Réduction de dimension - Analyses Factorielles

On désigne sous l'appellation analyse des données multidimensionnelles (data-mining, fouille de données) les différentes méthodes statistiques, descriptives ou prédictives principalement, s'appliquant à de grand tableaux de données pour lesquels les méthodes statistiques classiques ne permettent qu'une approche partielle.

Ces techniques ont été introduites au cours du XXème siècle (Pearson, Spearman, Fisher, Hotelling, Benzécri ...) et ont connu leur développement avec l'essor des ordinateurs à partir des années 1970.

En parallèle, les modèles linéaires gaussiens et généralisés seront abordés pour leurs applications inférentielles et prédictives.

Nous étudierons différents types de tableaux :

- tableau individus $\times$ variables quantitatives, en Analyse en Composantes Principales (ACP), Analyse Factorielle Discriminante (AFD), modèle de régression multiple. Chaque ligne représente un individu sur lequel sont mesurées p variables quantitatives,
- tableau de contingence en Analyse Factorielle des Correspondances (AC). Les lignes et les colonnes représentent les différentes modalités de deux variables qualitatives,
- tableau de distance en classification, k-means, Classification Ascendate Hiérarchique (CAH),
- tableau individus $\times$ différents types de variables, qualitatives ou quantitatives en Analyse des correspondances multiples (ACM), en régression logistique, en analyse de variance (ANOVA), analyse de covariance (ANCOVA).

On peut identifier trois grands groupes de méthodes dont seule le premier est l'objet de ce document.

- Méthodes factorielles
- Méthodes de classification supervisées et non-supervisées
- Modélisation statistiques

Le premier groupe a pour objectif de réduire les dimensions de l'espace dans lequel sont décrits les variables ou les individus. Il repose sur un schéma de dualité caractérisé par un triplet (X, Q, D), X le tableau de données proprement dit, Q la métrique utilisée et D la métrique associée au poids. Les méthodes abordées sont :

- ACP : étudier la liaison entre plusieurs variables quantitatives et décrire les individus à l'aide d'un nombre restreint de variables synthétiques,
- AFC et ACM : étudier la liaison entre plusieurs modalités de deux variables qualitatives (AFC) ou plus de deux variables qualitatives (ACM),
- AFD : définir des variables discriminantes permettant de discriminer des classes d'individus.

Table des matières

Chapitre 1

Réduction de dimension 1 : Bases de l'analyse factorielle

Ce premier chapitre présente les bases qui seront utilisées tout au long des chapitres et demande une bonne compréhension et connaissance pour la suite. Quelques références pourront compléter ce chapitre :

- Probabilités, analyse des données et statistique. Saporta G. Technip . 3eme ed. 2011. (version en ligne disponible).
- `http://jf-durand-pls.com/bibliography/polyalgmatc.pdf`

1.1 Décomposition en Valeurs Singulières

La SVD (singular value decomposition) est une méthode de réduction de matrice intervenant dans différents domaines en statistiques, inverse généralisée de matrice, réduction de dimension...

1.1.1 SVD simple

On se place dans le cas où les métriques des espaces $\mathbb{R}^n$ et $\mathbb{R}^p$ sont canoniques.

Définition 1 *Soit une matrice $X_{n\times p}$. On définit la norme de Frobenius dans $\mathcal{M}_{n\times p}(\mathbb{R})$ par :*

$$\| X \|_F^2 = tr(X^T X) = tr(X X^T).$$

Proposition 1 (Décomposition en valeurs singulières simple)
Soit une matrice réelle $X_{n\times p}$ de rang r. On note X^T sa transposée. Il existe :

- *r réels strictement positifs $s_1, \ldots, s_r$, appelés valeurs singulières, on pose* $\Lambda = \begin{pmatrix} s_1 & \ldots & 0 \\ 0 & s_s & 0 \\ 0 & \ldots & s_r \\ , & & \end{pmatrix}$
- *une famille orthonormée $U = [u_1, \ldots, u_r]$ de $\mathbb{R}^n$,*
- *une famille orthonormée $V = [v_1, \ldots, v_r]$ de $\mathbb{R}^p$,*

tels que :

$$X = U\Lambda V^T = \sum_{s=1}^{r} s_s u_s v_s^T$$

On montre que

- *$s_1^2, \ldots, s_r^2$ sont les valeurs propres non nulles des matrices $X^T X$ et $X X^T$.*
- *$V = [v_1, \ldots, v_r]$ une matrice de vecteurs propres orthonormés de $X^T X$ associés à $s_1^2, \ldots, s_r^2$,*
- *$U = XV\Lambda^{-1} = [u_1, \ldots, u_r]$ une matrice de vecteurs propres orthonormés de $X X^T$ associés à $s_1^2, \ldots, s_r^2$,*

preuve : Semestre 1 analyse matricielle + TD

Corollaire 1 *Pour réaliser une SVD simple :*

1. *on recherche les valeur propres $s_1^2, \dots, s_r^2$ de X^TX, on en déduit Λ,*
2. *on recherche les vecteurs propres normés de X^TX associés à $s_1^2, \dots, s_r^2$ de X^TX. On en déduit V,*
3. *on en déduit $U = XV\Lambda^{-1}$.*

Exemple 1 *Réaliser la SVD de $A = \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix}$. Calculer la distance entre A et $s_1 u_1 v_1^T$.*

On reprend les 3 étapes :

1. $A^TA = \begin{pmatrix} 11 & 1 \\ 1 & 11 \end{pmatrix}$.
   $\lambda_1 + \lambda 2 = tr(A^TA) = 22$ et $\lambda_1\lambda_2 = |A^TA| = 120$ dont on déduit $\lambda_1 = 12$ et $\lambda_2 = 10$. Soit $\Lambda = \begin{pmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \end{pmatrix}$.
2. On en déduit les vecteurs propres normés :
   Pour $\lambda_1 = 12$, on a $-x + y = 0$ soit $v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
   Pour $\lambda_2 = 10$, on a $x + y = 0$ soit $v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.
   Soit $V = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$
3. D'où $U = AV\Lambda^{-1} = \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{12} & 0 \\ 0 & 1/\sqrt{10} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{6} & 2/\sqrt{5} \\ 2/\sqrt{6} & -1/\sqrt{5} \\ 1/\sqrt{6} & 0 \end{pmatrix}$
   $A = U\Lambda V = s_1 u_1 v_1^T + s_2 u_2 v_2^T$.

On en déduit $Z_1 = s_1 u_1 v_1^T = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 1 & 1 \end{pmatrix}$ la meilleure approximation de A de rang 1.
$||X - Z_1||^2 = tr((X - Z1)^T(X - Z_1)) = tr((s_2 u_2 v_2^T)^T(s_2 u_2 v_2^T)) = \lambda_2 = 10$
En particulier on vérifie facilement que : $||X||^2 = \sum_{i=1}^r ||s_i u_i v_i^T||^2 = \sum_{i=1}^r \lambda_i$.
Donc la distance est $\sqrt{10}$.

1.1.2 SVD généralisée

En pratique, on verra que les problèmes de réduction de dimension conduiront à choisir différentes métriques dans les espaces $\mathbb{R}^n$ et $\mathbb{R}^p$. Nous noterons D et Q les matrices symétriques définies positives définissant ces métriques. On définit alors le triplet statistique (X, Q, D) :

- X représente le tableau de données,
- Q la métrique pour le calcul de distance entre lignes,
- D, le plus souvent la matrice des poids des individus assimilable à une métrique pour le calcul de distance entre colonnes de X.

Définition 2 *La norme de Frobenius se généralise alors par :*

$$|| X ||^2_{Q,D} = tr(X^T D X Q) = tr(X Q X^T D).$$

On dispose alors d'une généralisation de la propriété précédente.

Proposition 2

Soit une matrice réelle $X_{n\times p}$ de rang r. Soit D et Q les métriques dans $\mathbb{R}^n$ et $\mathbb{R}^p$. Il existe :

- $\mathsf{V}_{p\times r} = [v_1, ..., v_r]$, *dont les colonnes sont les vecteurs propres $Q-$ orthonormés associés aux valeurs propres $s_1^2 \geq s_2^2 \geq ... \geq s_r^2$ de $X^T D X Q$.*
- $\mathsf{U}_{n\times r} = [u_1, ..., u_r]$, *dont les colonnes sont les vecteurs propres $D-$ orthonormés associés aux valeurs propres $s_1^2 \geq s_2^2 \geq ... \geq s_r^2$ de $X Q X^T D$.*
- Λ *une matrice diagonales des r valeurs singulières non nulles du triplet (X, Q, D) rangées dans l'ordre décroissant $s_1 \geq s_2 \geq ... \geq s_r$.*

tels que X se décompose en :

$$X = \mathsf{U}\,\Lambda\,\mathsf{V}^T = \sum_{s=1}^{r} s_s\, u_s\, v_s^T.$$

On réalise la SVD simple de $D^{1/2} X Q^{1/2}$.
Ici $D^{1/2}$ et $Q^{1/2}$ ont un sens car D et Q sont des métriques donc des matrice symétriques définies positives dont toutes les valeurs propres sont strictement psitives. On a alors $D = PAP^T$ avec P une matrice orthogonale et A une matrice diagonale decoefficients strictement positifs. On pose $D^{1/2} = PA^{1/2}P^T$ avec $A^{1/2}$ la matrice diagonale dont les coefficients sont les racines carrés de ceux de A et $A^{-1/2}$ la matrice diagonale dont les coefficients sont les inverses des racines carrés de ceux de A.

On obtient $D^{1/2} X Q^{1/2} = U_* \Lambda_* V_*^T$ avec Λ_* les racines des valeurs propres de

$$(D^{1/2} X Q^{1/2})^T D^{1/2} X Q^{1/2} = Q^{1/2} X^T D X Q^{1/2}$$

et V_* les vecteurs propres orthonormés.

On a alors : $Q^{1/2} X^T D X Q^{1/2} V_* = V_* \Lambda_*^2$, par produit à gauche par $Q^{-1/2}$ on a :

$$X^T D X Q^{1/2} V_* = Q^{-1/2} V_* \Lambda_*^2$$

soit

$$X^T D X Q Q^{-1/2} V_* = Q^{-1/2} V_* \Lambda_*^2.$$

Par comparaison avec $X^T D X Q V = V \Lambda$ de la SVD de (X, Q, D), on en déduit que $\Lambda_* = \Lambda$ et que les vecteurs propres V et $Q^{-1/2} V_*$ sont proportionnels.
La Q norme des vecteurs V_* est

$$(Q^{-1/2} V_*)^T Q Q^{-1/2} V_* = V_*^T V_* = I_r,$$

donc $V = Q^{-1/2} V_*$ puis par symétrie $U = D^{-1/2} U_*$.

Conclusion : si Λ_*, V_*, U_* sont la DVS simple de alors $D^{1/2} X Q^{1/2}$ alors $\Lambda = \Lambda_*$, $V = Q^{-1/2} V_*$, $U = D^{-1/2} U_*$ est la DVS généralisée de (X, Q, D).

Ainsi pour réaliser une SVD généralisée :

1. on réalise la SVD simple de $D^{1/2}XQ1/2$, de décomposition U_*, Λ_*, V_*.

2. la SVD généralisée du triplet (X, Q, D) est alors $U = D^{-1/2}U_*$, Λ_*, $V = Q^{-1/2}V_*$

On définit un schéma de dualité.

$$\begin{array}{ccc} \mathbb{R}^p & \underset{\rightarrow}{Q} & \mathbb{R}^{p*} \\ {}^TX \uparrow & & \downarrow X \\ \mathbb{R}^{n*} & \underset{\leftarrow}{D} & \mathbb{R}^n \end{array}$$

Un seul des 4 systèmes d'axes permet d'obtenir les 3 autres :

$$\mathsf{V}^* = Q\mathsf{V} \qquad \mathsf{U} = X\mathsf{V}^*\Lambda^{-1} \qquad \mathsf{U}^* = D\mathsf{U} \qquad \mathsf{V} = X^T\mathsf{U}^*\Lambda^{-1}$$

1.1.3 optimisation

Le problème que nous posons est de déterminer dans $\mathcal{M}_{n\times p}(\mathbb{R})$ une matrice de rang donné la plus proche possible d'une matrice donnée X au sens de la norme de Frobenius.

On dispose alors du théorème suivant pour trouver la solution optimale :

Proposition 3 *Théorème d'approximation d'Eckart-Young.*
Soit une matrice réelle $X_{n\times p}$ de rang r et deux matrices réelles $Q_{p\times p}$ et $D_{n\times n}$ définissant toutes deux des métriques sur $\mathbb{R}^p$ et $\mathbb{R}^n$ respectivement.
On note $X = \mathsf{U}\,\Lambda\,\mathsf{V}^T = \sum_{i=1}^{r} s_i\, v_i\, u_i^T$ la DVS de $(X\,Q\,D)$ et k un entier inférieur ou égal à r.
On note $\mathsf{V^k} = [v_1, ..., v_k]$ et $\mathsf{U^k} = [v_1, ..., v_k]$ les matrices extraites de V et U et $\Lambda = diag(s_1, \ldots, s_k)$ la matrice diagonale des k premières valeurs singulières.

On cherche un élément Z_k de l'ensemble noté E_k des matrices réelles de $\mathcal{M}_{n\times p}(\mathbb{R})$ de rang k le plus proche de X au sens de la norme $\|X\|^2_{Q,D} = tr(X\,Q\,X^T\,D)$. On a alors :

- $min_{E_k}\|X - X_k\|^2_{Q,D} = \sum_{i=k+1}^{r} s_i^2$,
- *l'optimum est atteint par la DVS incomplète de rang k : $Z_k = \mathsf{U_k}\,\Lambda_k\,\mathsf{V_k}^T$.*

Preuve dans le cas $Q = I_p$ et $D = I_n$ (+ cours analyse matricielle) :

Soit une matrice $X_{n\times p}$, $p \leq n$, de rang $r \leq p$ (sinon on prend A^T), et de SVD complète $\Lambda_{n\times p}, U_{n\times n}, V_{p\times p}$ avec U et V des matrices orthogonales $U^TU = I_n$, avec $X = U\Lambda V^T$.
On recherche une matrice X_k de rang $k \leq r$ minimisant $||X - X_k||$.
Par invariance de la norme de Frobenius par produit par une matrice orthogonale, on en déduit :

$$||X - X_k|| = ||U^TXV - U^TX_kV|| = ||\Lambda - U^TX_kV|| = ||\Lambda - Y||$$

avec $Y = U^TX_kV$ de rang k.
$||\Lambda - Y||^2 = tr((\Lambda - Y)^T(\Lambda - Y))$ avec $\Lambda - Y = \begin{pmatrix} s_1 - y_{11} & y_{11} & \dots \\ y_{21} & s_2 - y_{22} & \dots \\ y_{31} & y_{32} & \dots\dots \end{pmatrix}$. Les éléments de la diagonale de sont $(\Lambda - Y)^T(\Lambda - Y)$ sont alors :
Pour $i \leq r$, $(s_i - y_{ii})^2 + \sum_{j\neq i} y_{ij}^2$
Pour $i > r$, $y_{ii}^2 + \sum_{j\neq i} y_{ij}^2$ car $s_i = 0$
Donc la trace est

$$\sum_{i=1}^{r}(s_i - y_{ii})^2 + \sum_{i=r+1}^{p} y_{ii}^2 + \sum_{i,j,i\neq j} y_{ij}^2$$

On admettra que le minimum sous la contrainte d'un rang k est bien atteint pour :
$\begin{cases} y_{ii} & = s_i \text{ pour } i \leq k \text{ minimisant } \sum_{i=1}^{r}(s_i - y_{ii})^2 \\ y_{ii} & = 0 \text{ pour } i > k \\ y_{ij} & = 0 \text{ pour } i \neq j \text{ annulant } \sum_{i,j,i\neq j} y_{ij}^2 \end{cases}$
On en déduit alors $Y = \begin{pmatrix} s_1 & 0 & \dots & \dots & \dots & \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & s_k & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \ddots & \vdots \\ \vdots & \dots & \dots & \ddots & \ddots & \vdots \end{pmatrix}$
Dont on déduit $Z_k = UYV^T = \sum_{i=1}^{k} s_iu_iv_i^T$ et $||X - Z_k||^2 = ||\sum_{i=k+1}^{r} s_ivu_iv_i^T||^2 = \sum_{i=k+1}^{r} s_i^2$ (calcul évident).

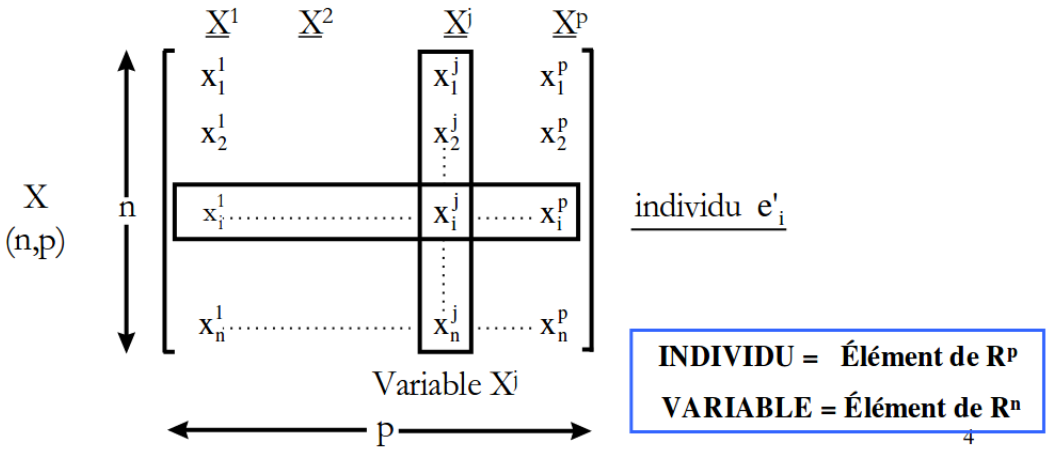
1.2 Inertie d'un nuage de points

On considère un triplet statistique (X, Q, D). Dans le cas classique :

- les lignes de X définissent des individus i associés à des points dans $\mathbb{R}^p$, de matrice colonne X_i,
- La distance entre points (individus) étant définie par Q, $d_Q^2(i, i') = (X_i - X_{i'})^TQ(X_i - X_{i'})$,
- D représente la matrice diagonale des poids du nuage,

Quand les colonnes sont des variables quantitatives centrées, l'interprétation devient particulière :

- les colonnes de X définissent des variables quantitatives centrées j associées à des vecteurs dans $\mathbb{R}^n$, de matrice colonne X^j,
- la D-norme de ces vecteurs est alors leur écart-type : $\sigma_j^2 = (X^j)^TDX^j = \sum_{i=1}^{n} p_i(x_i^j)^2$,
- le produit scalaire s'interprète comme la covariance entre variables : $COV(j, j') =< X^j, X^{j'} >_D$, la covariance est ainsi obtenue par projection orthogonale du vecteur X^j sur $X^{j'}$,
- si de plus les variables sont réduites, les covariances deviennent les corrélations entre variables.



1.2.1 Inertie totale et Matrice d'inertie

On s'intéresse dorénavant à un nuage de n points i de l'espace $\mathbb{R}^p$.

Définition 3 *Inertie d'un nuage*

On considère un nuage de points pondérés $\mathcal{N} = (i, p_i)_{i=1,\dots,n}$ dans un espace euclidien de métrique Q. On note D la matrice diagonale contenant les poids p_i des points i.

- *On appelle inertie totale du nuage de points $\mathcal{I}_g(\mathcal{N})$ la moyenne pondérée des carrés des distances des points au centre de gravité $g \sum_i p_i i$ de matrice colonne $X_g = \sum_i p_i X_i$:*

$$\mathcal{I}_g(\mathcal{N}) = \sum_{i=1}^n p_i d_Q^2(i,g) = \sum_{i=1}^n p_i ||X_i - X_g||_Q^2$$

- *Pour la métrique canonique de $\mathbb{R}^p$ et un nuage centré :*

$$\mathcal{I}_g(\mathcal{N}) = \sum_{i=1}^n p_i X_i^T X_i = \sum_{i=1}^n p_i (\sum_{j=1}^p (x_i^j)^2) = \sum_{j=1}^p \sigma_j^2$$

- *Plus généralement, on a pour un nuage centré :*

$$\mathcal{I}_g(\mathcal{N}) = || X ||_{Q,D}^2 = tr(X^T D X Q).$$

Preuve :

$I_T(\mathcal{N}) = \sum_{i=1}^n p_i d_Q^2(i,g) = \sum_{i=1}^n p_i X_i^T Q X_i$

$DX = \begin{pmatrix} p_1 X_1^T \\ \vdots \\ p_n X_n^T \end{pmatrix}$, $QX^T = \begin{pmatrix} QX_1 \dots QX_n \end{pmatrix}$, donc les coefficients de la diagonale de $DXQX^T$ sont $p_i X_i^T Q X_i$ et donc $\text{tr}(DXQX^T) = \sum_{i=1}^n p_i X_i^T Q X_i = I_T(\mathcal{N})$

Comme $\text{tr}(AB)=\text{tr}(BA)$ si les dimensions sont compatibles et $\text{tr}(A+B)=\text{tr}(A) + \text{tr}(B)$, on en déduit : $\text{tr}(DXQX^T) = \text{tr}(X^T D X Q)$.

Ainsi $I_T(\mathcal{N}) = \text{tr}(X^T D X Q) = ||X||_{Q,D}^2$. Cette égalité permettra de passer d'un problème d'optimisation d'inertie à un problème d'approximation de matrice (4.2).

Définition 4 *Matrice d'inertie*

- *Pour une métrique quelconque Q, on appellera matrice d'inertie du nuage de points $\mathcal{N} = (i, p_i)_{i=1,\dots,n}$ la matrice :*

$$M = X^T D X Q = SQ \text{ avec } S = X^T D X.$$

- *Pour un vecteur $Q-$unitaire u, ${}^t uQu = 1$, la grandeur $< u, Mu >_Q = u^T Q X^T D X Q u$ représente l'inertie projetée suivant cet axe u.*

- *Pour la métrique canonique de $\mathbb{R}^p$, la matrice d'inertie est la matrice de variance-covariance du nuage de points $\mathcal{N} = (i, p_i)_{i=1,\dots,n}$:*

$$M = S = X^T D X.$$

1.2.2 Propriété

Proposition 4 *Considérons un nuage de points de centre de gravité g. La moyenne des carrés de toutes les distances entre points est alors :*

$$\sum_{i,i'} p_i p_{i'} d_Q^2(i,i') = 2\mathcal{I}_g(\mathcal{N})$$

Preuve :

On a : $||x_i - x'_i||_Q^2 = ||x_i - g + g - x'_i||_Q^2 = ||x_i - g||_Q^2 + 2 < x_i - g, g - x'_i >_Q + ||g - x'_i||_Q^2$
Donc $\sum_{ii'} p_i p'_i ||x_i - g||_Q^2 + 2\sum_{ii'} p_i p'_i < x_i - g, g - x'_i >_Q + \sum_{ii'} p_i p'_i ||g - x'_i||_Q^2$
$= \sum_i p_i ||x_i - g||_Q^2 \sum_{i'} p'_i + 2\sum_{i'} p'_i < \sum_i p_i (x_i - g), g - x'_i >_Q + \sum_{i'} p'_i ||g - x'_i||_Q^2 \sum_i p_i = 2\mathcal{I}_g(\mathcal{N})$
car $\sum_i p_i = 1$ et $\sum_i p_i (x_i - g) = 0$.

