

UNIVERSITÉ PARIS-EST-CRÉTEIL
ESIPE

Analyse et fouille de données

Spécialité : INGÉNIEUR 2ème année

BÉATRICE DE TILIÈRE

Je tiens à remercier Belaïd Ghermani de m'avoir donné ses notes de cours et de nombreux livres qui m'ont aidés à construire ce polycopié.

TABLE DES MATIÈRES

1	ÉLÉMENTS MATHÉMATIQUES POUR L'ANALYSE DE DONNÉES	3
1.1	Matrices, vecteurs lignes et vecteurs colonnes	3
1.1.1	Définitions et notations	3
1.1.2	Opérations sur les matrices	4
1.1.3	Quelques matrices particulières	4
1.2	Matrices et applications linéaires	5
1.2.1	Écriture d'un vecteur dans une base	5
1.2.2	Représentation d'une application linéaire	5
1.2.3	Changement de bases	5
1.2.4	Noyau, Image d'une application linéaire	6
1.3	Trace et déterminant	6
1.3.1	Trace	6
1.3.2	Déterminant	7
1.4	Valeurs propres et vecteurs propres	7
1.5	Espaces euclidiens métriques	8
1.5.1	Formes bilinéaires, symétriques, définies positives	8
1.5.2	Produit scalaire, norme, distance	8
1.5.3	Orthogonalité	9
1.5.4	Projection M -orthogonale sur une droite	9
1.6	Cosinus	10
1.7	Quelques propriétés des matrices symétriques	10
1.8	Maximum d'une forme quadratique sous contrainte quadratique	12
1.9	Exercices	14
2	ANALYSE EN AXES PRINCIPAUX : PRINCIPES DE BASE	25
2.1	Idée générale	25
2.1.1	Les données	25
2.1.2	Représentation graphique	25
2.1.3	Critères de projection	26
2.2	Ajustement du nuage des individus dans l'espace des variables : principe des moindres carrés	27
2.2.1	Étape 1 : ajustement du nuage $N(I)$ par une droite \mathcal{D}_1	27
2.2.2	Étape 2 : Deuxième direction d'ajustement du nuage $N(I)$	29
2.2.3	Ajustement du nuage $N(I)$ sous le critère des moindres carrés	30
2.2.4	Projection des individus sur les nouvelles variables	31
2.3	Ajustement du nuage des variables dans l'espace des individus	31
2.3.1	Ajustement du nuage $N(J)$ sous le critère des moindres carrés	31
2.3.2	Projection des variables sur les nouveaux individus	32

2.4	Relation entre les ajustements dans les deux espaces	33
2.5	Reconstruction des données	33
2.5.1	Reconstruction exacte	34
2.5.2	Reconstitution approchée	34
2.6	Éléments supplémentaires	34
2.7	Généralisations : individus pondérés	35
2.7.1	Ajustement du nuage $N(I)$ sous le critère des moindres carrés	35
2.7.2	Projection des individus sur les nouvelles variables	36
2.7.3	Ajustement du nuage $N(J)$ sous le critère des moindres carrés	37
2.7.4	Projection des variables sur les nouveaux individus	37
2.7.5	Relation entre les ajustements dans les deux espaces	38
2.8	Exercices	38
3	STATISTIQUE DESCRIPTIVE ÉLÉMENTAIRE	45
3.1	Matrice des poids	45
3.2	Le cas d'une variable	45
3.2.1	Représentation graphique	46
3.2.2	Paramètres de position	46
3.2.3	Paramètres de dispersion	47
3.3	Le cas de deux variables	48
3.3.1	Représentation des données	48
3.3.2	Paramètres de position et de dispersion	49
3.3.3	Paramètres de relation	49
3.4	Le cas de p variables	51
3.4.1	Matrice des données centrées	52
3.4.2	Matrice des variances-covariances	52
3.4.3	Matrice des données centrées réduites	53
3.4.4	Matrice des corrélations	53
3.4.5	Inertie du nuage de points	54
3.5	Individus regroupés en classes	56
3.5.1	Notations	56
3.5.2	Matrice des variances inter et intra-classes	56
3.5.3	Inertie du nuage, inertie intra et inter-classes	58
4	ANALYSE EN COMPOSANTES PRINCIPALES	63
4.1	Ajustement du nuage $N(I)$ des individus	64
4.1.1	Distance et inertie	64
4.1.2	Ajustement du nuage des individus $N(I)$	65
4.1.3	Composantes principales : Projection des individus sur les nouvelles variables	66
4.1.4	Inertie revisitée	67
4.1.5	Aide à l'interprétation dans le plan factoriel	68
4.2	Ajustement du nuage des variables $N(J)$	69
4.2.1	Distance entre les points-variables	69
4.2.2	Ajustement du nuage des variables $N(J)$	70
4.2.3	Projection des variables sur les nouveaux individus	70
4.2.4	Relation entre les ajustements dans les deux espaces	71
4.2.5	Aide à l'interprétation	71

4.3	Exercice	76
5	ANALYSE FACTORIELLE DE CORRESPONDANCES	83
5.1	Distributions marginales et conditionnelles	83
5.1.1	Distributions marginales, tableaux de contigence	83
5.1.2	Distributions conditionnelles, tableaux des profils	85
5.1.3	Indépendance	87
5.2	Ajustement du nuage des profils-lignes et colonnes	88
5.2.1	Distance du χ^2	88
5.2.2	Inertie des nuages profils-lignes et colonnes	89
5.2.3	Ajustement du nuage des profils-lignes et colonnes	90
5.3	Lien entre les ajustements dans les deux espaces : représentation simultanée	92
5.3.1	Relation entre les ajustements dans les deux espaces	92
5.3.2	Relations quasi-barycentriques, représentation simultanée	93
5.4	Aide à l'interprétation	94
5.5	Calculs pour le théorème 5.1	95
5.6	Exercice	95
6	MÉTHODES DE CLASSIFICATION	103
6.1	Distances	103
6.2	Méthode de classification hiérarchique (CAH)	105
6.2.1	Algorithme	105
6.2.2	Distance de Ward entre classes	106
6.3	Méthodes non hiérarchiques : méthode des centres mobiles	114
7	ANALYSE DISCRIMINANTE	117
7.1	Introduction	117
7.1.1	Contexte	117
7.1.2	Notations	118
7.2	Discrimination linéaire	119
7.2.1	Principe de discrimination linéaire	119
7.2.2	Discrimination des données : approche 1	121
7.2.3	Évaluation des facteurs discriminants	124
7.2.4	Discrimination des données : approche 2	124
7.2.5	Lien entre les deux approches	125
7.3	Analyse factorielle discriminante : prise de décision	125
7.3.1	Règle géométrique d'affectation	125
7.3.2	Modèle Baysien d'affectation	126
7.3.3	Le cas de deux classes	127
	BIBLIOGRAPHY	129



CHAPITRE 1

ÉLÉMENTS MATHÉMATIQUES POUR L'ANALYSE DE DONNÉES

Le but de ce chapitre est de présenter les éléments mathématiques pour l'analyse de données. Les données sont représentées dans une matrice et l'un des principaux outils pour leur étude est l'algèbre linéaire. Nous rappelons donc les bases nécessaires d'algèbre linéaire, puis abordons un problème d'optimisation de forme quadratique sous contrainte quadratique, un des fondements des méthodes étudiées dans ce polycopié.

1.1 MATRICES, VECTEURS LIGNES ET VECTEURS COLONNES

1.1.1 DÉFINITIONS ET NOTATIONS

◦ Soit $A = (a_{ij})$ une matrice de taille $n \times p$.

$$A = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{np} \end{pmatrix}$$

◦ La *transposée* de la matrice A , notée A' , est de taille $p \times n$, elle est définie par : $(A')_{ij} = a_{ji}$.

◦ Un *vecteur ligne* de \mathbb{R}^p est une matrice de taille $1 \times p$. Un *vecteur colonne* de \mathbb{R}^n est une matrice de taille $n \times 1$.

◦ Les vecteurs colonnes sont écrits en lettres *grasses*, les vecteurs lignes sont écrits comme transposés de vecteurs colonnes. Par exemple,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{x}' = (x_1 \ \dots \ x_n).$$

- La $i^{\text{ème}}$ ligne de la matrice $A = (a_{ij})$ est notée \mathbf{a}'_i ; il s'agit d'un vecteur ligne de \mathbb{R}^p :

$$\mathbf{a}'_i = (a_{i1} \cdots a_{ip}), \quad \text{où } \mathbf{a}_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix} \text{ est la } i^{\text{ème}} \text{ ligne écrite sous forme de vecteur colonne.}$$

- La $j^{\text{ème}}$ colonne de la matrice A est notée \mathbf{a}^j ; il s'agit d'un vecteur colonne de \mathbb{R}^n :

$$\mathbf{a}^j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

- On note $\mathbf{1}_n$, ou simplement $\mathbf{1}$, le vecteur de taille $n \times 1$ formé de 1 uniquement :

$$\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

- On note I_n , ou I , la matrice identité de taille $n \times n$.

1.1.2 OPÉRATIONS SUR LES MATRICES

Voici les opérations usuelles sur les matrices. Soit A une matrice de taille $n \times p$.

- *Somme* : $(A + B)_{ij} = a_{ij} + b_{ij}$ pour B de même taille que A .
- *Multiplication par un scalaire*. Soit $\lambda \in \mathbb{R}$, $(\lambda A)_{ij} = \lambda a_{ij}$.
- *Multiplication de deux matrices* : $(AB)_{ij} = \mathbf{a}'_i \mathbf{a}_j = \sum_{k=1}^p a_{ik} b_{kj}$, pour B de taille $p \times q$; AB est alors de taille $n \times q$.
- *Inversion*. Supposons $n = p$. La matrice A est *inversible* s'il existe une matrice B de taille $n \times n$ telle que $AB = BA = I_n$. La matrice B est alors l'*inverse* de A et est notée $B = A^{-1}$.

1.1.3 QUELQUES MATRICES PARTICULIÈRES

Soit $A = (A_{ij})$ une matrice de taille $n \times n$.

- La matrice A est *symétrique* si :

$$A' = A.$$

- La matrice A est *orthogonale* ssi une des conditions équivalentes suivante est remplie :

— $A'A = I$.

— La matrice A est inversible et $A^{-1} = A'$.

- Si la matrice A est symétrique, elle est dite *définie positive* si,

$$\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}, \quad \mathbf{x}' A \mathbf{x} > 0.$$

1.2 MATRICES ET APPLICATIONS LINÉAIRES

1.2.1 ÉCRITURE D'UN VECTEUR DANS UNE BASE

Soit l'espace vectoriel \mathbb{R}^n muni d'une base $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. Un vecteur \mathbf{x} de \mathbb{R}^n s'écrit de manière unique sous la forme $\mathbf{x} = x_1\mathbf{e}_1 + \dots + x_n\mathbf{e}_n$ dans la base \mathcal{B} . Ceci justifie l'écriture sous forme de vecteur colonne \mathbf{x} , ou de vecteur ligne \mathbf{x}' ,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{x}' = (x_1 \ \cdots \ x_n).$$

Par défaut, un vecteur sera écrit dans la base canonique.

1.2.2 REPRÉSENTATION D'UNE APPLICATION LINÉAIRE

Considérons l'espace vectoriel \mathbb{R}^p muni d'une base $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ et l'espace vectoriel \mathbb{R}^n muni d'une base $\mathcal{B}' = \{\tilde{\mathbf{e}}_1, \dots, \tilde{\mathbf{e}}_n\}$.

○ Une application f de \mathbb{R}^p dans \mathbb{R}^n est une application linéaire, ssi il existe des scalaires $(a_{ij}, i \in \{1, \dots, n\}, j \in \{1, \dots, p\})$, tels que :

$$\forall \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p, \forall i \in \{1, \dots, n\}, \quad f(\mathbf{x})_i = \sum_{j=1}^p a_{ij}x_j.$$

Autrement dit, si on définit la matrice $A = (a_{ij})$, $f(\mathbf{x}) = A\mathbf{x}$. La matrice A est la *matrice de f dans les bases \mathcal{B} et \mathcal{B}'* ; la $j^{\text{ème}}$ colonne de A est l'image par f du $j^{\text{ème}}$ vecteur de base \mathbf{e}_j de \mathcal{B} écrit dans la base \mathcal{B}' .

○ Un *endomorphisme* f est une application linéaire de \mathbb{R}^n dans \mathbb{R}^n . La *matrice de f dans la base \mathcal{B}* est la matrice de f lorsque la base de \mathbb{R}^n est \mathcal{B} (en tant qu'espace de départ et espace d'arrivée).

1.2.3 CHANGEMENT DE BASES

○ Soient $\mathcal{B}_1, \mathcal{B}_2$ deux bases de \mathbb{R}^p . La *matrice de passage de base de \mathcal{B}_1 à \mathcal{B}_2* , notée $P_{\mathcal{B}_1}^{\mathcal{B}_2}$, est la matrice de l'application identité de \mathbb{R}^p muni de la base \mathcal{B}_2 dans \mathbb{R}^p muni de la base \mathcal{B}_1 . La $j^{\text{ème}}$ colonne est donc l'image du $j^{\text{ème}}$ vecteur de la base \mathcal{B}_2 écrit dans la base \mathcal{B}_1 .

○ La matrice $P_{\mathcal{B}_1}^{\mathcal{B}_2}$ est inversible et

$$P_{\mathcal{B}_2}^{\mathcal{B}_1} = (P_{\mathcal{B}_1}^{\mathcal{B}_2})^{-1}.$$

○ Soient aussi \mathcal{B}'_1 et \mathcal{B}'_2 deux bases de \mathbb{R}^n et soit A la matrice d'une application linéaire f de \mathbb{R}^p dans \mathbb{R}^n , dans les bases \mathcal{B}_1 et \mathcal{B}'_1 . Alors, la matrice B de l'application linéaire f dans les bases \mathcal{B}_2 et \mathcal{B}'_2 est :

$$B = Q^{-1}AP,$$

où P est la matrice de passage de \mathcal{B}_1 à \mathcal{B}_2 et Q est la matrice de passage de \mathcal{B}'_1 à \mathcal{B}'_2 .

◦ Considérons le cas particulier d'un endomorphisme f de \mathbb{R}^n . Soient deux bases \mathcal{B}_1 et \mathcal{B}_2 de \mathbb{R}^n , et soient A , respectivement B , la matrice de f dans \mathcal{B}_1 , respectivement \mathcal{B}_2 ; alors,

$$B = P^{-1}AP,$$

où P est la matrice de passage de base de \mathcal{B}_1 à \mathcal{B}_2 .

1.2.4 NOYAU, IMAGE D'UNE APPLICATION LINÉAIRE

Soit f une application linéaire de \mathbb{R}^p dans \mathbb{R}^n .

◦ L'*image* de f , notée $\text{Im}(f)$, est l'ensemble des vecteurs de \mathbb{R}^n de la forme $f(\mathbf{x})$ pour un certain \mathbf{x} dans \mathbb{R}^p , c'est un sous-espace vectoriel de \mathbb{R}^n :

$$\text{Im}(f) = \{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^p\}.$$

◦ Le *rang* de f , noté $\text{rg}(f)$, est la dimension de l'image de f .

◦ Le *noyau* de f , noté $\ker(f)$, est l'ensemble des vecteurs de \mathbb{R}^p que f annule, c'est un sous-espace vectoriel de \mathbb{R}^p :

$$\ker(f) = \{\mathbf{x} \in \mathbb{R}^p \mid f(\mathbf{x}) = \mathbf{0}\}.$$

Théorème 1.1 (Théorème du rang). *Soit f une application linéaire de \mathbb{R}^p dans \mathbb{R}^n . Alors,*

$$\text{rg}(f) + \dim(\ker(f)) = \dim \mathbb{R}^p = p.$$

Calcul du rang d'une application linéaire $f : \mathbb{R}^p \rightarrow \mathbb{R}^n$

◦ Le rang de l'application linéaire f est égal au rang de la matrice A représentant f dans un choix de base \mathcal{B} et \mathcal{B}' . Le *rang de la matrice* A est par définition le nombre maximal de vecteurs lignes (ou colonnes) indépendants. On calcule le rang de la matrice A de la manière suivante.

◦ Une matrice est sous forme *échelonnée* si le nombre de zéros précédants la première valeur non nulle d'une ligne augmente de ligne en ligne jusqu'à ce qu'il ne reste plus que des zéros.

◦ Toute matrice peut être transformée en une matrice échelonnée au moyen d'opérations élémentaires sur les lignes (méthode du Pivot de Gauss), à savoir :

- permuter deux lignes ;
- multiplier une ligne par une constante non nulle ;
- ajouter à une ligne le multiple d'une autre ligne.

◦ Le rang d'une matrice sous forme échelonnée est le nombre de lignes non nulles.

On a donc que le rang de f est le nombre de lignes non-nulles de la matrice A mise sous forme échelonnée.

1.3 TRACE ET DÉTERMINANT

1.3.1 TRACE

Soit A une matrice de taille $n \times n$. La *trace* de la matrice A , notée $\text{Tr}(A)$, est par définition :

$$\text{Tr}(A) = \sum_{i=1}^n a_{ii}.$$

La trace satisfait aux propriétés suivantes :

- $\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$, si B est de taille $n \times n$,
- soit $\lambda \in \mathbb{R}$, $\text{Tr}(\lambda A) = \lambda \text{Tr}(A)$,
- si A est de taille $n \times p$ et B est de taille $p \times n$, $\text{Tr}(AB) = \text{Tr}(BA)$.
- si A est de taille $n \times p$ et B est de taille $p \times q$ et C de taille $q \times n$,

$$\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB), \quad (\text{permutations circulaires}).$$

- Soit f un endomorphisme de \mathbb{R}^n , et soient A (respectivement B) la matrice de f dans une base \mathcal{B}_1 (respectivement \mathcal{B}_2) de \mathbb{R}^n . Alors, $\text{Tr}(A) = \text{Tr}(B)$.

1.3.2 DÉTERMINANT

Notons $\det(A)$ le déterminant d'une matrice A de taille $n \times n$.

- Pour calculer le déterminant, on peut utiliser le développement en colonnes (ou lignes). Le développement selon la $j^{\text{ème}}$ colonne est :

$$\det(A) = \sum_{i=1}^n a_{ij} (-1)^{i+j} \det(A_{i,j}),$$

où $A_{i,j}$ est la matrice A dont on a enlevé la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne.

Le déterminant satisfait aux propriétés suivantes :

- La matrice A est inversible si et seulement si son déterminant est non nul.
- Le déterminant d'une matrice triangulaire est le produit de ses éléments diagonaux.
- Si A est de taille $n \times n$ et B est de taille $n \times n$, $\det(AB) = \det(A) \det(B)$.
- Soit f un endomorphisme de \mathbb{R}^n , et soient A (respectivement B) la matrice de f dans une base \mathcal{B}_1 (respectivement \mathcal{B}_2) de \mathbb{R}^n . Alors, $\det(A) = \det(B)$.

Les opérations sur les lignes et les colonnes de la matrice A ont des conséquences sur son déterminant :

- si on permute deux lignes ou deux colonnes, le déterminant change de signe ;
- si deux lignes ou deux colonnes sont identiques, le déterminant est nul ;
- si on multiplie tous les termes d'une même ligne ou d'une même colonne par un réel k , le déterminant est multiplié par k ;
- on peut ajouter à une colonne (ou une ligne) un multiple d'une autre colonne (ou d'une autre ligne) sans changer la valeur du déterminant ; en conséquence, si une ligne ou une colonne est nulle, le déterminant est nul.

1.4 VALEURS PROPRES ET VECTEURS PROPRES

Soit A une matrice de taille $n \times n$.

- Un scalaire λ est *valeur propre* de A s'il existe un vecteur non nul \mathbf{u} de \mathbb{R}^n tel que $A\mathbf{u} = \lambda\mathbf{u}$. Si c'est le cas, \mathbf{u} est un *vecteur propre* associé à la valeur propre λ .
- Le *polynôme caractéristique* de la matrice A est $\det(A - \lambda I_n)$.
- Le scalaire λ est valeur propre de A si et seulement si $\det(A - \lambda I_n) = 0$.

- Soit λ une valeur propre de A . Le *sous-espace propre* de λ , noté E_λ , est formé des vecteurs propres de λ et $\mathbf{0}$. C'est un sous-espace vectoriel de \mathbb{R}^n .
- La dimension de E_λ est minorée par 1 et majorée par la multiplicité de λ dans le polynôme caractéristique $\det(A - \lambda I_n)$.
- Soient $\lambda_1, \dots, \lambda_r$ les valeurs propres de A distinctes 2 à 2. Considérons la famille \mathcal{F} obtenue par concaténation d'un choix de bases $\mathcal{B}_1, \dots, \mathcal{B}_r$ de chacun des $E_{\lambda_1}, \dots, E_{\lambda_r}$. Alors \mathcal{F} forme une famille libre.
- La matrice A est *diagonalisable* s'il existe une base de \mathbb{R}^n formée de vecteurs propres de A .
- Si la matrice A a n valeurs propres distinctes, alors elle est diagonalisable (condition suffisante mais non nécessaire).

1.5 ESPACES EUCLIDIENS MÉTRIQUES

1.5.1 FORMES BILINÉAIRES, SYMÉTRIQUES, DÉFINIES POSITIVES

- Soit $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ une base de \mathbb{R}^n . Une application f de $\mathbb{R}^n \times \mathbb{R}^n$ dans \mathbb{R} est une forme bilinéaire, ssi il existe des scalaires m_{ij} , $i \in \{1, \dots, n\}$, $j \in \{1, \dots, n\}$, tels que pour tout $\mathbf{x} = x_1\mathbf{e}_1 + \dots + x_n\mathbf{e}_n$ et tout $\mathbf{y} = y_1\mathbf{e}_1 + \dots + y_n\mathbf{e}_n$, $f(\mathbf{x}, \mathbf{y})$ s'écrit $f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \sum_{j=1}^n m_{ij}x_iy_j$, autrement dit, si on écrit $M = (m_{ij})$,

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}'M\mathbf{y}.$$

La matrice $M = (m_{ij})$ est la *matrice de f dans la base \mathcal{B}* ; on a $m_{ij} = f(\mathbf{e}_i, \mathbf{e}_j)$.

- Une forme bilinéaire est *symétrique*, ssi la matrice qui la représente dans une base \mathcal{B} est symétrique.
- Une forme bilinéaire symétrique est *définie positive*, ssi la matrice qui la représente dans une base \mathcal{B} est définie positive.

Remarquer que ces deux résultats ne dépendent pas de la base choisie.

1.5.2 PRODUIT SCALAIRE, NORME, DISTANCE

- Un *produit scalaire* sur \mathbb{R}^n est une forme bilinéaire, symétrique, définie positive. Au vu du résultat précédent, étant donné une base \mathcal{B} de \mathbb{R}^n (que l'on prendra comme étant la base canonique), il existe une bijection entre les produits scalaires et les matrices symétriques, définies positives. Pour toute telle matrice M , on note $\langle \cdot, \cdot \rangle_M$ le produit scalaire associé :

$$\langle \mathbf{x}, \mathbf{y} \rangle_M = \mathbf{x}'M\mathbf{y}.$$

- La *norme* associée au produit scalaire $\langle \cdot, \cdot \rangle_M$, notée $\|\cdot\|_M$ est définie par,

$$\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_M = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_M}.$$

Un vecteur \mathbf{x} de \mathbb{R}^n est dit *M -normé à 1* ou *M -unitaire* si,

$$\|\mathbf{x}\|_M = 1.$$

◦ La *distance* associée, notée $d_M(\cdot, \cdot)$, est définie par :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, d_M(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_M.$$

Remarque.

- L'espace \mathbb{R}^n muni du produit scalaire $\langle \cdot, \cdot \rangle_M$ est un *espace vectoriel euclidien*. Le produit scalaire $\langle \cdot, \cdot \rangle_M$ définit une distance d_M qui confère à \mathbb{R}^n une structure d'*espace métrique*. On parlera aussi de \mathbb{R}^n *muni de la métrique provenant de la matrice M*.
- Lorsque M est la matrice identité, on retrouve le produit scalaire usuel, etc. On ne note pas l'indice M dans ce cas.

1.5.3 ORTHOGONALITÉ

Soit M une matrice symétrique, définie positive, et soit $\langle \cdot, \cdot \rangle_M$ le produit scalaire associé.

◦ Deux vecteurs \mathbf{x}, \mathbf{y} de \mathbb{R}^n sont dits *M-orthogonaux* si,

$$\langle \mathbf{x}, \mathbf{y} \rangle_M = 0.$$

◦ Une base $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ est *M-orthonormée* si,

$$\forall i, j \in \{1, \dots, n\}, \langle \mathbf{e}_i, \mathbf{e}_j \rangle_M = \delta_{ij}.$$

◦ Soit \mathcal{B} une base *M-orthonormée* de \mathbb{R}^n . Une base \mathcal{B}_1 de \mathbb{R}^n est *M-orthonormée* si et seulement si la matrice de passage P de la base \mathcal{B} à la base \mathcal{B}_1 est *M-orthogonale*, c'est-à-dire,

$$P'MP = I, \text{ et } MPP' = I.$$

Dans le cas du produit scalaire usuel, on retrouve le fait que $P'P = PP' = I$, soit que la matrice P est orthogonale.

1.5.4 PROJECTION M-ORTHOGONALE SUR UNE DROITE

Soit M une matrice symétrique, définie positive, et soit $\langle \cdot, \cdot \rangle_M$ le produit scalaire associé.

◦ Soit \mathcal{D} une droite de \mathbb{R}^n qui passe par l'origine et soit \mathbf{u} un vecteur directeur de cette droite. Pour tout point \mathbf{x} de \mathbb{R}^n , la *projection M-orthogonale de x sur D*, notée $\pi(\mathbf{x})$, est le point appartenant à \mathcal{D} tel que le vecteur $\pi(\mathbf{x}) - \mathbf{x}$ est *M-orthogonal* à la droite \mathcal{D} .

◦ La projection *M-orthogonale* π est un endomorphisme de \mathbb{R}^n .

◦ La projection *M-orthogonale* $\pi(\mathbf{x})$ de \mathbf{x} sur \mathcal{D} est donnée par :

$$\pi(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{u} \rangle_M}{\|\mathbf{u}\|_M} \frac{\mathbf{u}}{\|\mathbf{u}\|_M}.$$

La coordonnée de la projection $\pi(\mathbf{x})$ de \mathbf{x} sur \mathcal{D} est : $\frac{\langle \mathbf{x}, \mathbf{u} \rangle_M}{\|\mathbf{u}\|_M}$. Remarquer que si \mathbf{u} est un vecteur *M-unitaire*, c'est-à-dire $\|\mathbf{u}\|_M = 1$, alors la projection est simplement donnée par

$$\pi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{u} \rangle_M \mathbf{u},$$

et la coordonnée est alors le produit scalaire $\langle \mathbf{x}, \mathbf{u} \rangle_M$.

1.6 COSINUS

Le *cosinus de l'angle* entre deux vecteurs \mathbf{x} et \mathbf{y} :

$$\cos_M \angle(\mathbf{0x}, \mathbf{0y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_M}{\|\mathbf{x}\|_M \|\mathbf{y}\|_M}.$$

Remarquer que $\frac{\langle \mathbf{x}, \mathbf{y} \rangle_M}{\|\mathbf{y}\|_M}$ est la coordonnée de la projection M -orthogonale de \mathbf{x} sur la droite $(\mathbf{0y})$.

Le cosinus est compris entre -1 et 1 comme conséquence de l'inégalité de Cauchy-Schwartz.

1.7 QUELQUES PROPRIÉTÉS DES MATRICES SYMÉTRIQUES

Rappelons qu'une matrice A est symétrique si $A' = A$. Un des théorème clé pour l'analyse factorielle est le théorème spectral pour les matrices symétriques.

Théorème 1.2 (Théorème spectral). *Soit A une matrice symétrique de taille $p \times p$. Alors, il existe une base orthonormale de \mathbb{R}^p formée de vecteurs propres de A . De plus, chacune des valeurs propres associée est réelle. Autrement dit, il existe une matrice orthogonale P telle que :*

$$P'AP = D,$$

où D est la matrice diagonale formée des valeurs propres de A .

Voici quelques autres faits sur les matrices symétriques qui nous seront utiles.

