hat{\Sigma} = \hat{\Gamma} \hat{\Lambda} \hat{\Gamma}$. Alors*

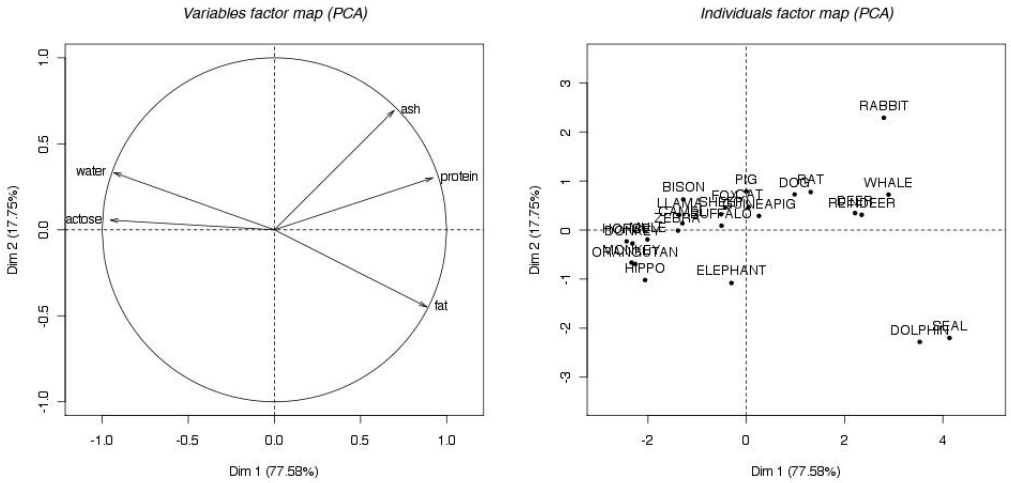


Figure 3.3: Composition du lait - Cercle des corrélations (à gauche) et graphe des individus (à droite) sur le premier plan principal de l'ACP réduite

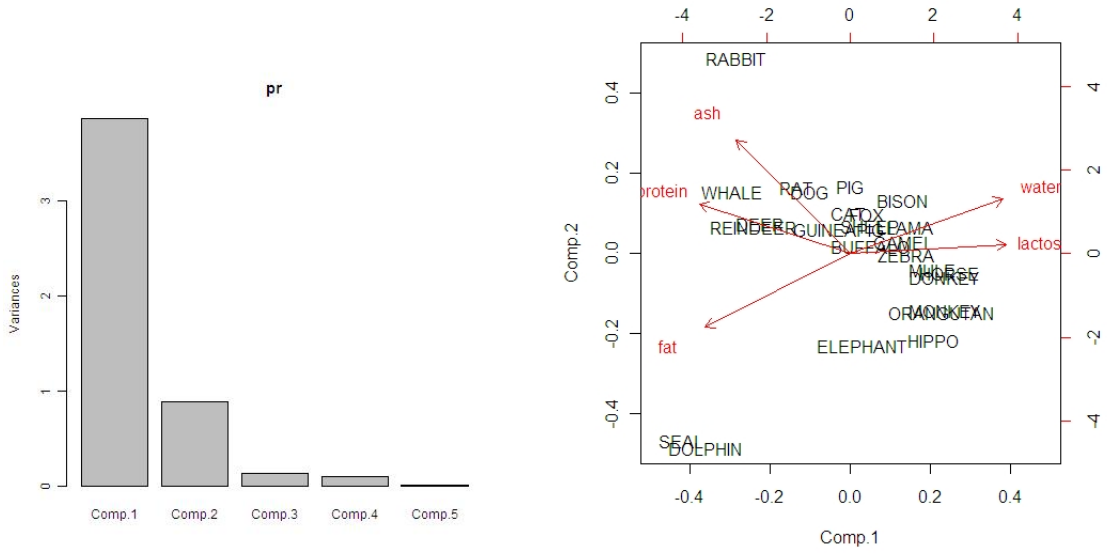


Figure 3.4: Composition du lait - Ébouli des valeurs propres (à gauche) et représentation simultanée sur le premier plan principal de l'ACP (à droite)

(a) $\sqrt{n}(\hat{\lambda} - \lambda) \to_d \mathcal{N}_p(0, 2\Lambda^2)$,
avec $\hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_p)^T$ *et* $\lambda = (\lambda_1, \cdots, \lambda_p)^T$ *sont les diagonales de* $\hat{\Lambda}$ *et* Λ.
(b) $\sqrt{n}(g_j - \gamma_j) \to_d \mathcal{N}_p(0, \mu_j)$,
avec $\mu_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \gamma_k \gamma_k^T$.
(c) les éléments de $\hat{\lambda}$ *sont asymptotiquement indépendants de ceux de* Γ.

$W_p(\Sigma, n)$ est la loi de Wishart de variance Σ à n degrés de liberté. C'est une généralisation de la loi du khi pour les matrices aléatoires.

Comme $n\hat{\Sigma} \sim W - p(\Sigma, n-1)$ si $X_1, = \cdots, X_n$ sont distribuées suivant une loi de Gauss de moyenne μ et de variance Σ, on déduit du théorème que

$$\sqrt{n-1}(\hat{\lambda}_j - \lambda_j) \to \mathcal{N}(0, 2\lambda_j^2),\ j = 1, \cdots, p$$

En appliquant une transformation log, on obtient par la delta méthode,

$$\sqrt{n-1}(\log(\hat{\lambda}_j) - \log(\lambda_j)) \to \mathcal{N}(0, 2),\ j = 1, \cdots, p$$

et on peut alors écrire un intervalle de confiance pour $\log(\lambda_j)$.

3.6 ACP par minimisation de l'erreur

On peut voir l'analyse en composantes principales comme un outil de synthèse d'information. L'idée est alors de chercher des facteurs latents sur lesquels se concentre l'information. Les facteurs latents jouent le même rôle que les composantes principales U. On peut alors écrire pour les variables centrées

$$\tilde{\mathbf{X}} = \sum_{j=k}^{q} c_q U_q + \epsilon$$

Dans le cas de l'ACP, on suppose que ϵ est un vecteur aléatoire gaussien centré dont les composantes sont indépendantes et de même variance : $\epsilon \sim \mathcal{N}(0, \sigma I)$. On a à faire à un modèle linéaire on peut donc réaliser l'inférence des paramètres inconnus z et U par minimisation de la variance des résidus. C'est à dire qu'on cherche les matrices $\mathbf{c}^*$ et $\mathbf{U}^*$ telles que

$$\begin{aligned}(\mathbf{c}^*, \mathbf{U}^*) &= arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T = Id\}} Var\left(\mathbf{X} - \sum_{k=1}^{q} c_k U_k\right) \\ &= arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T = Id\}} ||\mathbf{x} - \sum_{j=1}^{q} c_j U_j||^2 \end{aligned} \tag{3.4}$$

En pratique, on ne sait pas calculer cette variance. On l'estime à partir des observations. Et on montre que la solution unique est donnée par les composantes principales $\hat{U}$ et les axes principaux $\hat{z}$, vecteurs propres de la matrice de variance-covariance.

3.7 Changement de métrique dans l'espace des individus et poids sur les individus

Supposons maintenant que la mesure adéquate entre les individus n'est plus la distance euclidienne mais doive être basée sur une norme $||x||_M^2 = x^T M x$ où M est une matrice symétrique définie positive. La métrique M renormalise correctement les individus et il faut la prendre en compte dans le calcul d'inertie. Ceci est automatique si on considère la matrice des individus $\mathbf{x}' = \mathbf{x}M^{1/2}$. Dans la représentation des individus sur les axes factoriels c'est la nouvelle distance qui est approchée.

De manière analogue, si on veut donner des poids différents aux individus dans le calcul de l'inertie, on peut introduire une matrice de poids D qui est une matrice diagonale contenant les poids : $\mathbf{x}' = D^{1/2}\mathbf{x}M^{1/2}$.

Dans ce cas, on formule le problème (3.4) ainsi :

$$(\mathbf{c}^*, \mathbf{U}^*) = arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T=Id\}} ||\mathbf{x} - \sum_{k=1}^{q} c_k U_k||^2_{(M,D)}$$

Si l'espace est euclidien, par définition

$$||\mathbf{x} - \sum_{j=1}^{q} c_j U_j||^2_{(M,D)} = D(\mathbf{x} - \sum_{j=1}^{q} c_j U_j)^T M(\mathbf{x} - \sum_{k=1}^{q} c_k U_k)$$

La solution est donnée par

$$\sum_{k=1}^{q} c_k U_k = \sum_{k=1}^{q} \lambda_j^{1/2} u_k v_k^T$$

avec U et V des matrices unitaires. C'est la décomposition en valeurs singulières de la matrice des données centrées réduites. Les vecteurs u_k sont les vecteurs propres de la matrice de covariance $\mathbf{x}M\mathbf{x}^T D$, les valeurs propres étant rangées par ordre décroissant. Tandis que les vecteurs v_k sont les vecteurs propres de $\mathbf{x}^T M\mathbf{x}D$ correspondant aux mêmes valeurs propres. Ils sont correspondant aux *axes principaux*.

A partir de V_q, matrice construite à partir des q premiers vecteurs v_k, on construit la matrice de projection $P_q = V_q V_q^T M$.

Le choix de la métrique M et/ou de la matrice de pondération D a un impact sur les résultats et notamment sur les projections des individus sur les plans factoriels. Certaines métriques permettent par exemple de mettre en évidence les individus atypiques (voir TD). Le plus souvent, on choisi $D = \frac{1}{n}\mathbf{I}$ et $M = \mathbf{I}$ avec $\mathbf{I}$ la matrice identité. C'est à dire qu'on donné le même point à chaque individu et qu'on ne privilégie aucune variable.

Chapter 4

Analyse Canonique des Corrélation

Voir aussi la fiche wikistat de P. Besse (`www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acc`

4.1 Introduction

L'analyse canonique[1] permet les liaisons qui existent entre un groupe de variables à expliquer et un autre groupe de variables explicatives observées sur le même ensemble d'individus, c'est-à-dire de déterminer les corrélations existant entre les deux groupes de variables.

Par exemple, dans une étude de satisfaction de la clientèle de différents magasins, les variables à expliquer sont les suivantes

- Note de satisfaction obtenue sur l'accueil en magasin
- Note de satisfaction obtenue sur le conseil en magasin
- Note de satisfaction obtenue sur les délais de passage en caisse
- Note de satisfaction obtenue sur la largeur de l'assortiment

et les variables explicatives

- La taille du point de vente
- Le nombre de caisses ouvertes
- Le chiffre d'affaires quotidien du point de vente
- Le nombre de vendeuses dédiées à la surface de vente
- Le nombre de références dans la gamme A
- Le nombre de références dans la gamme B
- La surface de l'espace Loisirs

Toutes les variables sont mesurées dans toues les magasins. Et on obtient ainsi deux tables de données comportant le même nombre de lignes (nombre de magasins) l'une décrivant le magasin l'autre la satisfaction des clients. L'analyse canonique conclut sur le pouvoir explicatif de chacune des variables explicatives et le degré d'explication des variables à expliquer. Ainsi,

[1] En anglais : *Canonical Correlation Analysis*

dans cet exemple, il s'agit de comprendre le lien entre les critères de la surface de vente et la satisfaction des clients.

Le principe général de la méthode consiste à rechercher le couple de vecteurs, l'un lié aux magasin, l'autre à la satisfaction client, les plus corrélés possible. Ensuite, on recommence en cherchant un second couple de vecteurs non corrélés aux vecteurs du premier et le plus corrélés entre eux, et ainsi de suite. La démarche est donc similaire à celle utilisée en A.C.P. La représentation graphique des variables se fait soit par rapport aux vecteurs liés aux magasins, soit par rapport à ceux liés aux clients (en général, les deux sont équivalentes, au moins pour ce qui est de leur interprétation). Ces vecteurs, obtenus dans chaque espace associé à chacun des deux groupes de variables, sont analogues aux facteurs de l'A.C.P. et sont ici appelés variables canoniques. Comme en A.C.P., on peut tracer le cercle des corrélations sur le graphique des variables, ce qui en facilite l'interprétation (dont le principe est le même que pour le graphique des variables en A.C.P.). Des considérations techniques permettent de faire également un graphique pour les individus.

On note $\mathbf{X_1}$ et $\mathbf{X_2}$ les deux vecteurs de variables et $\mathbf{x_1}$ et $\mathbf{x_2}$ les tableaux des données observées. Les variables sont quantitatives ou qualitatives. Si les variables sont qualitatives alors, les colonnes du tableau de données sont constituées par les modalités des variables.

4.2 Interprétation géométrique de l'analyse canonique

L'objectif de l'analyse canonique est de trouver une représentation dans laquelle les proximités entre les deux ensembles de données soient maximisées. Autrement dit, on cherche des variables canoniques, transformées linéaires des variables d'origine, telles qu'en moyenne ces variables soient les plus proches possible (c'est à dire de corrélation maximum).

4.2.1 Analyse canonique ordinaire

Supposons dans un premier temps qu'on cherche seulement les deux premières variables canoniques. Si E et F sont les espaces engendrés par les colonnes de X_1 et X_2 respectivement, on cherche deux vecteurs unitaires u_E et u_F, un dans E et l'autre dans F, qui soient les plus proches possible. Soient P_E et P_F les projections orthogonales sur E et F.

$$P_E = X_1(X_1^T X_1)^{-1} X_1^T \text{ et } P_F = X_2(X_2^T X_2)^{-1} X_2^T$$

La minimisation de $||u_E - u_F||^2$ sous les contraintes $P_E u_E = u_E$, $P_F u_F = u_F$ et $||u_E|| = ||u_F|| = 1$ conduit à

$$P_E P_F u_E = \lambda^2 u_E$$
$$P_F P_E u_F = \lambda^2 U_F$$

pour un certain λ. C'est à dire que u_E est un vecteur propre à droite de $P_E P_F$ et u_F un vecteur propre à droite de $P_F P_E$. On en déduit facilement que

$$P_E u_F = \lambda u_E \text{ et } P_F u_E = \lambda u_F$$

de plus $\lambda = u_E^T u_F$.

Il est facile de vérifier que le vecteur $z = u_E + u_F$ est solution de $(P_E + P_F)z = (\lambda + 1)z$. C'est le point qui minimize $||z - P_E z||^2 + ||z - P_F z||^2$, la somme des carrés des distances à E et F. On a également $P_E z = (\lambda + 1)u_E$ et $P_F z = (\lambda + 1)u_F$.

C'est l'analogue d'une ACP avec $p = 2$, sauf que les variables ont été remplacées par des espaces. Les facteurs propres a_E et a_F sont tels que

$$a_E = X_1 u_E \text{ et } a_F = X_1 u_F$$

Les facteurs non normalisés sont donnés par

$$u = (X_1^T X_1)^{-1} X_1^T u_E = (1+\lambda)^{-1} (X_1^T X_1)^{-1} X_1^T z$$
$$v = (X_2^T X_2)^{-1} X_2^T u_F = (1+\lambda)^{-1} (X_2^T X_2)^{-1} X_2^T z$$

car en effet $X_1 u = (1+\lambda)^{-1} P z = u_E$. On remarque que u est alors l'estimateur aux moindres carrés de la régression linéaire permettant de prédire u_E à partir des variables de X_1.

On remarque que $||u_E - u_F||^2$ peut s'écrire

$$||u_E - u_F||^2 = E\left((u_E - u_F)^2\right) = E\left(u_E^2\right) + E\left(u_F^2\right) - 2E\left(u_E^T u_F\right)$$

Ainsi, si les données sont réduites, minimiser $||u_E - u_F||^2$ est équivalent à maximiser la covariance entre u_E et u_F. Ainsi, en analyse canonique des corrélations, on cherche des vecteurs u et v tels que les variables aléatoires $U_1 = X_1 u$ et $v_1 = X_2 v$ maximisent la corrélation

$$\rho = cor(X_1 u, X_2 v) = u^T \Sigma_{12} v$$

Proposition 3. *Les vecteurs $U_{E,s}$ sont les vecteurs propres normés de la matrice $P_E P_F$ respectivement associés aux valeurs propres λ_s rangées par ordre décroissant (on peut vérifier que ces valeurs propres sont comprises entre 1 et 0). De même, les vecteurs $U_{F,s}$ sont les vecteurs propres normés de la matrice $P_F P_E$ respectivement associés aux mêmes valeurs propres λ_s. De plus, les coefficients de corrélation canonique $\rho_s = \sqrt{\lambda_s}$ sont les racines carrées positives de ces valeurs propres.*

Les facteurs a_E^k et a_F^k ont les propriétés suivantes

- $a_E^k = X_1 u_{E,k}$ et $a_F^k = X_2 u_{F,k}$
- Les facteurs a_E sont solution de

  $$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a_{E,k} = R^2(u_{E,k}, u_{E,k}$$

  où $\Sigma_{ij} = \frac{1}{n} (X_i)^T X_j$
- Les facteurs a_F sont solution de

  $$\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} a_{F,k} = R^2(u_{F,k}, u_{F,k}$$

  où $\Sigma_{ij} = \frac{1}{n} (X_i)^T X_j$
- Les relations qui existent entre a_E et A_F sont

  $$\Sigma_{11}^{-1} \Sigma_{12} a_{F,k} = R(u_{E,k}, u_{F,k}) a_{E,k}$$

  et

  $$\Sigma_{22}^{-1} \Sigma_{21} a_{E,k} = R(u_{E,k}, u_{F,k}) a_{F,k}$$

  où R est le coefficient de détermination entre $u_{E,k}$ et $u_{F,k}$.

4.2.2 Analyse canonique généralisée

On a désormais une matrice $(X_1, \cdots, X_p)$, des espaces $E_1, \cdots, E_p$ et on veut faire une opération analogue. Il est difficile de chercher directement une famille $u_1, \cdots, u_p$. On cherche alors le vecteur z qui minimise $\sum_{j=1}^{p} ||z - P_j z||^2$. La solution est un vecteur propre de $\sum_{j=1}^{p} P_j$. Les composantes principales sont les $P_j z$ (la normalisation a changé), et les facteurs sont calculés de la même façon

$$\sum_j P_j z = \lambda z$$

$$c_j = P_j z$$

$$w_j = (X_j^T X_j)^{-1} X_j^T z, \quad X w_j = c_j$$

C'est l'analogue d'une ACP normalisée sauf que les variables ont été remplacées par des espaces. On obtient les autres axes factoriels en résolvant de nouveau ces équations. On cherche des axes orthogonaux $z_1, \cdots, z_q$ associés à des valeurs propres $\lambda_1, \cdots, \lambda_q$ décroissantes.

4.3 Représentations graphiques

Le but de l'analyse canonique est de mettre en évidence des proximités entre deux ensembles de données. Les représentations graphiques ont pour objectif de décrire les proximités entre variables et entre individus.

4.3.1 Représentation des variables

On note u les vecteurs propres liées à $X = X_1$ et v les vecteurs propres liées à $Y = X_1$.
S'intéresser aux $k$ième facteur (ou variable canonique), est équivalent à expliquer la corrélation entre u_k et v_k soit à expliquer la corrélation entre une combinaison linéaire de variables de $X = X_1$ et de variables de $Y = X_2$ est élevée. Il est donc nécessaire de faire figurer sur un même graphique l'ensemble des variables d'origine. Cette représentation se fait comme en ACP par un cercle des corrélations. L'axe correspondant au $k$ième facteur est une compromis entre u_k et v_k soit

$$F_k = \frac{1}{2}(u_k + v_k)$$

4.3.2 Représentation des individus

L'analyse canonique détermine des facteurs u et v tels qu'en moyenne les deux variables soient le plus proches possibles pour les n individus, c'est à dire de telle sorte que

$$\frac{1}{n} \sum_{i=1}^{n} (u_i - v_i)^2 \text{ pour tout } j = 1, \cdots, q \tag{4.1}$$

sous les mêmes contraintes que dans l'espace des variables.

Chacun des deux tableaux de données décrit un nuage pour les mêmes n individus. La représentation des individus de l'AC permet de cerner ce qui caractérise le mieux ces nuages d'individus dans les directions pour lesquelles ces nuages sont les plus ressemblants possibles. De plus la représentation des individus de l'AC permet de repérer les individus ayant un comportement particulier.

