Analyse des données

Joseph Rynkiewicz

11 décembre 2020

Chapitre 1

Analyse exploratoire des données

La pluspart du temps, les données se présentent sous la forme d'une table (data.frame en R). Il y a n individus ou observations et p variables. La synthèse de ces données se fait sous forme de résumés numériques et de graphiques.

1.1 Résumés numériques

Si on observe la réalisation $(x_1, \cdots, x_n)$ d'un n-échantillon de variables aléatoires, on peut résumé cet échantillon par plusieurs indicateurs :

- La moyenne empirique : $\bar{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$.
- La médiane empirique : m telle que $card(\{x_t,\, x_t \leq m\}) = card\left(\{x_t,\, x_t \geq m\}\right)$. Il y a autant de x_t plus petits que m que de x_t plus grands que m.
- Le mode : valeur la plus fréquente. Cela suppose que il y a des valeurs répétées dans l'échantillon.
- Des caractéristiques de la dispersion des observations :
  - La variance empirique : $S^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(x_t - \bar{x}_n\right)^2$
  - Les quartiles :
    - Q_1 tel que 25% de l'échantillon soit plus petit.
    - Q_2 tel que 50% de l'échantillon soit plus petit. Q_2 est égal à la médiane m.
    - Q_1 tel que 75% de l'échantillon soit plus petit.
  - Le minimum et la maximum de $(x_1, \cdots, x_n)$

Si les variables aléatoires sont vectorielles : $x_t = \begin{pmatrix} x_{1t} \\ \vdots \\ x_{dt} \end{pmatrix} \in \mathbb{R}^p$, alors la notion de variance est remplacée par la matrice de variance covariance :

$$\Sigma = \frac{1}{n}\sum_{t=1}^{n}\left(x_t - \bar{x}_n\right)\left(x_t - \bar{x}_n\right)^T = \left(Cov\left(x_i, x_j\right)\right)_{1\leq i,j\leq p}$$

De plus pour décrire les liens entre les variables on utilise aussi la matrice de correlation

$$R = (Corr\,(x_i, x_j))_{1 \leq i,j \leq p}$$

où $Corr\,(x_i, x_j) = \frac{Cov(x_i,x_j)}{\sqrt{V(x_i)}\sqrt{V(x_j)}}$.

1.2 Représentations graphiques

1.2.1 Barres et camemberts

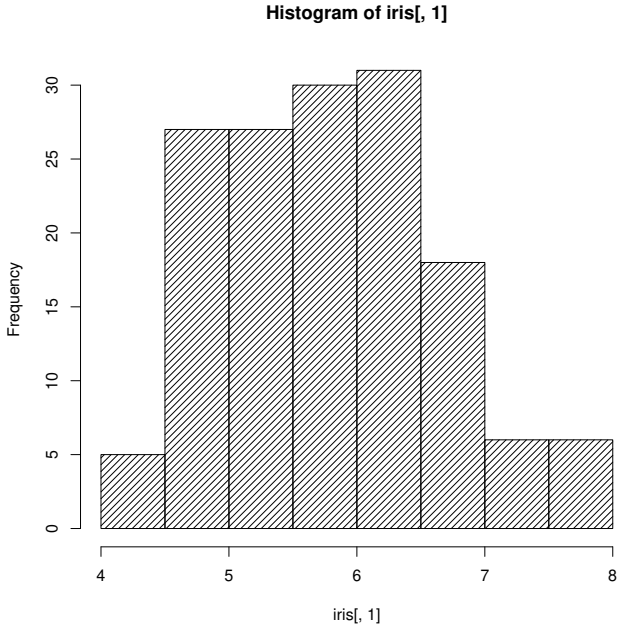
Pour des variables qualitatives à modalités non ordonnées, il existe une grand variété de diagrammes. Les plus répandus sont :

- Les diagrammes en barre : Les barres sont de longueurs proportionnelles aux fréquences des catégories, leur épaisseur est sans importance.
- Les camemberts (pie-chart en anglais) : chaque catégorie est représentée par une portion de superficie proportionnelle à sa fréquence.

1.2.2 Histogrammes

Un histogramme est un graphique à barre verticales accolées et de même largeur, obtenu après découpage en classe des observationsd'une variable continue. La hauteur (donc la surface) de chaque barre doit être proportionnelle à la fréquence de la classe. Par exemple, on pourra écrire sous R : $hist(iris[,1], density = 20)$ pour faire un histogramme de la première variable des données d'iris.

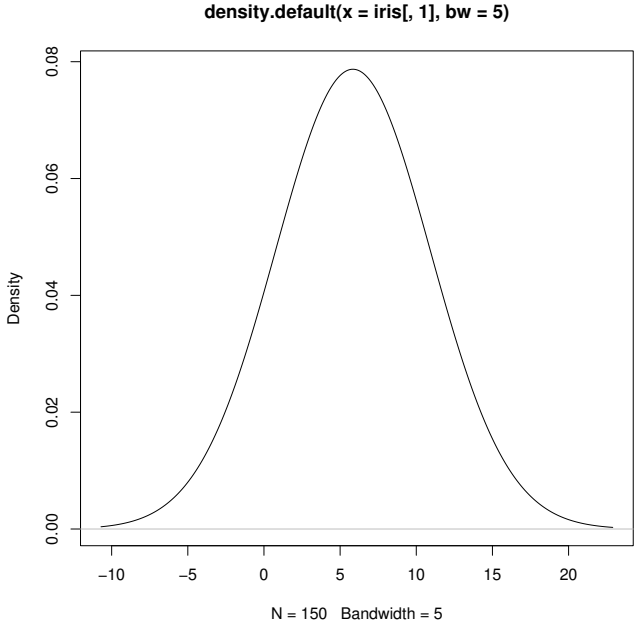
FIGURE 1.1 – Histogramme iris



La détermination du nombre de classes d'un histogramme est délicate, un trop faible nombre de classes fait perdre de l'information alors qu'un trop grand nombre de classes aboutit à des graphiques incohérents : certaines classes deviennent presque vides. On peut aussi critiquer le fait de représenter par une fonction en escalier la distribution d'une variable continue, c'est une approximation assez pauvre d'une fonction densité. La théorie de l'estimation de densité permet de proposer des solutions à ce problème.

Estimateurs à noyaux On approximera ainsi la fonction densité $f(x)$ par $\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$, où h est la fenêtre du noyau et K une fonction noyau, par exemple un noyau gaussien $K = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$. Par exemple en tapant la commande $d < -density(iris[,1], bw = 5)$, on obtiendra :

FIGURE 1.2 – Estimation de la densité



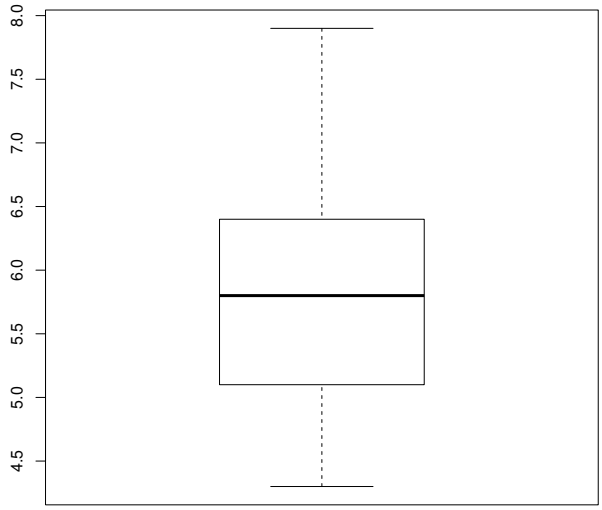
Cette méthode est facilement généralisable en dimension 2, c'est-à-dire si $x \in \mathbb{R}^2$, on aura alors :

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

1.2.3 La boîte à moustache ou box-plot

Ce diagramme est une représentation synthétique des principales caractéristiques d'une variable numérique. La boîte correspond à la partie centrale de la distribution : la moitié des valeurs comprises entre le premier et le troisième quartile Q_1 et Q_3. Les moustaches s'étendent de part et d'autre de la boîte jusqu'au valeurs $Q_1 - 1.5\,(Q_3 - Q_1)$ et $Q_3 + 1.5\,(Q_3 - Q_1)$ si il existe des valeurs au delà de cet intervalle sinon jusqu'aux valeurs extrêmes. Les points en dehors de cet intervalle sont aussi affichés. Par exemple, la commande $boxplot(iris[, 1])$ donnera :

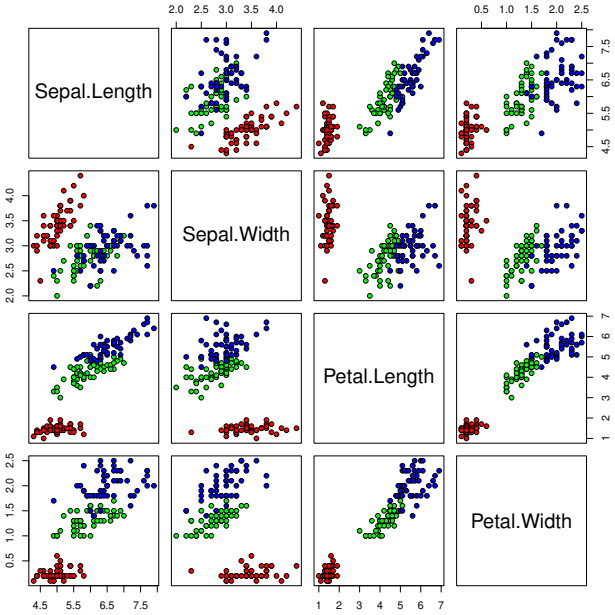
FIGURE 1.3 – Boxplot Iris



1.2.4 Le Scatterplot

Il s'agit de faire des graphiques en prenant en abcisse une variable et en ordonnée un autre (et cela pour tous les couples possibles). Cela permet de voir certaines relations deux à deux entre les variables. Par exemple la commande $plot(iris[, 1:4], pch = 21, bg = c("red","green","blue")[irisSpecies])$ donnera :

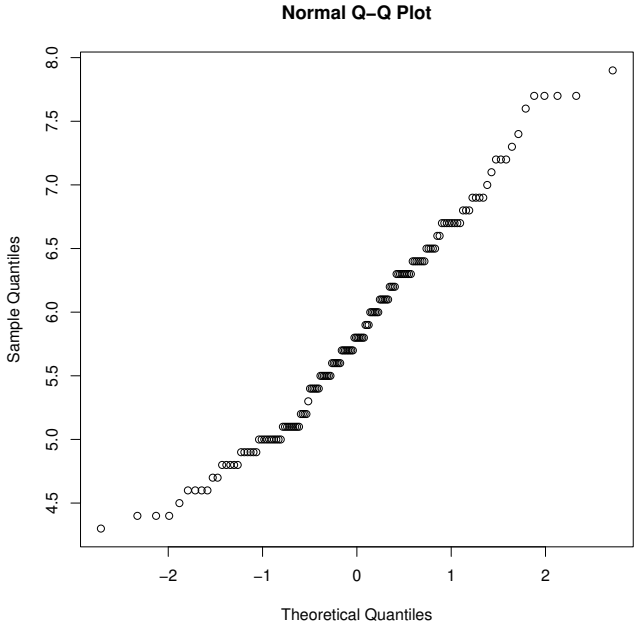
FIGURE 1.4 – Scatterplot iris



1.2.5 Le qqplot

Ce terme vient de quantile-quantile plot. Il permet de visualiser la distribution des données en comparant leur fonction de répartition empirique à une autre fonction de répartition comme, par exemple, la fonction de répartition théorique de la loi gaussienne. Ainsi, si la distribution est gaussienne la courbe de qqplot doit être proche d'une droite si elle est comparée a la fonction de répartition théorique de la loi gaussienne. La commande $qqnorm(iris[,1]$ donnera par exemple :

FIGURE 1.5 – qqplot Iris



Si on veut comparer la fonction de répartition de $x = (x_1, \cdots, x_n)$ avec $y = (y_1, \cdots, y_n)$, on écrira $qqplot(x, y)$, plus la courbe sera proche d'une droite plus les fonctions de répartition empiriques seront semblables.

Chapitre 2

Classification

2.1 Partitionnement d'un ensemble de données

Ces méthodes permettent de traiter rapidement des ensembles d'effectif assez élevé en optimisant localement un critère de type inertie. On supposera que les individus sont des points de $\mathbb{R}^p$ muni d'une distance euclidienne $\|.\|$.

2.1.1 Inertie interclasse et intraclasse

Soit $(x_1, \cdots, x_n)$ un ensemble de données avec $x_i \in \mathbb{R}^p$, partitionné en K groupes $I_1, \cdots, I_K$. Pour chaque groupe de cardinal n_k le centre de gravité sera $g_k = \frac{1}{n_k} \sum_{x_i \in I_k} x_i$. On aura alors

- L'inertie totale S : $S = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2$
- L'inertie intraclasse W : $\frac{1}{n} \sum_{k=1}^{K} \sum_{x_i \in I_k} \|x_i - g_k\|^2$
- L'inertie interclasse B : $\frac{1}{n} \sum_{k=1}^{K} n_k \|g_k - \bar{x}_n\|^2$

L'inertie totale S est alors la somme de W et B (Théorème de Huygens) :

$$S = W + B$$

La qualité d'une classification (ou partitionnement) des données dépendra de l'inertie intraclasse :

- Une classe est homogène si les observations des classes sont proches du centre de gravité des classes. Cela revient à dire que W est petite.
- Comme on a $S = W + B$ et que S est constante, c'est équivalent à dire que B est grande (les classes sont très différentes).

2.2 Agrégation autour des centres mobiles (K-means)

Cette méthode de classification est probablement la plus adaptée aux vastes recueils de données.

2.2.1 Base théorique de l'algorithme

Soit un ensemble I de n individus à partitionner, caractérisés par p variables. On suppose que l'espace $\mathbb{R}^p$ supportant les n points-individus est muni d'une distance appropriée notée d (souvent la distance euclidienne). On désire constituer au maximum q classes :

— Etape 0 : On détermine q centre provisoires de classes (par exemple par tirage aléatoire de q individus dans la population à classer). On obtient les centres

$$\left\{C_1^0, \cdots, C_q^0\right\}$$

qui induisent une partition P^0 de l'ensemble des individus en q classes

$$\left\{I_1^0, \cdots, I_q^0\right\}$$

Ainsi, l'individus i appartient à la classe I_k^0 s'il est plus proche de C_k^0 que tous les autres centres.

— Etape 1 : On détermine q nouveaux centres de classes

$$\left\{C_1^1, \cdots, C_q^1\right\}$$

en prenant les centres de gravité des classes $\left\{I_1^0, \cdots, I_q^0\right\}$ (i.e. $\left\{C_1^1, \cdots, C_q^1\right\}$ minimisent $\sum_{k=1}^{q} \frac{1}{n} \sum_{i \in I_k^m} \left\|i, C_k^1\right\|^2$). Ces nouveaux centres induisent une nouvelle partition P^1 de I construite selon la même règle que P^0 :

$$\left\{I_1^1, \cdots, I_q^1\right\}$$

— Etape m : On recommence cet algorithme jusqu'à ce que les classes se stabilisent.

2.2.2 Justification élémentaire de l'algorithme

On va montrer que la variance intra-classe ne peut que décroitre et conclure à la convergence de l'algorithme car le nombre de partitionnement possible de l'ensemble I est fini.

Nous nous interessons à la quantité "critère" :

$$v(m) = \sum_{k=1}^{q} \frac{1}{n} \sum_{i \in I_k^m} \left\|i, C_k^m\right\|^2$$

Rappelons qu'à l'étape m, la classe I_k^m est formé des individus les plus proches de C_k^m que de tous les autres centres (ces centres étant des centres de gravité des classes I_k^{m-1} de l'étape précédente). La variance intra-classe à l'étape m est la quantité :

$$V(m) = \sum_{k=1}^{q} \frac{1}{n} \sum_{i \in I_k^m} \left\|i, C_k^{m+1}\right\|^2$$

A l'étape $m+1$ la quantité critère est

$$v\left(m+1\right)=\sum_{k=1}^{q}\frac{1}{n}\sum_{i\in I_k^{m+1}}\left\|i,C_k^{m+1}\right\|^2$$

On va montrer que

$$v\left(m\right)\geq V\left(m\right)\geq v\left(m+1\right)$$

La première inégalité est vraie par définition. La seconde partie découle du fait que dans les sommes qui apparaissent dans les définitions de $V(m)$et $v(m+1)$, seules changent les affectations des points au centres et que les distances n'ont pu que décroître. Cet algorithme fourni un minimum local de la variance intra-classe. Les partitions obtenues dépendent des premiers centre choisis.

2.3 La Classification hiérarchique

Les principes généraux communs aux diverses techniques de classification ascendante hiérarchique sont extrêmement simples.

2.3.1 Principe

Le principe de l'algorithme consiste à créer, à chaque étape, une partition obtenue en agrégeant deux à deux les éléments les plus proches.

L'algorithme ne fournit pas une partition en q classes d'un ensemble de n objets mais une hiérarchie de partitions se présentant sous la forme d'arbres appelés également dendrogrammes et contenant $n-1$ partitions. Chaque coupure d'un arbre fournit une partition ayant d'autant moins de classes et des classes d'autant moins homogènes que l'on coupe plus haut.

2.3.2 Distance entre éléments et entre groupes

On suppose au départ que l'ensemble des individus à classer est muni d'une distance. On construit alors une première matrice de distances entre tous les individus. Une fois constitué un groupe d'individus, il convient de se demander ensuite sur quelle base on peut calculer une distance entre un individu et un groupe et par la suite une distance entre deux groupes. Cette distance entre groupe sera calculée à partir des distances des différents éléments impliqués dans le regroupement.

Par exemple, si x, y et z sont trois objets et si les objets x et y sont regroupés en un seul élément noté h, on peut définir la distance de ce groupement à z (qui est évantuellement un regroupement) par la plus petite distance des divers éléments de h à z :

$$d\left(h,z\right)=\min\left\{d\left(x,z\right),d\left(y,z\right)\right\}$$

Cette distance s'appelle le saut minimale (single linkage) et constitue un critère d'agrégation.

On peut définir aussi la distance du saut maximale :

$$d(h,z) = \max\{d(x,z), d(y,z)\}$$

Une autre règle simple et fréquemment employée est celle de la distance moyenne ; pour deux objets x et y regroupés en h :

$$d(h,z) = \frac{d(x,z) + d(y,z)}{2}$$

Plus généralement, si $x$et y désignent des sous-ensembles disjoints de l'ensemble des objets ayant respectivement $n_x$et n_y éléments, h est alors un sous-ensemble formé de $n_x + n_y$ éléments et on définit :

$$d(h,z) = \frac{n_x d(x,z) + n_y d(y,z)}{n_x + n_y}$$

2.3.3 Algorithme de classification

L'lagorithme fondamental de classification ascendante hiérarchique se déroule de la façon suivante :

Etape1 Il y a n éléments à classer (qui sont les n individus).

Etape2 On construit la matrice de distance entre les n éléments et l'on cherche les deux les plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à $n-1$classes.

Etape3 On construit une nouvelle matrice des distances qui résulte de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se retrouve dans les mêmes conditions que l'étape1 mais avec seulement $(n-1)$ éléments à classer et en ayant un critère d'agrégation. On cherche à nouveau les deux éléments les plus proches, que l'on agrège. On obtient une nouvelle partition avec $n-2$ classes et qui englobe la première.

