Analyse de Données

Mounir Bouguebina

Table des matières

Introduction

L'analyse de données est un ensemble de méthodes de statistique descriptive ayant pour objectif de résumer et visualiser l'information contenue dans un grand tableau de données. Ces données sont d'abord collectées par des recensements, des enquêtes ou diverses investigations, ce qui donne des tableaux (ou matrices) de nombres de plusieurs milliers de lignes et de colonnes. Elles sont ensuite étudiées géométriquement par des méthodes de projections pour résumer l'information et la représenter graphiquement. Les résultats sont enfin interpretés et les conclusions qui s'imposent tirées.

Dans ce cours nous présentons certaines méthodes pour étudier ces grands tableaux de données. Ce sont les methodes d'analyse factorielle. On en verra deux, l'analyse en composantes principales (ou ACP) et l'analyse factorielle des correspondances (ou AFC). Les techniques utilisées dans les deux sont les mêmes mais elles diffèrent essentiellement sur la nature des tableaux étudiés. L'ACP va porter sur les tableaux de données de variables quantitatives et l'AFC sur les tableaux de contingence obtenus par croisemment de modalités de deux ou plusieurs variables qualitatives. Les deux méthodes ont un but essentiellement descriptif en condensant l'information contenue dans ces grands tableaux dans des representations graphiques (et donc géométriques) le plus souvent à deux dimensions (lisibles donc sur une simple feuille de papier).

L'objectif de ce cours est donc de décrire de manière simple et d'illustrer par des exemples élémentaires les principes de base de ces deux méthodes. On commencera au tout debut et les étudiants ne sont censés connaître que quelques notions de Géométrie et d'Algèbre linéaire vues en première et deuxième année de Licence.

Chapitre 1

Analyse en composantes principales

1.1 Introduction

L'analyse en composantes principales est une méthode factorielle qui rentre dans le cadre de la statistique descriptive multidimensionnelle permettant de traiter simultanément un grand nombre de variables par opposition à la statistique descriptive univariée (une seule variable) ou bivariée (deux variables). Conçue par Karl Pearson en 1901, on l'utilise par exemple en biologie, en medecine ou encore en économie. Son usage a explosé dans les années soixante avec l'avénement des ordinateurs qui permettent le traitement d'un grand nombre de données en exploitant leurs aspects géométriques et leurs representations graphiques.

L'analyse en composantes principales ou ACP va porter sur des tableaux de données comportant l'étude de p variables quantitatives sur une population statistique de n individus.

Exemple 1.1.1. *On considère le tableau X de notes sur* 20 *obtenues par* 9 *élèves en Mathematiques, physique, Français et Anglais. On a donc* 9 *individus et* 4 *variables.*

	Maths	*Physique*	*Français*	*Anglais*
Amine	*6.0*	*6.0*	*5.0*	*5.5*
Moussa	*8.0*	*8.0*	*8.0*	*8.0*
Warda	*6.0*	*7.0*	*11.0*	*9.5*
Siham	*14.5*	*14.5*	*11.0*	*9.5*
Hanane	*14.0*	*14.0*	*15.5*	*15.0*
Farid	*11.0*	*10.0*	*5.5*	*7.0*
Randa	*5.5*	*7.0*	*14.0*	*11.5*
Faiza	*13.0*	*12.5*	*8.5*	*9.5*
Adel	*9.0*	*9.5*	*12.5*	*12.0*

On cherche des representations géométriques de ces individus et variables. Pour les individus, on cherche quels sont ceux qui se ressemblent et quels sont ceux qui se distinguent des autres ou quels sont ceux qui forment des groupes. Pour les variables, on cherche quelles sont celles qui sont correlées entre elles ou celles qui sont indépendantes les unes des autres.

On a donc en général un tableau (ou matrice) de données quantitatives de p variables $X_1, \ldots, X_p$ portant sur n individus $u_1, \ldots, u_n$

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

avec

$$x_{ij} = X_j(u_i)$$

la valeur de la variable X_j sur l'individu u_i. Les vecteurs lignes de la matrice X représentent les individus

$$u_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

vus comme point dans $\mathbb{R}^p$ et les vecteurs colonnes représentent les variables

vues comme points (ou vecteurs) dans $\mathbb{R}^n$. Les individus forment donc un nuage de n points dans $\mathbb{R}^p$ et les variables forment un nuage de p points dans $\mathbb{R}^n$. Dans l'exemple des notes on a donc 9 points individus dans $\mathbb{R}^4$ et 4 points ou vecteurs variables dans $\mathbb{R}^9$. Il est difficile de représenter graphiquement

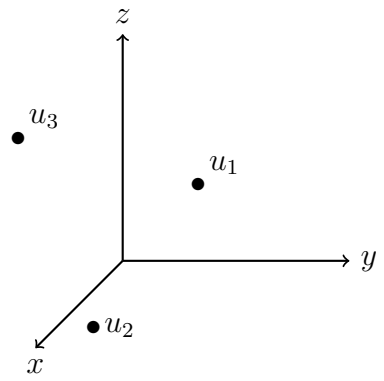

FIGURE 1.1 – Nuage de points individus

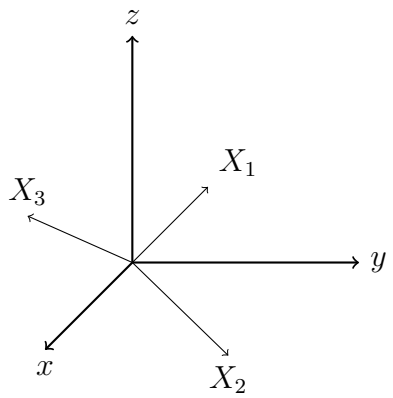

FIGURE 1.2 – Nuage de points-vecteurs variables

ces individus et variables dans des espaces de dimension supérieure à 3 ce qui est le cas ici. l'espace des individus est de dimension 4 et celui des variables est de dimension 9.

L'objectif de l'ACP est de simplifier l'étude de ces nuages de points en les projetant sur des sous-espaces de faible dimension $(1, 2$ ou $3)$ pour obtenir des représentations graphiques les plus simples possibles tout en gardant un maximum d'information. Pour n et p grands le nombre de données np est trés grand. Il s'agit de tirer le plus d'informations de ces nombres. L'ACP le fait de manière géométrique.

Remarque 1.1.1. *On préfère regarder les individus comme des points car ce qui nous importe ce sont les distances entre ces points. On cherche alors*

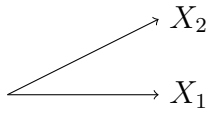

FIGURE 1.3 – Variables correlées positivement

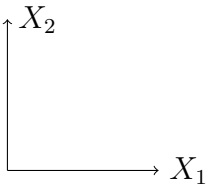

FIGURE 1.4 – Variables indépendantes

à savoir quels sont les points proches les uns des autres et quels sont ceux qui sont loin des autres ou encore quels sont ceux qui forment des groupes. Par contre on préfère regarder les variables comme des vecteurs car ce qui va nous importer c'est l'angle entre ces vecteurs. En effet on verra que ces angles décrivent le degré de liaison entre les variables. Un petit angle entre deux vecteurs-variables voudra dire que ces deux variables sont positivement correlées entre elle. Un angle droit voudra dire qu'elles sont indépendantes et un angle plat voudra dire qu'elles sont correlées négativement.

Exercice 1.1.1. *Soient les tableaux de données*

$$X = \begin{pmatrix} 11.0 & 13.5 \\ 12.5 & 9.0 \\ 2.0 & 7.5 \\ 4.5 & 11.5 \\ 11.0 & 10.0 \\ 9.5 & 12.0 \end{pmatrix}$$

FIGURE 1.5 – Variables correlées négativement

et

$$Y = \begin{pmatrix} 6 & 6 & 7 & 5.5 \\ 6 & 8 & 11 & 9.5 \end{pmatrix}$$

Dessiner les points individus de X et les vecteurs variables de Y. Peut-on faire la même chose pour le tableau des notes ?

On suppose que chaque individu u_i est muni d'un poids p_i avec

$$p_i \geq 0$$

et

$$\sum_{i=1}^{n} p_i = 1$$

En général les individus vont avoir le même poids

$$p_i = \frac{1}{n}$$

mais il arrive dans certains cas qu'on travaille avec des poids différents par exemple quand un individu représente une partie de la population statistique étudiée. Son poids sera d'autant plus grand que cette population est importante.

On regroupe les poids dans la matrice diagonale D_p de taille n

$$D_p = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{pmatrix}$$

qu'on appelle matrice des poids. Si $p_i = \frac{1}{n}$ pour tout i on a

$$D_p = \frac{1}{n} I_n$$

avec I_n la matrice identité de taille n

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Pour mesurer les distances entre les points individus dans $\mathbb{R}^p$ et les angles entre vecteurs variables dans $\mathbb{R}^n$ on a besoin d'un produit scalaire dans $\mathbb{R}^p$ et d'un produit scalaire dans $\mathbb{R}^n$.

1.2 Point moyen ou barycentre

Les individus $u_i, i = 1, \ldots, n$ forment un nuage de n points dans $\mathbb{R}^p$. Certains points du nuage vont être éloignés de l'origine des coordonnées $O = (0, \ldots, 0)$ et leur contribution à l'information contenue dans le nuage tout entier peut nous induire en erreur. Mais il y a un point qui va être par sa définition même au centre du nuage et donc proche de tous les points en même temps. Il est préferable de travailler avec ce point comme nouvelle origine. Ce point est le barycentre ou point moyen du nuage.

Définition 1.2.1. *Le barycentre G (ou point moyen) des n individus u_i munis des poids p_i est le point*

$$G = (\overline{X_1}, \overline{X_2}, \ldots, \overline{X_p})$$

avec

$$\overline{X_j} = \sum_{i=1}^{n} p_i x_{ij}$$

Si $p_i = \dfrac{1}{n}$ pour tout i, on a

$$\overline{X_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

qui est la moyenne de la variable X_j.

Exercice 1.2.1. *Montrer qu'on a en notation matricielle*

$$G = X^{'} D_p 1_n$$

avec $X^{'}$ la transposée de la matrice X et

$$1_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Exercice 1.2.2. *Calculer les points moyens du tableau des notes et des deux tableaux de l'exercice* 1.1.1 *avec $p_i = \dfrac{1}{n}$ pour i allant de* 1 *à n.*

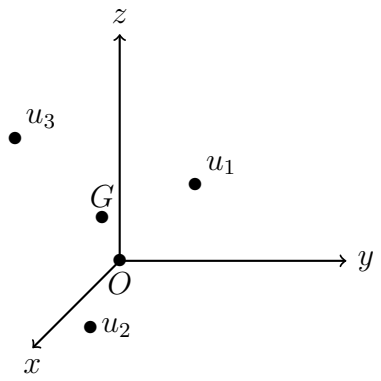

FIGURE 1.6 – Barycentre du nuage de points

Le tableau de données initial X est transformé en le tableau centré

$$Y = (y_{ij})$$

pour i allant de 1 a n et j allant de 1 à p. avec

$$y_{ij} = x_{ij} - \overline{X_j}$$

Exercice 1.2.3. *Vérifier qu'on a, en notation matricielle*

$$Y = X - 1_n G$$

Géométriquement travailler avec Y revient à faire une translation de l'origine des coordonnées de $O = (0, \ldots, 0)$ à G dans $\mathbb{R}^p$.