A chaque étape k, il s'agit de comparer la description des individus donnée par la variable canonique $u_{E,k}$ à celle donné par la variable canonique $u_{F,k}$. La proximité plus ou moins importante entre les deux descriptions des individus peut aussi être mise en évidence en calculant l'écart-type résiduel (4.1).

4.4 Exemple

Considérons l'exemple suivant dans lequel on cherche les relations entre des variables physiologiques et des exercices pratiqués dans des salles de sport pour 20 hommes d'âge moyen.

```
data Fit;
    input Weight Waist Pulse Chins  Situps Jumps;
    datalines;
 191  36  50   5  162   60
 189  37  52   2  110   60
 193  38  58  12  101  101
 162  35  62  12  105   37
 189  35  46  13  155   58
 182  36  56   4  101   42
 211  38  56   8  101   38
 167  34  60   6  125   40
 176  31  74  15  200   40
 154  33  56  17  251  250
 169  34  50  17  120   38
 166  33  52  13  210  115
 154  34  64  14  215  105
 247  46  50   1   50   50
 193  36  46   6   70   31
 202  37  62  12  210  120
 176  37  54   4   60   25
 157  32  52  11  230   80
 156  33  54  15  225   73
 138  33  68   2  110   43
 ;
 proc cancorr data=Fit all
      vprefix=Physiological vname='Physiological Measurements'
      wprefix=Exercises wname='Exercises';
    var Weight Waist Pulse;
    with Chins Situps Jumps;
    title 'Middle-Aged Men in a Health Fitness Club';
    title2 'Data Courtesy of Dr. A. C. Linnerud, NC State Univ';
 run;
```

On obtient les résultats suivants pour les coefficients normalisés et on fait l'interprétation suivante. Le premier facteur des variables physiologiques est une différence pondérée du tour de taille (1.58) et du poids (-0.78). Les corrélations entre la taille et le poids et la première variable canonique sont positives, 0.92 pour la taille et de 0.62 pour le poids. Le poids est donc une variable *suppressor*, ce qui signifie que son coefficient et sa corrélation sont de signes opposés.

	Physiological1	Physiological2	Physiological3
Poids	-0.78	-1.88	-0.19
Tour de taille	1.58	1.18	0.51
Poul	-0.06	-0.23	1.05

La première variable canonique pour les variables exercice montre également un mélange de signes: Situps (-1.05), Chins (-0.35), Jumps (0.72). Toutes les corrélations sont négatives, ce qui indique que Jumps est également une variable *suppressor*.

	Exercises1	Exercises2	Exercises3
Chins	-0.066	-0.071	-0.245
Situps	-0.017	0.002	0.020
Jumps	0.014	0.021	-0.008

Il peut sembler contradictoire qu'une variable ait un coefficient de signe opposé à celui de sa corrélation avec la variable canonique. Afin de comprendre comment cela peut arriver, considérons une situation simplifiée : la prédiction de Situps à partir du tour de taille et et du poids par régression multiple. En termes informels, il semble plausible que les gens gros fassent moins de situps (abdos) que les personnes maigres. Supposons que les hommes de l'échantillon aient tous environ la même taille, il y a donc une forte corrélation entre la le tour de taille et de poids (0.87).

Nous nous intéressons ensuite au coefficient de corrélation mutiple entre les mesures physiologiques et les M variables canoniques correspondant aux exercices et inversement. Les tableaux ci-dessous donnes les carrés des coefficients.

M	1	2	3
Poids	0.24	0.27	0.27
Tour de taille	0.54	0.55	0.55
Poul	0.07	0.07	0.07

M	1	2	3
Chins	0.33	0.34	0.34
Situps	0.42	0.44	0.44
Jumps	0.02	0.05	0.05

Les coefficients sont presque identiques pour les 3 variables canoniques. On en conclut, qu'un seule variable canonique suffit à caractériser chaque ensemble. Et que ce sont surtout le poids et le tour de taille qui sont corrélés aux exercices, principalement les exercices de tractions et d'abdominaux.

4.5 Interprétation probabiliste de l'analyse canonique

4.5.1 Rappel : analyse en composante principale

Comme nous l'avons évoqué dans la section précédente, on peut voir l'ACP comme une solution du maximum de vraisemblance d'une analyse en facteurs avec une covariance isotrope (c'est à dire identique dans toutes les directions soit pour toutes les variables). Plus précisément,

$$P(X|F = f) \sim \mathcal{N}(Qf + \mu, \sigma^2 I_p), \quad \sigma > 0$$

En pratique, Q et σ sont estimés par

$$\hat{Q} = U_q(\Lambda_q - \hat{\sigma}^2 I)^{1/2} R, \quad \hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^{q} \lambda_j$$

où U sont les vecteurs propres principaux de la matrice de covariance empirique $\hat{\Sigma}$ de $\mathbf{x}$ correspondants aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$, $q \leq p$ et R est une matrice orthogonale quelconque.

On a alors

$$E(F|\mathbf{X} = \mathbf{x}) = R^T(\Lambda - \sigma^2 I)^{1/2}\Lambda^{-1}U^T(\mathbf{x} - \hat{\mu}))$$

Comme nous l'avions précisé dans le chapitre précédent, ces équations conduisent au même sous espace linéaire que l'ACP et les mêmes projections des individus (à une rotation près) si les valeurs propres $\lambda_{q+1}, \cdots, \lambda_p$ sont égales à 0.

4.5.2 Modèle probabiliste pour l'analyse canonique

Dans l'ACP, on cherche une transformation linéaire des variables d'origine $\mathbf{X}$ telle que les composantes du vecteur transformé soient non corrélées. De la même façon, en analyse canonique, pour deux groupes de variables $\mathbf{X_1}$ et $\mathbf{X_2}$ de dimension p_1 et p_2, on cherche une paire de transformations linéaires telle que une composante de chaque vecteur transformé ne soit corrélée qu'avec une seule composante de l'autre vecteur.

Le modèle probabiliste s'écrit donc

$$\begin{aligned} &F \sim \mathcal{N}(0, I_q), \quad \min(p_1, p_2) \geq q \geq 1 \\ &P(\mathbf{X_1}|F = f) = \mathcal{N}(Q_1 f + \mu_1, \Psi_1) \\ &P(\mathbf{X_2}|F = f) = \mathcal{N}(Q_2 f + \mu_2, \Psi_2) \end{aligned}$$

Notons $\hat{\Sigma}_1$ et $\hat{\Sigma}_2$ les matrices de corrélation empiriques des observations $\mathbf{x_1}$ et $\mathbf{x_2}$ de $\mathbf{X_1}$ et $\mathbf{X_2}$. Les paramètres Q_1, Q_2, Ψ_1, Ψ_2 de ce modèle sont estimés par maximum de vraisemblance, et on obtient

$$\begin{aligned} \hat{Q}_1 &= \hat{\Sigma}_1 U_1 M_1 \\ \hat{Q}_2 &= \hat{\Sigma}_2 U_2 M_2 \\ \hat{\Psi}_1 &= \hat{\Sigma}_1 - \hat{Q}_1 \hat{Q}_1^T \\ \hat{\Psi}_2 &= \hat{\Sigma}_2 - \hat{Q}_2 \hat{Q}_2^T \end{aligned}$$

Les matrices M_1, M_2 sont des matrices carré arbitraires telles que $M_1 M_2 = P$ et telles que leur norme spectrale[2] soit inférieure à 1. Les matrices U_1 et U_2 sont telles que leurs colonnes sont les directions canoniques rangées par ordre de valeur propre décroissante. Et P est la matrice diagonale des correlations canoniques.

Notons $\hat{\Sigma}$ la matrice de correlation empirique des observations

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}$$

[2]La norme spectrale est la norme matricielle induite par la norme euclidienne et est définie pour ne matrice A carrée par $||A|| = \sqrt{\lambda_{max}(AA^T)}$

où $\hat{\Sigma}_{kl}$ est la matrice de corraltion empirique de du couple (X_k, X_l). On a alors

$$U_k = \Sigma_{kk}^{-1/2} V_k, \text{ pour } k = 1, 2$$

avec V_1, V_2 tels que de

$$\hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2} = V_1 P V_2$$

avec P la matrice diagonale des valeurs singulières. Ainsi P a sur sa diagonale les corrélations canoniques ρ_i, $i = 1, \cdots, q = \min(p_1, p_2)$ et des 0 ailleurs. Si la matrice de covariance $\hat{\Sigma}$ est inversible, on a

$$\begin{aligned} U_1^T \hat{\Sigma}_{11} U_1 &= I \\ U_2^T \hat{\Sigma}_{22} U_2 &= I \\ U_2^T \hat{\Sigma}_{21} U_1 &= P \end{aligned}$$

La solution n'est pas unique. Et on peut montrer que parmi toutes les solutions, celle qui minimize $-\log(det(\Psi)) = -\log(det(\Psi_1)) - \log(det(\Psi_2))$ (i.e. l'entropie conditionnelle de X sachant F), est telle que $M_1 = M_2 = P^{1/2} R$ avec R une matrice de rotation. La solution s'écrit alors

$$\begin{aligned} \hat{Q}_1 &= \hat{\Sigma}_1 U_1 P^{1/2} R \\ \hat{Q}_2 &= \hat{\Sigma}_2 U_2 P^{1/2} R \end{aligned}$$

Comme dans le cas de l'ACP, on peut en déduire les propriétés des facteurs sachant les observations $\mathbf{x_1}$ et $\mathbf{x_2}$,

$$\begin{aligned} &E(F|\mathbf{X_k} = \mathbf{x_k}) = M_k^T U_k^T (\mathbf{x_k} - \hat{\mu}_k), \quad k = 1, 2 \\ &Var(F|\mathbf{X_k} = \mathbf{x_k}) = I - M_k M_k^T \\ &E(F|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^T \begin{pmatrix} (I-P^2)^{-1} & (I-P^2)^{-1}P \\ (I-P^2)^{-1}P & (I-P^2)^{-1} \end{pmatrix} \begin{pmatrix} U_1^T(\mathbf{x_1} - \hat{\mu}_1) \\ U_1^T(\mathbf{x_1} - \hat{\mu}_1) \end{pmatrix} \\ &Var(F|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) = I - \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^T \begin{pmatrix} (I-P^2)^{-1} & (I-P^2)^{-1}P \\ (I-P^2)^{-1}P & (I-P^2)^{-1} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \end{aligned}$$

On remarque ici que M_1 et M_2 définissent des sous espaces dans lesquels $\mathbf{x_1}$ et $\mathbf{x_2}$ sont projettés.

Chapter 5

Analyse des Correspondances

5.1 Introduction

L'analyse factorielle des correspondances est un cas particulier de l'analyse canonique. Elle a été développée essentiellement par J.-P. Benzecri durant la période 1970-1990. L'analyse des correspondances est une technique d'analyse fatorielle destinée à mettre en évidence et décrire des associations entre deux variables qualitatives. On considère dans cette section deux variables qualitatives observées simultanément sur n individus de poids identiques $1/n$. En pratique, on va travailler avec une table de contingence qui est un tableau croisé contenant les effectifs des occurences simultannées de deux modalités.

Prenons des exemples,

1. Ponctuation dans l'oeuvre de Zola *(exemple emprunté M. Tenenhaus)* - L'étude de la ponctuation ou de la présence de certains mots dans des textes est utilisée pour reconnaitre l'auteur d'un document (article, roman, nouvelle, etc.). Les données se présentent selon le tableau Tab. **??**.

   Et une analyse factorielle des correspondances permet de faire le graphique suivant sur lequel on projette simultannément les modalités des deux variables (`Titre du roman` et `Ponctuation`) comme représenté dans la figure **??**.

2. Origine sociale des étudiants de première année et choix d'un secteur disciplinaire *(exemple emprunté à F.-G. Carpentier)*

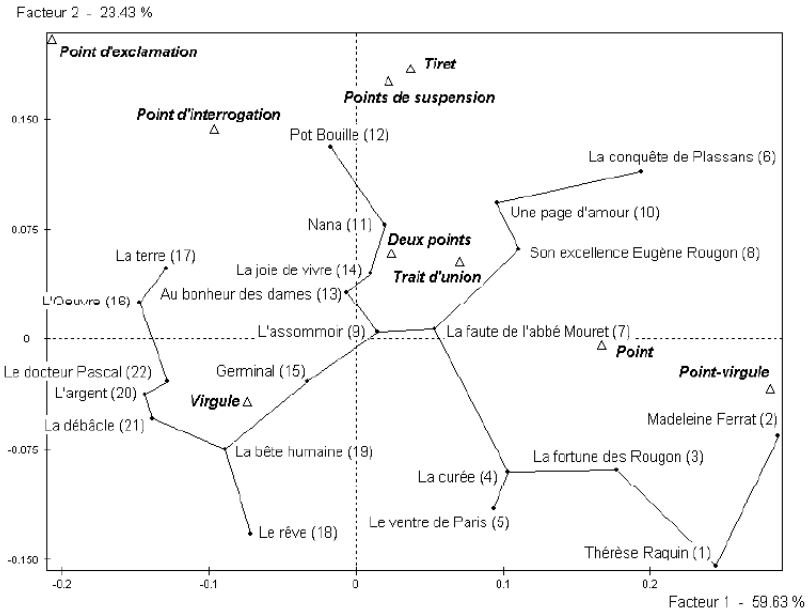
	Droit	Science	Médecine	IUT	**Total**
Exp. agri.	80	99	65	58	**302**
Patron	168	137	208	62	**575**
Cadre sup.	470	400	876	79	**1825**
Employé	145	133	135	54	**467**
Ouvrier	166	193	127	129	**615**
Total	**1029**	**962**	**1411**	**382**	**3784**

Soient[1] deux variables nominales X et Y, comportant respectivement p et q modalités. On a observé les valeurs de ces variables sur une population et on dispose d'un tableau de contingence

[1] Certaines parties de ce chapitre et notamment ce paragraphe sont fortement inspirées du cours de F.G. Carpentier

Roman	.	...	!	?	,	;	:	–	-
1. Thérèse Raquin	3468	236	138	76	6195	691	168	285	543
2. Madeleine Ferrat	5131	362	236	245	8012	922	291	518	1115
3. La fortune des Rougon	6157	238	534	229	11346	936	362	711	1301
4. La curée	4958	443	357	232	11164	738	364	679	1200
5. Le ventre de Paris	5538	534	426	232	13234	1015	318	734	1201
6. La conquête de Plassans	6292	943	756	512	11585	1285	402	1432	1916
7. La faute de l'abbé Mouret	6364	679	859	462	13948	634	377	1067	1564
8. Son excellence Eugène Rougon	7258	728	1002	496	14295	889	543	1469	1907
9. L'assommoir	7820	769	1929	443	19244	1399	436	995	2272
10 Une page d'amour	6206	843	918	492	11953	647	347	1235	1409
11. Nana	7821	1007	1796	611	17881	1087	509	1523	1797
12. Pot Bouille	6875	1045	1873	651	17044	912	675	1669	1935
13. Au bonheur des dames	6916	808	1313	651	18402	972	642	1531	2114
14. La joie de vivre	5803	710	972	623	13917	602	420	1142	1590
15. Germinal	7944	606	1463	729	21388	908	621	1362	2083
16. L'Œuvre	5000	774	1692	668	18292	811	566	1107	1489
17. La terre	6979	957	2307	796	23417	947	657	1681	2113
18. Le rêve	3052	292	385	237	9551	345	230	416	650
19. La bête humaine	5484	601	929	557	18264	673	467	957	1721
20. L'argent	5022	850	1235	569	19267	684	399	1049	1677
21. La débâcle	7440	860	1833	690	26482	832	564	1398	2197
22. Le docteur Pascal	4586	621	1072	464	15598	462	315	955	1218

Ponctuation dans les romans de Zola



Premier plan factoriel de l'ACM de la ponctuation dans la romans de Zola.

à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison d'une modalité i de X et d'une modalité j de Y. Les valeurs de ce tableau seront notées n_{ij} , l'effectif total sera noté N.

L'AFC vise à analyser ce type de tableaux en apportant des réponses à des questions telles que :

- Y a-t-il des lignes du tableau (modalités de X) qui se "ressemblent", c'est-à-dire telles que les distributions des modalités de Y soient analogues ?
- Y a-t-il des lignes du tableau (modalités de X) qui s'opposent, c'est-à-dire telles que les distributions des modalités de Y soient très différentes ?
- Mêmes questions pour les colonnes du tableau.
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

La méthode se fixe également comme but de construire des représentations graphiques mettant en évidence ces propriétés des données.

Notations

Soit $\mathbf{N} = (n_{ij})_{i=1,\cdots,p,j=1,\cdots,q}$ un tableau de contingence. On définit les marges du tableau par

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \ , \ n_{\bullet j} = \sum_{i=1}^{p} n_{ij}, \ n = n_{\bullet\bullet} \sum_{i,j} n_{ij}$$

Ceci correspond aux totaux en lignes et en colonne. Selon le mÃême principe, on peut définir les marges en fréquence avec $f_{ij} = n_{ij}/n$

$$f_{i\bullet} = \sum_{j=1}^{q} f_{ij}, \ f_{\bullet j} = \sum_{i=1}^{p} f_{ij}, \ f_{\bullet\bullet} \sum_{i,j} f_{ij} = 1$$

5.2 Modèle d'indépendance

5.2.1 Test du chi 2

Comme en ACP, on s'intéresse alors aux directions de "plus grande dispersion" de chacun de ces nuages de points, mais on utilise la distance du χ^2 entre ces deux variables (à la place de la distance euclidienne). Cette distance permet de comparer l'effectif de chacune des cellules du tableau de contingence à la valeur qu'elle aurait si les deux variables étaient indépendantes. Notons E_{ij} l'effectif attendu sous l'hypothèse d'indépendance ; par définition

$$E_{ij} = \frac{\text{Total ligne } i \times \text{Total ligne } j}{\text{Total général}} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}$$

ce qui correspondant bien au produit des probabilités marginales. Et la distance du χ^2 est définie par

$$d^2_{\chi^2}(\mathbf{N}, \mathbf{E}) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

On appelle *résidus standardisés*, les variables (centrées et de variance 1) :

$$c_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

Plus la distance $d^2_{\chi^2}(\mathbf{N}, \mathbf{E})$ est grande, plus le tableau observé est éloigné du tableau attendu sous l'hypothèse d'indépendance.

Pourquoi utiliser cette métrique plutôt que la métrique euclidienne ? Deux raisons fortes peuvent être avancées :

- Avec la métrique du χ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes. Ceci a pour conséquence, dans l'exemple, des étudiants de première année que les catégories socio-professionnelles sur-représentées ne prennent pas plus de poids que les autres dans le calcul de la distance.
- La métrique du χ^2 possède la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

Notons qu'en revanche, il n'existe pas d'outil mesurant une "distance" entre une ligne et une colonne.

Sous l'hypothèse d'indépendance des deux variables, la statistique $d^2_{\chi^2}$ suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté. Cette loi sert, par exemple, à définir une règle de décision du type : *On conclut que les variables sont indépendantes avec un risque α de se tromper si $d_{\chi^2}(\mathbf{N}, \mathbf{E}) < F^{-1}_{(p-1)(q-1)}(1-alpha)$ vec F la fonction de répartition de la loi du χ^2 à $(p-1)(q-1)$ degrés de liberté*

Dans l'exemple des étudiants de première année, la distance du χ^2 observée est

$$d^2_{\chi^2,\text{obs}}(\mathbf{N}, \mathbf{E}) = 320.2$$

et on la compare à $F^{-1}_{12}(.95) = 21.0$. La valeur de la statistique observée $d^2_{\chi^2,\text{obs}}(\mathbf{N}, \mathbf{E})$ étant supérieure au seuil, on conclut ici que le tableau observé est significativement éloigné du tableau attendu sous l'hypothèse d'indépendance et donc que les deux variables sont liées.

5.2.2 AFC et indépendance

L'analyse d'un tableau de contingence doit donc se faire en référence à la situation de d'indépendance. C'est ce que fait l'AFC en écrivant le modèle d'indépendance sous la forme suivante :

$$\forall i = 1, \cdots, p, \forall j = 1, \cdots, q, \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$$

La quantité $f_{ij}/f_{i\bullet}$ est la probabilité conditionnelle de posséder la modalité j de la variable X_2 sachant que l'on possède la modalité i de la variable X_1. De façon symétrique, on peut écrire

$$\forall i = 1, \cdots, p, \forall j = 1, \cdots, q, \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$$

Définition 7.