Proposition 5 *Théorème de Huygens*

L'inertie du nuage de centre de gravité g par rapport à un point quelconque a est définie par :

$$\mathcal{I}_a(\mathcal{N}) = \mathcal{I}_g(\mathcal{N}) + d_Q^2(a,g)$$

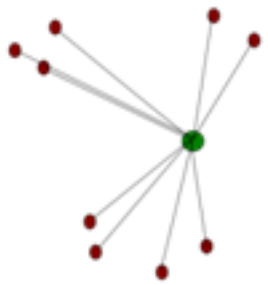
preuve : à faire

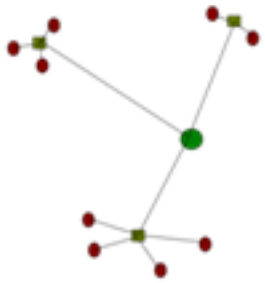
Proposition 6 *Décomposition de l'inertie*
Le nuage est décomposé en q groupes de points. Chaque groupe est noté $\mathcal{N}_k$ avec k variant de 1 à q. On note g_k le centre d'inertie du nuage $\mathcal{N}_k$ et $\pi_k = \sum_{i \in I(k)} p_i$ le poids de chaque groupe. $I(k)$ représente l'ensemble des indices i appartenant au groupe k.
L'inertie du nuage se décompose alors en :

$$\mathcal{I}_g(\mathcal{N}_n) = \sum_{k=1}^{q} \mathcal{I}_{g_k}(\mathcal{N}_k) + \sum_{k=1}^{q} \pi_k d_Q^2(g_k, g)$$

De même,la matrice de covariance se décompose en la somme de la matrice de covariance intra et la matrice de covariance inter.

preuve : à faire





1.2.3 Nuage gaussien

loi mulinormale

On suppose que les vecteurs lignes X_i suivent une loi multinormale à p dimensions, $\mathcal{N}(\mu, \Sigma)$.

Définition 5 *La fonction de densité d'une loi multinormale non dégénérée est :*

$$f_p(X_i) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} exp(\frac{^t(X_i - \mu)\Sigma^{-1}(X_i - \mu)}{2})$$

avce μ l'espérance et Σ la matrice de variance-covariance.

Proposition 7 *Pour un nuage gaussien, les estimateurs de μ et Σ sont respectivement g et $\frac{n}{n-1}S$, avec S la matrice de variance-covariance du nuage.*

Proposition 8 *Les lignes d'isodensité, $(X-\mu)^T\Sigma^{-1}(X-\mu) = k^2$, sont des ellipsoïdes dont les axes principaux sont les vecteurs propres de Σ et les longueurs des demi axes sont $k \times s_i$, avec s_i^2 les valeurs propres de Σ.*

Rappel : Soit Q une matrice symétrique définie positive. L'équation $x^TQx = 1$ définie une ellipse d'axe les vecteurs propres de Q et de demi axes les inverses des racines des valeurs propres. L'équation dans la base orthonormée de vecteurs propres et $\lambda_1 x_1^2 + \lambda_2 x_2^2 + .. = 1$.
Σ est une matrice de variance covariance supposée régulière (inversible). Σ et Σ^-1 sont donc symétriques définies positives. Σ et Σ^-1 admettent les mêmes vecteurs propres avec des valeurs propres inversées (vérification immédiate).
L'équation $(X-\mu)^T\Sigma^{-1}(X-\mu) = k^2$ est l'équation d'une ellipse dont les axes sont définis par les vecteurs propres et les demi axes par les racines des valeurs propres de Σ, l'équation dans un bon repère est $x_1^2/\lambda_1 + x_2^2/\lambda_2 + .. = k^2$.

Remarque 1 *Pour un vecteur u unitaire, $u^Tu = 1$, Xu représente l'image de u par X ou les affixes des projections du nuage sur l'axe u. $u^TSu = U^TX^TDXu$ représente ainsi l'inertie projetée suivant cet axe.*

L'image de la sphère unité, $u^Tu = 1$ par $S^{1/2} = V\Lambda V^T$ est alors une ellipse ou ellipsoïde d'équation $u^TS^{-1}u = 1$ donnant la forme générale du nuage de points ainsi que ses axes principaux et valeurs singulières. Chaque point de l'ellipse est l'image d'un point de la sphère et représente l'inertie projetée sur l'axe correspondant.

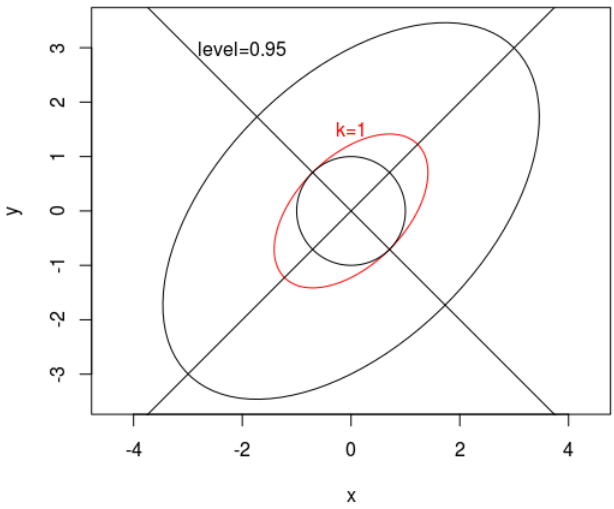
Exemple 2 *Etudier l'ellipse de concentration engendrée par*

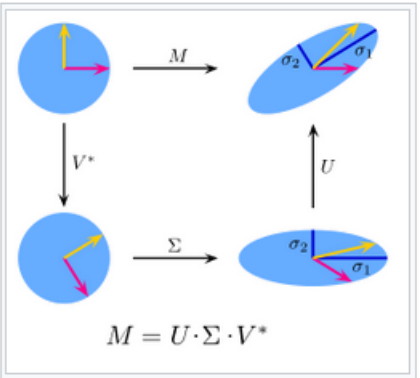
$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

> library(ellipse) ; S=matrix(c(2,1,1,2),ncol=2)

> plot(ellipse(S,level=0.95),type='l',asp=1) ; lines(ellipse(S,t=1),col='red')

> abline(0,1) ; abline(0,-1) ; text(0,1.5, labels='k=1',col='red') ; text(-2,3, labels='level=0.95')





Décomposition en valeurs singulières dans le cas d'une matrice réelle à 2 dimensions *M*. Cette transformation déforme, par exemple, un cercle unitaire bleu ci-dessus à gauche en une ellipse dans le coin supérieur droit de l'image. La transformation *M* peut alors être décomposée en une rotation V^* suivie d'une compression ou étirement Σ le long des axes de coordonnées suivie en fin d'une nouvelle rotation *U*. Les valeurs singulières σ_1 et σ_2 correspondent aux longueurs des grand et petit axes de l'ellipse.

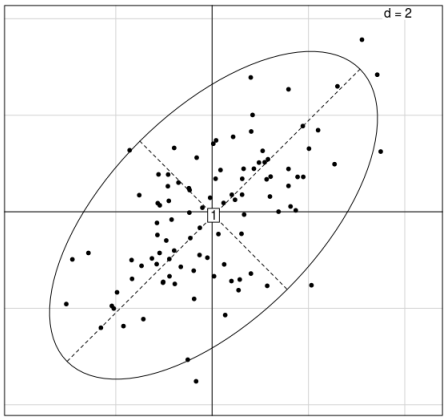
Ellipse de confiance

Proposition 9 *Soit n v.a. X_i suivant la loi normale $\mathcal{N}_p(\mu, \Sigma)$. On note g leur moyenne et S la matrice de variance-covariance :*

- $\sqrt{n}(g-\mu)$ *suit une loi* $\mathbb{N}_p(0, \Sigma)$,
- $n(g-\mu)^T\Sigma^{-1}(g-\mu)$ *suit une loi du* χ^2 *à* p *ddl,*
- $\dfrac{n-p}{p}(g-\mu)^T S^{-1}(g-\mu)$ *suit la loi* F *à* $p, n-p$ *ddl.*

Proposition 10 *On en déduit alors pour un niveau de confiance* $1-\alpha$*, les ellipsoïdes de confiance suivantes :*

- ${}^t(g-\mu)\Sigma^{-1}(g-\mu) = k$ *avec* $k = \dfrac{\chi^2(1-\alpha)}{n}$,
- ${}^t(g-\mu)S^{-1}(g-\mu) = k$ *avec* $k = \dfrac{p}{n-p}F_{p;n-p}(1-\alpha)$
- *En première approximation, en prenant* $k = 1$, $k = 2$, $k = 2.5$, *on obtient des ellipsoïdes de confiance à* 39%, 86%, 95% *respectivement* $(1-\alpha = 1-exp(-k^2/2)$ *pour* $p = 2)$.




```
> library(ade4) ; library(MASS)
> w=mvrnorm(100,c(0,0),S) ; s.class(w,as.factor(rep(1,100)),cell=2.5,cstar=0)
```

1.3 Réduction de dimension : application de la SVD

L'objectif de la réduction de dimension ici consiste à construire un nouveau nuage de points par projection orthogonale du nuage de points défini par X dans un sous espace de dimension réduite de manière à déformer le moins possible la position des points les uns par rapport aux autres. On construit ainsi une image simplifiée du nuage initiale la plus fidèle possible au nuage initial. En particulier, on revient toujours à une ou plusieurs images planes.

La détermination d'une solution optimale est obtenu par application de la SVD. Le choix de Q et D permet la construction de nombreuses méthodes : analyse en composantes principales, analyse des correspondances, analyse des correspondances multiples, analyse discriminante...

1.3.1 Rappels sur les projections

Proposition 11 *Soit F et G deux sous espaces vectoriels supplémentaires d'un espace vectoriel E.*
Les applications p et q de E dans E définies par :

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

sont linéaires et vérifient :

- $p^2 = p, q^2 = q$ *(idempotence),*
- $p \circ q = q \circ p = 0$,
- $p + q = Id_E$,
- $Im(p) = F = Ker q$ *et* $Im(q) = G = Ker p$.

On dit que p est la projection sur F parallèlement à G et inversement pour q.

Proposition 12 *Soit F un sous espace vectoriel de $E = R^n$*
Soit F=Vect$\{u_1, \ldots, u_r\}$ avec $U = [u_1, \ldots, u_r]$ la matrice d'une base orthonormée de F, $U^T U = I_r$, r dimension de F. On note $P = UU^T = \sum_{i=1}^{r} u_i u_i^T$.

- *P est l'opérateur de projection orthogonale sur F,*
- $P^2 = P$,
- *P admet exactement r valeurs propres non nulles égales à 1,*
- $Im(p) = F$ *et* $Ker(p) = F^{\perp}$,
- $P^T = P$.

Exemple 3 *Calculer l'opérateur de projection sur F=Vect$((1,1,0,0),(0,1,1,1)$. Déterminer les coordonnées de la projection orthogonale de $A(1,1,1,1)$ sur F.*

Construction d'une base orthonormée gram schmidt
e1=u1/||u1||
v2=u2 - <e1,u2> e1 et e2=v2/||u2||
> U
[,1] [,2]
[1,] 0.7071068 -0.3162278
[2,] 0.7071068 0.3162278
[3,] 0.0000000 0.6324555
[4,] 0.0000000 0.6324555
> P=Ut(U) ;P
[,1] [,2] [,3] [,4]
[1,] 0.6 0.4 -0.2 -0.2
[2,] 0.4 0.6 0.2 0.2
[3,] -0.2 0.2 0.4 0.4
[4,] -0.2 0.2 0.4 0.4

En statistiques, le calcul des distances est basé sur des métriques et des produits scalaires parfois différents de ceux usuels, la distance du χ^2 en analyse des correspondances, la distance de Mahalanobis en analyse discriminante...

En notant Q la métrique, on rappelle que :

- Q est symétrique, définie positive,
- le produit scalaire $< x, y >_Q = x^T Q y$,
- la norme $||x||^2 = x^T Q x$,
- $\cos(x, y) = \dfrac{< x, y >_Q]}{||x|| \times ||y||}$.

Définition 6 *Soit F un sous espace de $E = R^n$ muni de la métrique Q. On définit $F^\perp = \{u \in E, < x, u >_Q = 0 \ \forall x \in F\}$, avec $E = F \oplus F^\perp$.*
Pour tout u de E, $u = x_F + x_{F^\perp}$. L'application $p(x) = x_F$ définit la projection sur F $Q-$orthogonale.

Proposition 13 *Soit π_F l'opérateur de projection $Q-$orthogonale sur F. π_F est Q symétrique : $Q\pi_F = \pi_F^T Q$.*

1.3.2 Optimisation

Définition 7 *Soit F un sous espace de E, et Π_F l'opérateur de projection orthogonale sur F.*
On note $\mathcal{N}_F = (h_i, p_i)_{i=1,\dots,n}$ le nuage des projetés de $\mathcal{N}$ sur F.
L'inertie totale de $\mathcal{N}$ suivant F est égale à l'inertie totale de $\mathcal{N}_F$.

L'objectif est ici de maximiser l'inertie totale d'un nuage de points suivant un sev F_k, de dimension k, noté $\mathcal{N}_{Fk}$

Proposition 14 *L'inertie projetée suivant deux sous espaces orthogonaux, F et G, est égale à la somme des inerties suivant chaque sous espace (Pythagore) :*
$\mathcal{I}(\mathcal{N}_{F+G}) = \mathcal{I}(\mathcal{N}_F) + \mathcal{I}(\mathcal{N}_G)$

Preuve : Théorème de Pythagore

Proposition 15 *Soit F_k un sous-espace de dimension k portant l'inertie maximale alors le sous espace F_{k+1} de dimension $k+1$ portant l'inertie maximale et la somme directe de F_k et d'un sous espace de dimension 1, orthogonale à F_k. Les solutions sont emboîtées.*

Preuve :

$dim F_{k+1} = k+1$ et $dim F_k^\perp = n - k$, on a : $dim(F_{k+1} \cap F_k^\perp) \geq 1$.
Soit u un vecteur appartenant à $F_{k+1} \cap F_k^\perp$, posons $F_{k+1} = u \oplus G$ avec G le supplémentaire $Q-$ orthogonal de u dans F_{k+1}. On pose alors $F = F_k \oplus u$.
Donc $\mathcal{N}_{F_k \oplus u} = \mathcal{N}_{F_k} + \mathcal{N}_u \geq \mathcal{N}_G + \mathcal{N}_u = \mathcal{N}_{F_{k+1}}$

1.3.3 Application de la SVD

Proposition 16 *La matrice $Z_k = \sum_{i=1}^k s_i u_i v_i^T$ obtenue à l'aide du théorème d'Eckart Young représente la projection Q-orthogonale optimale du nuage défini par X dans un espace de dimension k. On a alors :*

- *les axes dits factoriels de projection ont pour vecteurs directeurs les vecteurs propres orthonormés $v_1, \dots, v_r$,*
- *l'inertie projetée suivant un axe v_s est égale à s_s^2,*
- *l'affixe des projections des points sur un axe v_s est appelée sième composante principale et est égale à $F_s = XQv_s$,*
- *l'opérateur de projection sur l'axe v_s est $\Pi_{v_s} = v_s v_s^T Q$, l'opérateur de projection sur le sous espace optimale de rang k est $\sum_{s=1}^k v_s v_s^T Q$.*

Preuve :

On recherche le sous espace F de dimension $k \leq p$ tel que l'inertie de la projection orthogonale suivant ce sous espace soit maximal. On pose π_F l'opérateur de projection. Par le théorème de Pythagore, on a :
$I_T(\mathcal{N}) = I_F(\mathcal{N}) + I_{F^\perp}(\mathcal{N})$
La projection sur F de chaque point i est $\pi_F X_i$, la matrice du nuage est ainsi : $X\pi_F^T = \begin{pmatrix} (\pi_F X_1)^T \\ \vdots \\ (\pi_F X_n)^T \end{pmatrix}$
La projection sur $F^\perp$ de chaque point i est $\pi_{F^\perp} X_i = (I_p - \pi_F)$, la matrice du nuage projeté sur $F^\perp$ est ainsi $X(I_p - \pi_F)^T$.
On en déduit ainsi par application du théorème de Pythagore à :

$$||X||^2_{Q,D} = ||X\pi_F^T||^2_{Q,D} + ||X(I_p - \pi_F)^T||^2_{Q,D} = ||X\pi_F^T||^2_{Q,D} + ||X - X\pi_F^T||^2_{Q,D}.$$

Maximiser suivant F revient à minimiser $||X - X\pi_F^T||^2_{Q,D}$. Nous avons vu avec le théorème d'Eckart Young que le minimum est atteint pour $Z_k = \sum_{i=1}^k s_i u_i v_i^T$.

Montrons que Z_k est bien une projection du nuage sur le sous espace vectoriel de dimension k, $F_k = vect\{v_1, \ldots, v_k\}$ et donc de la forme $X\pi_F$.
$\pi_{F_k} = \sum_{i=1}^k v_i\, v_i^T Q$ car vect$\{v_1, \ldots, v_k\}$ est une Q-b.o.n. de F_k, donc $\pi_{F_k}^T = \sum_{i=1}^k Q v_i\, v_i^T$.
Ainsi $X\pi_{F_k}^T = \sum_{i=1}^r s_i u_i v_i^T (\sum_{j=1}^r Q v_j v_j^T) = \sum_{i=1}^k s_i u_i v_i^T = Z_k$ car $v_i^T Q v_j = \delta_{ij}$.
$Z_k = X\pi_F^T$ **est bien le projeté du nuage dans un sous espace de dimension k maximisant l'inertie projetée.**
L'inertie projetée est alors $\sum_{i=1}^k s_i^2 = \sum_{i=1}^k \lambda_i$.

1.4 BILAN - Analyse factorielle

Soit le triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. On définit :

- Λ la matrice diagonale des r valeurs valeurs singulières non nulles du triplet (X, Q, D) rangées dans l'ordre décroissant $s_1 \geq ... \geq s_r$,
- $V_{p\times r} = [v_1, ..., v_r]$, matrice des vecteurs propres $Q-$ orthonormées associés aux valeurs propres $s_1^2 \geq ... \geq s_r^2$ de $X^T D X Q$:

$$\mathbf{X^T D X Q V = V \Lambda^2} \text{ et } \mathbf{V^T Q V = I_r},$$

- $U_{n\times r} = [u_1, ..., u_r]$, matrice des vecteurs propres $D-$ orthonormées associés aux valeurs propres $s_1^2 \geq ... \geq s_r^2$ de XQX^TD :

$$\mathbf{X Q X^T D U = U \Lambda^2} \text{ et } \mathbf{U^T D U = I_r},$$

tels que X se décompose en :

$$\mathbf{X = U\, \Lambda\, V^T} = \sum_{\mathbf{i=1}}^{\mathbf{r}} \mathbf{s_i u_i v_i^T}.$$

L'optimisation de la projection du nuage des lignes (individus) suivant un sous espace est donné par :

- V définit les axes factoriels, v_s, suivant lesquels les inerties projetées sont maximales,
- l'inertie projetée sur un axe v_s est s_s^2,
- QV définit les facteurs principaux correspondant aux formes linéaires (combinaisons linéaires des colonnes) permettant le calcul des affixes des projections des lignes ,
- $\mathbf{F_L = XQV}$ définit les affixes des projections sur les axes factoriels appelées composantes principales ligne,

Il est possible de procéder de même pour les colonnes de la matrice. Il suffit de traiter le triplet (X^T, D, Q). L'optimisation de la projection du nuage des colonnes (variables) suivant un sous espace est donné par :

- U définit les axes factoriels, u_s, suivant lesquels les inerties projetées sont maximales,
- l'inertie projetée sur un axe u_s est s_s^2,
- DU définit les cofacteurs principaux correspondant aux formes linéaires (combinaisons linéaires des lignes) permettant le calcul des affixes des projections des colonnes,
- $\mathbf{F^C = X^T DU}$ définit les affixes des projections sur les axes factoriels appelées composantes principales colonne,

Il existe des relations de dualité entre les lignes et les colonnes d'un tableau. La représentation de ces deux nuages dans des représentations planes permet une analyse plus approfondies du tableau. L'interprétation de ces relations dépend de la nature du tableau et sera abordée dans les chapitres suivants. On obtient alors des formules de transition entre les deux nuages :

- les formules de transition entre U et V :

$$\mathsf{U} = \mathbf{XQV\Lambda^{-1}} \qquad \mathsf{V} = \mathbf{X^T DU\Lambda^{-1}},$$

- $F_L = XQV = XQX^T DU\Lambda^{-1} = U\Lambda^2\Lambda^{-1}$ soit :

$$\mathbf{F_L = U\Lambda} \text{ et } \mathbf{F^C = V\Lambda}$$

- $X^T DF_L = X^T DU\Lambda$ soit $X^T DF_L = F^C\Lambda$, soit :

$$\mathbf{F^C = X^T DF_L\Lambda^{-1}} \text{ et } \mathbf{F_L = XQF^C\Lambda^{-1}}$$

- On définit ainsi des relations entre les projections des individus et les projections des variables sur un axe donné. Dans le cas de métriques canoniques :

$$\mathbf{F^s(j)} = \frac{\mathbf{1}}{\mathbf{s_s}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{F_s(i)} \text{ et } \mathbf{F_s(i)} = \frac{\mathbf{1}}{\mathbf{s_s}} \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{F^s(j)}$$

preuves :

Formules de transition
On a : $X^T DXQV = V\Lambda^2$ et $XQX^T DU = U\Lambda^2$. On multiplie à gauche la première par XQ, soit :
$XQX^T DXQV = XQV\Lambda^2$, XQV apparaît comme base de vecteurs propres de XQX^T donc proportionnels à U.
Calculons la D norme : $(XQV)^T D(XQV) = V^T Q\, X^T DXQV = V^T QV\Lambda^2 = \Lambda^2$. Ainsi $U = XQV\Lambda^{-1}$ et par symétrie $\mathsf{V} = X^T D\mathsf{U}\Lambda^{-1}$

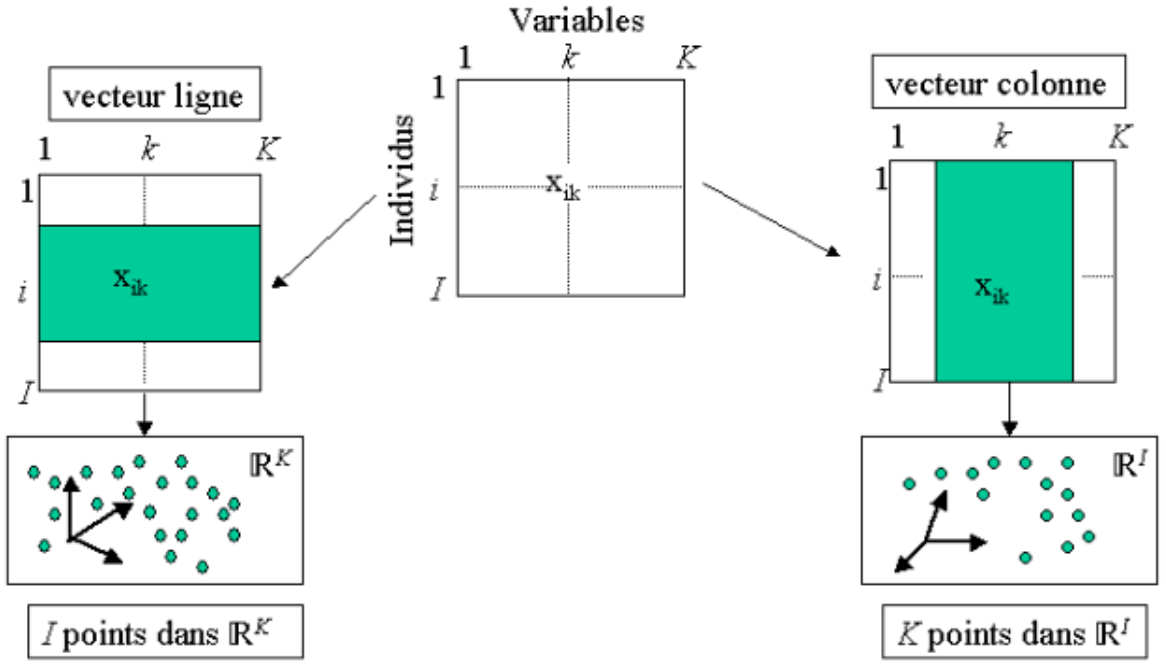
Formulaire Analyse factorielle

Soit le triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. On définit :

- Λ la matrice diagonale des r valeurs valeurs singulières non nulles du triplet (X, Q, D) rangées dans l'ordre décroissant $s_1 \geq ... \geq s_r$,
- $V_{p\times r} = [v_1, ..., v_r]$, matrice des vecteurs propres $Q-$ orthonormées associés aux valeurs propres $s_1^2 \geq ... \geq s_r^2$ de $X^T DXQ$: $\mathbf{X^T DXQV = V\Lambda^2}$ et $\mathbf{V^T QV = I_r}$,
- $U_{n\times r} = [u_1, ..., u_r]$, matrice des vecteurs propres $D-$ orthonormées associés aux valeurs propres $s_1^2 \geq ... \geq s_r^2$ de $XQX^T D$: $\mathbf{XQX^T DU = U\Lambda^2}$ et $\mathbf{U^T DU = I_r}$,

tels que X se décompose en :

$$\mathbf{X = U\,\Lambda\,V^T = \sum_{i=1}^{r} s_i u_i v_i^T}.$$



Composantes principales ligne
$F_L = XQV$

Composantes principales colonne
$F^C = X^T DU = V\Lambda$.