Exercice 1.2.4. *Calculer les tableaux centrés de l'exemple des notes et des tableaux de l'exercice* 1.1.1.

1.3 Réduction des données

La variance de la variable X_j est

$$var(X_j) = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{X_j})^2$$

Son écart type est

$$\sigma_j = \sigma(X_j) = \sqrt{var(X_j)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_j})^2}$$

L'écart type mesure la dispersion de la variable X_j autour de sa moyenne. Le tableau centré réduit est le tableau

$$Z = (z_{ij})$$

pour i allant de 1 à n et j allant de 1 à p avec

$$z_{ij} = \frac{x_{ij} - \overline{X_j}}{\sigma_j}$$

Exercice 1.3.1. *Montrer qu'on a en notation matricielle*

$$Z = YD_{\frac{1}{\sigma}}$$

avec

$$D_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_p} \end{pmatrix}$$

La matrice diagonale des inverses des écarts types.

Remarque 1.3.1. *L'écart type σ_j a la même unité de mesure que les valeurs de la variable X_j et donc z_{ij} est sans dimension. Réduire les données revient donc à les homogeiniser at à éliminer le problème des unités.*

Si par exemple on mesure la longueur d'une voiture en mètres et sa largeur en centimètres les largeurs vont contribuer beaucoup plus que les longueurs dans les calculs (comparer une longueur de 4 mètres à une largeur de 180 centimètres). En les réduisant on s'affranchit non seulement des unités de mesure (les centimètres et les mètres disparaissent) mais on uniforme en plus les nombres en jeu. Si on veut l'écart type devient lui-même l'unité homogène par laquelle on mesure l'écart des valeurs d'une variable par rapport à sa moyenne. Soit par exemple la variable X qui mesure la puissance en chevaux

de 30 voitures. Si la moyenne est 92 chevaux et l'écart type est 24 chevaux, alors une voiture de 140 chevaux aura

$$\frac{140-92}{24}=2$$

écarts-types au dessus de la moyenne

Exercice 1.3.2. *Calculer le tableau centré réduit Z pour le tableau des notes et les tableaux de l'exercice* 1.1.1 .

La covariance entre deux variables X_j et X_k est

$$Cov(X_j,X_k)=\sigma_{X_jX_k}=\sigma_{jk}$$

$$=\sum_{i=1}^{n}p_i(x_{ij}-\overline{X_j})(x_{ik}-\overline{X_k})$$

Le coefficient de corrélation est donné par

$$cor(X_j,X_k)=r_{X_jX_k}$$

$$=\frac{\sigma_{X_jX_k}}{\sigma_{X_j}\sigma_{X_k}}=\frac{cov(X_j,X_k)}{\sqrt{var(X_j)}\sqrt{var(X_k)}}$$

La covariance d'une variable avec elle même est sa variance

$$cov(X_j,X_j)=var(X_j)$$

et donc sa corrélation avec elle même est

$$cor(X_j,X_j)=r_{X_jX_j}=1$$

La covariance et la corrélation sont symétriques. Elles ne dépendent pas de l'ordre des variables

$$cov(X_j,X_k)=cov(X_k,X_j)$$
$$cor(X_j,X_k)=cor(X_k,X_j)$$

Exercice 1.3.3. *Montrer que* $\overline{Z_j}=0$, $\sigma_{Y_j}=\sigma_{X_j}$ *et* $\sigma_{Z_j}=1$.

Définition 1.3.1. *La matrice de covariance du tableau X est la matrice*

$$V = (\sigma_{jk})$$

pour j, k allant de 1 à p.

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{j1} & \sigma_{j2} & \dots & \sigma_{jp} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

C'est une matrice carrée symétrique $p \times p$. Sa matrice de corrélation est

$$R = (r_{jk})$$

pour j, k allant de 1 à p.

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{j1} & r_{j2} & \dots & r_{jp} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

C'est aussi une matrice carrée symétrique $p \times p$ avec des 1 sur la diagonale.

Exercice 1.3.4. *1. Montrer que*

$$V = Y^{'} D_p Y = X^{'} D_p X - G^{'} G$$

2. Montrer que

$$R = D_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}}$$

et que

$$R = \frac{1}{n} Z^{'} Z$$

si $p_i = \dfrac{1}{n}$.

3. Montrer que la matrice de covariance du tableau centré réduit Z est la matrice de corrélation du tableau initial X.

Exercice 1.3.5. *Calculer les matrices de covariance et de corrélation du tableau des notes et des tableaux de l'exercice* 1.1.1.

1.4 Interpretation géométrique

L'objectif de l'ACP est d'évaluer les ressemblancesentre individus et les liaisons entre variables. Pour savoir si deux individus sont proches ou éloignés l'un de l'autre on a besoin d'une notion de distance entre individus vus comme points dans $\mathbb{R}^p$. La liaison entre variables s'exprime par l'angle qu'elles font entre elles en tant que vecteurs dans $\mathbb{R}^n$. Du cours de Géométrie on sait que les distances et les angles sont donnés par un produit scalaire.

Définition 1.4.1. *Un produit scalire sur $\mathbb{R}^k$ est une application*

$$\phi : \mathbb{R}^k \times \mathbb{R}^k \longrightarrow \mathbb{R}$$

qui vérifie

1. *ϕ est bilinéaire.*
2. *ϕ est symétrique $\phi(u,v) = \phi(v,u)$.*
3. *ϕ est définie positive $\phi(u,u) \geq 0$ et $\phi(u,u) = 0$ si et seulement si $u = \vec{0}$.*

On note

$$\phi(u,v) = \langle u, v \rangle$$

Si $e_1 = (1,0,\ldots), \ldots, e_k = (0,\ldots,1)$ est la base canonique de $\mathbb{R}^k$, la matrice du produit scalaire dans cette base est $M = (a_{ij})$ avec $a_{ij} = \langle e_i, e_j \rangle$ pour i, j allant de 1 à k. Si $u = (u_1, \ldots, u_k)$et $v = (v_1, \ldots, v_k)$ on a

$$\langle u, v \rangle = \sum_{i,j} a_{ij} u_i v_j$$

$$= uMv^{'}$$

avec $v^{'}$ le vecteur colonne transposé de v. La matrice M est une matrice carrée symétrique $k \times k$. Par la définition du produit scalaire on a

$$\langle u, u \rangle = uMu^{'} > 0$$

pour tout vecteur u non nul. C'est la définition d'une matrice définie positive.

Exercice 1.4.1. *Soit dans $\mathbb{R}^2$ $\langle u, v \rangle = (u_2 - u_1)(v_2 - v_1) + (2u_1 - u_2)(2v_1 - v_2)$.*

1. *montrer que c'est un produit scalaire.*

2. *Donner sa matrice dans la base canonique.*

3. *Donner sa matrice dans la base* $\{(1,2),(1,1)\}$.

Définition 1.4.2. *Deux vecteurs* u, v *dans* $\mathbb{R}^k$ *sont dits orthogonaux si* $\langle u, v\rangle = 0$.

Le produit scalaire permet de définir une norme par la formule

$$\|u\|^2 = \langle u, u\rangle = uMu^{'}$$

une distance par la formule

$$d^2(u,v) = \|v-u\|^2 = (v-u)M(v-u)^{'}$$

et les angles par la formule

$$cos\theta = \frac{\langle u, v\rangle}{\|u\|\|v\|}$$

pour θ l'angle entre les vecteurs u et v.

Définition 1.4.3. *Une ACP normée est une ACP qui va porter sur des données centrées réduites et donc sur le tableau de données* Z.

Dans l'espace des individus $\mathbb{R}^p$ pour p le nombre de variables on considère le produit scalaire de matrice $M = I_p$ la matrice identité $p \times p$.

Exercice 1.4.2. *Montrer que le produit scalaire* $M = D_{\frac{1}{\sigma^2}}$ *(matrice diagonale des inverses des écarts types au carré) sur le tableau centré* Y *équivaut au produit scalaire* $M = I_p$ *sur le tableau centré réduit* Z.

Sur l'espace des variables $\mathbb{R}^n$ le produit scalaire a pour matrice $M = D_p$ la matrice des poids (toujours dans la base canonique). Si tous les poids $p_i = \frac{1}{n}$, on a $M = \frac{1}{n} I_n$.

Pour les données initiales centrées, le produit scalaire entre deux variables Y_j, Y_k est

$$\langle Y_j, Y_k\rangle = Y_j D_p Y_k^{'} = \sigma_{Y_j Y_k} = cov(Y_j, Y_k)$$

et

$$\|Y_j\|^2 = \langle Y_j, Y_j\rangle = \sigma^2_{Y_j} = var(Y_j)$$

et donc

$$\|Y_j\| = \sigma_{Y_j}$$

Si θ_{jk} est l'angle entre Y_j et Y_k, on a

$$cos\theta_{jk} = \frac{\langle Y_j, Y_k \rangle}{\|Y_j\|\|Y_k\|} = \frac{\sigma_{Y_j Y_k}}{\sigma_j \sigma_k}$$

et donc

$$cos\theta_{jk} = cor(Y_j, Y_k)$$

formule qui exprime le coefficient de corrélation de Y_j et Y_k (et donc aussi celui de X_j et X_k) comme le cosinus de l'angle entre les variables. En particulier on a

$$-1 \leq cor(X_j, X_k) \leq 1$$

Pour les variables centrées réduites on a donc

$$\|Z_j\| = 1$$

pour j allant de 1 à p et

$$cov(Z_j, Z_k) = cor(Z_j, Z_k) = \langle Z_j, Z_k \rangle$$

$$= cos\theta_{jk}$$

pour θ_{jk} l'angle entre les variables Z_j et Z_k. Quand cet angle est nul (correspondant à un cosinus égal à 1) ou plat (correspondant à un cosinus égal à -1) les deux variables sont sur la même droite et ont même sens dans le premier cas et opposées dans le second cas. Elles sont donc en liaison linéaire (ou plutôt affine)

$$Z_k = aZ_j + b$$

Quand cet angle est droit (correspondant à un cosinus nul) les deux variables sont sans liaison entre elles. Elles sont indépendantes.

1.5 Inertie

On va traviller sur les données centrés réduites et donc avec le tableau Z. Le barycentre ou point moyen est donc à l'origine des coordonnées

$$G = O$$

et on a un nuage de n points individus dans $\mathbb{R}^p$.

Définition 1.5.1. *L'inertie totale du nuage de points individus est*

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(w_i, O) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij}^2$$

C'est (au terme $\frac{1}{n}$ prés) la somme des carrés des distances des points individus w_i à l'origine des coordonnées $O = G$. L'inertie mesure la dispersion des points du nuage autour de son barycentre ou centre de gravité : plus l'inertie est grande , plus le nuage est dispersé et inversement une inertie petite indique que le nuage est concentré autour de son centre de de gravité.