EtapeM On calcule les nouvelles distances et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

2.3.4 Critère d'agrégation selon la variance

Les techniques de classification selon le saut minimal ont l'avantage de conduire à des calculs simples (pas de recalcul numérique des distances) et possèdent des propriétés mathématiques intéressantes. Les techniques d'agrégation selon la variance cherchent à optimiser, à chaque étape, selon des critères liés à des calculs d'inertie, la partition obtenue par agrégation de deux éléments.

2.3.5 Notations et principe

Nous considérons ici les n objets à classer comme un nuage de points (le nuage des individus) d'un espace à p dimensions (espace des variables). Chaque point x_i (vecteur à p composantes) est muni d'une masse m_i. On note m la masse totale du nuage

$$m = \sum_{i=1}^{n} m_i$$

Le carré de la distance entre les points x_i et x_j est noté $\|x_i - x_j\|^2$. L'inertie totale du nuage est la quantité

$$I = \sum_{i=1}^{n} m_i \|x_i - g\|$$

où g désigne le centre de gravité du nuage

$$g = \frac{1}{m} \sum_{i=1}^{n} m_i x_i$$

S'il existe une partition de l'ensemble des éléments en s classes, la $q$ième classe a pour masse :

$$m_q = \sum_{i \in q} m_i$$

et à pour centre de gravité

$$g_q = \frac{1}{m_q} \sum_{i \in q} m_i x_i$$

On cherche alors à minimiser l'inertie intra-classe

$$I_{intra} = \sum_{q} \sum_{i \in q} m_i \|x_i - g_q\|^2$$

Le principe de l'algorithme d'agrégation selon la variance consiste à rechercher à chaque étape une partition telle que la variance interne de chaque classe soit minimale (Critère de Ward).

2.4 Modèle de mélange et algorithme EM

Il s'agit d'une approche probabiliste, on suppose que l'on à K classes : $I_1 \cdots, I_K$ et pour chaque classes I_k, les observations suivent une loi dont le paramètre θ_k dépent de cette classe. Le plus souvent la loi est gaussienne et le vecteur paramètre sera $\theta_k = (\mu_k, \Sigma_k)$ et $X_i \sim \mathcal{N}(\mu_k, \Sigma_k)$ si $X_i \in I_k$. On parlera alors de mélange gaussien.

2.4.1 Le modèle

On dispose d'un échantillon $(x_1 \cdots, x_n)$, où $x_i \in \mathbb{R}^p$ et on suppose que la population est formée de K groupes et pour $k \in \{1, \cdots, K\}$ la proportion du groupe k sera π_k et les variables du groupe k a pour de densité $f(x; \theta_k)$, le vecteur paramètre sera donc $\theta = (\theta_1, \cdots, \theta_K, \pi_1, \cdots, \pi_K)$. Ce modèle correspond au modèle sous-jacent suivant : Il existe un n-échantillon i.i.d. $\left(\begin{pmatrix} Z_1 \\ X_1 \end{pmatrix}, \cdots, \begin{pmatrix} Z_n \\ X_n \end{pmatrix}\right)$ où Z_i est dans $\{1, \cdots, K\}$ et $P(Z_i = k) = \pi_k$. La vraisemblance d'une réalisation de cet échantillon sera donc

$$L_\theta\left(\begin{pmatrix} z_1 \\ x_1 \end{pmatrix}, \cdots, \begin{pmatrix} z_n \\ x_n \end{pmatrix}\right) = \prod_{t=1}^{n} f(x_t, z_t; \theta) = \prod_{t=1}^{n} f(x_t \,|z_t\,; \theta) \times P(Z_t = z_t; \theta)$$

On verra plus tard qu'il s'agit du modèle de l'analyse discriminante quadratique.

Comme on observe pas directement la variable Z_t (on dit que la varialbe Z_t est une variable latente), la densité de X_t sera $f(x_t; \theta) = \sum_{k=1}^{K} f(x_t, k; \theta) = \sum_{k=1}^{K} f(x_t, |Z_t = k\,; \theta) \times P(Z_t = k; \theta) = \sum_{k=1}^{K} \pi_k f(x; \theta_k)$ et $(X_1, \cdots, X_n)$ sera i.i.d. de densité :

$$f(x; \theta) = \sum_{k=1}^{K} \pi_k f(x; \theta_k).$$

Si les densité sont gaussiennes, on aura :

- $\theta_k = \left(\mu_k, \sigma_k^2\right)$ et $f(x; \theta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$ dans le cas uni-dimentionnel.
- $\theta_k = (\mu_k, \Sigma_k)$ et $f(x; \theta_k) = \left(\frac{1}{2\pi^p |\Sigma_k|}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k)\right)$ dans le cas multi-dimentionnel.

On pourra ainsi calculer la probabilité pour un point d'appartenir à un groupe k :

$$P(x \in I_k) = \frac{\pi_k f(x; \theta_k)}{f(x)}$$

On estimera le vecteur paramètre θ, par maximum de vraisemblance. Ce maximum est approximé numériquement par un algorithme itératif (descente de gradient ou bien algorithme E.M.) qui fait croître la vraisemblance à chaque étape.

2.4.2 Algorithme E.M.

1. On choisit un vecteur paramètre au hasard $\theta^0 = \left(\theta_1^0, \cdots, \theta_K^0, \pi_1^0, \cdots, \pi_K^0\right)$, c'est l'étape $i = 0$.
2. Etape E : Pour $k \in \{1, \cdots, K\}$, et pour chaque points x_j, on calcule la probabilité des points d'appartenir au groupes k :

$$\pi_k^{i+1}(x_j) = \frac{\pi_k^i f\left(x_j; \theta_k^i\right)}{f\left(x_j; \theta^j\right)}$$

3. Etape M : On estime les paramètres $\left\{\theta_1^{i+1}, \cdots, \theta_K^{i+1}\right\}$ grâce aux pondérations $\pi_k^{i+1}\left(x_j\right)_{1\leq k\leq K,\, 1\leq j\leq n}$. et on retourne à l'étape E.

On arrête l'algorithme quand la vraisemblance ne croît plus assez.

2.4.3 Determination du nombre de classes

Comme on aura une estimation de la vraisemblance $L_\theta\left(x_1, \cdots, x_n\right)$ on utilise un critère d'information de type BIC pour déterminer le nombre de classes : si $|\Theta_K|$ est le nombre de paramètres libres du modèle avec K classes et et $\hat{L}_K\left(x_1, \cdots, x_n\right)$ la vraisemblance maximale pour $\theta \in \Theta_K$

$$\hat{K} = \arg\min -\log \hat{L}_K\left(x_1, \cdots, x_n\right) + |\Theta_K| \log(n)$$

Cela revient à pénaliser la vraisemblance (qui ne peut que croître avec K) par la dimension du modèle.

2.5 Classification supervisée

On suppose que l'on dispose d'un échantillon de taille n pour lequel les p variables explicatives et la variable à expliquer ont été mesurées simultanément. Généralement cet échantillon est appelé échantillon d'apprentissage : $\left(\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_n \\ X_n \end{pmatrix}\right)$ de vecteurs aléatoires i.i.d. avec Y_t dans $\{1, \cdots, K\}$ qui est la classe et X_t un vecteur aléatoire dans $\mathbb{R}^p$. On veut définir à partir de cet échantillon d'apprentissage une règle de classification qui va permettre de prédire la valeur de Y pour un nouvel individu sur lequel on a mesuré uniquement les p variables explicatives. On parle de classification supervisée, chaque modalité de Y représentant une classe (un groupe) d'individus. Il faut faire attention au phénomène de surapprentissage, ou le modèle (la règle) de classification est très compliquée, suradaptée à l'ensemble d'apprentissage, mais très mauvaise pour classée de nouvelles observations, jamais "vu" par le modèle. Dans ce cours, les modèles considérés seront simples (linéaires) et on peut considérer que le surapprentissage est faible si la dimension des variables explicatives p n'est pas trop grande.

2.5.1 Analyse discriminante linéaire

On cherche de nouvelles variables (combinaisons linéaires des coordonnées de la variables explicative $X_t = \begin{pmatrix} X_{1t} \\ \vdots \\ X_{pt} \end{pmatrix}$) qui descriminent aux mieux les groupes. C'est-à-dire, qu'elles maximisent la variance interclasse B et minimise la variance intraclasse W. Notons $X = \begin{pmatrix} X_1^t \\ \vdots \\ X_n^t \end{pmatrix}$ le tableau qui regroupe toutes

les variables explicative. On commence par chercher un axe discrimant u telle que pour la variable $X \times u$ les projections des centres de gravité $g_1, \cdots, g_K$ soient les plus écartées possible, tandis que chaque sous-nuage des classes doit se projeter de manière groupée autour de la projection de son centre de gravité.

2.5.1.1 Notations

Si les indices des individus de la classe k sont regroupés dans I_k de cardinal n_k, on rappelle que la centre de gravité de la classe k est $g_k = \frac{1}{n_k} \sum_{j \in I_k} X_j$. Notons $M_{V_k} = \frac{1}{n_k} \sum_{j \in I_k} (X_j - g_k)(X_j - g_k)^t$ la matrice de variance-covariance de la classe k, $M_W = \frac{1}{n} \sum_{k=1}^{K} n_k V_k$ la matrice de variance-covariance intraclasse et $M_B = \frac{1}{n} \sum_{k=1}^{K} n_k (g_k - g)(g_k - g)^t$ la matrice de variance-covariance interclasse, avec $g = \frac{1}{n} \sum_{t=1}^{n} X_t$ le centre de gravité de tout l'échantillon. Si on note $M_S = \frac{1}{n} \sum_{t=1}^{n} (X_t - g)(X_t - g)^t$, on remarquera que l'on a une généralisation du théoreme de Théorème de Huygens :

$$M_S = M_W + M_B.$$

2.5.1.2 Axes et variables discriminantes

La matrice de variance covariance des centres de gravité étant M_B, la variance des projections des centre de gravité sur l'axe u sera $u^t M_B u$, on cherchera à la maximiser. Les variances de chaque sous-nuage seront $u^t M_{V_k} u$, pour $k \in \{1, \cdots, K\}$ et on cherchera à minimiser leur somme pondérée $u^t M_W u$, mais cela revient au même car $M_S = M_W + M_B = Cte$. On prendra donc comme critère la maximisation du rapport de la variance interclasse à la variance totale :

$$\max_u \frac{u^t M_B u}{u^t M_S u},$$

ce maximum sera atteint si $u = u_1$ est vecteur propre de $M_S^{-1} M_B$ associé à la plus grande valeur propre λ_1 : $M_S^{-1} M_B u_1 = \lambda_1 u_1$. Comme $u_1 = \begin{pmatrix} u_{11} \\ \vdots \\ u_{p1} \end{pmatrix} \in \mathbb{R}^p$, la combinaison linéaire des variables explicatives la plus discriminante sera donc $Xu_1 = \begin{pmatrix} \sum_{j=1}^{p} X_{j1} u_{j1} \\ \vdots \\ \sum_{j=1}^{p} X_{jn} u_{jn} \end{pmatrix}$.

Remarque En recommençant avec la nouvelle variable $X - Xu_1$ on pourra trouver le deuxième axe le plus discriminant etc...

On pourra par exemple faire l'analyse discriminante des données iris en tapant :

```
library(MASS)
lda(Species~.,data=iris)
```

Règle de classification On présente d'abord une approche purement géométrique (sans aucune hypothèse probabiliste). Notons x le vecteur des valeurs des p variables explicatives sur un nouvel individu dont que l'on veut classer. La règle géométrique consiste à calculer la distance de x à chacun des K centres de gravité $g_1, \cdots, g_K$ et à affecter x au groupe le plus proche. Pour cela, il faut préciser la métrique à utiliser dans le calcul des distances. La règle la plus utilisée est celle de Mahalanobis-Fisher qui consiste à prendre la métrique M_W^{-1} (ou M_S^{-1} ce qui est équivalent). La distance du nouvel individu au groupe k est alors :

$$d^2(x, g_k) = (x - g_k)^t M_W^{-1} (x - g_k)$$

La règle géométrique classe la nouvelle observation x dans le groupe k^* tel que :

$$k^* = \arg \min_{k \in \{1,\cdots,K\}} d^2(x, g_k)$$

ce qui se réécrit :

$$k^* = \arg \max_{k \in \{1,\cdots,K\}} L_k(x)$$

où

$$L_k(x) = x^t M_W^{-1} g_k - \frac{1}{2} g_k^t M_W^{-1} g_k$$

2.5.1.3 Analyse discriminante quadratique

Il s'agit du même modèle que le modèle de mélange gaussien, sauf que la classe des variable est observée. L'échantillon d'apprentissage est issu d'une population en K groupes $I_1, \cdots, I_K$ et :

— Y est une variable aléatoire qui prend ses valeurs dans $\{1, \cdots, K\}$.

— $X = (X_1, \cdots, X_p)^t$ est un vecteur de variables aléatoires réelles.

On notera :

— $(\pi_1, \cdots, \pi_K)$ la distribution de Y où $\pi_k = P(Y = k)$ est la proportion théorique de I_k encore appelée probabilité à priori de I_k.

— $f_k : \mathbb{R}^p \to [0, 1]$ la densité de X dans le groupe I_k.

La densité de X dans la population toute entière sera donc une densité de mélange :

$$X \sim \sum_{k=1}^{K} \pi_k f_k(x)$$

La règle de classement optimale de Bayes affecte une nouvelle observation x au groupe le plus probable sachant x :

$$k^* = \arg \max_{k \in \{1,\cdots,K\}} P(Y = k \,|x\,).$$

C'est la probabilité conditionnelle appelée la probabilité à posteriori de I_k. Cette règle se réécrit :

$$k^* = \arg \max_{k \in \{1,\cdots,K\}} \pi_k f_k(x).$$

apprentissage de cette règle de classification Ici on se place dans le cadre paramétrique gaussien. On suppose maintenant que $X \sim \mathcal{N}(\mu_k, \Sigma_k)$ dans chaque groupe I_k :

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

Dans ce cas, la règle de Bayes se réécrit :

$$k^* = \arg\min_{k\in\{1,\cdots,K\}} D_k^2(x),$$

où

$$D_k^2(x) = (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k) - 2\ln(\pi_k) + \ln(\det(\Sigma_k))$$

Estimation des paramètres A partir de l'échantillon d'apprentissage, on veut estimer le paramètre

$$\theta = (\pi_1, \cdots, \pi_{K-1}, \mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K).$$

La méthode du maximum de vraisemblance peut être utilisée, la vraisemblance s'écrit :

$$L_\theta\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \cdots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\right) = \prod_{k=1}^{K}\prod_{t=1}^{n} (\pi_k f_k(x_t))^{\mathbf{1}_{\{k\}}(y_t)}$$

et on en déduit que la log-vraisemblance s'écrit :

$$\ln(L_\theta(\cdots)) = \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbf{1}_{\{k\}}(y_t)\left(\ln(\pi_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(\Sigma_k)) - \frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\right).$$

En notant $n_k = \sum_{t=1}^{n} \mathbf{1}_{\{k\}}(y_t)$, le nombre d'observations qui appartiennent au groupe k, on obtient alors les estimateurs du maximum de vraisemblance suivant :

$$\begin{array}{l} \hat{\pi}_k = \frac{n_k}{n} \\ \hat{\mu}_k = \frac{1}{n_k}\sum_{t=1}^{n} \mathbf{1}_{\{k\}}(y_t)\, x_t \\ \hat{\Sigma}_k = \frac{1}{n_k}\sum_{t=1}^{n} \mathbf{1}_{\{k\}}(y_t)\,(x_t-\mu_k)^t(x_t-\mu_k) \end{array}$$

règle de classification d'analyse discriminante quadratique On aura

$$k^* = \arg\min_{k\in\{1,\cdots,K\}} Q_k(x),$$

où

$$Q_k(x) = (x-\hat{\mu}_k)^t \hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) - 2\ln(\hat{\pi}_k) + \ln\left(\det\left(\hat{\Sigma}_k\right)\right).$$

On pourra par exemple faire l'analyse discriminante quadratique des données iris en tapant :

```
library(MASS)
qda(Species~.,data=iris)
```

2.5.2 Arbres de décision

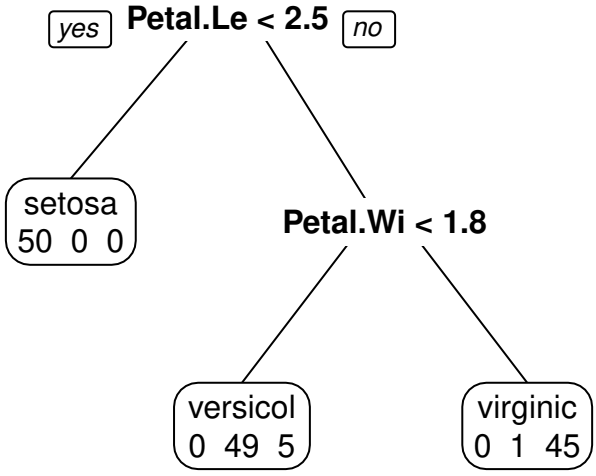
2.5.2.1 Introduction

Un arbre de décision est un modèle de classification sous forme d'arbre. Ces méthodes ont été développées dès les années 60 en marketing. Ces outils sont devenus populaires en raison de la lisibilité des résultats. On peut les utiliser pour prédire une variable Y quantitative (arbre de régression) ou qualitative (arbre de classification) à l'aide de prédicteurs X quantitatifs ou qualitatifs. Les différentes décisions possibles sont situées aux extrémités des branches (les feuilles). Un exemple d'arbre de décision sur les données des Iris de Fisher est le donnée figure 3.1.

Cette arbre correspond à la règle de classification suivante :

Si la longueur du pétale est plus petite que 2.5 alors l'espèce est Setosa. Sinon, si la largeur du pétale est plus petite que 1.8 l'espèce est Versicolor sinon l'espèce sera Virginica. Cette règle classe parfaitement les Setosa, mais elle classe par erreur 5 Virginica en Versicolor et 1 Versicolor en Virginica.

FIGURE 2.1 – Arbre de décision pour Iris



2.5.2.2 Construction de l'arbre

La variable à expliquer Y est binaire Dans ce cas $Y \in \{0, 1\}$. Le procédé consiste à diviser l'échantillon d'apprentissage en deux sous-ensemble à l'aide d'une des variables explicatives $X^1, \cdots, X^p$. Ensuite on recommence séparément dans chaque sous-ensemble etc... Pour chaque variable X^i il faut donc trouver la meilleur partition de ses valeurs en deux sous-ensemble de taille n_1 et n_2 qui soient le plus homogène possible vis à vis de Y. Le nombre de divisions en deux sous-ensembles que l'on peut réaliser en fonction de X^i dépend de sa nature :

- Si X^i est qualitative à m modalité, il y a $2^m - 1$ dichotomies possibles.
- Si X^i est ordonnée (quantitative ou qualitative ordinale) à m valeurs distinctes il y a $m - 1$ dichotomies possibles.