- les formules de transition entre U et V :

$$\mathsf{U} = \mathbf{XQV\Lambda^{-1}} \qquad \mathsf{V} = \mathbf{X^T DU\Lambda^{-1}},$$

- $F_L = XQV = XQX^T DU\Lambda^{-1} = U\Lambda^2\Lambda^{-1}$ soit :

$$\mathbf{F_L = U\Lambda} \text{ et } \mathbf{F^C = V\Lambda}$$

- $X^T DF_L = X^T DU\Lambda$ soit $X^T DF_L = F^C\Lambda$, soit :

$$\mathbf{F^C = X^T DF_L\Lambda^{-1}} \text{ et } \mathbf{F_L = XQF^C\Lambda^{-1}}$$

- On définit ainsi des relations entre les projections des individus et les projections des variables sur un axe donné. Dans le cas de métriques canoniques :

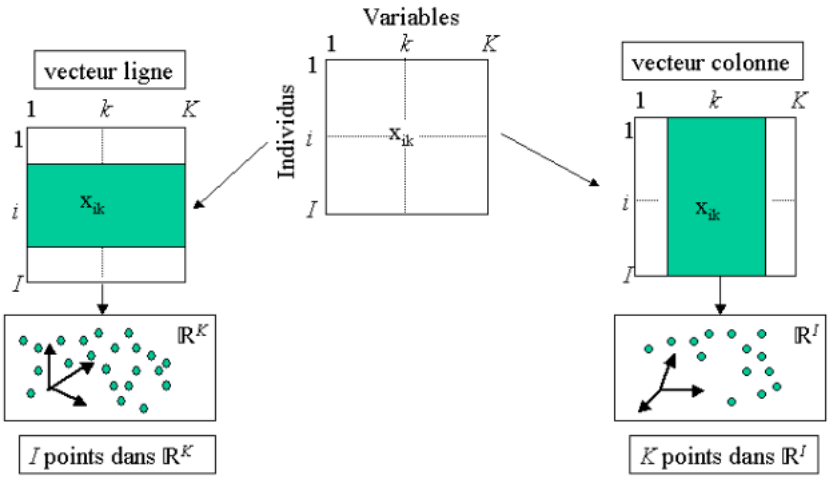
$$\mathbf{F^s(j) = \frac{1}{s_s}\sum_{i=1}^{n} F_s(i)} \text{ et } \mathbf{F_s(i) = \frac{1}{s_s}\sum_{j=1}^{p} F^s(j)}$$

Chapitre 2

Réduction de dimension 2 : Analyse en composantes principales

2.1 Introduction

Ce chapitre présente les principales étapes de l'analyse factorielle d'un tableau X, en s'intéressant plus particulièrement au cas particulier d'un tableau individus $\times$ variables quantitatives.



Pour étudier la dispersion des individus (lignes) dans un espace de grande dimension ou pour étudier simultanément les relations entre plusieurs variables, on doit se ramener à un espace de dimension réduite et les résultats sont représentés en dimension 2 (plans ou cartes factorielles) décrivant le mieux possible le tableau initial.

Comme nous l'avons vu au chapitre RD1, la solution mathématique revient à déterminer la matrice de rang k la plus proche de X, obtenue avec le théorème d'Eckart Young. La DVS du triplet (X, Q, D) nous donne ainsi directement les axes factoriels sur lesquels projeter les individus et les variables ainsi que la décomposition de l'inertie (l'information statistique du tableau) sur ces axes.

L'interprétation des résultats obtenus par l'ACP va permettre une analyse de la structure globale des données analysées. Cette interprétation est l'étape la plus importante de l'analyse et repose sur :

- l'utilisation d'aides à l'interprétation : τ, **qualité globale, cos2, qualité de représentation, ctr, contribution**,
- **les composantes principales** F_L **et** F^C, projections des lignes et colonnes sur les axes factoriels et dans des plans factoriels,
- l'utilisation de lignes et/ou colonnes en **supplémentaire**,
- **une approche pluridisciplinaire** reposant sur la bonne connaissance des domaines sur lesquels portent les données étudiées.

L'ACP est le fondement des autres analyses factorielles (AC, ACM, AFD) que nous aborderons spécifiquement. L'interprétation dépendra de la nature de X et du choix des métriques Q et D. Ses objectifs sont :

- décrire et visualiser les données dans des espaces de dimension réduite sur la base de cartes factorielles de dimension 2,
- construire un jeu de nouvelles variables décorrélées, appelées composantes principales, pouvant être utilisées pour d'autres méthodes,
- débruiter le tableau initial en ne conservant que les nouvelles variables significatives (analyse d'image, stockage,...).

L'ACP d'un tableau individus $\times$ variables quantitatives, notée ACP normée ou ACP non normée, que nous abordons dans ce chapitre conduit à des interprétations spécifiques :

- l'étude des lignes porte sur les distances entre individus (topologie), les points sont dans $\mathbb{R}^p$,
- l'étude des colonnes porte sur les relations entre variables quantitatives, en terme de corrélation, mais avec p variables simultanément.

Les références suivantes peuvent compléter ce chapitre :

- Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle. Dunod, Paris, 1995 (1ère édition), 2006 (4ème édition)
- `http://math.agrocampus-ouest.fr/infoglueDeliverLive/`
- `http://www.sthda.com`
- `https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acp.pdf`
- `http://www.arnaud.martin.free.fr/Doc/polyAD.pdf`

2.2 Nuage des individus et des variables

2.2.1 Nuage des individus

Définition 8 *Soit $\mathcal{N}_I$ le nuage de points pondérés, (i, p_i), $\sum_i p_i = 1$ avec $D = diag(p_1, \ldots, p_n)$. i représente la ième ligne du tableau de matrice colonne X_i. On définit ainsi l'espace des individus comme l'espace euclidien $\mathbb{R}^p$ de métrique Q dans lequel sont représentés les lignes du tableau.*

Définition 9 *En ACP normée ou non normée, la distance entre points est la distance euclidienne classique.*

On distinguera l'ACP normée et l'ACP non normée suivant la métrique Q utilisée, $Q = I_p$ en ACP non normée, $Q = diag(\frac{1}{\sigma_j^2})$ en ACP normée ce qui revient à réduire les variables. En pratique, pour l'ACP normée, on réduit au préalable les variables puis on réalise une ACP avec $Q = I_p$.

Remarque 2 *L'ACP normée, choix par défaut des logiciels, permet de donner le même poids à chacune des variables et d'éviter des biais du fait des unités ou de l'hétérogénéité des variables. Le poids d'une variable est égale à sa variance, si on change l'unité, par exemple kg en g, le poids est multiplié par 10^6. On utilisera donc l'ACP normée pour un tableau dont les variables sont hétérogènes.*

L'ACP non normée intervient éventuellement dans les cas où les variables sont toutes dans la même unité (notes, coût...).

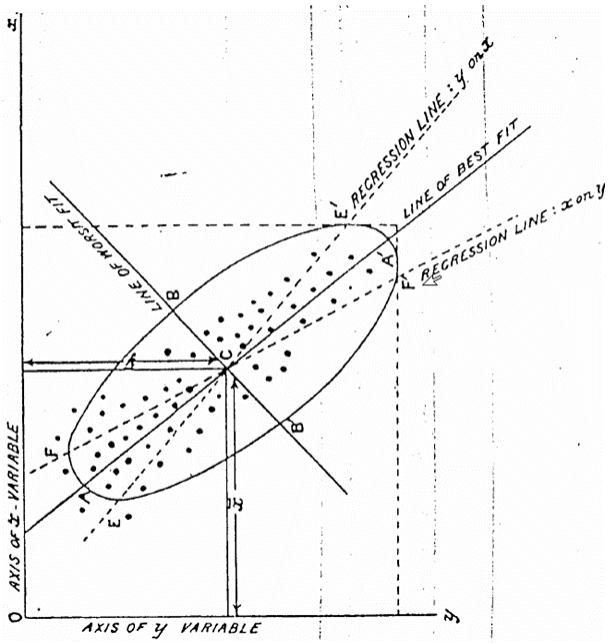
Proposition 17 *Soit le nuage $\mathcal{N}_I = (i, p_i)$. En ACP normée et non normée, on définit :*

- *son centre de gravité g de matrice $X_g = \sum_{i=1}^n p_i X_i = X^t D 1_n$, avec 1_n un vecteur colonne de 1.*

  Par la suite, on utilisera toujours le tableau centré $X - 1_n X_g^T$ que l'on notera également abusivement X. X **sera toujours centrée** *dans la suite du cours.*

- *sa matrice de variance-covariance $S = X^T D X$,*

- *l'inertie totale est la somme des variances des p variables,* $\sum_{j=1}^{p} \sigma_j^2$*, et de façon équivalente la trace de* S *soit la somme des valeurs propres de* S *donc* $\sum_{s=1}^{p} s_s^2 =$*.*
- **En ACP normée,** S **est alors la matrice des corrélations et l'inertie totale est le nombre de variables.**

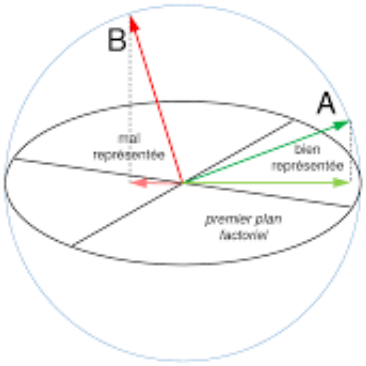


2.2.2 Nuage des variables

Définition 10 *Soit* $\mathcal{N}_J$ *le nuage des vecteurs colonnes* j *de matrice* X^j*. On définit l'espace des variables comme l'espace* $\mathbb{R}^n$ *de métrique* D *dans lequel sont représentés les colonnes du tableau.*

Proposition 18 *En ACP normée ou non normée, la métrique* D *conduit à une interprétation statistique spécifique des normes et des cosinus dans l'espace euclidien :*

- $|| X^j ||_D = \sigma_j$,
- $cov(j, j') =< X^j, X^{j'} >_D = (X^j)^T D X^{j'}$
- $\cos(X^j, X^{j'}) = cor(j, j')$



Proposition 19 *En ACP normée,l'interprétation devient :*

- $|| X^j ||_D = 1$,
- $cov(j, j') = \cos(X^j, X^{j'})$

- $\cos(X^j, X^{j'}) = cor(j, j')$

- *la projection orthogonale d'une variable sur une autre est aussi égale à la corrélation,*

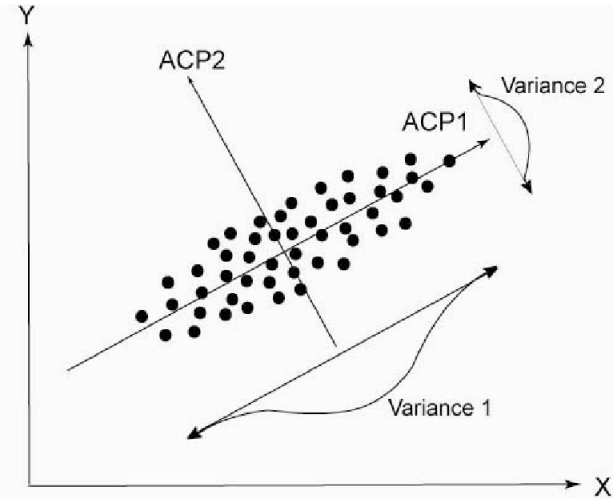
Remarque 3 *En ACP normée ou non normée, on peut définir une distance entre variables par* $d(j, j') = ||X^j - X^{j'}||_D = \sqrt{2(1 - cor(j, j'))}$*, utile en classification.*

2.3 ACP du triplet (X, Q, D)

2.3.1 Nuage des individus

Définition 11 *Critère d'ajustement*

L'objectif pour le nuage des individus est de projeter orthogonalement le nuage initial dans un sous espace de dimension réduite en conservant le mieux possible les distances entre points, ce qui est équivalent à conserver le mieux possible l'inertie totale du nuage projeté (RD1).



Proposition 20 *On appelle analyse en composantes principales du triplet* (X, Q, D) *les tableaux graphiques obtenus à partir de la décomposition en valeurs singulières de* (X, Q, D)*,* $X = U\Lambda V^t = \sum_i s_i u_i v_i^t$*.*

Les tableaux graphiques sont les représentations des individus dans les plans définis par les axes factoriels $(v_s, v_{s'})$*,* Q*-orthogonaux obtenus par SVD.*

On a alors :

- *le sous espace* $Vect\{v_1, \ldots, v_s\}$ *est le sous espace de dimension* s *optimisant l'inertie projetée suivant un sous espace de dimension* s*,*

- *l'inertie suivant ce sous espace est* $\lambda_1 + \cdots + \lambda_s = s_1^2 + \cdots + s_s^2$ *la somme des valeurs propres de la SVD, l'inertie suivant le sous espace* $Vect\{v_s, v_{s'}\}$ *est* $s_s^2 + s_{s'}^2$*,*

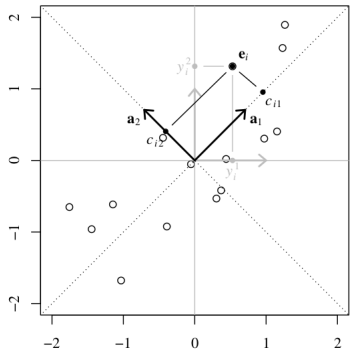
- v_s *et* $v_s^* = Qv_s$ *s'appellent les sième axe et facteur principaux,*

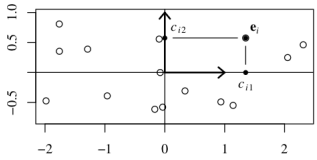
- *les projections sur* v_s*,* $F_s = XQu_s$ *s'appellent la sième composante principale et correspond aux affixes des points suivant cet axe,*

- *la variance de* F_s *est* s_s^2*,*

- *chaque composante principale correspond à une variable synthétique, obtenue par combinaison linéaire (facteur principal* v^**) des variables initiales,*

- *les nouvelles variables définies par les colonnes de* $F_L = XQV$ *sont non corrélées.*





2.3.2 Nuage des variables

Définition 12 *Critère d'ajustement*

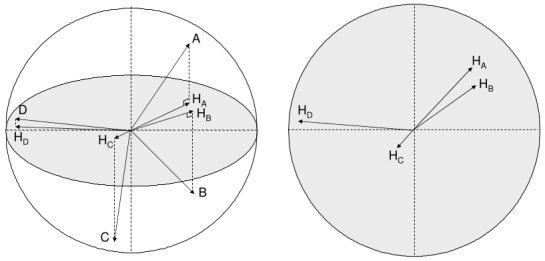
L'objectif pour le nuage des variables est de déterminer les combinaisons linéaires normées des variables non corrélées entre elles optimisant la variance, soit l'inertie projetée. Ce critère revient aussi à maximiser la somme des carrés des corrélations ou des covariances avec des axes de projection orthogonaux en ACP normée ou non normée.

Proposition 21 *Par analogie avec les lignes, la solution est obtenue par la DVS du triplet (X^T, D, Q). Par dualité, pour le nuage des vecteurs colonnes, représentant les variables quantitatives centrées, les combinaisons maximisant l'inertie projetée ont été déjà définies pour le nuage des individus, avec $v_s^* = Qv_s$.*

Les tableaux graphiques sont les représentations des variables dans les plans définis par les axes factoriels $(u_s, u_{s'})$, D-orthogonaux obtenus par SVD.

On a alors :

- *le sous espace Vect$\{u_1, \ldots, u_s\}$ est le sous espace de dimension s optimisant l'inertie projetée suivant un sous espace de dimension s,*
- *l'inertie suivant ce sous espace est $\lambda_1 + \cdots + \lambda_s = s_1^2 + \cdots + s_s^2$ la somme des valeurs propres de la SVD, l'inertie suivant le sous espace Vect$\{u_s, u_{s'}\}$ est $s_s^2 + s_{s'}^2$,*
- *u_s et $u_s^* = Du_s$ s'appellent les sième axe et facteur principaux,*
- *les projections sur u_s, $F^s = X^T D$, soit $\mathbf{F^s} = \mathbf{s_s v_s}$ s'appellent la sième composante principale et correspond aux affixes des points suivant cet axe,*
- *la composante principale d'une variable j sur un axe u_s est égale à la corrélation entre j et la sième composante principale F_s, $cor(j, F_s) = s_s v_s(j)$,*
- *la variance de F_s est s_s^2,*
- *chaque composante principale correspond à une variable synthétique, obtenue par combinaison linéaire (facteur principal v^*) des variables initiales,*
- *les nouvelles variables définies par les colonnes de $F^C = X^T DU = X^T DXQV\Lambda^{-1} = V\Lambda^2\Lambda^{-1} = V\Lambda$ sont non corrélées. On obtient ainsi directement $\mathbf{F^s} = \mathbf{s_s v_s}$.*



La corrélation entre PC^s et X^j en découle.

2.3.3 Bilan : calcul pratique d'une ACP normée

Pour réaliser l'ACP normée d'une matrice X, les étapes sont :

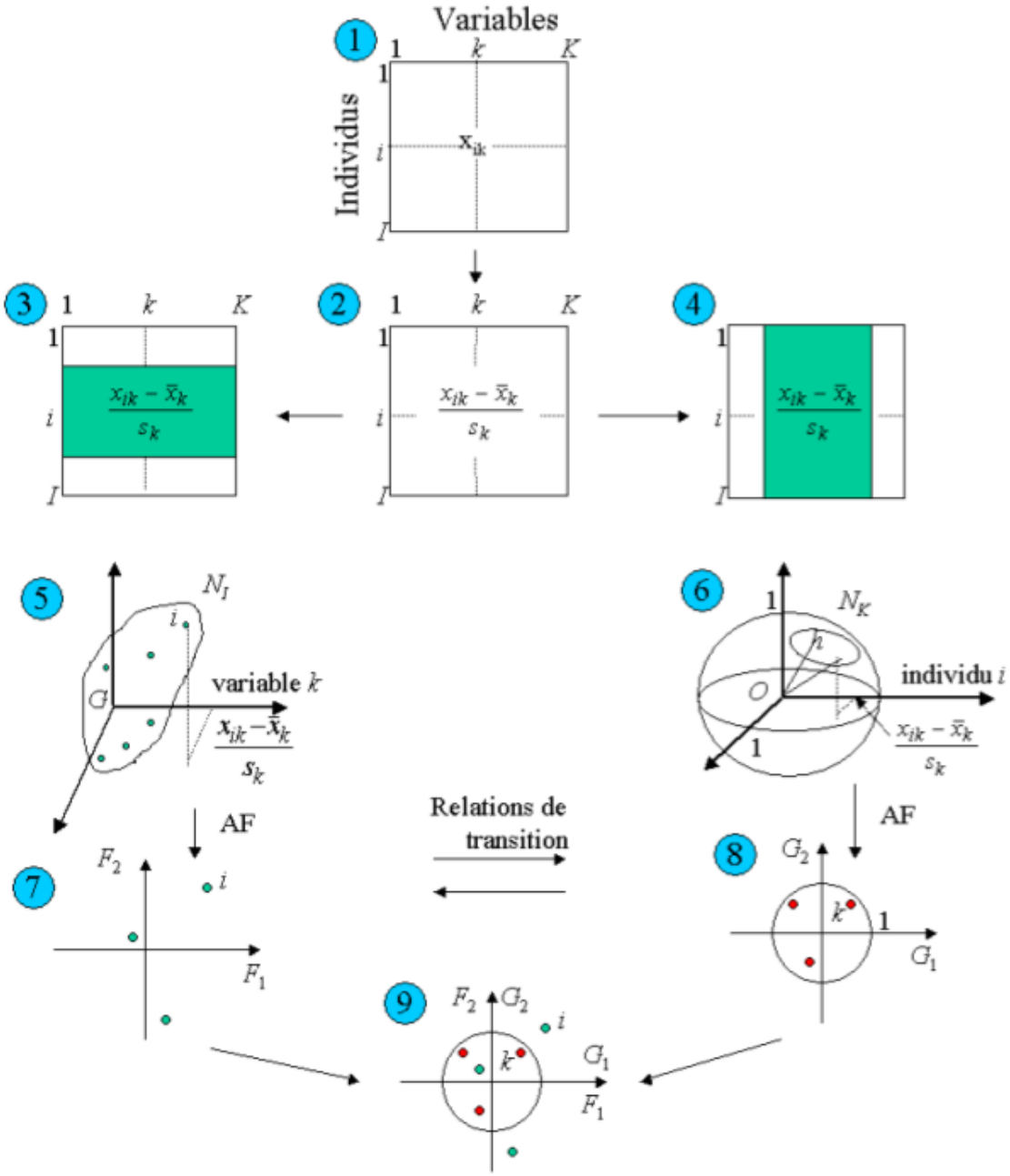
1. Calcul de la matrice centrée en ACP non normée (et réduite en ACP normée)
2. Calcul de la matrice de variance covariance (corrélation) $S = X^T DX$
3. Calcul des valeurs et vecteurs propres orthonormés de S, $\lambda_s = s_s^2$ et u_s pour s de 1 à k à définir.
4. Calcul des composantes principales lignes $F_s = Xv_s$
5. Calcul des composantes principales colonnes $F^s = s_s v_s$
6. application à :$X = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 2 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}$

2.3.4 Formules de transition en ACP normée et non normée

Proposition 22 *On déduit des formules de transition entre les deux nuages* $F_L = XF^C\Lambda^{-1}$ *et* $F^C = X^T DF_L\Lambda^{-1}$ *:*

- *Pour l'individu* i *:* $F_s(i) = \frac{1}{s_s}\sum_{j=1}^{p} x_i^j F^s(j)$*. Un individu* i *présente une forte valeur* $F_s(i)$ *si il possède de fortes valeurs* x_i^j *avec les variables* j *fortement corrélées* $(F^s(j))$ *à l'axe* u_s*.*
- *Pour la variable* j *:* $F^s(j) = \frac{1}{ns_s}\sum_{i=1}^{n} x_i^j F_s(i)$*. Une variable* j *présente une forte valeur* $F^s(j)$ *si il possède de fortes valeurs* x_i^j *avec les individus* i *présentant de fortes valeurs* $F_s(i)$ *sur l'axe* v_s*.*

Ces résultats justifient la dualité de l'interprétation sur les variables et les individus.