Exercice 1.5.1. *Montrer que*

$$I_T = \sum_{j=1}^{p} \sigma_{Z_j}^2 = \sum_{j=1}^{p} var(Z_j)$$

$$= trace(R) = p$$

avec R la matrice de corrélation.

L'inertie renseigne donc sur la forme du nuage de points. Nous voulons réduire la dimension de l'espace des individus $\mathbb{R}^p$ en le projetant sur des sous-espaces $F \subset \mathbb{R}^p$ de faible dimension. Dans ce processus nous voulons garder la forme du nuage aussi intacte que possible et donc garder l'inertie aussi intacte que possible. Soit F un sous-espace vectoriel (passant donc par l'origine) de $\mathbb{R}^p$.

Définition 1.5.2. *L'inertie du nuage de points w_i par rapport à F est*

$$I_F = \frac{1}{n}\sum_{i=1}^{n} d^2(w_i, H_i)$$

avec $H_i = proj_F(w_i)$ la projection orthogonale de w_i sur F

Cette inertie mesure la proximité à F du nuage de points. I_F petit signifie que F passe le plus prés possible de tous les points du nuage en même temps. L'objectif de l'ACP est de projeter les points individus w_i sur un sous-espace qui passe le plus prés possible de tous les points en même temps, ce qui équivaut à chercher à minimiser I_F. Autrement dit on veut trouver F tel que I_F soit minimal.

On obtient donc un nuage de points projetés sur F. Dans cette projection on doit s'assurer de ne pas trop déformer la forme du nuage et donc de ne pas trop alterer l'inertie. Or par Pythagore, on a

$$Ow_i^2 = OH_i^2 + w_iH_i^2$$

et donc on a toujours

$$OH_i^2 \leq Ow_i^2$$

Donc pour ne pas trop changer l'inertie, on cherche à maximiser les $OH_i^2 = d^2(w_i, H_i^{'})$ avec $H_i^{'} = proj_{F^\perp}(w_i)$ la projection orthogonale de w_i sur l'orthogonal de F. On cherche donc à rendre maximale la quantité

$$\frac{1}{n}\sum_{i=1}^{n} OH_i^2 = \frac{1}{n}\sum_{i=1}^{n} d^2(w_i, H_i^{'}) = I_{F^\perp}$$

C'est le second objectif de l'ACP. Trouver F tel que $I_{F^\perp}$ soit maximum. En conclusion on cherche F tel que I_F soit minimal et $I_{F^\perp}$ soit maximal.

Définition 1.5.3. *$I_{F^\perp}$ est l'inertie expliquée par F (ou portée par F). I_F est l'inertie résiduelle ou restante.*

Proposition 1.5.1. *On a I_F minimal $\Longleftrightarrow$ $I_{F^\perp}$ maximal.*

Preuve. Par hypothèse on a

$$Ow_i^2 = w_iH_i^2 + w_iH_i^{'2} = OH_i^2 + OH_i^{'2}$$

et donc I_F minimal équivaut à $I_{F^\perp}$ maximal. □

Cette proposition veut dire que les deux objectifs de l'ACP sont atteints d'un seul coup. L'inertie résiduelle de F indique le degré d'étirement du nuage de points par rapport à F. L'inertie portée par F est l'inertie du nuage de points projetés sur F.

Si $\mathbb{R}^p = D_1 \oplus D_2 \oplus \ldots \oplus D_p$ avec $D_1, \ldots, D_p$ des droites passant par l'origine (des sous-espaces vectoriels de dimension 1), on obtient par le même raisonnement et une récurrence sur p une décomposition de l'inertie totale comme somme des inerties expliquées ou portées par chaque axe ou droite D_i

$$I_T = I_{D_1^\perp} + I_{D_2^\perp} + \ldots I_{D_p^\perp}$$

Exercice 1.5.2. *Montrer que si D est une droite avec $I_{D^\perp}$ maximal alors le sous-espace de dimension 2 avec $I_{F^\perp}$ maximal doit contenir D.*

On cherchera donc à trouver le premier axe ou droite d'inertie portée maximum, puis le deuxième axe d'inetie portée restante maximum, et ainsi de suite.

Soit F un sous-espace de $\mathbb{R}^p$ d'orthogonal $F^\perp$. On a

$$\mathbb{R}^p = F \oplus F^\perp$$

Autrement dit tout vecteur w de $\mathbb{R}^p$ s'écrit de manière unique

$$w = u + v$$

avec $u \in F$ et $v \in F^\perp$. La projection orthogonale p_F sur F est l'application linéaire

$$p_F : \mathbb{R}^p \longrightarrow \mathbb{R}$$

définie par

$$p_F(w) = u$$

Si $F = D$ est une droite (sous-espace vectoriel de dimension 1) engendrée par le vecteur a, on a

$$p_D(w) = \frac{\langle w, a\rangle}{\langle a, a\rangle} a$$

En particulier si $\|a\|^2 = \langle a, a\rangle = 1$ on a

$$p_D(w) = \langle w, a\rangle a$$

La proposition suivante est trés importante. Elle fait le lien entre l'inertie et la matrice de corrélation R.

Proposition 1.5.2. *Soit D une droite dans $\mathbb{R}^p$ engendrée par le vecteur a avec $\|a\| = 1$. Alors on a*

$$I_{D^\perp} = aRa' = \langle a, (Ra')'\rangle$$

Preuve. On a

$$I_{D^\perp} = \frac{1}{n}\sum_{i=1}^{n} d^2(w_i, H_i') = \frac{1}{n}\sum_{i=1}^{n} d^2(O, H_i) = \frac{1}{n}\sum_{i=1}^{n} \|H_i\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\langle w_i, a\rangle a\|^2 = \frac{1}{n}\sum_{i=1}^{n}\langle w_i, a\rangle^2 = \frac{1}{n}\sum_{i=1}^{n}\langle a, w_i\rangle\langle w_i, a\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n} a w_i^{'} w_i a^{'} = a(\frac{1}{n}\sum_{i=1}^{n} w_i^{'} w_i)a^{'} = a\frac{1}{n}Z^{'}Za^{'} = aRa^{'}$$

□

Le premier axe (droite) portant le maximum d'inertie est donc la droite D de vecteur directeur a avec $\|a\| = 1$ et rendant maximum $I_{D^{\perp}} = aRa^{'}$ avec R la matrice de corrélation des variables initiales X_j. On cherche donc à résoudre le problème d'optimisation

$$maxaRa^{'}$$

$$aa^{'} = 1$$

Ce problème est un problème d'optimisation avec contraintes. Si on pose

$$f(a) = f(a_1, a_2, \ldots, a_n) = aRa^{'}$$

et

$$g(a) = g(a_1, a_2, \ldots, a_n) = aa^{'} - 1 = 0$$

ce problème est celui de la recherche des points critiques de f avec la contrainte g. Par la méthode des multiplicateurs de Lagrange, on sait qu'il est équivalent au problème de trouver les points critiques de

$$L(a, \lambda) = f(a) - \lambda g(a)$$

sans contrainte. Une condition necessaire pour un point critique de L est

$$\nabla(L) = (\frac{\partial L}{\partial a_1}, \frac{\partial L}{\partial a_2}, \ldots, \frac{\partial L}{\partial a_n}, \frac{\partial L}{\partial \lambda}) = (0, 0, \ldots, 0)$$

ce qui équivaut à

$$\frac{\partial f}{\partial a_i} = \lambda \frac{\partial g}{\partial a_i}$$

pour i allant de 1 à n et

$$g(a) = 0$$

qui est précisement la contrainte d'origine. Ici on a

$$\frac{\partial f}{\partial a} = 2Ra$$

et

$$\frac{\partial g}{\partial a} = 2a$$

avec $\frac{\partial f}{\partial a} = (\frac{\partial f}{\partial a_1}, \ldots, \frac{\partial f}{\partial a_n})$ et $\frac{\partial g}{\partial a} = (\frac{\partial g}{\partial a_1}, \ldots, \frac{\partial g}{\partial a_n})$. Une condition nécessaire pour notre problème avec contrainte est donc que

$$Ra^{'} = \lambda a^{'}$$

Autrement dit λ est une valeur propre de R et a est le vecteur propre unitaire correspondant. On a

$$Ra^{'} = \lambda a^{'}$$

et

$$I_{D^{\perp}} = aRa^{'} = a\lambda a^{'} = \lambda a a^{'} = \lambda \langle a, a \rangle = \lambda$$

Comme $I_{D^{\perp}}$ doit être maximal, la condition nécessaire sera suffisante si λ est la plus grande valeur popre de la matrice R. Le premier axe qui porte le maximum d'inertie est donc la droite D de vecteur directeur a le vecteur propre unitaire ($\|a\| = 1$) qui correspond à la plus grande valeur propre λ_1 de la matrice de corrélation R.

Remarque 1.5.1. *Les matrices de corrélation sont des matrices carrées symétriques semi-définies positives ($aRa^{'} \geq 0$) de rang $q \leq p$. Elles ont donc les propriétés suivantes :*

1. *Elles ont p valeurs propres réelles dont q sont non nulles et toutes positives avec $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ et $\sum_{j=1}^{p} \lambda_j = p$.*

2. *Les vecteurs propres correspondants sont orthogonaux deux à deux.*

Le second axe portant l'inertie maximum restante est la droite engendrée par le vecteur propre unitaire a_2 correspondant à la seconde plus grande valeur propre λ_2 de la matrice de corrélation et ainsi de suite. On obtient donc p axes $D_1, D_2, \ldots, D_p$ engendrés par les vecteurs propres unitaires $a_1, a_2, \ldots, a_p$ correspondants aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matrice de corrélation R et on a

$$D_1 \perp D_2 \perp \ldots \perp D_p$$

$$a_1 \perp a_2 \perp \ldots \perp a_p$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

$$I_{D_1^\perp} \geq I_{D_2^\perp} \geq \ldots \geq I_{D_p^\perp}$$

Remarquer que les vecteurs propres $a_1, \ldots, a_p$ forment une nouvelle base de $\mathbb{R}^p$.

Définition 1.5.4. *Les axes $D_1, \ldots, D_p$ sont appelés les axes principaux ou factoriels. Les vecteurs $a_1, \ldots, a_p$ sont les facteurs principaux.*

On a vu que l'inertie portée ou expliquée par l'axe D_j est

$$I_{D_j^\perp} = \lambda_j$$

Définition 1.5.5. *Le pourcentage de l'inertie expliquée par l'axe D_j (contribution relative de cet axe à l'inertie) est*

$$\frac{\lambda_j}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_j}{p}$$

De même le pourcentage de l'inertie expliquée par le sous-espace $F_k = D_1 \oplus \ldots \oplus D_k$ pour $k \geq 1$ est

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_1 + \ldots + \lambda_k}{p}$$

On va donc chercher le sous-espace F_k dont le pourcentage est proche de 1 (le plus grand possible). Dans la pratique les deux premiers sous-espaces $F_1 = D_1$ ou $F_2 = D_1 \oplus D_2$ suffisent.