- Soit X une matrice de taille $n \times p$. Alors,
 1. Les matrices $X'X$ et XX' sont symétriques.
 2. Les matrices $X'X$ et XX' ont les mêmes valeurs propres non nulles.
 3. Si \mathbf{u} est vecteur propre pour une valeur propre non-nulle λ de $X'X$, alors $X\mathbf{u}$ est vecteur propre pour la même valeur propre λ de XX' . De manière symétrique, si \mathbf{v} est vecteur propre pour une valeur propre non-nulle λ de XX' , alors $X'\mathbf{v}$ est vecteur propre pour la même valeur propre de $X'X$.

La démonstration de ces résultats est l'objet de l'exercice 1.3.

- Une matrice A symétrique est définie positive ssi toutes ses valeurs propres sont strictement positives.
- **Décomposition de Cholesky.** Soit M une matrice symétrique et définie positive. Alors il existe une matrice réelle triangulaire supérieure L telle que :

$$M = L'L.$$

- **Puissance $\frac{1}{2}$ d'une matrice symétrique définie positive.** Soit M une matrice symétrique définie positive. Alors, d'après le théorème spectral,

$$M = PDP',$$

où P est une matrice orthogonale, et D est la matrice diagonale formée des valeurs propres (réelle et positive) de M . Définissons $D^{\frac{1}{2}}$, la matrice diagonale formée des racines carrées (positives)

des valeurs propres de D , ainsi, $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$. Définissons $M^{\frac{1}{2}} = PD^{\frac{1}{2}}P'$. Alors, la matrice $M^{\frac{1}{2}}$ est symétrique et on a :

$$M^{\frac{1}{2}}M^{\frac{1}{2}} = M.$$

En effet,

$$M^{\frac{1}{2}}M^{\frac{1}{2}} = PD^{\frac{1}{2}}P'PD^{\frac{1}{2}}P' = PD^{\frac{1}{2}}D^{\frac{1}{2}}P' = PDP' = M,$$

où on a utilisé le fait que la matrice P est orthogonale. La matrice $M^{\frac{1}{2}} = PD^{\frac{1}{2}}P'$ s'appelle la *puissance* $\frac{1}{2}$ de la matrice M .

On a la version plus générale suivante du théorème spectral pour les matrices symétriques.

Théorème 1.3. *Soient M une matrice symétrique définie positive et A une matrice symétrique, les deux de taille $p \times p$. Alors, il existe une base M -orthonormée de \mathbb{R}^p formée de vecteurs propres de $M^{-1}A$. De plus, chacune des valeurs propres associée est réelle. Autrement dit, il existe une matrice M -orthogonale Q telle que,*

$$Q^{-1}M^{-1}AQ = D,$$

où D est la matrice diagonale formée des valeurs propres de $M^{-1}A$.

Première démonstration. Comme M est symétrique et définie positive, il existe une matrice L inversible telle que, $M = L'L$. Posons $C = (L')^{-1}AL^{-1}$. Alors, C est une matrice symétrique. Ainsi, il existe une matrice orthogonale P telle que,

$$P'CP = D$$

où D est la matrice diagonale formée des valeurs propres de C , qui sont réelles.

Posons $Q = L^{-1}P$. Montrons que la matrice Q est M -orthogonale, c'est-à-dire que $Q'MQ = I$ et $MQQ' = I$:

$$\begin{aligned} Q'MQ &= P'(L^{-1})'ML^{-1}P = P'(L^{-1})'L'LL^{-1}P = P'P = I, \text{ car } P \text{ est orthogonale} \\ MQQ' &= L'LL^{-1}PP'(L^{-1})' = L'PP'(L^{-1})' = L'(L^{-1})' = I. \end{aligned}$$

Il reste à montrer que $Q^{-1}M^{-1}AQ = D$. On sait que $P'CP = D$. De $Q = L^{-1}P$, on déduit $P = LQ$ d'où, $P^{-1} = Q^{-1}L^{-1}$. Ainsi, en utilisant aussi la définition de C , on a :

$$\begin{aligned} P'CP = D &\Leftrightarrow P^{-1}CP = D \Leftrightarrow Q^{-1}L^{-1}CLQ = D \Leftrightarrow Q^{-1}L^{-1}(L')^{-1}AL^{-1}LQ = D \\ &\Leftrightarrow Q^{-1}(L'L)^{-1}AQ = D \Leftrightarrow Q^{-1}M^{-1}AQ = D, \end{aligned}$$

ce qui termine la preuve. □

Conséquence (de la première démonstration). La matrice Q du théorème 1.3 est, $Q = L^{-1}P$, et la matrice D du théorème 1.3 est la matrice des valeurs propres (réelles) de $(L')^{-1}AL^{-1}$, où

- L est la matrice triangulaire supérieure de la décomposition de Cholesky de M : $M = L'L$,
- P est la matrice orthogonale de la décomposition de la matrice symétrique $(L')^{-1}AL^{-1}$.

Deuxième démonstration. Comme M est symétrique et définie positive, elle admet une puissance $\frac{1}{2}$, $M^{\frac{1}{2}}$, qui est symétrique. Posons $C = M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$. Alors, C est une matrice symétrique. Ainsi, il existe une matrice orthogonale P telle que,

$$P'CP = D$$

où D est la matrice diagonale formée des valeurs propres de C , qui sont réelles.

Posons $Q = M^{-\frac{1}{2}}P$. Montrons que la matrice Q est M -orthogonale, c'est-à-dire que $Q'MQ = I$ et $MQQ' = I$:

$$\begin{aligned} Q'MQ &= P'M^{-\frac{1}{2}}MM^{-\frac{1}{2}}P = P'P = I, \text{ car } P \text{ est orthogonale} \\ MQQ' &= M^{\frac{1}{2}}M^{\frac{1}{2}}M^{-\frac{1}{2}}PP'M^{-\frac{1}{2}} = M^{\frac{1}{2}}PP'M^{-\frac{1}{2}} = M^{\frac{1}{2}}M^{-\frac{1}{2}} = I. \end{aligned}$$

Il reste à montrer que $Q^{-1}M^{-1}AQ = D$. On sait que $P'CP = D$. De $Q = M^{-\frac{1}{2}}P$, on déduit $P = M^{\frac{1}{2}}Q$ d'où, $P^{-1} = Q^{-1}M^{-\frac{1}{2}}$. Ainsi, en utilisant aussi la définition de C , on a :

$$\begin{aligned} P'CP = D &\Leftrightarrow P^{-1}CP = D \Leftrightarrow Q^{-1}M^{-\frac{1}{2}}CM^{\frac{1}{2}}Q = D \Leftrightarrow Q^{-1}M^{-\frac{1}{2}}M^{-\frac{1}{2}}AM^{-\frac{1}{2}}M^{\frac{1}{2}}Q = D \\ &\Leftrightarrow Q^{-1}M^{-1}AQ = D, \end{aligned}$$

ce qui termine la preuve. □

Conséquence (de la deuxième démonstration). La matrice Q du théorème 1.3 est, $Q = M^{-\frac{1}{2}}P$, et la matrice D du théorème 1.3 est la matrice des valeurs propres (réelles) de $M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$, où

- $M^{\frac{1}{2}}$ est la puissance $\frac{1}{2}$ de la matrice M ,
- P est la matrice orthogonale de la décomposition de la matrice symétrique $M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$.

1.8 MAXIMUM D'UNE FORME QUADRATIQUE SOUS CONTRAINTE QUADRATIQUE

Soit l'espace vectoriel \mathbb{R}^p . Soit M une matrice de taille $p \times p$, symétrique et définie positive, définissant un produit scalaire $\langle \cdot, \cdot \rangle_M$ et soit A une matrice réelle symétrique.

On cherche \mathbf{u}_1 solution du problème suivant :

$$\begin{cases} \mathbf{u}'A\mathbf{u} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\|_M = 1 \text{ est satisfaite.} \end{cases}$$

Proposition 1.1. *La solution \mathbf{u}_1 est un vecteur propre M -unitaire correspondant à la plus grande valeur propre λ_1 de la matrice $M^{-1}A$. On a alors, $\mathbf{u}'_1A\mathbf{u}_1 = \lambda_1$.*

Remarque 1.1. La matrice $M^{-1}A$ est symétrique comme produit de deux matrices symétriques. Ses valeurs propres sont donc réelles et il fait sens de parler de plus grande valeur propre.

Démonstration. D'après les multiplicateurs de Lagrange, si la fonction $\mathbf{u}'A\mathbf{u}$ admet un extremum local au point \mathbf{u} sous la contrainte $\|\mathbf{u}\|_M^2 = \langle \mathbf{u}, \mathbf{u} \rangle_M = \mathbf{u}'M\mathbf{u} = 1$, alors il existe un nombre $\lambda_1 \in \mathbb{R}$, tel que $\nabla(\mathbf{u}'A\mathbf{u}) = \lambda_1 \nabla(\mathbf{u}'M\mathbf{u} - 1)$. On souhaite donc résoudre :

$$\begin{cases} \nabla(\mathbf{u}'A\mathbf{u}) = \lambda_1 \nabla(\mathbf{u}'M\mathbf{u} - 1) \\ \mathbf{u}'M\mathbf{u} - 1 = 0. \end{cases}$$

En calculant les gradients, ceci est équivalent à résoudre (voir exercice 1.7),

$$\begin{cases} A\mathbf{u} = \lambda_1 M\mathbf{u} \\ \mathbf{u}'M\mathbf{u} - 1 = 0. \end{cases}$$

La matrice M étant définie positive, elle est inversible. Le système devient

$$\begin{cases} (M^{-1}A)\mathbf{u} = \lambda_1 \mathbf{u} \\ \mathbf{u}'M\mathbf{u} - 1 = 0. \end{cases}$$

Les solutions de ce système sont les vecteurs propres M -unitaires de la matrice $M^{-1}A$; λ_1 est alors la valeur propre associée au vecteur propre \mathbf{u} . La valeur de la fonction $\mathbf{u}'A\mathbf{u}$ que l'on cherche à maximiser est $\mathbf{u}'A\mathbf{u} = \lambda_1 \mathbf{u}'M\mathbf{u} = \lambda_1$.

Étant donné que l'on cherche le maximum, on déduit que \mathbf{u}_1 est le vecteur propre correspondant à la plus grande valeur propre de $M^{-1}A$, notée λ_1 (afin de ne pas alourdir les notations). \square

On cherche maintenant \mathbf{u}_2 solution du problème suivant :

$$\begin{cases} \mathbf{u}'A\mathbf{u} \text{ est maximale} \\ \text{sous les contraintes } \|\mathbf{u}\|_M = 1, \langle \mathbf{u}_1, \mathbf{u} \rangle_M = 0. \end{cases}$$

Proposition 1.2. *La solution \mathbf{u}_2 est un vecteur propre M -unitaire correspondant à la deuxième plus grande valeur propre λ_2 (comptée avec multiplicité) de la matrice $M^{-1}A$. On a alors, $\mathbf{u}'_2 A \mathbf{u}_2 = \lambda_2$.*

Démonstration. D'après les multiplicateurs de Lagrange, si la fonction $\mathbf{u}'A\mathbf{u}$ admet un extremum local au point \mathbf{u} sous les contraintes $\|\mathbf{u}\|_M^2 = \langle \mathbf{u}, \mathbf{u} \rangle_M = \mathbf{u}'M\mathbf{u} = 1$, $\langle \mathbf{u}_1, \mathbf{u} \rangle_M = \mathbf{u}'_1 M \mathbf{u} = 0$, alors il existe un couple $(\lambda_2, \mu_2) \in \mathbb{R}^2$, tel que $\nabla(\mathbf{u}'A\mathbf{u}) = \lambda_2 \nabla(\mathbf{u}'M\mathbf{u} - 1) + \mu_2 \nabla(\mathbf{u}'_1 M \mathbf{u})$. On souhaite donc résoudre :

$$\begin{cases} \nabla(\mathbf{u}'A\mathbf{u}) = \lambda_2 \nabla(\mathbf{u}'M\mathbf{u} - 1) + \mu_2 \nabla(\mathbf{u}'_1 M \mathbf{u}) \\ \mathbf{u}'M\mathbf{u} - 1 = 0 \\ \mathbf{u}'_1 M \mathbf{u} = 0 \end{cases}$$

En calculant les gradients, ceci est équivalent à résoudre

$$\begin{cases} A\mathbf{u} = \lambda_2 M\mathbf{u} + \mu_2 M\mathbf{u}_1 \\ \mathbf{u}'M\mathbf{u} - 1 = 0 \\ \mathbf{u}'_1 M \mathbf{u} = 0 \end{cases}$$

En pré-multipliant la première équation par \mathbf{u}'_1 , on obtient :

$$\mathbf{u}'_1 A \mathbf{u} = \lambda_2 \mathbf{u}'_1 M \mathbf{u} + \mu_2 \mathbf{u}'_1 M \mathbf{u}_1 \Leftrightarrow \mathbf{u}'_1 A \mathbf{u} = \mu_2$$

Comme \mathbf{u}_1 est vecteur propre de $M^{-1}A$ pour la valeur propre λ_1 , on a $M^{-1}A\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$. En transposant, comme $M^{-1}A$ est symétrique, on obtient $\mathbf{u}'_1 A M^{-1} = \lambda_1 \mathbf{u}'_1$, autrement dit $\mathbf{u}'_1 A = \lambda_1 \mathbf{u}'_1 M$. Ainsi,

$$\mathbf{u}'_1 A \mathbf{u} = \mu_2 \Leftrightarrow \lambda_1 \mathbf{u}'_1 M \mathbf{u} = \mu_2 \Leftrightarrow \mu_2 = 0.$$

En pré-multipliant la première équation par M^{-1} , le système à résoudre devient,

$$\begin{cases} M^{-1}A\mathbf{u} = \lambda_2\mathbf{u} \\ \mathbf{u}'M\mathbf{u} - 1 = 0 \\ \mathbf{u}'_1M\mathbf{u} = 0 \end{cases}$$

Les solutions de ce système sont les vecteurs propres \mathbf{u} de la matrice $M^{-1}A$, normés à 1, tels que \mathbf{u} est orthogonal à \mathbf{u}_1 ; λ_2 est alors la valeur propre associée au vecteur propre \mathbf{u} . La valeur $\mathbf{u}'A\mathbf{u}$ que l'on cherche à maximiser est $\mathbf{u}'A\mathbf{u} = \lambda_2\mathbf{u}'M\mathbf{u} = \lambda_2$.

Si la multiplicité de la valeur propre λ_1 est 1, alors \mathbf{u}_2 est un vecteur propre normé à 1 correspondant à cette deuxième valeur propre, que l'on note λ_2 ; on sait qu'il est orthogonal à \mathbf{u}_1 car la matrice $M^{-1}A$ est symétrique. Si la multiplicité de la valeur propre λ_1 est supérieure à 1, alors \mathbf{u}_2 est un vecteur propre de la même valeur propre, orthogonal au premier. En écrivant les valeurs propres avec multiplicité, on la notera λ_2 . \square

En itérant ces étapes, le problème d'optimisation sous contrainte et sa solution peuvent s'exprimer ainsi. Soit A une matrice symétrique et M une matrice symétrique, définie positive, les deux étant de taille $p \times p$.

Trouver une base M -orthonormée $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ de \mathbb{R}^p telle que $\mathbf{u}'_1A\mathbf{u}_1$ soit maximale, puis $\mathbf{u}'_2A\mathbf{u}_2$, etc.

Théorème 1.4. *La solution au problème d'optimisation sous contrainte est la base¹ M -orthonormée $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ formée des vecteurs propres de $M^{-1}A$, telle que \mathbf{u}_1 correspond à la plus grande valeur propre λ_1 de $M^{-1}A$, puis \mathbf{u}_2 à la deuxième valeur propre λ_2 de $M^{-1}A$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, p\}, \quad \mathbf{u}'_kA\mathbf{u}_k = \lambda_k.$$

1.9 EXERCICES

Exercice 1.1. [Applications linéaires]

Soient les espaces vectoriels \mathbb{R}^3 et \mathbb{R}^4 , munis de leurs bases canoniques respectives $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ et $\mathcal{B}' = \{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \tilde{\mathbf{e}}_3, \tilde{\mathbf{e}}_4\}$ respectivement. Soit f l'application linéaire de \mathbb{R}^3 dans \mathbb{R}^4 définie par :

$$\begin{aligned} f : \quad \mathbb{R}^3 &\rightarrow \mathbb{R}^4 \\ (x_1, x_2, x_3) &\mapsto f(x_1, x_2, x_3) = (y_1, y_2, y_3, y_4) = (x_1 - x_2, x_2 + x_3, x_1 + x_3, x_1 + x_2 + 2x_3). \end{aligned}$$

1. Écrire la matrice A de l'application linéaire f dans les bases \mathcal{B} et \mathcal{B}' .
2. Mettre la matrice A sous forme échelonnée.
3. En déduire le rang de f et la dimension du noyau de f .
4. Déterminer le noyau de f .

1. Cette base existe d'après le théorème 1.3 qui est la forme généralisée du théorème spectral pour les matrices symétriques.

Solution de l'exercice 1.1.

1. On a,

$$f(\mathbf{e}_1) = f \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \tilde{\mathbf{e}}_1 + \tilde{\mathbf{e}}_3 + \tilde{\mathbf{e}}_4.$$

De manière analogue,

$$f(\mathbf{e}_2) = f \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix} = -\tilde{\mathbf{e}}_1 + \tilde{\mathbf{e}}_2 + \tilde{\mathbf{e}}_4, \quad f(\mathbf{e}_3) = f \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix} = \tilde{\mathbf{e}}_2 + \tilde{\mathbf{e}}_3 + \tilde{\mathbf{e}}_4.$$

Ainsi, par définition de la matrice A de f dans les bases \mathcal{B} et \mathcal{B}' , on a :

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \text{ et } f(\mathbf{x}) = A\mathbf{x}.$$

2. En effectuant les opérations élémentaires :

- 3^{ème} ligne \rightarrow 3^{ème} - 1^{ère} ligne,
- 4^{ème} ligne \rightarrow 4^{ème} - 1^{ère} ligne, on obtient,

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{pmatrix}.$$

Puis, en effectuant les opérations élémentaires :

- 3^{ème} ligne \rightarrow 3^{ème} - 2^{ème} ligne,
- 4^{ème} ligne \rightarrow 4^{ème} - 2 \times 2^{ème} ligne, on obtient,

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Cette matrice est sous forme échelonnée, donc le rang de A est $\text{rg}(A) = 2$.

3. On a, $\text{rg}(f) = \text{rg}(A) = 2$. D'après le théorème du rang,

$$\dim(\ker(f)) = \dim(\mathbb{R}^3) - \text{rg}(f) = 3 - 2 = 1.$$

4. Le noyau de f est formé des \mathbf{x} de \mathbb{R}^3 tels que $f(\mathbf{x}) = \mathbf{0}$, c'est-à-dire de l'ensemble des

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \mathbb{R}^3, \text{ solutions de } f \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \text{ En utilisant la matrice sous forme échelonnée,}$$

il s'agit des solutions de $\begin{cases} x_1 - x_2 = 0 \\ x_2 + x_3 = 0 \end{cases}$, d'où $\ker(f) = \{\alpha(1, 1, -1) \mid \alpha \in \mathbb{R}\}$.

Si l'on ne remarque pas que l'on peut utiliser la matrice sous forme échelonnée, on peut

$$\text{aussi partir du système de départ : } \begin{cases} x_1 - x_2 = 0 \Leftrightarrow x_2 = x_1 \\ x_2 + x_3 = 0 \Leftrightarrow x_3 = -x_2 \\ x_1 + x_3 = 0 \\ x_1 + x_2 + 2x_3 = 0, \end{cases}$$

et on retrouve $\ker(f) = \{\alpha(1, 1, -1) \mid \alpha \in \mathbb{R}\}$.

Exercice 1.2. [Diagonalisation d'endomorphisme symétrique, changement de base]

Soit f un endomorphisme de \mathbb{R}^3 muni de sa base canonique \mathcal{B}_1 . La matrice A de f dans la base \mathcal{B}_1 est :

$$\begin{pmatrix} 3 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

1. Quel est le cardinal de \mathcal{B}_1 ?
2. Montrer que la matrice A est régulière (de déterminant non nul).
3. Soit $\lambda \in \mathbb{R}$, et la matrice $A(\lambda)$:

$$A(\lambda) = \begin{pmatrix} 3 - \lambda & 1 & 1 \\ 1 & 5 - \lambda & 1 \\ 1 & 1 & 3 - \lambda \end{pmatrix}$$

Calculer les valeurs de λ pour lesquelles $A(\lambda)$ n'est pas inversible.

4. Déterminer les valeurs propres et vecteurs propres de la matrice A . On choisira les vecteurs propres de sorte à ce qu'ils soient de norme 1.
5. Notons $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ l'ensemble des trois vecteurs propres de A de la question précédente. Montrer que \mathcal{B}_2 est une base orthonormée de \mathbb{R}^3 .
6. Déterminer la matrice B de l'application linéaire f relativement à la base \mathcal{B}_2 .
7. Calculer $\det(B)$ et $\text{Tr}(B)$. Comparer ces résultats à $\det(A)$ et $\text{Tr}(A)$.

Solution de l'exercice 1.2.

1. Le cardinal de \mathcal{B}_1 est la dimension de l'espace vectoriel \mathbb{R}^3 . Ainsi, $\text{card}(\mathcal{B}_1) = 3$.
2. Montrons que le déterminant de A est non nul. En effectuant l'opération : 3^{ème} ligne \rightarrow 3^{ème} ligne -2 ^{ème} ligne, qui n'affecte pas le déterminant, puis en développant par rapport

à la première colonne, on obtient :

$$\det(A) = \det \begin{pmatrix} 3 & 1 & 1 \\ 1 & 5 & 1 \\ 0 & -4 & 2 \end{pmatrix} = 3(10 + 4) - 1(2 + 4) = 36 \neq 0.$$

Ainsi, la matrice A est régulière.

3. Effectuons les mêmes opérations qu'à la question précédente pour calculer le déterminant de $A(\lambda)$:

$$\begin{aligned} \det(A(\lambda)) &= \det \begin{pmatrix} 3 - \lambda & 1 & 1 \\ 1 & 5 - \lambda & 1 \\ 0 & -4 + \lambda & 2 - \lambda \end{pmatrix} \\ &= (3 - \lambda)[(5 - \lambda)(2 - \lambda) + 4 - \lambda] - 1(2 - \lambda + 4 - \lambda) \\ &= (3 - \lambda)[\lambda^2 - 8\lambda + 14] - 2(3 - \lambda) = (3 - \lambda)(\lambda^2 - 8\lambda + 12) \\ &= (3 - \lambda)(\lambda - 2)(\lambda - 6). \end{aligned}$$

Ainsi, la matrice $A(\lambda)$ est non inversible pour $\lambda \in \{2, 3, 6\}$.

4. Soit un réel λ . Alors, λ est valeur propre de A si et seulement si

$$\det(A - \lambda I) = 0 \Leftrightarrow \det(A(\lambda)) = 0 \Leftrightarrow \lambda \in \{2, 3, 6\}.$$

On retrouve le fait que les valeurs propres d'une matrice symétrique sont réelles.

Notons $\lambda_1 = 2$, $\lambda_2 = 3$, $\lambda_3 = 6$ ces trois valeurs propres. Étant donné qu'elles sont distinctes, le sous-espace associé à chacune est de dimension exactement égal à 1.

Pour $k \in \{1, 2, 3\}$, notons $\mathbf{u}_k = \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix}$ un vecteur propre correspondant à la valeur propre

λ_k (donc unique à une constante multiplicative près). Alors \mathbf{u}_k est solution de

$$(A - \lambda_k I)\mathbf{u}_k = \mathbf{0} \Leftrightarrow A(\lambda_k)\mathbf{u}_k = \mathbf{0}.$$

Pour $k = 1$, le système linéaire $A(\lambda_1)\mathbf{u}_1 = \mathbf{0}$ s'écrit :

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

En effectuant les opérations suivantes sur la matrice $A(\lambda_1)$:

2^{ème} ligne \rightarrow 2^{ème} ligne $-$ 1^{ère} ligne ; 3^{ème} ligne \rightarrow 3^{ème} ligne $-$ 1^{ère} ligne,

ce système est équivalent au système suivant :

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

d'où $y_1 = 0$ et $x_1 + z_1 = 0$. L'ensemble des solutions est donc $\{\alpha(1, 0, -1) \mid \alpha \in \mathbb{R}\}$. On souhaite que le vecteur \mathbf{u}_1 soit normé à 1, d'où

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Pour $k = 2$, le système linéaire $A(\lambda_2)\mathbf{u}_2 = \mathbf{0}$ s'écrit :

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

En effectuant les opérations suivantes sur la matrice $A(\lambda_2)$:

2^{ème} ligne \rightarrow 2^{ème} ligne $-$ 3^{ème} ligne ;

ce système est équivalent au système suivant :

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

d'où $y_2 + z_2 = 0$ et $x_2 + y_2 = 0$. L'ensemble des solutions est donc $\{\alpha(1, -1, 1) \mid \alpha \in \mathbb{R}\}$. On souhaite que le vecteur \mathbf{u}_2 soit normé à 1, d'où

$$\mathbf{u}_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Pour $k = 3$, le système linéaire $A(\lambda_3)\mathbf{u}_3 = \mathbf{0}$ s'écrit :

$$\begin{pmatrix} -3 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -3 \end{pmatrix} \begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

En effectuant les opérations suivantes sur la matrice $A(\lambda_3)$:

1^{ère} ligne \rightarrow 1^{ère} ligne $+3 \times$ 2^{ème} ligne ; 3^{ème} ligne \rightarrow 3^{ème} ligne -2 ^{ème} ligne ;

ce système est équivalent au système suivant :

$$\begin{pmatrix} 0 & -2 & 4 \\ 1 & -1 & 1 \\ 0 & 2 & -4 \end{pmatrix} \begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

d'où $y_3 - 2z_3 = 0$ et $x_3 - y_3 + z_3 = 0$. L'ensemble des solutions est donc $\{\alpha(1, 2, 1) \mid \alpha \in \mathbb{R}\}$. On souhaite que le vecteur \mathbf{u}_3 soit normé à 1, d'où

$$\mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

5. L'ensemble $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ des trois vecteurs propres de A forment une famille libre. Comme \mathbb{R}^3 est de dimension 3, ils forment une base de \mathbb{R}^3 . Ils sont normés à 1 par construction, il reste donc à vérifier qu'ils sont deux-à-deux orthogonaux. Pour cela il suffit d'effectuer les produits scalaires suivants :

$$\begin{aligned} \langle \mathbf{u}_1, \mathbf{u}_2 \rangle &= \mathbf{u}'_1 \mathbf{u}_2 = 0 \\ \langle \mathbf{u}_2, \mathbf{u}_3 \rangle &= \mathbf{u}'_2 \mathbf{u}_3 = 0 \\ \langle \mathbf{u}_1, \mathbf{u}_3 \rangle &= \mathbf{u}'_1 \mathbf{u}_3 = 0. \end{aligned}$$

Ces trois produits scalaires étant nuls, on en déduit que les vecteurs sont orthogonaux deux-à-deux. On conclut que \mathcal{B} est une base orthonormée de \mathbb{R}^3 .

6. Soit U la matrice dont les colonnes sont les vecteurs $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$. Alors U est la matrice de passage de la base \mathcal{B}_1 à la base \mathcal{B}_2 . Les bases \mathcal{B}_1 et \mathcal{B}_2 étant orthonormales, la matrice U est orthogonale et $U^{-1} = U'$. Ainsi, la matrice de l'endomorphisme f dans la base \mathcal{B}_2 est :

$$B = U'AU = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

La matrice B est diagonale, car U est formée de vecteurs propres de f .

On a diagonalisé la matrice A symétrique dans une base orthonormale. D'après le théorème spectral 1.2, on sait que c'est toujours possible car pour les matrices symétriques, il existe une base orthonormée de \mathbb{R}^p formée de vecteurs propres de A .

7. On a $\det(B) = \det(A) = 36$ et $\text{Tr}(B) = \text{Tr}(A) = 11$. Ceci n'est pas étonnant car :

$$\begin{aligned} \det(B) &= \det(U^{-1}AU) = \det(AUU^{-1}) = \det(AI) = \det(A) \\ \text{Tr}(B) &= \text{Tr}(U^{-1}AU) = \text{Tr}(AUU^{-1}) = \text{Tr}(AI) = \text{Tr}(A). \end{aligned}$$

Exercice 1.3. [Matrices symétriques]

Soit X une matrice de taille $n \times p$. Montrer les points suivants.