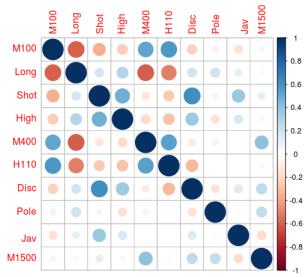
2.4 Résultats de l'ACP

Les résultats d'une ACP sont :

- la répartition de l'inertie projetée suivant les axes (eig),
- les paramètres cos2 et contrib pour les lignes (ind) et les colonnes (var),
- les projection des variables et des individus dans les plans factoriels,
- les projections des variables et individus supplémentaires

Nous étudierons les résultats à travers l'exemple decathlon décrivant les scores das 10 disciplines, ainsi que la note finale, le rang et le lieu de la compétition.

```
library(FactoMineR)
library(factoextra)
library("corrplot")
data(decathlon)
data=decathlon
rownames(data)=substr(rownames(data),1,4) #2 lettre par nom
colnames(data)=c('M100','Long','Shot','High','M400','H110','Disc','Pole',' Jav','M1500','Rank','Points','Comp')
corrplot(cor(data[,1 :10]))
```



La commande pour l'ACP normée est :

```
acp=PCA(data,quali.sup=13,quanti.sup=c(11,12),ind.sup=41)
```

indiquant que la colonne 13 est une variable qualitative supplémentaire, les colonnes 11 et 12 des variables quantitatives supplémentaires, la ligne 41 un individu supplémentaire.

2.4.1 qualité globale, eig, et choix des axes

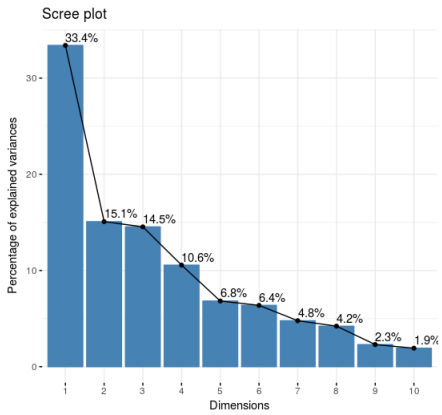
Définition 13 *On appelle qualité de représentation globale d'un sous espace E, noté τ_E la part d'inertie expliquée par la projection sur cet espace.*

On le note τ_s pour un axe v_s, avec $\tau_s = \frac{\lambda_s}{\sum_i \lambda_i^2}$. L'inertie dans un plan $(v_s, v_{s'}$ est alors $\tau_{s,s'} = \tau^s + \tau^{s'}$.

L'inertie expliquée par les k premiers axes factoriels est alors :

$$\tau_{1:k} = \frac{\sum_{s=1}^{k} \lambda_s}{\sum_{s=1}^{p} \lambda_s}$$

```
acp$eig
fviz_eig(acp, addlabels = TRUE)
```



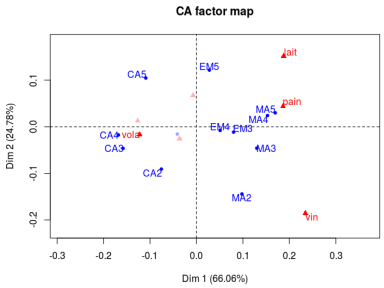
Le choix de la dimension retenue (nombre d'axes interprétés), q, dépend directement de ce paramètre. Il existe un certain nombre d'heuristique pour ce choix :

- **Part d'inertie** : on choisit les axes permettant d'expliquer une part d'inertie donnée,
- **Scree tree** : on détermine l'axe conduisant à une rupture (coude) dans la décroissance des valeurs propres (éboulis) à l'aide du changement de signe des différences secondes $(\lambda_{q-1} - \lambda_q) - (\lambda_q - \lambda_{+1}q) < 0$,
- **Valeur propre seuil** : on ne retient que les axes vérifiant $\lambda > 1$, règle de Kaiser, ou $\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$ règle de Karlis-Saporta-Spinaki pour limiter la permissivité de la règle de Kaiser,
- **Bootstrap** : on étudie par ré-échantillonnage la stabilité des axes.
- et d'autres encore

2.4.2 Paramètres d'interprétation

qualité de représentation : $\cos^2$

Les points ou vecteurs observés dans un plan de projection sont une image plus ou moins fidèle du point initial. Il est nécessaire de disposer d'un paramètre permettant de vérifier pour chaque individu ou variable si sa représentation est satisfaisante dans un plan donné.



Définition 14 *On appelle qualité de représentation, ou* $\cos^2$*, d'un individu ou d'une variable suivant un axe ou un plan sa part d'inertie projetée, égale au* $\cos^2$ *de l'angle de projection :*

$$\cos^2_s(i) = \frac{F_s^2(i)}{\sum_{k=1}^{p} F_k^2(k)} = \cos^2(\vec{OI}, O\vec{H_s}(i))$$

avec $\cos^2_{s,s'}(i) = \cos 2_s(i) + \cos 2_{s'}(i)$

Plus ce taux est proche de 1, plus le point est bien représenté dans l'espace étudié et est pertinent pour l'interprétation.

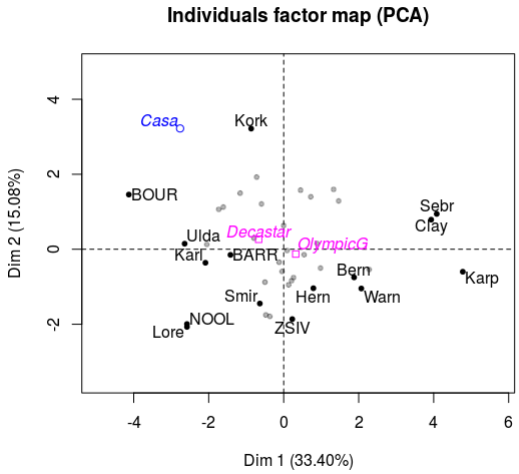
Remarque 4

- *En ACP normé,* $\cos^2_{s,s'}(j)$ *est égal au carré de la norme du vecteur, donc si le vecteur est proche du cercle des corrélations il est bien représenté, si il est de faible longueur, il est mal représenté.*
- *Pour les individus, il faut examiner les* $\cos^2$ *des individus utilisés pour l'interprétation. Certains logiciels utilisent une taille du point proportionnelle à* $\cos^2$*.*

  acpindcos2

  corrplot(acpindcos2, is.corr=FALSE)

  plot.PCA(acp,choix="ind",select="cos2 0.5")



- *Il n'existe pas de règle sur une valeur seuil, on se réfère à* $\tau_{s,s'}$.

Contribution : contrib

L'interprétation nécessite de donner une signification aux axes pour expliquer la position des individus en relation avec les variables. Dans cet objectif, on évalue pour chaque axe les variables et individus qui ont le plus d'influence à travers le paramètre contrib.

Définition 15 *On appelle contribution d'un individu ou d'une variable à un axe la part d'inertie du projeté par rapport à l'inertie de l'axe* λ_s.

$$contrib_s(i) = \frac{p_i F_s^2(i)}{\lambda_s}$$

On peut généraliser à un plan :

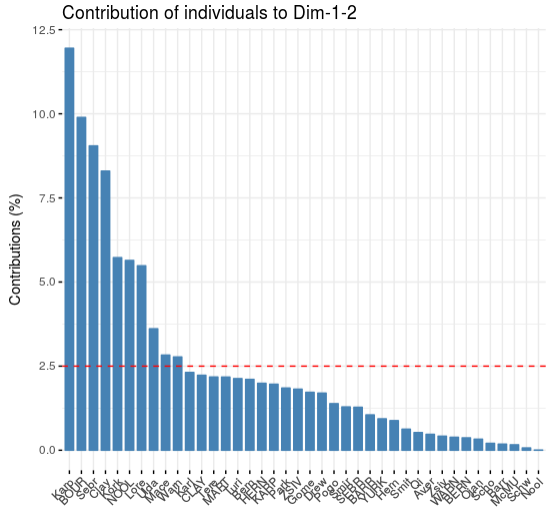
$$contrib_{s,s'}(i) = \frac{p_i(F_s^2(i) + F_{s'}^2(i))}{\lambda_s + \lambda_{s'}}$$

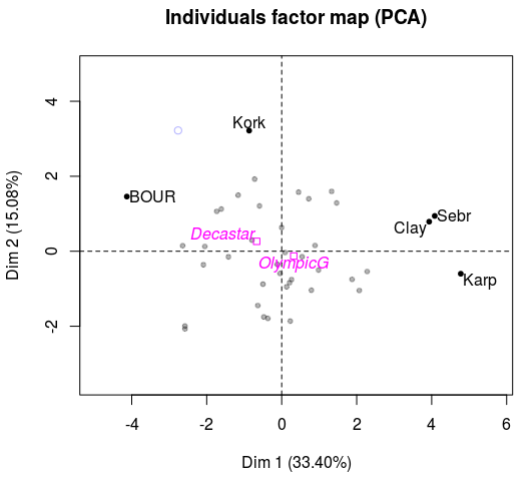
Plus ce taux est proche de 1, plus le point est influent dans l'espace étudié et permettra de lui donner une interprétation pertinente.

Remarque 5

- *En ACP normé, la contribution peut s'interpréter à l'aide de la corrélation avec l'axe.*
- *Parfois, un individu ou une variable présente une contribution très forte, le point ou la variable sont alors atypique et pourront éventuellement être mis en supplémentaire.*

```
acp$ind$contrib
corrplot(acp$ind$contrib, is.corr=FALSE)
fviz_contrib(acp, choice = "ind", axes = c(1,2))
plot.PCA(acp,choix="ind",select="contrib 5")
```





Individu, variable supplémentaires

Définition 16 *On appelle individu ou variable supplémentaires un individu ou variable qui n'est pas utilisé pour l'ajustement mais représenté a posteriori pour aider à l'interprétation.*

- *Un individu peut être mis en supplémentaire si il est atypique, sans rapport avec le jeu de données, calculer à partir des données (moyenne d'une classe d'individus)...*

  Après centrage et réduction, l'individu est projeté suivant u_s.

- *Pour une variable quantitative, on calcule les corrélations avec les composantes principales puis on la place dans le cercle des corrélations.*

- *Pour une variable qualitative, on peut représenter les individus moyens par modalité ou calculer les corrélations avec les axes .*

- *Il existe des tests permettant de vérifier si une variable qualitative présente des variations significatives suivant un axe donné. Nous y reviendrons en RD4 (ACM).*

  > acp$quali.sup$v.test

  Dim.1 Dim.2 Dim.3 Dim.4 Dim.5

  Decastar -1.582176 0.9320477 -0.55385 -0.9050013 2.053686

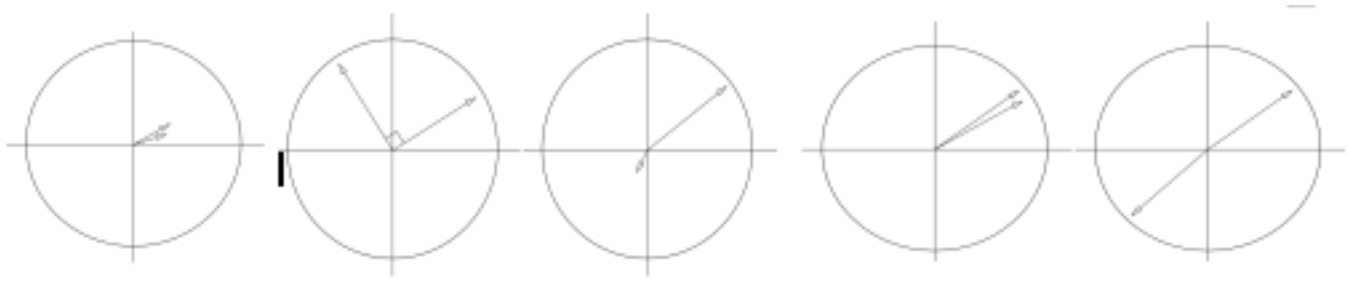
  OlympicG 1.582176 -0.9320477 0.55385 0.9050013 -2.053686

- *On utilise comme supplémentaires des éléments outlier, ou lorsque l'on dispose de trop de variables ou de variables qui n'ont pas de rapport direct avec les autres variables ou construites à partir d'elles.*

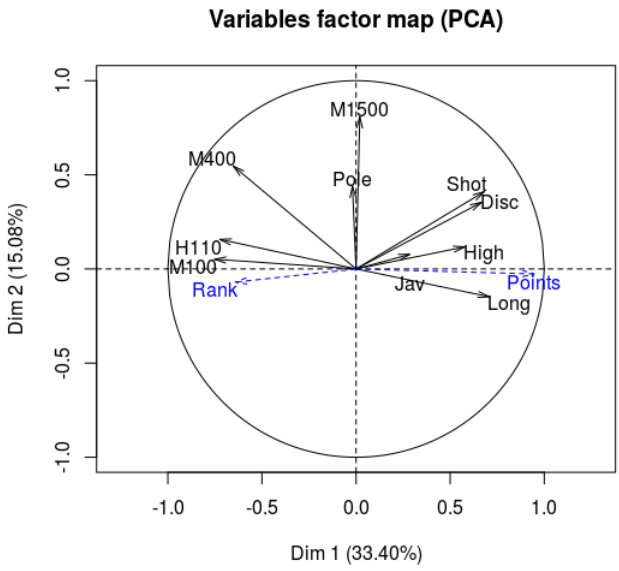
2.4.3 Projection des variables

En ACP normée, les variables sont représentées dans un cercle des corrélations pour faciliter l'interprétation :

- les variables bien représentées ont des vecteurs s'approchant du cercle ($||j||^2 = \cos 2(j)$ proche de 1),
- deux variables bien représentées ayant même direction et même sens sont corrélées positivement,
- deux variables bien représentées ayant même direction et sens contraire sont corrélées négativement,
- deux variables bien représentées ayant des directions orthogonales sont peu corrélées,
- une variable bien représentée et une variable mal représentées sont généralement peu corrélées,
- les variables quantitatives supplémentaires sont individualisées.



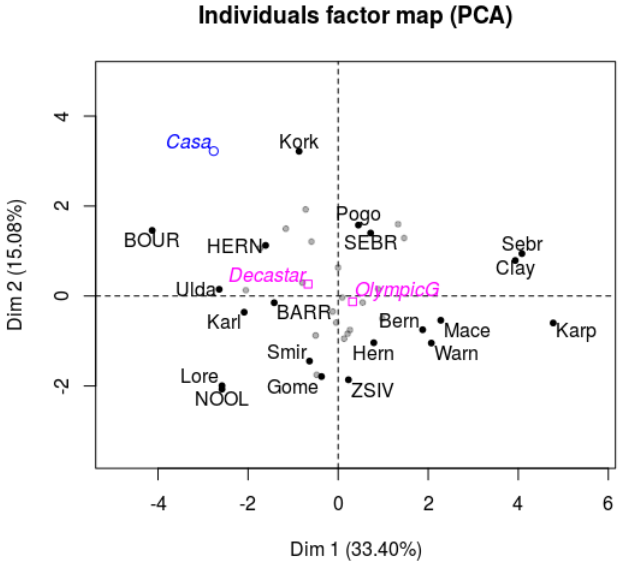
plot.PCA(acp,choix="var",select=")



2.4.4 Projection des individus

- On ne dispose plus du cercle des corrélations pour cos2 et il faut vérifier la qualité de représentation des individus,

  plot.PCA(acp,choix="ind",select="cos2 0.4")

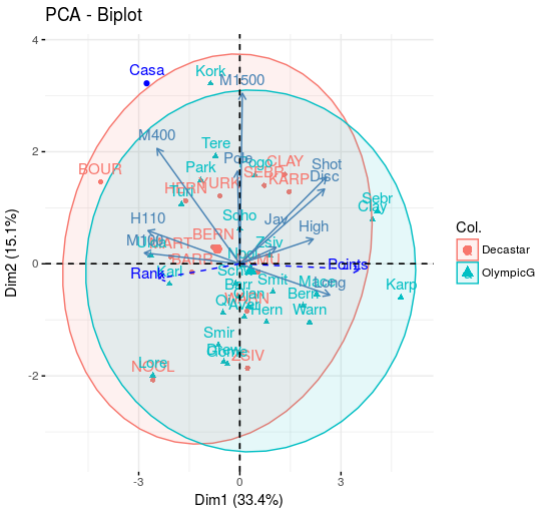


- les individus supplémentaires sont individualisés,

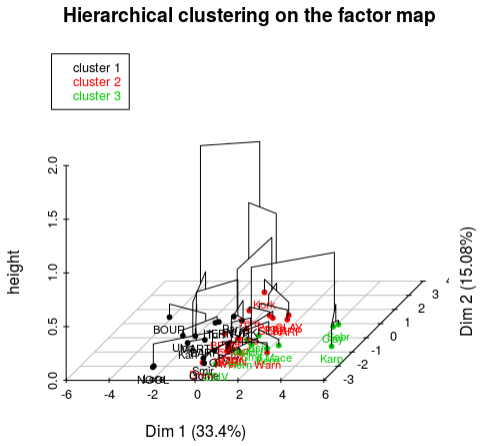
- les points moyens des classes des variables qualitativessont représentés,

- il est possible de différentier les classes d'une variable qualitative ou d'une classification par des ellipses et des couleurs différentes,

  fviz_pca_biplot (acp, col.ind = data$Comp[-41],addEllipses = TRUE)



- on est conduit souvent à réaliser une topographie explicative des individus en les regroupant par classe ou en réalisant une classification.



L'apprentissage de la méthode passe par l'étude de plusieurs exemples, à vous de jouer. On dispose également des commandes suivantes pour poursuivre l'exploration

```
library(explor)
explor(acp) # vous allez adorer
summary(acp) # résumé de l'ACP
dimdesc(acp) # aide à l'analyse
estim_ncp(data[,1 :10],scale=TRUE) # aide pour le choix des axes
```

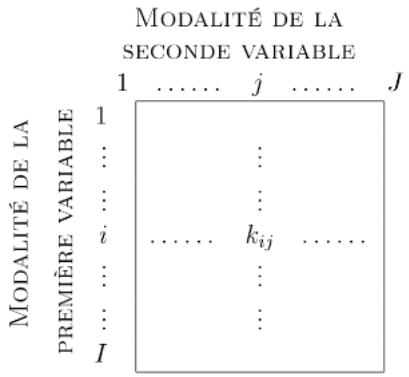
Chapitre 3

Réduction de dimension 3 : Analyse des correspondances

3.1 Introduction

Tableaux étudiés : L'AFC est utilisée pour l'étude de volumineux tableaux de contingence ou pour des tableaux dont les profils (distribution par ligne ou colonne) ont une signification :

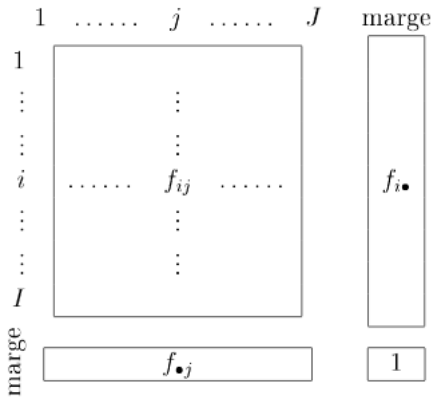
- des colonnes homogènes : dépenses, surface, prix, notes ...
- des questionnaires (ACM)
- des données logiques 0,1.



Notations :

- Les I lignes, i représentent les modalités d'une variable qualitative L,
- Les J colonnes, j représentent les modalités d'une variable qualitative C,
- Le tableau est constitué des effectifs $n_{i,j}$ des individus combinant les modalités i et j.
- On note n l'effectif total, N le tableau des effectifs, $\mathrm{F} = \frac{1}{n}N$ le tableau des fréquences relatives.

Remarque : La matrice F des fréquences relatives n'est pas à confondre avec les composantes principales notées F_L ou F^C.



- On note :
  - $\mathrm{F}_I = \mathrm{F}1_J = (f_{i.})_i$ avec $f_{i.} = \sum_j f_{ij}$ et $D_I = diag(f_{i.})$, les fréquences marginales lignes.
  - $\mathrm{F}_J = \mathrm{F}^T 1_I = (f_{.j})_j$ avec $f_{.j} = \sum_i f_{ij}$ et $D_J = diag(f_{.j})$, les fréquences marginales colonnes.

Objectifs : L'étude de tels tableaux a pour objectifs :

- d'étudier l'ensemble des liaisons entre les modalités des deux variables (plus ou moins forte association entre les modalités en regard de l'hypothèse d'indépendance), soit la correspondance entre les 2 variables,
- d'étudier les similarités entre les modalités d'une même variable au regard de l'autre variable,
- construire des variables quantitatives non corrélées à partir des variables qualitatives.

Les références suivantes peuvent compléter ce chapitre :

- Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle. Dunod, Paris, 1995 (1ère édition), 2006 (4ème édition)
- http ://www.sthda.com
- https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-explo-acp.pdf
- http ://www.arnaud.martin.free.fr/Doc/polyAD.pdf (dont sont issues les figures)

3.2 Nuage des profils

3.2.1 Etude élémentaire d'un tableau de contingence

La première question sur de tels tableaux pose sur l'existence d'une liaison entre variables qu'il est possible de tester par un test du Khi2. En absence de liaison, la méthode n'a pas d'intérêt.

On étudie alors l'écart à l'hypothèse d'indépendance : $e_{ij} = f_{ij} - f_{i.} \times f_{.j}$ et l'on peut quantifier la contribution de cet écart dans le calcul du Khi-2.

Exemple

On considère le tableau suivant sur la couleur des cheveux (B blond, C chatain, R roux) et la couleur des yeux (b bleu, m marron , v vert).