- *L'ensemble de probabilités* $\{f_{ij}/f_{i\bullet}; j = 1, \cdots, q\}$ *est appelée profil ligne.*
- *L'ensemble de probabilités* $\{f_{ij}/f_{\bullet j}; i = 1, \cdots, p\}$ *est appelée profil colonne.*
- $\{f_{i\bullet}; j = 1, \cdots, q\}$ *(resp.* $\{f_{\bullet j}; i = 1, \cdots, p\}$ *est le profil moyen correspondant au profil ligne (resp. colonne).*

Remarque - Si on a indépendance, le profil ligne d'une part et colonne d'autre part est égal au profil moyen correspondant.

5.3 Analyse factorielle des correspondances

On va voir que l'AFC est une double ACP : ACP des profils ligne et ACP des profils colonne.

5.3.1 Nuages de points

Intéressons nous aux profils ligne, l'analyse des profils colonne étant symétrique. On peut définir la notion de nuage d'individus (ou de modalité) partir du tableau de contingence en fréquence. En pratique, on construit un nuage de points dans l'espace $\mathbb{R}^q$ en définissant pour chaque ligne i, un point dont la coordonnées dans la dimension j est $f_{ij}/f_{i\bullet}$. Ce nuage est complété par le point moyen G_I dont la $j$ème coordonnée vaut $f_{\bullet j}$. Chaque point i est affecté du poids $f_{i\bullet}$.

On remarque que la distance entre les points i et i' (c'est à dire deux modalités de X_1) est

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

On utilise donc ici la métrique du χ^2 dans laquelle les inverses des fréquences marginales des modalités de Y sont introduites comme pondérations des écarts entre éléments de deux profils relatifs à X. Cette métrique attribue donc plus de poids aux écarts correspondants à des modalités de faible effectif (rares) pour Y . L'inertie du point i par rapport à G_I s'écrit

$$\begin{aligned} \text{Inertie}(i/G_I) &= f_{i\bullet} d^2_{\chi^2}(i, G_I) \\ &= f_{i\bullet} \sum_{j=1}^{q} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 \\ &= \sum_{j=1}^{q} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} \end{aligned}$$

5.3.2 l'AFC proprement dite

Pour étudier les lignes, on peut réaliser une ACP de la matrice A (telle que $a_{ij} = n_{ij}/n_{i\bullet}$)) puis de représenter les modalités de la première variable. En raison du changement de métrique, on introduit la matrice $M = D_C^{-1}$ avec $D_C = diag(n_{\bullet 1}, \cdots, n_{\bullet q})$ et on considère la matrice de poids $D = D_L^{-1}$ avec $D_C L = diag(n_{1\bullet}, \cdots, n_{q\bullet})$ (pour favoriser les gros effectifs, ce qui est discutable mais permet de faire facilement les calculs). On remarque $A = D_L^{-1}\mathbf{N}$. De façon symétrique, on peut définir $B = \mathbf{N} D_C^{-1}$.

Proposition 4. *Les éléments de l'ACP de* (A, D_C^{-1}, D_L)[2] *sont fournis pas l'analyse spectrale de la matrice carrée,* D_L^{-1}*-symétrique et semi-definie positive* AB*.*

Preuve - Elle se construit en remarque successivement que

- le barycentre du nuage des profils?colonnes est le vecteur g_C des fréquences marginales de X_2,
- la matrice $A'D_LA - g_CD_Lg_C'$ joue le rôle de la matrice des variances?covariances,
- la solution de l'ACP est fournie par la D.V.S. de $(A-1g_L', D_C^{-1}, D_L)$ qui conduit à rechercher les valeurs et vecteurs propres de la matrice (SM)

$$A'D_LAD_C^{-1} - G_CD_LG_C' = AB - G_CG_C'D_R^{-1}(\text{car } D_C^{-1}A' = BD_L^{-1})$$

- les matrices $AB - G_CG_C'D_R^{-1}$ et AB ont les mêmes vecteurs propres associées aux mêmes valeurs propres, à l'exception du vecteur g_L associé à la valeur propre $\lambda_0 = 0$ de $AB - G_CG_C'D_R^{-1}$ et à la valeur propre $\lambda_0 = 1$ de AB.

$\diamond$

On note U la matrice contenant les vecteurs propres D_C^{-1}-othonormés de AB. La représentation des ?individus? de l'ACP réalisée fournit une représentation des modalités de la variable X_1 . Elle se fait au moyen des lignes de la matrice des composantes principales (XMV) :

$$C_L = AD_C^{-1}U.$$

Les composantes principales permettent de représenter les modalités des variables sur les axes 2 et 3 (le premier est constant égal à 1). Une proximité de deux points i et i' indique que la distribution de la seconde variable sachant que la première vaut i est similaire à celle sachant i'.

Pour les colonnes, on fait les mêmes calculs en inversant les lignes et les colonnes. Il s'agit donc de l'ACP des 'individus' modalités de X_2 ou profils colonne (la matrice des données est B), pondérés par les fréquences marginales des lignes de N (la matrice diagonale des poids est D_C) et utilisant la métrique du χ^2. Il s'agit donc de l'ACP de (B, D_C^{-1}, D_L).

Proposition 5. *Les éléments de l'ACP de* (B, D_L^{-1}, D_C) *sont fournis par l'analyse spectrale de la matrice carrée,* D_L^{-1} *?symétrique et semi?définie positive* BA*.*

En notant V la matrice des vecteurs propres de la matrice BA ; les coordonnées permettant la représentation des modalités de la variable X_2 sont fournies par la matrice :

$$C_C = BD_L^{-1}V.$$

Sachant que V contient les vecteurs propres de BA et U ceux de AB, montre qu'il suffit de réaliser une seule analyse, car les résultats de l'autre s'en déduisent simplement :

$$U = A'V\Lambda^{-1/2},$$

$$V = B'U\Lambda^{-1/2};$$

Λ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) commune aux deux ACP.

$$C_C = BD_L^{-1}V = BD_L^{-1}B'V\Lambda^{-1/2} = D_C^{-1}A'B'U\Lambda^{-1/2} = D_C^{-1}U\Lambda^{1/2},$$

[2]Matrice, Métrique, Pondération

$$C_L = AD_C^{-1}U = D_L^{-1}V\Lambda^{1/2}.$$

On en déduit les formules de transition

$$C_C = BC_L\Lambda^{-1/2}$$

$$C_L = AC_C\Lambda^{-1/2}$$

.

On est alors tenté de mettre toutes les modalités sur un même graphique (option par défaut dans SAS). La proximité de modalités de variables différentes reste néanmoins difficile à interpréter.

5.4 Représentation graphique

5.4.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{N}$ se transforme encore en :

$$\frac{f_{ij} - f_{i\bullet}f_{\bullet j}}{f_{i\bullet}f_{\bullet j}} = \sum_{k=0}^{\min(p-1,q-1)} \sqrt{\lambda_k}\frac{v_{ik}}{f_{i\bullet}}\frac{u_{jk}}{f_{\bullet j}}$$

En se limitant au rang r, on obtient donc, pour chaque cellule (i,j) de la table $\mathbf{N}$, une approximtion de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs

$$\frac{v_{ik}}{f_{i\bullet}}\lambda^{1/4} \text{ et } \frac{u_{jk}}{f_{\bullet j}}\lambda^{1/4}$$

termes génériques respectifs des matrices

$$D_L^{-1}V\Lambda^{1/4} \text{ et } D_C^{-1}U\Lambda^{1/4}$$

Leur représentation (par exemple avec $r = 2$) illustre alors la correspondance entre les deux modalités x_{1i} et x_{2j} : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'AFC apparaît ainsi comme la meilleure reconstitution des fréquences f_{ij}, ou encore la meilleure représentation des écarts relatifs à l'indépendance.

5.4.2 Représentation barycentrique

La représentation graphique usuelle dite *représentation quasi-barycentrique*, place les points $(c_L(1,i), c_L(2,i))$ et $(c_C(1,i), c_C(2,i))$.

$$C_L = D_L^{-1}V\Lambda^{1/2} \text{ et } C_C = D_C^{-1}\Lambda^{1/2}$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de ?caution? puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double ACP est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

C'est cette représentation s'étend plus facilement au cas de plusieurs variables.

La *représentation barycentrique* est une autre représentation proposée par les logiciels. Elle utilise les matrices

$$D_L^{-1}V\Lambda^{1/2} \text{ et } D_C^{-1}U\Lambda$$

ou

$$D_L^{-1}V\Lambda \text{ et } D_C^{-1}U\Lambda^{1/2}.$$

Si l'on considère alors la formule de transition

$$C_L = AC_C\Lambda^{1/2} \Leftrightarrow C_L\Lambda^{1/2} = AC_C \Leftrightarrow D_L^{-1}V\Lambda = AD_C^{-1}U\Lambda^{1/2}$$

Dans cette représentation, chaque modalité j de la deuxième variable est représentée comme barycentre des modalités i de la première variable avec un poids qui est la probabilite de i sachant j.

La formule suivante

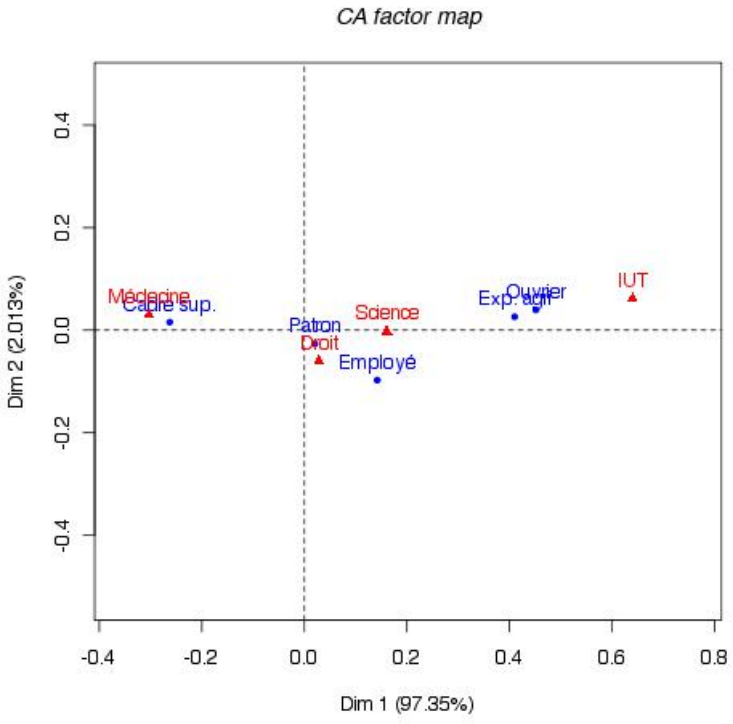
$$n_{ij} \simeq \frac{n_{i\bullet}n_{\bullet j}}{n}\left(1 + \frac{1}{\lambda_1}C_L(1,i)C_C(1,j) + \frac{1}{\lambda_2}C_L(2,i)C_C(2,j)\right)$$

indique que deux modalités formant un angle aigu (resp. obtus) s'attirent (resp. se repoussent) et ceci est d'autant plus marqué que les points sont éloignés du centre de gravité.

5.4.3 Exemples

Étudiants en première année

Dans l'exemple des étudiants en première année, on obtient le graphique suivant. On observe que toutes les modalités sont concentrées autour du premier axe. Ceci signifie qu'on a essentiellement une seule variable latente (ou facteur) structurante.

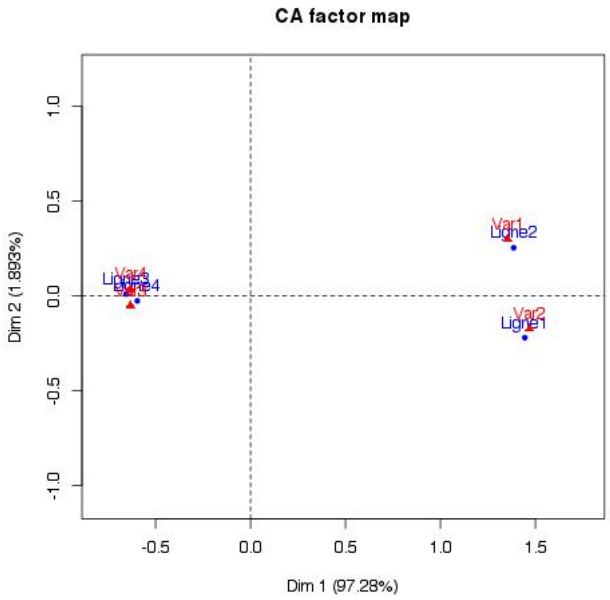


Sous groupes dans les données

Quand il existe des sous groupes dans les données, on obtient des résultats typiques. Par exemple, si on fait l'AFC du tableau suivant (tableau 1.)

	Var 1	Var 2	Var 3	Var 4
Ligne 1	20	45	2	0
Ligne 2	25	32	0	3
Ligne 3	1	0	78	112
Ligne 4	2	1	45	44

on obtient la projection ci-dessous sur le premier plan factoriel.

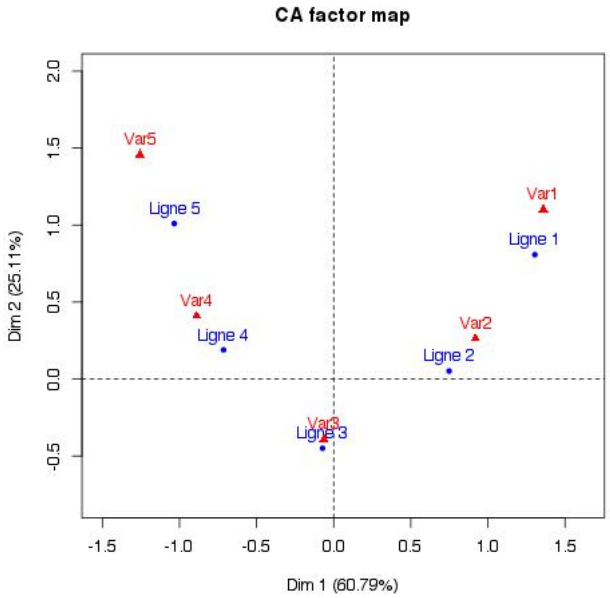


Effet Guttman

Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment lorsque les deux variables sont ordinales, et classent les sujets de la même façon. Dans ce cas, le premier axe oppose les valeurs extrêmes et classe les valeurs, tandis que le deuxième axe oppose les intermédiaires aux extrêmes.

	Var1	Var2	Var3	Var4	Var5
Ligne 1	10	30	7	0	0
Ligne 2	3	100	70	4	0
Ligne 3	2	32	200	35	1
Ligne 4	1	6	80	100	2
Ligne 5	0	3	5	25	5

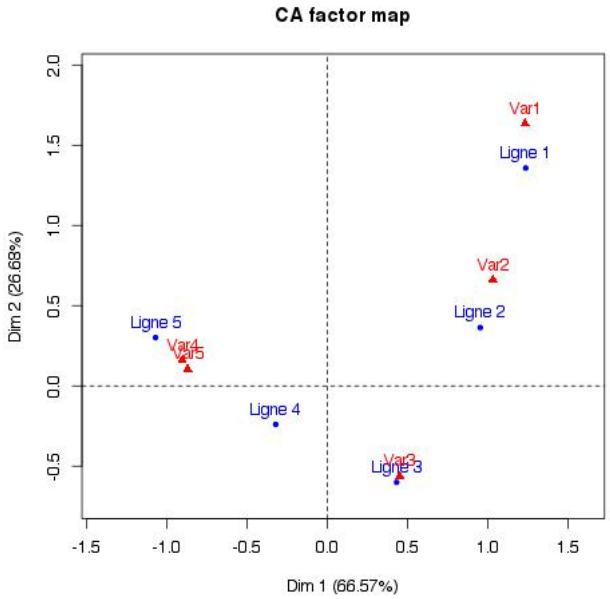
On obtient la projection ci-dessous sur le premier plan factoriel.



Exercice : Expliquer le graphique suivant associé au tableau ci-dessous en regard des résultats du jeux de données précédent.

	Var1	Var2	Var3	Var4	Var5
Ligne 1	10	30	7	0	0
Ligne 2	3	100	70	4	0
Ligne 3	2	32	200	35	1
Ligne 4	1	6	80	100	2
Ligne 5	0	3	5	250	5

On obtient la projection ci-dessous sur le premier plan factoriel.



5.5 Interprétation des résultats de l'AFC

5.5.1 Valeurs propres

On note tout d'abord que la la première valeur propre est une valeur propre triviale égale à 1. En général les logiciels l'ignorent.

On rappelle qu'on note

$$c_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}.$$

On remarque alors que

$$d^2_{\chi^2}(\mathbf{x}, \mathbf{E}) = \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2 = \mathrm{tr}(CC^T) = \sum_{k=1}^{\min(p-1,q-1)} \lambda_k$$

ce qui montre que la décomposition en valeurs singulières de C décompose le χ^2 total de même qu'en ACP on décompose l'inertie totale. La somme des valeurs propres non triviales multipliée par l'effectif total peut se comparer à un quantile de la loi du χ^2 à $(p-1)(q-1)$ degrés de liberté. La somme de toutes les valeurs propres est égale à l'inertie totale, c'est à dire à la distance $d^2(x, E)$. Elle donne donc une information sur l'écart à l'indépendance et on peut la comparer aux quantiles de la loi du χ^2.

Interprétation des valeurs propres -

- Si une valeur propre est proche de un, ça traduit le fait qu'il existe deux sous groupes de modalités dans les données. Il est alors intéressant de reconstruire la matrice N pour mettre en évidence ces deux sous groupes et de réaliser des AFC indépendamment sur les deux sous groupes.
  Par exemple l'analse factorielle des correspondances du tableau 1, renvoie les valeurs propres suivantes : 0.90, 0.01, 7e-3.

- De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités, les effectifs importants se trouvent alors le long de la diagonale.

Choix de la dimension - Comme en ACP, les valeurs propres peuvent être interprétées comme la proportion d'inertie expliquée par le facteur correspondant. On peut s'en servir pour aider au choix de la dimension $r < min(1-p, 1-q)$ de l'espace de projection. En pratique, on utilise le fait que

$$K_r = \sum_{i=1}^{p}\sum_{j=1}^{q} \left(\frac{n_{ij} - \widehat{n_{ij}^r})^2}{\widehat{n_{ij}^r}}\right. \simeq \sum_{k=r+1}^{\min(1-p,1-q)} \lambda_k$$

suit approximatiement une loi du χ^2 à $(p-r-1)(q-r-1)$ degrés de liberté. On peut donc retenir pour valeur de r la plus petite dimension pour laquelle K_r est inférieure à la valeur limite de cette loi. Le choix $r = 0$ correspond à la situation où les variables sont proches de l'indépendance en probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.

Dans l'exemple des étudiants en première année, on obtient le tableau de valeurs propres suivant :

Valeurs propres	8.24e-02	1.70e-03	5.40e-04	1.52e-34
Proportions	0.973	0.02	0.00	0.00
Prop. cumulées	0.973	0.994	1.00	1.00

On en déduit que le premier plan fatoriel explique presque toute l'inertie de la table de contingence. C'est souvent le cas en AFC.

5.5.2 Contribution des modalités

Pour chaque modalité de X_1 (resp. de X_2), la qualité de sa représentation en dimension r se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^p$ (resp. dans $\mathbb{R}^q$) et sa projection D_C^{-1}?orthogonale (resp. D_L^{-1}?orthogonale) dans le sous?espace principal de dimension r. Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de C_L (resp. de C_C).