1. Les matrices $X'X$ et XX' sont symétriques.
2. Les matrices $X'X$ et XX' ont les mêmes valeurs propres non nulles.
3. Si \mathbf{u} est vecteur propre pour une valeur propre non-nulle λ de $X'X$, alors $X\mathbf{u}$ est vecteur propre pour la même valeur propre λ de XX' . De manière symétrique, si \mathbf{v} est vecteur propre pour une valeur propre non-nulle λ de XX' , alors $X'\mathbf{v}$ est vecteur propre pour la même valeur propre de $X'X$.

Solution de l'exercice 1.3.

1. Par symétrie, il suffit de vérifier que XX' est symétrique. Pour cela, nous devons montrer que $(XX')' = XX'$. Or, d'après les propriétés de la transposition,

$$(XX')' = (X')' X' = XX',$$

d'où le résultat.

2. et 3. Soit λ une valeur propre non-nulle de $X'X$, alors il existe $\mathbf{u} \neq \mathbf{0}$, tel que :

$$X'X\mathbf{u} = \lambda\mathbf{u}.$$

En pré-multipliant les deux membres par X :

$$XX'X\mathbf{u} = \lambda X\mathbf{u} \Leftrightarrow XX'(X\mathbf{u}) = \lambda(X\mathbf{u}).$$

En remarquant que $X\mathbf{u}$ est non nul (sinon $\lambda = 0$ d'après la première équation, ce qui contredit notre hypothèse), on déduit que $X\mathbf{u}$ est vecteur propre de XX' relatif à la même valeur propre λ . Ainsi, λ est aussi valeur propre de XX' , de vecteur propre $X\mathbf{u}$. En répétant le même argument pour μ valeur propre non nulle de XX' , on déduit que μ est aussi valeur propre de $X'X$. Ainsi, les matrices $X'X$ et XX' ont mêmes valeurs propres non-nulles. De plus, si \mathbf{u} est vecteur propre de valeur propre λ de $X'X$, alors $X\mathbf{u}$ est vecteur propre de valeur propre λ de XX' , et par symétrie, si \mathbf{v} est vecteur propre de valeur propre λ de XX' , alors $X'\mathbf{v}$ est vecteur propre de valeur propre λ de $X'X$.

Remarquer que la matrice XX' est de taille $n \times n$ et la matrice $X'X$ est de taille $p \times p$. Elles sont les deux symétriques et sont donc les deux diagonalisables. Comme elles ont les mêmes valeurs propres non-nulles, il suffit de calculer celles de la plus petite des deux.

Exercice 1.4. [Métriques euclidiennes] Soit les vecteurs suivants de \mathbb{R}^3 :

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

1. Faire une représentation géométrique de ces deux vecteurs.
2. On suppose \mathbb{R}^3 muni de la métrique usuelle. Calculer la norme de ces vecteurs, et le cosinus de l'angle entre ces deux vecteurs.
3. Même question lorsque l'on munit \mathbb{R}^3 de la métrique provenant de la matrice M :

$$M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

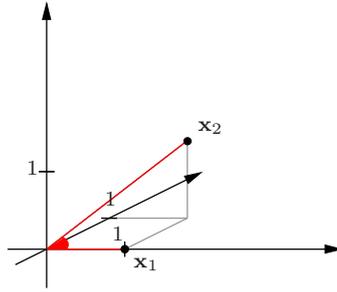
On justifiera que la matrice M est symétrique et définie positive.

Solution de l'exercice 1.4.

- 1.
2. Par définition,

$$\begin{aligned} \|\mathbf{x}_1\|^2 &= \langle \mathbf{x}_1, \mathbf{x}_1 \rangle = \mathbf{x}_1' \mathbf{x}_1 = 1 \\ \|\mathbf{x}_2\|^2 &= \langle \mathbf{x}_2, \mathbf{x}_2 \rangle = \mathbf{x}_2' \mathbf{x}_2 = 3 \end{aligned}$$

ainsi, $\|\mathbf{x}_1\| = 1$, $\|\mathbf{x}_2\| = \sqrt{3}$.



Le produit scalaire entre \mathbf{x}_1 et \mathbf{x}_2 est :

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 1,$$

d'où le cosinus

$$\cos \angle(\mathbf{0x}_1, \mathbf{0x}_2) = \frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} = \frac{1}{\sqrt{3}}.$$

3. La matrice étant diagonale, elle est symétrique. De plus, ses valeurs propres sont données par ses coefficients diagonaux. Étant donné qu'ils sont tous strictement positifs, on en déduit que la matrice M est définie positive.

Par définition,

$$\|\mathbf{x}_1\|_M^2 = \langle \mathbf{x}_1, \mathbf{x}_1 \rangle_M = \mathbf{x}_1' M \mathbf{x}_1 = 2$$

$$\|\mathbf{x}_2\|_M^2 = \langle \mathbf{x}_2, \mathbf{x}_2 \rangle_M = \mathbf{x}_2' M \mathbf{x}_2 = 6$$

ainsi, $\|\mathbf{x}_1\| = \sqrt{2}$, $\|\mathbf{x}_2\| = \sqrt{6}$.

Le produit scalaire entre \mathbf{x}_1 et \mathbf{x}_2 est :

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_M = \mathbf{x}_1' M \mathbf{x}_2 = 2,$$

d'où le cosinus

$$\cos_M \angle(\mathbf{0x}_1, \mathbf{0x}_2) = \frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_M}{\|\mathbf{x}_1\|_M \|\mathbf{x}_2\|_M} = \frac{2}{\sqrt{12}} = \frac{1}{\sqrt{3}}.$$

Exercice 1.5. [Projection orthogonale]

Démontrer que la projection M -orthogonale $\pi(\mathbf{x})$ d'un point \mathbf{x} de \mathbb{R}^n sur une droite \mathcal{D} de \mathbb{R}^n , passant par l'origine et de vecteur directeur \mathbf{u} , est donnée par :

$$\pi(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{u} \rangle_M}{\|\mathbf{u}\|_M} \frac{\mathbf{u}}{\|\mathbf{u}\|_M}.$$

Solution de l'exercice 1.5. Par définition, le point $\pi(\mathbf{x})$ satisfait aux conditions suivantes :

- Il appartient à la droite \mathcal{D} , ce qui se traduit mathématiquement par, $\pi(\mathbf{x}) = \lambda \mathbf{u}$, pour un $\lambda \in \mathbb{R}$.
- Le vecteur $\mathbf{x} - \pi(\mathbf{x})$ est M -orthogonal au vecteur \mathbf{u} , ce qui se traduit par,

$$\langle \mathbf{x} - \pi(\mathbf{x}), \mathbf{u} \rangle_M = 0.$$

La deuxième condition implique que, $\langle \mathbf{x}, \mathbf{u} \rangle_M = \langle \pi(\mathbf{x}), \mathbf{u} \rangle_M$. Remplaçant $\pi(\mathbf{x})$ par $\lambda \mathbf{u}$, on a alors :

$$\langle \mathbf{x}, \mathbf{u} \rangle_M = \langle \lambda \mathbf{u}, \mathbf{u} \rangle_M = \lambda \|\mathbf{u}\|_M^2,$$

d'où $\lambda = \frac{\langle \mathbf{x}, \mathbf{u} \rangle_M}{\|\mathbf{u}\|_M^2}$. On conclut en utilisant la première condition et en remplaçant λ par la valeur trouvée.

Exercice 1.6. [Projection orthogonale]

Soit les vecteurs suivants de \mathbb{R}^2 :

$$\mathbf{x} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Calculer la projection orthogonale du point \mathbf{x} sur la droite \mathcal{D} passant par l'origine et de vecteur directeur \mathbf{u} , lorsque \mathbb{R}^2 est muni

1. de la métrique usuelle,
2. de la métrique provenant de la matrice M :

$$M = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}.$$

Dans les deux cas, faire un dessin du projeté orthogonal.

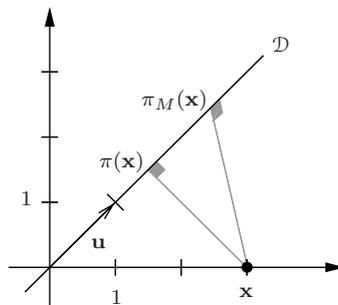
Solution de l'exercice 1.6.

1. On a, $\|\mathbf{u}\|^2 = 2$, $\langle \mathbf{x}, \mathbf{u} \rangle = 3$. Ainsi, la projection $\pi(\mathbf{x})$ de \mathbf{x} sur la droite \mathcal{D} est :

$$\pi(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} \mathbf{u} = \frac{3}{2} \mathbf{u}.$$

2. On a, $\|\mathbf{u}\|_M^2 = 5$, $\langle \mathbf{x}, \mathbf{u} \rangle_M = 12$. Ainsi, la projection $\pi_M(\mathbf{x})$ de \mathbf{x} sur la droite \mathcal{D} est :

$$\pi_M(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{u} \rangle_M}{\|\mathbf{u}\|_M^2} \mathbf{u} = \frac{12}{5} \mathbf{u}.$$



Exercice 1.7. [Gradient]

Soit A une matrice symétrique de taille $p \times p$, et soit f l'application de \mathbb{R}^p dans \mathbb{R} définie par, $f(\mathbf{u}) = \mathbf{u}'A\mathbf{u}$. Montrer que $\nabla f(\mathbf{u}) = 2A\mathbf{u}$, où ∇ représente le gradient, c'est-à-dire le vecteur de \mathbb{R}^p formé des dérivées partielles de f .

Solution de l'exercice 1.7. Pour tout $\mathbf{u}' = (u_1, \dots, u_p) \in \mathbb{R}^p$, on a

$$\begin{aligned} f(\mathbf{u}) &= \sum_{i,j=1}^p a_{ij}u_iu_j = \sum_{i=1}^p a_{ii}u_i^2 + \sum_{i \neq j} a_{ij}u_iu_j \\ &= \sum_{i=1}^p a_{ii}u_i^2 + 2 \sum_{i < j} a_{ij}u_iu_j, \text{ car la matrice } A \text{ est symétrique.} \end{aligned}$$

Cette application est \mathcal{C}^∞ , donc elle admet des dérivées partielles de tout ordre. Soit $k \in \{1, \dots, p\}$, alors :

$$\begin{aligned} \frac{\partial f}{\partial u_k}(\mathbf{u}) &= 2a_{kk}u_k + 2 \sum_{i \neq k} a_{ki}u_i \\ &= 2 \sum_{i=1}^p a_{ki}u_i = 2(A\mathbf{u})_k. \end{aligned}$$

Ainsi,

$$\nabla f(\mathbf{u}) = \begin{pmatrix} \frac{\partial f}{\partial u_1}(\mathbf{u}) \\ \vdots \\ \frac{\partial f}{\partial u_p}(\mathbf{u}) \end{pmatrix} = 2A\mathbf{u}.$$

CHAPITRE 2

ANALYSE EN AXES PRINCIPAUX : PRINCIPES DE BASE

2.1 IDÉE GÉNÉRALE

2.1.1 LES DONNÉES

On considère un échantillon I formé de n individus : $\text{card}(I) = n$. Sur cet échantillon, on s'intéresse à étudier p caractères/variables numériques, indexées par un ensemble J : $\text{card}(J) = p$. On suppose les données homogènes, c'est-à-dire exprimées dans la même unité.

Les données sont représentées dans une matrice $X = (x_{ij})$ de taille $n \times p$.

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

- Pour $i \in \{1, \dots, n\}$, la $i^{\text{ème}}$ ligne, notée \mathbf{x}'_i , représente les données de toutes les variables pour le $i^{\text{ème}}$ individu.
- Pour $j \in \{1, \dots, p\}$, la $j^{\text{ème}}$ colonne, notée \mathbf{x}^j , représente les données de la $j^{\text{ème}}$ variable pour tous les individus.
- Le coefficient x_{ij} est la donnée de la $j^{\text{ème}}$ variable pour le $i^{\text{ème}}$ individu.

2.1.2 REPRÉSENTATION GRAPHIQUE

Selon que l'on considère les lignes ou les colonnes de la matrice X , on peut définir deux nuages de points, un dans \mathbb{R}^p , l'autre dans \mathbb{R}^n . On note $N(I)$ le nuage des n points-individus défini dans l'espace des variables \mathbb{R}^p et $N(J)$ le nuage des p points-variables défini dans l'espace des individus \mathbb{R}^n :

$$N(I) = \{\mathbf{x}_i \in \mathbb{R}^p \mid i \in \{1, \dots, n\}\}, \quad N(J) = \{\mathbf{x}^j \in \mathbb{R}^n \mid j \in \{1, \dots, p\}\}.$$

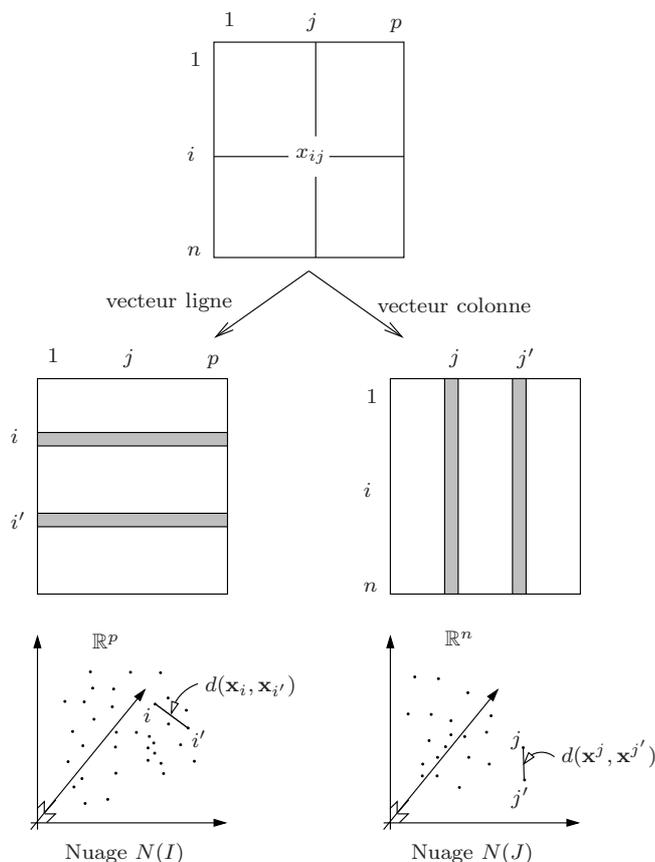


FIGURE 2.1 – Représentation graphique des données.

Les dimensions n et p sont grandes en général et les représentations ci-dessus ne sont donc pas visibles. Le but est d'obtenir des nuages de points dans des espaces de plus petite dimension (1, 2 ou 3) en perdant le moins d'information possible. Pour ce faire, nous allons projeter les données sur une droite, de sorte à perdre le moins d'information, puis sur une autre orthogonale à la première, etc. On observera alors la/les projections des données.

2.1.3 CRITÈRES DE PROJECTION

Supposons que l'on ait un nuage de points $N(I)$ dans \mathbb{R}^p , une droite \mathcal{D} et une notion de distance dans \mathbb{R}^p . Pour visualiser le nuage, on va en considérer une projection sur \mathcal{D} . Afin que la projection du nuage et le nuage soit le plus semblable possible, il est naturel de demander que la somme des carrés des distances entre les projections orthogonales de toutes les paires de points soit maximale. Ceci nous donne un critère pour déterminer la droite \mathcal{D} . Une fois obtenu une première droite de projection, on pourra en chercher une deuxième etc.

Nous verrons, lors du chapitre 4 sur l'Analyse en Composante Principale, que ceci revient à maximiser la somme des carrés des distances entre les projections orthogonales des points et du centre de gravité du nuage.

Dans la section suivante, nous supposons que la droite \mathcal{D} passe par l'origine, et considérons

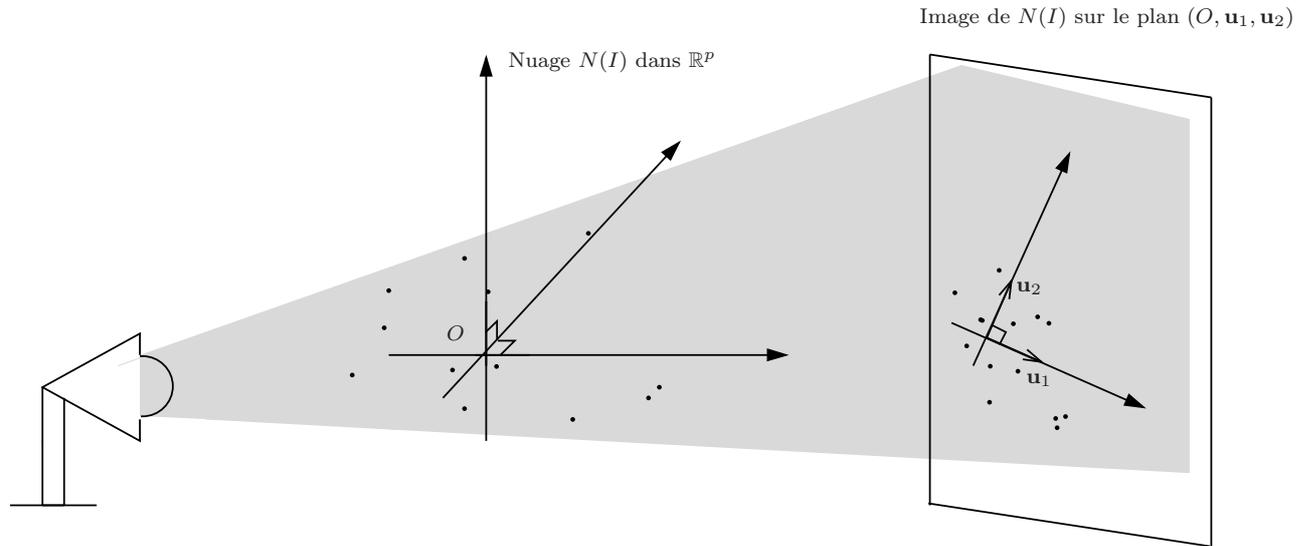


FIGURE 2.2 – Projection des données : on oriente la source de lumière de sorte à ce que le nuage projeté soit le moins déformé possible, c'est-à-dire de sorte à ce que les distances entre les projections des paires de points soient aussi proche que possible des distances entre les paires de points du nuage $N(I)$.

un autre critère de projection : le critère des moindres carrés. Nous verrons que cela revient à maximiser la somme des carrés des distances entre l'origine et les projections orthogonales des points. Lorsque le centre de gravité du nuage est l'origine du repère, nous nous retrouvons donc dans le cas précédent. Il n'est en fait par restrictif de se placer dans ce cadre, car si ce n'est pas le cas, il suffit de translater le nuage de points de sorte à ce que l'origine soit le centre de gravité, puis d'ajuster par une droite qui passe par l'origine.

Ce problème possède une symétrie. En effet, il semble tout aussi naturel de chercher à projeter le nuage de points $N(J)$ dans \mathbb{R}^p . Nous verrons que les projections des nuages $N(I)$ et $N(J)$ sont liées. Nous expliquerons également comment reconstruire les données à partir des projections.

2.2 AJUSTEMENT DU NUAGE DES INDIVIDUS DANS L'ESPACE DES VARIABLES : PRINCIPE DES MOINDRES CARRÉS

On considère le nuages des n points individus dans \mathbb{R}^p : $N(I) = \{\mathbf{x}_i \in \mathbb{R}^p \mid i \in \{1, \dots, n\}\}$.

2.2.1 ÉTAPE 1 : AJUSTEMENT DU NUAGE $N(I)$ PAR UNE DROITE \mathcal{D}_1

Supposons que la distance entre individus est la distance d induite par le produit scalaire usuel $\langle \cdot, \cdot \rangle$ sur \mathbb{R}^p . Pour toute paire d'individus $\mathbf{x}_i, \mathbf{x}_{i'}$:

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 = \sum_{j=1}^p (x_{ij} - x_{i'j})^2.$$

Supposons que l'on souhaite ajuster le nuage de points $N(I)$ par une droite \mathcal{D}_1 qui passe par l'origine. Notons \mathbf{u}_1 le vecteur unitaire qui engendre cette droite.

Le critère d'ajustement que nous retenons est le critère des *moindre carrés*. Nous demandons que la droite \mathcal{D}_1 soit le plus proche possible de l'ensemble des points, c'est-à-dire que la somme des carrés des distances entre les points et leur projection orthogonale sur la droite \mathcal{D}_1 soit minimale.

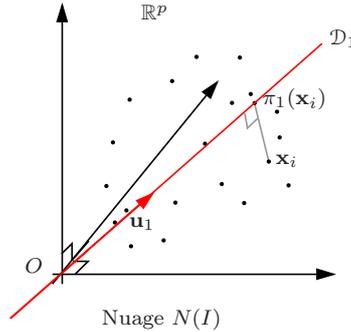


FIGURE 2.3 – Projection des données sur la droite \mathcal{D}_1 passant par l'origine et de vecteur directeur \mathbf{u}_1 .

Pour tout individu \mathbf{x}_i , notons $\pi_1(\mathbf{x}_i)$ la projection orthogonale de \mathbf{x}_i sur \mathcal{D}_1 . Alors, nous souhaitons trouver le vecteur unitaire \mathbf{u}_1 tel que la quantité suivante soit minimale :

$$\sum_{i=1}^n d^2(\mathbf{x}_i, \pi_1(\mathbf{x}_i)) = \sum_{i=1}^n \|\mathbf{x}_i - \pi_1(\mathbf{x}_i)\|^2. \quad (2.1)$$

D'après le théorème de Pythagore nous avons, pour chacun des individus,

$$\|\mathbf{x}_i\|^2 = \|\pi_1(\mathbf{x}_i)\|^2 + \|\mathbf{x}_i - \pi_1(\mathbf{x}_i)\|^2.$$

En conséquence, notre critère d'ajustement est équivalent à trouver le vecteur unitaire \mathbf{u}_1 tel que la quantité suivante soit maximale :

$$\sum_{i=1}^n \|\pi_1(\mathbf{x}_i)\|^2 = \sum_{i=1}^n d^2(\mathbf{0}, \pi_1(\mathbf{x}_i)).$$

Lemme 2.1. *On a l'égalité suivante :*

$$\sum_{i=1}^n d^2(\mathbf{0}, \pi_1(\mathbf{x}_i)) = \mathbf{u}'_1 X' X \mathbf{u}_1.$$

Démonstration. D'après la section 1.5.4 nous savons, puisque le vecteur \mathbf{u}_1 est unitaire, que pour tout $i \in \{1, \dots, n\}$,

$$\pi_1(\mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{u}_1 \rangle \mathbf{u}_1 = \mathbf{u}'_1 \mathbf{x}_i \cdot \mathbf{u}_1 = \mathbf{x}'_i \mathbf{u}_1 \cdot \mathbf{u}_1,$$

d'où, $\|\pi_1(\mathbf{x}_i)\|^2 = \langle \mathbf{x}_i, \mathbf{u}_1 \rangle^2 = \mathbf{u}'_1 \mathbf{x}_i \mathbf{x}'_i \mathbf{u}_1$. Ainsi,

$$\sum_{i=1}^n d^2(\mathbf{0}, \pi_1(\mathbf{x}_i)) = \sum_{i=1}^n \|\pi_1(\mathbf{x}_i)\|^2 = \sum_{i=1}^n \mathbf{u}'_1 \mathbf{x}_i \mathbf{x}'_i \mathbf{u}_1 = \mathbf{u}'_1 \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right) \mathbf{u}_1 = \mathbf{u}'_1 X' X \mathbf{u}_1,$$

où dans la dernière égalité nous avons écrit la somme sous forme matricielle. \square

En conclusion, nous cherchons \mathbf{u}_1 solution du problème d'optimisation sous contrainte suivant :

$$\begin{cases} \mathbf{u}'X'X\mathbf{u} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\| = 1 \text{ est satisfaite.} \end{cases}$$

En posant $M = I$ et $A = X'X$ dans la section 1.8, nous déduisons la proposition suivante.

Proposition 2.1. *Le vecteur \mathbf{u}_1 qui satisfait le problème d'optimisation sous contrainte est le vecteur propre unitaire correspondant à la plus grande valeur propre λ_1 de la matrice $X'X$. On a alors, $\mathbf{u}_1'X'X\mathbf{u}_1 = \lambda_1$.*

La droite \mathcal{D}_1 de vecteur directeur \mathbf{u}_1 est le *premier axe principal* ou *premier axe factoriel*. Comme,

$$\lambda_1 = \mathbf{u}_1'X'X\mathbf{u}_1 = \sum_{i=1}^n \|\pi_1(\mathbf{x}_i)\|^2,$$

la valeur propre λ_1 mesure la dispersion de la projection du nuage des points-individus sur le premier axe d'ajustement \mathcal{D}_1 .

2.2.2 ÉTAPE 2 : DEUXIÈME DIRECTION D'AJUSTEMENT DU NUAGE $N(I)$

Nous cherchons maintenant une droite \mathcal{D}_2 , de vecteur directeur unitaire \mathbf{u}_2 , orthogonale à la première, qui ajuste au mieux le nuage de points au sens des moindres carrés.

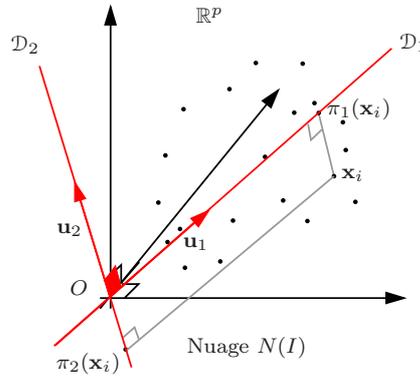


FIGURE 2.4 – Projection des données sur la droite \mathcal{D}_2 , orthogonale à \mathcal{D}_1 , passant par l'origine et de vecteur directeur \mathbf{u}_2 .

Notre problème d'optimisation devient : Trouver le vecteur \mathbf{u}_2 solution du problème suivant :

$$\begin{cases} \mathbf{u}'X'X\mathbf{u} \text{ est maximale} \\ \text{sous les contraintes } \|\mathbf{u}\| = 1, \langle \mathbf{u}_1, \mathbf{u} \rangle = 0. \end{cases}$$

En prenant $M = I$ et $A = X'X$ dans la section 1.8, nous déduisons la proposition suivante.

Proposition 2.2. *Le vecteur \mathbf{u}_2 qui satisfait le problème d'optimisation sous contrainte est le vecteur propre unitaire correspondant à la deuxième plus grande valeur propre (comptée avec multiplicité) de la matrice $X'X$. On a alors, $\mathbf{u}'_2 X'X \mathbf{u}_2 = \lambda_2$.*

La droite \mathcal{D}_2 de vecteur directeur \mathbf{u}_2 est la *deuxième axe principal* ou *deuxième axe factoriel*. La valeur propre λ_2 mesure la dispersion de la projection du nuage des points-individus sur le deuxième axe d'ajustement \mathcal{D}_2 .

2.2.3 AJUSTEMENT DU NUAGE $N(I)$ SOUS LE CRITÈRE DES MOINDRES CARRÉS

Pour tout $k \in \{1, \dots, p\}$, étant donné un vecteur unitaire \mathbf{u}_k , notons π_k la projection orthogonale sur la droite \mathcal{D}_k passant par l'origine et de vecteur directeur \mathbf{u}_k .

En itérant les étapes 1 et 2, la question de l'ajustement du nuage de points sous le critère des moindres carrés peut s'écrire de la manière suivante :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ tel que $\sum_{i=1}^n d^2(\mathbf{0}, \pi_1(\mathbf{x}_i))$ soit maximale, puis $\sum_{i=1}^n d^2(\mathbf{0}, \pi_2(\mathbf{x}_i))$, etc.

Autrement dit,

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ de \mathbb{R}^p tel que $\mathbf{u}'_1 X'X \mathbf{u}_1$ soit maximal, puis $\mathbf{u}'_2 X'X \mathbf{u}_2$, etc.