OBS	b	m	v	Total
B	20	20	10	
C	20	120	0	
R	10	0	10	
Total				

THEOR	b	m	v	Total
B				
C				
R				
Total				

χ^2	b	m	v	Total
B				
C				
R				
Total				

3.2.2 Nuage des profils

Définition 17 *On appelle profil ligne d'une modalité l_i, noté Li, la distribution conditionnelle à l_i de la variable C :*

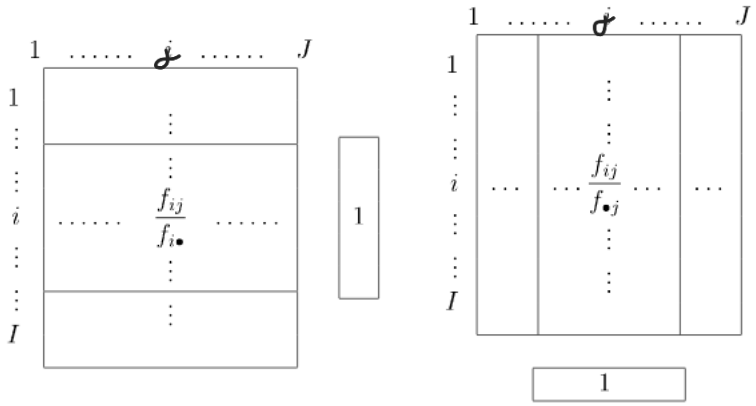
$$L_i = (f_{ij}/f_{i.})_j$$

La matrice des profils lignes est : $D_I^{-1}\mathrm{F}$

Par symétrie, on définit de même les profils colonnes.

$$C_j = (f_{ij}/f_{.j})_i$$

avec $C = D_J^{-1}\mathrm{F}^t$



Ligne	b	m	v	Total
B				
C				
R				
Total				

Colonne	b	m	v	Total
B				
C				
R				
Total				

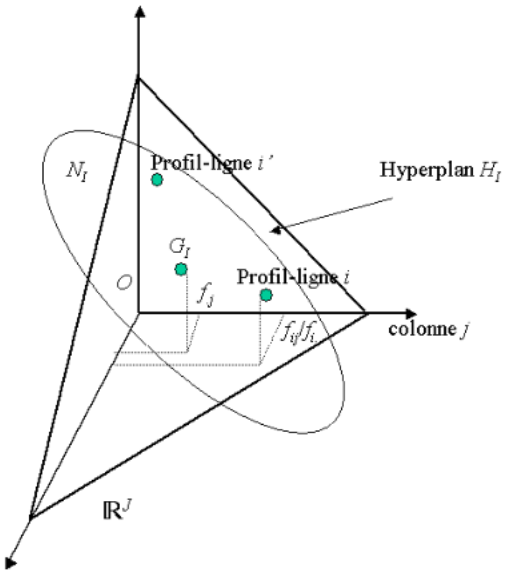
Définition 18 *Chaque profil ligne est associé à un point l_i de $\mathbb{R}^J$ avec un poids $f_{i.}$ et chaque profil colonne et associé à un point c_j de $\mathbb{R}^I$ avec un poids $f_{.j}$.*

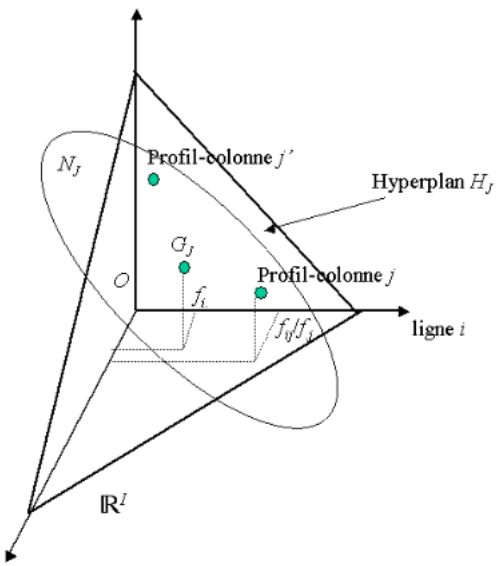
La métrique utilisé est dans chaque espace est la métrique du Khi 2 définie par D_J^{-1} pour l'espace des profils ligne et D_I^{-1} celui des colonnes.

La distance entre deux profils lignes est :

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

On définit ainsi deux triplets statistiques : (L, D_J^{-1}, D_I) *pour les profils lignes et* (C, D_I^{-1}, D_J) *pour les profils colonne.*





Remarque : En AFC, les tableaux L et C ne sont pas centrés, les profils sont dans un hyperplan du fait que la somme des colonnes vaut toujours 1 (chaque ligne d'un profil est une distribution conditionnelle).

3.2.3 Propriétés de la distance du χ^2.

Par analogie avec le test du chi-2, cette distance est choisie dans cette analyse pour calculer la différence entre distributions (les profils). Une des conséquences de cette distance est que pour une même différence, elle donnera plus de poids à une modalité faiblement représentée.

Par exemple pour une différence de 0.05, la distance sera $0.05^2/0.10 = 0.05$ pour une modalité à 0.10 alors qu'elle sera $0.05^2/0.80 = 0.006$ pour une modalité à 0.80.

Deux autres propriétés importantes justifient aussi son intérêt :

Proposition 23 *L'inertie totale du nuage est la valeur du test du chi-2 divisée par n :*

$$I_T = \frac{1}{n}\chi^2.$$

Contrairement à l'acp normée, l'inertie a une signification statistique en AFC. L'inertie expliquée représente les écarts à l ?hypothèse d'indépendance H_0.

Preuve

$$I_T = \sum_{i=1}^{I} f_{i.} d^2_{\chi^2}(i, g_I) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{f_{.j}} \frac{\left(f_{ij} - f_{i.}f_{.j}\right)^2}{f_{.i}} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i.} \frac{1}{n} \frac{\left(nf_{ij} - nf_{i.}f_{.j}\right)^2}{nf_{.i}f_{.j}}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} f_{i.} \frac{1}{n} \frac{n\left(n_{ij} - (n_{i.}n_{.j})/n\right)^2}{(n_{.i}n_{.j})/n} = \frac{1}{n} D_n^2$$

Proposition 24 *Equivalence distributionnelle*

Si deux modalités ont la même distribution, les nuages restent inchangés si on les regroupe en additionnant leur poids.

Les distances dans les deux nuages restent inchangées ainsi que l'inertie totale.

Preuve

Supposons que i et i' aient la même distribution, la distance entre $i + i'$ et i'' reste inchangée :

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

En effet, on se rappelle que par proportionnalité, on a :

$$\frac{f_{ij}}{f_{i.}} = \frac{f_{i'j}}{f_{i'.}} = \frac{f_{ij} + f_{i'j}}{f_{i.} + f_{i'.}}$$

3.3 Ajustement des nuages de profil

3.3.1 Ajustement

Définition 19 *Critère d'ajustement*

L'objectif pour chacun des nuages de profils est de projeter chi2-orthogonalement le nuage initial dans un sous espace de dimension réduite en conservant le mieux possible les distances entre points, ce qui est équivalent à conserver le mieux possible l'inertie totale du nuage projeté (RD1).

Proposition 25 *On appelle analyse des correspondances l'analyse en composantes principales conjointes des deux triplets statistiques (L, D_J^{-1}, D_I) pour les profils lignes et (C, D_I^{-1}, D_J) pour les profils colonne.*

Les tableaux graphiques sont les représentations des individus dans les plans définis par les axes factorielles $(v_s, v_{s'})$, Q-orthogonaux obtenus par svd.

On retrouve les différentes définitions et résultats de l'ACP.

Proposition 26 *En AFC, les tableaux L et F ne sont pas centrés, les profils sont dans un hyperplan et le premier axe factoriel relie l'origine du repère au centre de gravité du nuage égal au profil marginal d'inertie 1.*

En pratique, ce premier axe trivial, associé à la valeur propre 1, est éliminé de l'analyse automatiquement.

Preuve

Montrons que F_J est vecteur propre de X^TDXQ associé à la valeur propre 1.

$$X^TDXQ = F^TD_I^{-1}D_ID_I^{-1}FD_J^{-1} = F^TD_I^{-1}FD_J^{-1}$$

$$F^TD_I^{-1}FD_J^{-1}F_J = F^TD_I^{-1}F1_j = F^TD_I^{-1}F_I = F^T1_I = F_J$$

Le vecteur F_j est orthogonal à l'hyperplan contenant les profils lignes et n'intervient pas dans son analyse en composantes principales.
Ainsi au lieu de centré le nuage, on réalise l'ACP sur le nuage non centré et on élimine ce premier axe trivial.

3.3.2 Composantes principales F_L et F^C

Pour la dvs de (L, D_J^{-1}, D_I) :

Matrice d'inertie : $X^TDXQ = F^TD_I^{-1}D_ID_I^{-1}FD_J^{-1} = F^TD_I^{-1}FD_J^{-1}$

Vecteurs propres normés : $F^TD_I^{-1}FD_J^{-1} = V_L\Lambda^2$, $V_L^TD_J^{-1}V_L = I_r$

Les composantes principales sont alors : $F_L = XQV = LD_J^{-1}V_L = D_I^{-1}\mathrm{F}D_J^{-1}V_L$

Par analogie, pour la dvs de (C, D_I^{-1}, D_J) :

Matrice d'inertie : $XQX^tD = FD_J^{-1}F^tD_I^{-1}$
Vecteurs propres normés :$XQX^tDV^C = FD_J^{-1}F^tD_I^{-1}V^C = V^C\Lambda^2$ avec $V_C^TD_I^{-1}V_C = I$

Les composantes principales sont alors : $F^C = C^TDV = D_J^{-1}\mathrm{F}^TD_I^{-1}V_C$

3.3.3 Formules de transition

Proposition 27 *On obtient les formules de transition suivantes :*
$F^C = D_J^{-1}\mathrm{F}^TF_L\Lambda^{-1}$ *et* $F_L = D_I^{-1}\mathrm{F}F^C\Lambda^{-1}$, *dont on déduit :*

$$F_L^s(i) = \frac{1}{s_s}\sum_{j=1}^{J}\frac{f_{ij}}{f_{i.}}F_s^C(j)$$

$$F_s^C(j) = \frac{1}{s_s}\sum_{i=1}^{I}\frac{f_{ij}}{f_{.j}}F_L^s(i)$$

Preuve

$$FD_J^{-1}F^TD_I^{-1}V^C = V^C\Lambda^2$$

$$F^TD_I^{-1}FD_J^{-1}F^tD_I^{-1}V^C = F^TD_I^{-1}V^C\Lambda^2$$

Donc les valeurs propres sont les mêmes et :

$$V_C^TD_I^{-1}FD_J^{-1}F^TD_I^{-1}V^C = V_C^TD_I^{-1}V^C\Lambda^2 = \Lambda^2$$

Ainsi $V_L = F^TD_I^{-1}V^C\Lambda^{-1}$ et $F_L = D_I^{-1}FD_J^{-1}F^TD_I^{-1}V^C\Lambda^{-1} = D_I^{-1}V^C\Lambda$.

On en déduit que $D_J^{-1}F^TF_L = D_J^{-1}F^TD_I^{-1}V^C\Lambda = F^C\Lambda$

Proposition 28 *relations quasi barycentriques*

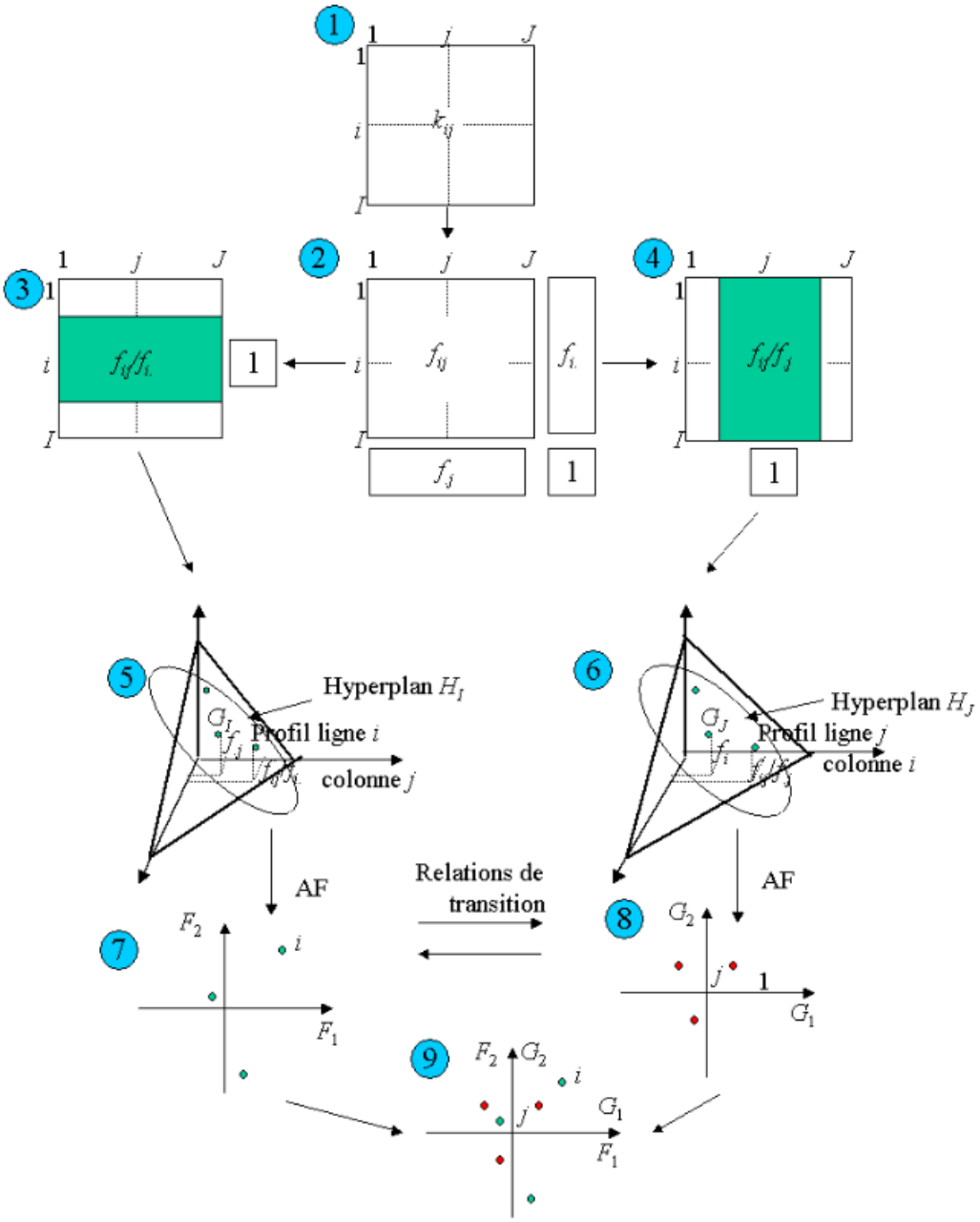
Les formules de transition montre que les coordonnées des projections d'une modalité sur un axes donnée sont les barycentres des coordonnées des projections des modalités de l'autre variable à un coefficient près $\frac{1}{s_s}$.

Ce résultat justifie la représentation simultanée des profils des 2 variables dans un même espace et l'interprétation des distances entre modalités des deux variables (correspondance) :

- *une faible distance indique une plus forte association par rapport à* H_0,
- *une forte distance indique une plus faible association par rapport à* H_0.

Remarque : Si le profil d'une modalité correspond à la fréquence marginale, le profil est confondu avec l'origine.

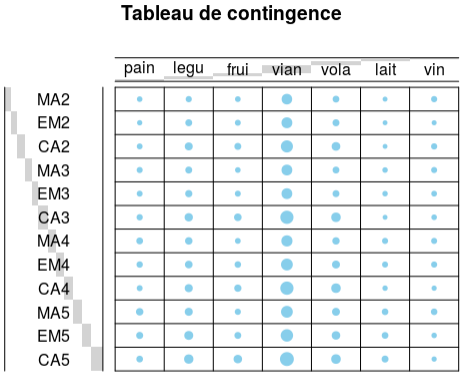
Bilan


1
2
3
4
5
6
7
8
9
Hyperplan H_I
Profil ligne i
colonne j
Hyperplan H_J
Profil ligne j
colonne i
AF
AF
Relations de transition

3.4 Interprétation

3.4.1 Présentation du tableau

```
tab=read.table('csp2.txt',h=T)
chisq=chisq.test (tab)
chisq$residuals
mosaicplot(tab,main='table')
mosaicplot(t(tab))
library("gplots")
ttab=as.table(as.matrix(tab))
balloonplot(t(ttab),xlab = "", ylab = "",
label=FALSE,show.margins=FALSE,main='Tableau de contingence')
```



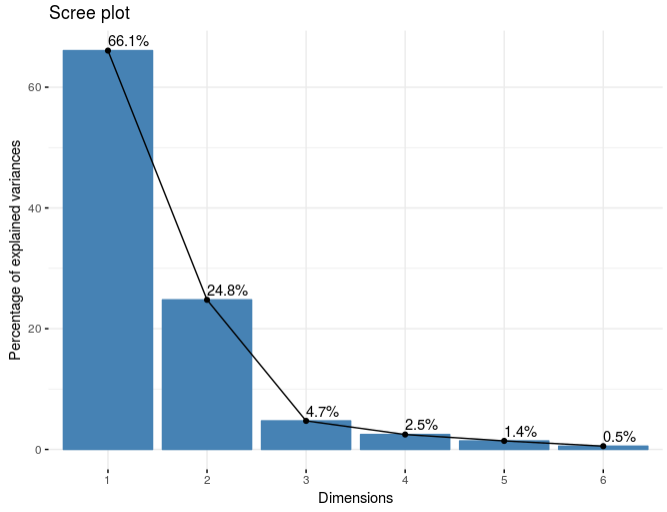

```
prop.table(as.matrix(tab),1) # ligne
prop.table(as.matrix(tab),2) # colonne
barplot(prop.table(as.matrix(tab),1),beside=TRUE)
barplot(t(prop.table(as.matrix(tab),2)),beside=TRUE)
```

3.4.2 AFC

```
afc=CA(tab)
```

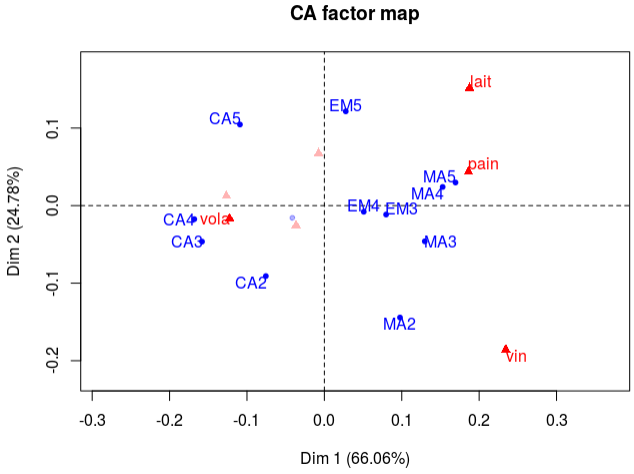
Qualité globale - Choix des axes

```
get_eigenvalue(afc)
fviz_eig(afc,addlabels = TRUE)
```

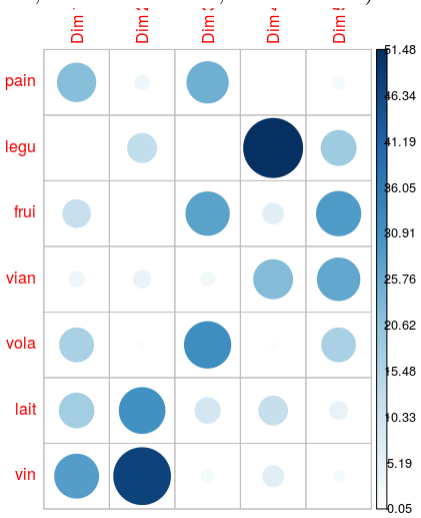


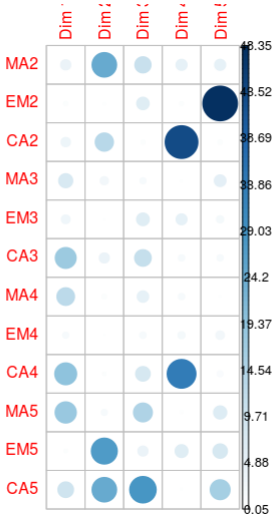
Signification des axes - cos2 + contrib

```
afc$row$cos2
get_ca_col(afc)
fviz_ca_row (afc, col.row = "cos2",repel = TRUE)
plot.CA(afc,selectCol="cos2 0.8",selectRow="cos2 0.8")
fviz_cos2(afc, choice = "row", axes = c(1,2))
corrplot(afc$row$cos2, is.corr = FALSE)
```




```
afccolcontrib
plot.CA(afc,selectCol="contrib 4", selectRow="contrib 4")
dimdesc(afc)
library("corrplot")
corrplot(afc$row$contrib, is.corr = FALSE)
corrplot(afc$col$contrib, is.corr = FALSE)
fviz_contrib (afc, choice = "col", axes = 1 :2)
```

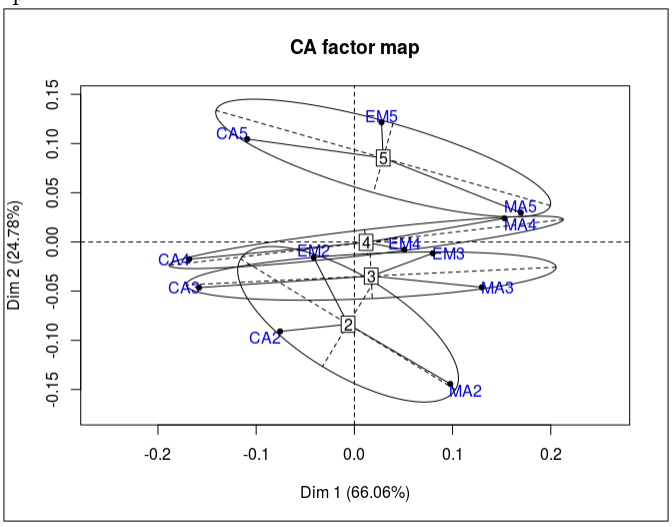


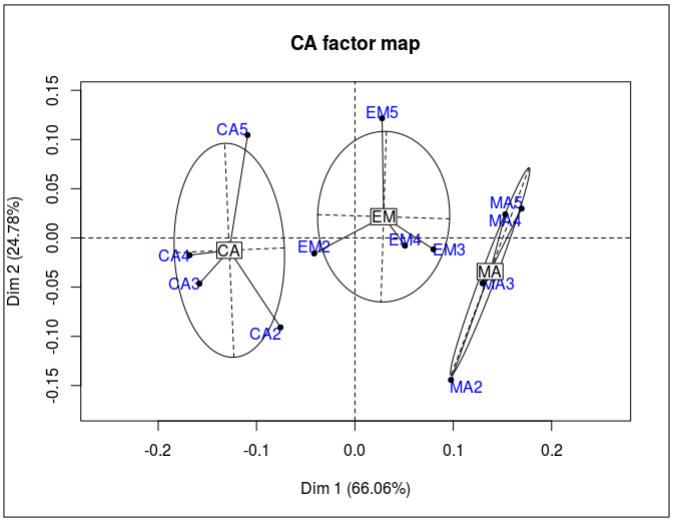


Variables supplémentaires - Interprétation

```
nb=factor(rep(2 :5,each=3))
sp=factor(rep(c('MA','EM','CA'),4))
afc3=CA(cbind(tab,nb,sp),quali.sup=8 :9)
library(ade4)#pour ellipses
plot.CA(afc3,invisible = c("col","quali.sup"))
s.class(dfxy=afc3$row$coord,fac=nb,add.plot=TRUE)
plot.CA(afc3,invisible = c("col","quali.sup"))
s.class(dfxy=afc3$row$coord,fac=sp,add.plot=TRUE)
afc3$quali.sup$v.test[,1 :2]
Dim 1 Dim 2
nb.2 -1.8719962 -10.7122669
nb.3 -0.9110741 -4.9430549
nb.4 -0.2144218 -0.2949769
nb.5 2.7015840 14.3257066
sp.CA -25.7854530 -0.9410175
sp.EM 5.1831306 4.9846737
sp.MA 22.4629624 -4.0523629

afc3$quali.sup$eta2[,1 :2]
Dim 1 Dim 2
nb 0.01001113 0.79028981
sp 0.91029874 0.09162391
```





Chapitre 4

Réduction de dimension 4 : Analyse des correspondances multiples

Tableaux étudiés : L'AFCM est utilisée pour l'étude de volumineux tableaux constitués de plusieurs variables qualitatives, principalement sous forme de questionnaires.
Le tableau se présente sous forme codé, disjonctif codé, ou de tableau de Burt. Il est mathématiquement traité sous forme d'un tableau disjonctif complet analysé comme un unique tableau de contingence.