La variable à expliquer Y a $K > 2$ classes On définit tout d'abord une mesure d'impureté vis à vis de Y. Cette mesure doit être nulle si tous les individus appartiennent à la même modalité de Y , maximale si les K catégories sont en proportion égales. Pour $i \in \{0, \cdots, K-1\}$, si on note $p_i = P(Y = i)$ les deux mesures les plus usuelles sont l'entropie $\sum_{i=0}^{K-1} p_i \ln(p_i)$ et l'indice de diversité de Gini $\sum_{i=0}^{K-1} p_i (1 - p_i)$. On cherche alors la division en deux sous-ensemble qui conduit à la diminution maximale de l'impureté. Le nombre de feuilles (ou noeuds terminaux) croit exponentiellement avec le niveau de l'arbre et il nécessaire de fixer des limites, sinon l'arbre va sur-apprendre les données d'apprentissage. En laissant croître indéfiniment l'arbre on finira avec des feuilles avec une seule observation et le taux d'erreur de classement sera nul. Mais le modèle risque fortement de généraliser très mal sur de nouvelles données. La méthodologie "CART" permet d'éviter ce problème.

Soit T_0 l'arbre maximal avec une erreur de classification nulle. L'objectif est de trouver un sous-arbre T obtenu en coupant certaines branches qui va un bon compromis entre son erreur d'apprentissage $C(T)$ et sa complexité mesurée par le nombre de feuilles $|T|$. On utilise une mesure pénalisée de performance : $C(T) + \alpha |T|$ où α est un hyper-paramètre que l'on détermine par validation croisée. Dans le package rpart de R, il est proportionel à l'option *cp*.

On pourra, par exemple, faire une classification et afficher l'arbre de classification des données iris en tapant sous R :

```
library(rpart)
library(rpart.plot)
iris.rpart<-rpart(Species~.,data=iris)
prp(iris.rpart)
```

2.5.3 Forêts aléatoires

Les forêts aléatoires consistent à agréger des modèles d'arbres pour rendre la prédiction plus robuste et plus précise. Il s'agit de bootstrapper les obervations en tirant N échantillons avec remise dans l'échantillons initial, en utilisant pour chaque tirage seulement b prédicteurs ($b < p$).

— Sur chaque échantillon, on entraine un arbre de décision en suivant la méthode de la section précédente.
— On stocke les N prédictions de la variable d'intérêt pour chaque observation d'origine.
— La prédiction de la forêt aléatoire est alors un simple vote majoritaire (Ensemble learning).

Cette méthode est plus efficace qu'avec un seul arbre, par contre on perd l'explicabilité simple d'un seul arbre. On obtient un modèle boîte noire.

On pourra, par exemple, faire une classification des données iris avec les forêts aléatoires en tapant sous R :

```
library(randomForest)
data(iris)
#  mtry = nombre de variable choisies pour chaque sous ensembles
iris_RandomForest <- randomForest(Species~.,data=iris, ntree = 100,
                                  mtry = 2, na.action = na.roughfix)
print(iris_RandomForest)
#Importance des variables
iris_RandomForest$importance[order(iris_RandomForest$importance[, 1],
decreasing = TRUE), ]
```

2.5.4 Régression logistique

Supposons que pour une population $\mathcal{P}$, on dispose d'une variable Y qui ne peut prendre que 2 valeurs. Par convention on associera à ces deux valeurs les valeurs canoniques 0 et 1. Ce genre de variables apparaissent naturellement dans de nombreux problèmes (mort/vie, activité/chômage, plus généralement toute réponse du type oui/non).

On dispose d'un n-uplet de cette population $(Y_1, \cdots, Y_n)$ de variables indépendantes, c'est-à-dire d'un tirage au hasard, avec remise de n individus de la population $\mathcal{P}$, dont on peux connaître la réponse Y_i pour chaque individu i.

On peut alors écrire

$$P(Y = 1) = \pi \text{ ou } P(Y = 0) = 1 - \pi$$

où π_i est la probabilité de "succès".

Dans la plupart des cas, les observations Y sont associées avec des variables explicatives (ou exogènes) $\left(X^1, \cdots, X^p\right)$. On dispose donc en réalité des n-uplets

$$\begin{matrix} \left(X_1^1, \cdots, X_1^p, Y_1\right) \\ \vdots \\ \left(X_n^1, \cdots, X_n^p, Y_n\right) \end{matrix}$$

où les variables $\left(X^1, \cdots, X^p\right)$ peuvent être numériques (quantitative), ou catégorielles (factors en anglais). Le principal objectif d'une analyse statistique est

d'étudier la relation entre la probabilité de réponse π_i et les variables explicatives $\left(X^1, \cdots, X^p\right)$.

2.5.4.1 Données explicatives catégorielles

On suppose ici que les variables explicatives sont catégorielles $\left(X^1, \cdots, X^p\right)$, c'est-à-dire qu'elles ne peuvent prendre qu'un nombre fini de valeurs possibles. Il est possible alors d'estimé les $P\left(Y = 1 \left| \left(X^1, \cdots, X^p\right)\right.\right)$ de façon efficace et sans biais (c'est-à-dire de façn optimale).

Prenons un exemple, on suppose que $p = 2$, et que les valeurs possibles pour X^1 et X^2 soient 0 et 1. Les combinaisons possibles pour les variables $\left(X^1, \cdots, X^p\right)$ sont alors forcément fini, il est facile de les énumérer :

$$\begin{array}{l}\left(X^1 = 1,\, X^2 = 1\right) := C_1 \\ \left(X^1 = 1,\, X^2 = 0\right) := C_2 \\ \left(X^1 = 0,\, X^2 = 1\right) := C_3 \\ \left(X^1 = 0,\, X^2 = 0\right) := C_4\end{array}$$

alors pour notre n-uplet

$$\begin{array}{c}\left(X_1^1, \cdots, X_1^p, Y_1\right) \\ \vdots \\ \left(X_n^1, \cdots, X_n^p, Y_n\right)\end{array}$$

Lorsque tous les C_i apparaissent dans l'échantillon plus d'une fois, on parle de données répétées. Dans ce cas particulier, on peut étudier le modèle saturé :

$$\hat{\pi}_i = \frac{\sum_{j/\left(X_j^1, X_j^2\right) = C_i} Y_j}{\sum_{j/\left(X_j^1, X_j^2\right) = C_i} 1}$$

qui est la proportion conditionnelle au classe C_i des réponses Y. On note alors

$$P\left(Y = 1 \left| \left(X^1, X^2\right) = C_i\right.\right) = \pi_i \simeq \hat{\pi}_i = \frac{\sum_{j/\left(X_j^1, X_j^2\right) = C_i} Y_j}{\sum_{j/\left(X_j^1, X_j^2\right) = C_i} 1}$$

ce qui revient à dire que $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ est une fonction de $\left(X^1, X^2\right)$ que l'on pourra donc noter

$$\begin{array}{l}\pi\left(X^1, X^2\right) = \\ \left(\pi\left(X^1 = 1, X^2 = 1\right), \pi\left(X^1 = 1, X^2 = 0\right), \pi\left(X^1 = 0, X^2 = 1\right), \pi\left(X^1 = 0, X^2 = 0\right)\right) \\ = (\pi_1, \pi_2, \pi_3, \pi_4)\end{array}$$

2.5.4.2 Modèle Logit

Paramétrisation On considère, à priori, que les données explicatives sont vectorielles. Si jamais elles sont qualitatives, elle seront forcément codées sur des simplexes (des vecteurs, avec que des 0 et un 1). Notons que si une variable qualitative a K états possibles, alors $\beta_0 + \beta^T X$ aura $K + 1$ paramètres si X est

codé sur le simplexe de $\mathbb{R}^K$, il faudra donc, en réalité, coder X sur le simplexe de $\mathbb{R}^{K-1}$ et lui permettre de prendre la valeur nulle. De même, si il y a plusieurs variables qualitatives il faudra imaginer que toutes ces variables et leurs interactions sont regroupées dans une seule variables de bonne dimension pour que les notations du reste de la section reste valide. Nous verrons ultérieurement (modèles log-linéaires) comment on peut décomposer ces croisements de variables en interactions. Le principe est exactement le même que l'analyse de variance.

Dans la suite on suppose que l'on regroupe toutes les variables quantitatives dans un vecteur X_{quant}, et les modalités possibles des variables qualitatives dans la variable X_{qual}. Pour étudier la relation entre la réponse π et les variables explicatives (X_{quant}, X_{qual}), il est pratique de construire un modèle formel de la fonction $\pi(X_{quant}, X_{qual})$. Les modèles linéaires ont un rôle prépondérant en économétrie et en statistique car leur propriétés sont généralement bien connue et ils sont relativement facilement interprétables. On va donc supposer que la dépendance entre Y et (X_{quant}, X_{qual}) intervient grâce à une combinaison linéaire de (X_{quant}, X_{qual}) :

$$\eta = \beta_0 + \beta_{quant}^T X_{quant} + \beta_{qual}^T X_{qual}$$

pour des coefficients $\beta_0, \beta_{quant}, \beta_{qual}$ inconnus. On remarque que généralement on aura $-\infty < \eta < \infty$, donc pour exprimer π à l'aide de η il faut utiliser une transformation g de π qui va de $]0,1[$ dans $\mathbb{R}$ et modéliser alors

$$g(\pi_i) = \eta_i = \beta_0 + \beta_{quant}^T X_{quant} + \beta_{qual}^T X_{qual}$$

pour $i = 1, \cdots, n$. Un choix standard de fonction g est la fonction logistique :

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

Si, par exemple, on étudie un modèle logistique avec deux variables explicatives scalaires X_1 et X_2, on aura le modèle

$$\ln \frac{\pi\left(X^1, X^2\right)}{1-\pi\left(X^1, X^2\right)} = \beta_0 + \beta_1 X^1 + \beta_2 X^2$$

d'où la rapport des chances conditionnellement à $X = \left(X^1, X^2\right)$ (odd ratio) vaudra

$$\frac{\pi\left(X^1, X^2\right)}{1-\pi\left(X^1, X^2\right)} = \exp\left(\beta_0 + \beta_1 X^1 + \beta_2 X^2\right)$$

soit

$$\pi\left(X^1, X^2\right) = \left(1 - \pi\left(X^1, X^2\right)\right) \exp\left(\beta_0 + \beta_1 X^1 + \beta_2 X^2\right)$$

d'où

$$\pi\left(X^1, X^2\right) = \frac{\exp\left(\beta_0 + \beta_1 X^1 + \beta_2 X^2\right)}{1 + \exp\left(\beta_0 + \beta_1 X^1 + \beta_2 X^2\right)}$$

par définition cette fonction est donc l'inverse de $g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$.

Interprétation des paramètres En supposant que X^1 et X^2 sont indépendants, on peut dire les choses suivantes :

— Si on augmente X^2 d'une unité, on "augmente" le rapport des chances en le multipliant par un facteur $\exp(\beta_2)$. Il est alors important que X^1 reste constant (d'où l'hypothèse d'indépendance de X^1 et X^2).

— La dérivé de $\pi\left(X^1, X^2\right)$ par rapport à X^2 vaudra

$$\frac{\pi}{\partial x^2} = \pi(1-\pi)\beta_2$$

donc un petit changement en x^2 aura un effet d'autant plus grand que la probabilité π est proche de 0.5.

— La façon la plus simple de représenter les résultats est certainement de donner les graphiques de

$$\pi(\eta) = \frac{\exp(\eta)}{1+(\exp(\eta))}$$

en faisant varié les X^i.

2.5.4.3 Modèles polytomiques

Si la réponse d'un individu ou une caractéristique appartient à un ensemble fixé fini on dit que la réponse est polytomique. Souvent les réponse peuvent être qualitatives (par exemple des groupes sanguins). On leur associera souvent des valeurs arbitraires $\{0, \cdots, K-1\}$. On notera alors

$$\pi_k = P(Y=k),\ k = 0, \cdots, K-1$$

Paramétrisation posons $a_k = \ln(\pi_k)$. Le modèle logistique polytomique nominale spécifie les rapports $\frac{\pi_k}{\pi_0}$:

$$a_k - a_0 = \ln\frac{\pi_k}{\pi_0} = X^T\beta_k$$

sous la contrainte $\beta_0 = 0$, le modèle est identifiable. Ce modèle peut aussi s'écrire

$$P(Y=k\,|X\,;\beta) = \pi_k(X;\beta) = \frac{\exp\left(X^T\beta_k\right)}{1+\sum_{l=1}^{K-1}\exp\left(X^T\beta_l\right)}$$

on remarque alors que pour $K=2$, on retrouve le modèle logit. Dans cette expression β_j est le changement de la $j$ème probabilité par unité de changement dans chaque composante de X. Ainsi, le rapport des chances en faveur de la catégorie j sur la catégorie j' sont augmentées d'un facteur

$$\frac{\pi_j(X)}{\pi_{j'}(X)} = \frac{\exp\left(X^T\beta_j\right)}{\exp\left(X^T\beta_{j'}\right)} \times \frac{1+\sum_{l=1}^{K-1}\exp\left(X^T\beta_l\right)}{1+\sum_{l=1}^{K-1}\exp\left(X^T\beta_l\right)} = \exp\left(X^T\left(\beta_j - \beta_{j'}\right)\right)$$

c'est donc la différence entre les vecteurs β_j plutôt que les vecteurs eux-mêmes qui sont intéressantes, c'est pourquoi on a pu, sans perdre en généralité, prendre $\beta_0 = 0$.

2.5.4.4 Décomposition en facteurs.

Supposons que nous avons les variables explicatives qualitatives X puisse être décomposés en facteurs $X = \left(X^1, X^2\right)$. On cherche simplement une analyse des influences des différentes facteurs.

L'indépendance des deux influences X^1, X^2 correspond au modèle $X^1 + X^2$. soit

$$\ln \frac{\pi_k(i,j)}{\pi_{00}} = \mu + \alpha_k(i) + \beta_k(j)$$

où α_i (resp. β_j) est le coefficient associé au niveau i (resp. j) de la variable X^1 (resp. X^2).

Maintenant, un modèle plus compliqué pourrait être : $X^1 * X^2$, qui signifie que l'influence du facteur X^2 dépend du niveau de X^1.

Paramétrisation des modèles à deux facteurs X^1 et X^2 Dans toute la suite on notera de la même façon l'ensemble de facteur I et son cardinal lorsqu'il n'y aura pas d'ambiguïté.

Pour chaque k, on choisit $\pi_k(0,0)$ comme état de référence $\left(X^1, X^2\right) = (0,0)$ et on pose

$$\theta := \left(\theta_k(i,j) = \ln \frac{\pi_k(i,j)}{\pi_k(0,0)}\right)_{i,j \in I \times J, k \in \{1,\cdots,K-1\}}$$

le nombre de paramètre libre est donc $(IJ-1)(K-1)$. Définissons les effets principaux :

- Les $(I-1)$ effets principaux $\alpha^k := \left(\alpha_i^k = \theta_k(i,0)\right)_{i=1,\cdots,I-1}$
- Les $(J-1)$ effets principaux $\beta^k := \left(\beta_j^k = \theta_k(0,j)\right)_{j=1,\cdots,J-1}$
- Les $(I-1)\times(J-1)$ interactions $(\alpha\beta)^k := \left((\alpha\beta)_{ij}^k\right) = \theta_k(i,j) - \theta_k(i,0) - \theta_k(0,j)$, $i \times j \neq 0$

La représentation des $\theta_k(i,j)$ est alors

$$\theta_k(i,j) = \mu_k + \alpha_i^k + \beta_i^k + (\alpha\beta)_{ij}^k$$

L'avantage de la Paramétrisation en $\left(\alpha^k, \beta^k, (\alpha\beta)^k\right)_{k \in \{1,\cdots,K-1\}}$ de pouvoir dégager des sous-modèles traduisant des hypothèses d'indépendances. Notons que cette paramétrisation en facteurs peut se généraliser à un nombre quelconques de facteurs et est couramment utilisée par le logiciel R.

2.5.5 Estimation des modèles

Dans cette section, toutes les variables explicatives (qualitative, quantitative et la constante) sont regroupé dans une variable générique X.

Estimation par maximum de vraisemblance

Régression logistique On note $\pi_i = \pi(X_i;\beta)$ la probabilité conditionnelle de l'observation i , on aura alors

$$L(X_1, Y_1, \cdots, X_n Y_n; \beta) = \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i}$$

d'où la log-vraisemblance

$$\ln(L(X_1, Y_1, \cdots, X_n Y_n; \beta)) = \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)$$

Il est alors aisé de calculer la dérivée de cette fonction paramétrique par rapport à chaque composante de β_k.

On montre que si la fonction de lien et log-concave alors il existe généralement un unique maximum $\hat{\beta}$ à la log-vraisemblance. Si le modèle est régulier (i.e. la matrice de Fisher du modèle est inversible) alors on a les résultats asymptotiques classiques de statistique paramétrique :

— Consistance du paramètre :$\hat{\beta} \to \beta_0$
— Normalité asymptotique : $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to \mathcal{N}\left(0, I(\beta_0)^{-1}\right)$
— Test du rapport de vraisemblance, de Wald etc...

Par exemple, si on test on modèle $\mathcal{M}_1$ de dimension p_1, contre un modèle $\mathcal{M}_2$ de dimension $p_2 > p_1$, on aura la statistique :

$$2(l_n(\mathcal{M}_2) - l_n(\mathcal{M}_1)) \overset{Loi}{\to} \chi^2_{p_2-p_1}$$

Modèle polytomique Pour simplifier les notations on note β le vecteur regroupant tous les paramètres du modèle. On peut alors écrire$\pi_k(X_i) := \pi_k(X_i;\beta)$, on aura alors

$$L(X_1, Y_1, \cdots, X_n Y_n; \beta) = \prod_{i=1}^{n} \prod_{k=0}^{K-1} \pi_k(X_i;\beta)^{1_{\{k\}}(Y_i)}(X_i)$$

d'où la log-vraisemblance

$$\ln(L(X_1, Y_1, \cdots, X_n Y_n; \beta)) = \sum_{i=1}^{n} \sum_{k=0}^{K-1} 1_{\{k\}}(Y_i) \ln(\pi_k(X_i;\beta))$$

Il est alors aisé de calculer la dérivée de cette fonction paramétrique par rapport à chaque composantes de β.

Si le modèle est identifiable (un seul vrai paramètre) et régulier (i.e. la matrice de Fisher du modèle est inversible) alors on a les résultats asymptotiques classiques de statistique paramétrique :

— Consistance du paramètre :$\hat{\beta} \to \beta_0$
— Normalité asymptotique : $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to \mathcal{N}\left(0, I^{-1}\right)$
— Test du rapport de vraisemblance, de Wald etc...

Si on test on modèle $\mathcal{M}_1$ de dimension p_1, contre un modèle $\mathcal{M}_2$ de dimension $p_2 > p_1$, on aura la statistique :

$$2\left(l_n\left(\mathcal{M}_2\right) - l_n\left(\mathcal{M}_1\right)\right) \overset{Loi}{\rightarrow} \chi^2_{p_2-p_1}$$

Chapitre 3

Analyse en composantes principales

L'analyse en composantes principales est une méthode puissante pour explorer la structure des données. C'est également la "mère" de la plupart des méthodes descriptive multidimensionnelles.

3.1 Tableau de données.

L'utilisateur de l'analyse en composante principale se trouve dans la situation suivante : il possède un tableau $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ rectangulaire de mesures, dont les colonnes figurent des variables à valeur numériques continues et dont les lignes représentent les individus sur lesquels ces variables sont mesurées. Une variable $\mathbf{x}^j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$ est la liste des n valeurs qu'elle prend sur les n individus. De même on identifiera l'individu i au vecteur $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$.