D'après le théorème 1.4 de la section 1.8, nous avons :

Théorème 2.1. *La solution au problème d'optimisation sous contrainte est le repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ formée des vecteurs propres de $X'X$, telle que \mathbf{u}_1 correspond à la plus grande valeur propre λ_1 de $X'X$, puis \mathbf{u}_2 à la deuxième valeur propre λ_2 de $X'X$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, p\}, \quad \sum_{i=1}^n d^2(\mathbf{0}, \pi_k(\mathbf{x}_i)) = \mathbf{u}'_k X'X \mathbf{u}_k = \lambda_k.$$

Définition 2.1. Les axes $\mathcal{D}_1, \dots, \mathcal{D}_p$, passant par l'origine, de vecteur directeur $\mathbf{u}_1, \dots, \mathbf{u}_p$ respectivement, sont les *axes principaux* ou *axes factoriels*, respectivement le premier, ..., le p -ième. Chaque vecteur \mathbf{u}_k est combinaison linéaire des vecteurs $\mathbf{x}^1, \dots, \mathbf{x}^p$, ce qui signifie que \mathbf{u}_k représente une nouvelle variable "abstraite" combinaison des variables d'origine.

Remarque 2.1. En itérant le théorème de Pythagore, nous avons, pour tout $i \in \{1, \dots, n\}$,

$$\|\mathbf{x}_i\|^2 = d^2(\mathbf{0}, \mathbf{x}_i) = \sum_{k=1}^p d^2(\mathbf{0}, \pi_k(\mathbf{x}_i)) = \sum_{k=1}^p \|\pi_k(\mathbf{x}_i)\|^2.$$

Pour tout $k \in \{1, \dots, p\}$, la valeur propre λ_k mesure la dispersion de la projection du nuage des points-individus sur le $k^{\text{ème}}$ axe d'ajustement \mathcal{D}_k .

2.2.4 PROJECTION DES INDIVIDUS SUR LES NOUVELLES VARIABLES

Pour $i \in \{1, \dots, n\}$, $k \in \{1, \dots, p\}$, la coordonnée ψ_{ik} de l'individu i sur l'axe k est

$$\psi_{ik} = \langle \mathbf{x}_i, \mathbf{u}_k \rangle = \mathbf{x}'_i \mathbf{u}_k.$$

Définissons la matrice $U = (\mathbf{u}_1 \dots \mathbf{u}_p)$ dont les colonnes sont les vecteurs de la nouvelle base. Alors, on peut décrire les coordonnées des individus dans le nouveau repère dans la matrice Ψ :

$$\Psi = XU,$$

où la $i^{\text{ème}}$ ligne $\Psi'_i = \mathbf{x}'_i U$ représente les coordonnées du $i^{\text{ème}}$ individu sur les p axes factoriels $\mathbf{u}_1, \dots, \mathbf{u}_p$; la $k^{\text{ème}}$ colonne $\Psi^k = X\mathbf{u}_k$ représente les coordonnées de tous les individus sur le $k^{\text{ème}}$ axe factoriel.

Remarquons que la base $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ étant orthonormale, la matrice U est orthogonale et donc $UU' = U'U = I_p$.

2.3 AJUSTEMENT DU NUAGE DES VARIABLES DANS L'ESPACE DES INDIVIDUS

On considère le nuage des p points variables dans \mathbb{R}^n : $N(J) = \{\mathbf{x}^j \in \mathbb{R}^n \mid j \in \{1, \dots, p\}\}$. Tout se passe de manière analogue au cas du nuage des n points individus dans \mathbb{R}^p .

2.3.1 AJUSTEMENT DU NUAGE $N(J)$ SOUS LE CRITÈRE DES MOINDRES CARRÉS

Supposons que la distance entre variables soit la distance d induite par le produit scalaire usuel $\langle \cdot, \cdot \rangle$ sur \mathbb{R}^n . Pour toute paire de variables $\mathbf{x}^j, \mathbf{x}^{j'}$:

$$d^2(\mathbf{x}^j, \mathbf{x}^{j'}) = \|\mathbf{x}^j - \mathbf{x}^{j'}\|^2 = \sum_{i=1}^n (x_{ij} - x_{ij'})^2. \quad (2.2)$$

On souhaite ajuster le nuage de points $N(J)$ par une droite \tilde{D}_1 , de vecteur directeur \mathbf{v}_1 qui passe par l'origine, sous le critère des moindres carrés. Pour toute variable \mathbf{x}^j , notons $\tilde{\pi}_1(\mathbf{x}^j)$ la projection orthogonale de \mathbf{x}^j sur \tilde{D}_1 . Nous souhaitons trouver le vecteur unitaire \mathbf{v}_1 tel que la quantité suivante soit minimale :

$$\sum_{j=1}^p d^2(\mathbf{x}^j, \tilde{\pi}_1(\mathbf{x}^j)) = \sum_{j=1}^p \|\mathbf{x}^j - \tilde{\pi}_1(\mathbf{x}^j)\|^2.$$

Utilisant le théorème de Pythagore pour chacune des variables, puis écrivant la projection avec le produit scalaire, nous déduisons que notre critère d'ajustement est équivalent à trouver le vecteur unitaire \mathbf{v}_1 tel que la quantité suivante soit maximale :

$$\sum_{j=1}^p \|\tilde{\pi}_1(\mathbf{x}^j)\|^2 = \sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_1(\mathbf{x}^j)) = \sum_{j=1}^p \langle \mathbf{x}^j, \mathbf{v}_1 \rangle^2 = \mathbf{v}'_1 X X' \mathbf{v}_1.$$

En conclusion, nous cherchons \mathbf{v}_1 solution du problème suivant :

$$\begin{cases} \mathbf{v}'XX'\mathbf{v} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{v}\| = 1 \text{ est satisfaite.} \end{cases}$$

En prenant $M = I$ et $A = XX'$, on se retrouve dans le cadre de l'optimisation d'une forme quadratique sous contrainte quadratique de la section 1.8.

En itérant les étapes, la question de l'ajustement du nuage de points-variables sous le critère des moindres carrés peut s'écrire de la manière suivante :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ tel que $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_1(\mathbf{x}^j))$ soit maximale, puis $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_2(\mathbf{x}^j))$, etc.

Autrement dit,

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ tel que $\mathbf{v}'_1XX'\mathbf{v}_1$ soit maximal, puis $\mathbf{v}'_2XX'\mathbf{v}_2$, etc.

D'après le théorème 1.4 de la section 1.8, nous avons :

Théorème 2.2. *La solution au problème d'optimisation sous contrainte est le repère orthonormé $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ formée des vecteurs propres de XX' , telle que \mathbf{v}_1 correspond à la plus grande valeur propre μ_1 de XX' , puis \mathbf{v}_2 à la deuxième valeur propre μ_2 de XX' , etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, n\}, \quad \sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_k(\mathbf{x}^j)) = \mathbf{v}'_kXX'\mathbf{v}_k = \mu_k.$$

2.3.2 PROJECTION DES VARIABLES SUR LES NOUVEAUX INDIVIDUS

Pour $j \in \{1, \dots, p\}$, $k \in \{1, \dots, n\}$, la coordonnée ϕ_{jk} de la variable j sur l'axe k est

$$\phi_{jk} = \langle \mathbf{x}^j, \mathbf{v}_k \rangle = (\mathbf{x}^j)' \mathbf{v}_k$$

Définissons la matrice $V = (\mathbf{v}_1 \dots \mathbf{v}_n)$ dont les colonnes sont les vecteurs de la nouvelle base. Alors, on peut décrire les coordonnées des variables dans le nouveau repère dans la matrice Φ :

$$\Phi = X'V,$$

où la $j^{\text{ème}}$ ligne $\Phi'_j = (\mathbf{x}^j)'V$ représente les coordonnées de la $j^{\text{ème}}$ variable sur les n axes factoriels $\mathbf{v}_1, \dots, \mathbf{v}_n$; la $k^{\text{ème}}$ colonne $\Phi^k = X'\mathbf{v}_k$ représente les coordonnées de toutes les variables sur le $k^{\text{ème}}$ axe factoriel.

La base $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ étant orthonormale, la matrice V est orthogonale et donc $VV' = V'V = I_n$.

2.4 RELATION ENTRE LES AJUSTEMENTS DANS LES DEUX ESPACES

Il existe des relations fortes entre les ajustements dans les deux espaces. Le but de cette section est de les décrire.

Nous avons montré les résultats suivants dans l'exercice 1.3.

1. Les matrices $X'X$ et XX' ont les mêmes valeurs propres non-nulles.
2. Si \mathbf{u} est vecteur propre pour une valeur propre non-nulle λ de $X'X$, alors $X\mathbf{u}$ est vecteur propre pour la même valeur propre λ de XX' . De manière symétrique, si \mathbf{v} est vecteur propre pour une valeur propre non-nulle λ de XX' , alors $X'\mathbf{v}$ est vecteur propre pour la même valeur propre de $X'X$.

Notons $\lambda_1 \geq \dots \geq \lambda_q > 0$ les valeurs propres non-nulles de $X'X$ (et donc de XX'), comptées avec multiplicité. Remarquer que l'on a forcément $q \leq \min\{p, n\}$.

Pour $k \in \{1, \dots, q\}$, soit \mathbf{u}_k le vecteur propre unitaire correspondant à la valeur propre λ_k de $X'X$, et soit \mathbf{v}_k le vecteur propre unitaire correspondant à la valeur propre λ_k de XX' .

D'après le point 2 ci-dessus, on a que $X\mathbf{u}_k$ est vecteur propre pour la valeur propre λ_k de XX' . Comme \mathbf{v}_k est le vecteur propre unitaire de XX' , correspondant à la même valeur propre λ_k , on déduit que :

$$\mathbf{v}_k = \frac{X\mathbf{u}_k}{\|X\mathbf{u}_k\|}.$$

Or, $\|X\mathbf{u}_k\|^2 = \mathbf{u}_k' X' X \mathbf{u}_k = \lambda_k$, d'après le théorème 2.1, d'où $\mathbf{v}_k = \frac{X\mathbf{u}_k}{\sqrt{\lambda_k}}$. Par un argument symétrique, on a $\mathbf{u}_k = \frac{X'\mathbf{v}_k}{\sqrt{\lambda_k}}$.

Utilisant de plus les notations des sections 2.2.4, 2.3.2 pour les projections, nous avons démontré le résultat suivant.

Théorème 2.3. *Soient $\lambda_1 \geq \dots \geq \lambda_q > 0$ les valeurs propres non-nulles de $X'X$. Pour $k \in \{1, \dots, q\}$, soit \mathbf{u}_k , respectivement \mathbf{v}_k , le vecteur propre unitaire correspondant à la valeur propre λ_k de $X'X$, respectivement XX' . Alors, ces vecteurs propres sont liés par les relations :*

$$\begin{cases} \mathbf{v}_k = \frac{X\mathbf{u}_k}{\sqrt{\lambda_k}} = \frac{\Psi^k}{\sqrt{\lambda_k}} \\ \mathbf{u}_k = \frac{X'\mathbf{v}_k}{\sqrt{\lambda_k}} = \frac{\Phi^k}{\sqrt{\lambda_k}}. \end{cases}$$

On déduit les relations suivantes pour les coordonnées des projections :

$$\Psi^k = X\mathbf{u}_k = \sqrt{\lambda_k} \mathbf{v}_k, \quad (2.3)$$

$$\Phi^k = X'\mathbf{v}_k = \sqrt{\lambda_k} \mathbf{u}_k. \quad (2.4)$$

Si les vecteurs propres correspondent à des valeurs propres nulles, on les choisit de sorte à compléter la base orthonormée formée par les axes précédents.

2.5 RECONSTRUCTION DES DONNÉES

Nous repartons de l'égalité, pour tout $k \in \{1, \dots, p\}$,

$$\Psi^k = X\mathbf{u}_k = \sqrt{\lambda_k} \mathbf{v}_k.$$

2.5.1 RECONSTRUCTION EXACTE

On poste multiplie les deux membres de cette égalité par \mathbf{u}'_k et on somme sur k , alors :

$$X \left(\sum_{k=1}^p \mathbf{u}_k \mathbf{u}'_k \right) = \sum_{k=1}^p \sqrt{\lambda_k} \mathbf{v}_k \mathbf{u}'_k.$$

Or, $\sum_{k=1}^p \mathbf{u}_k \mathbf{u}'_k = UU'$ et la matrice U étant orthogonale, on a $UU' = I_p$. On obtient donc la formule de reconstitution du tableau des données X :

$$X = \sum_{k=1}^p \sqrt{\lambda_k} \mathbf{v}_k \mathbf{u}'_k. \quad (2.5)$$

Il suffit de considérer la somme sur les indices correspondants aux valeurs propres non-nulles. La matrice X est écrite de façon unique comme somme de matrices de rang 1 (produit d'une matrice ligne par une matrice colonne). Il s'agit de la décomposition aux valeurs singulières de la matrice.

2.5.2 RECONSTITUTION APPROCHÉE

Les valeurs propres étant rangées dans l'ordre décroissant, on peut approcher la matrice X par la somme précédente prise sur ℓ termes uniquement, $\ell \in \{1, \dots, q\}$.

La qualité de la reconstitution τ_ℓ est donnée par le rapport :

$$\tau_\ell = \frac{\sum_{k=1}^{\ell} \lambda_k}{\sum_{k=1}^p \lambda_k}.$$

Ce rapport représente la proportion de la dispersion des données expliquée par les données approchées.

2.6 ÉLÉMENTS SUPPLÉMENTAIRES

La construction des nouveaux repères se base sur les éléments dits *actifs*. L'analyse factorielle générale permet aussi de positionner des éléments n'ayant pas participé à la construction, appelés éléments *supplémentaires* ou *illustratifs*. Par exemple, pour les variables, il peut s'agir de variables qui n'ont pas la même unité que les autres, et donc empêche d'avoir un ensemble de variables homogènes.

Étant donné un individu supplémentaire \mathbf{x}_+ , ses coordonnées sur le p axes factoriels $\mathbf{u}_1, \dots, \mathbf{u}_p$ (représentant des combinaisons linéaires des variables d'origine) sont données dans le vecteur :

$$\Psi'_+ = \mathbf{x}'_+ U. \quad (2.6)$$

Étant donné une variable supplémentaire \mathbf{x}^+ , ses coordonnées sur les n axes factoriels $\mathbf{v}_1, \dots, \mathbf{v}_n$ (représentant des combinaisons linéaires des individus d'origine) sont données dans le vecteur :

$$\Phi'_+ = (\mathbf{x}^+)' V. \quad (2.7)$$

2.7 GÉNÉRALISATIONS : INDIVIDUS PONDÉRÉS

Si les données n'ont pas été récoltées dans les mêmes conditions pour tous les individus, on peut souhaiter les pondérer. Dans ce cas, on leur attribue des poids relatifs $(p_i)_{i=1}^n$, tels que $\sum_{i=1}^n p_i = 1$. Ces poids sont encodés dans la matrice diagonale D , de taille $n \times n$,

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & p_n \end{pmatrix}.$$

Notons $D^{\frac{1}{2}}$ la matrice dont la diagonale contient les $(\sqrt{p_i})_{i=1}^n$. Alors, on a $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$.

2.7.1 AJUSTEMENT DU NUAGE $N(I)$ SOUS LE CRITÈRE DES MOINDRES CARRÉS

Le critère d'ajustement que nous retenons est encore le critère des moindres carrés, mais nous pondérons maintenant chaque individu par son poids relatif.

Soit \mathcal{D}_1 la droite passant par l'origine et de vecteur directeur \mathbf{u}_1 . Nous souhaitons trouver le vecteur unitaire \mathbf{u}_1 tel que la quantité suivante soit minimale :

$$\sum_{i=1}^n p_i d^2(\mathbf{x}_i, \pi_1(\mathbf{x}_i)) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \pi_1(\mathbf{x}_i)\|^2.$$

De manière analogue au cas non pondéré, utilisant le théorème de Pythagore pour chacun des individus, notre critère d'ajustement est équivalent à trouver le vecteur unitaire \mathbf{u}_1 tel que la quantité suivante soit maximale :

$$\sum_{i=1}^n p_i \|\pi_1(\mathbf{x}_i)\|^2 = \sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_1(\mathbf{x}_i)).$$

Lemme 2.2. *On a l'égalité suivante :*

$$\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_1(\mathbf{x}_i)) = \mathbf{u}_1' X' D X \mathbf{u}_1.$$

Démonstration. Exprimant la projection avec le produit scalaire, voir section 1.5.4, nous obtenons

nous :

$$\begin{aligned}
 \sum_{i=1}^n p_i \|\pi_1(\mathbf{x}_i)\|^2 &= \sum_{i=1}^n p_i \langle \mathbf{x}_i, \mathbf{u}_1 \rangle^2 \\
 &= \sum_{i=1}^n \langle \sqrt{p_i} \mathbf{x}_i, \mathbf{u}_1 \rangle^2, \text{ d'après la linéarité du produit scalaire} \\
 &= \sum_{i=1}^n \langle (D^{\frac{1}{2}} X)_i, \mathbf{u}_1 \rangle^2 = \sum_{i=1}^n \mathbf{u}_1' (D^{\frac{1}{2}} X)_i (D^{\frac{1}{2}} X)_i' \mathbf{u}_1 \\
 &= \mathbf{u}_1' \left[\sum_{i=1}^n (D^{\frac{1}{2}} X)_i (D^{\frac{1}{2}} X)_i' \right] \mathbf{u}_1 \\
 &= \mathbf{u}_1' (D^{\frac{1}{2}} X)' (D^{\frac{1}{2}} X) \mathbf{u}_1 = \mathbf{u}_1' X' D X \mathbf{u}_1. \quad \square
 \end{aligned}$$

Nous nous retrouvons donc dans le cadre l'analyse générale non pondérée, avec comme matrice de travail $D^{\frac{1}{2}} X$, et nous cherchons \mathbf{u}_1 solution du problème suivant :

$$\begin{cases} \mathbf{u}' X' D X \mathbf{u} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\| = 1 \text{ est satisfaite.} \end{cases}$$

En itérant les étapes, la question de l'ajustement du nuage de points sous le critère des moindres carrés, lorsque les données sont pondérées, peut s'écrire de la manière suivante :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ tel que $\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_1(\mathbf{x}_i))$ soit maximale, puis $\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_2(\mathbf{x}_i))$, etc.

Autrement dit,

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ tel que $\mathbf{u}_1' X' D X \mathbf{u}_1$ soit maximal, puis $\mathbf{u}_2' X' D X \mathbf{u}_2$, etc.

D'après le théorème 1.4 de la section 1.8, nous avons :

Théorème 2.4. *La solution au problème d'optimisation sous contrainte est le repère orthonormé $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ formé des vecteurs propres de $X' D X$, telle que \mathbf{u}_1 correspond à la plus grande valeur propre λ_1 de $X' D X$, puis \mathbf{u}_2 à la deuxième valeur propre λ_2 de $X' D X$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, p\}, \quad \sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_k(\mathbf{x}_i)) = \mathbf{u}_k' X' D X \mathbf{u}_k = \lambda_k.$$

2.7.2 PROJECTION DES INDIVIDUS SUR LES NOUVELLES VARIABLES

Les coordonnées des individus dans le nouveau repère sont décrites dans la matrice Ψ :

$$\Psi = XU,$$

où la $i^{\text{ème}}$ ligne $\Psi_i' = \mathbf{x}_i' U$ représente les coordonnées du $i^{\text{ème}}$ individu sur les p axes factoriels $\mathbf{u}_1, \dots, \mathbf{u}_p$; la $k^{\text{ème}}$ colonne $\Psi^k = X \mathbf{u}_k$ représente les coordonnées de tous les individus sur le $k^{\text{ème}}$ axe factoriel.

Remarque 2.2. Les projections des individus dans le nouveau repère sont définies comme dans le cas non-pondéré. En effet, la pondération change l'importance que l'on donne aux individus dans le critère d'optimisation, mais ne change pas les données elles-mêmes.

2.7.3 AJUSTEMENT DU NUAGE $N(J)$ SOUS LE CRITÈRE DES MOINDRES CARRÉS

Nous avons vu qu'ajuster les individus pondérés revenait à l'ajustement non-pondéré de la matrice de travail $W = D^{\frac{1}{2}}X$.

La distance entre variables (transformées avec la matrice $D^{\frac{1}{2}}$ devient :

$$\begin{aligned} d^2((D^{\frac{1}{2}}X)^j, (D^{\frac{1}{2}}X)^{j'}) &= \sum_{i=1}^n [(D^{\frac{1}{2}}X)_{ij} - (D^{\frac{1}{2}}X)_{ij'}]^2 \\ &= \sum_{i=1}^n (\sqrt{p_i}x_{ij} - \sqrt{p_i}x_{ij'})^2 = \sum_{i=1}^n p_i(x_{ij} - x_{ij'})^2. \end{aligned}$$

Le problème d'ajustement devient :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ tel que $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_1[(D^{\frac{1}{2}}\mathbf{x})^j])$ soit maximale, puis $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_2[(D^{\frac{1}{2}}\mathbf{x})^j])$, etc.

Autrement dit,

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ tel que $\mathbf{v}'_1 D^{\frac{1}{2}} X X' D^{\frac{1}{2}} \mathbf{v}_1$ soit maximal, puis $\mathbf{v}'_2 D^{\frac{1}{2}} X X' D^{\frac{1}{2}} \mathbf{v}_2$, etc.
--

Et la solution est le théorème 1.4 avec la matrice de travail $D^{\frac{1}{2}}X$:

Théorème 2.5. *La solution au problème d'optimisation sous contrainte est la repère orthonormé $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ formé des vecteurs propres de $D^{\frac{1}{2}} X X' D^{\frac{1}{2}}$, telle que \mathbf{v}_1 correspond à la plus grande valeur propre μ_1 de $D^{\frac{1}{2}}$, puis \mathbf{u}_2 à la deuxième valeur propre μ_2 de $D^{\frac{1}{2}}$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, n\}, \quad \sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_k[(D^{\frac{1}{2}}\mathbf{x})^j]) = \mathbf{v}'_k D^{\frac{1}{2}} X X' D^{\frac{1}{2}} \mathbf{v}_k = \mathbf{v}'_k D^{\frac{1}{2}} X X' D^{\frac{1}{2}} \mathbf{v}_k = \mu_k.$$

2.7.4 PROJECTION DES VARIABLES SUR LES NOUVEAUX INDIVIDUS

Les coordonnées des variables dans le nouveau repère sont décrites dans la matrice Φ :

$$\Phi = (D^{\frac{1}{2}}X)'V,$$

où la $j^{\text{ème}}$ ligne $\Phi'_j = (D^{\frac{1}{2}}\mathbf{x}^j)'V$ représente les coordonnées de la $j^{\text{ème}}$ variable sur les n axes factoriels $\mathbf{v}_1, \dots, \mathbf{v}_n$; la $k^{\text{ème}}$ colonne $\Phi^k = (D^{\frac{1}{2}}X)' \mathbf{v}_k$ représente les coordonnées de toutes les variables sur le $k^{\text{ème}}$ axe factoriel.

Remarque 2.3. Lors de la projection des variables, nous considérons la matrice de travail $D^{\frac{1}{2}}X$. Ceci permet d'avoir des relations de transition entre les ajustements dans les deux espaces, fait qui sera utilisé en Analyse en Composantes Principales, voir chapitre 4.

2.7.5 RELATION ENTRE LES AJUSTEMENTS DANS LES DEUX ESPACES

Avec les choix que nous avons faits, il existe encore des relations entre les ajustements dans les deux espaces.

Le théorème 2.3 énoncé avec la matrice de travail $D^{\frac{1}{2}}X$ devient,

Théorème 2.6. *Soient $\lambda_1 \geq \dots \geq \lambda_q > 0$ les valeurs propres non-nulles de $X'DX$. Pour $k \in \{1, \dots, q\}$, soit \mathbf{u}_k , respectivement \mathbf{v}_k , le vecteur propre unitaire correspondant à la valeur propre λ_k de $X'DX$, respectivement $D^{\frac{1}{2}}XX'D^{\frac{1}{2}}$. Alors, ces vecteurs propres sont liés par les relations :*

$$\begin{cases} \mathbf{v}_k = \frac{D^{\frac{1}{2}}X\mathbf{u}_k}{\sqrt{\lambda_k}} = \frac{D^{\frac{1}{2}}\Psi^k}{\sqrt{\lambda_k}} \\ \mathbf{u}_k = \frac{(D^{\frac{1}{2}}X)'\mathbf{v}_k}{\sqrt{\lambda_k}} = \frac{\Phi^k}{\sqrt{\lambda_k}}. \end{cases}$$

On déduit les relations suivantes pour les coordonnées des projections :

$$\Psi^k = X\mathbf{u}_k = \sqrt{\lambda_k}D^{-\frac{1}{2}}\mathbf{v}_k, \quad (2.8)$$

$$\Phi^k = (D^{\frac{1}{2}}X)'\mathbf{v}_k = \sqrt{\lambda_k}\mathbf{u}_k. \quad (2.9)$$

2.8 EXERCICES

Exercice 2.1. La consommation de lait, de pain et de beurre (exprimée en dizaines d'euros) du mois dernier, observée sur un échantillon de 7 ménages, est donnée dans le tableau suivant :

	\mathbf{x}^1 (lait)	\mathbf{x}^2 (pain)	\mathbf{x}^3 (beurre)
\mathbf{x}'_1	2	1	1
\mathbf{x}'_2	6	6	2
\mathbf{x}'_3	5	3	1
\mathbf{x}'_4	3	3	1
\mathbf{x}'_5	3	5	3
\mathbf{x}'_6	1	2	2
\mathbf{x}'_7	3	9	10

On suppose que la variable \mathbf{x}^3 et l'individu \mathbf{x}_7 sont illustratifs.

1. Posons

$$X = \begin{pmatrix} 2 & 1 \\ 6 & 6 \\ 5 & 3 \\ 3 & 3 \\ 3 & 5 \\ 1 & 2 \end{pmatrix}.$$

Pourquoi la matrice $X'X$ est-elle diagonalisable ?

2. Calculer les valeurs propres et les vecteurs propres de la matrice $X'X$.
3. En déduire une nouvelle base orthonormée $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2\}$ telle que les droites $\mathcal{D}_1, \mathcal{D}_2$ passant pas l'origine, de vecteurs directeurs $\mathbf{u}_1, \mathbf{u}_2$, soient les plus proches de tous les points du nuage des individus $N(I)$ au sens des moindres carrés.
4. Calculer les coordonnées du nuage de points-individus $N(I)$ dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$.
5. Représenter le nuage des points-individus $N(I)$ relativement aux deux repères.
6. Ajuster, au sens des moindres carrés, le nuage $N(J)$ des points-variables dans un repère orthonormé $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$.
7. Déterminer les coordonnées du nuage des points-variables $N(J)$ dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$.
8. Déterminer les coordonnées de l'individu illustratif dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$ et de la variable supplémentaire dans le repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$.
9. Reconstituez les données X de manière exacte, puis de manière approchée.

Solution de l'exercice 2.1.

1. On a montré dans l'exercice 1.3 que la matrice $X'X$ est symétrique, donc diagonalisable. Ses valeurs propres sont réelles et il existe une base de vecteurs propres orthogonaux. Si les valeurs propres sont toutes distinctes, alors les vecteurs propres associés sont orthogonaux. La matrice $X'X$ est :

$$X'X = \begin{pmatrix} 84 & 79 \\ 79 & 84 \end{pmatrix}.$$

2. On remarque que la somme de chacune des lignes de $X'X$ est égale à 163, ce qui revient à écrire,

$$X'X \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 84 & 79 \\ 79 & 84 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 163 \\ 163 \end{pmatrix} = 163 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

D'où $\lambda_1 = 163$ est valeur propre de $X'X$ et $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)'$ est un vecteur propre unitaire correspondant.

Pour déterminer la deuxième valeur propre λ_2 , on se rappelle que $\text{Tr}(X'X) = \lambda_1 + \lambda_2$, d'où

$$\lambda_2 = \text{Tr}(X'X) - \lambda_1 = 168 - 163 = 5.$$

Un vecteur propre \mathbf{u}_2 associé est orthogonal à \mathbf{u}_1 , soit de la forme $\alpha(-1, 1)$. En imposant que \mathbf{u}_2 soit normé à 1, on obtient $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(-1, 1)'$ (ou son opposé).