Les références suivantes peuvent compléter ce chapitre :

- http ://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/75-acm-analyse-des-correspondances-multiples-avec-r-l-essentiel/
- Principal component analysis (article) (Abdi and Williams 2010). https ://goo.gl/1Vtwq1.
- https ://eric.univ-lyon2.fr/ ricco/cours/slides/ACM.pdf

Notations :

- n individus notés i, q variables X^k, ayant chacune p_k modalités avec $p = p_1 + \cdots + p_s$ modalités au total.
- $Z = (Z_1, \ldots, Z_s)$ représente le tableau disjonctif complet. $z_{ij} = 1$ si l'individu i a la modalité j, 0 sinon. On a $\sum_i z_{ij} = z_{.j}$, $\sum_j z_{ij} = q$, $\sum_{i,j} z_{ij} = nq$.
- $z_{.j}$ définit l'effectif marginal de la modalité j variant de 1 à p.
- On note D =diag$(z_{.j})$.
- On appelle tableau de Burt le tableau $Z^T Z$. Il correspond aux tableaux de contingence de tous les couples de variables.

Exemple : Prenons le tableau suivant comprenant $n = 5$ individus notés i, i de 1 à n, $q = 3$ variables X^k, k de 1 à q, ayant $p_1 = 3$, $p_2 = 2$ et $p_3 = 2$ modalités m_l^k soit $p = p_1 + p_2 + p_3$ modalités au total.
A partir du tableau codé, construire le tableau disjonctif noté Z et de Burt $Z^T Z$.

$$\begin{array}{ccc} X^1 & X^k & X^s \\ \left(\begin{array}{ccc} 1 & 2 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 2 \\ 3 & 2 & 2 \\ 2 & 1 & 1 \end{array}\right) \end{array} \quad \left(\begin{array}{ccc|cc|cc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}\right) \quad \left(\begin{array}{ccc|cc|cc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 2 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 1 & 1 & 0 & 2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 3 & 1 & 2 \\ \hline 1 & 1 & 0 & 1 & 1 & 2 & 0 \\ 1 & 1 & 1 & 1 & 2 & 0 & 3 \end{array}\right)$$

Objectifs :

- Comme l'ACP, on peut envisager d'étudier la typologie des individus. Les individus sont considérés comme proches si ils ont de nombreuses modalités communes. Le plus souvent le nombre d'individus est très grand et anonyme. On se limite alors à l'étude de certaines classes (la ménagère de moins de 50 ans, les teen-agers...)
- Comme l'AFC, on peut étudier la liaison entre les modalités :
  - Pour deux modalités d'une même variable (exclusives), elles se ressemblent si les modalités pour les autres variables sont similaires.
  - Pour des modalités de deux variables, leur proximité indique une plus forte association que sous l'hypothèse d'indépendance et des modalités communes.
- L'ACM permet aussi de construire des variables quantitatives (composantes principales F_L à partir de variables qualitatives exploitables dans d'autres méthodes comme la méthode DISQUAL en analyse discriminante.

4.1 Nuage des profils et ajustement

L'idée de départ est d'utiliser une approche analogue à l'AFC pour l'étude des liaisons entre modalités. L'utilisation des programmes de l'AFC sur les tableaux disjonctifs complets (ou les tableaux de Burt) permet une étude de ces tableaux avec des caractéristiques spécifiques ici :

- le tableau de contingence N de l'AFC est remplacé ici par le tableau disjonctif Z, le tableau ne contient donc que des 0 et 1,
- l'effectif total est ns, la somme d'une ligne est toujours s,
- la matrice des fréquence relative F de l'AFC devient $\mathrm{F} = \frac{1}{np} Z$ en ACM,
- les colonnes peuvent être regroupées en variables,
- le profil marginal ligne est $\frac{s}{np}\mathbf{1_n}$, le profil marginal colonne est $(\frac{z_{.j}}{nq})$,
- par analogie avec l'AFC, on en déduit :

$$D = (z_{.j}) \quad D_I = \frac{1}{n} I_n \quad D_J = (\frac{z_{.j}}{nq}) = \frac{1}{nq} D \quad F = \frac{1}{nq} Z.$$

Définition 20 *On appelle ACM d'un tableau l'AFC du tableau disjonctif complet.*

Remarque : On obtient une représentation des modalités identiques en réalisant l'AFC du tableau de Burt.

Proposition 29 *On en déduit par analogie :*

- $L = D_I^{-1}\mathrm{F} = \frac{1}{q} Z$ *de poids* $D_I = \frac{1}{n} I_n$ *et de métrique :*

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

  Deux individus sont donc proches si ils ont de nombreuses modalités communes ($\left(\frac{z_{ij}}{q} - \frac{z_{i'j}}{q} \right) = 0$). *Ils seront d'autant plus éloignés qu'ils présentent des modalités rares différentes* ($\frac{nq}{z_{.j}}$ *élevé).*

- $C = D^{-1} Z^t$ *de poids* $D_J = \frac{1}{nq} D$ *et de métrique*

$$d^2_{\chi^2}(j, j') = \sum_{i=1}^{n} n \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2$$

  Deux modalités sont proches si les individus les possédant ont des modalités communes.

- *On réalise ainsi les dvs des triplets* $(L = \frac{1}{q}Z, nqD^{-1}, \frac{1}{n}I_n)$ *et* $(C = D^{-1}Z^t, nI_n, \frac{1}{nq}D^{-1})$
- *Les matrices d'inertie sont :*

$$Z^tZD^{-1} \text{ pour } L \quad ZD^{-1}Z^t \text{ pour } C$$

- *Les formules de transition sont :*

$$F_L = \frac{1}{q}ZF^C\Lambda^{-1} \quad F^C = D^{-1}Z^tF_L\Lambda^{-1}$$

$$F_L^s(i) = \frac{1}{\sqrt{\lambda_s}}\sum_{j=1}^{q}\frac{z_{ij}}{s}F_s^C(j) = \frac{1}{s\sqrt{\lambda_s}}\sum_{j\in I(i)}F_s^C(j)$$

avec $I(i)$ *l'ensemble des modalités* j *de l'individu* e_i

$$F_s^C(j) = \frac{1}{\sqrt{\lambda_s}}\sum_{i=1}^{n}\frac{z_{ij}}{z_{.j}}F_s^C(j) = \frac{1}{z_{.j}\sqrt{\lambda_s}}\sum_{i\in I(j)}F_L^s(i)$$

avec $I(j)$ *l'ensemble des individus* i *qui possèdent la modalité* j*.*

Preuve : Bon exercice en reprenant les résultats de RD3 sur l'AFC avec F = $\frac{1}{nq}Z$.

4.2 Paramètres d'interprétation de l'ACM

4.2.1 Inertie

Proposition 30 *L'inertie totale se décompose par modalité et par variable :*

- *L'inertie d'une modalité est :* $I(j) = \frac{1}{q}(1 - \frac{z_{.j}}{n})$
- *L'inertie d'une variable est :* $I(X^k) = \frac{p_k}{q} - \frac{1}{q}$
- *L'inertie totale est :* $I_T = \frac{p}{q} - 1$

  avec q *le nombre de variables,* p_k *le nombre de modalités de la variable* k*,* p *le nombre total de modalités,* $z_{.j}$ *l'effectif des individus possédant la modalité* j *et* n *le nombre d'individus.*

Remarque :

- Le poids d'une modalité est d'autant plus grand qu'elle est peu représentée ($z_{.j}$ faible). On regroupera éventuellement certaines modalités peu représentées.
- Le poids d'une variable est d'autant plus grand qu'elle a beaucoup de modalité (p_k grand). On limitera éventuellement le nombre de modalités.
- L'inertie totale est une grandeur abstraite, sans interprétation statistique. Elle est lié à la structure du tableau et une partie de cette inertie ne représente pas une information statistique. L'interprétation de cette grandeur et de celles en découlant (λ_s, cos2...) est donc délicate et négative.

Preuve : En remarquant que $F_J = (\frac{1}{n})$, le poids d'un individu $\frac{q}{nq}$ et d'une modalité $\frac{z_{.j}}{nq}$ de poser le calcul, la distance entre j et j' est :

$$d^2_{\chi^2}(j,j') = \sum_i \frac{nq}{q}(\frac{z_{ij'}}{z.j'} - \frac{z_{ij}}{z.j})^2 \text{ donc } d^2_{\chi^2}(j,g) = \sum_i n(\frac{z_{ij}}{z.j'} - \frac{1}{n})^2$$

Soit en remarquant que z_{ij} ne prend la valeur 1 que pour $z_{.j}$ individus, l'inertie de j :

$$I(j) = \frac{z_{.j}}{nq}\sum_i n(\frac{z_{ij}}{z.j} - \frac{1}{n})^2 = \frac{z_{.j}}{q}\sum_i(\frac{z_{ij}^2}{z.j^2} - 2\frac{z_{ij}}{nz.j} + \frac{1}{n^2}) = \frac{z_{.j}}{q} - \frac{1}{q}.$$

De même en sommant sur les modalités de la variable k et en sommant sur les variables pour l'inertie totale.

4.2.2 Représentations graphiques

Les relations quasi-barycentriques justifient la représentation simultanée mais le nombre d'individus est souvent très important.

- La proximité de deux individus indique qu'ils ont de nombreuses modalités en commun.
- La proximité entre deux modalités de deux variables différentes indique que ces modalités apparaissent souvent ensemble (association)
- La proximité entre deux modalités d'une même variable indique que les autres variables présentent des modalités similaires pour les individus concernés.

Proposition 31 *Les modalités d'une même variables ont l'origine comme centre de gravité dans les cartes factorielles. En conséquence, l'ACM ne possède au plus que $p-q+1$ valeurs propres non nulles, soit $p-q$ axes pour l'interprétation en supprimant le premier axe trivial.*

Preuve : Bon exercice en calculant le centre de gravité des modalités d'un même variable. On en déduit une conséquence sur la dimension de l'image.

4.2.3 Interprétation des paramètres cos2 et contrib

L'inertie totale comme les cos2 et contribution sont en ACM des indicateurs difficiles à interpréter car ce sont des indicateurs en général pessimistes et qui ne représentent pas une information statistique exclusivement. L'interprétation demande une bonne expérience. Pour la sélection des axes, Benzecri a proposé de recalculer les valeurs propres pour $\lambda > \frac{1}{q}$, q le nombre de variables, par :

$$(\frac{q}{q-1})^2(\lambda - \frac{1}{q})^2$$

se révélant un critère de sélection des axes plus adapté. On réalise alors le diagramme de ce critère.

4.2.4 Corrélation d'une variable avec l'axe factoriel

L'interprétation porte sur les modalités mais aussi sur les variables en définissant pour chacune d'elle un coefficient de corrélation. On doit ainsi définir une corrélation entre une variable qualitative et une variable quantitative.

Définition 21 *Pour chaque variable X^k, on définit le carré d'un coefficient de corrélation avec l'axe s comme le rapport de la variance intermodalité sur la variance totale de F_L^s :*

$$\eta^2(s,k) = \frac{Var(E(F_L^s|j))}{Var(F_L^s)} = \sum_{j\in I(k)} \frac{1}{\lambda_s}\frac{z_{.j}}{n}(F_L^s(B_j))^2 = \sum_{j\in I(k)} \frac{z_{.j}}{n}(F_s^C(j))^2$$

avec B_j le barycentre des individus $I(j)$ possédant la modalité j et $F_L^s(B_j) = \sqrt{\lambda_s}F_s^C(j)$ par les relations de transition. On en déduit que :

$$\eta^2(s,k) = q \sum_{j\in I(k)} \textit{ inertie de } j \textit{ sur } s$$

soit la contribution relative de la variable ($\sum_{j\in I(k)} contrib_s(j)$) à l'axe multipliée par l'inertie de l'axe (λ_s) et par le nombre de variables q.

Remarque : Sous H_0 "absence de lien entre l'axe s et la variable k, $(n-2)\frac{\eta^2}{1-\eta^2}$ suit la loi F à 1 et $n-2$ ddl.

Preuve et interprétation :

L'inertie totale de l'axe est λ_s, l'inertie inter s'obtient en remplaçant chaque individu par la moyenne des affixes des individus de sa modalité.
On remarque que :

$$F_L^s(B_j) = \frac{1}{z_{.j}} \sum_{i \in I(j)} F_L^s(i) = \sqrt{\lambda_s} \frac{1}{\sqrt{\lambda_s} \frac{1}{z} .j} \sum_{i \in I(j)} \sum_{i \in I(j)} F_L^s(i) = \sqrt{\lambda_s} F_s^C(j)$$

en utilisant l'expression quasi-barycentrique de $F_s^C(j)$
De plus, l'inertie de j suivant s est $\frac{z_{.j}}{nq} F_s^C(j)^2$.

4.2.5 test de signification d'une modalité

Il est possible de construire un test sur la signification d'une modalité pour un axe donné.

Proposition 32 *La statistique* $t_s(j) = \sqrt{z_{.j} \frac{n-1}{n-z_{.j}}} F_s^C(j)$ *tend vers une loi normale centrée.*

On a défini $F_s^C(j) = \frac{1}{z_{.j}\sqrt{\lambda_s}} \sum_{i \in I(j)} F_L^s(i)$ égal au facteur près $\frac{1}{\sqrt{\lambda_s}}$ à la moyenne arithmétique des individus possédant j, $\frac{1}{z_{.j}} \sum_{i \in I(j)} PC_L^s(i)$.
On note $X_s(j)$ la variable $\frac{1}{z_{.j}} \sum_{i \in I(j)} F_L^s(i)$. Si la variable est indépendante de s le choix des individu $I(j)$ n'a pas d'influence et $E(X_s(j) = 0$(variable centrée et $\mathrm{Var}(X_s(j) = \frac{n-z_{.j}}{n-1} \frac{\lambda_s}{z_{.j}}$ (tirage avec remise).

Remarque :

- le test est approximatif et purement indicatif. En particulier les comparaisons multiples ne permettent pas d'estimer ici le risque de 1ère espèce.
- le test n'a pas de sens pour les variables actives car elles ont contribué aux calculs. Elles sont néanmoins présentes à titre indicatif.

4.2.6 Variables supplémentaires

quantitative

Il est possible d'utiliser une variable quantitative comme variable active en la transformant en une variable qualitative avec des classes.

On peut aussi l'introduire comme suplémentaire en calculant ses corrélations avec les composantes principales F_L.

qualitative

Par projection après ajustement comme en AFC.

4.3 Exemple d'interprétation

L'étude porte sur le tableau poison de FactoMineR : The data used here refer to a survey carried out on a sample of children of primary school who suffered from food poisoning. They were asked about their symptoms and about what they ate. A data frame with 55 rows and 15 columns.

Les variables sont : "Age" "Time" "Sick" "Sex" "Nausea" "Vomiting" "Abdominals" "Fever" "Diarrhae" "Potato" "Fish" "Mayo" "Courgette" "Cheese" "Icecream" .

Les deux premières sont quantitatives, les deux suivantes seront utilisées en supplémentaires, la premère sexe étant a priori sans rôle et sick correspondant ici à une variable réponse.

```
library(FactoMineR)
data(poison)
?poison
summary(poison)

for (i in 5 :15)
plot(poison[,i], main = colnames(poison)[i],
ylab = "Count", col="steelblue", las = 2)
```

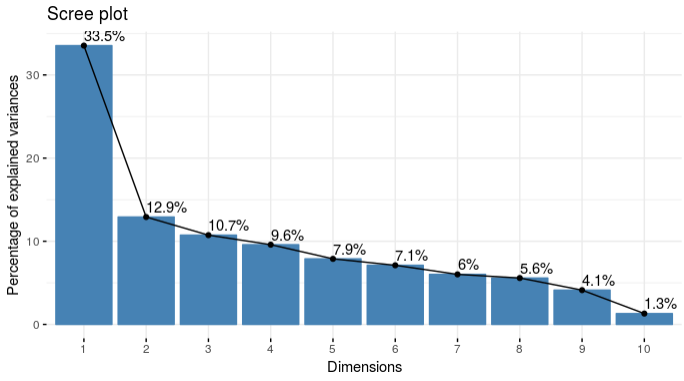
On peut construire le tableau disjonctif complet Z et le tableau de Burt :

```
Z=tab.disjonctif(poison[,c(5 :15)]) #tableau disjonctif complet
t(Z)%*%Z # tableau de Burt
```

	Nausea_n	Nausea_y	Vomit_n	Vomit_y
Nausea_n	43	0	28	15
Nausea_y	0	12	5	7
Vomit_n	28	5	33	0
Vomit_y	15	7	0	22

4.3.1 sélection des axes

```
library(factoextra)
fviz_eig(acm, addlabels = TRUE)
```



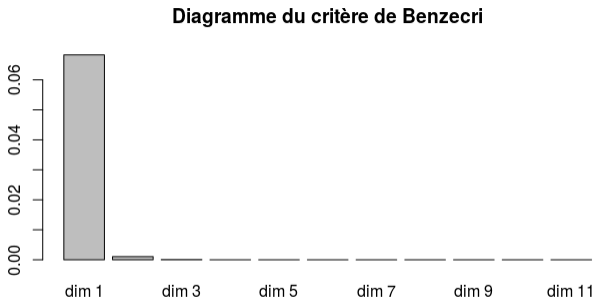
$q = 10$.

Les 2 premiers axes de l'ACM expriment 46.4 % de l'inertie totale du jeu de données. L'interprétation de ce % est à prendre avec précaution, une partie de l'inertie est artificielle liée à la nature du tableau codé. La règle du coude indique cependant que le premier axe semble pertinent. Pour pallier les difficultés d'interprétation, on peut utiliser le critère de Benzécri qui confirme clairement que seul le premier axe est pertinent. En fait, cet axe explique la plus grande partie de l'"inertie" intéressante alors que sa valeur n'est que de 33%.

```
# critère de Benzecri de sélection des axes
E=acm$eig[,1]
E[acm$eig[,1]<1/10]=1/10
E = (10/9)^2 * (E - 1/10)^2
barplot(E)
```

$= \left(\frac{q}{q-1}\right)^2 \left(\text{valeur propre} - \frac{1}{q}\right)^2$

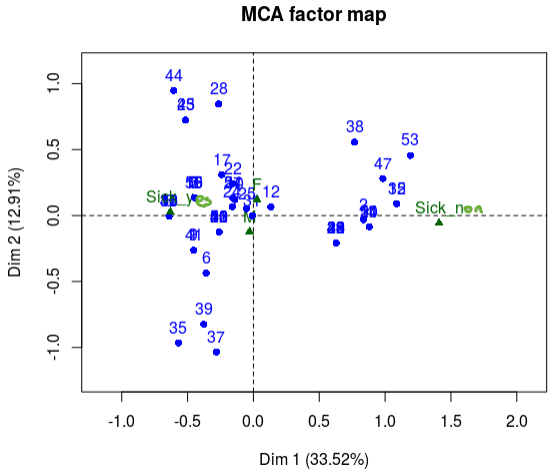
title("Diagramme du critère de Benzecri")

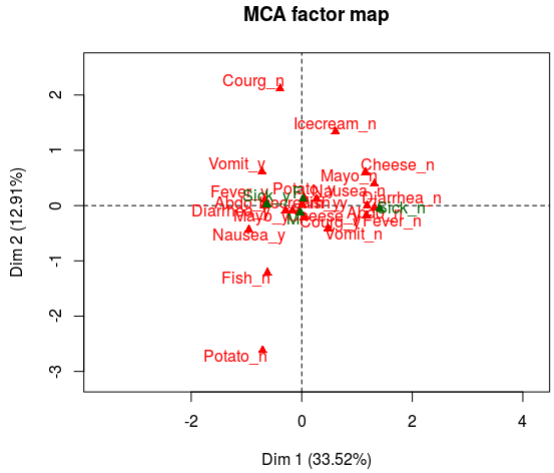


1 seul axe permettrait de tout expliquer
mais par principe on en prend au moins 2 tjs

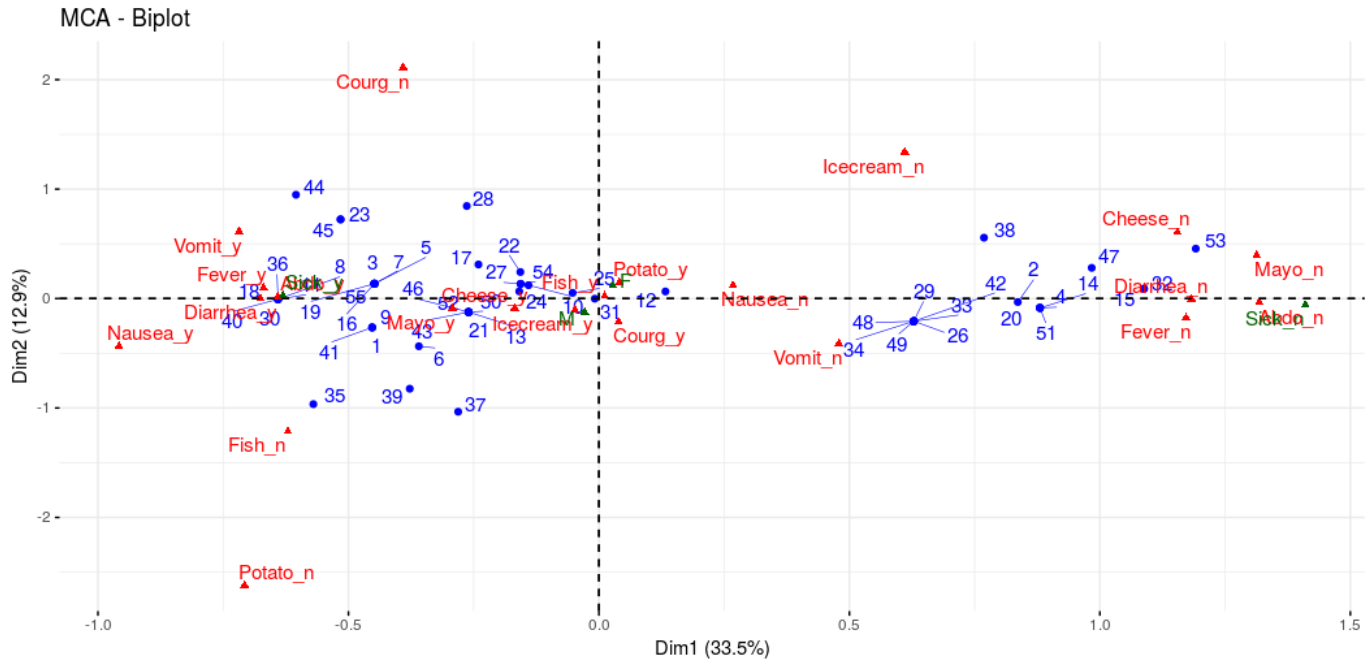
4.3.2 projection des individus, modalités et variables

plot(acm) ;plot(acm, invisible= "var") ;plot(acm, invisible = "ind")