3.1.1 Poids et centre de gravité

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales les n individus ont tous la même importance et chacun aura pour poids $\frac{1}{n}$ dans les calculs. Il n'en est pas toujours ainsi et il est peut être utile de travailler avec des poids p_i éventuellement différents d'un individu à l'autre. Ces poids, qui sont des nombres positifs de somme 1 comparables à des fréquences, sont

regroupés, dans une matrice diagonale D de taille n :

$$D = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_n \end{bmatrix}$$

Dans le cas le plus usuel de poids égaux $D = \frac{1}{n} I$.

Le vecteur $\mathbf{g} = X^T D\mathbf{1}$ où $\mathbf{1}$ est le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1 est tel que $\mathbf{g}^T = \left(\bar{x}^1, \cdots, \bar{x}^p\right)$ avec $\bar{x}^j = \sum_{i=1}^n p_i x_{ij}$. $\mathbf{g}$ est le centre de gravité du nuage. Le tableau Y tel que $y_{ij} = x_{ij} - \bar{x}^j$ est le tableau centré associé à X. On aura ainsi

$$Y = X - \mathbf{1}\mathbf{g}^T$$

3.1.2 Matrice de variance-covariance et matrice de corrélation

On aura la matrice de covariance qui vaudra :

$$V = Y^T DY = X^T DX - gg^T$$

De plus, si on note T la matrice diagonale des inverses des écarts-types des variables :

$$T = \begin{bmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{s_p} \end{bmatrix}$$

avec $s_j = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(x_{ij} - \bar{x}^j\right)^2}$, on aura le tableau des données centrées et réduites :$Z = YT$, soit $Z = (z_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ avec $z_{ij} = \frac{\left(x_{ij} - \bar{x}^j\right)}{s_j}$. La matrice des corrélations linéaires sera alors :

$$R = Z^T DZ = TVT.$$

3.1.3 Données actives et supplémentaires

Le tableau X ne représente souvent qu'une partie de l'information disponible, les variables disponiblent se partagent en deux ensembles : Les variables actives qui serviront au calcul des axes principaux et les variables supplémentaires qui seront reliées à postériori aux résultats de l'ACP. On peut également n'utiliser qu'une partie des individus, soit pour valider les résultats, soit parce qu'ils n'auront leur données disponible qu'ultèrieurement. Mettre des individus en supplémentaire revient à leur attribuer un poids nul.

3.2 L'espace des individus

chaque individu est un point de $\mathbb{R}^p$, l'ensemble des individus est alors un "nuage" de point dans $\mathbb{R}^p$ et $\mathbf{g}$en est le centre de gravité.

3.2.1 Le rôle de la métrique

La distance la plus classique entre les individus est la distance euclidienne :

$$d^2(\mathbf{x}_l, \mathbf{x}_k) = \sum_{j=1}^{p} (x_{lj} - x_{kj})^2$$

Cette distance n'est pas forcément la meilleure en statistique, surtout qu'elle dépend des unités choisies pour les variables. On utilisera donc la formulation générale :

$$d^2(\mathbf{x}_l, \mathbf{x}_k) = (\mathbf{x}_l - \mathbf{x}_k)^T M (\mathbf{x}_l - \mathbf{x}_k)$$

où M est une matrice symétrique de taille p définie positive. En théorie le choix de la matrice M dépend de l'utilisateur, en pratique les métriques les plus utilisées sont $M = I$, ce qui revient à utiliser la matrice de variance-covariance des individus ou $M = T^2$ ce qui revient à utiliser la matrice des corrélations. Ainsi

$$d^2(\mathbf{x}_l, \mathbf{x}_k) = (T(\mathbf{x}_l - \mathbf{x}_k))^T T(\mathbf{x}_l - \mathbf{x}_k) = (\mathbf{z}_l - \mathbf{z}_k)^T (\mathbf{z}_l - \mathbf{z}_k)$$

avec $\mathbf{z} = T\mathbf{x}$, ainsi travailler avec la métrique T^2 revient à travailler avec la distance euclidienne classique sur les données transformées $\mathbf{z}$.

3.2.2 L'inertie

On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des points au centre de gravité :

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i (\mathbf{x}_i - \mathbf{g})^T M (\mathbf{x}_i - \mathbf{g}) = trace\left(DYMY^T\right)$$

Ainsi si $M = I$, l'inertie est égale à la somme des variances. Si $M = T^2$, on aura

$$trace\left(DYMY^T\right) = trace\left(TVT\right) = trace\left(T^2V\right)$$

l'inertie est égale à p. On remarquera aussi que maximiser la trace de TVT est équivalent à maximiser la trace de T^2V.

3.3 l'espace des variables

Chaque variable $\mathbf{x}^j$ est un point de $\mathbb{R}^n$, on peut donc observer un nuage de p points-variables.

3.3.1 La métrique des poids

Pour étudier la proximité des variables entre elles, il faut munir cet espace d'une métrique, c'est-à-dire trouver une matrice d'ordre n définie positive. Ici la matrice D s'impose d'elle même :

— Le produit scalaire de deux variables centrée x^jet x^k vaudra

$$\left(x^j\right)^T Dx^k = \sum_{i=1}^{n} p_i x_i^j x_i^k = Cov\left(x^j, x^k\right) := s_{jk}$$

— Le norme d'une variable est alors $\|x^j\|_D = \sqrt{s_j^2}$. La longueur d'une variable est son écart-type.

— L'angle θ_{jk} entre deux variables centrées est donné par :

$$\cos\theta_{jk} = \frac{\left\langle x^j, x^k \right\rangle_D}{\|x^j\|_D \|x^k\|_D} = \frac{s_{jk}}{s_j s_k}$$

Le cosinus de l'angle entre deux variables centrées n'est autre que leur coefficient de corrélation linéaire.

3.3.2 Droites d'ajustement

Considérons un axe Δ de l'espace des individus engendré par un vecteur unitaire a, i.e. $a^T a = 1$ et projetons les individus sur cet axe (projection orthogonale). La liste des coordonnées c_i des individus sur Δ forme une nouvelle variable ou composante $\mathbf{c}$. Ainsi $c_i = a^T \mathbf{x}_i = \mathbf{x}_i^T a$.

$$\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_3 \end{pmatrix} = Xa = \sum_{j=1}^{p} x^j a_j$$

A la variable $\mathbf{c}$ sont donc associés trois êtres mathématiques :

— un axe Δ de vecteur unitaire a.

— un vecteur $\mathbf{c}$ de l'espace des variables.

L'ensemble des variables $\mathbf{c}$ que l'on peut engendrer par combinaison linéaire des vecteurs-colonnes de X forme un sous-espace vectoriel de l'espace des variables de dimension égale ou inférieure à p.

La variance de $\mathbf{c}$ sera $V(\mathbf{c}) = (Ya)^T D(Ya) = a^T Y^T DYa = a^T Va$.

3.4 L'analyse

3.4.1 Meilleure projection sur une droite

Commençons par chercher un sous-espace de dimension 1 qui réalise le meilleur ajustement possible du nuage. Pour maximiser $V(\mathbf{c})$ la variance de la projection. Il suffit de maximiser $a^T Va = \langle a, Va \rangle$. Ainsi, si on considère la

base des vecteurs propres orthonormés, a sera donc le vecteur propre associer à la plus grande valeur propre de V : λ_1. Plus généralement, le sous-espace à q dimensions qui ajuste au mieux le nuage de points dans $\mathbb{R}^p$ est engendré par les q premiers vecteurs propres de la matrice V.

3.4.2 Facteurs principaux

A l'axe principale a,est associé la forme linéaire $u =< , a >$, normé, qui définit une combinaison linéaire des variables $\mathbf{x}^1, \cdots, \mathbf{x}^p$, puisque a est vecteur propre de V, $Va = \lambda a$. La forme linéaire u peut donc être identifié au vecteur propre de V, normé, on l'appelle facteur principal. On remarquera que si a est de norme $1u$ est de norme 1.

3.4.3 Composante principale

Ce sont les variables $\mathbf{c}_i = X\mathbf{u}_i$. $\mathbf{c}_i$ est le vecteur renfermant les coordonnées des projections M-orthogonales des individus sur l'axe défini par a_i avec a_i unitaire. La variance d'une composante principale sera égale à la valeur propre

$$V(\mathbf{c}_i) = \lambda_i$$

En effet $V(\mathbf{c}) = \mathbf{c}^T D\mathbf{c} = u^T X^T DXu = u^T Vu$ où $Vu = \lambda u$, donc $V(c) = \lambda u^T u = \lambda$.

Les c_i sont les combinaisons linéaires de $x_1, x_2, \cdots, x_p$ de variance maximale sous la contrainte $u_i^T u = 1$. Les composantes principales sont elles-mêmes vecteurs propres d'une matrice de taille n : En effet $Vu = \lambda u \Leftrightarrow X^T DXu = \lambda u$ en multipliant à gauche par X et en remplaçant Xu par c on a :

$$XX^T D\mathbf{c} = \lambda \mathbf{c}$$

Pour résumer :

Facteurs Principaux u	$Vu = \lambda u$	orthonormés
Axes principaux a	$Va = \lambda a$	orthonormés
Composantes principales $c = Xu$	$XMX^T D\mathbf{c} = \lambda \mathbf{c}$	D-orthogonales

3.4.4 Reconstitution des données de départ

Comme $X\mathbf{u}_j = \mathbf{c}_j$, en post-multipliant les deux membres par $\mathbf{u}_j^T$ et en sommant sur j, il vient :

$$X \sum_j \mathbf{u}_j \mathbf{u}_j^T = \sum_j \mathbf{c}_j \mathbf{u}_j^T$$

or $\sum_j \mathbf{u}_j \mathbf{u}_j^T = I$ car les $\mathbf{u}_j$ sont orthonormées, il suffit en effet de vérifier que :

$$\left(\sum_j \mathbf{u}_j \mathbf{u}_j^T \right) \mathbf{u}_i = \mathbf{u}_i \text{ car } \mathbf{u}_j^T \mathbf{u}_i = \delta_{ij}.$$

Ainsi

$$X = \sum_j \mathbf{c}_j \mathbf{u}_j^T$$

Si l'on se contente de la somme des k premiers termes, on obtient alors la meilleure approximation de X par une matrice de rang k au sens des moindres carrés (théorème d'Eckart-Young). On aura également

$$V = \sum_{j=1}^{p} \lambda_j u_j u_j^T$$

et

$$V = \sum_{j=1}^{p} \lambda_j a_j a_j^T$$

$X = \sum_{j=1}^{p} c_j u_j^T = \sum_{j=1}^{p} \sqrt{\lambda_j} z_j v_j^T$ où les z_j sont les vecteurs propres de norme 1 de XX^T et les v_j les vecteurs propres de X^TX de norme 1.

3.5 Interprétation des résultats

3.5.1 Qualité de l'approximation

La qualité de la reconstitution peut être évaluée par le pourcentage d'inertie totale expliquée : la quantité :

$$\tau_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha} = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{I_{\mathbf{g}}}$$

Le coefficient τ_q inférieur ou égal à 1 sera appelé taux d'inertie relatif aux q premiers facteur. Si par exemple $\frac{\lambda_1+\lambda_2}{I_{\mathbf{g}}} = 0.9$, on concoit clairement que le nuage de points est presque aplati sur un sous-espace de dimension 2 et que la projection du nuage sur le premier plan principale est très satisfaisante.

L'appréciation du pourcentage d'inertie doit faire intervenir le nombre de variables initiales, 10% d'inertie expliquée n'a pas le même intérêt sur un tableau de 20 variables que sur un tableau de 100 variables.

3.5.2 Mesure locale

Pour qu'un individu soit bien représenté sur le plan factoriel, il faut que le cosinus de ce point-individu avec le plan factoriel soit assez grand. Si on note $\left\|P_\alpha^M(\mathbf{x}_i)\right\|_D$ la longueur de la projection de l'individu i sur l'axe α et $\|\mathbf{x}_i\|_D$ sa longueur initiale, on aura $\cos(\theta_{\alpha i}) := \frac{\left\|P_\alpha^M(\mathbf{x}_i)\right\|_D}{\|\mathbf{x}_i\|_D}$. Il est plus pratique de regarder les $\cos^2$, car par le théorème de pythagore, on aura la longueur au carré de la projection $P_{\alpha_1,\alpha_2}^M(\mathbf{x}_i)$ sur le plan engendré par les facteurs α_1 et α_2 qui sera donné par :

$$\left\|P_{\alpha_1,\alpha_2}^M(\mathbf{x}_i)\right\|_D^2 = \left\|P_{\alpha_1}^M(\mathbf{x}_i)\right\|_D^2 + \left\|P_{\alpha_2}^M(\mathbf{x}_i)\right\|_D^2$$

3.5.3 Choix de la dimension

Le principale intérêt de l'ACP consistant à réduire la dimension des l'espace des individus, le choix du nombre d'axes à retenir est un point essentiel qui n'a pas de solution rigoureuse. Le critère le plus connu est celui de Kaiser : On ne retient que les valeurs propres supérieurs à 1, cela suppose évidemment qu'on travaille avec la matrice des corrélations.

3.5.4 Analyse du nuage des points-variables

La méthode la plus naturelle pour donner une signification à une composante principale $\mathbf{c}$ est de la relier aux variables initiales $\mathbf{x}^j$, en calculant les coefficients de corrélation linéaire $r\left(\mathbf{c},\mathbf{x}^j\right)=\frac{Cov\left(\mathbf{c},\mathbf{x}^j\right)}{V(\mathbf{c})V(\mathbf{x}^j)}$. Lorsqu'on choisit la métrique $M=D_{\frac{1}{s^2}}$ ce qui revient à travailler avec la matrice des corrélations, le calcul de $r\left(\mathbf{c},\mathbf{x}^j\right)$ est simple :

$$r\left(\mathbf{c},\mathbf{x}^j\right)=r\left(\mathbf{c},\mathbf{z}^j\right)=\frac{\mathbf{c}^T D\mathbf{z}^j}{s_c}$$

et comme $V\left(\mathbf{c}\right)=\lambda$

$$r\left(\mathbf{c},\mathbf{x}^j\right)=\frac{\mathbf{c}^T D\mathbf{z}^j}{\sqrt{\lambda}}$$

et comme $\mathbf{c}=Z\mathbf{u}$ où $\mathbf{u}$ est facteur principal associé à $\mathbf{c}$, vecteur propre de la matrice des corrélations R associé à la valeur propre λ :

$$r\left(\mathbf{c},\mathbf{x}^j\right)=\frac{\mathbf{u}^T Z^T D\mathbf{z}^j}{\sqrt{\lambda}}=\frac{\left(\mathbf{z}^j\right)^T DZ\mathbf{u}}{\sqrt{\lambda}}$$

Or, $\left(\mathbf{z}^j\right)^T DZ$ est la $j$ième ligne de $Z^T DZ=R$ donc $\left(\mathbf{z}^j\right)^T DZ\mathbf{u}$ est la $j$ième composante de $R\mathbf{u}$ et comme $R\mathbf{u}=\lambda\mathbf{u}$,il vient

$$r\left(\mathbf{c},\mathbf{x}^j\right)=\sqrt{\lambda}\mathbf{u}_j$$

Ces calculs s'effectuent pour chaque composante principale. Pour un couple de composante principale $\mathbf{c}^1$ et $\mathbf{c}^2$, on synthétise usuellement les corrélations sur une figure appelée "cercle des corrélations", où chaque $\mathbf{x}^j$ est repérée par un point d'abscisse $r\left(\mathbf{c}^1,\mathbf{x}^j\right)$ et d'ordonnée $r\left(\mathbf{c}^2,\mathbf{x}^j\right)$.

3.5.5 Place et importance des individus

Si les individus ne sont pas anonymes, ils aident à l'interprétation des composantes principales : On recherchera par exemple les individus opposés le long d'un axe.

De plus, pour la composante $\mathbf{c}_k$, on a

$$\sum_{i=1}^{n} p_i c_{ki}^2=\lambda_k$$

On définie ainsi la contribution de l'individu i à la composante $\mathbf{c}_k$ par :

$$\frac{p_i c_{ki}^2}{\lambda_k}$$

Les considérations des contributions aident à l'interprétation des axes. Cependant, il n'est pas souhaitable qu'un individu ait une contribution excessive car cela serait un facteur d'instabilité, le fait de retirer cet individu modifiant profondément le résultat de l'analyse. Lorsque les poids des individus sont tous égaux $\frac{1}{n}$, les contributions n'apportent pas plus d'information que les coordonnées.

3.5.6 Effet "taille"

Lorsque toutes les variables $\mathbf{x}^j$ sont corrélés positivement entre elles et avec la première composante principale, les individus seront rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne). On parle alors d'effet de taille.

La deuxième composante principale différencie alors des individus de "taille" semblable, on l'appelle facteur de "forme".

3.5.7 Individus et variables supplémentaires

Les interprétations fondées sur les remarques précédentes présente le défaut d'être tautologique : on explique les résultats à l'aide des données qui ont servi à les obtenir. Par exemple, il n'est pas étonnant par exemple de trouver de fortes corrélations entre la première composante principale $\mathbf{c}^1$ et certaines variables puisque $\mathbf{c}^1$ maximise :

$$\sum_{j=1}^{p} r^2\left(\mathbf{c}, \mathbf{x}^j\right)$$

On n'est donc pas sûr d'avoir découvert un phénomène significatif. Par contre, si l'on trouve une forte corrélation entre une composante principale et une variable qui n'a pas servi à l'analyse, le caractère probant de ce phénomène sera bien plus élevé. Ainsi, une pratique courante consiste à partager en deux groupes l'ensemble des variables : d'une part les variables "actives" qui servent à déterminer les axes principaux, d'autre part les variables "supplémentaires" que l'on relie à postériori aux composantes principales. Les variables numérique supplémentaires peuvent être placées dans les cercles de corrélation : Il suffit de calculer le coefficient de corrélation entre ces variables et les composantes principales.

Une variable qualitative supplémentaire correspond à la donnée d'une partition des n individus en k catégories. En général on se contente de représenter chaque catégories par son centre de gravité. On peut aussi calculer ce qu'on appelle la "valeur test" associée à chaque modalité, qui mesure sur chaque axe la différence divisée par l'écart-type correspondant on raisonnement suivant : Si les n_i individus de la catégories i étudiée avaient été tirés au hasard avec probabilité égales, parmi les n étudié, la moyenne de leur coordonnées serait une variable aléatoire d'espérance la projection du centre de gravité sur l'axe

k et de variance $\frac{\lambda_k}{n_i}\frac{n-n_i}{n-1}$ car le tirage est sans remise. Si le nuage est centré la valeur test de la coordonnée a_{ik} du centre de gravité est alors :

$$\frac{a_{ik}}{\sqrt{\frac{\lambda_k}{n_i}\left(\frac{n-n_i}{n-1}\right)}}$$

En faisant une approximation gaussienne, on décidera qu'une modalité occupe une position significativement différente de la moyenne si, en valeur absolue, sa valeur test dépasse 2.

La coordonnée du nouveau point-individu $\mathbf{x}^{sup}$ sur la composante $\mathbf{u}_\alpha$ sera simplement $\left(\mathbf{x}^{sup}\right)^T \mathbf{u}_\alpha$.

3.5.8 Représentation simultanée

Les deux nuages ne sont pas dans le même repère, ce qui rend impossible la représentation simultanée des individus et des variables. Les proximités entre individus s'interprètent en termes de similitudes de comportement vis-à-vis des variables et les proximité entre variables en termes de corrélations.