3. D'après le cours, la base $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2\}$ formée des deux vecteurs propres associés aux valeurs propres de la matrice $X'X$ est la base recherchée. Soit \mathcal{D}_1 , respectivement \mathcal{D}_2 , la droite passant par l'origine et de vecteur directeur unitaire \mathbf{u}_1 , respectivement \mathbf{u}_2 . Alors, la base \mathcal{B}_2 est telle que la somme des carrés des distances des points-individus à la droite \mathcal{D}_1 est minimale, puis la même chose pour \mathcal{D}_2 .

4. La coordonnée ψ_{ik} de l'individu \mathbf{x}_i sur l'axe \mathcal{D}_k est $\langle \mathbf{x}_i, \mathbf{u}_k \rangle$. Par exemple, les coordonnées du premier individu sont,

$$\psi_{11} = \langle \mathbf{x}_1, \mathbf{u}_1 \rangle = \frac{1}{\sqrt{2}}(2 + 1) = \frac{3}{\sqrt{2}}, \quad \psi_{12} = \langle \mathbf{x}_1, \mathbf{u}_2 \rangle = \frac{1}{\sqrt{2}}(-2 + 1) = -\frac{1}{\sqrt{2}}.$$

On peut résumer toutes les coordonnées dans la matrice $\Psi = XU$, où U est la matrice dont les colonnes sont les vecteurs \mathbf{u}_1 et \mathbf{u}_2 :

$$\Psi = XU = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 1 \\ 6 & 6 \\ 5 & 3 \\ 3 & 3 \\ 3 & 5 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 3 & -1 \\ 12 & 0 \\ 8 & -2 \\ 6 & 0 \\ 8 & 2 \\ 3 & 1 \end{pmatrix}$$

La $i^{\text{ème}}$ ligne représente les coordonnées de l'individu i dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$.

5. Représentation graphique.

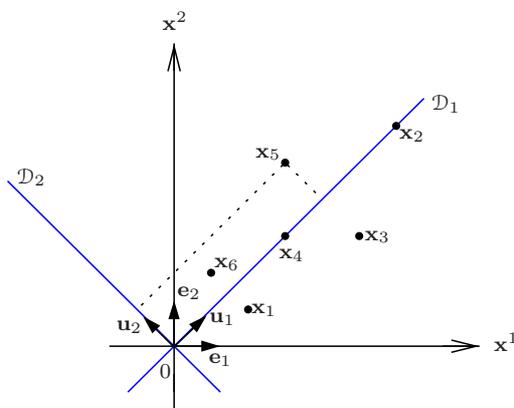


FIGURE 2.5 – Représentation graphique.

6. La matrice de travail est XX' . D'après le cours, ses valeurs propres non nulles sont les mêmes que celles de la matrice $X'X$, soit λ_1 et λ_2 de la question 2.

Les vecteurs unitaires $\{\mathbf{v}_1, \mathbf{v}_2\}$ de la nouvelle base obtenue par ajustement des points-variables $N(J)$ sous le critère des moindres carrés, sont les vecteurs propres de la matrice XX' . D'après le théorème 2.3 donnant les formules de transitions, ces vecteurs peuvent être calculés ainsi :

$$\forall k \in \{1, 2\}, \quad \mathbf{v}_k = \frac{\Psi^k}{\sqrt{\lambda_k}}.$$

Ainsi, d'après les questions 2 et 4 :

$$\mathbf{v}_1 = \frac{\Psi^1}{\sqrt{163}} = \frac{1}{\sqrt{2 \cdot 163}} \begin{pmatrix} 3 \\ 12 \\ 8 \\ 6 \\ 8 \\ 3 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{\Psi^2}{\sqrt{5}} = \frac{1}{\sqrt{2 \cdot 5}} \begin{pmatrix} -1 \\ 0 \\ -2 \\ 0 \\ 2 \\ 1 \end{pmatrix}.$$

7. Les coordonnées du nuage des points-variables $N(J)$ dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$ sont répertoriées dans la matrice $\Phi = X'V$. D'après la formule (2.4), pour $k \in \{1, 2\}$,

$$\Phi^k = \sqrt{\lambda_k} \mathbf{u}_k.$$

Ainsi,

$$\Phi^1 = \frac{\sqrt{163}}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Phi^2 = \frac{\sqrt{5}}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \text{et}$$

$$\Phi = (\Phi^1 \ \Phi^2) = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{163} & -\sqrt{5} \\ \sqrt{163} & \sqrt{5} \end{pmatrix},$$

où la première (resp. deuxième) ligne est formée des coordonnées de la première (resp. deuxième) variable dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$, chaque axe étant combinaison linéaire des individus de départ.

8. D'après la formule (2.6), les coordonnées de l'individu supplémentaire \mathbf{x}_7 dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$ sont données par :

$$\Psi'_7 = \mathbf{x}'_7 U = \frac{1}{\sqrt{2}} (3 \ 9) \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{\sqrt{2}} (12 \ 6).$$

D'après la formule (2.6), les coordonnées de la variable supplémentaire \mathbf{x}^3 dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$ sont données par :

$$(\Phi_3)' = (\mathbf{x}^3)'V = \begin{pmatrix} 1 & 2 & 1 & 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} \frac{3}{\sqrt{326}} & \frac{-1}{\sqrt{10}} \\ \frac{12}{\sqrt{326}} & 0 \\ \frac{8}{\sqrt{326}} & \frac{-2}{\sqrt{10}} \\ \frac{6}{\sqrt{326}} & 0 \\ \frac{8}{\sqrt{326}} & \frac{2}{\sqrt{10}} \\ \frac{3}{\sqrt{326}} & \frac{1}{\sqrt{10}} \end{pmatrix} = \begin{pmatrix} \frac{71}{\sqrt{326}} & \frac{5}{\sqrt{10}} \end{pmatrix}.$$

9. D'après la formule de reconstitution (2.5), et puisqu'il n'y a que deux valeurs propres λ_1, λ_2 , non-nulles, on a :

$$X = \sum_{k=1}^2 \sqrt{\lambda_k} \mathbf{v}_k \mathbf{u}'_k = \sum_{k=1}^2 \Psi^k \mathbf{u}'_k.$$

Explicitement, ceci donne :

$$\begin{aligned} \sum_{k=1}^2 \sqrt{\lambda_k} \mathbf{v}_k \mathbf{u}'_k &= \frac{1}{\sqrt{2}} \begin{pmatrix} 3 \\ 12 \\ 8 \\ 6 \\ 8 \\ 3 \end{pmatrix} \frac{1}{\sqrt{2}} (1 \ 1) + \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ -2 \\ 0 \\ 2 \\ 1 \end{pmatrix} \frac{1}{\sqrt{2}} (-1 \ 1) \\ &= \frac{1}{2} \begin{pmatrix} 3 & 3 \\ 12 & 12 \\ 8 & 8 \\ 6 & 6 \\ 8 & 8 \\ 3 & 3 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 0 & 0 \\ 2 & -2 \\ 0 & 0 \\ -2 & 2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 6 & 6 \\ 5 & 3 \\ 3 & 3 \\ 3 & 5 \\ 1 & 2 \end{pmatrix} = X. \end{aligned}$$

La reconstitution à l'ordre 1 est :

$$\frac{1}{2} \begin{pmatrix} 3 & 3 \\ 12 & 12 \\ 8 & 8 \\ 6 & 6 \\ 8 & 8 \\ 3 & 3 \end{pmatrix}.$$

La qualité de cette reconstruction est :

$$\tau_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{163}{163 + 5} \sim 0,97.$$

Exercice 2.2. On considère 4 individus $\mathbf{x}_1, \dots, \mathbf{x}_4$ et 2 variables $\mathbf{x}^1, \mathbf{x}^2$ supposées homogènes (elles s'expriment dans la même unité). On suppose qu'à chaque individu est associé un poids, et que l'individu 4 est illustratif (il a poids 0).

	\mathbf{x}^1	\mathbf{x}^2	Poids
\mathbf{x}'_1	5	3	2
\mathbf{x}'_2	2	6	1
\mathbf{x}'_3	4	1	7
\mathbf{x}'_4	7	4	0

Ajuster, au sens des moindres carrés, le nuage des points-individus $N(I)$ dans un nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$. Pour cela on abordera les points suivants :

1. Déterminer les poids relatifs de chacun des individus.

2. Déterminer la matrice X à considérer et la racine $D^{\frac{1}{2}}$ de la matrice des poids relatifs.
3. Déterminer la matrice de travail et la matrice à diagonaliser.
4. Calculer les valeurs propres et vecteurs propres de la matrice à diagonaliser.
5. En déduire une nouvelle base orthonormée $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2\}$ telle que les droites $\mathcal{D}_1, \mathcal{D}_2$ passant pas l'origine, de vecteurs directeurs $\mathbf{u}_1, \mathbf{u}_2$, soient les plus proches de tous les points du nuage des individus $N(I)$ au sens des moindres carrés.
6. Calculer les coordonnées du nuage des points-individus $N(I)$ dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$. Calculer aussi les coordonnées de l'individu supplémentaire \mathbf{x}_4 .
7. Calculer les coordonnées du nuage des points-variables $N(J)$ dans le repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$ provenant de l'ajustement des points-variables $N(J)$ sous le critère des moindres carrés.

Solution de l'exercice 2.2.

1. La somme des poids est égale à $2+1+7=10$. Donc les poids relatifs sont :

$$p_1 = \frac{2}{10}, p_2 = \frac{1}{10}, p_3 = \frac{7}{10}, p_4 = 0.$$

2. Dans la matrice X que l'on retient, on ne tient compte, ni de la troisième variable car elle est illustrative, ni du quatrième individu car il a poids nul. Les poids relatifs ont été calculés à la question précédente, ce qui nous permet de déterminer la matrice $D^{\frac{1}{2}}$. On a,

$$X = \begin{pmatrix} 5 & 3 \\ 2 & 6 \\ 4 & 1 \end{pmatrix}; D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{0.2} & 0 & 0 \\ 0 & \sqrt{0.1} & 0 \\ 0 & 0 & \sqrt{0.7} \end{pmatrix}.$$

3. La matrice de travail est $W = D^{\frac{1}{2}}X$ et la matrice à diagonaliser est $W'W = X'DX$:

$$W = \begin{pmatrix} \sqrt{0.2} & 0 & 0 \\ 0 & \sqrt{0.1} & 0 \\ 0 & 0 & \sqrt{0.7} \end{pmatrix} \begin{pmatrix} 5 & 3 \\ 2 & 6 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 5\sqrt{0.2} & 3\sqrt{0.2} \\ 2\sqrt{0.1} & 6\sqrt{0.1} \\ 4\sqrt{0.7} & 1\sqrt{0.7} \end{pmatrix}$$

$$W'W = \begin{pmatrix} 5 & 2 & 4 \\ 3 & 6 & 1 \end{pmatrix} \begin{pmatrix} 0.2 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \begin{pmatrix} 5 & 3 \\ 2 & 6 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 16.6 & 7 \\ 7 & 6.1 \end{pmatrix}$$

4. Les valeurs propres de la matrice $W'W$ sont solutions de l'équation, $\det(W'W - \lambda I) = 0$, c'est-à-dire :

$$\det \begin{pmatrix} 16.6 - \lambda & 7 \\ 7 & 6.1 - \lambda \end{pmatrix} = (16.6 - \lambda)(6.1 - \lambda) - 49 = \lambda^2 - 22.7\lambda + 52.26 = 0.$$

Les deux valeurs propres sont : $\lambda_{1,2} = \frac{1}{2}(22.7 \pm \sqrt{22.7^2 - 4 \cdot 52.26}) \simeq \frac{1}{2}(22.7 \pm 17.5)$. Ainsi, on a $\lambda_1 \simeq 20.1$, $\lambda_2 \simeq 2.6$.

Remarque : la somme des deux valeurs propres est égale à la trace de la matrice :

$$\lambda_1 + \lambda_2 = \text{Tr}(W'W) = 16.6 + 6.1 = 22.7.$$

C'est une manière de vérifier que l'on ne s'est pas trompé dans le calcul des valeurs propres, ou de déduire la deuxième une fois que l'on connaît la première. Pour $k \in \{1, 2\}$, soit $\mathbf{u}_k = (x_k, y_k)'$ le vecteur propre de norme 1 associé à la valeur propre λ_k . Alors, \mathbf{u}_k est solution de $(A - \lambda_k I)\mathbf{u}_k = 0$. Pour $k = 1$, on a,

$$(A - \lambda_1 I)\mathbf{u}_1 = 0 \Leftrightarrow \begin{pmatrix} -3.5 & 7 \\ 7 & -14 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow x_1 = 2y_1.$$

Comme on veut que le vecteur \mathbf{u}_1 soit normé à 1, on a $\mathbf{u}_1 = \frac{1}{\sqrt{3}}(2, 1)'$ (ou son opposé).

La matrice $W'W$ étant symétrique, on sait que le vecteur \mathbf{u}_2 est orthogonal à \mathbf{u}_1 , soit de la forme $\alpha(1, -2)$. En imposant que \mathbf{u}_2 soit normé à 1, on obtient $\mathbf{u}_2 = \frac{1}{\sqrt{3}}(1, -2)'$ (ou son opposé).

- D'après le cours, la base $\mathcal{B}_2 = \{\mathbf{u}_1, \mathbf{u}_2\}$ formée des deux vecteurs propres associés aux valeurs propres de la matrice $X'X$ est la base recherchée. Soit \mathcal{D}_1 , respectivement \mathcal{D}_2 , la droite passant par l'origine et de vecteur directeur unitaire \mathbf{u}_1 , respectivement \mathbf{u}_2 . Alors, la base \mathcal{B}_2 est telle que la somme des carrés des distances des points-individus à la droite \mathcal{D}_1 est minimale, puis la même chose pour \mathcal{D}_2 .
- Les coordonnées du nuage des points-individus $N(I)$ dans les nouveau repère sont encodées dans la matrice $\Psi = XU$, où U est la matrice dont les colonnes sont les vecteurs $\mathbf{u}_1, \mathbf{u}_2$:

$$\Psi = XU = X = \frac{1}{\sqrt{3}} \begin{pmatrix} 5 & 3 \\ 2 & 6 \\ 4 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} 13 & -1 \\ 10 & -10 \\ 9 & 2 \end{pmatrix}.$$

La $i^{\text{ème}}$ ligne représente les coordonnées de l'individu i dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$. Par exemple, les coordonnées du premier individu dans le nouveau repère sont,

$$\Psi'_1 = \frac{1}{\sqrt{3}}(13, -1).$$

Les coordonnées Ψ'_4 de l'individu supplémentaire sont données par :

$$\Psi'_4 = \mathbf{x}'_4 U = \frac{1}{\sqrt{3}}(7, 4) \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} = \frac{1}{\sqrt{3}}(18, -1).$$

- Les coordonnées du nuage des points-variables $N(J)$ dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$ sont répertoriées dans la matrice $\Phi = (D^{\frac{1}{2}}X)'V$. D'après la formule (2.9), pour $k \in \{1, 2\}$,

$$\Phi^k = \sqrt{\lambda_k} \mathbf{u}_k.$$

Ainsi,

$$\Phi^1 = \frac{\sqrt{20.1}}{\sqrt{3}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \Phi^2 = \frac{\sqrt{2.6}}{\sqrt{3}} \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \text{ et}$$

$$\Phi = (\Phi^1 \ \Phi^2) = \frac{1}{\sqrt{3}} \begin{pmatrix} 2\sqrt{20.1} & \sqrt{2.6} \\ \sqrt{20.1} & -2\sqrt{2.6} \end{pmatrix},$$

où la première (resp. deuxième) ligne est formée des coordonnées de la première (resp. deuxième) variable dans le nouveau repère $(\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2)$, chaque axe étant combinaison linéaire des individus de départ.

CHAPITRE 3

STATISTIQUE DESCRIPTIVE ÉLÉMENTAIRE

On considère un échantillon I formé de n individus : $\text{card}(I) = n$. Sur cet échantillon, on s'intéresse à étudier p caractères/variables numériques, indexées par un ensemble J : $\text{card}(J) = p$. Cette section a pour objectif de rappeler le premier ensemble de manipulations faites sur les données. Il s'agit de calculer les paramètres de position, dispersion et relation entre les variables.

3.1 MATRICE DES POIDS

Si les données ont été récoltées dans les mêmes conditions pour les n individus, on leur attribue un même poids relatif $\frac{1}{n}$. Il se peut que ce ne soit pas le cas, auquel cas on leur attribue des poids relatifs $(p_i)_{i=1}^n$, tels que $\sum_{i=1}^n p_i = 1$.

Notons D la matrice diagonale de taille $n \times n$ dont les éléments diagonaux sont p_1, \dots, p_n . Remarque que lorsque, pour tout $i \in \{1, \dots, n\}$, $p_i = \frac{1}{n}$, on retrouve les poids relatifs homogènes.

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & p_n \end{pmatrix}.$$

La matrice D étant symétrique et définie positive, elle induit un produit scalaire $\langle \cdot, \cdot \rangle_D$ sur \mathbb{R}^n . Pour tout $\mathbf{x} = (x_1, \dots, x_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$ de \mathbb{R}^n ,

$$\langle \mathbf{x}, \mathbf{y} \rangle_D = \mathbf{x}' D \mathbf{y} = \sum_{i=1}^n p_i x_i y_i.$$

Associé à ce produit scalaire, il y a une norme, une distance et un cosinus, voir chapitre 1.

3.2 LE CAS D'UNE VARIABLE

Supposons que l'on étudie un *caractère/variable* de type numérique sur l'échantillon I formé de n individus ($p = 1$). On dispose du vecteur $\mathbf{x}' = (x_1, \dots, x_n)$ des données de la variable pour chacun des individus, où x_i est la donnée de la variable pour le $i^{\text{ème}}$ individu.

3.2.1 REPRÉSENTATION GRAPHIQUE

Les données de la variable peuvent être représentées par un nuage de points, noté $N(I)$, formé de n points dans \mathbb{R} :

$$N(I) = \{x_i \in \mathbb{R} \mid i \in \{1, \dots, n\}\}.$$

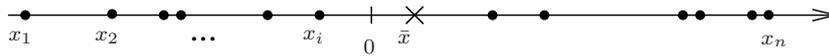


FIGURE 3.1 – Représentation du nuage de points $N(I)$ et de la moyenne arithmétique \bar{x} .

Remarque 3.1. Selon le type des données, il y a d'autres types de représentations : camembert, diagramme en bâtons, histogramme, boîte à moustaches aussi connue sous le nom de box-plot.

3.2.2 PARAMÈTRES DE POSITION

Il s'agit de définir une valeur autour de laquelle se répartissent les observations. Les plus utilisées sont la *moyenne arithmétique*, la *médiane* et le *mode*. Nous nous restreignons ici à présenter la moyenne arithmétique, notée \bar{x} .

Si chacun des individus a même poids relatif $\frac{1}{n}$, la *moyenne arithmétique* est définie par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Si les individus ont des poids relatifs $(p_i)_{i=1}^n$, $\sum_{i=1}^n p_i = 1$, la *moyenne arithmétique* est :

$$\bar{x} = \sum_{i=1}^n p_i x_i.$$

Notations avec le produit scalaire sur \mathbb{R}^n :

$$\boxed{\bar{x} = \mathbf{x}' D \mathbf{1} = \langle \mathbf{x}, \mathbf{1} \rangle_D .}$$

La moyenne arithmétique est le *barycentre* ou *centre de gravité* du nuage de points $N(I)$, voir Figure 3.1.

Les *données centrées* sont les données dont on a enlevé la moyenne. Elles sont enregistrées dans le vecteur \mathbf{y} :

$$\mathbf{y} = (x_1 - \bar{x}, \dots, x_n - \bar{x})' = (\mathbf{x} - \bar{x} \mathbf{1}).$$

Propriété 3.1. *Les données centrées ont moyenne arithmétique nulle.*

Exercice 3.1. Démontrer la propriété 3.1 en utilisant les notations faisant intervenir le produit scalaire.

Solution de l'exercice 3.1. Par définition,

$$\begin{aligned} \bar{y} &= \langle \mathbf{y}, \mathbf{1} \rangle_D = \langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{1} \rangle_D \\ &= \langle \mathbf{x}, \mathbf{1} \rangle_D - \bar{x} \langle \mathbf{1}, \mathbf{1} \rangle_D, \text{ d'après la bilinéarité du produit scalaire} \\ &= \bar{x} - \bar{x} \left(\sum_{i=1}^n p_i \right) \\ &= \bar{x} - \bar{x} = 0, \text{ car les poids relatifs satisfont } \sum_{i=1}^n p_i = 1. \end{aligned}$$

3.2.3 PARAMÈTRES DE DISPERSION

La moyenne ne donne qu'une information partielle sur les données. Par exemple les données (10, 10, 10) et (8, 9, 11, 12) ont même moyenne arithmétique, mais sont de nature différente.

Il est important de pouvoir mesurer la dispersion de ces données autour de leur moyenne. Il existe différents indicateurs de dispersion : l'*étendue*, l'*intervalle inter-quartile*, la *variance* et l'*écart-type*. Nous considérons ici la variance, notée $v_{\mathbf{x}}$, et l'écart-type $s_{\mathbf{x}}$ qui est la racine carrée de la variance.

Si chacun des individus a même poids relatif $\frac{1}{n}$, la *variance* est définie par :

$$v_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2$$

Si les individus ont des poids relatifs $(p_i)_{i=1}^n$, $\sum_{i=1}^n p_i = 1$, la *variance* est :

$$v_{\mathbf{x}} = \sum_{i=1}^n p_i (x_i - \bar{x})^2 = \sum_{i=1}^n p_i y_i^2.$$

Notations avec le produit scalaire sur \mathbb{R}^n :

$\begin{aligned} v_{\mathbf{x}} &= (\mathbf{x} - \bar{x}\mathbf{1})' D (\mathbf{x} - \bar{x}\mathbf{1}) = \langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle_D = \ \mathbf{x} - \bar{x}\mathbf{1}\ _D^2 \\ v_{\mathbf{x}} &= \mathbf{y}' D \mathbf{y} = \langle \mathbf{y}, \mathbf{y} \rangle_D = \ \mathbf{y}\ _D^2. \end{aligned}$

Propriété 3.2 (Théorème de König-Huygens). *La variance est égale à la (moyenne des carrés) moins (le carré de la moyenne) :*

$$v_{\mathbf{x}} = \sum_{i=1}^n p_i (x_i)^2 - (\bar{x})^2 = \|\mathbf{x}\|_D^2 - (\bar{x})^2.$$

Exercice 3.2. Démontrer la propriété 3.2 en utilisant les notations faisant intervenir le produit scalaire.

Solution de l'exercice 3.2. Par définition,

$$\begin{aligned} v_{\mathbf{x}} &= \langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle_D \\ &= \langle \mathbf{x}, \mathbf{x} \rangle_D - 2\bar{x} \langle \mathbf{x}, \mathbf{1} \rangle_D + \bar{x}^2 \langle \mathbf{1}, \mathbf{1} \rangle_D \text{ par bilinéarité et symétrie du produit scalaire} \\ &= \|\mathbf{x}\|_D^2 - 2\bar{x}^2 + \bar{x}^2 \langle \mathbf{1}, \mathbf{1} \rangle_D, \text{ d'après la définition de la moyenne} \\ &= \|\mathbf{x}\|_D^2 - \bar{x}^2, \text{ car } \sum_{i=1}^n p_i = 1. \end{aligned}$$

L'écart-type, noté $s_{\mathbf{x}}$, est la racine carrée de la variance, soit $s_{\mathbf{x}} = \sqrt{v_{\mathbf{x}}}$. Il a l'avantage d'être dans la même unité que la variable, alors que la variance est dans l'unité au carré.

Les *données centrées-réduites* sont les données centrées divisées par l'écart-type. Elles sont enregistrées dans le vecteur \mathbf{z} :

$$\mathbf{z} = \left(\frac{x_1 - \bar{x}}{s_{\mathbf{x}}}, \dots, \frac{x_n - \bar{x}}{s_{\mathbf{x}}} \right)' = \frac{1}{s_{\mathbf{x}}} \mathbf{y} = \frac{1}{\|\mathbf{y}\|_D} \mathbf{y}. \quad (3.1)$$

Propriété 3.3. *Les données centrées réduites ont moyenne nulle et variance égale à 1.*

Exercice 3.3. Démontrer la propriété 3.3 en utilisant les notations faisant intervenir le produit scalaire.

Solution de l'exercice 3.3. Par définition,

$$\begin{aligned} \bar{z} &= \langle \mathbf{z}, \mathbf{1} \rangle_D \\ &= \frac{1}{s_{\mathbf{x}}} \langle \mathbf{y}, \mathbf{1} \rangle_D, \text{ d'après la bilinéarité du produit scalaire} \\ &= 0, \text{ d'après la propriété 3.1.} \end{aligned}$$

De plus, la variance de \mathbf{z} est égale à :

$$\begin{aligned} v_{\mathbf{z}} &= \|\mathbf{z} - \bar{z}\mathbf{1}\|_D^2 = \|\mathbf{z}\|_D^2, \text{ d'après le point précédent.} \\ &= \frac{1}{\|\mathbf{y}\|_D^2} \|\mathbf{y}\|_D^2 = 1, \text{ d'après une propriété de la norme et (3.1).} \end{aligned}$$

3.3 LE CAS DE DEUX VARIABLES

On étudie maintenant deux caractères/variables sur l'échantillon I formé de n individus. On dispose de deux vecteurs $\mathbf{x}^1 = (x_{11}, \dots, x_{n1})'$ et $\mathbf{x}^2 = (x_{12}, \dots, x_{n2})'$ représentant la donnée de chacune des deux variables pour tous les individus.

3.3.1 REPRÉSENTATION DES DONNÉES

Les données des deux variables peuvent être représentées par un nuage de points, noté $N(I)$, formé de n points dans \mathbb{R}^2 :

$$N(I) = \{(x_{i1}, x_{i2}) \in \mathbb{R}^2 \mid i \in \{1, \dots, n\}\}.$$

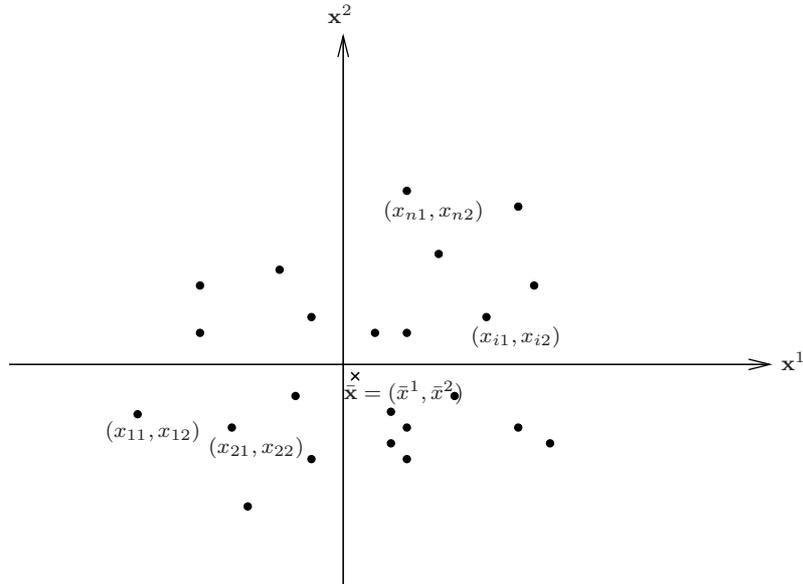


FIGURE 3.2 – Représentation du nuage de points $N(I)$ et du barycentre $\bar{\mathbf{x}}$.

3.3.2 PARAMÈTRES DE POSITION ET DE DISPERSION

De manière analogue au cas d'une variable, on peut définir les moyennes arithmétiques \bar{x}^1, \bar{x}^2 pour chacune des variables.

Le *barycentre* ou *centre de gravité* du nuage de point, noté $\bar{\mathbf{x}}$, est le point de \mathbb{R}^2 dont les coordonnées sont \bar{x}^1, \bar{x}^2 , voir Figure 3.2 :

$$\bar{\mathbf{x}} = (\bar{x}^1, \bar{x}^2)'$$

On a également les variances pour chacune des variables, que l'on notera simplement v_1, v_2 au lieu de $v_{\mathbf{x}_1}, v_{\mathbf{x}_2}$.