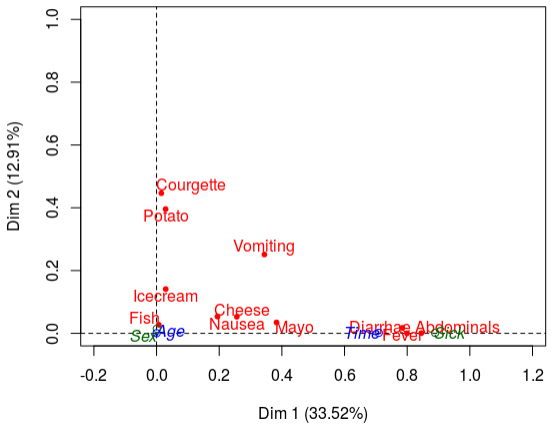




fviz_mca_biplot (acm, repel = TRUE,ggtheme = theme_minimal())



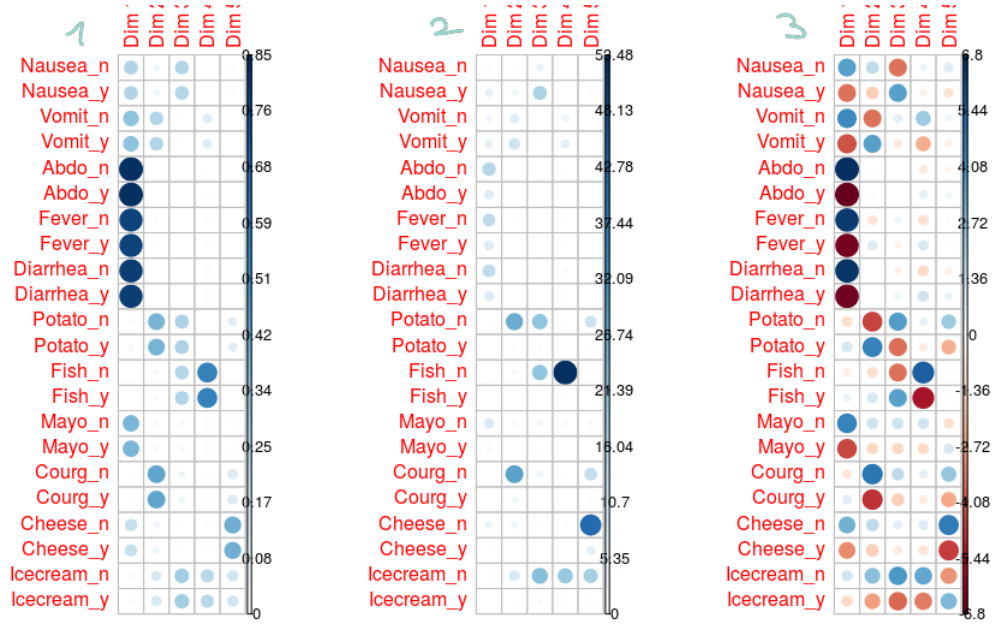
On peut aussi représenter le carré de la corrélation des variables avec les axes :
plot(acm, choix = "var")



4.3.3 cos2 et contrib

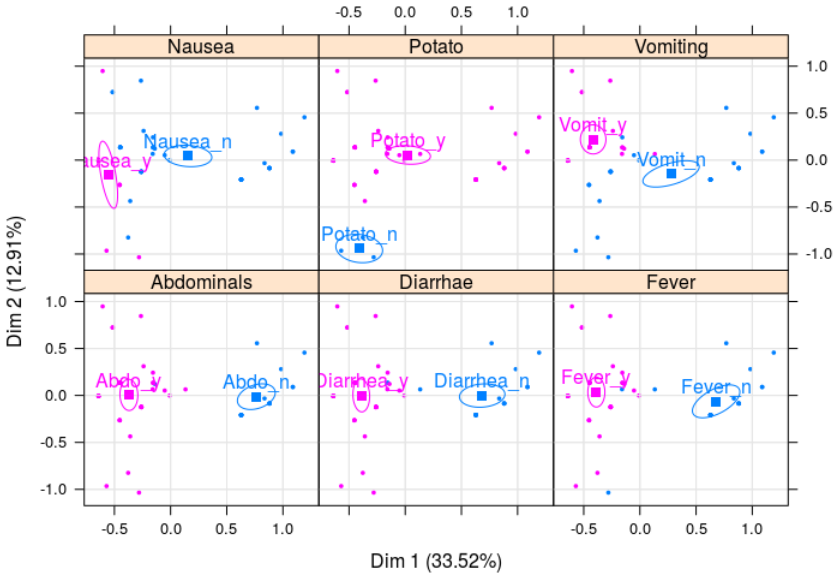
Souvent les individus sont très nombreux et anonymes. L'étude porte principalement sur les modalités et les variables. On utilise les composantes principales des individus dans d'autres méthodes (classification, disqual).

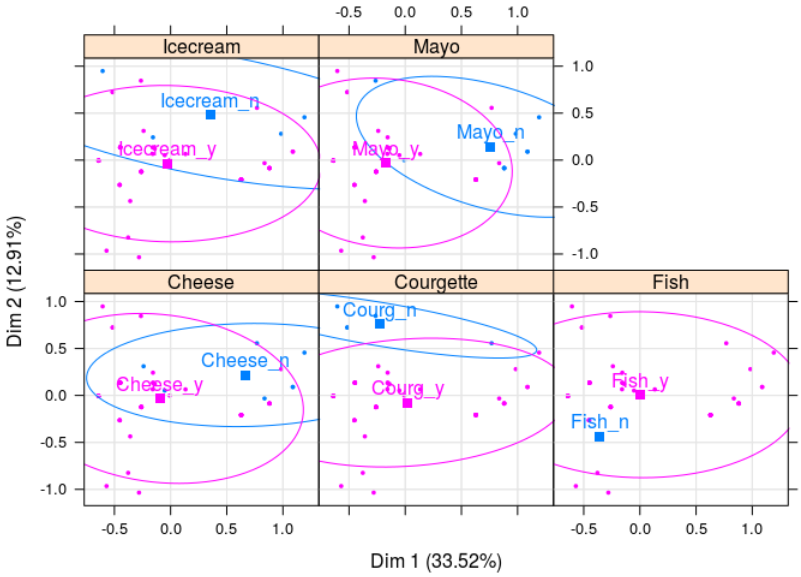
```
library(corrplot)
1 corrplot(acm$var$cos2, is.corr = FALSE)
2 corrplot(acm$var$contrib, is.corr = FALSE)
3 round(acm$var$v.test,1)
corrplot(round(acm$var$v.test,1), is.corr = FALSE)
```



On peut aussi représenter la répartition des individus par modalité d'une variable. L'ellipse de confiance du centre de gravité de chaque modalité est également représentée ou celle des observations (means=FALSE).

```
plotellipses(acm,keepvar=c(5 :10),level = 0.95)
plotellipses(acm,keepvar=c(11 :15),means=FALSE))
```





4.3.4 Interprétation du plan F1 F2

Il ressort des résultats précédents que les variables les plus pertinentes pour F1 sont Fever, Diarrhea, Abdominal principalement à travers leurs modalités _n. Les individus sans ces symptômes d'intoxication alimentaire se retrouve dans la partie positive de F1, les autres dans la partie négative. Cet axe semble indiquer la gravité de l'intoxication, du plus grave au moins grave. Ce résultat est confirmé par les tests réalisés sur les variables et modalités pour lesquels les valeurs obtenus sont fortes. On retrouve pour la variable silk la même répartition selon l'axe F1 confirmant son interprétation en terme d'intensité de l'intoxication.

L'axe F2 est conditionné principalement par les variables courgette et potato, et dans une moindre mesure, vomit, par leur modalités négatives, _n. La signification de cet axe est plus incertaine, les 2 modalités s'opposant selon ces deux axes sans relation de causalité évidente avec d'autres variables. Il faut garder en mémoire que cet axe a une signification très faible dans cette étude.

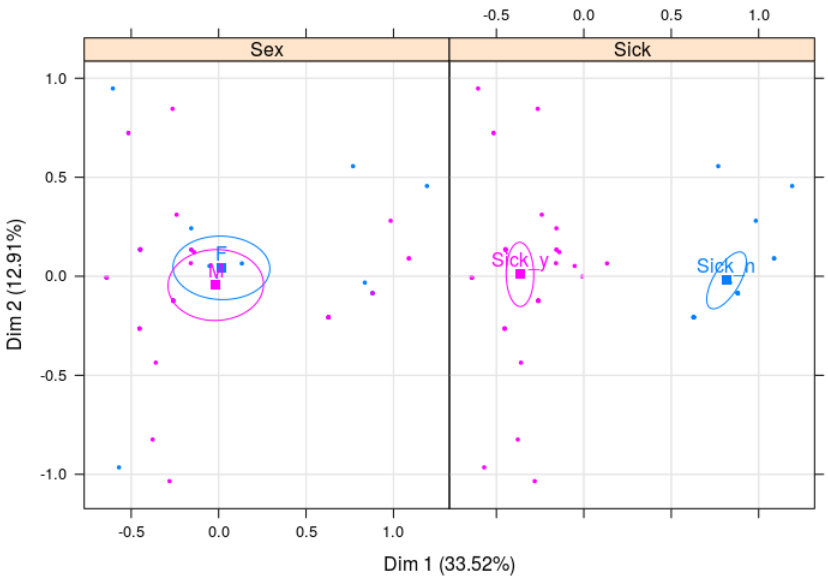
On peut remarquer un groupe d'individus à droite associés aux modalités cheese, mayo, diarrhea, fever, abdo, vomit, icecream nausea tous négatifs associés aux faibles intensités d'intoxication, opposé à un groupe à gauche présentant les modalités positives.

Les aliments cheese, mayo et icecream semblent liés à la sévérité de l'intoxication.

4.3.5 éléments supplémentaires

variable qualitative supplémentaire

plotellipses(acm,keepvar=c(3 :4))



acm$quali.sup$v.test

Dim 1 Dim 2

Sickn 6.9294293 -0.2841346

Sicky -6.9294293 0.2841346

F 0.2047705 0.8860047

M -0.2047705 -0.8860047

acm$quali.sup$eta_2

Dim 1 Dim 2

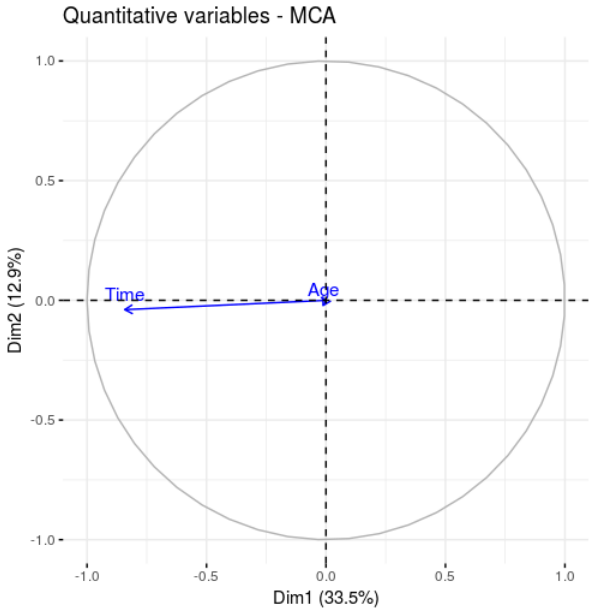
Sick 0.8892035226 0.001495045

Sex 0.0007764989 0.014537117

Seule la variable silk est fortement liée à l'axe 1 (v-test »2, eta2 proche de 1) comme l'on pouvait s'y attendre, le sexe n'intervenant pas dans cette étude.

variable quantitative supplémentaire

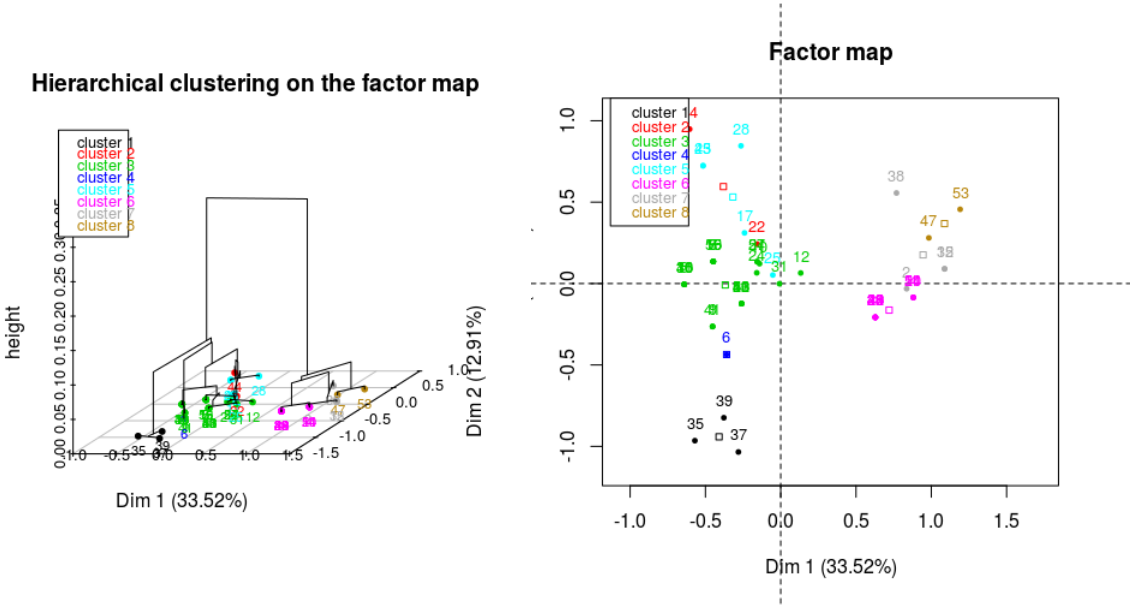
fviz_mca_var(res.mca, choice = "quanti.sup",ggtheme = theme_minimal())



Si time signifie le temps d'indisposition, il est logique de trouver ici une corrélation positive avec la gravité de l'intoxication.

4.3.6 compléments

C'est tellement plus simple avec :
library(FactoInvestigate)
Investigate(acm)
dimdesc(acm, axes = 1 :1)
HCPC(acm)



La CAH du tableau corrobore les observations précédentes. On observe 8 groupes.

La classe 1 est composée d ?individus tels que 35, 37 et 39 : caractérisé par une forte fréquence de la modalité Potato=Potato_n.

La classe 2 contient l'individu 44, par une forte fréquence de la modalité Icecream=Icecream_n.

La classe 3 est composé d ?individus partageant une forte fréquence des modalités Sick=Sick_y, Abdominals=Abdo_y, Fever=Fever_y, Diarrhae=Diarrhea_y, Vomiting=Vomit_y, Cheese=Cheese_y, Mayo=Mayo_y et Courgette=Courg_y (du plus commun au plus rare) et une faible fréquence des modalités Sick=Sick_n, Abdominals=Abdo_n, Fever=Fever_n, Diarrhae=Diarrhea_n, Vomiting=Vomit_n, Cheese=Cheese_n, Mayo=Mayo_n

et Courgette=Courg_n (du plus rare au plus commun).

La classe 4 contient l'individu 6 caractérisé par une forte fréquence de la modalité Fish=Fish_n.

Ce groupe est caractérisé par23, 28 et 45, avec une forte fréquence de la modalité Courgette=Courg_nCourgette=Courg_y.

La classe 6 est caractérisée par4, 14, 20, 26, 29 et 51, avec une forte fréquence des modalités Sick=Sick_n, Abdominals=Abdo_n, Diarrhae=Diarrhea_n, Fever=Fever_n, Vomiting=Vomit_n et Nausea=Nausea_n (du plus commun au plus rare) et une faible fréquence des modalités Sick=Sick_y, Abdominals=Abdo_y, Fever=Fever_y, Diarrhae=Diarrhea_y, Vomiting=Vomit_y et Nausea=Nausea_y (du plus rare au plus commun).

La classe 7 est caractérisé par2, 15, 32 et 38, avec une forte fréquence des modalités Cheese=Cheese_n, Sick=Sick_n, Abdominals=Abdo_n, Diarrhae=Diarrhea_n et Fever=Fever_n (du plus commun au plus rare) et une faible fréquence des modalités Cheese=Cheese_y, Sick=Sick_y, Abdominals=Abdo_y, Fever=Fever_y et Diarrhae=Diarrhea_y (du plus rare au plus commun).

La classe 8 est caractérisée par 47 et 53. avec une forte fréquence des modalités Icecream=Icecream_n et Mayo=Mayo_n (du plus commun au plus rare) et une faible fréquence des modalités Icecream=Icecream_y et Mayo=Mayo_y (du plus rare au plus commun).

On pourrait aussi compléter en faisant une analyse DISQUAL pour expliquer silk.

Chapitre 5

Réduction de dimension 5 : Analyse factorielle discriminante

5.1 Introduction

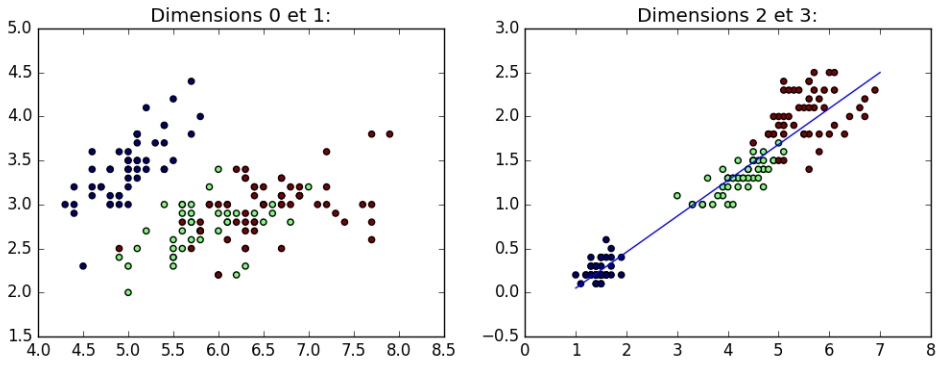
On considère ici q classes d'individus sur lesquels p variables quantitatives sont mesurées. On note T la matrice des indicatrices et $X = (X^1, \ldots, X^p)$ le tableau des variables quantitatives.

L'objectif de l'analyse factorielle discriminante (AFD) est double :

- **descriptif**, en cherchant les combinaisons linéaires des p variables permettant de séparer au mieux les individus et en donner une représentation graphique satisfaisant cet objectif,
- **décisionnel**, en permettant de classer de nouveaux individus dans les classes préexistantes connaissant les p variables.

Pour illustrer le cours, nous utiliserons l'exemple iris caractérisant $q = 3$ variétés d'iris par $p = 4$ variables quantitatives, longueur et klargeur des sépales et des pétales.