AUtrement dit, la "qualité" de la représentation d'une modalité contribution de la modalité i de la variable X sur l'axe k est donnée par le cosinus carré de l'angle formé avec l'axe.

$$\cos_k^2(i) = \frac{d_k^2(i,G)}{d^2(i,G)}$$

avec G le centre de gravité et $d^2(i,G) = \sum_k d_k^2(i,G)$

5.5.3 Interprétation en terme de reconstruction des effectifs

La décomposition de la matrice $\mathbf{N}$ est (formule $X = CU^T M^{-1}$))

$$\mathbf{N} = \frac{1}{n} D_L \left(1 + \sum_{k=2}^{r} c_k u_k^T\right) D_C$$

où 1 est la matrice de 1. La terme de première approximation, $n^{-1} D_L 1 D_C$, correspond à deux variables indépendantes. Si on approche par les trois premiers axes :

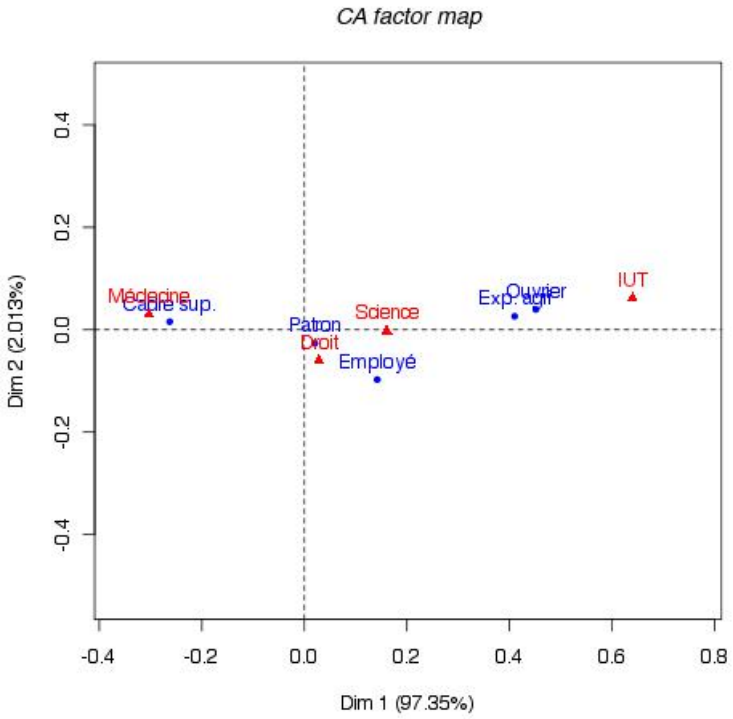
$$n_{ij} \simeq \frac{n_{i\bullet} n_{\bullet j}}{n} \left(1 + \frac{1}{\lambda_1} c_1(i) d_1(j) + \frac{1}{\lambda_2} c_2(i) d_2(j)\right) \tag{5.1}$$

5.6 Exemple

Dans l'exemple des étudiants de première année, on obtient le graphique ci-dessous. D'autre part, on obtient les contributions suivantes des modalités aux axes factoriels

En ligne		
	Dim 2	Dim 3
Exp. agri	16.29	3.22
Patron	0.07	6.30
Cadre sup.	40.40	6.89
Employé	3.02	68.63
Ouvrier	40.21	14.95

En colonne		
	Dim 2	Dim 3
Droit	0.26	58.76
Science	7.94	0.11
Médecine	41.58	19.26
IUT	50.21	21.87



Récapitulatif

Dans $\mathbb{R}^p$ (Lignes)		Dans $\mathbb{R}^q$ Colonnes)
$S = N^T D_L^{-1} N D_C^{-1}$ $Su_k = \lambda_k u_k$	Matrice à diagonaliser Axe facoriel	$T = N D_C^{-1} N^T D_L^{-1}$ $Tv_k = \lambda_k v_k$
$\psi_k = D_L^{-1} N D_C^{-1} u_k$ $\psi_{ki} = \sum_{j=1}^{p} \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}} u_{ki}$	Coordonnées	$\phi_k = D_C^{-1} N^T D_L^{-1} v_k$ $\phi_{ki} = \sum_{i=1}^{q} \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}} v_{ki}$

Chapter 6

Analyse des Correspondances Multiples

6.1 Introduction

L'analyse factorielle des correspondances multiples (ACM ou AFCM) est la généralisation de l'analyse des correspondances au cas de plusieurs variables. Elle consiste donc à représenter les modalités de variables qualitatives dans un espace euclidien dans lequel les distances du χ^2 entre deux modalités d'une même variable sont préservées au mieux. On considère donc dans cette section p variables qualitatives observées simultanément sur n individus de poids identiques $1/n$.

Exemple - Considérons le jeu données de la Table 6.1 dans lequel on caractérise différentes races de chien en fonction de 7 variables portant sur des caractéristiques de physique, sur des points de caractère et une variable d'utilité.
La plupart les tableaux et figures liées à cet exemple sont empruntées à M. Tenenhaus.

6.2 Definitions et notations

6.2.1 Tableau disjonctif complet

Il est difficile de travailler directement avec un tableau de données comme celui de l'utilité des races de chien. En effet, on ne peut pas considérer ces données comme des données quantitatives. Par exemple, ça n'a pas de sens de considérer qu'il y a une distance équivalente entre les classes - et + de la variable Poids et de la variable Intelligence. En conséquence, il est d'usage de recoder les données et de construire le *tableau disjonctif complet*.

Le *tableau disjonctif complet* est tel que chaque ligne correspond à un individu et chaque colonne à une modalité. On note K le nombre total de modalités. Et les observations x_{ij} sont codées 1 si l'individu i a la modalité j et 0 sinon. Notons X le tableau disjonctif complet.

Dans l'exemple, on obtient alors le tableau de la Table 6.2.

6.2.2 Tableau de Burt

On appelle tableau de Burt le tableau $\mathcal{B} = X^T X$. On peut écrire $\mathcal{B} = (B_{k,k'})_{k,k'=1,\cdots,p}$ où

	Race	Taille	Poids	Vitesse	Intell.	Affect.	Agress.	Fonction
1	Beauceron	TA++	PO+	V++	INT+	AF+	AG+	Utilité
2	Basset	TA-	PO-	V-	INT-	AF-	AG+	Chasse
3	Berger-Allemand	TA++	PO+	V++	INT++	AF+	AG+	Utilité
4	Boxer	TA+	PO+	V+	INT+	AF+	AG+	Compagnie
5	Bull-Dog	TA-	PO-	V-	INT+	AF+	AG-	Compagnie
6	Bull-Mastiff	TA++	PO++	V-	INT++	AF-	AG+	Utilité
7	Caniche	TA-	PO-	V+	INT++	AF+	AG-	Compagnie
8	Chihuahua	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
9	Cocker	TA+	PO-	V-	INT+	AF+	AG+	Compagnie
10	Colley	TA++	PO+	V++	INT+	AF+	AG-	Compagnie
11	Dalmatien	TA+	PO+	V+	INT+	AF+	AG-	Compagnie
12	Doberman	TA++	PO+	V++	INT++	AF-	AG+	Utilité
13	Dogue Allemand	TA++	PO++	V++	INT-	AF-	AG+	Utilité
14	Epagneul Breton	TA+	PO+	V+	INT++	AF+	AG-	Chasse
15	Epagneul Français	TA++	PO+	V+	INT+	AF-	AG-	Chasse
16	Fox-Hound	TA++	PO+	V++	INT-	AF-	AG+	Chasse
17	Fox-Terrier	TA-	PO-	V+	INT+	AF+	AG+	Compagnie
18	Grd Bleu de Gascogne	TA++	PO+	V+	INT-	AF-	AG+	Chasse
19	Labrador	TA+	PO+	V+	INT+	AF+	AG-	Chasse
20	Lévrier	TA++	PO+	V++	INT-	AF-	AG-	Chasse
21	Mastiff	TA++	PO++	V-	INT-	AF-	AG+	Utilité
22	Pékinois	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
23	Pointer	TA++	PO+	V++	INT++	AF-	AG-	Chasse
24	Saint-Bernard	TA++	PO++	V-	INT+	AF-	AG+	Utilité
25	Setter	TA++	PO+	V++	INT+	AF-	AG-	Chasse
26	Teckel	TA-	PO-	V-	INT+	AF+	AG-	Compagnie
27	Terre-Neuve	TA++	PO++	V-	INT+	AF-	AG-	Utilité

Table 6.1: Caractéristiques (physique, caractère, utilité) de différentes races de chien

Race	T-	T+	T++	P-	P+	P++	V-	V+	V++	I-	I+	I++	Af-	Af+	Ag-	Ag+
Beauceron	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	1
Basset	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
Berger all	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1
Boxer	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1
Bull-dog	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0
Bull Mastiff	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1
Caniche	1	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0
Chihuahua	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0
Cocker	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	1
Colley	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0
Dalmatien	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0
Doberman	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1
Dogue all	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1
Epagneul br	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0
Epagneul fr	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0
Fox-Hound	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1
Fox-Terrier	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1
Grd Bl de G	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1
Labrador	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0
Lévrier	0	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0
Mastiff	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1
Pékinois	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0
Pointer	0	0	1	0	1	0	0	0	1	0	0	1	1	0	1	0
St-Bernard	0	0	1	0	0	1	1	0	0	0	1	0	1	0	0	1
Setter	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0
Teckel	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0
Terre neuve	0	0	1	0	0	1	1	0	0	0	1	0	1	0	1	0

Table 6.2: tableau disjonctif complet des caractéristiques (physique, caractère, utilité) de différentes races de chien.

- p est nombre total de variables
- si $k \neq k'$, $B_{k,k'}$ est la table de contingence des variables X_k et $X_{k'}$,
- si $k = k'$, B_{kk} est une matrice diagonale contenant les effectifs marginaux de X_k dans la diagonale, notés $n^k_{c_1}, \cdots, n^k_{c_k}$.

Propriétés :

- $\mathcal{B}$ est symétrique.
- La somme des lignes (resp. des colonnes) de $\mathcal{B}$ est $pn^k_l, l = c_1, \cdots, c_k$.
- La somme des éléments de $\mathcal{B}$ est $p^2 n$.

Remarque : si on considère les données du tableau disjonctif X comme des observations de variables qualitatives, alors le tableau de Burt représente la variance de X à un facteur multiplicatif près.

Dans l'exemple des chiens, le tableau de Burt prend la forme suivante. On observe que la diagonale représente les profils (ou distribution en effectif) des différentes variables tandis que les termes extra diagonaux donnent les effectifs croisés entre deux modalités.

TABLEAU DE BURT

	TA-	TA+	TA++	PO-	PO+	PO++	VE-	VE+	VE++	INT-	INT+	INT++	AF-	AF+	AG-	AG+	Comp	Chas	Util
TA-	7	0	0																
TA+	0	5	0																
TA0++	0	0	15																
PO-	7	1	0	8	0	0													
PO+	0	4	10	0	14	0													
PO++	0	0	5	0	0	5													
VE-	5	1	4	6	0	4	10	0	0										
VE+	2	4	2	2	6	0	0	8	0										
VE++	0	0	9	0	8	1	0	0	9										
IN-	3	0	5	3	3	2	4	1	3	8	0	0							
IN+	3	4	6	4	7	2	5	5	3	0	13	0							
IN++	1	1	4	1	4	1	1	2	3	0	0	6							
AF-	1	0	12	1	7	5	5	2	6	6	4	3	13	0					
AF+	6	5	3	7	7	0	5	6	3	2	9	3	0	14					
AG-	5	3	6	5	8	1	5	5	4	3	8	3	5	9	14	0			
AG+	2	2	9	3	6	4	5	3	5	5	5	3	8	5	0	13			
Comp	6	3	1	7	3	0	5	4	1	2	7	1	0	10	7	3	10	0	0
Chas	1	2	6	1	8	0	1	4	4	4	3	2	7	2	6	3	0	9	0
Util	0	0	8	0	3	5	4	0	4	2	3	3	6	2	1	7	0	0	8
	TA-	TA+	TA++	PO-	PO+	PO++	VE-	VE+	VE++	INT-	INT+	INT++	AF-	AF+	AG-	AG+	Comp	Chas	Util

6.2.3 Tableau des χ^2

Avant d'aller plus loin et pour aider à l'interprétation des résultats qu'on obtiendra par la suite, il est utile de générer aussi le tableau des statistiques du χ^2 entre les différentes variables. Ce tableau garde un sens, en effet la distance entre deux modalités j et j'

$$d^2(j, j') = \sum_{i=1}^{n} n \left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{ij'}}{x_{\bullet j'}} \right)^2$$

Ainsi deux modalités choisies par les mêmes individus coïncident. Par ailleurs, les modalités de faible effectif sont éloignées des autres.

La distance entre deux individus i et i' s'exprime

$$d^2(i,i') = \frac{1}{p}\sum_{j=1}^{K} \frac{n}{x_{\bullet j}} \left(x_{ij} - x_{i'j}\right)^2$$

Deux individus sont proches s'ils ont répondu de la même manière.

Dans le tableau, les chiffres entre parenthèses représentent les degrés de significativité (p-value) du test du χ^2. On remarque par exemple que la taille et le poids sont liés à la vélocité tandis que seule l'agressivité est lié à la fonction.

	Poids	Vélocité	Intelligence	Affection	Agressivité	Fonction
Taille	25.3 (.000)	15.9 (.000)	3.6 (.46)	14.0 (.001)	2.1 (.36)	16.35 (.003)
Poids		18.4 (.001)	1.35 (.85)	9.5 (.008)	2.6 (.28)	24.41 (.000)
Vélocité			3.16 (.53)	3.0 (.23)	.57 (.75)	8.49 (.08)
Intelligence				3.9 (.14)	1.15 (.56)	4.14 (.39)
Affection					1.8 (.18)	14.76 (.000)
Agressivité						7.07 (.03)

6.3 Analyse Factorielle des Correspondances Multiples

L'Analyse Factorielle des Correspondances Multiples des variables $x_1, \cdots, x_p$ est l'analyse factorielle des correspondances du tableau disjonctif complet ou du tableau de Burt.

On rappelle les notations définies plus haut.

- n est le nombre d'individus.
- On a p variables qualitatives.
- La variable X_j admet $n_{c_j}^j$ modalités.
- $K = n_{c_1}^1 + \cdots + n_{c_j}^j$ est le nombre total de modalités.
- La modalité jl a une fréquence absolue $n_{jl} = n_l^j$ et une fréquence relative $\frac{n_{jl}}{np}$ dans le tableau de Burt.

6.3.1 AFC du tableau disjonctif complet relatif à 2 variables

On note toujours X_1 et X_2 les 2 variables qualitatives et on note r et c leur nombre respectif de modalités. Les matrices intervenant dans l'AFC usuelle sont reprises ici selon les mêmes notations que dans le chapitre précédent mais surlignées. Ici, D_L (reps. D_C) est la matrice

diagonale qui contient les profils lignes (resp. colonnes) en fréquence.

$$\begin{aligned}
\bar{N} &= X = [X_1|X_2] \\
\bar{D}_L &= \frac{1}{n}\mathbb{I}_n \\
\bar{D}_C &= \frac{1}{2}\begin{bmatrix} D_L & 0 \\ 0 & D_C \end{bmatrix} = \frac{1}{2}\Delta \\
\bar{A} &= \frac{1}{2n}\bar{N}^T\bar{D}_L^{-1} = \frac{1}{2}X^T, \text{ avec } N \text{ la table de contingence.} \\
\bar{B} &= \frac{1}{2n}\bar{N}\bar{D}_C^{-1} = \frac{1}{2}X\Delta^{-1}
\end{aligned}$$

L'AFC est considérée comme une double ACP : celle des profils lignes de $\bar{A}$ puis celle des profils colonne de $\bar{B}$.

Proposition 6. - *ACP des profils lignes*
L'ACP des profils lignes issue de l'AFC réalisée sur le tableau disjonctif complet relatif à 2 variables qualitatives conduit à l'analyse spectrale de la matrice $\bar{D}_C^{-1}$-symétrique et positive :

$$\bar{A}\bar{B} = \frac{1}{2}\begin{bmatrix} \mathbb{I}_r & B \\ A & \mathbb{I}_c \end{bmatrix}.$$

Les $r+c$ valeurs propres de $\bar{A}\bar{B}$ s'écrivent

$$\mu_k = \frac{1 \pm \sqrt{\lambda_k}}{2}$$

où les $|\lambda_k$ sont les valeurs propres de la matrice AB (celle de l'AFC classique de X).
Les vecteurs propres $\bar{D}_C^{-1}$-orthonormés associés peuvent se mettre sous la forme

$$\bar{V} = \frac{1}{2}\begin{bmatrix} U \\ V \end{bmatrix}$$

où U et V sont les matrices de vecteurs propres obtenues en faisant l'AFC de la table de contingence associée à X_1 et X_2.
La matrice des composantes principales s'écrit

$$\bar{C}_L = \frac{1}{2}[X_1C_L + X_2C_C]\Lambda^{-1/2}$$

où C_L et C_C sont les matrices de l'AFC classique.

Dans la pratique on ne considère que les $d = \inf(r-1, c-1)$ plus grandes valeurs propres différentes de 1.

$$M = diag(\mu_1, \cdots, \mu_d) = \frac{1}{2}[\mathbb{I}_d + \Lambda^{1/2}]$$

Les autres valeurs propres non nulles sont des artéfacts liés à la construction de la matrice à diagonaliser. Elles n'ont donc pas de sens statistique.

Proposition 7. - *ACP des profils colonnes*
L'ACP des profils colonnes issue de l'AFC réalisée sur le tableau disjonctif complet relatif à 2 variables qualitatives conduit à l'analyse spectrale de la matrice $\bar{D}_L^{-1}$-symétrique et positive :

$$\bar{B}\bar{A} = \frac{1}{2n}\left[X_1D_L^{-1}X_1^T + X_2D_C^{-1}X_2^T\right]$$

Les $r+c$ valeurs propres non nulles de $\bar{B}\bar{A}$ sont les μ_k. Les vecteurs propres $\bar{D}_L^{-1}$-orthonormés associés peuvent se mettre sous la forme

$$\bar{U} = \frac{1}{n}\bar{C}_L M^{-1/2}.$$

La matrice des composantes principales s'écrit

$$\bar{C}_C = \begin{bmatrix} C_L \\ C_C \end{bmatrix} \Lambda^{-1/2} M^{1/2}.$$

L'AFC du tableau disjonctif complet permet, grâce aux coordonnées contenues dans $\bar{C}_C$, la représentation simultanée des modalités des deux variables. Cette représentation est très proche de celle de l'AFC classique. De plus cette approche permet une représentation des individus avec les coordonnées de la matrice $\bar{C}_L$. A un facteur près, l'individu apparait comme le barycentre des deux modalités qu'il a présentées.

6.3.2 AFC du tableau disjonctif complet

Comme dans le cas où $p = 2$, on reprend les notations de l'AFC classique en les surlignant

$$\begin{aligned} \bar{T} &= X = [X_1|\cdots|X_p] \\ \bar{D}_L &= \frac{1}{n}\mathbb{I}_n \\ \bar{D}_C &= \frac{1}{p}\Delta \\ \bar{A} &= \frac{1}{p}X^T \\ \bar{B} &= \frac{1}{n}X\Delta^{-1} \end{aligned}$$

Proposition 8. - *ACP des profils lignes*
L'ACP des profils lignes issue de l'AFC réalisée sur le tableau disjonctif complet relatif à p variables qualitatives conduit à l'analyse spectrale de la matrice $\bar{D}_C^{-1}$-symétrique et positive :

$$\bar{A}\bar{B} = \frac{1}{np}\mathcal{B}\Delta^{-1}$$

Il y a m ($m \leq c-p$) valeurs propres notées μ_k comprises entre 0 et 1 rangées dans la matrice diagonale M. La matrice des vecteurs propres $\bar{D}_C^{-1}$-orthonormés associés se décompose par blocs de la façon suivante

$$\bar{V} = \begin{bmatrix} V_1 \\ \cdots \\ V_p \end{bmatrix}$$

La matrice des composantes principales s'écrit

$$\bar{C}_L = \sum_{j=1}^{p} X_j D_j^{-1} V_j$$

Comme dans le cas où $p = 2$, chaque individu est positionné au barycentre des modalités qu'il a représentée. De plus, il faut noter que les modalités d'une même variable sont centrées : les facteurs opposent les modalités d'une même variable.

Proposition 9. - *ACP des profils colonnes*
L'ACP des profils lignes issue de l'AFC réalisée sur le tableau disjonctif complet relatif à p variables qualitatives conduit à l'analyse spectrale de la matrice $\bar{D}_L^{-1}$-symétrique et positive :

$$\bar{B}\bar{A} = \frac{1}{np}\sum_{j=1}^{p} X_j D_j^{-1} X_j^T$$

La matrice des vecteurs propres $\bar{D}_L^{-1}$-orthonormés vérifie

$$\bar{U} = \bar{B}\bar{V}M^{-1/2}$$

La matrice des composantes principales s'écrit

$$\bar{C}_C = p\Delta^{-1}\bar{V}M^{1/2}$$

Chaque bloc C_j de $\bar{C}_C$ fournit en lignes les coordonnées des modalités de la variable X_j et permet la repsésentation graphique simultanée.