On définit aussi les écarts-types correspondants s_1, s_2 , les données centrées $\mathbf{y}^1, \mathbf{y}^2$, et les données centrées-réduites $\mathbf{z}^1, \mathbf{z}^2$. Pour $j \in \{1, 2\}$,

$$\begin{aligned} \mathbf{y}^j &= (x_{1j} - \bar{x}^j, \dots, x_{nj} - \bar{x}^j)' \\ \mathbf{z}^j &= \left(\frac{x_{1j} - \bar{x}^j}{s_j}, \dots, \frac{x_{nj} - \bar{x}^j}{s_j} \right)' = \frac{1}{s_j} \mathbf{y}^j = \frac{1}{\|\mathbf{y}^j\|_D} \mathbf{y}^j. \end{aligned}$$

3.3.3 PARAMÈTRES DE RELATION

La forme du nuage de points dans \mathbb{R}^2 peut suggérer qu'il existe une liaison entre les deux variables $\mathbf{x}^1, \mathbf{x}^2$. Le *coefficient de corrélation (linéaire)*, aussi connu sous le nom de *coefficient de Bravais-Pearson*, noté r_{12} , permet de mesurer le caractère linéaire du nuage de points. Attention : s'il y a absence de corrélation, cela ne signifie pas que les variables sont indépendantes, cela signifie seulement qu'elles ne sont pas liées linéairement. Définissons dans un premier temps la covariance des variables \mathbf{x}^1 et \mathbf{x}^2 , notée v_{12} .

Si chacun des individus a même poids relatif $\frac{1}{n}$, la *covariance de \mathbf{x}^1 et \mathbf{x}^2* est définie par :

$$v_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \bar{x}^1)(x_{i2} - \bar{x}^2) = \frac{1}{n} \sum_{i=1}^n y_{i1}y_{i2}.$$

Si les individus ont des poids relatifs $(p_i)_{i=1}^n$, $\sum_{i=1}^n p_i = 1$, la *covariance de \mathbf{x}^1 et \mathbf{x}^2* est :

$$v_{12} = \sum_{i=1}^n p_i (x_{i1} - \bar{x}^1)(x_{i2} - \bar{x}^2) = \sum_{i=1}^n p_i y_{i1}y_{i2}.$$

Notations avec le produit scalaire.

$$\begin{aligned} v_{12} &= (\mathbf{x}^1 - \bar{x}^1 \mathbf{1})' D (\mathbf{x}^2 - \bar{x}^2 \mathbf{1}) = \langle \mathbf{x}^1 - \bar{x}^1 \mathbf{1}, \mathbf{x}^2 - \bar{x}^2 \mathbf{1} \rangle_D \\ v_{12} &= (\mathbf{y}^1)' D \mathbf{y}^2 = \langle \mathbf{y}^1, \mathbf{y}^2 \rangle_D. \end{aligned}$$

Propriété 3.4. *La covariance satisfait aux propriétés suivantes :*

1. Elle est symétrique : $v_{12} = v_{21}$.
2. (Théorème de König-Huygens) :

$$v_{12} = \sum_{i=1}^n p_i x_{i1} x_{i2} - (\bar{x}^1)(\bar{x}^2).$$

Si chacun des individus a même poids relatif $\frac{1}{n}$, le *coefficient de corrélation de Bravais-Pearson* est défini par :

$$r_{12} = \frac{v_{12}}{s_1 s_2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i1} - \bar{x}^1)(x_{i2} - \bar{x}^2)}{s_1 s_2} = \frac{1}{n} \sum_{i=1}^n z_{i1} z_{i2}.$$

Si les individus ont des poids relatifs $(p_i)_{i=1}^n$, $\sum_{i=1}^n p_i = 1$, le *coefficient de corrélation de Bravais-Pearson* est :

$$r_{12} = \frac{v_{12}}{s_1 s_2} = \frac{\sum_{i=1}^n p_i (x_{i1} - \bar{x}^1)(x_{i2} - \bar{x}^2)}{s_1 s_2} = \sum_{i=1}^n p_i z_{i1} z_{i2}.$$

Notations avec le produit scalaire.

$$r_{12} = \frac{\langle \mathbf{x}^1 - \bar{x}^1 \mathbf{1}, \mathbf{x}^2 - \bar{x}^2 \mathbf{1} \rangle_D}{\|\mathbf{x}^1 - \bar{x}^1 \mathbf{1}\|_D \cdot \|\mathbf{x}^2 - \bar{x}^2 \mathbf{1}\|_D} = \langle \mathbf{z}^1, \mathbf{z}^2 \rangle_D$$

Propriété 3.5. *Le coefficient de corrélation satisfait aux propriétés suivantes :*

1. Il est symétrique : $r_{12} = r_{21}$.
2. r_{12} représente le cosinus de l'angle formé par les vecteurs $\mathbf{y}^1 = \mathbf{x}^1 - \bar{x}^1 \mathbf{1}$ et $\mathbf{y}^2 = \mathbf{x}^2 - \bar{x}^2 \mathbf{1}$.
3. $|r_{12}| = 1$ si et seulement si il existe a, b, c réels tels que, pour tout $i \in \{1, \dots, n\}$

$$ax_{i1} + bx_{i2} + c = 0.$$

Autrement dit, les n points $(x_{i1}, x_{i2})_{i=1}^n$ de \mathbb{R}^2 se trouvent sur une droite, i.e., il existe une dépendance linéaire entre les variables \mathbf{x}^1 et \mathbf{x}^2 , voir Figure 3.3.

Remarque 3.2. Attention, si le coefficient de corrélation n'est pas proche de 1 en valeur absolue, cela signifie seulement qu'il n'y a pas de dépendance linéaire entre les données, mais il peut y avoir une dépendance d'une autre nature : quadratique, etc.

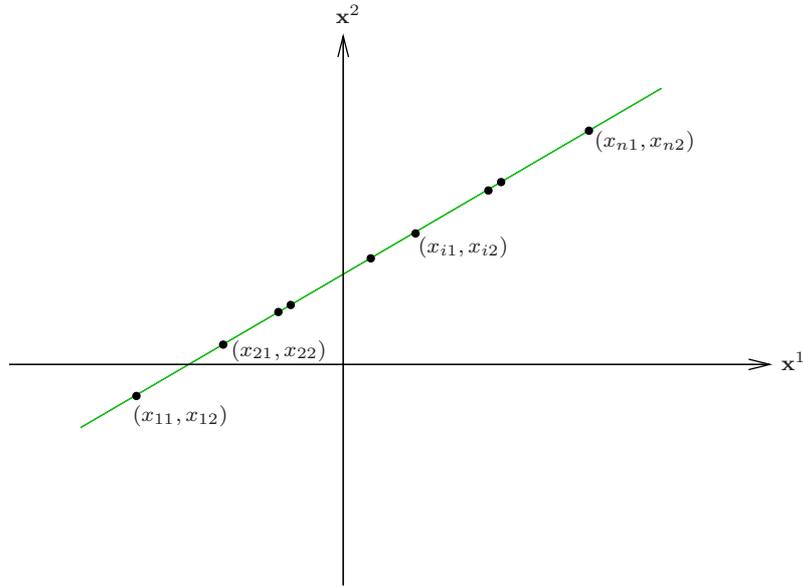


FIGURE 3.3 – Variables ayant un coefficient de corrélation égale à 1 en valeur absolue.

3.4 LE CAS DE p VARIABLES

Sur l'échantillon I formé de n individus, on étudie maintenant p caractères/variables numériques, indexées par un ensemble $J : \text{card}(J) = p$. Les données sont représentées dans une matrice $X = (x_{ij})$ de taille $n \times p$.

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}.$$

- Pour $i \in \{1, \dots, n\}$, la $i^{\text{ème}}$ ligne, notée \mathbf{x}'_i , représente les données de toutes les variables pour le $i^{\text{ème}}$ individu.
- Pour $j \in \{1, \dots, p\}$, la $j^{\text{ème}}$ colonne, notée \mathbf{x}^j , représente les données de la $j^{\text{ème}}$ variable pour tous les individus.
- Le coefficient x_{ij} est la donnée de la $j^{\text{ème}}$ variable pour le $i^{\text{ème}}$ individu.

On note $N(I)$ le nuage des n points-individus défini dans l'espace des variables \mathbb{R}^p et $N(J)$ le nuage des p points-variables défini dans l'espace des individus \mathbb{R}^n :

$$N(I) = \{\mathbf{x}_i \in \mathbb{R}^p \mid i \in \{1, \dots, n\}\}, \quad N(J) = \{\mathbf{x}^j \in \mathbb{R}^n \mid j \in \{1, \dots, p\}\}.$$

Un exemple de représentation graphique est donné à la figure 2.1.

3.4.1 MATRICE DES DONNÉES CENTRÉES

On définit les moyennes arithmétiques $\bar{x}^1, \dots, \bar{x}^p$ pour chacune des variables :

$$\forall j \in \{1, \dots, p\}, \quad \bar{x}^j = \sum_{i=1}^n p_i x_{ij}.$$

Le barycentre du nuage de point, noté $\bar{\mathbf{x}}$, est le point de \mathbb{R}^p dont les coordonnées sont $\bar{x}^1, \dots, \bar{x}^p$:

$$\bar{\mathbf{x}} = (\bar{x}^1, \dots, \bar{x}^p)'$$

En notation matricielle :

$$\boxed{\bar{\mathbf{x}} = X'D\mathbf{1}.}$$

La matrice des données *centrées* est de taille $n \times p$, elle est notée $Y = (y_{ij})$ et est définie par $y_{ij} = x_{ij} - \bar{x}^j$. On a $Y = X - \mathbf{1}\bar{\mathbf{x}}'$:

$$Y = \begin{pmatrix} x_{11} - \bar{x}^1 & \cdots & x_{1p} - \bar{x}^p \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}^1 & \cdots & x_{np} - \bar{x}^p \end{pmatrix}.$$

3.4.2 MATRICE DES VARIANCES-COVARIANCES

Soit $V(X) = (v_{jj'})$ la matrice de taille $p \times p$,

$$V(X) = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ \cdot & v_{22} & \cdots & v_{2p} \\ & & \ddots & \vdots \\ \cdot & & \cdot & v_{pp} \end{pmatrix},$$

dont les coefficients sont, pour tout $j, j' \in \{1, \dots, p\}$:

$$v_{jj'} = \sum_{i=1}^n p_i (x_{ij} - \bar{x}^j)(x_{ij'} - \bar{x}^{j'}) = \sum_{i=1}^n p_i y_{ij} y_{ij'}. \quad (3.2)$$

- Pour $j \neq j'$, $v_{jj'}$ est la covariance des variables \mathbf{x}^j et $\mathbf{x}^{j'}$,
- Pour $j = j'$, v_{jj} est la variance de la variable \mathbf{x}^j : $v_{jj} = v_j$.

La matrice $V(X)$ est appelée *matrice des variances-covariances*, ou simplement, *matrice des variances*.

Propriété 3.6. *Les coefficients de la matrice des variances satisfont aux propriétés suivantes. Pour tout $j, j' \in \{1, \dots, p\}$,*

1. $v_{jj'} = v_{j'j}$.
2. (Théorème de König-Huygens),

$$v_{jj'} = \sum_{i=1}^n p_i x_{ij} x_{ij'} - (\bar{x}^j)(\bar{x}^{j'}).$$

La matrice des variances-covariances satisfait aux propriétés suivantes :

1. elle est symétrique ;
2. on a les identités,

$$V(X) = Y'DY. \text{ En particulier, si } p_i \equiv \frac{1}{n}, V(X) = \frac{1}{n}Y'Y.$$

$$V(X) = \sum_{i=1}^n p_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \sum_{i=1}^n p_i \mathbf{y}_i \mathbf{y}_i'.$$

3.4.3 MATRICE DES DONNÉES CENTRÉES RÉDUITES

La matrice des données *centrées réduites* est de taille $n \times p$, elle est notée $Z = (z_{ij})$ et ses coefficients sont définis par, $z_{ij} = \frac{x_{ij} - \bar{x}^j}{s_j}$:

$$Z = \begin{pmatrix} \frac{x_{11} - \bar{x}^1}{s_1} & \dots & \frac{x_{1p} - \bar{x}^p}{s_p} \\ \vdots & \vdots & \vdots \\ \frac{x_{n1} - \bar{x}^1}{s_1} & \dots & \frac{x_{np} - \bar{x}^p}{s_p} \end{pmatrix}.$$

Introduisons la matrice diagonale $S^{\frac{1}{2}}$, de taille $p \times p$, dont les entrées sont les inverses des écarts-types :

$$S^{\frac{1}{2}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \frac{1}{s_p} \end{pmatrix}.$$

Alors, $S^{\frac{1}{2}}S^{\frac{1}{2}} = S$, où S est la matrice diagonale dont les coefficients sont les inverse des variances. Les données centrées et les données centrées réduites sont liées par la relation :

$$Z = YS^{\frac{1}{2}}.$$

3.4.4 MATRICE DES CORRÉLATIONS

Soit $R(X)$ la matrice de taille $p \times p$,

$$R(X) = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ \cdot & r_{22} & \dots & r_{2p} \\ & & \ddots & \vdots \\ \cdot & & \cdot & r_{pp} \end{pmatrix},$$

dont les coefficients sont, pour tout $j, j' \in \{1, \dots, p\}$:

$$r_{jj'} = \frac{v_{jj'}}{s_j s_{j'}} = \frac{\sum_{i=1}^n p_i (x_{ij} - \bar{x}^j)(x_{ij'} - \bar{x}^{j'})}{s_j s_{j'}} = \sum_{i=1}^n p_i z_{ij} z_{ij'}.$$

- Pour $j \neq j'$, $r_{jj'}$ est le coefficient de corrélation entre \mathbf{x}^j et $\mathbf{x}^{j'}$,
- Pour $j = j'$, $r_{jj} = 1$, en effet, $r_{jj} = \frac{v_{jj}}{s_j^2} = 1$.

Ainsi, les données centrées réduites sont comparables du point de vue de leur dispersion. La matrice $R(X)$ est appelée *matrice des corrélations*.

Propriété 3.7. *Les coefficients de la matrice des corrélations satisfont aux propriétés suivantes. Pour tout $j, j' \in \{1, \dots, p\}$,*

1. $r_{jj'} = r_{j'j}$.
2. $r_{jj'}$ représente le cosinus de l'angle formé par les vecteurs $\mathbf{y}^j = \mathbf{x}^j - \bar{x}^j$ et $\mathbf{y}^{j'} = \mathbf{x}^{j'} - \bar{x}^{j'}$.
3. $|r_{jj'}| = 1$ si et seulement si il existe une dépendance linéaire entre les variables \mathbf{x}^j et $\mathbf{x}^{j'}$.

La matrice des corrélations satisfait aux propriétés suivantes :

1. elle est symétrique, et possède des 1 sur sa diagonale,
2. on a les identités :

$$R(X) = Z'DZ = S^{\frac{1}{2}}V(X)S^{\frac{1}{2}}. \text{ En particulier, si } p_i \equiv \frac{1}{n}, R(X) = \frac{1}{n}Z'Z.$$

$$R(X) = \sum_{i=1}^n p_i \mathbf{z}_i \mathbf{z}_i'$$

3.4.5 INERTIE DU NUAGE DE POINTS

Considérons une distance d_M entre individus sur \mathbb{R}^p . Cette section a pour but de définir l'inertie, qui mesure la dispersion du nuage $N(I)$ des individus autour d'un point donné de \mathbb{R}^p . Cette notion joue un rôle important dans les chapitres 4, 5 et 7.

Soit \mathbf{a} un point de \mathbb{R}^p . L'*inertie du nuage de points* $N(I)$ par rapport au point \mathbf{a} , notée $J_{\mathbf{a}}^M(X)$, est la moyenne pondérée des carrés des distances au point \mathbf{a} .

$$J_{\mathbf{a}}^M(X) = \sum_{i=1}^n p_i d_M^2(\mathbf{x}_i, \mathbf{a}) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \mathbf{a}\|_M^2.$$

Nous omettrons le M et le X de la notation lorsqu'il n'y a pas de confusion possible.

L'*inertie totale du nuage de points* $N(I)$, notée $J_{\text{tot}}^M(X)$, est l'inertie par rapport au centre de gravité $\bar{\mathbf{x}}$:

$$J_{\text{tot}}^M(X) = \sum_{i=1}^n p_i d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|_M^2.$$

Le théorème de Huygens ci-dessous dit que l'inertie du nuage de points est maximale si elle est mesurée par rapport au centre de gravité du nuage $N(I)$.

Théorème 3.1 (Théorème de Huygens). *Pour tout point \mathbf{a} de \mathbb{R}^p , l'inertie du nuage de points par rapport au point \mathbf{a} satisfait à l'égalité suivante,*

$$J_{\mathbf{a}}^M(X) = J_{\text{tot}}^M(X) + d_M^2(\bar{\mathbf{x}}, \mathbf{a}).$$

Démonstration. Par définition,

$$\begin{aligned}
 \mathcal{J}_{\mathbf{a}}^M(X) &= \sum_{i=1}^n p_i d_M^2(\mathbf{x}_i, \mathbf{a}) \\
 &= \sum_{i=1}^n p_i \langle \mathbf{x}_i - \mathbf{a}, \mathbf{x}_i - \mathbf{a} \rangle_M = \sum_{i=1}^n p_i \langle \mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{a}, \mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{a} \rangle_M \\
 &= \sum_{i=1}^n p_i [\|\mathbf{x}_i - \bar{\mathbf{x}}\|_M^2 + \|\bar{\mathbf{x}} - \mathbf{a}\|_M^2 - 2 \langle \bar{\mathbf{x}} - \mathbf{a}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle_M] \\
 &= \mathcal{J}_{\text{tot}}^M(X) + 2 \sum_{i=1}^n p_i \langle \bar{\mathbf{x}} - \mathbf{a}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle_M + \|\bar{\mathbf{x}} - \mathbf{a}\|_M^2, \text{ car } \sum_{i=1}^n p_i = 1.
 \end{aligned}$$

Il reste à montrer que $\sum_{i=1}^n p_i \langle \bar{\mathbf{x}} - \mathbf{a}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle_M = 0$. D'après la bilinéarité du produit scalaire,

$$\begin{aligned}
 \sum_{i=1}^n p_i \langle \bar{\mathbf{x}} - \mathbf{a}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle_M &= \langle \bar{\mathbf{x}} - \mathbf{a}, \sum_{i=1}^n p_i (\mathbf{x}_i - \bar{\mathbf{x}}) \rangle_M \\
 &= \langle \bar{\mathbf{x}} - \mathbf{a}, \sum_{i=1}^n p_i \mathbf{x}_i - \bar{\mathbf{x}} \rangle_M, \text{ car } \sum_{i=1}^n p_i = 1 \\
 &= \langle \bar{\mathbf{x}} - \mathbf{a}, \bar{\mathbf{x}} - \bar{\mathbf{x}} \rangle_M, \text{ par définition de } \bar{\mathbf{x}} \\
 &= 0. \quad \square
 \end{aligned}$$

Le théorème suivant relie l'inertie totale du nuage de points $N(I)$ et la somme des distances entre tous les individus. Il sera utilisé dans le chapitre 4.

Théorème 3.2. *Deux fois l'inertie totale du nuage de points est égale à la moyenne des carrés de toutes les distances entre les n individus :*

$$2\mathcal{J}_{\text{tot}}^M(X) = \sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} d_M^2(\mathbf{x}_i, \mathbf{x}_{i'}).$$

CAS PARTICULIER DE LA MÉTRIQUE USUELLE

Supposons que la métrique choisie est la métrique usuelle d de \mathbb{R}^p . Omettant l'indice M , l'inertie totale du nuage de points est :

$$\mathcal{J}_{\text{tot}}(X) = \sum_{i=1}^n p_i d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2.$$

La propriété suivante montre que pour ce choix de métrique, l'inertie totale est égale à la somme des variances des variables. Ceci donne une autre interprétation de l'inertie, comme mesurant la variabilité totale des données.

Propriété 3.8. *L'inertie totale du nuage de points est égale à la somme des variances :*

$$\mathcal{J}_{\text{tot}}(X) = \sum_{j=1}^p v_j = \text{Tr}(V(X))$$

Démonstration. Par définition,

$$\begin{aligned} J_{\text{tot}}(X) &= \sum_{i=1}^n p_i d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \sum_{j=1}^p (x_{ij} - \bar{x}^j)^2 \\ &= \sum_{j=1}^p \left(\sum_{i=1}^n p_i (x_{ij} - \bar{x}^j)^2 \right), \text{ en échangeant les sommes} \\ &= \sum_{j=1}^p v_j, \text{ par définition de la variance.} \end{aligned}$$

La deuxième égalité découle de la définition de la trace. □

3.5 INDIVIDUS REGROUPÉS EN CLASSES

Lors du chapitre 7 sur l'analyse discriminante et du chapitre 6 sur les méthodes de classification, nous allons considérer le cas où les individus sont regroupés en classes. Nous aurons besoin de la notion de variance intra- et inter-classes et d'inertie totale, intra- et inter-classe. Ces notions préliminaires s'insèrent naturellement dans ce chapitre, raison pour laquelle nous les abordons ici.

3.5.1 NOTATIONS

Considérons la matrice des données X comme précédemment. Supposons de plus que les individus sont regroupés en k classes (ou groupes) disjoint(e)s : C_1, \dots, C_k .

On a pour chaque classe, c'est-à-dire pour chaque $\ell \in \{1, \dots, k\}$:

- I_ℓ est l'ensemble des individus de la classe C_ℓ , avec $\text{card}(I_\ell) = n_\ell$. Ainsi, $\sum_{\ell=1}^k n_\ell = n$.
- Le poids relatif $p_{(\ell)}$ de la classe C_ℓ :

$$p_{(\ell)} = \sum_{i \in I_\ell} p_i. \quad \text{Ainsi,} \quad \sum_{\ell=1}^k p_{(\ell)} = 1.$$

- Le *centre de gravité* ou *barycentre de la classe* C_ℓ , noté $\bar{\mathbf{x}}_{(\ell)}$:

$$\bar{\mathbf{x}}_{(\ell)} = (\bar{x}_{(\ell)}^1, \dots, \bar{x}_{(\ell)}^p)',$$

où pour tout $j \in \{1, \dots, p\}$,

$$\bar{x}_{(\ell)}^j = \frac{1}{p_{(\ell)}} \sum_{i \in I_\ell} p_i x_{ij}.$$

3.5.2 MATRICE DES VARIANCES INTER ET INTRA-CLASSES

On se souvient de la définition de la matrice $V(X)$ des variances de la section 3.4.2. Comme il n'y a pas de confusion possible dans cette section, on a la notera simplement $V = (v_{ij})$.

Maintenant que les individus sont divisés en classes, il est naturel d'introduire la notion analogue pour les individus à l'intérieur d'une même classe avec leurs centres de gravités respectifs, et pour les centres de gravité des différentes classes.

Soit $W = (w_{ij})$ la matrice de taille $p \times p$,

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ \cdot & w_{22} & \cdots & w_{2p} \\ & & \ddots & \vdots \\ \cdot & & \cdot & w_{pp} \end{pmatrix},$$

dont les coefficients sont, pour tout $j, j' \in \{1, \dots, p\}$:

$$w_{jj'} = \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (x_{ij} - \bar{x}_{(\ell)}^j)(x_{ij'} - \bar{x}_{(\ell)}^{j'}).$$

La matrice W s'appelle *matrice des variances intra-classes*, où le W est pour "within".

Remarque. Pour tout $j, j' \in \{1, \dots, p\}$, le coefficient $w_{jj'}$ s'écrit :

$$w_{jj'} = \sum_{\ell=1}^k p^{(\ell)} \sum_{i \in I_\ell} \frac{p_i}{p^{(\ell)}} (x_{ij} - \bar{x}_{(\ell)}^j)(x_{ij'} - \bar{x}_{(\ell)}^{j'}).$$

En particulier, si $p_i \equiv \frac{1}{n}$,

$$w_{jj'} = \sum_{\ell=1}^k \frac{n_\ell}{n} \left[\frac{1}{n_\ell} \sum_{i \in I_\ell} (x_{ij} - \bar{x}_{(\ell)}^j)(x_{ij'} - \bar{x}_{(\ell)}^{j'}) \right].$$

Le coefficient $w_{jj'}$ est la moyenne arithmétique des variances/covariances $v_{jj'}^{(\ell)}$ de chacune des classes.

Soit $B = (b_{ij})$ la matrice de taille $p \times p$,

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ \cdot & b_{22} & \cdots & b_{2p} \\ & & \ddots & \vdots \\ \cdot & & \cdot & b_{pp} \end{pmatrix},$$

dont les coefficients sont, pour tout $j, j' \in \{1, \dots, p\}$:

$$b_{jj'} = \sum_{\ell=1}^k p^{(\ell)} (\bar{x}_{(\ell)}^j - \bar{x}^j)(\bar{x}_{(\ell)}^{j'} - \bar{x}^{j'}). \quad (3.3)$$

La matrice B s'appelle *matrice des variances inter-classes*, où le B est pour "between".

Remarque 3.3. Le coefficient $b_{jj'}$ est la variance/covariance de la matrice de taille $k \times p$, dont les individus sont les centres de gravité de chacune des classes. À noter que le centre de gravité des centres de gravité des classes est aussi $\bar{\mathbf{x}}$.

Propriété 3.9. Les matrices W et B sont symétriques, et on a les identités :

$$W = \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)}) (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)})'$$

$$B = \sum_{\ell=1}^k p_{(\ell)} (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}})'$$

Théorème 3.3. Les matrices W , B et V sont reliées par l'égalité suivante,

$$V = W + B.$$

Démonstration. D'après la propriété 3.6

$$\begin{aligned} V &= \sum_{i=1}^n p_i (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})' \\ &= \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})', \text{ (réécriture de la somme)} \\ &= \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)} + \bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)} + \bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}})' \\ &= \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)}) (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)})' + \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}})' + \\ &\quad + \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)}) (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}})' + \sum_{\ell=1}^k \sum_{i \in I_\ell} p_i (\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)})'. \end{aligned}$$

Le premier terme est la matrice W , le deuxième terme est la matrice B car $\sum_{i \in I_\ell} p_i = p_{(\ell)}$. Les deux derniers termes sont nuls car :

$$\sum_{i \in I_\ell} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)}) = \sum_{i \in I_\ell} p_i \mathbf{x}_i - \left(\sum_{i \in I_\ell} p_i \right) \bar{\mathbf{x}}_{(\ell)} = p_{(\ell)} \bar{\mathbf{x}}_{(\ell)} - p_{(\ell)} \bar{\mathbf{x}}_{(\ell)} = \mathbf{0}.$$

Ceci achève de démontrer le résultat. □

3.5.3 INERTIE DU NUAGE, INERTIE INTRA ET INTER-CLASSES

Avec des individus séparés en classe, il est naturel d'introduire la notion d'inertie intra- et inter-classe. On se restreint au cas où la distance choisie est la distance usuelle d sur \mathbb{R}^p , mais les définitions se généralisent facilement à un autre choix de métrique. Étant donné qu'il n'y a pas de confusion possible, on omettra l'argument X dans les notations des inerties.

Pour tout $\ell \in \{1, \dots, k\}$, l'inertie de la classe C_ℓ , notée $J_{\text{tot}}^{(\ell)}$ est l'inertie totale des individus de la classe ℓ :

$$J_{\text{tot}}^{(\ell)} = \sum_{i \in I_\ell} \frac{p_i}{p_{(\ell)}} d^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(\ell)}) = \sum_{i \in I_\ell} \frac{p_i}{p_{(\ell)}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{(\ell)}\|^2.$$

L'inertie intra-classe du nuage de points $N(I)$, notée J_{intra} , est la moyenne arithmétique des inerties des classes :

$$J_{\text{intra}} = \sum_{\ell=1}^k p_{(\ell)} J_{\text{tot}}^{(\ell)} = \sum_{\ell=1}^k \sum_{i \in I_{\ell}} p_i d^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(\ell)}).$$

L'inter inter-classe du nuage de points $N(I)$, notée J_{inter} , est l'inertie totale de la matrice dont les individus sont les centres de gravité des classes :

$$J_{\text{inter}} = \sum_{\ell=1}^k p_{(\ell)} d^2(\bar{\mathbf{x}}_{(\ell)}, \bar{\mathbf{x}}) = \sum_{\ell=1}^k p_{(\ell)} \|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}\|^2.$$

La propriété ci-dessous est l'analogie pour l'inertie intra- et inter-classe de la propriété 3.8.