Exercice 1.5.3. *Soit le tableau de données*

$$X = \sqrt{10} \begin{pmatrix} 2 & 2 & 3 \\ 3 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 4 \\ 2 & 1 & 3 \end{pmatrix}$$

portant sur 5 individus de poids égaux $\frac{1}{5}$ et 3 variables. On veut faire une ACP centrée réduite (normée) sur ce tableau.

1. *Calculer le tableau centré réduit Z.*

2. *Calculer la matrice de corrélation R.*

3. *Determiner les valeurs propres de R.*

4. *Calculer les pourcentages en inertie des differents axes.*

1.6 Composantes principales

A chaque axe principal (ou factoriel) D_k est associée une nouvelle variable C_k appelée composante principale.

Définition 1.6.1. *La composante principale C_k est la variable dont la valeur $C_k(w_i)$ en l'individu w_i est la coordonnée du projeté H_i de w_i sur l'axe D_k*

$$C_k(w_i) = \langle H_i, a_k \rangle = c_{ik}$$

Comme $H_i = \langle w_i, a_k \rangle a_k$, on a bien $C_k(w_i) = \langle w_i, a_k \rangle$ et donc

$$c_{ik} = \langle w_i, a_k \rangle = \sum_{j=1}^{p} z_{ij} a_{kj}$$

En notation matricielle on a

$$C_k = Z a_k^{'}$$

si C_k est le vecteur colonne des c_{ik} avec Z le tableau centré réduit.

Exercice 1.6.1. *Montrer que $w_i = \sum_{k=1}^{p} c_{ik} a_k$. Autrement dit les c_{ik} sont les coordonnées de w_i dans la nouvelle base $\{a_1, \ldots, a_p\}$.*

Ces nouvelles variables C_k ont des propriétés interessantes. Elles sont par exemple toutes centrées. En effet on a

$$\overline{C_k} = \frac{1}{n} \sum_{i=1}^{n} c_{ik}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} z_{ij} a_{kj}$$

$$= \sum_{j=1}^{p} a_{kj} \frac{1}{n} \sum_{i=1}^{n} z_{ij} = 0$$

De plus on a

$$var(C_k) = \lambda_k$$

et les variables C_k sont décorrelées entre elles.

Exercice 1.6.2. *Montrer que*

$$var(C_k) = \lambda_k$$

et que

$$cov(C_k, C_l) = cor(C_k, C_l) = 0$$

Exercice 1.6.3. *On reprend les données de l'exercice* 1.5.3.

1. *Verifier que* $a_1 = \frac{1}{2}(\sqrt{2}, 1, -1), a_2 = \frac{1}{\sqrt{2}}(0, 1, 1)$ *sont les vecteurs propres correspondants aux valeurs propres* $\lambda_1 = 1 + \frac{\sqrt{2}}{2}$ *et* $\lambda_2 = 1$.
2. *Determiner les deux premières composantes principales* C_1 *et* C_2.
3. *Verifier qu'on a bien* $var(C_1) = \lambda_1$ *et* $var(C_2) = \lambda_2$ *et que* $cov(C_1, C_2) = cor(C_1, C_2) = 0$.

1.7 Nuage des p variables

Rappelons que l'espace $\mathbb{R}^n$ des variables est muni du produit scalaire de matrice (dans la base canonique) $M = \frac{1}{n} I_n$ avec I_n la matrice identité $n \times n$.

Définition 1.7.1. *L'inertie totale du nuage des variables* $Z_j, j = 1, \ldots, p$ *est*

$$I_T = \sum_{j=1}^{p} d^2(O, Z_j)$$

La variable $Z_j = (z_{1j}, \ldots, z_{nj})$ est vue ici comme un vecteur ligne. Autrement dit, on travaille avec le tableau Z' et non pas Z.

Proposition 1.7.1. *L'inertie des variables est égale à l'inertie des individus.*

Preuve. L'inertie des variables est

$$I = \sum_{j=1}^{p} d^2(O, Z_j) = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} z_{ij}^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} z_{ij}^2$$

qui est exactement l'inertie des individus. □

Rappelons que $d^2(O, Z_j) = \|Z_j\|^2 = var(Z_j) = 1$ et donc $I_T = p$ comme pour le nuage des individus. On procède de la même manière que dans le cas des individus pour trouver les axes factoriels des variables. Le premier axe E_1 de vecteur directeur unitaire (variable) b_1 doit passer le plus proche possible de tous les Z_j en même temps. Autrement dit les angles des Z_j avec b_1 doivent être aussi petits que possible. On sait que les cosinus de ces angles expriment les coefficients de corrélation entre variables. On doit donc chercher une nouvelle variable b_1 qui soit liée le plus possible aux variables $Z_1, \ldots, Z_j$ en même temps.

Comme pour les individus b_1 est le vecteur unitaire propre correspondant à la plus grande valeur propre α_1 de la matrice $\frac{1}{n} ZZ^{'}$ qui est une matrice $n \times n$ (pour les individus c'était $R = \frac{1}{n} Z^{'} Z$).

Proposition 1.7.2. *Les matrices $Z^{'}Z$ et $ZZ^{'}$ ont mêmes valeurs propres non nulles. En particulier elles ont le même rang.*

Preuve. Soit b un vecteur propre de $ZZ^{'}$ correspondant à la valeur propre $\alpha \neq 0$. Donc $ZZ^{'}b^{'} = \alpha b^{'}$ et donc $Z^{'}(ZZ^{'}b^{'}) = Z^{'}(\alpha b^{'}) = \alpha Z^{'}b^{'}$ et donc $(Z^{'}Z)Z^{'}b^{'} = \alpha Z^{'}b^{'}$. Le vecteur $a^{'} = Z^{'}b^{'}$ est donc vecteur propre de $Z^{'}Z$ pour la même valeur propre α de $ZZ^{'}$. Toute valeur propre non nulle de $ZZ^{'}$ est donc valeur propre de $Z^{'}Z$. Par symétrie toute valeur propre de $Z^{'}Z$ est aussi valeur propre de $ZZ^{'}$. On finit en remarquant que le rang d'une matrice carrée est le nombre de ses valeurs propres non nulles. □

Remarquer que la matrice $ZZ^{'}$ est beaucoup plus grande que $Z^{'}Z$ et qu'il vaut mieux travailler avec cette dernière. La proposition nous dit que c'est la même chose.

Le premier axe E_1 est donc la droite engendrée par le vecteur propre unitaire correspondant à la plus grande valeur propre λ_1 de $\frac{1}{n}ZZ^{'}$. Le second axe E_2 est la droite engendrée par b_2 le vecteur propre unitaire correspondant à la seconde plus grande valeur propre λ_2 et ainsi de suite. La proposition suivante exprime les composantes principales C_k en termes des nouvelles variables b_k

Proposition 1.7.3. *On a* $C_k = \sqrt{\lambda_k} b_k$.

Preuve. On a $(\frac{1}{n}ZZ^{'})C_k = \frac{1}{n}ZZ^{'}Za^{'} = Z(\frac{1}{n}Z^{'}Za_k^{'}) = Z(\lambda_k a_k^{'}) = \lambda_k Za_k^{'} = \lambda_k C_k$. C_k est donc vecteur propre de $\frac{1}{n}ZZ^{'}$ pour la même valeur propre λ_k que b_k. Il doit donc exister β tel que

$$C_k = \beta b_k$$

On sait que $\|b_k\| = 1$ et donc $\beta = \|C_k\|$. Mais

$$\|C_k\|^2 = \langle C_k, C_k \rangle$$

$$= \langle Za_k^{'}, Za_k^{'} \rangle = \frac{1}{n}(Za_k^{'})^{'}Za_k^{'}$$

$$= a_k \frac{1}{n} Z^{'}Za_k^{'} = \lambda_k a_k a_k^{'} = \lambda_k$$

Donc $\beta = \sqrt{\lambda_k}$ et

$$C_k = \sqrt{\lambda_k} b_k$$

□

La composante principale C_k est donc portée par l'axe E_k. L'exercice suivant permet de calculer les coefficients de covariance et de corrélation des composantes principales C_k avec les variables Z_j. Le suivant exprime les Z_j comme combinaison linéaire des C_k.

Exercice 1.7.1. *1. Montrer que* $cov(C_k, Z_j) = \lambda_k a_{kj}$.

2. Montrer que $cor(C_k, Z_j) = \sqrt{\lambda_k} a_{kj}$

Exercice 1.7.2. *Montrer que* $Z_j = \sum_{k=1}^{p} a_{kj} C_k$.

On sait que $var(Z_j) = 1$. Mais on a aussi

$$var(Z_j) = cov(Z_j, Z_j)$$

$$= cov(Z_j, \sum_{k=1}^{p} a_{kj} C_k) = \sum_{k=1}^{p} a_{kj} cov(Z_j, C_k)$$

$$= \sum_{k=1}^{p} \lambda_k a_{kj}^2 = \sum_{k=1}^{p} cor(Z_j, C_k)^2 = 1$$

et donc $cor(C_k, Z_j)^2 + cor(C_l, Z_j)^2 \leq 1$. Ceci veut dire que le point $(cor(C_k, Z_j), cor(C_l, Z_j))$ est à l'interieur du cercle de centre O et de rayon 1 dessiné sur le plan d'axes C_k et C_l.

Définition 1.7.2. *Le cercle des corrélations de deux composantes principales C_k, C_l est le cercle de centre O et de rayon 1 sur le plan des deux composantes dans lequel la variable Z_j est représentée par le point de coordonnées $(cor(C_k, Z_j), cor(C_l, Z_j))$.*

Remarquer que la projection de la variable Z_j sur l'axe E_k est

$$p_{E_k}(Z_j) = \langle Z_j, b_k \rangle b_k = cor(Z_j, C_k) b_k$$

puisque $C_k = \sqrt{\lambda_k} b_k$ et donc $\langle Z_j, b_k \rangle = \frac{1}{\sqrt{\lambda_k}} \langle Z_j, C_k \rangle = cor(Z_j, C_k)$. Le point $P_{Z_j} = (cor(Z_j, C_k), cor(Z_j, C_l))$ n'est donc autre que le projeté de Z_j sur le plan factoriel (de l'espace des variables) engendré par b_k et b_l (et donc aussi par C_k et C_l).

Exercice 1.7.3. *Representer les variables Z_1, Z_2 et Z_3 dans le cercle des corrélations des composantes C_1 et C_2 de l'exercice* 1.6.3.

1.8 Qualité de représentation

La formule

$$d^2(w_i, w_k) = d^2(H_i, H_k) + t^2$$

montre que si H_i et H_k sont éloignés sur l'axe de projection D alors w_i et w_k sont éloignés dans $\mathbb{R}^p$. L'inverse n'est pas vrai. Si H_i et H_k sont proches sur D, on ne peut pas conclure que w_i et w_k vont être proches dans l'espace. Il nous faut donc nous assurer de la qualité de representation des individus ou des variables sur le sous-espace de projection.