6.3.3 AFC du tableau de Burt

Cas où $p = 2$

Prenons le cas où $p = 2$ et étudions ce que donne, dans ce cas, l'AFC du tableau de Burt. On se rappelle que l'AFC est une double ACP sur les profils-ligne d'une part et sur les profils-colonne d'autre part. Le tableau de Burt est symétrique, les profils ligne et colonne sont identiques : on s'intéresse donc à une seule des ACP.

On note

$$\tilde{T} = \mathcal{B} = \begin{bmatrix} nD_L & N \\ N^T & nD_C \end{bmatrix}$$

$$\tilde{D}_L = \tilde{D}_c = \frac{1}{2}\begin{bmatrix} D_L & 0 \\ 0 & D_C \end{bmatrix} = \frac{1}{2}\Delta = \bar{D}_c$$

$$\tilde{A} = \tilde{B} = \frac{1}{2}\begin{bmatrix} \mathbb{I}_L & B \\ A & \mathbb{I}_C \end{bmatrix} = \bar{A}\bar{B}$$

On fait l'AFC comme l'ACP des profils lignes de $\tilde{A}$.

Proposition 10. *L'ACP des profils-lignes issue de l'AFC réalisée sur le tableau de Burt relatif à deux variables qualitatives conduit à l'analyse spectrale de la matrice $\tilde{D}_C$-symétrique et positive :*

$$\tilde{A}\tilde{B} = (\bar{A}\bar{B})^2.$$

Elle admet pour matrice de vecteurs propres $\tilde{D}_C^{-1}$-orthonormés

$$\tilde{U} = \tilde{V} = \bar{V}$$

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$. La matrice des composantes principales s'écrit

$$\tilde{C}_L = \tilde{C}_C = \begin{bmatrix} C_L \\ C_C \end{bmatrix}\Lambda^{1/2}M.$$

La matrice $\tilde{C}_L$ permet de représenter simultanément les modalités des deux variables.

Remarques

- Les différentes AFC présentées ci-dessus conduisent à la même représentation simultanée des modalités des 2 variables.
- Dans l'AFC du tableau disjonctif complet comme dans celle du tableau de Burt, on obtient des valeurs propres non nulles qui n'ont pas de sens statistique. Ainsi les valeurs propres ne peuvent plus être interprétées comme une part d'inertie.
- L'AFC du tableau de Burt ne considère que des croisements de variables deux à deux, si on veut étudier des interactions d'ordre plus élevé, il faut recoder les variables.

Cas où p est quelconque

Le tableau de Burt est symétrique, on ne fera donc qu'une ACP. On note

$$\tilde{T} = \mathcal{B}$$
$$\tilde{D}_L = \tilde{D}_C = \frac{1}{p}\Delta = \bar{D}_C$$
$$\tilde{A} = \tilde{B} = \frac{1}{np}\mathcal{B}\Delta^{-1} = \bar{A}\bar{B}$$

Proposition 11. *L'ACP des profils-lignes issue de l'AFC réalisée sur le tableau de Burt relatif à pvariables qualitatives conduit à l'analyse spectrale de la matrice $\tilde{D}_C$-symétrique et positive :*

$$\tilde{A}\tilde{B} = (\bar{A}\bar{B})^2.$$

Elle admet pour matrice de vecteurs propres $\tilde{D}_C^{-1}$-orthonormés

$$\tilde{U} = \tilde{V} = \bar{V}$$

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$. La matrice des composantes principales s'écrit

$$\tilde{C}_L = \tilde{C}_C = \bar{C}_C M^{1/2}.$$

La matrice $\tilde{C}_L$ permet de représenter simultanément les modalités de toutes les variables. En revanche on ne peut pas faire la représentation des individus quand on fait l'AFC du tableau de Burt.

6.3.4 Interprétation

Comme en ACM, on définit le nuage de points associé aux profils-ligne. L'inertie totale est $K/p - 1$ et la dimension maximum du nuage de points $K - p$. La moyenne des valeurs propres sera égale à

$$\frac{K/p - 1}{K - p} = 1/p$$

et on retient les axes associées à des valeurs propres supérieures à $1/p$; on peut aussi utiliser la règle du coude. Attention, les valeurs propres ne peuvent pas être interprétées comme des parts d'inertie.

Dans l'exemple, l'inertie totale est égale à 1.67 et il y a 3 valeurs propres supérieures à $1/p = 1/6$ (voir Table 6.3).

VALEURS PROPRES
INERTIE TOTALE 1.6667

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
1	0.4816	28.90	28.90
2	0.3847	23.08	51.98
3	0.2110	12.66	64.64
4	0.1576	9.45	74.09
5	0.1501	9.01	83.10
6	0.1233	7.40	90.50
7	0.0815	4.89	95.38
8	0.0457	2.74	98.12
9	0.0235	1.41	99.54
10	0.0077	0.46	100.00

Table 6.3: Valeurs propres associées à l'AFCM des caractéristiques de différentes races de chiens.

6.3.5 Représentation des individus

La variance de l'axe h est λ_h ce qui est classique en analyse factorielle et la contribution de l'individu i à l'axe h est donnée par

$$\frac{\frac{1}{n}c_{ih}}{\lambda_h}$$

où c_{ih} est la coordonnée de l'individu i sur l'axe h (voir 6.4).

6.3.6 Représentation des variables

L'inertie apportée par une modalité jl au nuage de points est

$$\frac{1}{p}\left(1-\frac{n_{jl}}{n}\right)$$

Elle est donc d'autant plus forte que l'effectif de la modalité est faible. De nombreuses modalités à faible effectif peuvent donc déséquilibrer une AFCM. Et il est préférable de limiter le nombre de modalités à faible effectif, quitte à redéfinir les modalités.

Par ailleurs, l'inertie apportée par une variable j est

$$\frac{c_j-1}{p}$$

Elle est donc d'autant plus importante que le nombre de modalités de la variable est important. Il est donc conseillé de travailler avec des variables ayant des modalités en nombre comparable. La contribution de la modalité k de la variable X_j à l'inertie de l'axe h est donnée par

$$\frac{\frac{n_{jk}}{pn}c_{jk}^2}{\lambda_h}$$

Voir Table 6.5.

En pratique,

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS
AXES 1 A 3

INDIVIDUS			COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
IDENTIFICATEUR	P.REL	DISTO	1	2	3	1	2	3	1	2	3
beauceron	3.70	1.14	-0.32	0.42	0.10	0.8	1.7	0.2	0.09	0.15	0.01
basset	3.70	1.91	0.25	-1.10	0.19	0.5	11.7	0.6	0.03	0.63	0.02
berger allemand	3.70	1.54	-0.49	0.46	0.50	1.8	2.1	4.4	0.15	0.14	0.16
boxer	3.70	1.80	0.45	0.88	-0.69	1.5	7.5	8.4	0.11	0.43	0.27
bull-dog	3.70	1.64	1.01	-0.55	0.16	7.9	2.9	0.5	0.62	0.18	0.02
bull-mastiff	3.70	2.09	-0.75	-0.55	-0.50	4.4	2.9	4.3	0.27	0.14	0.12
caniche	3.70	2.16	0.91	0.02	0.58	6.4	0.0	5.8	0.39	0.00	0.15
chihuahua	3.70	1.86	0.84	-0.84	0.47	5.4	6.9	3.9	0.38	0.38	0.12
cocker	3.70	1.93	0.73	-0.08	-0.66	4.1	0.1	7.7	0.28	0.00	0.23
colley	3.70	1.11	-0.12	0.53	0.33	0.1	2.7	2.0	0.01	0.25	0.10
dalmatien	3.70	1.77	0.65	0.99	-0.46	3.2	9.4	3.7	0.24	0.55	0.12
doberman	3.70	1.56	-0.87	0.32	0.45	5.9	1.0	3.6	0.49	0.06	0.13
dogue allemand	3.70	1.95	-1.05	-0.51	-0.17	8.4	2.5	0.5	0.56	0.13	0.01
epagneul breton	3.70	2.18	0.48	1.04	-0.06	1.8	10.4	0.1	0.10	0.49	0.00
epagneul français	3.70	1.20	-0.14	0.52	-0.12	0.2	2.6	0.2	0.02	0.22	0.01
fox-hound	3.70	1.38	-0.88	-0.03	0.36	5.9	0.0	2.3	0.56	0.00	0.10
fox-terrier	3.70	1.78	0.88	-0.14	-0.05	6.0	0.2	0.1	0.44	0.01	0.00
grand bleu de gascogne	3.70	1.44	-0.52	0.11	-0.04	2.1	0.1	0.0	0.19	0.01	0.00
labrador	3.70	1.77	0.65	0.99	-0.46	3.2	9.4	3.7	0.24	0.55	0.12
levrier	3.70	1.35	-0.68	0.08	0.60	3.5	0.1	6.2	0.34	0.01	0.26
mastiff	3.70	1.90	-0.76	-0.89	-0.59	4.4	7.6	6.1	0.30	0.41	0.18
pékinois	3.70	1.86	0.84	-0.84	0.47	5.4	6.9	3.9	0.38	0.38	0.12
pointer	3.70	1.54	-0.67	0.42	0.69	3.5	1.7	8.3	0.29	0.12	0.31
saint-bernard	3.70	1.69	-0.58	-0.59	-0.89	2.6	3.4	14.0	0.20	0.21	0.47
setter	3.70	1.14	-0.50	0.38	0.29	2.0	1.4	1.5	0.22	0.13	0.07
teckel	3.70	1.64	1.01	-0.55	0.16	7.9	2.9	0.5	0.62	0.18	0.02
terre-neuve	3.70	1.66	-0.38	-0.49	-0.66	1.1	2.3	7.7	0.09	0.14	0.26

Table 6.4: Coordonnées et contributions des individus (modalités) associées à l'AFCM des caractéristiques de différentes races de chiens.

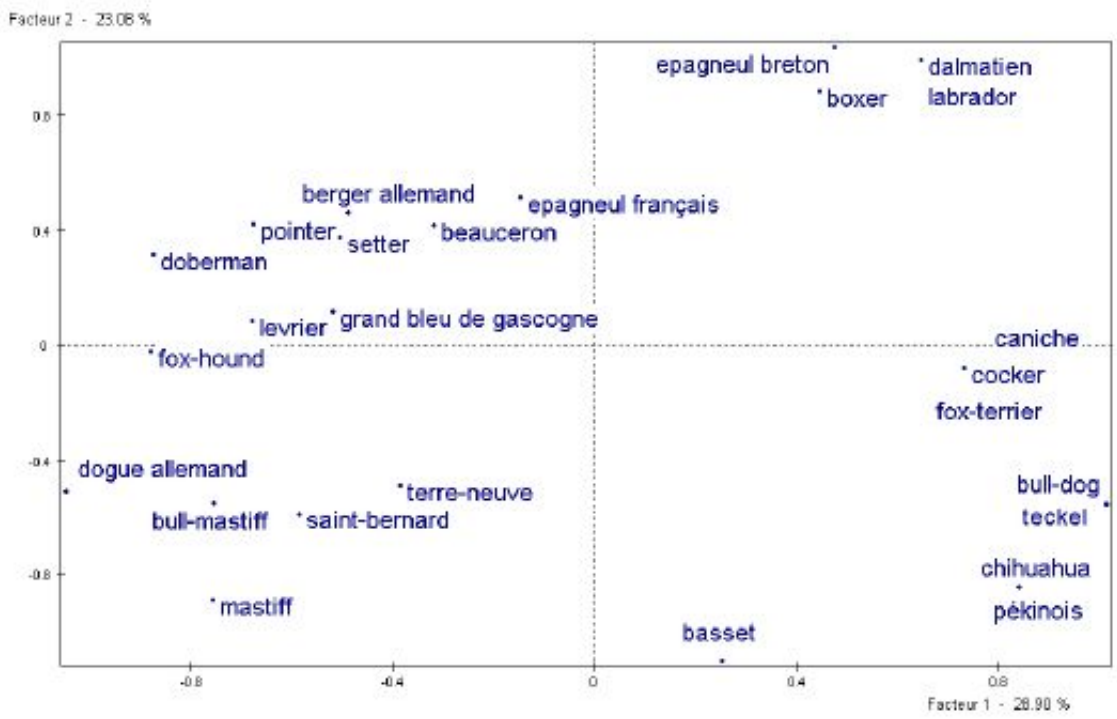


Figure 6.1: Représentation des différentes races de chiens (individus).

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES
AXES 1 A 3

MODALITES			COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
LIBELLE	P.REL	DISTO	1	2	3	1	2	3	1	2	3
1 . Taille											
ta-	4.32	2.86	1.18	-0.92	0.62	12.6	9.6	7.8	0.49	0.30	0.13
ta+	3.09	4.40	0.85	1.23	-1.02	4.6	12.2	15.1	0.16	0.34	0.23
ta++	9.26	0.80	-0.84	0.02	0.05	13.5	0.0	0.1	0.88	0.00	0.00
			CONTRIBUTION CUMULEE =			30.7	21.8	23.0			
2 . Poids											
po-	4.94	2.38	1.17	-0.82	0.36	14.0	8.7	3.0	0.58	0.29	0.05
po+	8.64	0.93	-0.31	0.82	0.23	1.7	15.1	2.2	0.10	0.72	0.06
po++	3.09	4.40	-1.02	-0.97	-1.22	6.6	7.6	21.8	0.23	0.22	0.34
			CONTRIBUTION CUMULEE =			22.3	31.4	27.0			
3 . Vélocité											
v-	6.17	1.70	0.32	-1.04	-0.40	1.3	17.5	4.7	0.06	0.64	0.09
v+	4.94	2.38	0.60	0.89	-0.36	3.7	10.1	3.0	0.15	0.33	0.050
v++	5.56	2.00	-0.89	0.37	0.76	9.2	2.0	15.3	0.40	0.07	0.29
			CONTRIBUTION CUMULEE =			14.2	29.6	23.0			
4 . Intelligence											
int-	4.94	2.38	-0.35	-0.81	0.35	1.2	8.4	2.9	0.05	0.28	0.05
int+	8.02	1.08	0.37	0.29	-0.49	2.3	1.7	9.3	0.13	0.08	0.23
int++	3.70	3.50	-0.34	0.46	0.60	0.9	2.0	6.3	0.03	0.06	0.10
			CONTRIBUTION CUMULEE =			4.4	12.1	18.5			
5 . Affection											
af-	8.02	1.08	-0.84	-0.29	-0.07	11.6	1.7	0.2	0.65	0.08	0.00
af+	8.64	0.93	0.78	0.27	0.06	10.8	1.6	0.2	0.65	0.08	0.00
			CONTRIBUTION CUMULEE =			22.4	3.3	0.3			
6 . Aggressivité											
ag-	8.64	0.93	0.40	0.19	0.31	2.9	0.8	3.9	0.17	0.04	0.10
ag+	8.02	1.08	-0.43	-0.21	-0.33	3.1	0.9	4.2	0.17	0.04	0.10
			CONTRIBUTION CUMULEE =			6.0	1.8	8.2			

Table 6.5: Coordonnées et contributions des variables associées à l'AFCM des caractéristiques de différentes races de chiens.

- On interprète les proximités et les oppositions entre les modalités des différentes variables
- On privilégie les interprétations sur les modalités suffisamment éloignées du centre du graphique
- Les rapports de valeurs propres ne sont pas interprétables mais on peut regarder la décroissance des valeurs propres pour choisir la dimension.
- Seules les contributions des modalités à l'inertie selon les axes sont interprétables

Une modalité jl a une position sur l'axe h qui est significativement différente du centre de gravité 0 si

$$\left| C_h(jl) \sqrt{\frac{n_{jl}(n-1)}{n-n_{jl}}} \right| > 2$$

où $C_h(jl)$ est la coordonnée de la modalité l de la variable j sur l'axe h. Dans l'exemple des chiens ont obtient le tableau ci-dessous. On observe par exemple que seule la modalité `int-` de l'intelligence est significative et seulement sur l'axe 2.

Le premier axe factoriel oppose les chiens de grande taille (à gauche) aux chiens de petite taille (à droite), il oppose aussi l'affection faible (à gauche) à l'affectivité (à droite). L'axe deux oppose les chiens lents et légers en bas aux chiens assez rapides en haut et gros.

COORDONNEES ET VALEURS-TEST DES MODALITES
AXES 1 A 3

MODALITES	VALEURS-TEST		
LIBELLE	1	2	3
1 . Taille			
ta-	3.6	-2.8	1.9
ta+	2.1	3.0	-2.5
ta++	-4.8	0.1	0.3
2 . Poids			
po-	3.9	-2.7	1.2
po+	-1.6	4.3	1.2
po++	-2.5	-2.4	-3.0
3 . Vélocité			
v-	1.3	-4.1	-1.6
v+	2.0	2.9	-1.2
v++	-3.2	1.3	2.8
4 . Intelligence			
int-	-1.2	-2.7	1.2
int+	1.8	1.4	-2.4
int++	-0.9	1.3	1.6
5 . Affection			
af-	-4.1	-1.4	-0.3
af+	4.1	1.4	0.3
6 . Aggressivité			
ag-	2.1	1.0	1.6
ag+	-2.1	-1.0	-1.6

Table 6.6: Coordonnées et valeurs tests des associées à l'AFCM des caractéristiques de différentes races de chiens.

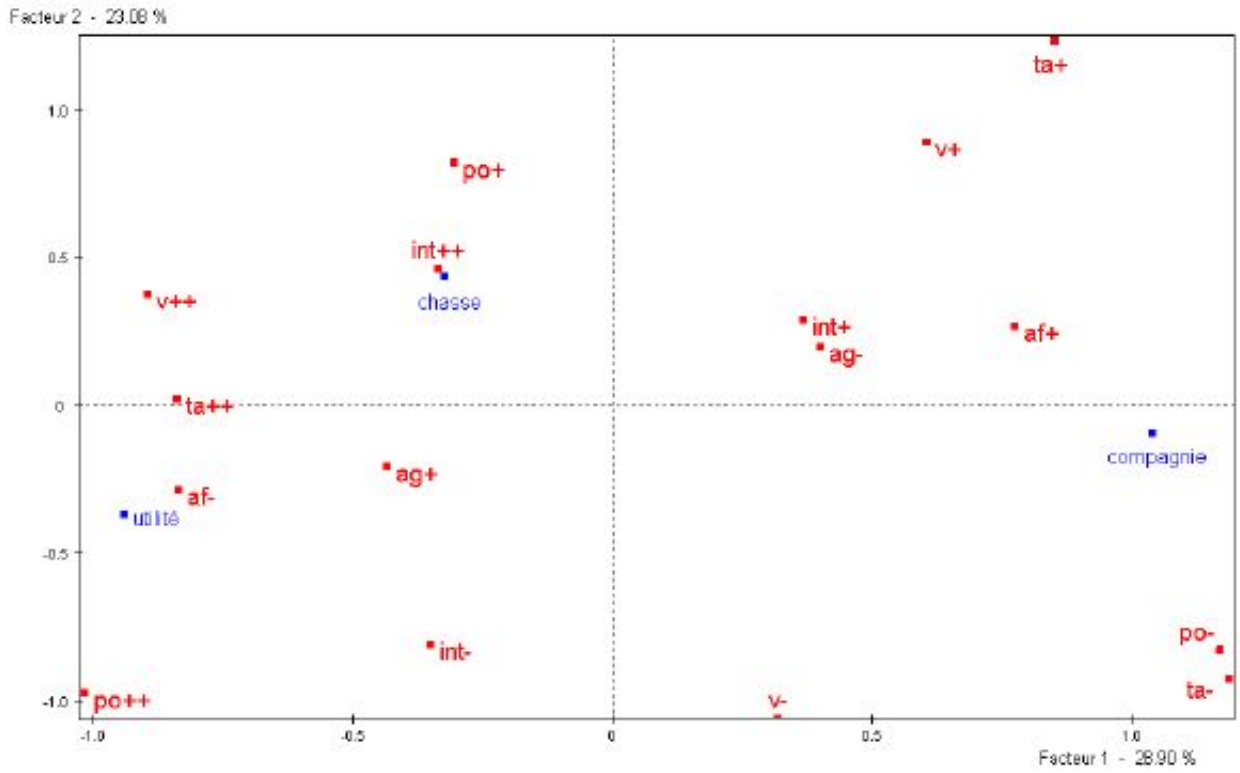


Figure 6.2: Représentation des caractéristiques (variables).