Si on considère non pas les points-variables mais des directions de variables dans $\mathbb{R}^p$, on peut envisager de représenter simultanément à la fois des points individus et des vecteurs représentant les variables. Dans l'espace $\mathbb{R}^p$ des n points-individus, après transformation du tableau de données, on dispose de deux systèmes d'axes :

— Les anciens axes unitaires $(e_1, \cdots, e_p)$ correspondant aux p variables avant l'analyse.

— Les nouveaux axes unitaires (factoriels) $(u_1, \cdots, u_p)$

La possibilité d'une représentation simultanée réside alors dans la projection de l'ancien axe e_j sur le nouvel axe $u_\alpha$grâce au cercle des corrélations.

3.6 Un exemple d'ACP : Dépense des ménages.

3.6.1 Les données

Le jeu de données est issue d'une enquête budget des familles 2006 de l'INSEE. Ces enquêtes permettent de connaître le poids des grands postes de consommation dans le budjet des ménages. Pour obtenir une typologie des classes d'âges fondée sur leurs dépenses, on définit les distances entre deux classes d'âge uniquement sur la base de leurs dépenses dans les différentes rubriques. En ce qui concerne les individus, les classes d'âge sont des individus actifs et les déciles du revenu sont considérés comme des individus supplémentaires. On décide de centré-normé les variables.

3.6.2 Mise en oeuvre de l'analyse

Pour effectuer l'analyse, on utilise la fonction PCA de la librairie R FactoMineR.

```
res.pca<-PCA(don,ind.sup=8:18,quanti.sup=27:30)
```

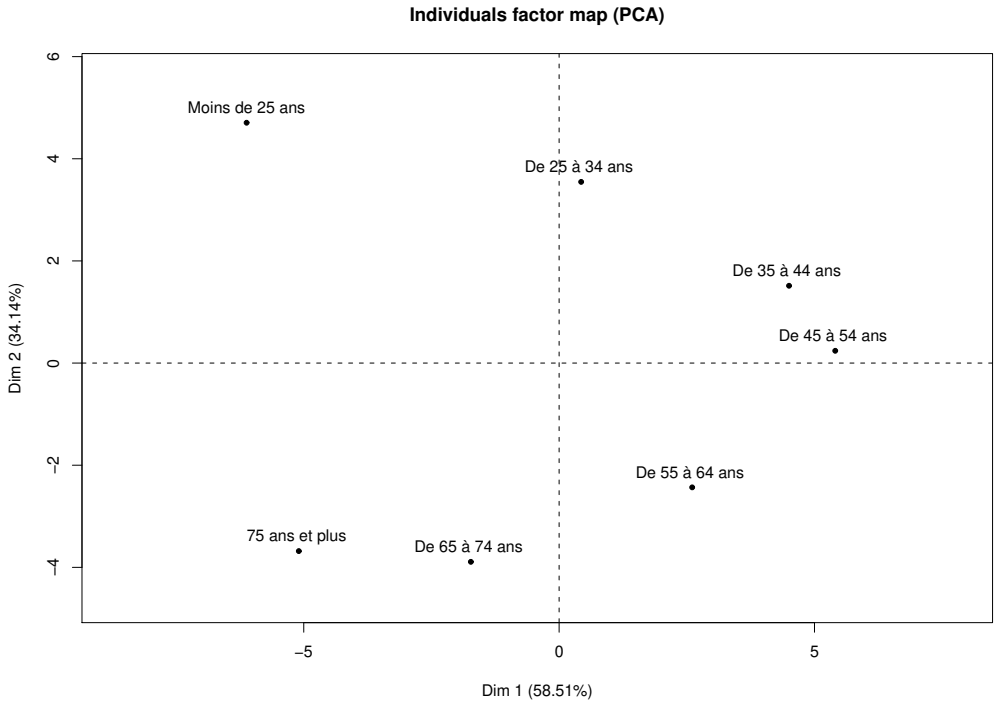
La fonction PCA fournit le graphe des individus et le graphes des variables. L'inertie des trois premiers axes factoriels est la suivante :

$\lambda_1 = 15.52$, $\lambda_2 = 8.67$, $\lambda_3 = 1.22$, ce qui représente 97.71%.

3.6.3 Etude sur le plan engendré par les deux premiers axes

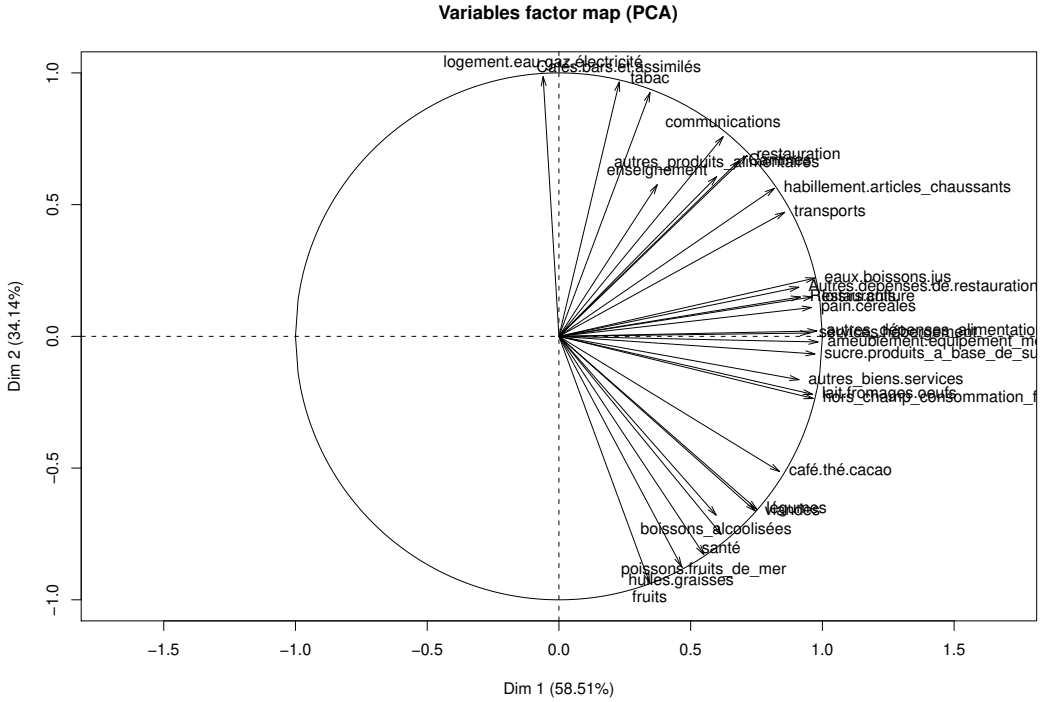
Etude du nuage des individus actifs. On peut construire le graphe des individus actifs en sélectionnant (choix="ind") et (invisible="ind.sup").

Ce graphe des individus présente une disposition remarquable : le premier axe oppose les tranches d'âge extrêmes aux tranches d'âge moyennes. le second range les tranches d'âge de la plus haute à la plus basse.

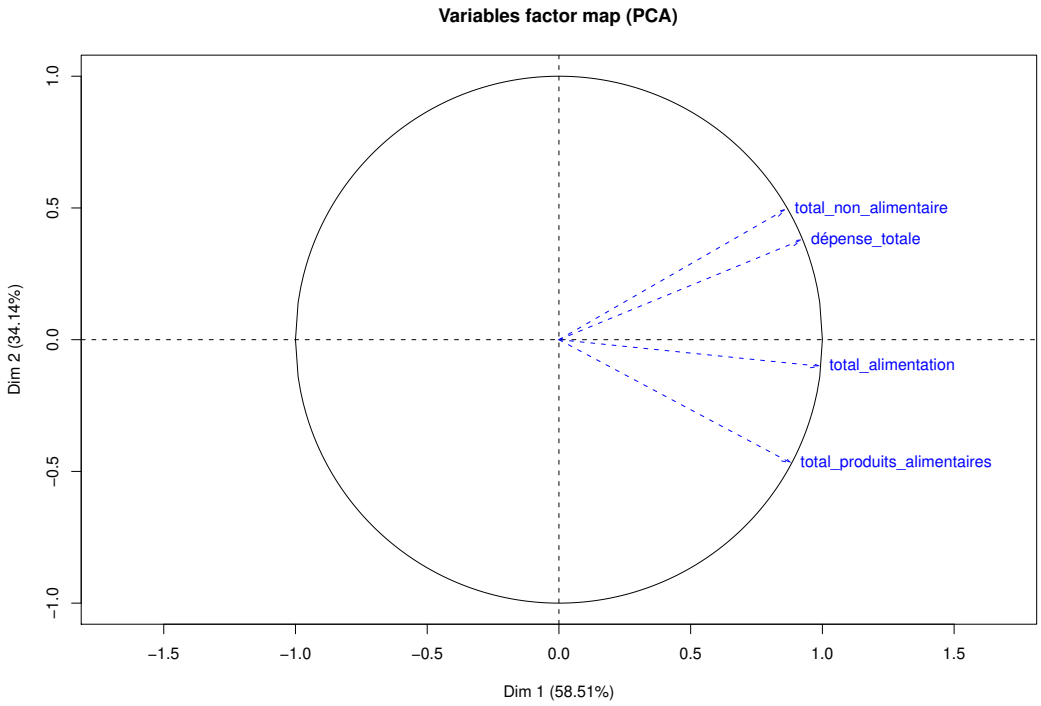


Etude du nuage des variables Les représentations du nuage des variables permettent de visualiser rapidement les corrélations entre les variables. On peut construire le graphes des variables actives à l'aide de la fonction plot.PCA : (choix="var") (invisible="quanti.sup"). Le graphe met en évidence un corrélation positive entre la première composante et toute les variables sauf une (logement eau gaz etc...) Cet axe oppose donc les tranches d'âge qui consomment peu (celle qui ont des coordonnées négatives sur cet axe) à celle qui consomment

beaucoup. Toutes les variables sont bien représentée sur le premier plan factoriel à l'exception de la variable enseignement. On lit la qualité de représentation sur le graphe grâce à la proximité de l'extrèmité de la flèche et le cercle de rayon 1.



Les variables étant bien représentée, il en est de même de l'angle entre deux variables donc de la corrélation entre les variables. Ainsi (pain et céréales) est très corrélée à (lait,frimage,oeufs) . Logement,eau,gaz est orthogonale (donc indépendante) de ces deux variables. Ici les variables supplémentaires (des totaux qui résument plusieurs variables) sont utiles pour simplifier la lecture du graphe.



Description automatique des dimensions Il est possible d'obtenir une description automatique des axes factoriels à l'aide de la fonction "dimdesc" :

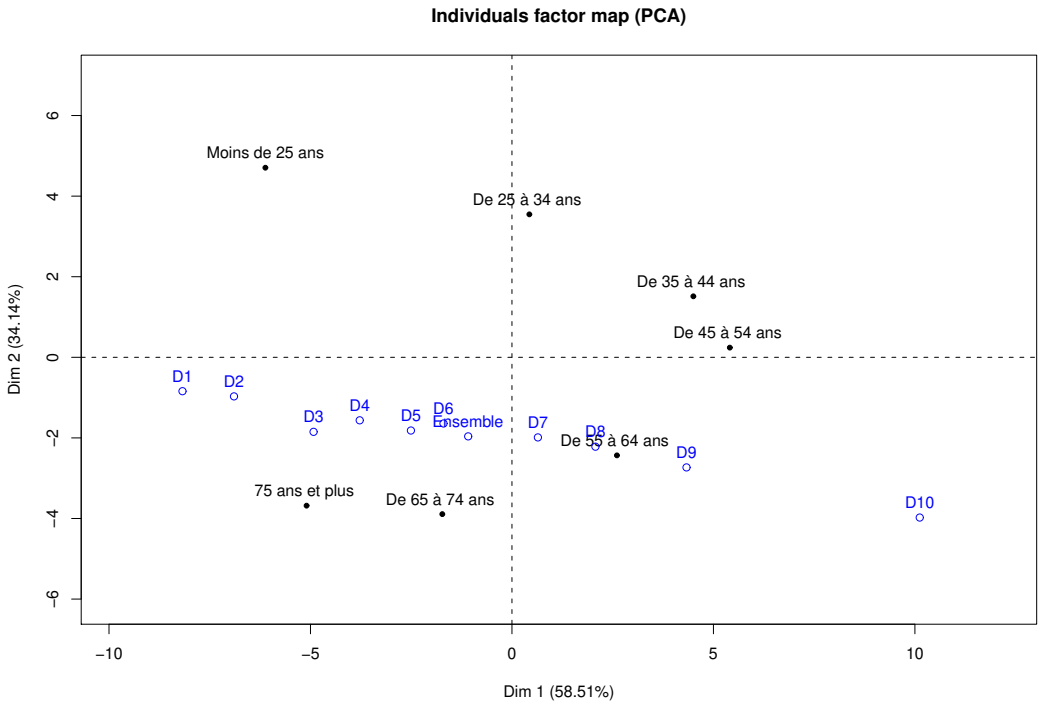
dimdesc(res.pca)

On voit que le premier axe est très lié à la variable (total alimentation) (corrélation : 0.99),

ainsi une grande part des dépenses qui différencient les classes d'âge peut être résumé par la seule variable (total alimentation).

le deuxième axe est très lié à (logement eau gaz) (café bars) et (tabac) (corrélation 0.98,0.97 et 0.93)(qu'il oppose aux dépenses (fruits, huiles)(corrélation -0.87 et -0.92). Cet axe sépare entre elles les classes qui dépense relativement peu (les classes d'axe extrêmes). Les moins de 25 ans dépense plus que la moyenne en communication, logement restauration et moins que la moyenne en poisson fruit graisse. Les personnes agées présente un profil opposé.

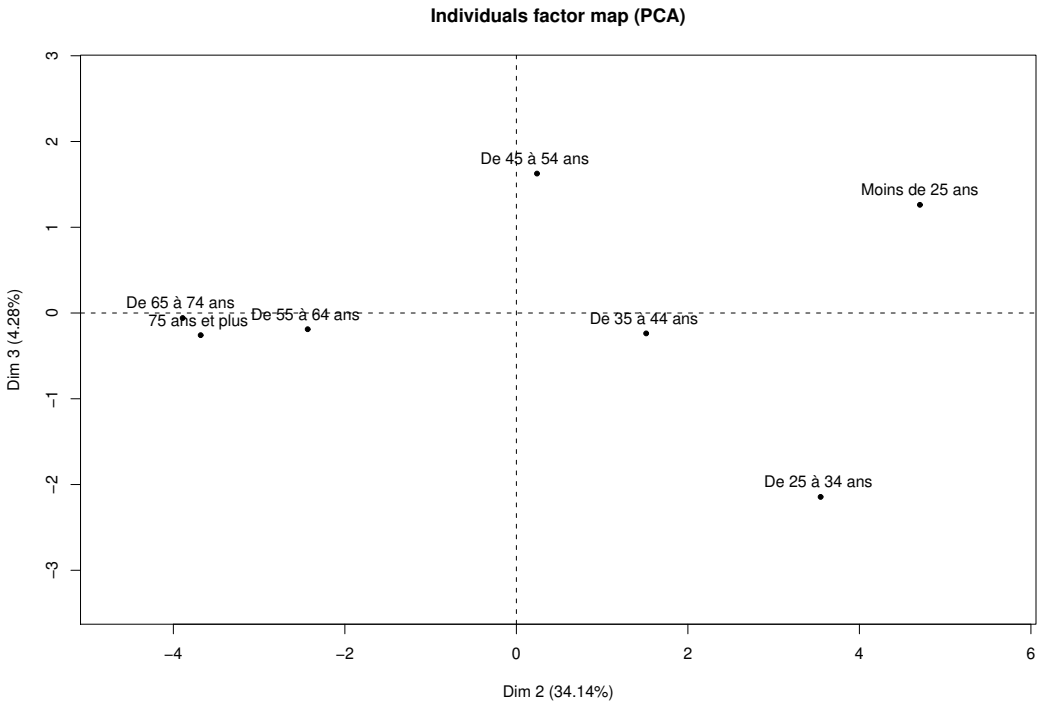
Individus supplémentaires Nous étudions le profil des dépenses en fonction de l'age, mais la variable "revenue" joue un rôle majeur , ainsi tout au long de l'axe 1 les déciles de revenus sont rangés de l'ordre croissant, cet axe oppose donc les pauvres au riches. Il y a un saut élévé entre le 9ème et le 10ème décile qui correspond à un plus grand écart entre les revenus.

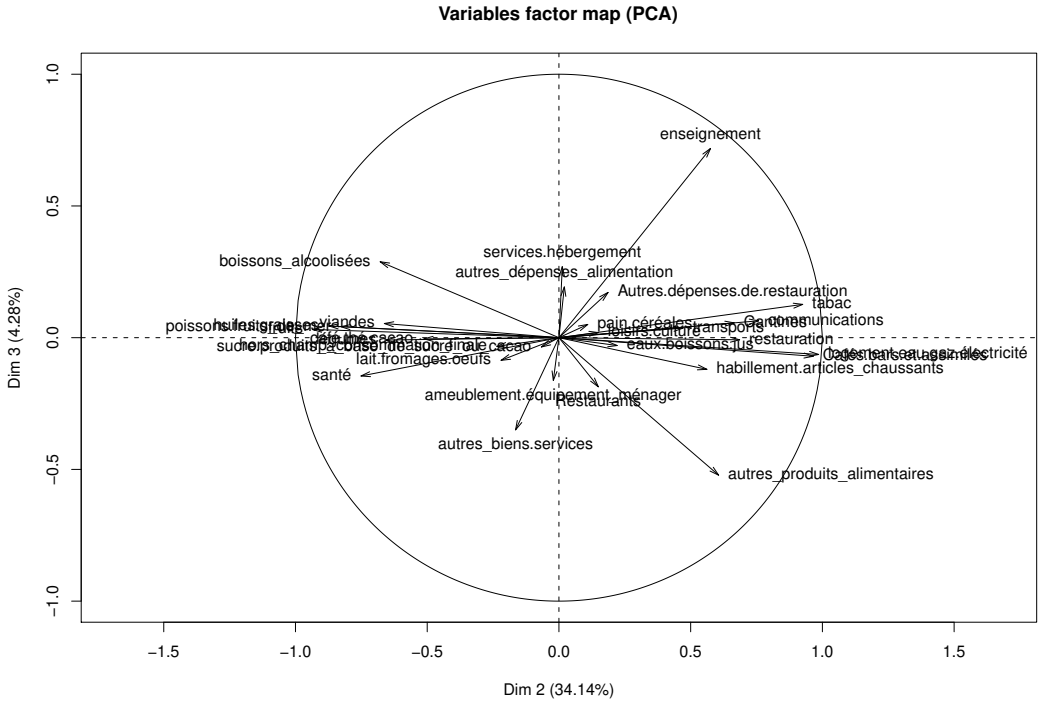


3.6.4 Plan engendré par les axes 2 et 3

On peut construire les graphes des variables et des individus sur le plan engendré par les axes 2 et 3 par :

```
plot(res.pca,choix=''ind'', axes=2:3)
plot(res.pca,choix=''var'', axes=2:3)
```


Individuals factor map (PCA)
De 45 à 54 ans
Moins de 25 ans
De 65 à 74 ans
75 ans et plus
De 55 à 64 ans
De 35 à 44 ans
De 25 à 34 ans
Dim 3 (4.28%)
Dim 2 (34.14%)
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
6



Analyse conjointe du nuage des individus et des variables L'axe 3 est essentiellement lié à la variable "enseignement" (corrélation positive) et à un moindre degré à la variable (autres produits alimentaires) (corrélation négative). Il oppose principalement la tranche d'age 25-34 aux tranches moins de 25 et 45-54. Les classes moins de 25 et 45-54 dépensent plus que les autres classes. On peut supposer que ce sont soit des étudiants, soit les parents d'étudiants ou d'enfants scolarisés.