Définition 1.8.1. *La qualité de representation de l'individu w_i sur l'axe D_k est le rapport*

$$q_{D_k}(w_i) = \frac{OH_i^2}{Ow_i^2} = cos\theta_k^2$$

avec θ_k l'angle que fait w_i avec D_k

Donc plus l'angle entre w_i et D_k est petit (D_k passe le plus prés de w_i), plus la representation de l'individu sur l'axe est bonne. Si les projections de deux individus bien représentés sont proches alors les individus eux-mêmes sont proches dans la réalité. En conclusion deux individus bien représentés sur un sous-espace auront les mêmes distances avant et aprés la projection.

Remarque 1.8.1.

1. *Si par exemple $F = D_1 + D_2$, la qualité de representation de w_i sur F est la somme des qualités de représentation de w_i sur D_1 et D_2*
$$q_F(w_i) = q_{D_1}(w_i) + q_{D_2}(w_i)$$

2. *Tout individu proche de $O = G$ sera bien representé dans tout sous-espace. En effet dans ce cas Ow_i^2 sera trés petit et donc $cos^2\theta_k$ sera trés grand.*

3. *Aprés la projection sur un sous-espace F les points trop proches de l'origine seront mal représentés (OH_i^2 petit). Les points loin de l'origine seront bien représentés(OH_i^2 grand).*

On fait la même chose pour les variables. La qualité de représentation de la variable Z_j sur l'axe E_k est donnée par

$$q_{E_k}(Z_j) = cos^2\theta_k$$

avec θ_k l'angle entre OZ_j et l'axe E_k. Cette qualité de représentation est là aussi additive. Si $E = E_k \oplus E_l$, alors on a

$$q_E(Z_j) = q_{E_k}(Z_j) + q_{E_l}(Z_j)$$

Rappelons que l'axe E_k porte la composante principale C_k et donc θ_k est aussi l'angle entre Z_j et C_k. On sait que

$$cos\theta_k = cor(Z_j, C_k)$$

La qualité de représentation de la variable Z_j sur le plan d'axes C_k et C_l peut donc être vue sur le cercle de corrélation. La variable Z_j est bien représentée sur le plan d'axes C_k et C_l si son point $(cor(Z_j, C_k), cor(Z_j, C_l))$ est proche du bord du cercle de corrélation. Dans un cercle donné on traitera uniquement les variables bien représentées.

Exercice 1.8.1. *Le tableau de données suivant mesure la tension artérielle diastolique et systolique et le taux de cholesterol de* 6 *patients (Ali, Warda, Farid, walid, Loubna, Nahla).*

$$X = \begin{pmatrix} 90 & 140 & 6.0 \\ 60 & 85 & 5.9 \\ 75 & 135 & 6.1 \\ 70 & 145 & 5.8 \\ 85 & 130 & 5.4 \\ 70 & 145 & 5.0 \end{pmatrix}$$

1. *Calculer le tableau centré réduit Z et la matrice de corrélation R.*

2. *Sachant que les valeurs propres de R sont $\lambda_1 = 1.58, \lambda_2 = 1.05$ et $\lambda_3 = 0.37$, vérifier que $a_1 = (0.641, 0.72, -0.265)$ et $a_2 = (0.4433, -0.0652, 0.894)$ sont les vecteurs propres unitaires correspondants aux deux premières valeurs propres.*

3. *Determiner les pourcentages en inertie des différents axes factoriels.*

4. *Calculer les deux premières composantes principales et dessiner les projections des* 6 *points individus sur le plan des deux premiers axes factoriels.*

5. *Calculer la qualité de représentation de chaque individu sur les deux premiers axes factoriels et sur le plan des deux premiers axes factoriels.*

6. *Calculer les qualités de représentation des* 3 *variables et les représenter dans le cercle des corrélations des deux premières composantes principales.*

1.9 Contribution

On peut aussi parler de la contribution d'un individu ou d'une variable à un axe. La contribution de l'individu w_i à l'axe D_k est

$$Ct_{D_k}(w_i) = \frac{OH_i^2}{\lambda_k}$$

avec H_i le projeté de w_i sur D_k. La contribution est d'autant plus grande que H_i est à l'extrémité de l'axe. Cette quantité permet aussi de detecter les points isolés dans la direction de l'axe. Si un individu contribue trés fortement à un axe, il est preferable de supprimer cet axe de l'analyse. La contribution de l'individu w_i au plan $F = D_k \oplus D_l$ est

$$Ct_F(w_i) = \frac{OH_i^2}{\lambda_k + \lambda_l}$$

avec H_i le projeté de w_i sur F. De même la contribution de la variable Z_j à l'axe E_k (qui rappelons le porte la composante principale C_k) est

$$Ct_{E_k}(Z_j) = \frac{cor^2(Z_j, C_k)}{\lambda_k}$$

Si on veut c'est la contribution de la variable Z_j à la définition de la nouvelle variable C_k. Cette dernière est définie par l'ensemble des variables initiales qui lui sont le plus correlées.

Remarquer que la contribution d'un individu ou une variable peut se lire directement sur les graphiques. Un individu va contribuer d'autant plus à un axe que sa position est à l'extrémité de cet axe. Une variable va fortement contribuer à une composante principale si l'angle qu'elle fait avec la composante est petit.

Exercice 1.9.1. *On reprend les données de l'exercice* 1.8.1. *Calculer les contributions des individus et des variables aux différents axes.*

1.10 Interpretation

On peut résumer les étapes d'une ACP en quelques points simples :

1. Centrer et réduire les données pour passer du tableau X au tableau Z.

2. Calculer la matrice de corrélation $R = \frac{1}{n} Z^{'} Z$.

3. Calculer les valeurs propres λ_k et les vecteurs propres unitaires correspondants a_k.

4. Calculer les pourcentages d'inertie et décider du nombre d'axes à retenir.

5. Calculer les composantes principales C_k.

6. Calculer les qualités de représentation des individus et des variables et leur contribution aux différents axes.

7. Dessiner les individus sur les axes ou les plans factoriels.

8. Tracer les cercles de corélation.

9. Interpreter.

La première composante est la nouvelle variable qui a la plus grande variance ($var(C_1) = \lambda_1$). C'est elle qui va le mieux résumer et expliquer les données de départ. Comme elle est centrée cela veut dire que les (carrés des) valeurs qu'elle prend en les divers individus sont les plus grandes parmi toutes les nouvelles variables (les composantes principales). Si par exemple les variables initiales étaient la quantité de produits vendus, C_1 serait la variable représentant le produit le plus vendu. C_2 serait la variable representant le second produit le plus vendu et ainsi de suite.

L'interpretation d'une ACP se fait de la manière suivante. On commence d'abord par tracer le cercle des deux premières composantes principales C_1 et C_2 et on repère quelles sont les variables initiales qui sont fortement correlées avec une des deux composantes en examinant simplement les angles qu'elles (ou plutôt leurs projections) font avec les composantes. On ne traitera bien sûr que les variables qui sont bien représentées sur ce cercle. Les variables fortement corrélés avec C_1 vont fortement contribuer à la fabrication et la définition de la composante principale à laquelle on essaiera de donner un nom adequat. On fait la même chose pour la deuxième composante principale C_2. Si une variable n'est pas bien représentée dans ce premier cercle on fera intervenir la troisième composante principale C_3 et on essaiera les deux cercles de corrélation C_1C_3 ou C_2C_3 pour voir dans lequel la représentation est bonne

et ainsi de suite. Remarquer comment on est passé à chaque fois de plusieurs variables à uniquement deux variables explicatives.

On fait intervenir ensuite les points individus (bien représentés). Il faut comprendre que l'espace des individus ($\mathbb{R}^p$) et l'espace des variables ($\mathbb{R}^n$) sont des espaces complètement différents. Les sous-espaces de projections

$$F = D_k \oplus D_l$$

et

$$E = E_k \oplus E_l$$

bien que de dimension 2 tous les deux ne contiennent pas les mêmes points ou vecteurs. Dans F les points ont p coordonnées et dans E les vecteurs ont n coordonnées (Il y a isomorphisme à cause de la même dimension mais il n y a pas égalité). On ne peut donc normalement pas représenter les individus et les variables sur le même graphique. Mais remarquer qu'on a

$$p_{D_k}(w_i) = H_i = \langle w_i, a_k \rangle a_k = c_{ik} a_k$$

et que

$$c_{ik} = C_k(w_i)$$

Autrement dit la position du projeté H_i de w_i sur l'axe D_k est complètement determinée par la valeur de la nouvelle variable C_k (la composante principale) en l'individu w_i et donc la position du projeté H_i de w_i sur le premier plan factoriel $F = D_1 \oplus D_2$ est complètemet determinée par les valeurs $C_1(w_i)$ et $C_2(w_i)$ des deux premières composantes principales en ce point.

On va donc juxtaposer le graphique du cercle de corrélation des deux premières composantes princpales au graphique des projetés H_i des w_i sur le plan factoriel $F = D_1 \oplus D_2$ et faire une étude simultanée des individus et variables. On lira ce double graphique de gauche à droite et de bas en haut. Plus on va vers la droite et plus on parcourt les individus qui ont de fortes valeurs pour la première composante principale et donc qui y contribuent le plus ainsi qu' à toutes les variables initiales correlées avec cette dernière (et qui ont donc servi à la définir). Aller de bas en haut permet de faire la même chose pour la seconde composante principale et les variables initiales qui sont avec elles. On pourra ainsi determiner les différences ou ressemblances entre individus ou groupes d'individus à travers leurs contributions aux composantes principales (et par ricochet aux variables initiales les définissant).

Si une variable n'est pas bien représentée dans le cercle C_1C_2 et qu'elle l'est par exemple dans le cercle C_1C_3 (avec une corrélation par exemple avec C_3) il faudra juxtaposer le gaphique de ce cercle de corrélations avec le graphique des points projetés sur le plan factoriel $F = D_1 \oplus D_3$ et ainsi de suite.

Exercice 1.10.1. *Le tableau suivant représente la consommation en aliments de* 8 *catégories socio-professionnelles : les exploitants agricoles (AGRI), les salariés agricoles (SAAG), les professionels indépendants (PRIN), les cadres supérieurs (CSUP), les cadres moyens (CMOY), les employés (EMPL), les ouvries (OUVR) et les inactifs (INAC). Les aliments sont le pain, les croissants, l'eau, le coca, la pomme de terre, les pattes, la viande et les plats préparés.*

$$X = \begin{pmatrix} 167 & 1 & 163 & 23 & 41 & 8 & 6 & 6 \\ 162 & 2 & 141 & 12 & 40 & 12 & 4 & 15 \\ 119 & 6 & 69 & 56 & 39 & 5 & 13 & 41 \\ 87 & 11 & 63 & 111 & 27 & 3 & 18 & 39 \\ 103 & 5 & 68 & 77 & 32 & 4 & 11 & 30 \\ 111 & 4 & 72 & 66 & 34 & 6 & 10 & 28 \\ 130 & 3 & 76 & 52 & 43 & 7 & 7 & 16 \\ 138 & 7 & 117 & 74 & 53 & 8 & 12 & 20 \end{pmatrix}$$

Effectuer une ACP normée sur ces données et interpreter les résultats.