6.3.7 Représentation simultanée

En AFCM, la représentation simultanée est assez naturelle puisqu'on cherche à représenter les individus au centre des modalités qu'il ont choisies et les madalités au centre des individus qui les ont choisies, aux facteurs de dilatation $1/\sqrt{\lambda}$ près.

Quand les individus sont nombreux et anonymes cette représentation n'est pas très utile.

Pour l'exemple des races de chiens, nous obtenons la figure 6.3.

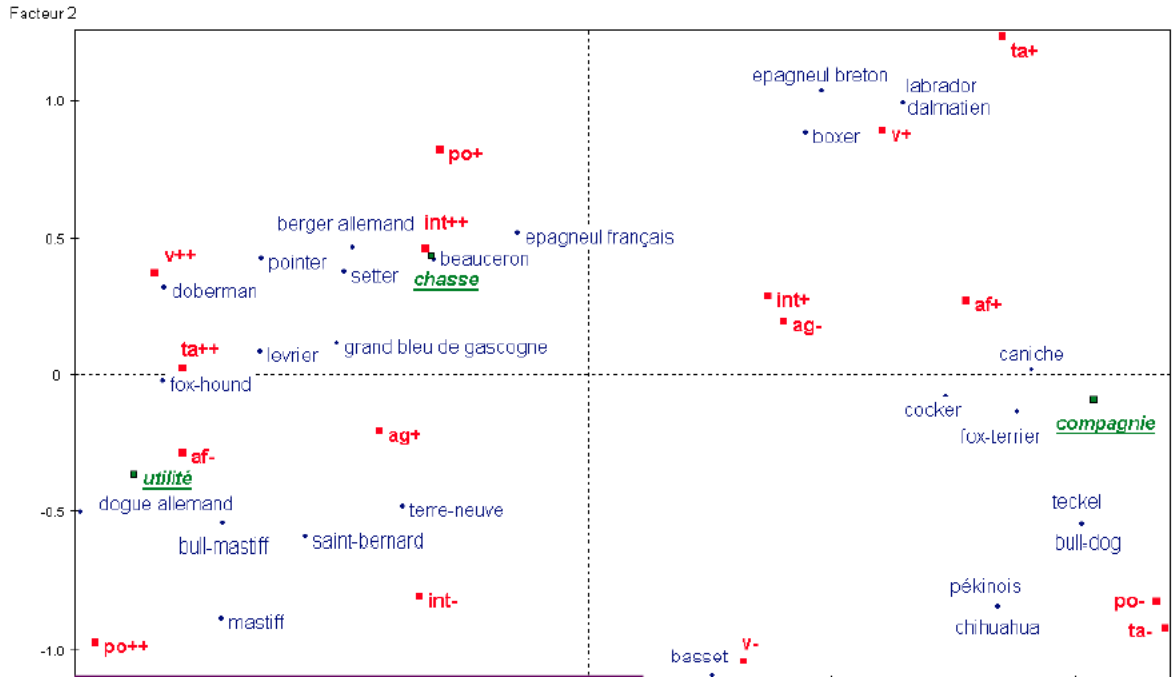


Figure 6.3: AFCM des données sur les races de chiens, la variable d'utilité (en bleu) est considérée comme une variable illustrative (ou supplémentaire).

6.4 Individus et variables suplémentaires

La prise en compte d'éléments supplémentaires permet d'aider à la compréhension et à l'interprétation des résultats. Ceci est généralement intéressant quand les variables se décomposent en thèmes. En pratique, les modalités des variables supplémentaires seront prises en compte en projetant, dans les sous espaces factoriels, le centre des groupes d'individus ayant choisi ces modalités. On pourra aussi, dans une optique de prévision, projeter des individus ou des variables supplémentaires pour les situer par rapport aux individus et variables actives.

6.5 Les variables continues

Pour être actives dans une analyse factorielle des correspondances multiples, les variables continues doivent être rendues nominales. On devra s'attacher à retenir un nombre de modalités proche de celui des variables discrètes actives dans l'analyse et à avoir une répartition de l'effectif équilibré entre les différentes modalités (ie faire un découpage équi-réparti). Cependant, on pourra éventuellement retenir une modalité à faible effectif si celle ci est importante pour l'étude, ou

respecter des seuils naturels s'ils existent.
Quand une variable continue est introduite comme variable supplémentaire, on peut avoir intérêt à faire un découpage fin.

On doit être extrêmement vigilant quand on discrétise une variable continue car ceci conduit à une perte d'information brute. Néanmoins, ce type de transformation peut présenter certains avantages :

- ça permet de traiter conjointement des variables continues et discrètes en correspondances multiples,
- ça permet d'observer une éventuelle contiguïté de classes voisines et de valider a posteriori les données,
- et surtout ça permet de mettre en évidence d'éventuelles liaisons non linéaires entre les variables continues.

Chapter 7

Classification non supervisée

7.1 Introduction

La classification[1] ou segmentation recouvre l'ensemble des méthodes permettant de regrouper des individus selon leurs similarités en un nombre fini de classes. On constitue ainsi une partition des individus. De façon un peu schématique, on peut dire qu'il existe deux grandes familles de méthodes de classification :

- les méthodes de classification globale reposant, implicitement, sur la construction d'un modèle probabiliste et l'optimisation de critères globaux comme des rapports de variance par exemple,
- les méthodes itératives reposant sur des regroupements locaux d'individus voisins reposant uniquement sur des notions de distances entre individus (ou groupes d'individus).

Dans la première famille, la méthode la plus connue est la méthode des nuées dynamiques ou centres mobiles tandis que dans la seconde famille la méthode la plus répandue est la méthode de classification hiérarchique ascendante.

7.2 Distances et similarités

Les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

- un tableau de distances (ou dissimilarités, ou mesures de dissemblance) entre les individus pris deux à deux ;
- les observations de p variables quantitatives sur ces n individus ;
- les observations, toujours sur ces n individus, de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se ramener au tableau des distances deux à deux entre les individus (c'est-à-dire au premier cas).

Les mesures de distance ou de similarités sont utilisées pour déterminer la proximité entre des objets (ou individus). Une distance mesure une dissimilarité. les notions de distance et

[1] en anglais : clustering

de similarité sont duales. Par exemple di d_{ij} est la distance entre deux objets$i$ et j alors $d'_{ij} = \max_{i,j} d_{ij} - d_{ij}$ décrit une similarité entre les deux objets.

Définition 8. *Soit $I = \{1, \cdots, n\}$ un ensemble d'indices.*

- *Un indice de ressemblance ou similarité est une mesure s définie de $I \times I$ dans $\mathbb{R}^+$ et vérifiant*

$$\begin{aligned} &s(i,j) = s(j,i) \\ &s(i,i) = S > 0 \\ &s(i,j) \leq S, \ \forall (i,j) \in I \times I \end{aligned}$$

- *Un indice de dissemblance ou dissimilarité est une application d de $I \times I$ dans $\mathbb{R}^+$ vérifiant*

$$\begin{aligned} &d(i,j) = d(j,i) \\ &d(i,i) = 0 \end{aligned}$$

- *Un indice de distance est un indice de dissemblance vérifaint en plus la propriété*

$$d(i,j) = 0 \text{ implique } i = j$$

- *Une distance est une indice de distance vérifiant en plus une inégalité triangulaire*

$$d(i,j) \leq d(i,k) + d(k,j), \forall (i,j) \in I \times I \times I$$

La nature des observations joue un rôle prépondérant dans le choix de la mesure de proximité. Les données nominales vont en général conduite à travailler avec une mesure de similarité alors que les données quantitatives sont plus souvent traitées via des distances.

7.2.1 Similarité entre des objets à structure binaire

Pour mesurer la similarité entre des objets, on compare toujours des paires d'observations (x_i, x_j) avec $x_i^T = (x_{i1}, \cdot, x_{ip})$. Ici on suppose que $x_{ik} \in \{0, 1\}$. Définissons

$$\begin{aligned} a_1 &= \sum_{k=1}^{p} \mathbb{I}(x_{ik} = x_{jk} = 1) \\ a_2 &= \sum_{k=1}^{p} \mathbb{I}(x_{ik} = 0, x_{jk} = 1) \\ a_3 &= \sum_{k=1}^{p} \mathbb{I}(x_{ik} = 1, x_{jk} = 0) \\ a_4 &= \sum_{k=1}^{p} \mathbb{I}(x_{ik} = x_{jk} = 0) \end{aligned}$$

Les mesures de proximité sont en général définies sous la forme :

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$$

où δ et λ sont des pondérations qui permettent de donner plus ou moins d'importance aux ressemblances (présence d'un caractères commun) ou aux différences (absence de caractère commun).

Dans le cas où $\delta = 0$ et $\lambda = 1$ on obtient l'indice de Jaccard, utilisé en génétique et en écologie.

7.2.2 Distance entre des objets à variables nominales

Quand les objets sont décrits par des variables nominales, on se ramène à un tableau disjonctif complet ou à un tableau de contingence. Dans le premier cas, on compare des objets à structure binaire. Dans le second, on travaille généralement avec la distance du χ^2.

7.2.3 Distance entre des objets à variables continues

Une grande variété de distance peut être générée à partir des normes L_r.

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^p \right)^{1/p}$$

Dans cette définition on sous entend que les variables sont toutes mesurées selon la même échelle, ce qui est rarement vrai. Si on a des différences d'échelle en tre les variables, il est usuel d'utiliser une distance corrigée par la variance de chaque variable.

$$d_{ij} = \left(\sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|^2}{\sigma_k^2} \right)^{1/2}$$

7.3 Classification hiérarhique ascendante

La classification hiérarchique ascendante est une méthode itérative qui consiste, à chaque étape, à regrouper les classes les plus proches. A la première étape chaque individu constitue une classe. L'algorithme démarre donc de la partition triviale des n singletons. Et l'algorithme s'arrête avec l'obtention d'une seule classe. Les regroupement successifs sont représentés sous la forme d'un arbre ou dendogramme.

Algorithme de la classification hiérarchique ascendante

Initialisation Les classes initiales sont les singletons. Calculer la matrice des distances 2 à 2.

Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe.

- regrouper les deux classes les plus proches au sens de la distance entre groupe choisie,
- mettre à jour la matrice des distances 2 à 2.

Un des points délicats est de définir $d(A, B)$ la distance entre deux éléments d'une partition de I : - *Cas d'une dissemblance*
Les stratégies ci-dessous s'accomodent d'un simple indice de dissemblance défini entre les individus. Elles s'appliquent également à des indices plus structurés (distance) mais n'en utilisent

pas toutes les propriétés.

$$d(A,B) = \min_{i\in A, j\in B} d_{ij} \quad \text{saut minimum, } \textit{single linkage}$$
$$d(A,B) = \sup_{i\in A, j\in B} d_{ij} \quad \text{saut maximum ou diamètre, } \textit{complete linkage}$$
$$d(A,B) = \frac{1}{card(A)card(B)} \sum_{i\in A, j\in B} d_{ij} \quad \text{saut moyen, } \textit{group average linkage}$$

- Cas d'une distance euclidienne
Les stratégies suivantes nécéssitent la connaissance de représentations euclidiennes des individus : matrice $n \times p$ des individus afin, au minimum de pouvoir définir les barycentres notés g_A et g_B des classes. On note w_A et w_B le poids de chacune des classes.

$$d(A,B) = d(g_A, g_B) \quad \text{distance des barycentres, centroïd}$$
$$d(A,B) = \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{saut de Ward}$$

Le saut de Ward joue un rôle particulier et est la stratégie la plus courante ; c'est même l'option par défaut dans (SAS) dans le cas d'une distance euclidienne entre individus. En effet, ce critère induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse.

On peut tracer un graphique représentant la décroissance du rapport de la variance intra classe sur la variance totale (R^2 partiel) en fonction du nombre de classes. La présence d'un rupture importante dans cette décroissance aide au choix du nombre de classes. D'autre part, on trace généralement aussi le *dendogramme*. C'est une représentation graphique des aggrégations successives sous la forme d'un arbre binaire. La hauteur d'une branche est proportionnalle à l'indice de dissemblance ou distance entre les deux objets regroupés. Dans le cas du saut de Ward, c'est la perte de variance interclasses.

Une CAH est souvent utilisée pour initialiser une méthode des centres mobiles (nombre de classes, centre des classes). Si le nombre d'observations est grand, il est d'usage de réaliser la CAH sur un échantillon tiré au hasard dans la base de données.

7.4 Méthode des centres mobiles

La méthode des centres mobiles ou k-moyennes [2] est un algorithme de réallocation dynamique qui repose sur la maximisation d'un critère global construit comme étant le rapport de l'inertie intraclasse sur l'inertie interclasse. Ce critère sous entend que l'on cherche une partition telle que les individus d'une même classe soient le plus semblables possible (variance intra classe faible) et que les classes différent le plus possible entre elles (variance interclasse élevée).

Soit $\mathbf{x} = \{x_{ij}\}_{i=1,\cdots,n, j=1\cdots,p}$ une matrice d'observations. On choisit a priori le nombre de classes K. On note g_k le centre de gravité de la classe k.

Algorithme des kmeans

Initialisation Choisir le nombre de classes K puis choisir K points (individus) au hasard parmi les observations

[2] en anglais : kmeans

Itérer jusqu'à ce que le critère de variance interclasse ne croisse plus de manière significative.
Pour $i = 1, ..., n$,

- Allouer l'individu i à la classe k telle que $\text{dist}(x_i, g_k) \leq \text{dist}(x_i, g_l)$ pour tout $l \neq k$.
- Calculer les centres de gravités g_k des K classes.

Propriétés Le critère (variance interclasses) est majoré par la variance totale. Il est simple de vérifier qu'l ne peut que croitre à chaque étape de l'algortihme, ce qui en assure la convergence. Concrêtement, une dizaine d'itération suffit gnéralement à atteindre la convergence. Lz solution obtenue est un optimum local. La partition obtenue par l'algorithme des k-moyennes dépend des représentants initialement choisis (essayez de vous en convaincre sur un exemple simple). De façon à s'affranchir en partie de cette dépendance, on exécute l'algorithme des k-moyennes (K et dist étant fixés) avec des initialisations différentes, et on retient la meilleure partition.

La qualité d'une partition est mesurée par la quantité

$$\sum_{k=1}^{K} \sum_{i \in C_k} \text{dist}(x_i, g_k)$$

qui mesure la cohésion des classes obtenues.

7.4.1 Généralisations

On remarque que la distance dist peut-être définie en fonction du type de variables observées. Cependant, dans la version la plus usuelle de l'algorithme des k-moyennes la distance considérée est la distance euclidienne. Dans le cas où les variables ne sont pas toutes quantitatives, on travaille généralement directement avec un tableau de distances. Dans ce cas, on ne calculera plus le centre de gravité de la classe mais il sera remplacé par le mode de la distribution conditionnellement à la classe.

7.4.2 Modèles de mélange

Modèle

Un modèle de mélange caractérise la distribution de la variable X d'un couple (S, X) tel que

- S est une variable aléatoire discrète définie sur $\{1, \cdots, M\}$; S n'est pas observée (on dit aussi que S est *cachée* ou *latente*).
- X est une variable aléatoire à valeurs sur $\mathbb{R}^d, d \leq 1$ telle que la loi conditionnelle $P(X|S = m)$ admet une densité $g_m(.)$ pour tout $m \in \{1, \cdots, M\}$.

D'après le théorème de Bayes, pour tout ensemble A, la loi marginale de la variable X vérifie

$$P(X \in A) = \sum_{m=1}^{M} P(X \in A|S = m)P(S = m)$$

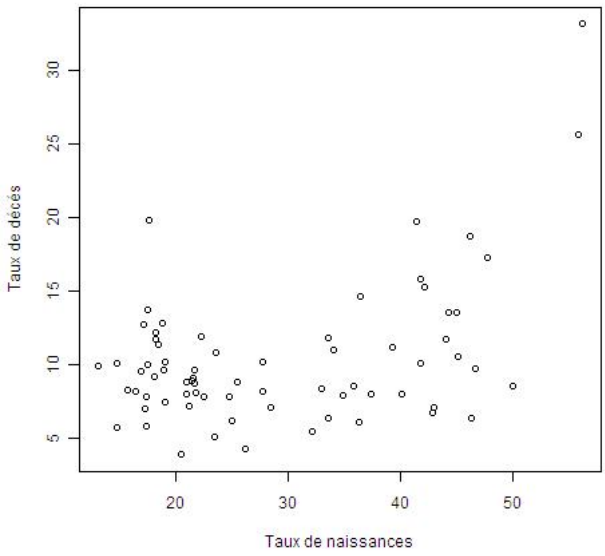
On écrit alors la densité de X comme une combinaison convexe des densités $g_m(.), m \in \{1, \cdots, M\}$

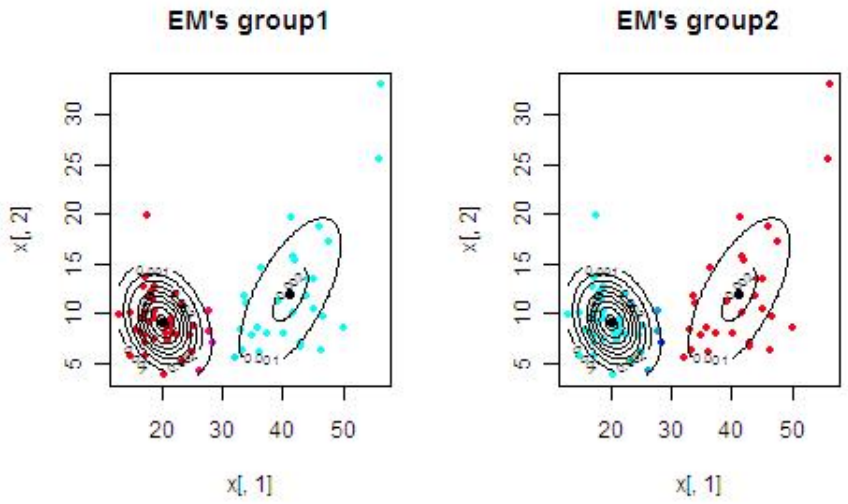
$$g(x) = \sum_{m=1}^{M} \pi_m g_m(x)$$

où $\pi_m = P(S = m)$ avec

$$\pi_1 + \cdots + \pi_M = 1$$

Exemple - Taux de naissances et de décès pour 70 pays du monde.





Inférence : algorithme EM

L'algorithme EM (Estimation-Maximisation) a été proposé par Arthur Dempster, Nan Laird et Donald Rubin en 1977. C'est une méthode qui permet d'approcher l'estimateur du maximum

de vraisemblance quand les données sont incomplètes (cas d'une variable cachée par exemple) ou quand une partie des données est manquante (cas d'une censure par exemple).

Algorithme EM pour un mélange de lois de Gauss

Initialisation Choisir le nombre de classes K puis initialiser le vecteur de paramètres

Itérer jusqu'à ce que la log-vraisemblance ne croisse plus de manière significative.
Pour $i = 1, ..., n$,

- **E-step** Estimer la probabilité a postériori qu'un individu i appartienne à la classe k

$$T_{k,i} = P(S = k|X = x_i; \theta^{(t)}) = \frac{\pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, \Sigma_k^{(t)}}{\sum_{l=1}^{K} \pi_l^{(t)} \phi(x_i; \mu_l^{(t)}, \Sigma_l^{(t)}}$$

- **M-step** Estimer les paramètres pour $k = 1, \cdots, K$

$$\begin{aligned}
\pi_k^{(t+1)} &= \frac{1}{n} \sum_{i=1}^{n} T_{k,i} \\
\mu_k^{(t+1)} &= \frac{\sum_{i=1}^{n} T_{k,i} x_i}{\sum_{i=1}^{n} T_{k,i}} \\
\Sigma_k^{(t+1)} &= \frac{\sum_{i=1}^{n} T_{k,i} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} T_{k,i}}
\end{aligned}$$

7.5 Exemple : composition du lait chez différents mammifères

Nous considérons de nouveau le jeu de données dans lequel on a la composition du lait pour 25 mammifères.

Choisissons tout d'abord le nombre de classes. La figure 7.3 représente la décroissance du R^2 partiel en fonction du nombre de classes. On en déduit qu'il semble raisonnable de considérer 3 ou 4 classes.

7.6 Combinaison de différentes méthodes de classification

Il est courant de combiner les méthodes de classification introduites précédemment. En effet, la méthode de classification hiérarchique n'est raisonnablement applicable que si le nombre d'observations est relativement faible. Son résultat constitue néanmoins souvent une intialisation intéressante pour une méthode des k-moyennes. Il fournit en effet à la fois des critères pour sélectionner le nombre de classes et une initialisation des centres de classe.