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn import datasets
iris = datasets.load_iris()
Xiris = iris.data
yiris = iris.target
plt.figure(figsize=(12,4))
plt.subplot(1, 2, 1)
plt.title("Longueur et largeur des sépales")
plt.scatter(Xiris[ :, 0], Xiris[ :, 1], c=yiris)
plt.subplot(1, 2, 2)
plt.title("Longueur et largeur des pétales")
plt.scatter(Xiris[ :, 2], Xiris[ :, 3], c=yiris)
plt.plot([1,7],[0.05,2.5])
plt.show()
```



L'AFD est une méthode très utilisée et diversifiée. On la rencontre en contrôle qualité, en diagnostic, en prévision des risques. L'approche décisionnelle conduit à la construction de scores permettant la construction de règles de classement. Par exemple, on peut ainsi prévoir le risque d'avalanche à partir de mesures météo, d'exposition, de pentes, faire un diagnostic médical à partir de mesures accessibles.

http ://cedric.cnam.fr/vertigo/Cours/ml/tpAfd.html

https ://archive.ics.uci.edu/ml/index.php

5.2 Détermination des fonctions linéaires discriminantes

Principe : Le principe général est de construire une première variable dite discriminante comme combinaison linéaire des variables initiales. Cette variable doit minimiser la variance intra-classe et maximiser la variance inter-classes (critère d'ajustement). La seconde variable discriminante est construite comme non corrélée à la première et vérifiant le même critère, et ainsi de suite.

5.2.1 Notation

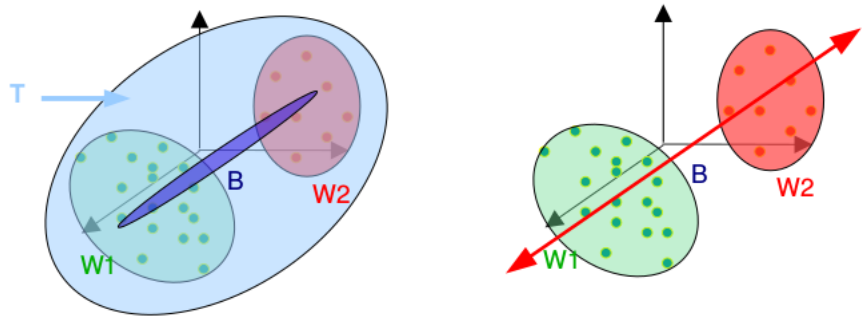
On considère un n échantillon constitué de n individus appartenant à des classes données, 1 à q, sur lesquels sont mesurés p variables quantitatives.

Les n individus sont représentés par des points $X_i^T = (x_i^1, \ldots, x_i^p)$ dans l'espace $\mathbb{R}^p$.

Pour simplifier les calculs tout en préservant la généralité de la démonstration, nous considérerons que tous les individus ont le même poids $\omega_i = \frac{1}{n}$ et que le tableau a été préalablement **centré**.

On note :

- ω_i le poids d'un individu i, D la matrice diagonale des poids,
- $I(k)$ représente l'ensemble des indices i des n_k individus de la classe k,
- n_k le nombre d'individus dans la classe k et $\omega_k = \sum_{i \in I(k)} \omega_i = \frac{n_k}{n}$ le poids de la classe k,
- $g = \sum_{i=1}^{n} \omega_i X_i$ le centre de gravité du nuage (l'origine ici car centré),
- $g_k = \sum_{i \in I(k)} \frac{\omega_i}{\omega_k} X_i$ le centre de gravité du sous-nuage $\mathcal{N}_k$ constitué des individus de la classe k,
- $W_k = \sum_{i \in I(k)} \frac{\omega_i}{\omega_k} (X_i - g_k)(X_i - g_k)^T)$ la matrice de variance-covariance du nuage $\mathcal{N}_k$,
- $B = \sum_{k=1}^{q} \omega_k (g_k - g)(g_k - g)^T$ la matrice de variance-covariance inter-classes (between),
- $W = \sum_{k=1}^{q} \omega_k W_k$ la matrice de variance-covariance intra-classes (within),
- $T = \sum_{i=1}^{n} \omega_i (X_i - g)(X_i - g)^T$ la matrice de variance-covariance totale.



5.2.2 Propriétés du nuage de points

Proposition 33 *Théorème de Huygens*

La matrice de variance covariance totale, T, est égale à la somme de la matrice de variance covariance intra groupe, W, et de la matrice de variance covariance inter groupe, B On obtient ainsi la relation : $T = B + W$.

Preuve Théorème de Huygens (RD1)

Proposition 34 *Cas d'une population multinormale*

Dans le cas d'une population multinormale, chaque individu suit une loi multinormale $\mathcal{N}(\mu_k, \Sigma_k)$.
Sous cette hypothèse, les estimateurs non biaisés utilisés sont :

- $\hat{\mu}_k = g_k$,
- $\hat{\Sigma}_k = \dfrac{n_k}{n_k - 1} W_k$,
- $\hat{\Sigma} = \dfrac{n}{n-q} W_k$,
- $\hat{T} = \dfrac{n}{n-1} T$
- $\hat{B} = \dfrac{n}{q} T$

Remarque : On n'a plus l'égalité $\hat{T} = \hat{B} + \hat{B}$.

Les hypothèses de loi multinormale et d'homoscédasticité sont à la base des principaux tests statistiques utilisés.

Définition 22 *La métrique* $M = \Sigma^{-1}$ *engendre la distance de Mahalanobis. On note* $\Delta_k^2 = (X_i - g_k)^T \Sigma^{-1} (X_i - g_k)$ *la distance de Mahalanobis entre un point et le centre de gravité. Cette distance tient compte de la forme du nuage de points dans le calcul de distance.*

Sous l'hypothèse gaussienne, Δ_k^2 *suit un* χ^2 *à* p *ddl. On en déduit ainsi la forme de l'ellipsoïde de confiance.*

$$(X_i - g_k)^T \Sigma^{-1} (X_i - g_k) = \chi^2_{1-\alpha}(p).$$

Cette distance est estimé par $D_p^2 = (X_i - g_k)^T \hat{\Sigma}^{-1} (X_i - g_k)$.

Exemple :

On étudie une population constituée de 2 classes (T=1 ou 2) de 5 individus chacune caractérisés par deux variables quantitatives X^1 et X^2.

X^1	0	2	4	6	8	5	7	9	11	13
X^2	3	1	5	9	7	2	0	4	8	6
T	1	1	1	1	1	2	2	2	2	2

Déterminer $g, g_1, g_2, T, B, W_1, W_2, W, \hat{W}$

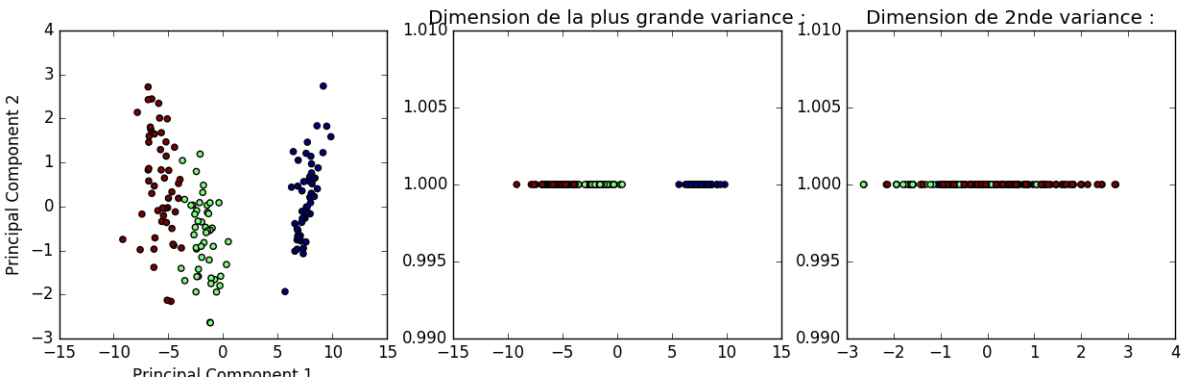
```
X=np.array([[0,3],[2,1],[4,5],[6,9],[8,7],[5,2],[7,0],[9,4],[11,8],[13,6]])
C=np.array([[1,0],[1,0],[1,0],[1,0],[1,0],[0,1],[0,1],[0,1],[0,1],[0,1]])
n=len(C) ;n
In=np.ones([n,1]) ;In # vecteur 1n
g=np.transpose(X).dot(In)/n ;g # gravité
X0=X-In.dot(np.transpose(g)) ;X0 # centrée
Dq=np.transpose(C).dot(C)/n ;Dq # poids des groupes
nq=np.transpose(C).dot(C) ;nq # effectif groupes
Xq=np.linalg.inv(nq).dot(np.transpose(C).dot(X0)) ;Xq# centre des groupes
B=np.transpose(Xq).dot(Dq.dot(Xq)) ;B
T=np.transpose(X0).dot(X0)/10 ;T
W=T-B ;W
Xc=C.dot(Xq) ;Xc #matrice des espérances
XX=X0-Xc ;XX # Ecart à gk pour calculer Wk
X1c=XX[0 :nq[1,1], :] ;X1c
W1=np.transpose(X1c).dot(X1c)/nq[0,0] ;W1
X2c=XX[nq[1,1] :(n+1), :] ;X2c
W2=np.transpose(X2c).dot(X2c)/nq[1,1] ;W2
```

5.2.3 Critère d'ajustement

Définition 23 *Une variable discriminante, F_D, est une combinaison linéaire des variables initiales, soit Xa, cette variable avec a un vecteur colonne. a est la fonction linéaire discriminante correspondante (forme linéaire).*

$$F_D = Xa, \quad F_D(i) = \sum_{j=1}^{p} a_j X_i^j.$$

```
lda = LinearDiscriminantAnalysis(n_components=2)
XirisLDA = lda.fit(Xiris, yiris).transform(Xiris)
def graph.acp2(XPC2, y) :
plt.figure(figsize=(15,4))
plt.subplot(1, 3, 1)
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.scatter(XPC2[ :, 0], XPC2[ :, 1], c=2*y)
plt.subplot(1, 3, 2)
plt.title("Dimension de la plus grande variance :")
plt.scatter(XPC2[ :, 0], np.ones(XPC2.shape[0]), c=y)
plt.subplot(1, 3, 3)
plt.title("Dimension de 2nde variance :")
plt.scatter(XPC2[ :, 1], np.ones(XPC2.shape[0]), c=y)
plt.show()
graph.acp2(XirisLDA, Yiris)
```



Proposition 35 *Les variables discriminantes sont centrés et de somme des carrés :*

$$F_D^T D F_D = \sum_{i=1} n\omega_i F_D(i)^2 = a^T T a$$

La somme des carrés totale $a^T T a$ se décompose alors en une somme inter $a^T B a$ et une somme intra $a^T W a$:

$$a^T T a = a^T B a + a^T W a.$$

L'objectif de l'analyse discriminante est de définir de nouvelles variables à partir de combinaisons linéaires des variables initiales et :

- rendre maximale la variance inter-classe $a^T B a$ et minimales l'inertie intra-classe $a^T W a$,
- ou de façon équivalente, rendre maximale la variance interclasse $a^T B a$ par rapport à l'inertie totale $a^T T a$.

Ces deux critères définissent des variables discriminantes équivalentes.

Proposition 36 *Critère d'ajustement*

Les fonctions discriminantes a sont les fonctions qui permettent de maximiser le quotient $\frac{a^T B a}{a^T W a}$ équivalent à maximiser $\frac{a^T B a}{a^T T a}$.

5.2.4 Fonctions linéaire discriminantes

Proposition 37 *Le quotient* $\frac{a^T Ba}{a^T Wa}$ *est invariant si* a *est changé en* λa, $\lambda > 0$. *Maximiser* $\frac{a^T Ba}{a^T Wa}$ *revient à maximiser* $a^T Ba$ *sous la contrainte* $a^T Wa = 1$ *par exemple. Dans la pratique, on utilisera la normalisation de lda (package MASS, R) :*

$$a^T \hat{W} a = 1 \quad \textit{équivalent à } a^T W a = \frac{n-q}{n}.$$

Remarque : Dans la littérature, les fonctions discriminantes peuvent être définies différemment suivant :

- le critère utilisé $\frac{a^T Ba}{a^T Wa}$ ou $\frac{a^T Ba}{a^T Ta}$,
- la normalisation de a.

Proposition 38 *Ajustement*

L'ajustement revient à rechercher les valeurs propres non nulles, λ_s *de* $W^{-1}B$ *et ses vecteurs propres* $\hat{W}$ *normés* a_s.

Preuve

Retour à l'exemple : Trouver a et F_D.

$W^{-1}B = \begin{bmatrix} 2.52 & -0.50 \\ -2.17 & 0.43 \end{bmatrix}$

```
WiB=np.linalg.inv(W).dot(B) ;WiB # matrice W**-1B
vap=np.linalg.eig(WiB)[0] ;vap # une seule vap non nulle
vep=np.linalg.eig(WiB)[1] ;vep # a non normé
np.diag(np.transpose(vep).dot(Wc.dot(vep)))**(-0.5)
a=vep.dot(np.diag(np.diag(np.transpose(vep).dot(Wc.dot(vep)))**(-0.5)))
np.transpose(a).dot(Wic.dot(a))
a=a[ :,0] # on ne retient qu'une variable discriminante
y=np.array([1,1,1,1,1,2,2,2,2,2])
# avec scikrit learn
lda = LinearDiscriminantAnalysis(n_components=2)
XLDA=lda.fit(X0,y).transform(X0)
lda.scalings_
a
```

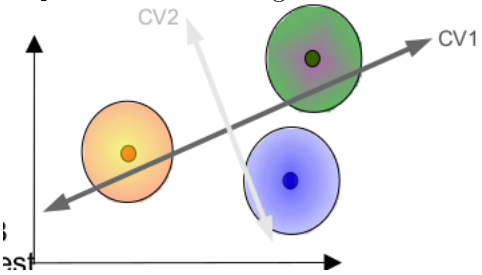
Déterminer les fonctions et variables discriminantes dans l'exemple.

5.2.5 Nombre de fonctions discriminantes

Proposition 39 *$W^{-1}B$ est une matrice de rang r=min$(p, q-1)$.*

Preuve

Dans le graphique suivant, on dispose de 3 classes soit 2 axes discriminants possibles dans l'espace engendré par les 3 centres de gravité.



Interprétation des valeurs propres : Soit $\lambda_1 \geq \cdots \geq \lambda_r > 0$ les r valeurs propres non nulles :

- si λ est très grand, l'inertie inter est très grandes par rapport à l'inertie intra donc les q nuages sont dans des plans M-orthogonaux
- si $\lambda \approx 0$, les q nuages se projettent en un même point (concentrique) et l'axe n'a aucun intérêt pour la discrimination (axe orthogonal au plan formé par les q centres g_k).

5.2.6 Equivalence des critères utilisés

Proposition 40 *Equivalence entre les critères d'ajustement*

En notant λ_s et a_s, les valeurs propres et fonctions discriminantes obtenues avec $\frac{a^TBa}{a^TWa}$ et μ_s et u_s, les valeurs propres et fonctions discriminantes obtenues avec $\frac{a^TBa}{a^TTa}$. On a alors les équivalences :

- $\lambda_s = \frac{\mu_s}{1-\mu_s}$ *et* $\mu_s = \frac{\lambda_s}{1+\lambda_s}$,
- a_s *et* u_s *sont égaux à un facteur près.*

Remarque : Dans les calculs, nous choisirons le critère et la normalisation $t\hat{W}a = 1$ pour normaliser a et ainsi être en adéquation avec la fonction lda de R. La fonction de ade4, discrimin, utilise au contraire le critère $\frac{a^TBa}{a^TTa}$ et la normalisation $u^T\hat{T}u = 1$. Les interprétations changent mais le résultat final est le même.

5.3 AFD et ACP

Soit X le tableau initial décrivant n individus en fonction de p variables quantitatives.

On note C la matrice des indicatrices de l'appartenance aux q groupes. $C = (c_{ik})$, $c_{ik} = 1$ si l'individu i appartient au groupe k, 0 sinon. On note D la matrice diagonale des poids des individus.

Proposition 41 *On a :*

- $g^T = 1_n^T DX$, *on centre alors* $X = X - 1_n G^T$, X **est centrée à partir de là**,

- *le poids de chaque classe est* $D_q = C^T DC$,
- *la matrice* X_q *des barycentres des* q *classes est* $X_q = C^T DX$,
- *on note* $\hat{X}$ *la matrice qui associe à chaque individu le centre de gravité de la classe correspondante, c'est l'espérance conditionnelle de* X_i *sachant sa classe :*
  - $\hat{X} = CX_q = HX$ *avec* $H = C(C^T DC)^{-1}C^T D$ *l'opérateur de projection des variables sur les indicatrices de classes,*
  - *on en déduit que* $X = HX + (I_n - H)X$ *et :*

    $T = X^T DX = X^T(H^T + (I_n - H)^T)D(H + I_n - H)X = \hat{X}^T D\hat{X} + (X - \hat{X})^T D(X - \hat{X}) = B + W = X_q^T D_q X_q$

Proposition 42 *L'AFD de* $(Y|C, D)$ *est l'ACP du triplet* (X_q, W^{-1}, D_q) *équivalent au triplet* (HX, W^{-1}, D).

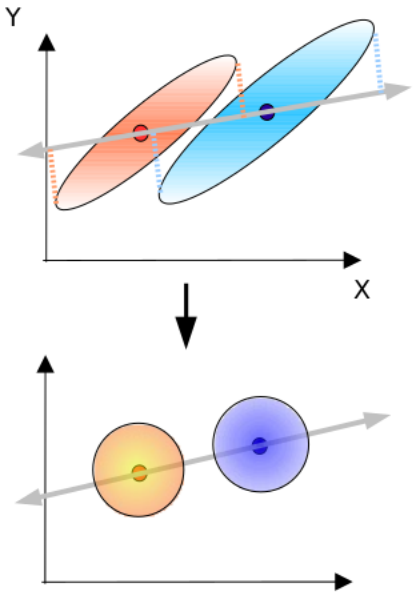
Preuve TD

Remarques :

Réaliser une AFD revient à réaliser l'ACP du nuage des centre de gravité des classes (dimension $q-1$ maximum) avec la métrique W^{-1}.

On réalise ensuite la projection W^{-1} orthogonale des points sur les axes factorielles ainsi définis. On examine alors la discrimination des individus et des centres de gravité dans les différents espaces ainsi définis.

Géométriquement, l'utilisation de la métrique de Mahalanobis revient à normaliser les matrice de variance covariance des groupes et ainsi de passer de nuages ellipsoides à des nuages sphériques permettant de discriminer au mieux les nuages :



5.4 Cas particuliers de deux classes : fonction de Fisher

Dans ce cas il n'existe qu'une valeur propre non nulle ($q - 1 = 1$) et donc une seule fonction discriminante. On obtient alors pour B :

Proposition 43 *On considère ici deux groupes. On a alors :*

- $n_1 g_1 + n_2 g_2 = 0$,
- $B = \frac{n_1 \times n_2}{n^2}(g_2 - g_1)(g_2 - g_1)^T$,
- $W^{-1}B$ *admet une unique valeur propre non nulle* $\lambda = \frac{n_1 \times n_2}{n^2} D_p^2$ *avec*

  $$D_p^2 = (g_2 - g1)^T \hat{W}^{-1}(g_2 - g_1)$$

  l'estimation de la distance de Mahalanobis entre les deux centres de gravité,
- *le facteur discriminant,* a, *appelée fonction de Fisher, est :*

  $$a = \hat{W}^{-1}(g_2 - g_1).$$

Preuve TD

5.5 Inférence dans le cas de populations suivant une loi multinormale

http ://eric.univ-lyon2.fr/∼ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf
https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-modmixt5-manova.pdf

5.5.1 Estimation des matrices de variances

Supposons maintenant que dans chaque groupe, les individus suivent une loi multinormale. Sous l'hypothèse où les lois multinormales ont toute la même matrice de covariance, Σ, un estimateur non biaisé de Σ est donné par $\hat{W}$.

5.5.2 Pseudo F

Pour une fonction discriminante a donnée, permettant le calcul des coordonnées $F_D = Xa$, il est courant de calculer le pseudo F, $F*$, correspondant au test F d'analyse de variance qui teste l'égalité des moyennes :

$$F* = \frac{\frac{a^T Ba}{q-1}}{\frac{a^T Wa}{n-q}}.$$

Il permet d'apprécier la qualité de discrimination de l'axe en comparaison des qualités des variables initiales. F* ne suit pas exactement F.

5.5.3 Egalité des matrices de variances intra groupes

Il est possible de tester l'hypothèse d'égalité des Σ_k, avec le test de Kullback, de Cox, de Bartlett. Cette égalité est nécessaire pour le test de Bartlett et le classement des individus.

5.5.4 Test de Bartlett (différences entre groupes)

L'AFD repose sur l'existence d'une différence des moyennes μ_k entre classes. L'hypothèse H_0 est alors ""$\mu_1 = \cdots = \mu_q$".

Proposition 44 *Le test de Bartlett permet de tester H_0 : Il repose sur la statistique*

$$[n-1-\frac{p+q}{2}]\sum_{s=1}^{r} ln(1+\lambda_s) = -[n-1-\frac{p+q}{2}]\ln\Lambda$$

qui suit asymptotiquement un χ^2 avec $p(q-1)$ ddl.
On appelle lambda de Wilks la statistique $\Lambda = \frac{|W|}{|B|}$ avec $\Lambda^{-1} = \prod_{s=1}^{r}(1+\lambda_s)$.

Un test analogue permet de déterminer parmi les fonctions discriminantes celles significatives.
Pour tester l'apport des fonctions r' à r, on utilise la statistique

$$[n-1-(p+q)/2]\sum_{s=r'}^{r} ln(1+\lambda_s)$$

qui suit asymptotiquement un χ^2 avec $(p-r'+1)(q-r')$ ddl.
Certaines procédures permettent également de choisir les variables initiales permettant la meilleure discrimination et limitant ainsi le nombre de variables initiales nécessaires.

5.5.5 Cas de deux groupes : distance de Mahalanobis

Dans le cas de deux groupes, on a le carré de la distance de Mahalanobis qui est :

$$\delta_p^2 = (\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1) \text{ estimé par} D_p^2\frac{n_1+n_2-2}{n^2}(g_2-g_1)^T\hat{W}^{-1}(g_2-g_1)$$

Proposition 45 *Sous H_0 : "$\mu_1 = \mu_2$,$\frac{n_1 n_2}{n^2}\frac{n-p-1}{p(n-2)}D_p^2$ suit la loi $F(p; n-p-1)$.*

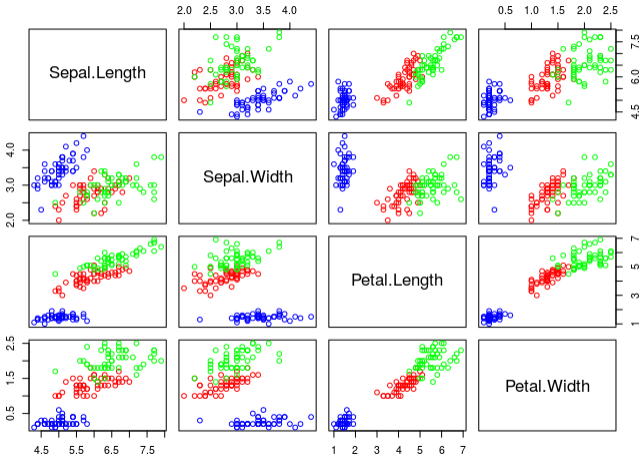
5.6 Interprétation

Reprenons l'exemple iris mais avec R, MASS et ade4 :

library(ade4)

data(iris)

plot(iris[,1 :4],col=c("blue","red","green")[iris$Species])



Recherche de la variable la plus discriminante

for (i in 1 :4) print(anova(lm(iris[,i] iris$Species)))

Response : iris[, 1]

Df Sum Sq Mean Sq F value Pr(>F)

iris$Species 2 63.212 31.606 119.26 < 2.2e-16 ***

Residuals 147 38.956 0.265

Response : iris[, 2]

Df Sum Sq Mean Sq F value Pr(>F)

iris$Species 2 11.345 5.6725 49.16 < 2.2e-16 ***

Residuals 147 16.962 0.1154

Response : iris[, 3]

Df Sum Sq Mean Sq F value Pr(>F)

iris$Species 2 437.10 218.551 1180.2 < 2.2e-16 ***

Residuals 147 27.22 0.185

Response : iris[, 4]

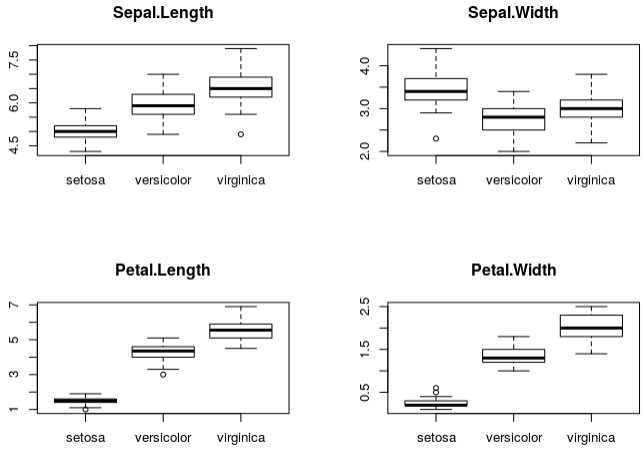
Df Sum Sq Mean Sq F value Pr(>F)

iris$Species 2 80.413 40.207 960.01 < 2.2e-16 ***

Residuals 147 6.157 0.042

par(mfrow=c(2,2))

for (i in 1 :4) boxplot(iris[,i] iris$Species,main=names(iris)[i])



par(mfrow=c(1,1))

Existe-t-il une différence entre les centres de gravités ?

summary(manova(as.matrix(iris[,1 :4]) iris$Species), test="Wilks")
Df Wilks approx F num Df den Df Pr(>F)
iris$Species 2 0.023439 199.15 8 288 < 2.2e-16 ***
Residuals 147

Analyse discriminante

library(MASS)

iris.lda=lda(iris$Species .,data=iris[,1 :4])

iris.lda$scaling

LD1 LD2
Sepal.Length 0.8293776 0.02410215
Sepal.Width 1.5344731 2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width -2.8104603 2.83918785

names(iris.lda)

[1] "prior" "counts" "means" "scaling" "lev" "svd" "N" "call" "terms" [10] "xlevels"

library(ade4)
iris.dis=discrimin(dudi.pca(iris[,1 :4],scan=FALSE),iris$Species,scan=FALSE)

names(iris.dis)

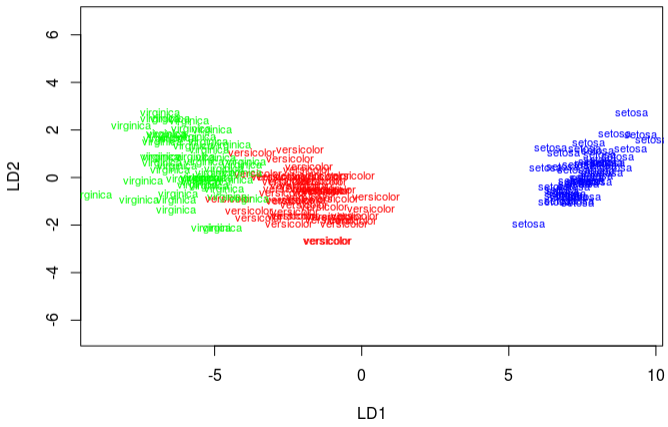
[1] "eig" "nf" "fa" "li" "va" "cp" "gc" "call"

#Qualité de l'AFD F*

anova(lm(iris.dis$li[,1] iris$Species))

Response : iris.dis$li[, 1]
Df Sum Sq Mean Sq F value Pr(>F)
iris$Species 2 145.481 72.740 2366.1 < 2.2e-16 ***
Residuals 147 4.519 0.031

plot(iris.lda,col=c("blue","red","green")[iris$Species])



plot(iris.dis)