Chapitre 2

Analyse factorielle des correspondances

2.1 Tableau de contingence

L'analyse factorielle des correspondances est une analyse factorielle développée par J.P. Benzecri dans les années 1970. Cette analyse permet l'étude des liaisons (ou correspondances) entre deux variables qualitatives nominales via leur tableau de contingence ou encore de fréquences (voir plus loin). C'est un tableau de n lignes et m colonnes. Les lignes représentent les modalités de la première variable et les colonnes ceux de la seconde variable. Les deux variables qualitatives V_1 et V_2 sont définies sur la même population de N individus. A l'intersection de la ligne i et de la colonne j on retrouve le nombre k_{ij} qui donne le nombre d'individus ayant à la fois la modalité i de V_1 et la modalité j de V_2. On a donc

$$\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij} = N$$

Dans ce tableau les lignes et les colonnes jouent un rôle symétrique.

Exemple 2.1.1. *Soient les deux variables V_1 = couleur des yeux et V_2 = couleur des cheveux portant sur une population de 592 femmes. V_1 a quatre modalités qui sont marron, noisette, vert et bleu. V_2 a quatre modalités qui sont brun, chatain, roux et blond. On a donc $N = 592, n = 4$ et $m = 4$. Le tableau des modalités est*

	brun	*chatain*	*roux*	*blond*
marron	*68*	*119*	*26*	*7*
noisette	*15*	*54*	*14*	*10*
vert	*5*	*29*	*14*	*16*
bleu	*20*	*84*	*17*	*94*

Par exemple il y a 29 femmes ayant les cheveux chatains et les yeux verts. A partir de ce tableau de contingence on fabrique le tableau des fréquences relatives X.

Définition 2.1.1. *La fréquence relative des deux variables est*

$$f_{ij} = \frac{k_{ij}}{N}$$

Les marges (ou fréquences marginales) sont les

$$f_{i\bullet} = \sum_{j=1}^{m} f_{ij}$$

et les

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

On a donc

$$\sum_{i=1}^{n} f_{i\bullet} = \sum_{j=1}^{m} f_{\bullet j} = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} = 1$$

En quelque sorte $f_{i\bullet}$ est la probabilité d'obtenir la modalié i de la variable V_1

$$f_{i\bullet} = P(V_1 = i)$$

et $f_{\bullet j}$ est la probabilité d'obtenir la modalité j de la variable V_2

$$f_{\bullet j} = P(V_2 = j)$$

f_{ij} est la probabilité d'obtenir à la fois la modalité i de V_1 et la modalité j de V_2

$$f_{ij} = P(V_1 = i, V_2 = j)$$

Le tableau des fréquences X est le tableau des f_{ij}. La somme de la ligne i de ce tableau est $f_{i\bullet}$ et la somme de la colonne j est $f_{\bullet j}$. Le tableau des fréquences de l'exemple précedant est

	brun	chatain	roux	blond	marge $f_{i\bullet}$
marron	1.14	2.01	0.43	0.11	0.371
noisette	0.25	0.91	0.23	0.16	0.157
vert	0.08	0.48	0.23	0.27	0.108
bleu	0.33	0.141	0.28	0.158	0.363
marge $f_{\bullet j}$	0.182	0.483	0.119	0.214	1

Ce tableau est le plus souvent représenté en pourcentage en multipliant tous les nombres par 100.

Les tableaux de contingence ou de fréquences peuvent être vus comme des ensembles de lignes ou de colonnes. Le but est de savoir quelles sont les lignes ou les colonnes qui se ressemblent et quelles sont celles celles qui s'opposent, ou encore de savoir si il ya des groupes de lignes ou de colonnes homogènes. On veut étudier la relation entre les lignes et les colonnes et le degré de dependance des deux variables. Dans l'exemple précedent on cherche à savoir s'il y a dépendance ou non entre la couleur des yeux et la couleur des cheveux. Cet exemple est simple mais il faut imaginer des tableaux de dimension beaucoup plus grande.

Définition 2.1.2. *Les deux variables V_1 et V_2 sont indépendantes si*

$$P(V_1 = i, V_2 = j) = P(V_1 = i)P(V_2 = j)$$

ou encore si

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

Elles sont dites liées si elles ne sont pas indépendantes.

Pour deux variables indépendantes V_1 et V_2, on a donc

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

pour i allant de 1 à n. Cela veut dire que pour deux variables indépendantes les lignes du tableau des fréquences relatives sont toutes proportionnelles. le même argument appliqué aux colonnes montre que pour deux variables

indépendantes toutes les colonnes du tableau de fréquences relatives sont proportionnelles.

Si $f_{ij} > f_{i\bullet}f_{\bullet j}$ les modalités i et j s'associent plus que sous l'hypothèse d'indépendance. On dira qu'elles s'attirent. Si $f_{ij} < f_{i\bullet}f_{\bullet j}$, les modalités i et j s'associent moins que sous l'hypothèse d'indépendance. On dira qu'il y a repulsion entre les deux modalités.

Exercice 2.1.1. *Dans l'exemple de la couleur des yeux et des cheveux, calculer le tableau des féquences correspondant à l'indépendance des deux variables et comparer au tableau des fréquences réel.*

Exercice 2.1.2. *Soit le tableau de contingence suivant portant sur $N = 3784$ individus de deux variables V_1= catégorie socio-professionnelle des parents et V_2= choix de la filière d'études à l'université des enfants. V_1 a cinq modalités : exploitant agricole, patron, cadre superieur, employé et ouvrier. V_2 a quatre modalités : droit, science, médecine et technologie.*

	droit	*science*	*medecine*	*technologie*
exp.agricole	*80*	*99*	*65*	*58*
patron	*168*	*137*	*208*	*62*
cadre.sup	*470*	*400*	*876*	*79*
employé	*145*	*133*	*135*	*54*
ouvrier	*166*	*193*	*127*	*129*

1. *Calculer le tableau des fréquences relatives.*
2. *Les deux variables sont-elles indépendantes ?*

2.2 Transformation des données

Les données du tableau des fréquences relatives X sont normalisées en divisant la ligne i par $f_{i\bullet}$ et la colonne j par $f_{\bullet j}$. On obtient des profils lignes $\dfrac{f_{ij}}{f_{i\bullet}}$ et des profils colonnes $\dfrac{f_{ij}}{f_{\bullet j}}$ et donc deux tableaux de données, celui des profils lignes X_L et celui des profils colonnes X_C

	brun	chatain	roux	blond
marron	30.9	54.0	11.8	3.1
noisette	16.1	58.0	15.0	10.7
vert	7.8	45.3	21.8	25.0
bleu	9.3	39.0	7.9	43.7

	brun	chatain	roux	blond
marron	62.9	41.6	36.6	5.5
noisette	13.8	18.8	19.7	7.8
vert	4.6	10.1	19.7	12.5
bleu	18.5	29.3	23.9	74.0

Remarquer que $\frac{f_{ij}}{f_{i\bullet}}$ représente la probabilité d'avoir la modalité j sachant qu'on a la modalité i. C'est donc une probabilité conditionnelle. De même $\frac{f_{ij}}{f_{\bullet j}}$ est la probabilité d'obtenir la modalité i sachant qu'on a la modalité j. Dans l'exemple de la couleur des yeux et des cheveux, on sait que 11 femmes sur 100 ont les yeux marrons et les cheveux bruns. On a 31 chances sur 100 que les yeux marrons donnent des cheveux bruns et on a 63 chances sur 100 que les cheveux bruns donnent des yeux marrons.

Exercice 2.2.1. *On reprend les données de l'exercice* 2.1.2. *Calculer les tableaux des profils lignes X_L et profils colonnes X_C*

En notation matricielle on a

$$X_L = D_L^{-1} X$$

et

$$X_C = D_C^{-1} X'$$

Les profils lignes définissent un nuage de points dans $\mathbb{R}^m$. Chaque point x_{iL} est affecté du poids $f_{i\bullet}$. La matrice des poids est donc la matrice diagonale D_L des $f_{i\bullet}$. Le barycentre est le point G_L dont les coordonnées sont

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

et donc

$$G_L = (f_{\bullet 1}, \ldots, f_{\bullet m})$$

Le barycentre s'interprète comme un profil ligne moyen.

Les profils colonnes définissent un nuage de points dans $\mathbb{R}^n$. Chaque point x_{jC} est affecté du poids $f_{\bullet j}$. La matrice des poids est donc la matrice diagonale D_C des $f_{\bullet j}$. Le barycentre est le point G_C de coordonnées

$$\sum_{j=1}^{m} f_{\bullet j} \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$$

et on a donc

$$G_C = (f_{1\bullet}, \ldots, f_{n\bullet})$$

Le barycentre s'interprète comme un profil colonne moyen.

Exercice 2.2.2. *Montrer que si les deux variables X et Y sont indépendantes alors tous les profils lignes sont égaux au barycentre G_L et tous les profils colonnes sont égaux au barycentre G_C.*

En notation matricielle on a

$$G_L = (X_L^{'} D_L 1_n)^{'} = (X^{'} 1_n)^{'}$$

et

$$G_C = (X_C^{'} D_C 1_q)^{'} = (X 1_m)^{'}$$

Exercice 2.2.3. *Demontrer ces formules. Ne pas oublier que nos vecteurs sont en ligne.*

Les deux nuages sont centrés autour des barycentres. Le point x_{iL} devient

$$y_{iL} = (\frac{f_{i1}}{f_{i\bullet}} - f_{\bullet 1}, \ldots, \frac{f_{im}}{f_{i\bullet}} - f_{\bullet m})$$

et le point x_{jC} devient

$$y_{jC} = (\frac{f_{1j}}{f_{\bullet j}} - f_{1\bullet}, \ldots, \frac{f_{nj}}{f_{\bullet j}} - f_{n\bullet})$$

On notera Y_L le tableau des profils lignes centrés et Y_C le tableau des profils colonnes centrés. Géometriquement le centrage revient à faire une translation de l'origine des coordonnées O vers le barycentre G_L ou G_C. En notation matricielle on a

$$Y_L = X_L - 1_n G_L$$

et

$$Y_C = X_C - 1_m G_C$$

Exercice 2.2.4. *Calculer les tableaux centrés de l'exemple de la couleur des yeux et des cheveux et celui de l'exercice* 2.1.2

Pour comparer les lignes et le colonnes des tableaux Y_L et Y_C respectivement, on a besoin d'une distance dans $\mathbb{R}^m$ et une autre dans $\mathbb{R}^n$. Elles seront données par des produits scalaires sur ces espaces. Dans $\mathbb{R}^m$ le produit scalaire a pour matrice M_L (dans la base canonique) la matrice diagonale des inverses des $f_{\bullet j}$ et dans $\mathbb{R}^n$, la matrice M_C du produit scalaire est la matrice diagonale des inveres des $f_{i\bullet}$. On a donc

$$M_L = D_C^{-1}$$

et

$$M_C = D_L^{-1}$$

Exercice 2.2.5. *Montrer que* $\|G_L\| = 1$ *et* $\|G_C\| = 1$.