Le pourcentage d'inertie expliqué par l'axe 3 est petit, mais celui-ci reste facilement intepprétable. Le critère ultime du choix de garder un axe ou pas reste donc sa faculté d'interprétation.

Chapitre 4

Analyse des correspondances

L'analyse des correspondance est une méthode adaptée au tableaux de contingence et permet d'étudier les éventuelles relations existant entre deux variables nominales.

4.1 Tableaux de contigence

Ce tableau est obtenu en ventilant une population selon deux variables nominales. L'ensemble des colonnes du tableau désigne les modalités d'une variable et l'ensemble des lignes correspond à celle de l'autre variable. Considérons le tableau T de contigence obtenu en ventilant une population de 592 femmes suivant leurs couleurs des yeux et cheveux :

		Couleur	des	cheveux		
		brun	châtain	roux	blond	total
Couleur	marron	68	119	26	7	220
des	noisette	15	54	14	10	93
yeux	vert	5	29	14	16	64
	bleu	20	84	17	94	215
	total	108	286	71	127	592

En ligne est représenté la variable "Couleur des yeux" à $n = 4$ modalités et en colonne la variable "couleur des cheveux" à $p = 4$ modalité. A l'intersection d'une ligne et d'une colonne, nous avons le nombre k_{ij} de femmes ayant simultanément la couleur i des yeux et la couleur j de cheveux. Le total marginal $k_{i.}$ est le nombre de femmes ayant les yeux de couleur i, alors que le total marginal $k_{.j}$ est le nombre de femmes ayant les cheveux de couleur j. On a les relations suivantes :

$$k_{i.} = \sum_{j=1}^{p} k_{ij},\ k_{.j} = \sum_{i=1}^{n} k_{ij},\ k = \sum_{i,j} k_{ij}$$

qui, en terme de fréquences relatives, donnent lieu aux relations :

$$f_{ij} = \frac{k_{ij}}{k},\ f_{i.} = \sum_{j=1}^{p} f_{ij},\ f_{.j} = \sum_{i=1}^{n} f_{ij},\ \sum_{i,j} f_{ij} = 1$$

Transformation du tableau de contingence Pour analyser un tableau de contingence, ce n'est pas le tableau d'effectifs bruts qui nous intéresse mais les tableaux des profils-lignes et celui des profils-colonnes, c'est-à-dire les répartitions en pourcentage à l'intérieur d'une ligne ou d'une colonne.

On note les profils-lignes :

$$\frac{f_{ij}}{f_{i.}} = \frac{k_{ij}}{k_{i.}}$$

et les profils colonnes

$$\frac{f_{ij}}{f_{.j}} = \frac{k_{ij}}{k_{.j}}$$

On obtient alors le tableau des profils-lignes :

		Couleur	des	cheveux		
		brun	châtain	roux	blond	total
Couleur	marron	31	54	12	3	100
des	noisette	16	58	15	11	100
yeux	vert	8	45	22	25	100
	bleu	9	39	8	44	100
profil	moyen	18	48	12	22	100

et le tableau des profils-colonnes :

		Couleur	des	cheveux		
		brun	châtain	roux	blond	Profil moyen
Couleur	marron	63	42	37	6	37
des	noisette	14	19	20	8	16
yeux	vert	5	10	20	13	11
	bleu	19	29	24	74	36
	total	100	100	100	100	100

Si on note E_n et E_p les matrices diagonales des effectifs marginaux des deux variables :

$$E_n = \begin{pmatrix} k_{1.} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & k_{n.} \end{pmatrix} \text{ et } E_p = \begin{pmatrix} k_{.1} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & k_{.p} \end{pmatrix}$$

Le tableau de contigence étant noté $T = (k_{ij})_{1 \le i \le n, 1 \le j \le p}$, on aura le tableau des profils lignes :$T_l = E_n^{-1} T$ et le tableau des profils colonnes : $T_c = T E_p^{-1}$.

Pour l'analyse d'un tableau de contingence, nous raisonnerons en terme de profils.

Nuage des n lignes L'ensemble des profils-lignes forme un nuage de n points dans l'espace des p colonnes et représente le nuage des 4 modalités de couleurs des yeux. Chaque point i a pour coordonnées dans $\mathbb{R}^p$:

$$\left(\frac{f_{i1}}{f_{i.}}, \cdots, \frac{f_{ip}}{f_{i.}}\right)$$

il est affecté d'une masse $f_{i.}$ qui est sa fréquence relative. Puisque $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} = 1$, les n points du nuage sont situés dans un sous-espace à $p-1$ dimension. Le centre de gravité de ce nuage est la moyenne des profils-lignes affectés de leur masse et correspond au profil moyen. Sa $j$ième composante vaut :

$$\sum_{i=1}^{n} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$

c'est la fréquence marginale des colonnes. Autrement dit, les profils ligne forment donc un nuage de n points dans R^p, chacun de des points étant affecté d'un poids proportionnel à sa fréquence marginal $\frac{D_n}{k}$. Le centre de gravité de ce nuage est donc

$$g_l = \frac{1}{n} \left(E_n^{-1} T\right)^T E_n \mathbf{1} = \begin{pmatrix} \frac{k_{.1}}{k} \\ \vdots \\ \frac{k_{.p}}{k} \end{pmatrix}$$

avec $k = \sum_{i=1}^{n} k_{i.} = \sum_{j=1}^{p} k_{.j}$ le total du tableau.

Nuage des p colonnes De la même façon, l'ensemble des p profils-colonnes constitue un nuage de p points dans l'espace des n lignes et représente ici le nuage des 4 modalités de couleurs des cheveux. Les coordonnées dans $\mathbb{R}^n$du point j sont données par

$$\left(\frac{f_{1j}}{f_{.j}}, \cdots, \frac{f_{nj}}{f_{.j}}\right)$$

chaque point est affecté de la masse $f_{.j}$. Les p points du nuage sont situés dans un sous-espace à $n-1$ dimensions puisque $\sum_{j=1}^{n} \frac{f_{ij}}{f_{.j}} = 1$. Le centre de gravité du nuage des profils-colonnes est le profil moyen de la couleur des yeux. Sa $i$ième composante vaut :

$$\sum_{j=}^{p} f_{.j} \frac{f_{ij}}{f_{.j}} = f_{i.}$$

c'est la fréquence marginale des lignes. Autrement dit les profils-colonnes forment un nuage de p points dans $\mathbb{R}^n$ avec des poids $\frac{E_p}{k}$, et leur centre de gravité sera

$$g_c = \begin{pmatrix} \frac{k_{1.}}{k} \\ \vdots \\ \frac{k_{n.}}{k} \end{pmatrix}.$$

Hypothèse d'indépendance On s'intéresse aux liens éventuels entre couleur des yeux et des cheveux. il y a indépendance si

$$f_{ij} = f_{i.} \times f_{.j} \Leftrightarrow k_{ij} = n \times f_{i.} \times f_{.j}$$

Le classique test du χ^2 de Pearson permet d'apprécier l'écart entre les lois empiriques f_{ij} et $f_{i.} \times f_{.j}$.

$$\sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(k_{ij} - n \times f_{i.} \times f_{.j})^2}{n \times f_{i.} \times f_{.j}} \overset{H_0}{\sim} \chi^2 ((n-1) \times (p-1))$$

Cette hypothèse s'exprime aussi sur les profils-lignes. En effet, il en découle que, quelque soit j :

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

Si tous les profils "couleurs des yeux" sont identiques entre eux et par conséquent identique au profil moyen. Il y aura indépendance entre la couleur des cheveux et la couleur des yeux puisque la connaissance d'une couleur des cheveux ne change par la répartition de la couleur des yeux. Il en est de même pour les profils-colonnes :

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

Ainsi, examiner les proximités entre les profils revient à examiner la proximité entre chaque profil et le profil moyen, ce qui permet d'étudier la liaison entre deux variables nominale, c'est-à-dire l'écart à l'indépendance. Nous allons voir comment la construction du nuage, le choix du critère d'ajustement et celui de la distance s'imposent par la nature même des données analysées.

4.1.1 Critère d'ajustement

On cherche les proximités (ou les différences) entre les profils et le profil moyen, cela nous amène, comme en analyse en composantes principales dans le cas des points-individus, à considérer le nuage de points centré sur centre de gravité. Pour le calcul de l'ajustement, la quantité à rendre maximale sera donc la somme pondérée des carrés des distances entre les points et le centre de gravité du nuage, en utilisant une distance qu'il reste à définir.

4.1.2 Une généralisation de l'ACP

4.1.2.1 Droites d'ajustement

Considérons un axe Δ de l'espace des individus engendré par un vecteur unitaire a pour la métrique M, i.e. $a^T M a = 1$ et projetons les individus sur cet axe (projection M-orthogonale). La liste des coordonnées c_i des individus sur

Δ forme une nouvelle variable ou composante $\mathbf{c}$. Ainsi $c_i = a^T M \mathbf{x}_i = \mathbf{x}_i^T M a = \mathbf{x}_i^T u$ en posant $u = Ma \Leftrightarrow M^{-1}u = a$.

$$\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_3 \end{pmatrix} = XMa = Xu = \sum_{j=1}^{p} x^j u_j$$

A la variable $\mathbf{c}$ sont donc associés trois êtres mathématiques :
- un axe Δ de vecteur unitaire a.
- un vecteur $\mathbf{c}$ de l'espace des variables.
- une forme linéaire u appelée facteur.

L'ensemble des variables $\mathbf{c}$ que l'on peut engendrer par combinaison linéaire des vecteurs-colonnes de X forme un sous-espace vectoriel de l'espace des variables de dimension égale ou inférieure à p.

Remarquons que si a appartient à l'espace des individus, u appartient à son dual et que si a est M-normé à 1, u est M^{-1} normé à 1 :

$$a^T Ma = \left(M^{-1}u\right)^T MM^{-1}u = u^T M^{-1}u = 1$$

La variance de $\mathbf{c}$ sera $V(\mathbf{c}) = (Yu)^T D(Yu) = u^T Y^T DYu = u^T Vu = a^T MVMa$.

Commençons par chercher un sous-espace de dimension 1 qui réalise le meilleur ajustement possible du nuage. Pour maximiser $V(\mathbf{c})$ la variance de la projection. Il suffit de maximiser $a^T MVMa = \langle a, VMa \rangle_M$ où $\langle ., . \rangle_M$ est le produit scalaire associé à la métrique M. Ainsi, si on considère la base des vecteurs propres M-orthonormés, a sera donc le vecteur propre associer à la plus grande valeur propre de VM : λ_1. Plus généralement, le sous-espace à q dimensions qui ajuste au mieux le nuage de points dans $\mathbb{R}^p$ est engendré par les q premiers vecteurs propres de la matrice MV.

4.1.2.2 Facteurs principaux

A l'axe principale a, M-normé est associé la forme linéaire $u = Ma$, M^{-1}-normé, qui définit une combinaison linéaire des variables $\mathbf{x}^1, \cdots, \mathbf{x}^p$, puisque a est vecteur propre de VM, $VMa = \lambda a \Rightarrow MVMa = \lambda Ma$ soit $MVu = \lambda u$. La forme linéaire u peut donc être identifié au vecteur propre de MV, M^{-1} normé, on l'appelle facteur principal. On remarquera que si a est de norme 1 pour la mesure M, u est de norme 1, pour la mesure M^{-1}.

4.1.2.3 Composante principale

Ce sont les variables $\mathbf{c}_i = X\mathbf{u}_i$. $\mathbf{c}_i$ est le vecteur renfermant les coordonnées des projections M-orthogonales des individus sur l'axe défini par a_i avec a_i unitaire. La variance d'une composante principale sera égale à la valeur propre

$$V(\mathbf{c}_i) = \lambda_i$$

En effet $V(\mathbf{c}) = \mathbf{c}^T D\mathbf{c} = u^T X^T DXu = u^T Vu$ où $Vu = \lambda M^{-1}u$, donc $V(c) = \lambda u^T M^{-1}u = \lambda$.

Les c_i sont les combinaisons linéaires de $x_1, x_2, \cdots, x_p$ de variance maximale sous la contrainte $u_i^T M^{-1}u = 1$. Les composantes principales sont elles-mêmes vecteurs propres d'une matrice de taille n : En effet $MVu = \lambda u \Leftrightarrow MX^T DXu = \lambda u$ en multipliant à gauche par X et en remplaçant Xu par c on a :

$$XMX^T D\mathbf{c} = \lambda \mathbf{c}$$

Pour résumer :

Facteurs Principaux u	$MVu = \lambda u$	M^{-1}-orthonormés
Axes principaux a	$VMa = \lambda a$	M-orthonormés
Composantes principales c	$XMX^T D\mathbf{c} = \lambda \mathbf{c}$	D-orthogonales
$\mathbf{c} = Xu$	$u = MA$	

4.1.3 Reconstitution des données de départ

Comme $X\mathbf{u}_j = \mathbf{c}_j$, en post-multipliant les deux membres par $\mathbf{u}_j^T M^{-1}$ et en sommant sur j, il vient :

$$X \sum_j \mathbf{u}_j \mathbf{u}_j^T M^{-1} = \sum_j \mathbf{c}_j \mathbf{u}_j^T M^{-1}$$

or $\sum_j \mathbf{u}_j \mathbf{u}_j^T M^{-1} = I$ car les $\mathbf{u}_j$ sont M^{-1}-orthonormées, il suffit en effet de vérifier que :

$$\left(\sum_j \mathbf{u}_j \mathbf{u}_j^T M^{-1} \right) \mathbf{u}_i = \mathbf{u}_i \text{ car } \mathbf{u}_j^T M^{-1} \mathbf{u}_i = \delta_{ij}.$$

Ainsi

$$X = \sum_j \mathbf{c}_j \mathbf{u}_j^T M^{-1}$$

Si l'on se contente de la somme des k premiers termes, on obtient alors la meilleure approximation de X par une matrice de rang k au sens des moindres carrés (théorème d'Eckart-Young). On aura également

$$MV = \sum_{j=1}^{p} \lambda_j u_j u_j^T M^{-1}$$

et

$$VM = \sum_{j=1}^{p} \lambda_j a_j a_j^T M$$

Lorsque $M = I$, $X = \sum_{j=1}^{p} c_j u_j^T = \sum_{j=1}^{p} \sqrt{\lambda_j} z_j v_j^T$ où les z_j sont les vecteurs propres de norme 1 de XX^T et les v_j les vecteurs propres de $X^T X$ de norme 1.

4.1.4 Application à l'analyse des correspondances

4.1.4.1 Choix des distances pour l'analyse des correspondances

La distance euclidienne usuelle entre deux profils-lignes traduit bien la ressemblance ou la différence entre les deux couleurs des yeux sans tenir compte des effectifs totaux de ces modalités : $\sum_{k=1}^{p}\left(\frac{f_{ik}}{f_{i.}}-\frac{f_{jk}}{f_{j.}}\right)^2$. Cependant, cette distance favorise les colonnes qui ont une masse $f_{.k}$ importante, c'est-à-dire les couleurs de cheveux qui sont bien représentées dans la population étudiée. Pour remédier à cela, on pondère cheque écart par l'inverse de la masse de la colonne et l'on calcule une nouvelle distance appelée "distance du χ^2" :

$$d^2\left(i,i'\right)=\sum_{j=1}^{p}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}}-\frac{f_{i'j}}{f_{i'.}}\right)^2=\sum_{k=1}^{j}\frac{k}{k_{.j}}\left(\frac{k_{ij}}{k_{i.}}-\frac{k_{i'j}}{k_{i'.}}\right)^2$$

Il s'agit donc de la métrique diagonale kD_p^{-1}.

On définit de la même manière la distance entre les profils-colonnes par :

$$d^2\left(j,j'\right)=\sum_{i=1}^{n}\frac{1}{f_{i.}}\left(\frac{f_{ij}}{f_{.j}}-\frac{f_{ij'}}{f_{.j'}}\right)^2=\sum_{i=1}^{n}\frac{k}{k_{i.}}\left(\frac{k_{ij}}{k_{.j}}-\frac{k_{ij'}}{k_{.j'}}\right)^2$$

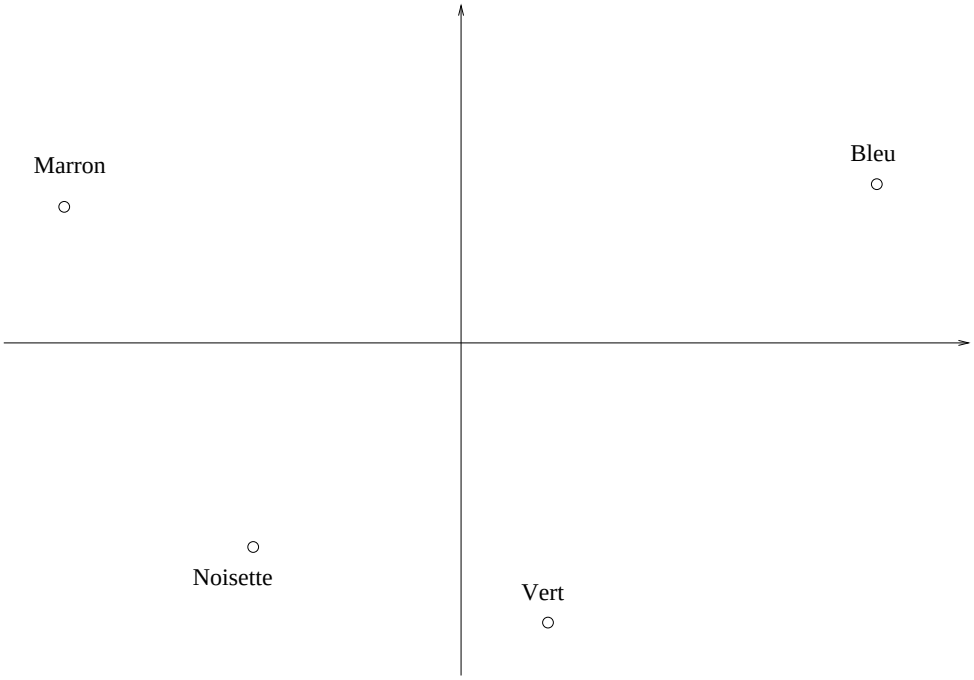
Il s'agit donc de la métrique diagonale kD_n^{-1}.