Pour les grand ensembles de données, comme on en rencontre fréquemment en data mining, on peut mettre en place la stratégie suivante :

1. Réaliser une classification par nuées dynamique sur un sous échantillon tiré au hasard et de taille environ 10% de n. On choisit un nombre de classes grand.

2. Exécuter une classification hiérarchique ascendante sur les barycentres des classes obtenus puis déterminer un nombre de classe optimal K.

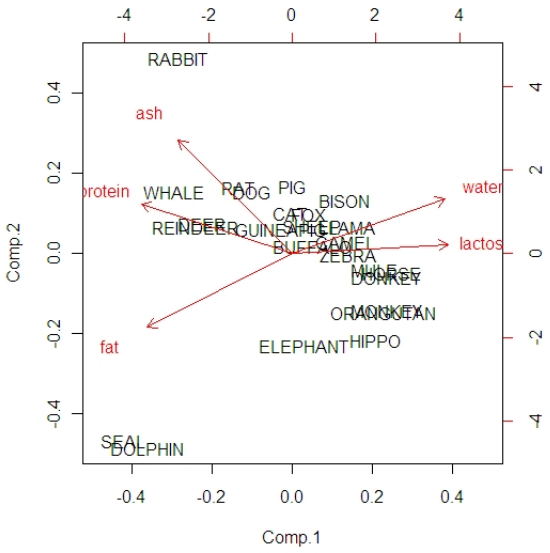


Figure 7.1: Analyse en composante principale, biplot.

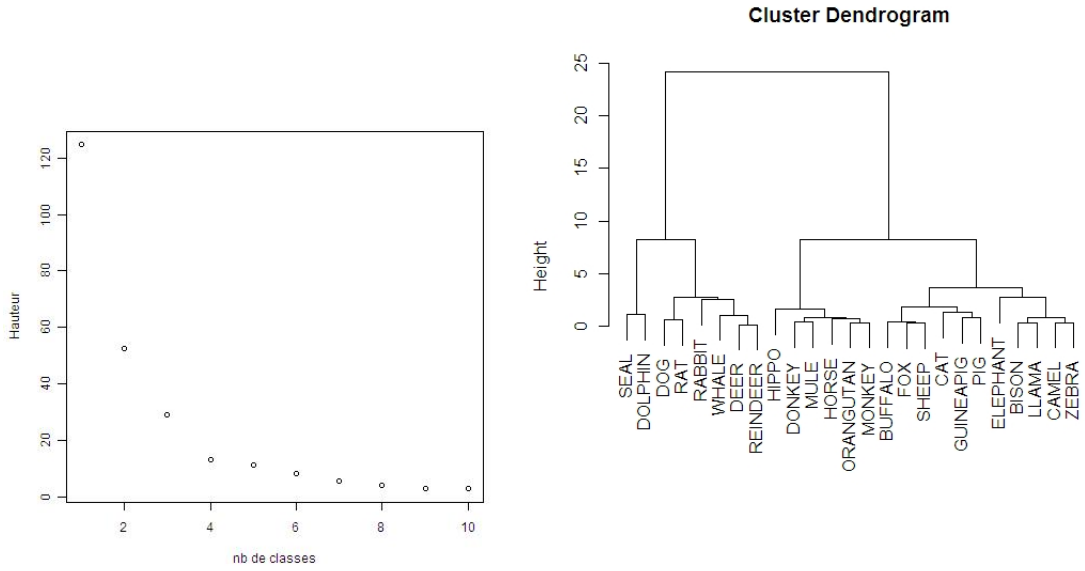


Figure 7.2: Décroissance du R^2 partiel en fonction du nombre de classes (à gauche) et dendogramme (à droite)

3. Réaliser une classification par k-moyennes pour K classes et en choisissant comme valeurs initiales des centres de classe les barycentres des classes de l'étape précédente. On pourra pondérer ces centres par le nombre d'individu dans les classes.

Dans un second temps, on enchaîne généralement d'autres analyses telles que

- Une analyse en composantes principales qui permet de représenter les classes dans un sous espace factoriel et de se faire une idée de la pertinence de la classification obtenue.

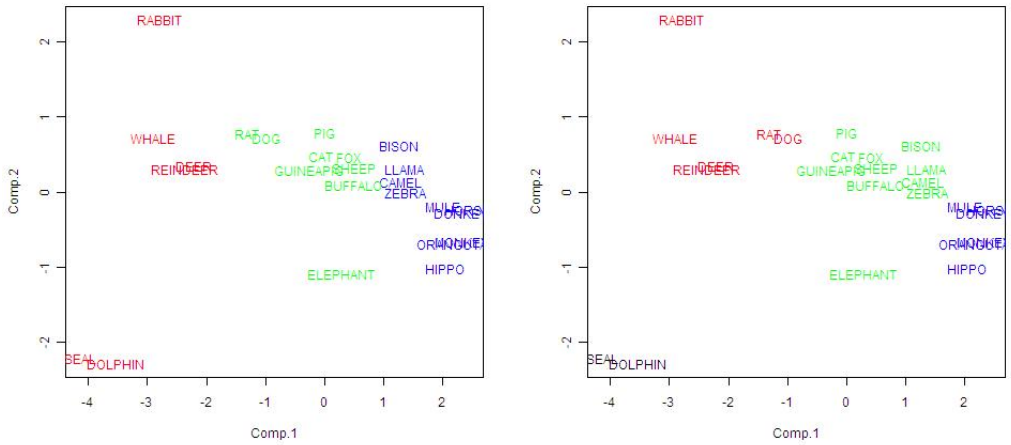


Figure 7.3: Projection des individus sur le premier plan factoriel de l'ACP avec matérialisation de 3 classes (à gauche) et 4 classes (à droite) par un code couleur.

- Une analyse discriminante qui permet d'aider à l'interprétation des classes.

Chapter 8

Analyse discriminante

8.1 Introduction

L'analyse discriminante est utilisée pour identifier, dans une population, des caractéristiques permettant de séparer deux groupes naturels. En pratique, il s'agit de définir une règle de décision pour classer un individu dans un groupe connaissant ses caractéristiques. L'analyse discriminante vise donc à résoudre des problèmes de classement. On dit que c'est une méthode de *classification supervisée*. Elle se différencie des méthodes de *classification*[1], dans la mesure où les classes sont définies a priori, c'est à dire qu'on dispose d'un jeu de données incluant une variable de classe qui est observée. Dans certains cas, on peut voir l'analyse discriminante comme une extension de l'analyse de la variance.

Donnons quelques exemples de problèmes de classification supervisée.

- **En biologie** - Un des exemples les plus fréquemment utilisés est celui des Iris de Fisher (1936). Il s'agit de discriminer 3 espèces d'Iris à partir de la mesure de la longueur et la largeur des pétales et des sépales des 50 fleurs de chacune des espèces. A partir de telles mesures on propose de définir une règle permettant d'affecter à une espèce déterminée un iris dont on ne connaitrait pas l'espèce. (voir Figure 8.1)

  Un autre exemple concerne le problème de la détermination de la sous espèce de certains poissons. Une des méthodes pour déterminer l'espèce consiste à disséquer le poisson. Debouche et al. 1979 proposent une règle de décision basée sur des mesures faciles à réaliser sur un poisson vivant.

- **En marketing** - Identifier un bon client/un mauvais client ou encore un client à qui on peut faire un prêt.

- **En médecine** - Maladie du coeur (analyse de sang), présence de tumeur (analyse d'image)

- **En multimédia** - Reconnaissance de forme : retrouver un visage dans une banque de données, etc.

Il existe de nombreuses méthodes ou modèles permettant de construire des régles de classement. Citons les plus communes :

[1] En anglais: *clustering*

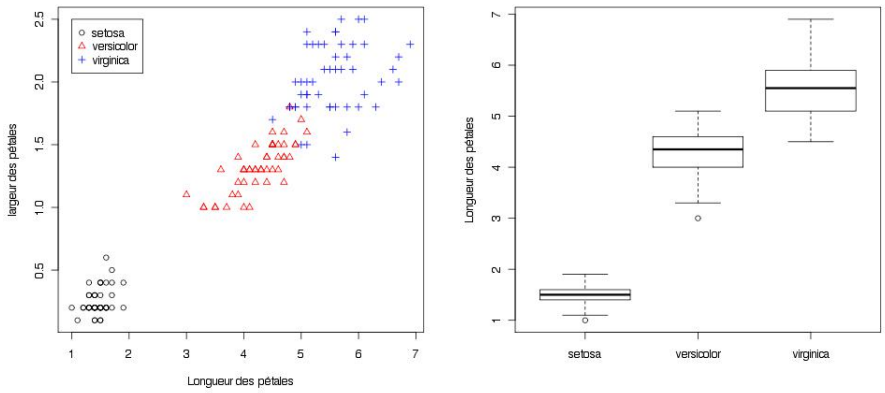


Figure 8.1: Iris de Fisher

- analyse discriminante décisionnelle[2]
- analyse discriminante factorielle, appelée aussi analyse des variables canoniques[3]
- modèle linéaire généralisé (régression logistique)
- arbres de décision
- réseaux de neurones artificiels
- machines à vecteurs supports

Certains modèles peuvent être vus comme des cas particulier des autres. Par exemple, l'analyse discriminante factorielle est un cas particulier d'analyse discriminante décisionnelle et c'est aussi un modèle linéaire généralisé particulier. Ce cours portera essentiellement sur l'analyse discriminante décisionnelle et sur l'analyse discriminante factorielle.

8.2 Analyse discriminante décisionnelle

8.2.1 Règle de décision

Commençons par poser le problème et introduire les notations. Soient $X \in \mathbb{R}^p$ une matrice de covariables et $Y \in \{1, \cdots, K\}$ une variable discrète à prédire sachant $X = x$. K définit le nombre de classes. Soit de plus un échantillon $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ de n observations du couple (X, Y).

Définition 9. *Une règle de décision δ est une application de $\mathbb{R}^p$ à valeurs dans l'ensemble $\{1, \ldots, K\}$. Le résultat $\delta(\mathbf{x}) = k$ signifie que l'individu associé au vecteur $\mathbf{x}$ est affecté au groupe numéro k.*

Une règle de décision engendre une partition de $\mathbb{R}^p$ en K régions $R_1, \cdots, R_K$ avec pour tout $k = 1, \ldots, K$

$$R_k = \{\mathbf{x} \in \mathbb{R}^p | \delta(\mathbf{x}) = k\}$$

[2]En anglais: *discriminant analysis, predictive discriminant analysis, classification procedure*
[3]En anglais: *factorial discriminant analysis, descriptive discriminant analysis, canonical variate analysis*

Il s'agit de définir une règle de décision (ou règle de classement) δ telle que le risque de se tromper quand on classera un nouvel individu soit minimal.

8.2.2 Risque de Bayes

A une règle de décision δ, on associe le risque de Bayes qui représente le l'espèrance de la fonction de coût et s'écrit

$$\mathcal{R}(\delta) = \sum_{k=1}^{K} P(\delta(X) = k | Y \neq k) = \sum_{k=1}^{K} \pi_k \sum_{l \neq k} \int_{R_l} f_k(x_1, \ldots, x_p) dx_1 \ldots dx_p \tag{8.1}$$

avec f_k la densité de X dans la classe k et π_k la probabilité que Y soit égal à k ; $\pi_1 + \cdots + \pi_k = 1$. Les probabilités $\pi_k,\ k = 1, \cdots, K$ sont appelées *probabilités a priori.*

Dans l'équation (8.1), on a supposé que le coût pour le décideur d'affecter un individu de la classe k à la classe l était le même pour pour tout les couples $k \neq l$. Lorsque les coûts d'un mauvais classement sont différents, on introduit les coefficients $C(k, l)$ et on pose par définition $C(k, k) = 0$ (cela ne coûte rien de ne pas se tromper). Dans ce cas le coût moyen s'écrit

$$R(\delta) = \sum_{k=1}^{K} \pi_k \sum_{l \neq k} C(l, k) \int_{R_l} f_k(x_1, \ldots, x_p) dx_1 \ldots dx_p$$

Proposition 12. *Lorsque $C(k, l) = C$ pour tout $k \neq l$, la règle de décision qui minimise le risque de Bayes est la suivante*

$$\delta(\mathbf{x}) = \arg \max_{k=1,\ldots,K} P(Y = k | X = x) = \arg \max_{k=1,\ldots,K} \frac{\pi_k f_k(\mathbf{x})}{\sum_{l=1}^{K} \pi_l f_l(\mathbf{x})}$$

On affecte l'individu au groupe de probabilité a posteriori $\pi_k f_k$ maximum.
Cette règle est parfois appelée règle de Bayes.

On remarque que

$$P(Y = k | X = x) = \frac{P(Y = k, X = x)}{P(X = x)} = \frac{P(X = x | Y = k) P(Y = k)}{\sum_{l=1}^{K} P(X = x, Y = l)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{8.2}$$

et $\delta(x) == \arg\max_{k=1,\ldots,K} \pi_k f_k(\mathbf{x})$.
Les probabilités $P(Y = k | X = x)$ sont appelées *probabilités a posteriori.*
Ainsi la règle de Bayes affecte l'invidu associé à x à la classe qui a la plus forte probabilité a posteriori.

Preuve - *à faire?*

Lorsque les probabilités a priori sont inconnues, on utilise un critère minimax. Il s'agit de minimiser le risque maximum de mauvais classement. La règle de décision qui minimise le critère minimax est la suivante

$$\delta(\mathbf{x}) = \arg \max_{k=1,\ldots,K} f_k(\mathbf{x})$$

En général, les densités f_k sont inconnues et il faut les estimer. Selon les problèmes, on choisit des estimateurs paramétriques ou non paramétriques.

8.2.3 Cas de variables aléatoires gaussiennes

Nous considérons dans cette partie le cas où les variables dépendantes suivent une loi de Gauss. La loi jointe (X, Y) est donc caractérisée par

- Y est une variable nominale définie sur $\{1, \cdots, K\}$ et on note $\pi_k = P(Y = k)$,
- pour tout $k \in \{1, \cdots, K\}$, $X|Y = k$ suit une loi de Gauss de moyenne μ_k et de variance $\Sigma_k \in \mathbb{R}^{p,p}$ et on a donc

$$f_k(x) = \frac{1}{(2\pi)^{p/2} det(\Sigma_k)} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

et

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \text{ (par Bayes)}$$

Ainsi la règle de Bayes devient ici

On attribue un individu x_i à la classe k si pour tout $l \neq k$

$$\frac{\pi_k}{(2\pi)^{p/2} det(\Sigma_k)} \exp\left(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right) \geq \frac{\pi_l}{(2\pi)^{p/2} det(\Sigma_l)} \exp\left(-\frac{1}{2}(x_i-\mu_l)^T \Sigma_l^{-1}(x_i-\mu_l)\right)$$

Pour simplifier cette expression, on peut la transformer par passage au logrithme et la fonction log étant croissante, on obtient

$$\log(\pi_k) - \frac{p}{2}\log(2\pi) - \log(det(\Sigma_k)) - \left(\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right) \geq \log(\pi_l) - \frac{p}{2}\log(2\pi) - \log(det(\Sigma_l)) - \left(\frac{1}{2}(x_i-\mu_l)^T \Sigma_l^{-1}(x_i-\mu_l)\right)$$

soit encore

$$2\log(\pi_k)) - 2\log(det(\Sigma_k)) - (x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k) \geq 2\log(\pi_l) - 2\log(det(\Sigma_l)) - (x_i-\mu_l)^T \Sigma_l^{-1}(x_i-\mu_l) \quad (8.3)$$

Définition 10. *Les fonctions*

$$g_k(x) = \log(\pi_k) - \frac{p}{2}\log(2\pi) - \log(det(\Sigma_k)) - \left(\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

sont appelées fonctions discriminantes.

Quand les variables dépendantes sont gaussiennes, les fonctions discriminantes sont quadratiques :

$$g_k(x) = x^T W_k x + w_k^T x + \omega_0$$

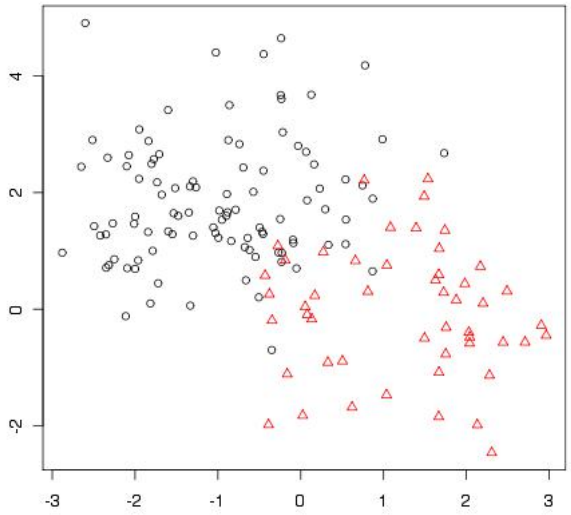
avec le vecteur de poids $w_k = \Sigma^{-1}\mu_k$, la métrique $W_i = -\frac{1}{2}\Sigma^{-1}$ et le seuil (ou biais)

$$\omega_0 = -\frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k - \frac{1}{2}\log det(\Sigma_k) + \log \pi_k$$

La frontière entre deux classes est une surface quadratique.

L'inégalité (8.3) donne la régle de Bayes dans le cas général de variables dépendantes gaussiennes. On peut considérer plusieurs cas particuliers.

Cas 1 : Cas homoscédastique avec matrices de covariances sphériques. On suppose que pour tout $k \in \{1, \cdots, K\}$, $\Sigma_k = \sigma^2 I_p$.



$$\pi_1 = 2/3, \mu_1 = (-1, 2)^T, \mu_2 = (1, 0)^T, \sigma^2 = 1, n = 150$$

La fonction discriminante g_k devient alors

$$g_k(x) = -\frac{||x - \mu_k||^2}{2\sigma^2} + \log(\pi_k) + constante$$

En développant g_k et en substituant l'expression obtenue dans l'inégalité (8.3), on obtient

$$-\frac{x^T x + 2x^T \mu_k + \mu_k^T \mu_k}{2\sigma^2} + \log(\pi_k) + constante \geq$$
$$-\frac{x^T x + 2x^T \mu_l + \mu_l^T \mu_l}{2\sigma^2} + \log(\pi_l) + constante$$

La constante est bien la même dans les membres de gauche et de droite. Après simplification, on a

$$2x^T \mu_k + \mu_k^T \mu_k - 2\sigma^2 \log(\pi_k) \leq 2x^T \mu_l + \mu_l^T \mu_l - 2\sigma^2 \log(\pi_l)$$

Dans le cas homoscédastique, la règle de décision est linéaire. La frontière de décision qui correspond au cas où on a égalité entre les deux fonctions discriminantes se met alors sous la forme

$$(\mu_k - \mu_l)(x - x_0) = 0$$

avec

$$x_0 = \frac{1}{2}(\mu_k + \mu_l) + \left(\frac{\sigma^2}{|||\mu_k - \mu_l||^2} \log \frac{\pi_k}{\pi_l}\right)(\mu_l - \mu_k)$$

Ainsi dans le cas homoscédastique avec matrice de covariance sphérique, la frontière est linéaire.