Remarquer que si les deux variables sont indépendantes, alors tous les profils lignes sont égaux au barycentre G_L et tous les profils colonnes sont égaux au barycentre G_C. Les distances entre profils lignes et profils colonnes vont donc mesurer en quelque sorte le degré de liaison entre les deux variables.

2.3 Inertie

L'inertie du nuage de points profils lignes centrés va mesurer leur dispersion par rapport à leur centre de gravité qui est l'origine $O = G_L$ et aussi le degré de liaison entre les deux variables. De même l'inertie du nuage de points profils colonnes centrés va mesurer leur dispersion par rapport à leur barycentre qui est l'origine $O = G_C$.

Définition 2.3.1. *L'inertie des profils lignes du tableau* Y_L *est*

$$I_L = \sum_{i=1}^{n} f_{i\bullet} d^2(O, i)$$

De même l'inertie des profils colonnes du tableau Y_C *est*

$$I_C = \sum_{j=1}^{m} f_{\bullet j} d^2(O, j)$$

Remarquer la présence des poids $f_{i\bullet}$ des lignes et les poids $f_{\bullet j}$ des colonnes. Au chapitre précedant les poids étaient $p_i = \dfrac{1}{n}$.

Exercice 2.3.1. *1. Montrer que $I_L = I_C$.*

2. Calculer l'inertie des nuages de points de l'exemple de la couleur des cheveux et des yeux et de l'exercice 2.1.2.

L'AFC a pour objectif de résumer les liaisons entre les deux variables par un processus de réduction de la dimension et de visualisation graphique. En ce sens ce n'est autre qu'une double ACP, une portant sur le tableau des profils lignes Y_L et l'autre portant sur le tableau des profils colonnes Y_C. La démarche est donc la même que dans le premier chapitre . Il s'agit de trouver le meilleur sous-espace F de $\mathbb{R}^m$ ou $\mathbb{R}^n$ qui conserve le maximum d'inertie possible et sur lequel on va projeter les données de départ. Traitons le cas des profils lignes. La même approche se fera pour les profils colonnes.

Pour effectuer une ACP sur le tableau des profils lignes Y_L, on va commencer par trouver le premier axe factoriel E de vecteur directeur unitaire u. Ce premier axe est celui pour lequel $I_{E^\perp}$ est maximal. Rappelons qu'on a

$$I_{E^\perp} = \sum_{i=1}^{n} f_{i\bullet} d^2(y_i, H_i^{'})$$

avec $H_i^{'}$ le projeté du profil ligne y_i sur $E^\perp$. Le projeté H_i de y_i sur E est

$$H_i = \langle y_i, u \rangle u$$

et on a donc la relation

$$y_i - H_i^{'} = H_i$$

Proposition 2.3.1. *On a $I_{E^\perp} = uM_L V_L M_L u^{'}$ avec $V_L = Y_L^{'} D_L Y_L$.*

Preuve. On a $d^2(y_i, H_i^{'}) = \|y_i - H_i^{'}\|^2 = \|\langle y_i, u\rangle u\|^2 = \langle y_i, u\rangle^2 = \langle u, y_i\rangle\langle y_i, u\rangle = (uM_L y_i^{'})(y_i M_L u^{'}) = uM_L y_i^{'} y_i M_L u^{'}$ et donc

$$I_{E^\perp} = \sum_{i=1}^{n} f_{i\bullet} uM_L y_i^{'} y_i M_L u^{'} = uM_L (\sum_{i=1}^{n} f_{i\bullet} y_i^{'} y_i) M_L u^{'} = uM_L V_L M_L u^{'}$$

□

On cherche donc u avec

$$uM_L V_L M_L u^{'}$$

maximum et

$$uM_Lu^{'} = 1$$

C'est un problème de recherche d'un maximum avec contraintes. On le résoud par la méthode des multiplicateurs de Lagrange comme dans le premier chapitre. Rappelons que si le problème est $uAu^{'}$ maximum et $uMu^{'} = 1$, on forme $L(u, \lambda) = uAu^{'} - \lambda(uMu^{'} - 1)$. Le probl'eme devient ainsi un problème d'optimisation sans contraintes et consiste en la recherche des points critiques de L. Une condition necessaire pour un point critique de L est $\frac{\partial L}{\partial u} = 0$ et $\frac{\partial L}{\partial \lambda} = 0$ ce qui équivaut à $Au' = \lambda Mu^{'}$ et $uMu^{'} = 1$ respectivement. On a donc $M^{-1}Au^{'} = \lambda u^{'}$ et $\lambda = uAu^{'}$ qui est la quantité à maximiser. La condition necessaire est donc que λ est valeur propre de $R = M^{-1}A$ et cette condition sera suffisante si λ est la plus grande valeur propre de R.

Définition 2.3.2. *$R = M^{-1}A$ est la matrice d'inertie. C'est la matrice dont on cherche les valeurs propres et les vecteurs propres.*

Pour les profils lignes centrés on a

$$A = M_L V_L M_L$$

et

$$M = M_L$$

Donc la matrice d'inertie des profils lignes centrés est

$$R_L = M_L^{-1}A = V_L M_L$$

Le premier axe principal sera donc engendré par le vecteur propre unitaire correspondant à la plus grande valeur propre de la matrice d'inertie $R_L = V_L M_L$. Le second axe principal sera engendré par le vecteur propre unitaire correspondant à la seconde valeur propre de R_L et ainsi de suite. De manière analogue la matrice $R_C = V_C M_C$ est la matrice d'inertie des profils colonnes centrés avec $V_C = Y_C^{'} D_C Y_C$. Les vecteurs propres unitaires correspondant aux valeurs propres de R_C vont donner les axes factoriels de l'ACP sur les profils colonnes centrés.

Remarque 2.3.1. *1. Au premier chapitre la matrice d'inertie du tableau centré réduit Z était la matrice des corrélations $R = \frac{1}{n}Z^{'}Z = Z^{'}D_pZ$*

avec D_p la matrice des poids $p_i = \frac{1}{n}$. La matrice du produit scalaire est la matrice identié. Pour le tableau centré Y, la matrice d'inertie est la matrice de covariance $V = \frac{1}{n}Y'Y = Y'D_pY$ avec l'identité comme matrice du produit scalaire ($A = R$ et $M = Id$ pour le tableau centré réduit Z et $A = V, M = Id$ pour le tableau centré Y).

2. *les barycentres G_L et G_C sont vecteurs propres unitaires de R_L et R_C pour la valeur propre $\lambda = 0$. Ils determinent donc chacun un axe factoriel inutile qu'on élimine de l'étude (contribution à l'inertie nulle).*

3. *R_L est une matrice $m \times m$ et R_C est une matrice $n \times n$. Ces matrices ont même rang qui est le nombre de leurs valeurs propres non nulles. On sait par la seconde remarque que $\lambda = 0$ est une valeur propre des deux matrices. Le rang de R_L et R_C est donc au plus $min(m-1, n-1)$. On cherchera donc les valeurs propres de la plus petite des deux matrices.*

Exercice 2.3.2. *Montrer que $V_L = X_L'D_LX_L - G_L'G_L$ et $V_C = X_C'D_CX_C - G_C'G_C =$. En déduire que $R_LG_L' = 0 = 0.G_L'$ et $R_CG_C' = 0 = 0.G_C'$.*

Les profils lignes sont dans l'hyperplan

$$W_L = \{(x_1, \ldots, x_m) \in \mathbb{R}^m / x_1 + \ldots + x_m = 1\}$$

De même les profils colonnes sont dans l'hyperplan

$$W_C = \{(x_1, \ldots, x_n) \in \mathbb{R}^n / x_1 + \ldots + x_n = 1\}$$

Remarquer que le barycentre $G_L \in W_L$ et le barycentre $G_C \in W_C$. W_L et W_C ne sont pas des sous-espaces vectoriels de $\mathbb{R}^m$ et $\mathbb{R}^n$ (ils ne passent pas par l'origine $O = (0, \ldots, 0)$). Ce sont ce qu'on appelle des espaces affines. Ils vont devenir des sous-espaces vectoriels de dimension $m-1$ et $n-1$ respectivement dés qu'on choisit un point origine O_L dans W_L et un point origine O_C dans W_C. Le choix le plus simple et le plus naturel est de choisir O_L comme le projeté orthogonal de O sur W_L et de choisir O_C comme le projeté orthogonal de O sur W_C. Mais on a

Proposition 2.3.2. *Le projeté orthogonal de O sur W_L est G_L et le projeté orthogonal de O sur W_C est G_C.*

Preuve. Il s'agit de montrer que pour tout $P = (x_1, \ldots, x_m) \in W_L$, on

$$\langle \overrightarrow{OG_L}, \overrightarrow{G_L P} \rangle = 0$$

Mais on a

$$\langle \overrightarrow{OG_L}, \overrightarrow{G_L P} \rangle = \overrightarrow{G_L P} M_L \overrightarrow{OG_L}' = (P - G_L) M_L (G - O)'$$

$$= (P - G_L) M_L G_L' = (P - G_L) D_C^{-1} G_L' = (P - G_L) 1_m = P 1_m - G_L 1_m$$

$$= 1 - 1 = 0$$

On fait la même chose pour G_C. □

Autrement dit les points profils lignes et profils colonnes sont déja dans les sous-espaces W_L et W_C de dimension $m-1$ et $n-1$ (il y a déja une réduction de la dimension) et les origines de ces espaces sont le barycentre G_L et le barycentre G_C. On peut donc complètement oublier les espaces $\mathbb{R}^m$ et $\mathbb{R}^n$ et considerer que les nuages de points sont dans $W_L = \mathbb{R}^{m-1}$ et $W_C = \mathbb{R}^{n-1}$. Dans ces nouveaux espaces les points lignes et les points colonnes sont déja centrés. Il est donc possible de travailler directement avec les tableaux X_L et X_C sans passer par les tableaux Y_L et Y_C. Ceci est en accord avec le fait que les axes $\overrightarrow{OG_L}$ et $\overrightarrow{OG_C}$ perpendiculaires à W_L et W_C n'apportent aucune contribution à l'inertie (G_L et G_C sont vecteurs propres de R_L et R_C pour la valeur propre $\lambda = 0$). Toute l'information sur les points est donc contenue dans W_L et W_C. Il faut juste faire attention à la définition de l'inertie qui est par rapport aux deux barycentres

$$I_L = \sum_{i=1}^{n} f_{i\bullet} d^2(x_{iL}, G_L)$$

et

$$I_C = \sum_{j=1}^{m} f_{\bullet j} d^2(x_{jC}, G_C)$$

avec x_{iL} et x_{jC} les lignes des tableaux X_L et X_C. On a vu que le premier axe factoriel E est engendré par le vecteur unitaire u qui vérifie $u M_L V_L M_L u'$ maximum. Mais on a

$$u M_L V_L M_L u' = u M_L (X_L' D_L X_L - G_L' G_L) M_L u'$$

$$= uM_L X_L^{'} D_L X_L M_L u^{'} - uM_L G_L^{'} G_L M_L u^{'}$$
$$= uM_L X_L^{'} D_L X_L M_L u^{'}$$

car $G_L M_L u^{'} = 0$ puisque G_L et u sont orthogonaux en tant que vecteurs propres de $V_L M_L$ correspondant à deux valeurs propres différentes. Si on travaille avec des données non centrées et donc avec le tableau X_L au lieu de Y_L alors on a $A = M_L X_L^{'} D_L X M_L$ et donc la matrice d'inertie est

$$S_L = M_L^{-1} A = X_L^{'} D_L X_L M_L = X_L^{'} D_L X_L D_C^{-1}$$
$$= X^{'} D_L^{-1} D_L D_L^{-1} X D_C^{-1}$$
$$= X^{'} D_L^{-1} X D_C^{-1}$$
$$= X_L^{'} X_C^{'}$$

qui est une matrice $m \times m$. En faisant la même choses pour les profils colonnes la matrice d'inertie quand on travaille avec X_C au lieu de Y_C est

$$S_C = X D_C^{-1} X^{'} D_L^{-1} = X_C^{'} X_L^{'}$$

qui est une matrice $n \times n$.