4.1.4.2 équivalence distributionnelle

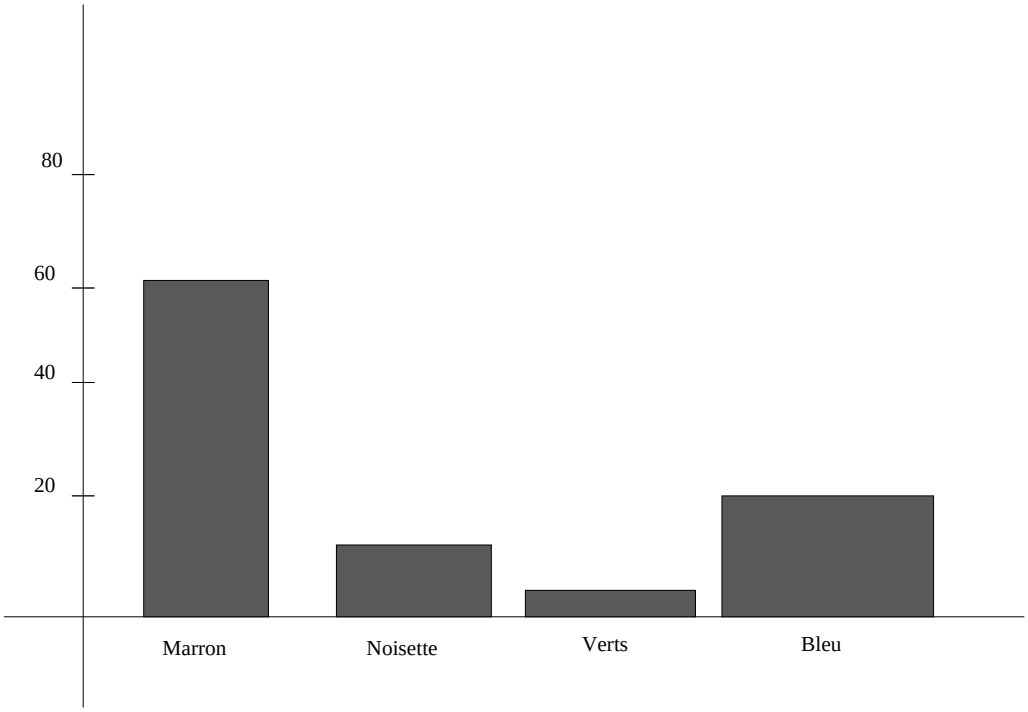
La propriété d'équivalence distributionnelle permet d'agréger deux modalités d'une même variable ayant des profil identique en une nouvelle modalité affectée de la somme de leur masses, sans rien changer, ni aux distances entre les modalités de cette variable, ni aux distances entre les modalités de l'autre variable. Si par exemple, les deux profils-lignes i et j sont identiques dans $\mathbb{R}^p$, on les agrège en un profil-ligne i dont la masse sera la somme des fréquences des deux profils i et j et les distances entre colonnes restent inchangées. Cette propriété est fondamentale puisqu'elle garantit une certaine invariance des résultats vis-à-vis de la nomenclature choisie pour la construction des modalités.

4.1.4.3 Relation de transition ou quasi-barycentrique

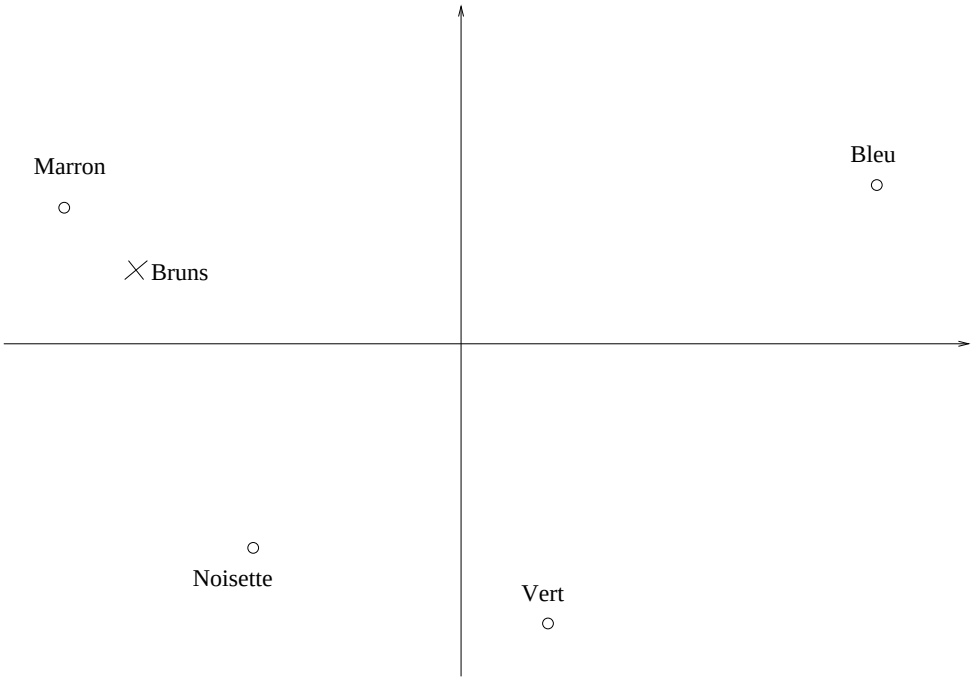
Une des caractéristique de l'analyse des correspondances est l'existence de type barycentrique qui lient graphiquement les deux variables représentées en ligne et en colonne. L'idée est simple et revient à représenter les histogrammes des profils-colonnes dans le nuage des profils-lignes et réciproquement. Supposons fixé le nuage des couleurs des yeux (nuage des profils-lignes) dans un espace à deux dimensions. Le centre du graphique représente le profil moyen des couleurs des yeux.



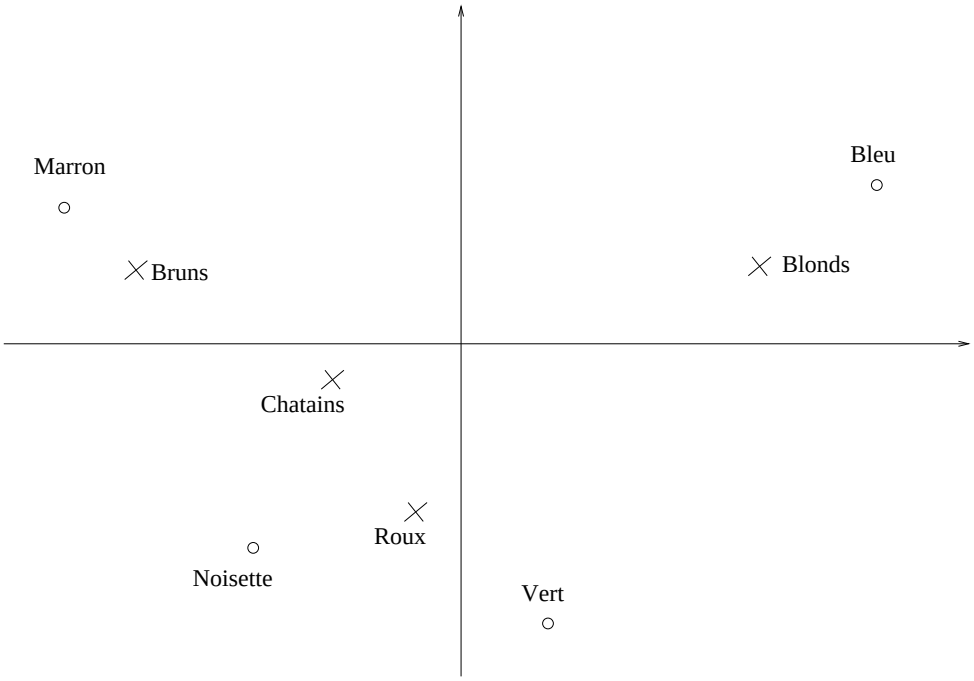
Considérons maintenant l'histogramme décrivant le profil des cheveux bruns suivant la couleur des yeux :



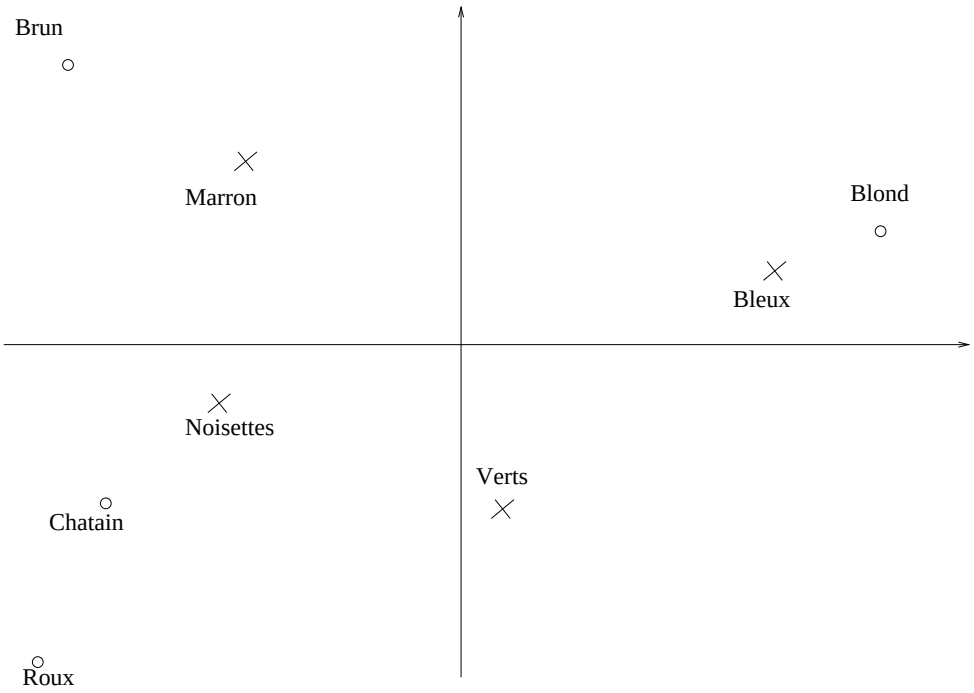
Cet histogramme va permettre de positionner le point-colonne “Cheveux bruns” dans le nuage des points lignes :



On construit ainsi le barycentre de ces points qui correspond au point "cheveux brun". Il est contenu dans une enveloppe convexe constituée par l'ensemble des points pondérés. Cette modalité sera attirée par les yeux marrons, compte tenu de sa masse plus élevée. Elle sera par contre éloigné des yeux verts. On peut ainsi représenter chaque point de la couleur des cheveux :



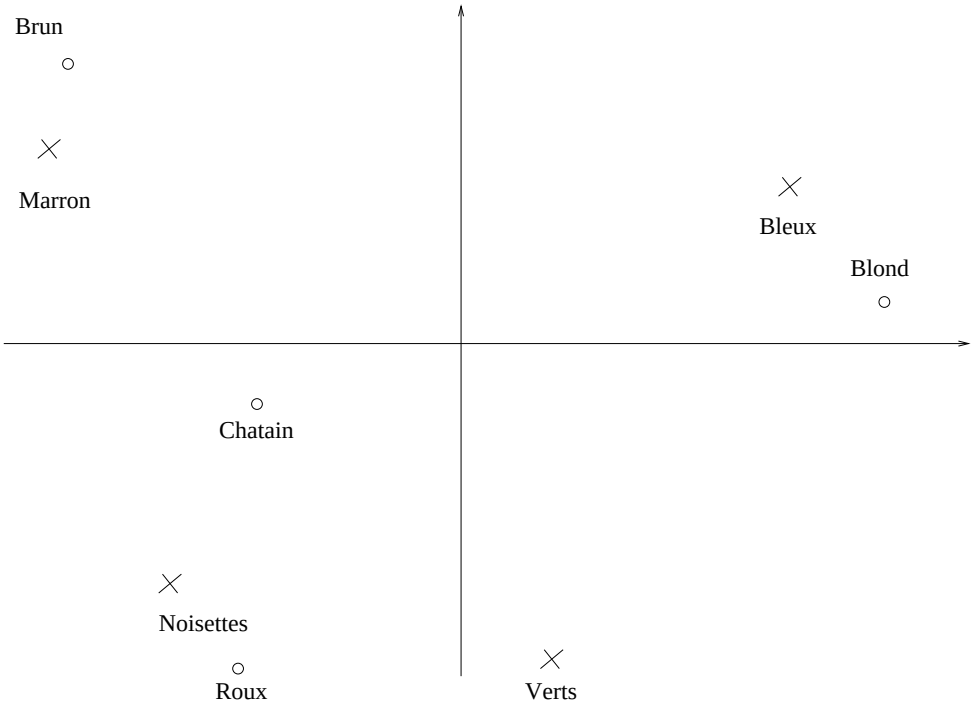
Si l'on considère maintenant le nuage des profils-colonnes, c'est-à-dire le nuage des couleurs de cheveux, il est naturel de procéder de la même façon que précedemment :



Les relations barycentriques vont justifier et donner un sens à la représentation simultanée des nuages définis dans les deux espaces.

4.1.5 Justification de la représentation simultanée

Il est possible de forcer une représentation simultanée en dilatant sur chaque axe les centres de gravité. On pourra alors représenter sur de mêmes axes l'ensemble des lignes et des colonnes afin d'approcher au mieux la situation idéale. Les relations seront quasi-barycentriques. Les yeux bleux s'associent aux cheveux blonds, les yeux marrons aux cheveux bruns. Les cheveux roux sont attirés par les yeux noisettes et verts. La catégories des cheveux chatains est assez proche de l'origine du plan représentant le profil moyen et n'est spécifique d'aucune couleur des yeux. :



4.1.6 Schéma général de l'analyse des correspondances

4.1.6.1 Eléments de base

Contrairement à l'ACP le tableau de données subit deux transformations, l'une en profils-lignes, l'autre en profils-colonnes, à partir desquelles vont être construits les nuages de points dans $\mathbb{R}^p$ et dans $\mathbb{R}^n$. Les transformations opérées sur le tableau des données peuvent s'écrire à partir des trois matrices F, D_n et D_p.

F d'ordre (n, p) désigne le tableau des fréquence relatives ; D_n d'ordre (n, n) est la matrice diagonale dont les éléments diagonaux sont les marges en lignes $f_{i.}$; D_p est la matrice diagonale d'ordre (p, p) des marges en colonnes $f_{.j}$. Les deux nuages de points (dans l'espace des colonnes et dans l'espace des lignes) sont construit de manière analogue :

Nuage de n points-lignes Dans l'espace $\mathbb{R}^p$	Eléments de base	Nuage de p points-colonnes dans l'espace $\mathbb{R}^n$
$X = D_n^{-1}F$ $\frac{f_{ij}}{f_{i.}}$, $j = 1, \cdots, p$	Acp du tableau X	$X = D_p^{-1}F^T$ $\frac{f_{ij}}{f_{.j}}$, $i = 1, \cdots, n$
$M = D_p^{-1}$ $d^2(i,j) = \sum_{k=1}^{p} \frac{1}{f_{.k}} \left(\frac{f_{ik}}{f_{i.}} - \frac{f_{jk}}{f_{j.}} \right)^2$	avec la métrique M	$M = D_n^{-1}$ $d^2(i,j) = \sum_{k=1}^{p} \frac{1}{f_{k.}} \left(\frac{f_{ki}}{f_{.i}} - \frac{f_{kj}}{f_{.j}} \right)^2$
$N = D_n$ masse du point i : $f_{i.}$	et le critère N	$N = D_p$ masse du point j : $f_{.j}$

La matrice N des masses dans un espace est liée à la métrique M utilisée dans l'autre espace.

4.1.6.2 Critère à maximiser et matrice à diagonaliser

Nous voulons représenter graphiquement les proximités entre profils. Nous commencerons par effectuer l'analyse générale par rapport à l'origine (on peut montrer que c'est équivalent à l'analyse effectuée par rapport aux centres de gravité). Plaçons dans l'espace des colonnes (compte tenu de la symétrie du tableau de contingence, les démonstrations dans l'autre espace se déduisent par permutation des indices i et j). On cherche l'axe d'inertie maximum du nuage des points lignes passant par l'origine O, engendré par un vecteur-unitaire u pour la métrique D_p^{-1}. Ceci nous amène à maximiser la somme pondérée des carrés des projections sur l'axe :

$$\max_u \left\{ \sum_i f_{i.} d^2(i, O) \right\}$$

et à rendre maximale la quantité :

$$uD_p^{-1}F^T D_n^{-1} F D_p^{-1} u$$

avec la contrainte :

$$u^T D_p^{-1} u = 1$$

u est vecteur propre de la matrice :

$$S = F^T D_n^{-1} F D_p^{-1}$$

associé à la plus grande valeur propre λ différente de 1.

De la même façon, on doit rendre maximum dans $\mathbb{R}^n$, la quantité :

$$v^T D_n^{-1} F D_p^{-1} F^T D_n^{-1} v$$

avec la contrainte

$$v^T D_n^{-1} v = 1$$

v est vecteur propre de la matrice :

$$T = F D_p^{-1} F^T D_n^{-1}$$

4.1.6.3 Axes factoriels et facteurs

Après avoir écarté la valeur propre triviale égale à 1 et le vecteur propre associé, nous retenons de la diagonalisation de la matrice les $p-1$ valeurs propres non nulles et les vecteur propres associés. Nous obtenons ainsi au plus $p-1$ axes factoriels.

Dans $\mathbb{R}^p$	Eléments de construction	Dans $\mathbb{R}^n$
$S = F^T D_n^{-1} F D_p^{-1}$	Matrice à diagonaliser	$T = F D_p^{-1} F^T D_n^{-1}$
$Su_\alpha = \lambda_\alpha u_\alpha$	Axe factoriel	$Tv_\alpha = \lambda_\alpha v_\alpha$
$\psi_\alpha = D_n^{-1} F D_p^{-1} u_\alpha$ $\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.} f_{j.}} u_{\alpha j}$	Coordonnées factorielles	$\phi_\alpha = D_p^{-1} F^T D_n^{-1} v_\alpha$ $\phi_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.} f_{.j}} v_{\alpha i}$

Les coordonnées factorielles sont centrées et de variance égales à λ_α.

4.1.6.4 Relation entre les deux espaces

L'analyse générale a montré que les matrice S et T ont les mêmes valeurs propres non nulles λ_α (cf Lebart, Morineau, Piron p. 25-26) et qu'entre le vecteur propre unitaire u_α de S associé à λ_α et le vecteur propre v_α de T relatif à la même valeur propre, il existe les relations dites de transition :

$$\begin{cases} v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F D_p^{-1} u_\alpha \\ u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F^T D_n^{-1} v_\alpha \end{cases}$$

En comparant ces formules avec les coordonnées factorielles on obtient :

$$\begin{cases} \psi_\alpha = \sqrt{\lambda_\alpha} D_n^{-1} v_\alpha \\ \phi_\alpha = \sqrt{\lambda_\alpha} D_p^{-1} u_\alpha \end{cases}$$

c'est-à-dire

$$\begin{cases} \psi_{\alpha i} = \frac{\sqrt{\lambda_\alpha}}{f_{i.}} v_{\alpha i} \\ \phi_{\alpha j} = \frac{\sqrt{\lambda_\alpha}}{f_{.j}} u_{\alpha j} \end{cases}$$

4.1.6.5 Relations de transition (ou quasi-barycentrique)

Les équations précédentes conduisent aux relations fondamentales existant entre les coordonnées des points-lignes et des points-colonnes sur l'axe α, les relations quasi-barycentriques :

$$\begin{cases} \psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} \phi_{\alpha j} \\ \phi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} \psi_{\alpha i} \end{cases}$$

Ainsi, au coefficient de dilatation $\frac{1}{\sqrt{\lambda_\alpha}}$ près, les projections des points représentatifs d'un nuage sont, sur un axe les barycentres des projections des points représentatifs de l'autre nuage. Les relations quasi-barycentrique justifient la représentation simultanée des lignes et des colonnes.