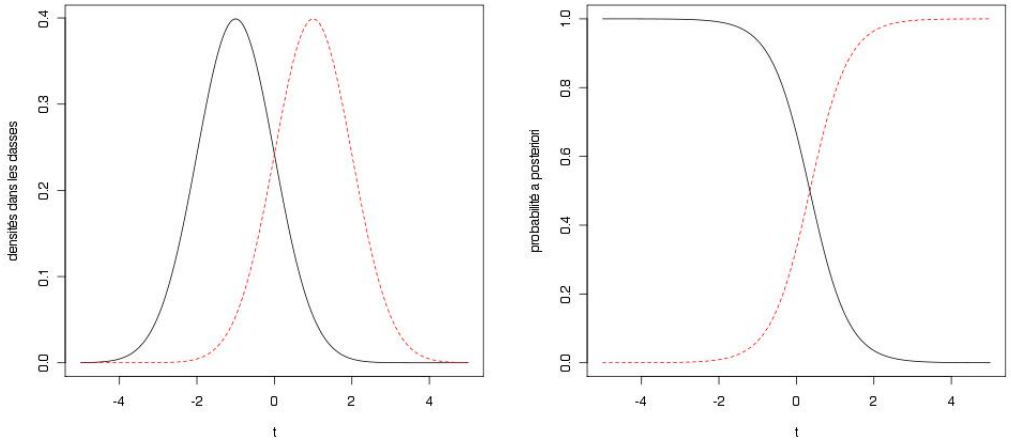
Si, de plus, les probabilités a priori sont égales, $\log \frac{\pi_k}{\pi_l} = 0$ et on obtient la règle de décision de la plus proche moyenne selon la distance euclidienne :

On attribue un individu x_i à la classe k
si pour tout $l \neq k$

$$(\mu_k - \mu_l)^T (x - \frac{1}{2}(\mu_k + \mu_l)) \geq 0$$

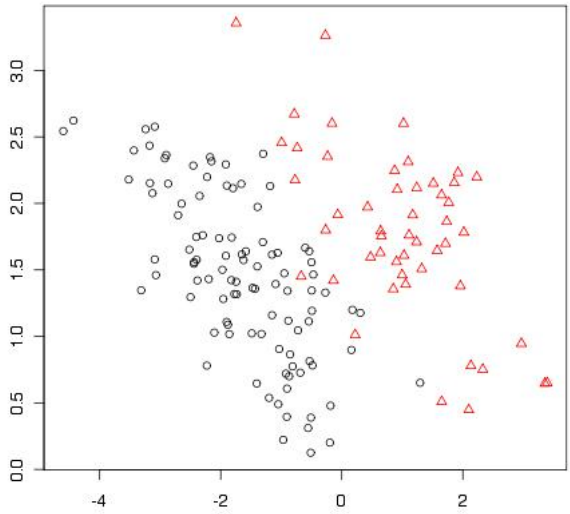
c'est à dire si

$$(x - \mu_k)^2 \leq (x - \mu_l)^2$$



Variances égales à 1 et $\pi_1 = 2/3$

Cas 2 : Cas homoscédastique. On suppose que pour tout $k \in \{1, \cdots, K\}$, $\Sigma_k = \Sigma$.



$$\pi_1 = 2/3, \Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 2 \end{pmatrix}, n = 150$$

Dans ce cas, la frontière de décision se met sous la forme

$$\Sigma^{-1}(\mu_k - \mu_l)^T(x - x_0) = 0$$

avec

$$x_0 = \frac{1}{2}(\mu_k + \mu_l) + \left(\frac{1}{(\mu_k - \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l)} \log \frac{\pi_k}{\pi_l} \right) (\mu_l - \mu_k)$$

Ainsi dans le cas homoscédastique, la frontière est linéaire.

Si les probabilités a priori sont égales, $\log \frac{\pi_k}{\pi_l} = 0$ et on obtient la règle de décision de la plus proche moyenne selon la distance de Mahalonobis.

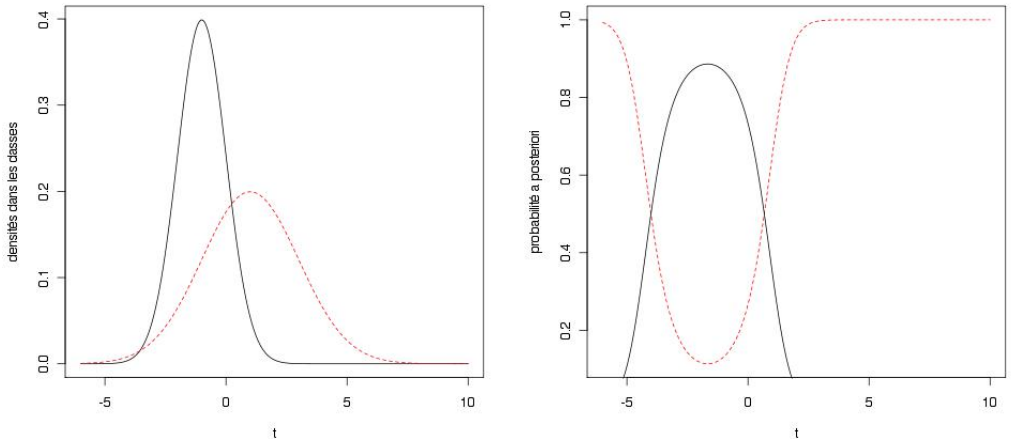
Définition 11. *La distance de Mahalanobis est la distance euclidienne corrigée par la variance :*

$$d(x, \mu) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

On remarque, qu'au sens de la métrique de Mahalanobis, l'hyperplan qui sépare deux classes est la médiatrice du segment qui joint les centres de gravité des deux classes. En effet si x appartient à la frontière entre les classes k et l, alors $d(x, \mu_k) = d(x, \mu_l)$.

On attribue un individu x_i à la classe k,
si pour tout $l \neq k$

$$d(x_i, \mu_k) \leq d(x_i, \mu_l)$$



$$\sigma_1 = 1, \sigma_2 = 2, \pi_1 = 2/3$$

8.2.4 Cas de variables dépendantes quelconques

Jusqu'à présent nous avons supposé que les covariables X étaient de distribution gaussienne conditionnellement à la variable réponse Y. Or cette hypothèse est rarement vérifiée en pratique. Cependant, rien n'empèche d'ajuster une régle de décision linéaire ou quadratique qui sera alors une approximation de la règle de décision bayésienne optimale.

Lorsque les co-variables sont quantitatives et ont une distribution symétrique (ou faiblement dissymétrique) alors il est usuel de supposer que l'hypothèse de normalité est approximativement vérifiée. Dans ce cas en effet les méthodes basées sur des hypothèses de normalité sont généralement robustes et les résultats restent interprétables. Mais quand les co-variables sont de distribution fortement dissymétrique ou si elles sont qualitatives, il faut envisager d'autres approches.

Une méthode consiste à transformer les variables de façon à les rendre marginalement gausiennes. La transformation la plus courante est la transformation de Box-Cox :

$$\begin{aligned} \varphi_0(x) &= \ln(x) \\ \varphi_\lambda(x) &= \frac{x^\lambda - 1}{\lambda} \text{ si } \lambda > 0 \end{aligned}$$

Le paramètre λ est généralement estimé par validation croisée (ou graphique). On peut faire les remarques suivantes :

1. Si X prend des valeurs négatives ou nulles, il est loisible de lui rajouter arbitrairement une constante afin d'obtenir des observations positives avant d'effectuer une transformation de Box-Cox.

2. Pour $\lambda = 1$, la transformation de Box-Cox ne modifie pas l'échantillon, en dehors d'une translation par 1 qui n'a aucune incidence sur l'étude statistique.

3. Pour $\lambda < 1$, l'effet de la transformation de Box-Cox est de d'affaiblir un skewness positif. Plus λ est proche de 0, plus cet effet est important.

4. Pour $\lambda > 1$, l'effet est opposé.

Cependant, quand n'a pas de connaissance a priori sur la loi des co-variables, la méthode la plus naturelle consiste à considérer des estimateurs non paramétriques des densités.

Estimateur des plus proches voisins

Si la densité de probabilité f d'une variable ou d'un vecteur aléatoire X est continue au point x, elle peut être estimée en x par l'estimateur des plus proches voisins

$$\hat{f}_n(x) = \frac{k_n}{nV_n(x)}$$

où n est la taille de l'échantillon observé et $V_n(x)$ le volume de la région contenant les k_n plus proches voisins de x.

On peut montrer que pour que $\hat{f}_n(x)$ converge en probabilité vers $f(x)$ avec f continue en x, il suffit que $k_n \to \infty$ et $k_n/n \to 0$ lorsque $n \to \infty$.

Règle de décision
Supposons qu'une région R de volume V centrée en x contienne k individus parmi lesquels k_j appartiennent à la classe j. Alors la probabilité conjointe de $X = x$ et de $Y = j$ est estimée par $\hat{f}_n(x, j) = \frac{k_j}{nV}$ et la densité a posteriori est

$$P(\text{individu } I \in j | X = x) = \frac{\pi_j \hat{f}_n(x, j)}{\sum_{l=1}^{M} \pi_l \hat{f}_n(x, l)}$$

En particulier, si les probabilités a priori sont égales, on a

$$P(\text{individu } I \in j | X = x) = \frac{k_j}{k}$$

Ainsi on classe l'individu I dans la classe la plus représentée (la plus nombreuse) dans le voisinage de x. Si les probabilités a priori sont différentes, on obtient la règle de décision de Bayes naive[4]

On attribue un individu x_i à la classe j,
si pour tout $l \neq j$

$$\pi_k k_j \geq \pi_l k_l$$

Estimateurs à noyau

Quand la dimension p de l'espace des variables dépendantes X n'est pas trop grande (de l'ordre de quelques unités), on peut alternativement utiliser des estimateurs à noyau

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Pour que l'estimateur $\hat{f}$ de f définisse une densité et ait de bonnes propriétés de convergence, on choisit en général un noyau K ayant les propriétés suivantes :

[4]En anglais : naive Bayes decision rule

1. K est une densité : $\int K(u)du = 1$, $\lim_{|u|\to 0} K(u) = 0$
2. K est une fonction paire
3. K est deux fois différentiable
4. $\int u^2 K(u)du \neq 0$, $\int K(u)^2 du < \infty$

Les paramètres importants des algorithmes d'analyse discriminante basés sur des estimateurs non paramétriques de densités sont le nombre de voisins dans le cas de l'estimateur des plus proches voisins et la largeur de fenêtre dans le cas de l'estimateur à noyau. En effet ce sont ces paramètres qui vont faire que l'on obtient des frontières plus ou moins lisses. On les choisit généralement par validation croisée : on fait la classification pour plusieurs valeurs du nombre de voisins (resp. de la largeur de fenêtre), on compare le classement $\hat{y}_i = \hat{k}$ obtenu aux observations y_i pour $i \in I$ et on retient la valeur qui conduit à la plus faible erreur de classement.

8.3 Analyse factorielle discriminante

L'analyse factorielle discriminante (AFD) est une méthode géométrique et essentiellement descriptive qui ne repose que sur des notions de distance et ne fait pas intervenir d'hypothèses probabilistes. Comme dans l'ACP et les autres méthodes factorielles on cherche un espace dans lequel on va projeter le nuage de point tout en mettant préservant au mieux des distances choisies ; ici on veut mettre en évidence les groupes, autrelent dit préserver les distance à l'intérieur des groupes et entre les centres de gravité des groupes.

8.3.1 Variances interclasse et intraclasse

Soient n individus x_i de l'échantillon constituant K groupes $R_1, \cdots, R_K$. On note $\bar{x}_k$ le centre de gravité du groupe E_k pour tout i. Chaque individu x_i est affecté d'un poids p_i. Le plus souvent, on a $p_i = \frac{1}{n}$. On note de plus q_k le poids du groupe k donné par

$$q_k = \sum_{i=1}^{n_k} p_i$$

avec n_k l'effectif du groupe E_k.

Définition 12. *La* **variance interclasse**[5] *B est estimée par la variance empirique des K centres de gravité.*

$$B = \sum_{k=1}^{K} q_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})' \tag{8.4}$$

où q_k représente le poids relatif de la classe k. On a $q_1 + \cdots + q_K = 1$.

Définition 13. *La* **variance intraclasse**[6] *W est estimée par la moyenne des variances empiriques de chaque classe.*

$$W = \sum_{k=1}^{K} q_k V_k \text{ avec } V_k = \frac{1}{q_k} \sum_{i=1^{n_k}} p_i (x_i - \bar{x}_k)(x_i - \bar{x}_k)' \tag{8.5}$$

La variance W est en général inversible.

[5] En anglais : variance between
[6] En anglais : variance within

Proposition 13. *La* **variance totale** V *se décompose*

$$V = B + W \tag{8.6}$$

Preuve - la variance empirique s'écrit

$$\begin{aligned}
\hat{\Sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i - \bar{x})^T(x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{k=1}^{K} SS(k)
\end{aligned}$$

On note SS pour *sum of squares* et on a la décomposition suivante :

$$\begin{aligned}
SS(k) &= \sum_{i\in C_k}(x_i - \bar{x})^T(x_i - \bar{x}) \\
&= \sum_{i\in C_k}(x_i - \bar{x}_k + \bar{x}_k - \bar{x})^T(x_i - \bar{x}_k + \bar{x}_k - \bar{x}) \\
&= \sum_{i\in C_k}(||x_i - \bar{x}_k||^2 + ||\bar{x}_k - g||^2) \\
&= \left(\sum_{i\in C_k}||x_i - \bar{x}_k||^2\right) + n_k||\bar{x}_k - \bar{x}||^2
\end{aligned}$$

Et donc la variance totale se met sous la forme

$$\begin{aligned}
\hat{\Sigma}^2 &= \frac{1}{n}\left\{\sum_{k=1}^{K}\sum_{i\in C_k}||x_i - \bar{x}_k||^2 + \sum_{k=1}^{K} n_k||\bar{x}_k - \bar{x}||^2\right\} \\
&= \frac{1}{n}\sum_{k=1}^{K} n_k\left\{\frac{1}{n_k}\sum_{i\in C_k}||x_i - \bar{x}_k||^2 + ||\bar{x}_k - \bar{x}||^2\right\} \\
&= \frac{1}{n}\sum_{k=1}^{K} n_k(W_k + B_k) = W + B
\end{aligned}$$

◇

8.3.2 Axes et variables discriminantes

L'Analyse Factorielle Discriminante (AFD) consiste à rechercher de nouvelles variables (les variables discriminantes) correspondant à des directions de $\mathbb{R}^p$ qui séparent le mieux possible, en projection, les K groupes d'observations. Une variable discriminante est intéressante pour expliquer la classification induite par Y si cette variable :

- regroupe bien les individus d'un même groupe,

- sépare bien les K groupes.

Notons b_1 le premier axe discriminant. En projection sur l'axe b_1, les K centres de gravités doivent être aussi séparés que possible tandis que chaque groupe doit se projeter de manière groupée autour de la projection de son centre de gravité. En d'autres termes l'inertie B du nuage des centres de gravité doit être maximale. Et la variance intra classe W doit être minimale.

La première variable discriminante est donc telle que le rapport de la variance entre les groupes (variance interclasse) à la variance totale soit maximum. La seconde variable vise aussi à maximiser ce rapport sous la contrainte de non corrélation avec la première variable. Et ainsi de suite. Le nombre total de variables discriminantes ne peut dépasser ni p ni $K-1$.
Plus précisement, le premier axe discriminant b_1 est solution de :

$$\max_{b\in\mathbb{R}^p/b'MVMb=1} \frac{b'MBMb}{b'MVMb} = \max_{b\in\mathbb{R}^p/b'MVMb=1} b'MBMb \tag{8.7}$$

avec M une métrique choisie. On peut par exemple choisir M égale à l'identité.
Il est facile de vérifier que chercher le maximum de $\frac{b'MBMb}{b'MVMb} = \frac{b'MBMb}{b'M(W+B)Mb}$ est équivalent à chercher le maximum de $\frac{b'MBMb}{b'MWMb}$. Par ailleurs, on peut montre, comme pour l'ACP, que la solution est la plus grande valeur propre λ_1 de $M^{-1}V^{-1}BM$ (soit $V^{-1}B$ si la métrique est l'identité) ou de manière équivalente de $M^{-1}W^{-1}BM$. $\lambda_w = \lambda_v/(1-\lambda_v)$.

On a toujours $0 \leq \lambda_1 \leq 1$.

- $\lambda_1 = 1$ correspond au cas où en projection sur b les variances intra classe sont nulles. Les K sous-nuages appartiennent donc à des hyperplans orthogonaux à b

- $\lambda_1 = 0$ au cas où les centres de gravités sont confondus (exemple : les nuages sont concentriques).

La valeur propre λ est une mesure pessimiste du pouvoir discriminant de l'axe b.

Exemple : *le cas de 2 groupes.*
Dans le cas de deux groupes, il n'y a qu'une seule variable discriminante car $\min(p, K-1) = 1$.
Le facteur discriminant vaut alors

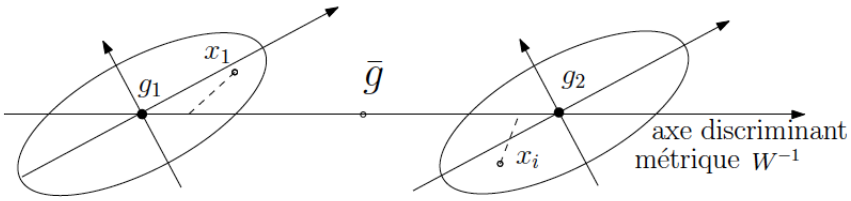
$$b = W^{-1}(\bar{x}_1 - \bar{x}_2)$$

et la variable discriminante vaut Xb.

- Pour l'individu i de l'échantillon initial, cette variable prend la valeur

$$(\mathbf{x}b)_i = x_i^T W^{-1}(\bar{x}_1 - \bar{x}_2) = d_{\text{Mahalanobis}}(x_i, \bar{x}_1 - \bar{x}_2)$$

  La variable discriminante s'obtient donc en projetant les observations sur l'axe reliant les deux centres de gravité pour la métrique de Mahalanobis.



- Pour un individu $z \in \mathbb{R}^p$, on obtient la fonction de Fisher (qui est aussi, à une constante près, la fonction discriminante dans le cas gaussien homoscédastique) :

$$z^T W^{-1}(\bar{x}_1 - \bar{x}_2)$$

8.3.3 Une ACP particulière

L'analyse factorielle discriminante apparaît comme une analyse en composantes principales du nuage G des K centres de gravité pour la métrique W^{-1} qui est la métrique de Mahalanobis avec les poids $\hat{\pi}_1, \cdots, \hat{\pi}_K$. La métrique de Mahalonobis permet icide prendre en compte la variance intra dans le calcul des distances.

Comme en ACP, on peut projeter les individus sur les plans factoriels et interpréter les variables discriminantes au moyen d'un cercle des corrélations.

Une représentation simultanée des individus et des barycentres des classes par rapport aux axes discriminants est obtenue dans l'espace des individus au moyen des coordonnées :

$$C = XW^{-1}U \text{ pour les individus}$$
$$C_G = GW^{-1}U \text{ pour les barycentres}$$

avec G la matrice centrée des barycentres.

Chaque variable X_j est représentée par un vecteur dont les coordonnées dans le système des axes factoriels est une ligne de la matrice $U(\Lambda_v/(1-\Lambda_v))^{1/2}$.

8.3.4 Sélection de modèle et MANOVA

En pratique, on utilise différentes méthodes pour sélection puis valider un modèle. En premier lieu on s'assure que la variable classe a bien un effet sur les autres variables par une analyse de la variance.

ANOVA - MANOVA

L'analyse de la variance (ANOVA) permet de tester l'effet de chacun des facteurs sur la différence en moyenne entre les groupes. Elle permet ainsi de sélectionner les variables qui ont un pouvoir discriminant.

L'analyse de la variance multivariée (MANOVA) permet de tester l'hypothèse selon laquelle le vecteur des variables explicatives apporte de l'information sur le fait qu'au moins deux des centres de gravité des groupes sont significativement différents. L'hypothèse H_0 est que les centres

de gravité sont égaux. On peut considérer plusieurs statistiques de test. L'une des plus commune est le *lambda de Wilks* $\Lambda = \frac{det(W)}{det(V)}$. Le Lambda de Wilks est une mesure directe de la proportion de l'inertie des groupes qui n'est pas expliquée par la variable indépendante (qui identifie les groupes) dans un schéma de décomposition de la variance totale des observations. C'est donc le rapport de l'inertie intraclasses, à l'inertie totale. On remarque d'après cette définition que plus Λ est petit plus la différence entre les centres de gravité est marquée. Si les variables sont distribuées dans chaque groupe suivant une loi de Gauss centrée, le lambda de Wilks suit une loi de Wishart, qui est la généralisation de la loi du χ^2 au cadre multivarié.

Définition 14. *Soit V une matrice aléatoire dans $\mathbf{R}^{p\times p}$.Alors $V*$ est de loi de Wishart à m degrés de liberté (notation $V \simeq W_p(m)$) est équivalent à $V = \sum_{i=1}^{m} Z_i Z_i^T$, où les Z_i sont i.i.d. $\mathcal{N}_p(0, \mathbb{I}_p)$.*

Sélection de variables

Forward stepwise analysis. In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

Backward stepwise analysis. One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups.

8.4 Validation de modèle

Validation croisée, calcul du risque de Bayes.