Proposition 2.3.3. *1. Les matrices R_L et S_L ont les mêmes vecteurs propres pour les mêmes valeurs propres sauf que $G_L^{'}$ est vecteur propre de S_L pour la valeur propre $\lambda = 1$ (alors qu'il est vecteur propre de R_L pour la valeur propre $\lambda = 0$).*

2. Les matrices R_C et S_C ont les mêmes vecteurs propres pour les mêmes valeurs propres sauf $G_C^{'}$ est vecteur propre de S_C pour la valeur propre $\lambda = 1$ (alors qu'il est vecteur propre de R_L pour la valeur propre $\lambda = 0$).

Preuve. Montrons la première propriété. La seconde est analogue. On sait que $V_L M_L G_L^{'} = 0 G_L^{'}$. Comme $V_L = X_L^{'} D_L X_L - G_L^{'} G_L$ on a $(X_L^{'} D_L X_L M_L - G_L^{'} G_L) M_L G_L^{'} = 0$. Donc $X_L^{'} D_L X_L M_L G_L^{'} = G_L^{'} G_L M_L G_L^{'} = 1.G_L^{'}$. Ceci montre que $G_L^{'}$ est vecteur propre de S_L pour la valeur propre $\lambda = 1$. On montre de même que si λ est valeur propre de R_L de vecteur propre unitaire $u^{'}$ alors λ est valeur propre de S_L pour le même vecteur propre $u^{'}$ (utiliser le fait que $G_L M_L u^{'} = 0$ comme produit scalaire de deux vecteurs orthogonaux). □

Comme R_L et R_C ont les mêmes valeurs propres non nulles on en déduit que S_L et S_C ont aussi les mêmes valeurs propres non nulles. Autrement dit elles ont le même rang (égal au rang de R_L ou R_C plus 1). On travaillera donc avec la plus petite des deux matrices en n'oubliant pas d'enlever l'axe trivial correspondant au barycentre et donc d'enlever la valeur propre $\lambda = 1$ de l'étude. Remarquer que les projection des points profils lignes ou profils colonnes sur l'axe donné par le barycentre sont toutes égales au barycentre lui même qui est l'origine des coordonnées dans W_L ou W_C.

2.4 Facteurs principaux et Composantes principales

Les valeurs propres des matrices d'inertie λ_k nous donnent les axes principaux qui sont les vecteurs propres unitaires a_k (profils lignes) et b_k (profils colonnes) correspondants. On notera les doites (axes) qu'ils engendrent par E_k et F_k respectivement.

Définition 2.4.1. *Les facteurs principaux pour les profils lignes sont les* $u_k = M_L a_k^{'}$. *Pour les profils colonnes ce sont les* $v_k = M_C b_k^{'}$

Au chapitre premier les facteurs principaux étaient confondus avec les axes principaux car la matrice du produit scalaire était $M = Id$ la matrice identité. Les facteurs principaux vont servir à définir les composantes principales dont les coordonnées vont nous donner les positions des projetés des points profils lignes et profils colonnes sur les axes principaux. On sait que ces positions servent de manière essentielle dans l'interpretatation des résultats.

Définition 2.4.2. *Les composantes principales des profils lignes sont les* $C_k = Y_L u_k = X_L u_k$. *Pour les profils colonnes ce sont les* $D_k = Y_C v_k = X_C v_k$.

Remarquer que pour calculer les composantes principales on peut utiliser les tableaux centrés Y_L et Y_C ou les tableaux non centrés X_L et X_C. En effet on a par exemple pour les profils lignes

$$C_k = Y_L u_k = (X_L - 1_n G_L) M_L a_k^{'} = X_L M_L a_k^{'} - 1_N G_L M_L a^{\prime} = X_L u_k$$

puisque $G_L M_L a_k^{'} = \langle G_L, a_k \rangle = 0$. C'est une autre manière de dire qu'en AFC le centrage n'est pas vraiment nécessaire.

Les composantes principales des profils lignes (les modalités de la variable V_1) définissent en quelque sorte de nouvelles modalités de la variable V_2 qui vont nous aider à comprendre le tableau de données initial et donc la liaison entre les deux variables. On n'aura bien sûr besoin que des deux ou trois premières composantes qui vont résumer et concentrer l'information. On essaiera de leur donner un nom adequat. De même les composantes principales des profils colonnes (les modalités de la variable V_2) définissent de nouvelles modalités de la variable V_1.

L'égalité

$$I_{E_k^\perp} = \lambda_k$$

montre que la part de l'inertie expliquée par l'axe E_k est égale à la valeur propre λ_k. Comme on l'a vu l'inertie renseigne sur la nature du nuage de points profils lignes ou profils colonnes. Mais ici en AFC elle renseigne aussi sur le degré de liaison entre les deux variables. En effet posons

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(Nf_{ij} - Nf_{i\bullet}f_{\bullet j})^2}{Nf_{i\bullet}f_{\bullet j}}$$

On remarque que les deux variables V_1 et V_2 sont indépendantes si et seulement si $\chi^2 = 0$. De plus on a

$$\frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} = f_{i\bullet}\frac{(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j})^2}{f_{\bullet j}} = f_{\bullet j}\frac{(\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet})^2}{f_{i\bullet}}$$

Ceci montre que

$$I_L = I_C = \frac{\chi^2}{N}$$

Autrement dit plus l'inertie est grande et plus on s'ecarte de l'hypothèse d'indépendance des deux variables qui est donnée par $\chi^2 = 0$. Soit r le nombre de valeurs propres non nulles des matrices d'inertie R_L et R_C (profils centrés) ou non nulles et différentes de 1 des matrices d'inertie S_L et S_C (profils non centrés). On sait que $r \leq min(n-1, m-1)$. L'inertie totale est la somme des inerties expliquée par chacun des axes E_k pour $k = 1, \ldots, r$. On a donc

$$I_L = I_C = I_{E_1^\perp} + \ldots + I_{E_r^\perp} = \lambda_1 + \ldots + \lambda_r$$

Chaque axe et donc chaque composante principale explique une partie de l'inertie qui est égale à

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_r}$$

et qui exprime aussi le degré de liaison entre les deux variables expliquée par l'axe k. En général les deux premiers axes vont contribuer á la plus grande partie de l'inertie et on projette alors les profils sur le plan engendré par ces deux premiers axes correspondant donc à λ_1 la plus grande valeur propre de la matrice d'inertie et λ_2 la seconde plus grande valeur propre.

2.5 Contribution et qualité de représentation

La contribution d'un profil ligne x_{iL} à l'axe E_k est

$$Ct_{E_k}(x_{iL}) = \frac{f_{i\bullet}c_{ik}^2}{\lambda_k}$$

avec c_{ik} la coordonnée i de la composante principale C_k. De même la contribution du profil colonne x_{jC} à l'axe F_k est

$$Ct_{D_k}(x_{jC}) = \frac{f_{\bullet j}d_{jk}^2}{\lambda_k}$$

avec d_{jk} la coordonnée j de la composante principale D_k.

La qualité de représentation du profil ligne x_{iL} (ou y_{iL}) sur l'axe E_k est le cosinus au carré de l'angle que fait le profil ligne avec E_k

$$q_{E_k}(x_{iL}) = \frac{OH_{iL}^2}{Oy_{iL}^2} = \frac{c_{ik}^2}{d^2(O, y_{iL})}$$

avec H_{iL} le projeté de x_{iL} sur E_k. De même la qualité de représentation du profil colonne x_{jC} (ou y_{jC}) sur l'axe F_k est le cosinus au carré de l'angle entre le profil colonne et l'axe

$$q_{F_k}(x_{jC}) = \frac{OH_{jC}^2}{Oy_{jC}^2} = \frac{d_{jk}^2}{d^2(O, y_{jC})}$$

avec H_{jC} le projeté du profil colonne sur l'axe F_k. Les qualités de représentation s'ajoutent sur les axes.

2.6 Interpretation

Une fois qu'on a les valeurs propres et les vecteurs propres unitaires des matrices d'inertie des profils lignes et des profils colonnes, on peut calculer les composantes principales , les contributions des modalités aux axes

et leur qualité de représentation. On ne fera confiance qu'aux profils bien représentés. On a vu qu'une AFC consiste en une double ACP : une sur les profils lignes et une autre sur les profils colonnes. La première ACP va essayer d'expliquer la variable V_1 par la variable V_2. On fera comme si les lignes (et donc la modalités de V_1) sont les individus et les colonnes (donc les modalités de V_2) sont les variables. Les composantes principales sont donc de nouvelles modalités (virtuelles) de la variable V_2. La seconde ACP va expliquer la variable V_2 par la variable V_1. On fera comme si les lignes (les modalités de V_2) sont les individus et les colonnes (les modalités de V_1) sont les variables. Les composantes principales sont donc de nouvelles modalités (virtuelles) de la variable V_1. Les différentes contributions vont nous renseigner sur les modalités qui rentrent dans la fabrication d'une composante principale donnée.

Les composantes principales vont nous donner les coordonnées des projetés des profils lignes et colonnes sur les axes factoriels et donc sur le plan factoriel. On pourra donc dessiner ces points sur le plan. La proximité des projetés de deux points profils lignes ou profils colonnes voudra dire que ces deux profils (ou modalités) ont les mêmes caractéristiques (presque les mêmes valeurs ou distributions pour les modalités de l'autre variable). la proximité des projetés d'un profil ligne et d'un profil colonne voudra dire qu'il ya liason trés forte entre les modalités correspondantes des deux variables (on pourra jauger cette liaison par exemple par l'angle que font les deux projetés).

Exercice 2.6.1. *Faire une AFC complète (avec interpretation des résultats) sur l'exemple de la couleur des cheveux et des yeux et sur celui de l'exercice* 2.1.2.