Propriété 3.10. *L'inertie intra-classe, respectivement inter-classe, est égale à la somme des variances intra-classes, respectivement inter-classe :*

$$J_{\text{intra}} = \sum_{j=1}^p w_{jj} = \text{Tr}(W)$$

$$J_{\text{inter}} = \sum_{j=1}^p b_{jj} = \text{Tr}(B).$$

En conséquence, on obtient :

Théorème 3.4 (Théorème de König-Huygens). *L'inertie totale est égale à l'inertie intra-classe plus l'inertie inter-classe :*

$$J_{\text{tot}} = J_{\text{intra}} + J_{\text{inter}}.$$

Démonstration. Ce théorème découle du théorème 3.3, des propriétés 3.8, 3.10 et de l'additivité de la trace. □

Exercice 3.4.

Cet exercice est à faire sous excel.

1. Télécharger les données de l'OCDE sur les principales zones protégées à l'adresse suivante : <http://www.oecd-ilibrary.org/fr/environment/panorama-de-l-environnement-2013/principales-zones-protégees-derniere-annee>
On suppose que chacun des individus a poids $1/n$, où n est le nombre de pays considérés.
2. En extraire le tableau de données suivant et affecter un groupe (1 ou 2) à chaque individu selon qu'il appartienne à l'Union Européenne ou non.

	Réserves naturelles intégrales 2010	Réserves de biosphère 2012	Zones humides 2012	Groupes UE/hors UE
	Superficie, milliers km ²	nombre de sites	Superficie km ²	1-2
Allemagne	1	15	8682	1
Australie	601	14	81117	2
Autriche	1	6	1200	1
Canada	513	16	130666	2
Chili	82	10	3590	2
Corée	3	4	177	2
Danemark	0	1	23031	1
Espagne	2	39	2966	1
Estonie	2	1	3048	1
Finlande	11	2	7995	1
France	3	10	33083	1
Grèce	1	2	1635	1
Hongrie	2	5	2450	1
Irlande	1	2	670	1
Israël	0	2	4	2
Italie	14	8	602	1
Japon	6	4	1370	2
Mexique	3	40	88264	2
Pays-Bas	1	1	8189	1
Pologne	2	10	1451	1
Portugal	1	7	866	1
République slovaque	4	4	407	1
République tchèque	2	6	547	1
Royaume-Uni	0	8	12757	1
Slovénie	1	3	82	1
Suède	41	4	5147	1
Suisse	0	2	1469	2
Turquie	4	1	1799	2

3. Pour chacun des variables quantitatives, calculer de deux manières la moyenne arithmétique, la variance, l'écart-type.
4. Donner la matrice des données centrées et des données centrées-réduites.
5. Calculer de deux manières la matrice des variances-covariances et la matrice des corrélations.
6. On considère aussi la variable qualitative qui sépare les pays en deux groupes.
 - (a) Calculer le poids relatif de chaque classe et le centre de gravité de chaque classe.
 - (b) Calculer la matrice W des variances intra-classes et la matrice B des variances inter-classes. Vérifier la formule : $V = W + B$.

Solution de l'exercice 3.4. Le détail des solutions est donné dans le corrigé excel. Voici quelques uns des résultats.

3. Le vecteur des moyennes arithmétiques (ou centre de gravité du nuage de points) est,

$$\bar{\mathbf{x}} = (\bar{x}^1, \bar{x}^2, \bar{x}^3)' = (47, 8, 15117)'$$

Le vecteur des variance est,

$$(v_1, v_2, v_3)' = (20472, 94, 968530611)'$$

Le vecteur des écarts-types est,

$$(s_1, s_2, s_3)' = (143, 10, 31121)'$$

5. La matrice des variances-covariances est,

$$V = \begin{pmatrix} 20472 & 269 & 3469254 \\ 269 & 94 & 149160 \\ 3469254 & 149160 & 968530611 \end{pmatrix}.$$

La matrice des corrélations est,

$$R = \begin{pmatrix} 1,00 & 0,19 & 0,78 \\ 0,19 & 1,00 & 0,49 \\ 0,78 & 0,49 & 1,00 \end{pmatrix}.$$

6. La matrice des variances intra-classes est,

$$W = \begin{pmatrix} 16801 & 135 & 2672103 \\ 135 & 89 & 119944 \\ 2672103 & 119944 & 795406142 \end{pmatrix}.$$

La matrice des variances inter-classes est,

$$B = \begin{pmatrix} 3670 & 135 & 797152 \\ 135 & 5 & 29216 \\ 797152 & 29216 & 173124469 \end{pmatrix}.$$

On retrouve $W + B = V$.

CHAPITRE 4

ANALYSE EN COMPOSANTES PRINCIPALES

On considère un échantillon I formé de n individus, sur lequel on étudie p variables numériques, indexées par un ensemble J . Comme lors des précédents chapitres, les données sont représentées dans une matrice $X = (x_{ij})$ de taille $n \times p$.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Rappelons que $N(I)$ est le nuage des n points-individus dans \mathbb{R}^p et $N(J)$ est le nuage des p points-variables dans \mathbb{R}^n . Les dimensions n et p étant grandes en général, la visualisation des données n'est pas possible à partir des nuages entiers.

Comme pour l'analyse factorielle générale présentée au chapitre 2, le but de l'Analyse en Composantes Principales (ACP) est de projeter le nuage des points individus (puis aussi le nuage des points variables) sur un espace de plus petite dimension (1, 2 ou 3), de sorte à pouvoir visualiser les données. Le critère retenu lors de la projection est que l'on perde le moins d'information possible, l'information étant encodée dans la dispersion des données.

Les fondateurs de l'ACP sont Pearson (1901), Hotelling (1933), Anderson (1958), Kchirsagar (1972), Horst (1965) et Harman (1966).

Rappelons que si les données ne sont pas récoltées dans les mêmes conditions, on attribue des poids relatifs $(p_i)_{i=1}^n$ aux individus, avec $\sum_{i=1}^n p_i = 1$. Lorsque, pour tout $i \in \{1, \dots, n\}$, $p_i = \frac{1}{n}$, on retrouve les poids relatifs homogènes. La matrice D est la matrice diagonale de ces poids :

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & p_n \end{pmatrix}.$$

4.1 AJUSTEMENT DU NUAGE $N(I)$ DES INDIVIDUS

Lors de la projection du nuage des individus, nous souhaitons garder la dispersion des données. La section suivante précise ces notions.

4.1.1 DISTANCE ET INERTIE

Pour mesurer la dispersion, nous avons besoin d'une notion de distance entre individus, ce qui revient à choisir une métrique M . Rappelons qu'alors la distance est définie, pour tout $i, i' \in \{1, \dots, n\}$, par :

$$d_M^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_M^2 = (\mathbf{x}_i - \mathbf{x}_{i'})' M (\mathbf{x}_i - \mathbf{x}_{i'}).$$

- **Données réduites.** Lorsque les variables s'expriment dans des unités non comparables (par exemple : âge, salaire, taille de l'habitat, etc.), il vaut mieux considérer les données réduites, car alors, elles n'ont plus d'unité et ont même variance égale à 1. Ceci revient à prendre pour M la matrice diagonale S , voir section 3.4.3, dont les coefficients diagonaux sont les inverses des variances. On a alors,

$$d_S^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^p \left(\frac{x_{ij} - x_{i'j}}{s_j} \right)^2.$$

- **Données non-réduites.** Si les unités des variables sont comparables, il suffit de prendre $M = I$. Alors,

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2.$$

La dispersion des données est encodée dans la somme des distances entre toutes les paires d'individus :

$$\sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} d_M^2(\mathbf{x}_i, \mathbf{x}_{i'}).$$

D'après le théorème 3.2 du chapitre 3, à un facteur 2 près, cette quantité est égale à l'inertie totale du nuage de points :

$$\sum_{i=1}^n \sum_{i'=1}^n p_i p_{i'} d_M^2(\mathbf{x}_i, \mathbf{x}_{i'}) = 2J_{\text{tot}}^M(X),$$

où rappelons que l'inertie totale du nuage de points est la somme pondérée des distances des individus au centre de gravité :

$$J_{\text{tot}}^M(X) = \sum_{i=1}^n p_i d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|_M^2.$$

Ainsi, lors de la projection des données, nous souhaitons que l'inertie totale des données projetées reste maximale. Regardons ce que vaut l'inertie avec les deux choix possibles de métrique.

- **Données réduites.** Avec le choix de la métrique $M = S$, nous avons :

$$\begin{aligned} J_{\text{tot}}^S(X) &= \sum_{i=1}^n p_i d_S^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \sum_{j=1}^p \left(\frac{x_{ij} - \bar{x}^j}{s_j} \right)^2 \\ &= \sum_{i=1}^n p_i d^2(\mathbf{0}, \mathbf{z}_i) = J_{\text{tot}}(Z), \end{aligned}$$

où rappelons que \mathbf{z}_i est la $i^{\text{ème}}$ ligne de la matrice Z des données centrées-réduites, représentant le $i^{\text{ème}}$ individu centré-réduit. En conséquence, travailler avec la métrique S revient à *considérer la métrique usuelle sur les données centrées réduites* Z . On parle alors d'*ACP normée*. C'est le cas que nous traitons dans ce cours.

- **Données non-réduites.** Avec le choix de la métrique usuelle, $M = I$, nous avons :

$$\begin{aligned} J_{\text{tot}}(X) &= \sum_{i=1}^n p_i d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \sum_{j=1}^p (x_{ij} - \bar{x}^j)^2 \\ &= \sum_{i=1}^n p_i d^2(\mathbf{0}, \mathbf{y}_i) = J_{\text{tot}}(Y), \end{aligned}$$

où rappelons que \mathbf{y}_i est la $i^{\text{ème}}$ ligne de la matrice Y des données centrées, représentant le $i^{\text{ème}}$ individu centré. En conséquence, travailler avec la métrique usuelle revient à *considérer la métrique usuelle sur les données centrées* Y . On parle alors d'*ACP non-normée*. L'analyse est analogue que pour des données centrées réduites, mais nous ne la détaillons pas dans ce cours.

4.1.2 AJUSTEMENT DU NUAGE DES INDIVIDUS $N(I)$

Soit une droite \mathcal{D}_1 , passant par l'origine de vecteur directeur unitaire \mathbf{u}_1 . Pour tout individu centré-réduit \mathbf{z}_i , notons $\pi_1(\mathbf{z}_i)$ la projection orthogonale de \mathbf{z}_i sur \mathcal{D}_1 . D'après la section précédente, nous souhaitons trouver le vecteur unitaire \mathbf{u}_1 , tel que l'inertie totale des données projetées,

$$\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_1(\mathbf{z}_i)).$$

soit maximale. De manière analogue au chapitre 2, nous itérons ensuite cette étape. Pour tout $k \in \{1, \dots, p\}$, étant donné un vecteur unitaire \mathbf{u}_k , notons π_k la projection orthogonale sur la droite \mathcal{D}_k passant par l'origine et de vecteur directeur \mathbf{u}_k . La question de l'ajustement du nuage des individus de sorte à ce que la dispersion des données soit maximale s'écrit alors :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ tel que $\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_1(\mathbf{z}_i))$ soit maximale, puis $\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_2(\mathbf{z}_i))$, etc.

La réponse est donnée par le théorème 2.4 du chapitre 2, écrit avec la matrice des données centrées-réduites Z .

Théorème 4.1. *La solution au problème d'optimisation sous contrainte de l'ACP est le repère orthonormé $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ formé des vecteurs propres de $Z'DZ$, tel que \mathbf{u}_1 correspond à la plus*

grande valeur propre λ_1 de $Z'DZ$, puis \mathbf{u}_2 à la deuxième valeur propre λ_2 de $Z'DZ$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,

$$\forall k \in \{1, \dots, p\}, \quad \sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_k(\mathbf{z}_i)) = \mathbf{u}'_k Z'DZ \mathbf{u}_k = \lambda_k.$$

Remarque 4.1. D'après le chapitre 3, nous savons que la matrice $Z'DZ$ est la matrice des corrélations $R(X)$ des données X :

$$Z'DZ = R(X).$$

Ainsi, la matrice à diagonaliser est la matrice $R(X)$.

Définition 4.1. Les axes passant par l'origine et de vecteurs directeurs $\mathbf{u}_1, \dots, \mathbf{u}_p$ sont les *axes principaux* ou *axes factoriels*¹, respectivement le premier, ..., le p -ième. Chaque vecteur \mathbf{u}_k est combinaison linéaire des vecteurs $\mathbf{x}^1, \dots, \mathbf{x}^p$, ce qui signifie que \mathbf{u}_k représente une nouvelle variable "abstraite" combinaison des variables d'origine. Ce sont les variables abstraites qui permettent le mieux de visualiser la dispersion des données.

4.1.3 COMPOSANTES PRINCIPALES : PROJECTION DES INDIVIDUS SUR LES NOUVELLES VARIABLES

Pour $i \in \{1, \dots, n\}$, $k \in \{1, \dots, p\}$, la coordonnée ψ_{ik} de l'individu i sur le $k^{\text{ème}}$ axe factoriel est

$$\psi_{ik} = \langle \mathbf{z}_i, \mathbf{u}_k \rangle = \mathbf{z}'_i \mathbf{u}_k.$$

Définissons la matrice $U = (\mathbf{u}_1 \dots \mathbf{u}_p)$ dont les colonnes sont les vecteurs de la nouvelle base. Alors, on peut décrire les coordonnées des individus dans le nouveau repère dans la matrice Ψ :

$$\Psi = ZU.$$

La $i^{\text{ème}}$ ligne $\Psi'_i = \mathbf{z}'_i U$ représente les coordonnées du $i^{\text{ème}}$ individu sur les p axes factoriels $\mathbf{u}_1, \dots, \mathbf{u}_p$. La $k^{\text{ème}}$ colonne $\Psi^k = Z\mathbf{u}_k$ représente les coordonnées de tous les individus sur le $k^{\text{ème}}$ axe factoriel \mathbf{u}_k , autrement dit, Ψ^k représente la donnée de la nouvelle variable \mathbf{u}_k pour tous les individus.

Définition 4.2. Le vecteur $\Psi^k = Z\mathbf{u}_k$ est appelé la $k^{\text{ème}}$ *composante principale*.

Nous déduisons le résultat suivant.

Proposition 4.1. *La matrice des variances-covariances $V(\Psi)$ des composantes principales (des données pour les nouvelles variables) est :*

$$V(\Psi) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}.$$

1. Lorsque la métrique n'est pas la métrique usuelle, on distingue la notion d'axe et de facteur.

En particulier, la variance de la $k^{\text{ème}}$ composante principale est égale à λ_k :

$$v_{\Psi^k} = \lambda_k.$$

et les composantes principales sont non-corrélées linéairement :

$$\forall k \neq k', \quad v_{\Psi^k, \Psi^{k'}} = 0.$$

De plus, les composantes principales sont telles que la variance de la première est maximale, puis celle de la deuxième, etc.

Démonstration. Les composantes principales sont centrées car elles sont égales à ZU , et les données Z le sont. Ainsi, d'après la propriété 3.6 du chapitre 3, la matrice des variances-covariances de Ψ est égale à :

$$V(\Psi) = \Psi' D \Psi = U' Z' D Z U = U' R(X) U.$$

Étant donnée que la matrice U est formée d'une base de vecteurs propres orthonormaux de la matrice $R(X)$, on obtient la première partie du résultat. La maximalité découle du principe d'ajustement. \square

4.1.4 INERTIE REVISITÉE

Rappelons, voir chapitre 3, que l'inertie totale est aussi la trace de la matrice des variances, donnant lieu à l'interprétation de l'inertie comme mesurant la variabilité totale des données. Cette vision de l'inertie nous permet de déduire les résultats suivants.

Proposition 4.2.

1. L'inertie totale du nuage de points centrés-réduits Z est égale à :

$$\mathcal{J}_{\text{tot}}(Z) = p.$$

2. L'inertie totale des données $\Psi = ZU$ dans le nouveau repère est égale à la somme des inerties des données projetées sur tous les axes :

$$\mathcal{J}_{\text{tot}}(\Psi) = \sum_{k=1}^p \mathbf{u}'_k R(X) \mathbf{u}_k = \sum_{k=1}^p \lambda_k.$$

3. L'inertie totale est conservée :

$$\mathcal{J}_{\text{tot}}(Z) = \mathcal{J}_{\text{tot}}(\Psi).$$

Démonstration.

1. D'après la propriété 3.8, $\mathcal{J}_{\text{tot}}(Z) = \text{Tr}(V(Z))$. Les données Z étant en particulier centrées, on a d'après la propriété 3.4, $V(Z) = Z' D Z$, matrice qui est égale à $R(X)$. Comme la matrice des corrélations est de taille $p \times p$ et possède des 1 sur sa diagonale, on déduit le résultat.
2. D'après la preuve de la proposition 4.1, nous savons que $V(\Psi) = U' R(X) U$ qui est aussi la matrice diagonale formée des $\lambda_1, \dots, \lambda_p$. Les égalités découlent de ceci et de la définition de la trace.

3. Utilisant à nouveau la proposition 4.1, nous avons :

$$J_{\text{tot}}(\Psi) = \text{Tr}(V(\Psi)) = \text{Tr}(U'R(X)U) = \text{Tr}(UU'R(X)) = \text{Tr}(R(X)).$$

Nous avons utilisé le fait que la trace est invariante par commutation des matrices, et le fait que la matrice U est orthogonale, ce qui implique que $UU' = I$.

□

4.1.5 AIDE À L'INTERPRÉTATION DANS LE PLAN FACTORIEL

Définition 4.3. Pour tout $k \in \{1, \dots, p\}$, le *pourcentage d'inertie expliquée* par les k premiers axes est :

$$F_k = \frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p} = \frac{\lambda_1 + \dots + \lambda_k}{p}.$$

Rappelons que la valeur propre λ_k est égale à l'inertie des données projetées sur le $k^{\text{ème}}$ axe, d'où le nom de la définition ; λ_k représente aussi la variance de la $k^{\text{ème}}$ composante principale. La deuxième égalité utilise la proposition 4.2, qui montre que la somme des valeurs propres est égale à p . Les nombres $(F_k)_{k=1}^p$ servent à déterminer le nombre d'axes à garder.

Le logiciel SPAD donne, pour chacun des individus i , $i \in \{1, \dots, n\}$:

- **Coordonnées dans le nouveau repère** : $(\psi_{ik} : k \in \{1, \dots, p\})$.
- **Distance au carré de l'individu i à l'origine** : $\text{Disto}(i)$.

$$\text{Disto}(i) = d^2(\mathbf{0}, \mathbf{z}_i).$$

Il s'agit de la somme des coordonnées au carré de l'individu i dans le repère. Ceci ne dépend pas du repère choisi (celui d'origine ou celui des axes principaux si on les prend tous).

- **Contribution de l'individu i à l'inertie sur l'axe principal k** : $\text{CRT}_k(i)$.

$$\text{CRT}_k(i) = \frac{p_i d^2(\mathbf{0}, \pi_k(\mathbf{z}_i))}{\lambda_k} = \frac{p_i \psi_{ik}^2}{\lambda_k}.$$

On a $\sum_{i=1}^n \text{CRT}_k(i) = 1$ car $\sum_{i=1}^n p_i d^2(\mathbf{0}, \pi_k(\mathbf{z}_i)) = \lambda_k$.

La considération des contributions, quand elles sont excessives, aide à l'interprétation des axes. Normalement, il n'est pas souhaitable qu'un individu ait une contribution excessive car cela ferait un facteur d'instabilité, surtout si les individus constituent un échantillon et ne présentent pas d'intérêt en eux-même. Il faudrait alors refaire l'analyse sans eux.

Lorsque les poids des individus sont tous les mêmes, cette quantité ne donne pas plus d'information que les coordonnées.

- **Cosinus carré entre le vecteur $(\mathbf{0z}_i)$ et la droite \mathcal{D}_k** : $\cos_k^2(i)$

$$\cos_k^2(i) = \frac{d^2(\mathbf{0}, \pi_k(\mathbf{z}_i))}{d^2(\mathbf{0}, \mathbf{z}_i)} = \frac{\psi_{ik}^2}{d^2(\mathbf{0}, \mathbf{z}_i)}.$$

Un cosinus carré proche de 1 signifie que l'individu est proche de l'axe factoriel \mathcal{D}_k et par conséquent bien représenté par cet axe. Un cosinus carré proche de 0 signifie que l'individu est loin de l'axe seulement si la distance à l'origine est grande.

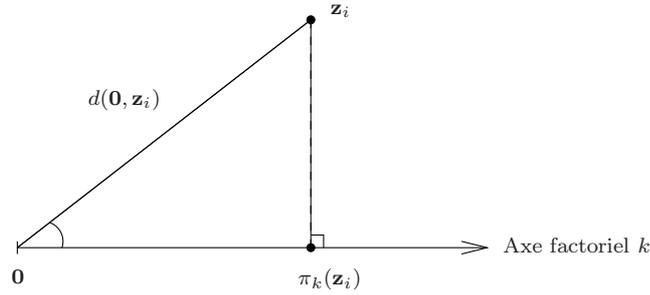


FIGURE 4.1 – Angle définissant le $\cos_k^2(i)$.

Remarque 4.2. Effet taille : Si toutes les variables sont toutes de mêmes ordre, (comme dans l'exercice à venir des voitures) alors, sur le premier axe principal, les individus sont rangés par valeurs croissantes de la moyenne des variables d'origine. La deuxième composante principale différencie alors les individus de taille semblable : on l'appelle le facteur forme.

4.2 AJUSTEMENT DU NUAGE DES VARIABLES $N(J)$

Nous avons vu lors du chapitre 2 qu'à un ajustement du nuage des individus correspond un ajustement du nuage des variables. Nous sommes dans le cadre pondéré de la section 2.7, avec pour matrice de travail $W := D^{\frac{1}{2}}Z$.

4.2.1 DISTANCE ENTRE LES POINTS-VARIABLES

La distance entre points-variables (transformés avec la matrice $D^{\frac{1}{2}}$) devient :

$$\begin{aligned}
 d^2((D^{\frac{1}{2}}Z)^j, (D^{\frac{1}{2}}Z)^{j'}) &= \sum_{i=1}^n [(D^{\frac{1}{2}}Z)_{ij} - (D^{\frac{1}{2}}Z)_{ij'}]^2 \\
 &= \sum_{i=1}^n (\sqrt{p_i}z_{ij} - \sqrt{p_i}z_{ij'})^2 = \sum_{i=1}^n p_i(z_{ij} - z_{ij'})^2 \\
 &= \sum_{i=1}^n p_i z_{ij}^2 + \sum_{i=1}^n p_i z_{ij'}^2 - 2 \sum_{i=1}^n p_i z_{ij} z_{ij'}, \text{ en développant,} \\
 &= r_{jj} + r_{j'j'} - 2r_{jj'}, \text{ d'après la section 3.4.4} \\
 &= 2(1 - r_{jj'}), \text{ car } r_{jj} = r_{j'j'} = 1, \text{ d'après la section 3.4.4.}
 \end{aligned}$$

Les coefficients de corrélations étant compris entre -1 et 1 , on en déduit que la distance entre variables est comprise entre 0 et 2 . Elle s'interprète à partir de la corrélation entre les deux variables. Cette distance est proche de 0 si les variables sont très corrélées linéairement.

Regardons maintenant la distance d'un point-variable à l'origine $\mathbf{0}$ de \mathbb{R}^n :

$$d^2((D^{\frac{1}{2}}Z)^j, \mathbf{0}) = \sum_{i=1}^n p_i z_{ij}^2 = r_{jj} = 1.$$

Ainsi, tous les points-variables sont sur une hyper-sphère de rayon 1 dans \mathbb{R}^n , appelée *sphère des corrélations*. Si l'on garde deux axes, un plan d'ajustement coupera la sphère selon un grand cercle, appelé *cercle des corrélations*, à l'intérieur duquel se trouvera les projections des points-variables.

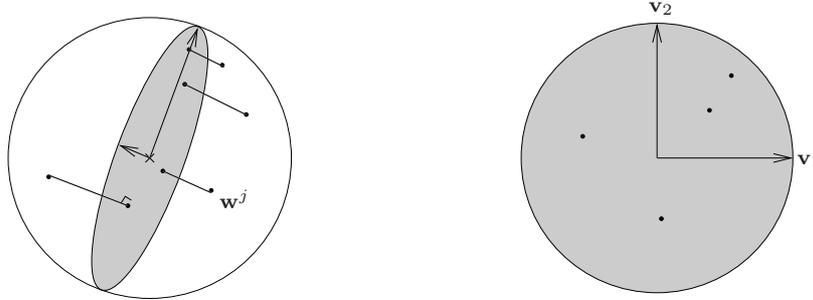


FIGURE 4.2 – Hyper-sphère des points-variables et cercle des corrélations dans le cas de deux axes factoriels.

4.2.2 AJUSTEMENT DU NUAGE DES VARIABLES $N(J)$

Pour le nuage des variables, le problème d'ajustement devient :

Trouver un repère orthonormal $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ tel que $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_1[(D^{\frac{1}{2}}Z)^j])$ soit maximale, puis $\sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_2[(D^{\frac{1}{2}}Z)^j])$, etc.

La solution est le théorème 2.5 du chapitre avec la matrice de travail $W = D^{\frac{1}{2}}Z$:

Théorème 4.2. *La solution au problème d'optimisation sous contrainte est le repère orthonormé $(\mathbf{0}, \mathbf{v}_1, \dots, \mathbf{v}_n)$ formé des vecteurs propres de $D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}$, telle que \mathbf{v}_1 correspond à la plus grande valeur propre μ_1 de $D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}$, puis \mathbf{u}_2 à la deuxième valeur propre μ_2 de $D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, n\}, \quad \sum_{j=1}^p d^2(\mathbf{0}, \tilde{\pi}_k[(D^{\frac{1}{2}}Z)^j]) = \mathbf{v}'_k D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}\mathbf{v}_k = \mathbf{v}'_k D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}\mathbf{v}_k = \mu_k.$$

Chacun des vecteurs \mathbf{v}_k est combinaison linéaire des vecteurs $\mathbf{x}_1, \dots, \mathbf{x}_n$, ce qui signifie que \mathbf{v}_k représente un nouvel individu "abstrait" combinaison des individus d'origine.

4.2.3 PROJECTION DES VARIABLES SUR LES NOUVEAUX INDIVIDUS

Les coordonnées des variables dans le nouveau repère sont décrites dans la matrice Φ :

$$\Phi = (D^{\frac{1}{2}}Z)'V,$$

où la $j^{\text{ème}}$ ligne $\Phi'_j = ((D^{\frac{1}{2}}Z)^j)'V = (D^{\frac{1}{2}}\mathbf{z}^j)'V$ représente les coordonnées de la $j^{\text{ème}}$ variable sur les n axes factoriels $\mathbf{v}_1, \dots, \mathbf{v}_n$; la $k^{\text{ème}}$ colonne $\Phi^k = (D^{\frac{1}{2}}Z)'\mathbf{v}_k$ représente les coordonnées de toutes les variables sur le $k^{\text{ème}}$ axe factoriel (combinaison d'individus cette fois-ci).

4.2.4 RELATION ENTRE LES AJUSTEMENTS DANS LES DEUX ESPACES

Rappelons que les matrices $Z'DZ$ et $D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}$ ont les mêmes valeurs propres non nulles, notées $\lambda_1, \dots, \lambda_q$, avec $q \leq \min\{p, n\}$.

Le théorème 2.6 énoncé avec la matrice de travail $D^{\frac{1}{2}}Z$, donne les relations entre les ajustements dans les deux espaces.