4.1.6.6 Formule de reconstitution des données

En notant que les vecteurs u_α et v_α sont maintenant orthonormés pour les métriques D_p^{-1} et D_n^{-1}, on obtient la formule pour des vecteurs ϕ_α et ψ_α normés à 1 :

$$f_{ij} = f_{i.}f_{.j}\sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\phi_{\alpha j}\psi_{\alpha i}$$

ce qui s'écrit aussi (en remarquant que la plus grande valeur propre vaut 1) :

$$f_{ij} = f_{i.}f_{.j}\left(1+\sum_{\alpha=2}^{p}\sqrt{\lambda_\alpha}\phi_{\alpha j}\psi_{\alpha i}\right)$$

4.1.7 Règle d'interprétation

Les nuages de points-lignes et de points-colonnes vont être représentés dans les plans de projection formés par les premiers axes factoriels pris deux à deux.

4.1.7.1 Inertie et test d'indépendance

L'inertie totale I du nuage de points par rapport au centre de gravité s'écrit :

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

L'inertie s'exprime également par :

$$I = \sum_{\alpha=2}^{p}\lambda_\alpha$$

La somme des valeurs propres non triviales d'une analyse des correspondances a donc une interprétation statistique simple. D'une façon générale, deux variables sont indépendantes si les profils de leurs modalités sont identiques. L'inertie est faible et il n'existe pas de direction privilégiée. Géométriquement : Tous les points sont concentrés autour du centre de gravité du nuage suivant une forme sphérique.

4.1.8 Règles d'interprétation : contributions et cosinus

Deux séries de coefficients apportent une information supplémentaire par rapport aux coordonnées factorielles :

— Les contributions qui expriment la part prise par une modalité de la variable dans la variance (inertie) expliquée par un facteur

— Les cosinus carrés, qui expriment la part prise par un facteur dans la dispersion d'une modalité de la variable

4.1.8.1 Les contributions

On cherche à connaitre les éléments responsables de la construction de l'axe α. Le coefficient

$$Cr_\alpha(i) = \frac{f_{i.}\psi_{\alpha i}^2}{\lambda_\alpha}$$

mesure la part de l'élément i dans la variance prise en compte sur l'axe α. De la même façon, on définit la contribution de l'élément j à l'axe α par :

$$Cr_\alpha(j) = \frac{f_{.j}\phi_{\alpha j}^2}{\lambda_\alpha}$$

Pour trouver une éventuelle signification à un axe, on s'intéresse d'abord aux points ayant une forte contribution. Ce sont eux qui fixent la position de l'axe (dans $\mathbb{R}^p$ pour les points i et dans $\mathbb{R}^n$ pour les points j).

4.1.8.2 Cosinus carrés

On cherche à apprécier si un point est bien représenté sur un sous espace factoriel. Les axes factoriels de chaqsue espace constituent des bases orthonormées. Le carré de la distance d'un point au centre de gravité se décompose en somme de carré des coordonnées sur ces axes. Pour un point i de $\mathbb{R}^p$, on a :

$$d^2(i,G) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

On remarque que la distance s'annule lorsque le profil du point est égal au profil moyen. Le carré de la projection de la variable i sur l'axe α vaut :

$$d_\alpha^2(i,G) = \psi_{\alpha i}^2$$

Notons que

$$\sum_{\alpha=2}^{p} d_\alpha^2(i,G) = d^2(i,G)$$

La "qualité" de la représentation du point i sur l'axe α peut être évalué par le cosinus de l'angle entre l'axe et le vecteur joignant le centre de gravité du nuage au point i :

$$\cos_\alpha^2(i) = \frac{d_\alpha^2(i,G)}{d^2(i,G)} = \frac{\psi_{\alpha i}^2}{d^2(i,G)}$$

Cette quantité, appelée "cosinus carré" représente la part de la distance au centre prise en compte dans la direction α. On l'appelle aussi la "contribution relative" du facteur à la position du point i. Plus le cosinus carré est proche de 1, plus la position du point observé en projection est proche de la position réelle du point dans l'espace. On apprécie la qualité de la représentation d'un point dans un plan en faisant la somme des cosinus carrés sur les axes étudiés. On notera que

$$\sum_{\alpha=2}^{p} Cos_\alpha^2(i) = 1$$

Ce qui vient d'être dit des n points-lignes peut être transposé aux p éléments de l'autre ensemble. On mesure la contribution relative du facteur α à la position du point j par le cosinus carré de j :

$$Cos^2_\alpha(j) = \frac{\phi^2_{\alpha j}}{d^2(j,G)}$$

et l'on a également

$$\sum_{\alpha=2}^{p} Cos^2_\alpha(j) = 1$$

4.1.9 Eléments supplémentaires

On dispose par exemple de p_s colonnes supplémentaires qui concernent des modalités de variables nominales, analogues aux colonnes de la table de contingence. Il s'agit de situer ces nouveaux points-colonnes par rapport aux p points analysés. Soit k^+_{ij} la $i$ème coordonnée de la $j$ème colonne supplémentaire. Son profil est donné par :

$$\left\{ \frac{k^+_{ij}}{k^+_{.j}} ;\ i = 1, \cdots, n \right\} \text{ et } k^+_{.j} = \sum_{i=1}^{n} k^+_{ij}$$

On projette ce pont j sur l'axe α en utilisant la même formule que dans les relation quasi-barycentriques :

$$\phi^+_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{k^+_{ij}}{k^+_{.j}} \psi_{\alpha i}$$

pour une modalité i d'une variable portée en ligne supplémentaire, on aura de façon analogue :

$$\psi^+_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{k^+_{ij}}{k^+_{i.}} \phi_{\alpha j}$$

Chapitre 5

Analyse des correspondances multiples

5.1 Domaine d'application

L'analyse des correspondances multiples est une analyse des correspondances simple appliquée non plus à une table de contingence mais à un tableau disjonctif complet. L'extension du domaine d'application de l'analyse des correspondances se fonde sur l'équivalence suivante : Si pour n individus, on dispose des valeurs prises par deux variables nominales ayant respectivement p_1 et p_2 modalités, il est alors équivalent de soumettre à l'analyse des correspondances le tableau de contingence p_1, p_2 croisant les deux variables ou d'analyser le tableau binaire à n lignes et $(p_1 + p_2)$ colonnes décrivant les réponses.

5.2 Notations et définitions

5.2.1 Tableau disjonctif complet

On désigne par I l'ensemble des n sujets ayant répondu au questionnaire et par p le nombre total des modalités des s questions. On a :

$$p = \sum_{q=1}^{s} p_q$$

On construit, à partir du tableau de données $\mathbb{R}$, le tableau Z à n lignes et p colonnes décrivant les s réponses des n individus par un codage binaire. Le tableau Z est la juxtaposition des s sous-tableaux

$$Z = [Z_1, \cdots, Z_s]$$

Le sous-tableau Z_q à n lignes et p_q colonnes, sa $i$ème ligne contient $p_q - 1$ fois la valeur 0 et une fois la valeur 1. La tableau Z est appelé tableau disjonctif

complet dont le terme général s'écrit :

$$z_{ij} = 1 \text{ ou } z_{ij} = 0$$

selon que le sujet i a choisi la modalité j de la question q ou non. Les marges en ligne du tableau disjonctif complet sont constantes et égales au nombre s de questions :

$$z_{i.} = \sum_{j=1}^{p} z_{ij} = s$$

Les marges colonnes : $z_{.j} = \sum_{i=1}^{n} z_{ij}$ correspondent au nombre de sujets ayant choisi la modalité j de la question q.

5.2.2 Tableau de contingence de Burt

On construit, à partir du tableau disjonctif complet Z, le tableau symétrique (contigence de Burt) B d'ordre (p, p) qui rassemble les croisements deux à deux de toutes les variables :

$$B = Z^T Z$$

Le terme général de Bs'écrit

$$b_{ij} = \sum_{k=1}^{n} z_{ki} z_{kj}$$

B est une juxtaposition de tableau de contingence.

Le tableau B est formé de s^2bloc où l'on distingue :

- Le bloc $Z_q^T Z_{q'}$ indicé par (q, q') d'ordre $(p_q, p_{q'})$ qui n'est autre que la table de contingence croisant les réponses aux question q et q'.
- Le $q$ième bloc $Z_q^T Z_q$obtenu par le croisement d'une variable avec elle-même. C'est une matrice d'ordre (p_q, p_q)diagonale. Les termes diagonaux sont les effectifs des modalités de la question q.

5.2.3 Principe de l'analyse des correspondances multiples

L'analyse des correspondance multiples est l'analyse des correspondances d'un tableau disjonctif complet. Ses principes sont donc ceux de l'analyse des correspondances, à savoir :

- Mêmes transformation du tableau de données en profil-lignes et en profils-colonnes.
- Même critère d'ajustement avec pondération des points par leurs profils marginaux
- Même distance du χ^2.

5.2.3.1 Critère d'ajustement

Les individus sont tous affectés d'une masse identique égale à $m_i = \frac{1}{n}$ et chacune des modalités j est pondéré par sa fréquence $m_j = \frac{z_{.j}}{ns}$. La distance du χ^2 appliquée appliquée à un tableau disjonctif complet conserve un sens. En effet, dans $\mathbb{R}^n$, la distance entre modalité s'écrit :

$$d^2(i,j) = \sum_{k=1}^{n} n \left(\frac{z_{ki}}{z_{.i}} - \frac{z_{kj}}{z_{.j}} \right)^2$$

Ainsi deux modalités choisies par les mêmes individus coincident.

Dans $\mathbb{R}^p$, la distance entre deux individus i et j s'exprime par

$$d^2(i,j) = \frac{1}{s} \sum_{k=1}^{p} \frac{n}{z_{.k}} (z_{ik} - z_{jk})^2$$

Deux individus sont proches s'ils ont choisi les mêmes modalités.

5.2.3.2 Axes factoriel et facteurs

Nous désignons par D la matrice diagonale d'ordre (p,p) ayant les mêmes élements diagonaux que la tableau de Burt. Ces éléments sont les effectifs correspondant à chacune des modalités :

$$d_{jj} = b_{jj} = z_{.j}$$

Les autres éléments étant nuls. On pose :

- $F = \frac{1}{ns} Z$ de terme général $f_{ij} = \frac{z_{ij}}{ns}$
- $D_p = \frac{1}{ns} D$ de terme général $f_{.j} = \delta_{ij} \frac{z_{.j}}{ns}$
- $D_n = \frac{1}{n} I_n$ de terme général $f_{i.} = \frac{\delta_{ij}}{n}$

Pour trouver les axes factoriels u_α on diagonalise la matrice :

$$S = F^T D_n^{-1} F D_p^{-1} = \frac{1}{s} Z^T Z D^{-1}$$

Dans $\mathbb{R}^p$, l'équation du αième axe factoriel u_α est :

$$\frac{1}{s} Z^T Z D^{-1} u_\alpha = \lambda_\alpha u_\alpha$$

L'équation du αième facteur $\phi_\alpha = D^{-1} u_\alpha$ s'écrit :

$$\frac{1}{s} D^{-1} Z^T Z \phi_\alpha = \lambda_\alpha \phi_\alpha$$

De même, l'équation du αième facteur ψ_α dans $\mathbb{R}^n$ s'écrit :

$$\frac{1}{s} Z D^{-1} Z^T \psi_\alpha = \lambda_\alpha \psi_\alpha$$

Les facteur ϕ_α et ψ_α de norme λ_α représentent les coordonnées de points-lignes et des points-colonnes sur l'axe factoriel α. Les relations de transition entre les facteur ϕ_α et $\psi_\alpha$sont :

$$\begin{cases} \phi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D^{-1} Z^T \psi_\alpha \\ \psi_\alpha = \frac{1}{s\sqrt{\lambda_\alpha}} Z \phi_\alpha \end{cases}$$

5.2.3.3 Facteurs et relations quasi-barycentriques

La coordonnée factorielle de l'individu i sur l'axe α est donnée par :

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{z_{ij}}{z_{i.}} \phi_{\alpha j} = \frac{1}{s\sqrt{\lambda_\alpha}} \sum_{j \in p(i)} \phi_{\alpha j}$$

où $p(i)$ désigne l'ensemble des modalités choisies par l'individu i. Au coefficient $\frac{1}{\sqrt{\lambda_\alpha}}$ près, l'individu i se trouve au point moyen du nuage des modalités qu'il a choisies.

De même, la coordonnée de la modalité j sur l'axe α est donné par :

$$\phi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{z_{ij}}{z_{.j}} \psi_{\alpha i} = \frac{1}{z_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in I(j)} \psi_{\alpha i}$$

où $I(j)$ désigne l'ensemble des individus ayant choisi la modalité j. Avant la dilatation sur l'axe α, la modalité j se trouve au point moyen du nuage des individus qui l'ont choisie comme réponse.

La représentation simultanée des individus et des modalités est rarement utilisée au cause de l'encombrement du graphique.

5.2.3.4 Inertie du nuage des modalités

On rappelle que la distance du χ^2 dans $\mathbb{R}^n$ est la métrique D_n^{-1}. La distance entre la modalité j et le centre de gravité du nuage G, dont toutes les n coordonnées valent $\frac{1}{n}$, s'écrit :

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

La distance d'une modalité au centre de gravité est d'autant plus grande que l'effectif est plus faible.

Inertie d'une modalité L'inertie $I(j)$ de la modalité j vaut :

$$I(j) = m_j d^2(j, G)$$

avec

$$m_j = \frac{z_{.j}}{ns}$$

d'où

$$I(j) = \frac{1}{s} \left(1 - \frac{z_{.j}}{n} \right)$$

La part d'inertie due à une modalité de réponse est d'autant plus grande que l'effectif dans cette modalité est plus faible. Le maximum $\frac{1}{s}$ serait atteint par une modalité d'effectif nul. En conséquence, on évite, au moment du codage, les modalité à faibles effectifs susceptibles de perturber les directions des premiers axes factoriels.

Inertie d'une question L'inertie de la question q, notée $I(q)$, vaut :

$$I(q) = \sum_{j=1}^{p_q} I(j) = \frac{1}{s}(p_q - 1)$$

Ainsi, la part d'inertie due à une question est fonction croissante du nombre de modalité de réponse. La part minimale $\frac{1}{s}$ correspond aux questions à 2 modalités. D'où l'intérêt d'équilibrer le système des questions, c'est-à-dire le découpage des variables en modalités, si on veut faire jouer le même rôle à toutes les questions.

Inertie totale On en déduit que l'inertie totale I vaut :

$$I = \sum_{q} I(q) = \sum_{j=1}^{p} \frac{z_{.j}}{ns} d^2(j, G) = \frac{p}{s} - 1$$

En particulier, elle vaut 1 dans le cas où toutes les questions ont deux modalités de réponse ($p = 2s$). L'inertie totale dépend uniquement du nombre de variables et de modalités et non des liaisons entre les variables. C'est une quantité qui n'a pas de signification statistique.

5.2.4 Règles d'interprétation

On retrouve les règles de l'analyse des correspondances en ce qui concerne les cosinus carrés et les contributions.

De plus, dire qu'il existe des affinités entre réponse, c'est dire aussi qu'il existe des individus qui ont choisi simultanément toutes ou presque toutes ces réponses. Ainsi, on exprime :

- La proximité entre individus en terme de ressemblances : Deux individus se ressemblent s'ils ont choisi globalement les même modalités.
- La proximité entre modalités de variables différentes en terme d'association : Ces modalités correspondent aux points moyens des individus qui les ont choisis et sont proches parce qu'elles concernent des individus semblables.
- La proximité entre deux modalité d'une même variable en termes de ressemblance : Par construction, les modalités d'une même variable s'excluent. Si elles sont proches, cette proximité s'interprète en terme de ressemblance entre les groupes d'individus qui les ont choisis.

5.2.5 Exemples : Le thé

5.2.5.1 Description des données

On considère le jeu de donnée du thé : (http ://factominer.free.fr/livre/the.csv), data(tea). Il s'agit d'une enquête auprès de 300 consommateurs de thé. Les différentes questions qui leur ont été posées concernent la façon dont ils consomment le thé, l'image qu'ils ont du produit et enfin leur signalétique.

5.2.5.2 Descriptions des variables

Dans l'analyse, les variables de comportement de consommation sont introduites seules actives, les variables d'images et de signalétiques sont supplémentaires. Dix neuf questions concernent la façon dont ils consomment le thé, douze questions concernent l'image, quatre la signalétique (sexe, CSP, ...).

L'ACM est donc obtenue en précisant que la variable 22 est supplémentaire et que les variables 19 à 21 et 23 à 36 sont qualitatives supplémentaires :

```
library(FactoMineR)
data(tea)
res.mca=MCA(tea,quanti.sup=19,quali.sup=20:36)
```

Cette commande fournie le nuage des individus et des variables.

5.2.5.3 Le nuage des individus

Comme dans la pluspart des enquêtes le nuage des individus comportent beaucoup de points et on cherche uniquement à voir si il se dégage une forme particulière, voir des groupes d'individus. Dans l'exemple le nuage a plutôt une forme homogène.

5.2.5.4 Le nuage des variables

Ce graphe est obtenue en calculant les les rapport de corrélations entre les coordonnées des individus sur un axe et chacune des variables qualitatives. Si le rapport de corrélation entre la variable j et l'axe s est proche de 1, les individus possédant la même modalité (pour cette variable qualitative) ont des coordonnées voisines pour l'axe s .

Ici les variables (type), (forme) et (lieu d'achat) sont très liées à chacun des deux premiers axes, mais on ne sait pas encore comment.

5.2.5.5 Nuage des modalités

On peut construire le nuage des toutes les modalités actives par :

```
plot(res.mca,invisible=c("ind","quali.sup","quanti.sup"),cex=0.8)
```

Le premier axe oppose les modalités (salon de thé) (GMS et magasin spécialisé), (sachet+vrac), (bar) (resto) (travail) aux modalités (Pas.amis) (Pas.resto), (Pas.travail), (Pas.maison). Ce premier axe oppose donc les buveurs régulier aux occasionnels. Quant au deuxième il oppose les modalités (magasin spécialisé), (vrac), (type haut de gamme) et dans une moindre mesure (vert) et (après-diner) de l'ensemble des autres modalités. Il s'agirait donc plus des vrais amateurs "puristes" aux autres.

5.2.5.6 Description automatiques des axes

On peut aussi obtenir le tableau des valeurs propres et la description automatique :

```
dimdesc(res.mca)
```

Ici le premier axe est caractérisé par les variables (lieux d'achat), (salon de thé) et certaine variable supplémentaires sont bien liées à cet axe :(sexe) (convivialité). Comme la pluspart des variables ont deux modalités, la caractérisation par les modalités est similaire à celle des variables, car les deux modalités sont exclusives.

5.2.5.7 Ellipses de confiance

Des ellipse de confiance peuvent être tracée autour des modalités d'une variable quantitative (i.e. autour du barycentre des individus qui possèdent la modalité). Ces ellipses sont adaptées à des représentations planes et permettent de visualiser si deux modalités sont significativement différentes. Il est possible de construire des ellipses de confiance pour l'ensemble des modalités de plusieurs variables quantitatives grâce à la fonction plotellipse :

```
plotellipses(res.mca,keepvar=1:4)
```