Théorème 4.3. *Soient $\lambda_1 \geq \dots \geq \lambda_q > 0$ les valeurs propres non-nulles de $Z'DZ$. Pour $k \in \{1, \dots, q\}$, soit \mathbf{u}_k , respectivement \mathbf{v}_k , le vecteur propre unitaire correspondant à la valeur propre λ_k de $Z'DZ$, respectivement $D^{\frac{1}{2}}ZZ'D^{\frac{1}{2}}$. Alors, ces vecteurs propres sont liés par les relations :*

$$\begin{cases} \mathbf{v}_k = \frac{D^{\frac{1}{2}}Z\mathbf{u}_k}{\sqrt{\lambda_k}} = \frac{D^{\frac{1}{2}}\Psi^k}{\sqrt{\lambda_k}} \\ \mathbf{u}_k = \frac{(D^{\frac{1}{2}}Z)'\mathbf{v}_k}{\sqrt{\lambda_k}} = \frac{\Phi^k}{\sqrt{\lambda_k}}. \end{cases}$$

On déduit les relations suivantes pour les coordonnées des projections :

$$\begin{aligned} \Psi^k &= Z\mathbf{u}_k = \sqrt{\lambda_k}D^{-\frac{1}{2}}\mathbf{v}_k, \\ \Phi^k &= (D^{\frac{1}{2}}Z)'\mathbf{v}_k = \sqrt{\lambda_k}\mathbf{u}_k. \end{aligned}$$

4.2.5 AIDE À L'INTERPRÉTATION

Les relations de transitions ci-dessus permettent d'interpréter les coordonnées Φ^k de toutes les variables sur le $k^{\text{ème}}$ axe factoriel.

Proposition 4.3. *Pour tout $j \in \{1, \dots, p\}$, tout $k \in \{1, \dots, q\}$, le coefficient ϕ_{jk} de la matrice Φ est égale à la corrélation entre la $j^{\text{ème}}$ variable \mathbf{z}^j et la $k^{\text{ème}}$ composante principale Ψ^k :*

$$\phi_{jk} = r_{j, \Psi^k}.$$

Démonstration. D'après le théorème 4.3, on a,

$$\Phi^k = (D^{\frac{1}{2}}Z)'\mathbf{v}_k = (D^{\frac{1}{2}}Z)' \frac{D^{\frac{1}{2}}\Psi^k}{\sqrt{\lambda_k}} = \frac{1}{\sqrt{\lambda_k}}(DZ)'\Psi^k.$$

Ainsi, pour les composantes ϕ_{jk} , $j \in \{1, \dots, p\}$, cela donne

$$\phi_{jk} = \sum_{i=1}^n p_i z_{ij} \frac{\Psi_{ik}}{\sqrt{\lambda_k}}.$$

La variable \mathbf{z}^j est centrée-réduite. D'autre part, la composante principale Ψ^k (nouvelle variable abstraite) est centrée et a variance λ_k , d'où la variable $\frac{\Psi_{ik}}{\sqrt{\lambda_k}}$ est aussi centrée-réduite. Ainsi, la quantité ci-dessus est la corrélation entre l'ancienne variable \mathbf{z}^j et la composante principale Ψ^k . \square

Définition 4.4. Les coefficients $\phi_{jk} = r_{j, \Psi^k}$, $j \in \{1, \dots, p\}$, $k \in \{1, \dots, q\}$, s'appellent les *corrélations variables facteurs*.

En pratique, pour calculer Φ^k , le vecteur qui contient les corrélations entre la composante principale Ψ^k et toutes les variables $\mathbf{z}^1, \dots, \mathbf{z}^p$, on utilise la relation de transition :

$$\Phi^k = \sqrt{\lambda_k} \mathbf{u}_k.$$

Définition 4.5. Si l'on garde deux axes et que l'on a donc deux composantes principales Ψ^1, Ψ^2 , on représente les corrélations variables facteurs dans le *cercle des corrélations*.

- Les deux axes du repère représentant les deux composantes principales Ψ^1, Ψ^2 .
- Il y a p points, correspondant aux p variables $\mathbf{z}^1, \dots, \mathbf{z}^p$.
- Le point correspondant à la variable \mathbf{z}^j a pour coordonnées les corrélations entre la variable \mathbf{z}^j et les composantes principales Ψ^1 et Ψ^2 :

$$(r_{j,\Psi^1}, r_{j,\Psi^2}) = (\phi_{j1}, \phi_{j2}).$$

Remarque 4.3. On se souvient que les point-variables $\mathbf{z}^1, \dots, \mathbf{z}^p$ se trouvent sur une hyper-sphère de rayon 1 dans \mathbb{R}^n et que la distance entre deux points variables $\mathbf{z}^j, \mathbf{z}^{j'}$ est proche de 0 ou 2 si les variables sont très corrélées linéairement. Selon la qualité de l'ajustement, cette propriété sera plus ou moins bien gardée en projection ; on se gardera donc d'interpréter la proximité entre les points proches sur le cercle des corrélations, sauf lorsqu'ils sont près du cercle (dans ce cas les coordonnées "cachées" par la projection sont presque nulles).

EXEMPLE 4.1. Effectuons l'ACP normée de l'exemple des voitures, issu du livre de Saporta [Sap11], p.428, avec le logiciel SPAD.

Les variables « finition » et « prix » sont mises en variables supplémentaires.

	cyl.	puiss.	long.	larg.	poids	vites.	finit.	prix
Alfasud-ti-135	1350	79	393	161	870	165	1	30570
Audi-100-l	1588	85	468	177	1110	160	2	39990
Simca-1307-gls	1294	68	424	168	1050	152	0	29600
Citroen-gc-club	1222	59	412	161	930	151	0	28250
Fiat-132-1600gls	1585	98	439	164	1105	165	1	34900
Lancia-beta-1300	1297	82	429	169	1080	160	2	35480
Peugeot-504	1796	79	449	169	1160	154	1	32300
R16-tl	1565	55	424	163	1010	140	1	32000
R30-ts	2664	128	452	173	1320	180	2	47700
Toyota-corolla	1166	55	399	157	815	140	0	26540
Alfetta-1.66	1570	109	428	162	1060	175	2	42395
Princess-1800-hl	1798	82	445	172	1160	158	2	33990
Datsun-200l	1998	115	469	169	1370	160	2	43980
Taunus-2000-gl	1993	98	438	170	1080	167	1	35010
Rancho	1442	80	431	166	1129	144	2	39450
Mazda-9295	1769	83	440	165	1095	165	0	27900
Opel-rekord-l	1979	100	459	173	1120	173	1	32700
Lada-1300	1294	68	404	161	955	140	0	22100

- **Matrice des corrélations** R (matrice à diagonaliser).

$$R = \begin{pmatrix} 1,000 & & & & & \\ 0,797 & 1,000 & & & & \\ 0,701 & 0,641 & 1,000 & & & \\ 0,630 & 0,521 & 0,849 & 1,000 & & \\ 0,789 & 0,765 & 0,868 & 0,717 & 1,000 & \\ 0,665 & 0,844 & 0,476 & 0,473 & 0,478 & 1,000 \end{pmatrix}$$

- **Valeurs propres de la matrice** R .

Numéro	Valeur propre	% inertie expliqué	% inertie cumulé
1	4,4209	73,68	73,68
2	0,8561	14,27	87,95
3	0,3731	6,22	94,17
4	0,2139	3,57	97,73
5	0,0928	1,55	99,28
6	0,0433	0,72	100

Voici quelques critères pour déterminer le nombre de valeurs propres à garder :

- Se rappelant que les valeurs propres représentent la variance des composantes principales (nouvelles variables), ne garder que celles qui sont supérieures à 1, *i.e.*, celles qui ont une plus grande variabilité que les données d'origine qui est de 1.
- Tracer les pourcentages d'inertie sur un graphe et regarder où il y a un point de rupture.
- Garder au moins deux axes afin de pouvoir extraire suffisamment d'informations de l'ACP.

Dans le cas de cet exemple, même si nous ne sommes pas en accord avec le premier critère, nous décidons de garder deux valeurs propres, soit deux axes.

- **Plan factoriel.**

Pour chacun des individus i , on obtient : son poids relatif p_i , la distance à l'origine (au carré) $\text{Disto}(i)$. Pour $k = 1, 2$ (on garde deux axes), on a aussi la coordonnée Ψ_{ik} sur l'axe k , la contribution $\text{CRT}_k(i)$ de l'individu à l'inertie de l'axe k , le cosinus carré $\cos_k^2(i)$ entre le vecteur $(\mathbf{0z}_i)$ et l'axe k .

Individus	poids rel.	Disto	Coordonnées		CRT		cos ²	
			1	2	1	2	1	2
Alfasud-ti-135	5,556	8,225	-2,139	-1,786	5,749	20,693	0,556	0,388
Audi-100-l	5,556	6,674	1,561	1,527	3,064	15,133	0,365	0,349
Simca-1307-gls	5,556	2,159	-1,119	0,675	1,575	2,953	0,580	0,211
Citroen-gc-club	5,556	6,780	-2,574	-0,113	8,324	0,083	0,977	0,002
Fiat-132-1600gls	5,556	1,169	0,428	-0,696	0,230	3,140	0,157	0,414
Lancia-beta-1300	5,556	1,135	-0,304	0,196	0,116	0,250	0,082	0,034
Peugeot-504	5,556	1,513	0,684	0,933	0,588	5,650	0,309	0,575
R16-tl	5,556	5,637	-1,948	0,980	4,771	6,238	0,674	0,171
R30-ts	5,556	21,790	4,410	-1,064	24,437	7,342	0,892	0,052
Toyota-corolla	5,556	16,290	-3,986	-0,236	19,964	0,362	0,975	0,003
Alfetta-1.66	5,556	4,457	0,438	-1,912	0,241	23,736	0,043	0,821
Princess-1800-hl	5,556	1,953	1,018	0,842	1,303	4,598	0,531	0,363
Datsun-200l	5,556	11,113	2,941	0,559	10,870	2,029	0,778	0,028
Taurus-2000-gl	5,556	2,453	1,315	-0,487	2,173	1,536	0,705	0,096
Rancho	5,556	1,963	-0,691	0,898	0,600	5,230	0,243	0,410
Mazda-9295	5,556	0,685	0,386	-0,356	0,187	0,823	0,217	0,185
Opel-rekord-l	5,556	6,083	2,290	-0,104	6,589	0,071	0,862	0,002
Lada-1300	5,556	7,922	-2,709	0,144	9,219	0,134	0,926	0,003

Par exemple, pour le premier individu :

- $p_1 = \frac{1}{18} \cdot 100 \simeq 5,56$.

- Nouvelles coordonnées sur les deux premiers axes : $\psi_{11} = -2,139, \psi_{12} = -1,786$.

- $\text{Disto}(1) \simeq 2.139^2 + 1.786^2 + \dots + 0.054^2 \simeq 8.225$.

- $\text{CRT}_1(1) = \frac{p_1 \psi_{11}^2}{\lambda_1} \simeq \frac{1}{18} \cdot \frac{(-2.139)^2}{4.4209} \simeq 5.749\%$.

- $\cos^2_1(1) = \frac{\psi_{11}^2}{d^2(\mathbf{0}, \mathbf{z}_i)} \simeq \frac{(-2.139)^2}{8.225} = 0.556$.

On obtient le graphique suivant du plan factoriel pour les individus.

• **Corrélations variables-facteurs**

Le tableau suivant donne les corrélations variables-facteurs. Les deux premières colonnes sont représentées dans le cercle des corrélations :

	Ψ^1	Ψ^2	Ψ^3	Ψ^4	Ψ^5	Ψ^6
cyl	0,893	-0,115	0,216	-0,374	-0,046	0,012
puiss	0,887	-0,385	0,113	0,165	0,089	0,132
long	0,886	0,381	-0,041	0,129	-0,223	0,040
larg	0,814	0,413	-0,369	-0,098	0,146	0,023
poids	0,905	0,225	0,296	0,140	0,093	-0,121
vites	0,755	-0,574	-0,297	0,034	-0,057	-0,095

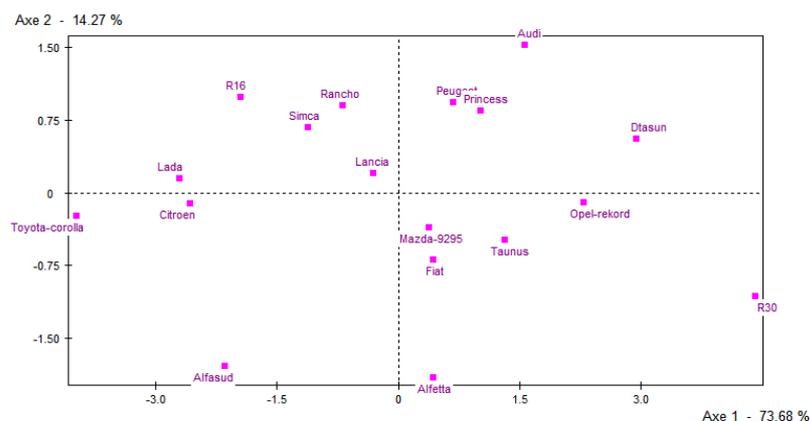


FIGURE 4.3 – Projection des individus sur le plan principal obtenu avec SPAD. Il explique $\sim 87,95\%$ de l’inertie (de la variabilité) des données.

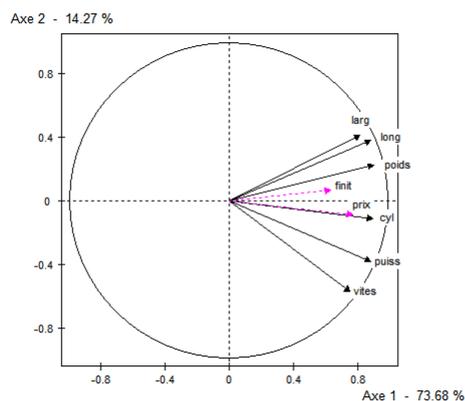


FIGURE 4.4 – Cercle des corrélations obtenu avec SPAD.

• **Analyse.**

Dans le cercle des corrélations, si les variables sont bien représentées (c'est-à-dire qu'elles ont une forte corrélation avec les axes, ou encore, qu'elles sont situées vers l'extrémité du cercle), on a l'interprétation suivante : les variables proches sont corrélées positivement, deux variables opposées sont corrélées négativement, deux variables orthogonales ne sont pas corrélées.

Dans notre cas, le cercle des corrélations montre un effet "taille" sur le premier axe qui va donc classer les individus selon leur taille. Le deuxième axe oppose longueur/largeur et vitesse puissance, il s'interprète comme opposant les véhicules sportifs et les véhicules familiaux. Les variables supplémentaires montre en outre que la première composante principale est liée à la qualité et au prix.

Pour l'interprétation des individus, on utilise la règle suivante. Si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés

par une valeur élevée de la variable.

Ainsi, les individus se trouvant dans le cadran de droite (resp. gauche) sont de grands (resp. petits) véhicules ; les individus se trouvant en haut du demi-plan supérieur (resp. inférieur) sont des véhicules familiaux (resp. sportifs).

Si la représentation des individus dans le plan factoriel est bonne (c'est à dire que le \cos^2 est proche de 1), la proximité entre individus est l'image d'une proximité réelle. Cela permet alors de créer des groupes/classes d'individus qui se ressemblent. Par exemple, Mazda, Fiat etc. On peut ensuite confirmer cette répartition en groupes grâce aux méthodes de classification, voir chapitre 6.

4.3 EXERCICE

Exercice 4.1. On considère 6 ménages $\mathbf{x}_1, \dots, \mathbf{x}_6$ et 3 variables $\mathbf{x}^1, \dots, \mathbf{x}^3$, où \mathbf{x}^1 représente la consommation d'eau, \mathbf{x}^2 la consommation d'électricité, \mathbf{x}^3 la consommation de gaz (unité 100 euros). On suppose que la variable \mathbf{x}^3 est illustrative, que les individus $\mathbf{x}_1, \dots, \mathbf{x}_5$ ont poids $\frac{1}{5}$ et que l'individu \mathbf{x}_6 a poids 0 (il est illustratif).

	\mathbf{x}^1	\mathbf{x}^2	\mathbf{x}^3	Poids
\mathbf{x}'_1	2	5	3	$\frac{1}{5}$
\mathbf{x}'_2	4	2	4	$\frac{1}{5}$
\mathbf{x}'_3	6	3	3	$\frac{1}{5}$
\mathbf{x}'_4	3	4	3	$\frac{1}{5}$
\mathbf{x}'_5	5	1	6	$\frac{1}{5}$
\mathbf{x}'_6	3	6	7	0

1. Calculer la matrice des corrélations R et celle des données centrées-réduites Z .
2. Trouver les vecteurs propres et valeurs propres de la matrice des corrélations R . En déduire les axes factoriels obtenus par l'ACP et le pourcentage d'inertie expliqué par le premier axe, par les deux premiers axes.
3. Déterminer les coordonnées des individus actifs et illustratif dans le nouveau repère.
4. Détermine les coordonnées des variables actives et illustratives dans le repère provenant de l'ajustement du nuage des points-variables. Que représente les coordonnées des variables actives ?
5. Faire l'ACP normée de ces données avec SPAD. Expliquer toutes les issues (sauf celles impliquant un test statistique).

Solution de l'exercice 4.1.

1. Calcul la matrice des corrélations R et des données centrées-réduites Z .
 - Calcul de la moyenne des variables :

$$\bar{x}^1 = \frac{2 + 4 + 6 + 3 + 5}{5} = 4, \quad \bar{x}^2 = \frac{5 + 2 + 3 + 4 + 1}{5} = 3.$$

- Tableau intermédiaire :

	\mathbf{x}^1	\mathbf{x}^2	$\mathbf{x}^1 - \bar{x}^1$	$\mathbf{x}^2 - \bar{x}^2$	$(\mathbf{x}^1 - \bar{x}^1)^2$	$(\mathbf{x}^2 - \bar{x}^2)^2$	$(\mathbf{x}^1 - \bar{x}^1)(\mathbf{x}^2 - \bar{x}^2)$
\mathbf{x}'_1	2	5	-2	2	4	4	-4
\mathbf{x}'_2	4	2	0	-1	0	1	0
\mathbf{x}'_3	6	3	2	0	4	0	0
\mathbf{x}'_4	3	4	-1	1	1	1	-1
\mathbf{x}'_5	5	1	1	-2	1	4	-2

- Nous déduisons les variances et covariances :

$$v_1 = \frac{4+0+4+1+1}{5} = 2, \quad v_2 = \frac{4+1+0+1+4}{5} = 2, \quad v_{12} = \frac{-4+0+0-1-2}{5} = -\frac{7}{5}$$

et les écarts-types : $s_1 = \sqrt{2}$, $s_2 = \sqrt{2}$.

- Nous pouvons maintenant calculer la corrélation linéaire entre \mathbf{x}^1 et \mathbf{x}^2 :

$$r_{12} = \frac{v_{12}}{s_1 s_2} = \frac{-\frac{7}{5}}{\sqrt{2}\sqrt{2}} = -\frac{7}{10}.$$

et déduire la matrice des corrélations R , qui est la matrice à diagonaliser :

$$R = \begin{pmatrix} 1 & -\frac{7}{10} \\ -\frac{7}{10} & 1 \end{pmatrix}$$

- Nous obtenons aussi le tableau Z des données centrées-réduites, et le tableau de travail $W = D^{\frac{1}{2}}Z$. (Souvenez-vous que $R = W'W = Z'DZ$).

$$Z = \begin{pmatrix} \frac{-2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \\ 0 & \frac{-1}{\sqrt{2}} \\ \frac{2}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-2}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -2 & 2 \\ 0 & -1 \\ 2 & 0 \\ -1 & 1 \\ 1 & -2 \end{pmatrix}, \quad W = D^{\frac{1}{2}}Z = \begin{pmatrix} \frac{-2}{\sqrt{5}\sqrt{2}} & \frac{2}{\sqrt{5}\sqrt{2}} \\ 0 & \frac{-1}{\sqrt{5}\sqrt{2}} \\ \frac{2}{\sqrt{5}\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{5}\sqrt{2}} & \frac{1}{\sqrt{5}\sqrt{2}} \\ \frac{1}{\sqrt{5}\sqrt{2}} & \frac{-2}{\sqrt{5}\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{10}} \begin{pmatrix} -2 & 2 \\ 0 & -1 \\ 2 & 0 \\ -1 & 1 \\ 1 & -2 \end{pmatrix}.$$

- Calcul des valeurs propres et vecteurs propres de la matrice des corrélations R .

On remarque que la somme de chacune des lignes de R est égale à $\frac{3}{10}$, ce qui implique,

$$R \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{7}{10} \\ -\frac{7}{10} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{10} \\ \frac{3}{10} \end{pmatrix} = \frac{3}{10} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

D'où $\lambda = \frac{3}{10}$ est valeur propre de R et $\mathbf{u} = -\frac{1}{\sqrt{2}}(1, 1)'$ est un vecteur propre unitaire correspondant.

Pour déterminer la deuxième valeur propre, on se rappelle que la somme des valeurs propres est égale à la trace de la matrice R , d'où la deuxième valeur propre est égale à

$$\text{Tr}(R) - \lambda = 2 - \frac{3}{10} = \frac{17}{10}.$$

Un vecteur propre associé est orthogonal à \mathbf{u} , soit de la forme $\alpha(-1, 1)'$. En imposant qu'il soit normé à 1, on obtient $\frac{1}{\sqrt{2}}(-1, 1)'$ (ou son opposé).

Ainsi, $\lambda_1 = \frac{17}{10}$ est la plus grande valeur propre de R , avec vecteur propre unitaire $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(-1, 1)'$ et $\lambda_2 = \frac{3}{10}$ est la deuxième valeur propre de R , avec vecteur propre unitaire $\mathbf{u}_2 = -\frac{1}{\sqrt{2}}(1, 1)'$.

D'après un résultat du cours, le repère donné par l'ACP est formé des vecteurs propres de la matrice des corrélations R , ordonnés de sorte à ce que les valeurs propres correspondantes soient classées de la plus grande à la plus petite. Ainsi, les axes factoriels de l'ACP sont les axes passant par l'origine et de vecteur directeur unitaire \mathbf{u}_1 et \mathbf{u}_2 respectivement. Le pourcentage d'inertie expliquée par le premier axe est :

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\frac{17}{10}}{\frac{17}{10} + \frac{3}{10}} = \frac{17}{20} = 0,85.$$

Le pourcentage d'inertie expliquée par les deux premiers axes est :

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2} = 1.$$

On retrouve toute la dispersion des données avec deux axes.

3. Calcul des coordonnées des données centrées-réduites dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$.

Ménages actifs

Soit $U = (\mathbf{u}_1 \mathbf{u}_2)$ la matrice dont les colonnes sont les vecteurs propres de la matrice des corrélations :

$$U = \begin{pmatrix} \frac{-1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}$$

Alors les coordonnées des données centrées-réduites dans le nouveau repère $(\mathbf{0}, \mathbf{u}_1, \mathbf{u}_2)$ sont données par la matrice $\Psi = ZU$:

$$\Psi = ZU = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \begin{pmatrix} -2 & 2 \\ 0 & -1 \\ 2 & 0 \\ -1 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 4 & 0 \\ -1 & 1 \\ -2 & -2 \\ 2 & 0 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ -0,5 & 0,5 \\ -1 & -1 \\ 1 & 0 \\ -1,5 & 0,5 \end{pmatrix}.$$

Pour $i \in \{1, \dots, 5\}$, la $i^{\text{ème}}$ ligne Ψ_i représente les coordonnées du $i^{\text{ème}}$ ménage centré-réduit sur les deux axes factoriels $\mathbf{u}_1, \mathbf{u}_2$.

Ménage illustratif

Nous pouvons aussi calculer les coordonnées du ménage illustratif centré-réduit $\mathbf{z}'_6 = \left(\frac{3-4}{\sqrt{2}}, \frac{6-3}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}}(-1, 3)$:

$$\psi_6 = \mathbf{z}'_6 U = \frac{1}{2} \begin{pmatrix} -1 & 3 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 4 & -2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \end{pmatrix}.$$

4. Calcul des coordonnées pour les variables dans le nouveau repère.

Variables actives

On cherche à calculer la matrice $\Phi = (D^{\frac{1}{2}}Z)'V = W'V$. Pour cela, on utilise les formules de transition qui nous disent que, pour $k = 1, 2$, $\Phi^k = \sqrt{\lambda_k} \mathbf{u}_k$. D'où :

$$\Phi = \left(\sqrt{\lambda_1} \mathbf{u}_1 \quad \sqrt{\lambda_2} \mathbf{u}_2 \right) = \begin{pmatrix} -\sqrt{\frac{1,7}{2}} & -\sqrt{\frac{0,3}{2}} \\ \sqrt{\frac{1,7}{2}} & -\sqrt{\frac{0,3}{2}} \end{pmatrix} \simeq \begin{pmatrix} -0,92 & -0,39 \\ 0,92 & -0,39 \end{pmatrix}.$$

Les coordonnées des variables actives dans le nouveau repère représente les corrélations variables-facteurs. Le coefficient ϕ_{jk} est la corrélation entre la $j^{\text{ème}}$ variable et le $k^{\text{ème}}$ facteur.

Le premier axe représente donc bien les deux variables eau et électricité.

Variable illustrative

Pour calculer les coordonnées de la variable illustrative "gaz", nous centrons et réduisons le vecteur $\mathbf{x}^3 = (3, 4, 3, 3, 6)'$ pour obtenir $\mathbf{z}^3 = (-0,686; 0,171; -0,686; -0,686; 1,886)'$. Nous le multiplions ensuite par $D^{\frac{1}{2}}$, pour obtenir $\mathbf{w}^3 = (-0,31; 0,08; -0,31; -0,31; 0,84)'$. Nous utilisons la formule de transition pour calculer la matrice V , et déduisons :

$$\begin{aligned} \Phi'_3 &= (\mathbf{w}^3)'V = [(D^{\frac{1}{2}}Z)^3]' \begin{pmatrix} \frac{D^{\frac{1}{2}}\Psi^1}{\sqrt{\lambda_1}} & \frac{D^{\frac{1}{2}}\Psi^2}{\sqrt{\lambda_2}} \end{pmatrix} \\ &= [(D^{\frac{1}{2}}Z)^3]' \begin{pmatrix} \frac{2}{\sqrt{5 \cdot 1,7}} & 0 \\ -\frac{0,5}{\sqrt{5 \cdot 1,7}} & \frac{0,5}{\sqrt{5 \cdot 0,3}} \\ -\frac{1}{\sqrt{5 \cdot 1,7}} & -\frac{1}{\sqrt{5 \cdot 0,3}} \\ \frac{1}{\sqrt{5 \cdot 1,7}} & 0 \\ -\frac{1,5}{\sqrt{5 \cdot 1,7}} & \frac{0,5}{\sqrt{5 \cdot 0,3}} \end{pmatrix} \simeq \begin{pmatrix} -0,66 & 0,63 \end{pmatrix}. \end{aligned}$$

5. Avec SPAD, nous obtenons le plan factoriel et le cercle des corrélations.

Après l'analyse, les seules issues qu'il nous reste à expliquer sont les distances carrées à l'origine, les contributions des individus à l'inertie des axes, les cosinus carrés. Nous les résumons dans le tableau ci-dessous, et les calculons seulement pour le premier individu.

Individus	Disto	CRT		cos ²	
		1	2	1	2
1	4,000	47,059	0,000	1,000	0,000
2	0,500	2,941	16,667	0,500	0,500
3	2,000	11,765	66,667	0,500	0,500
4	1,000	11,765	0,000	1,000	0,000
5	2,500	26,471	16,667	0,900	0,100

- Distance au carré de l'individu à l'origine : Disto(1).

$$\text{Disto}(1) = d^2(\mathbf{0}, \mathbf{z}_1) = \frac{1}{2}[2^2 + 2^2] = 4.$$

- Contribution de l'individu 1 à l'inertie sur les axes principaux : (CRT₁(1), CRT₂(1))

$$(\text{CRT}_1(1), \text{CRT}_2(1)) = \left(\frac{p_1 \psi_{11}^2}{\lambda_1}, \frac{p_1 \psi_{12}^2}{\lambda_2} \right) = \left(\frac{2^2}{5 \cdot 1,7}, \frac{(0)^2}{5 \cdot 0,3} \right) \simeq (0,471; 0).$$

- Cosinus carré entre le vecteur $(\mathbf{0z}_1)$ et les axes principaux : $(\cos_1^2(1), \cos_2^2(1))$

$$(\cos_1^2(1), \cos_2^2(1)) = \left(\frac{\Psi_{11}^2}{d^2(\mathbf{0}, \mathbf{z}_1)}, \frac{\Psi_{12}^2}{d^2(\mathbf{0}, \mathbf{z}_1)} \right) = \left(\frac{2^2}{4}, \frac{0^2}{4} \right) = (1, 0).$$

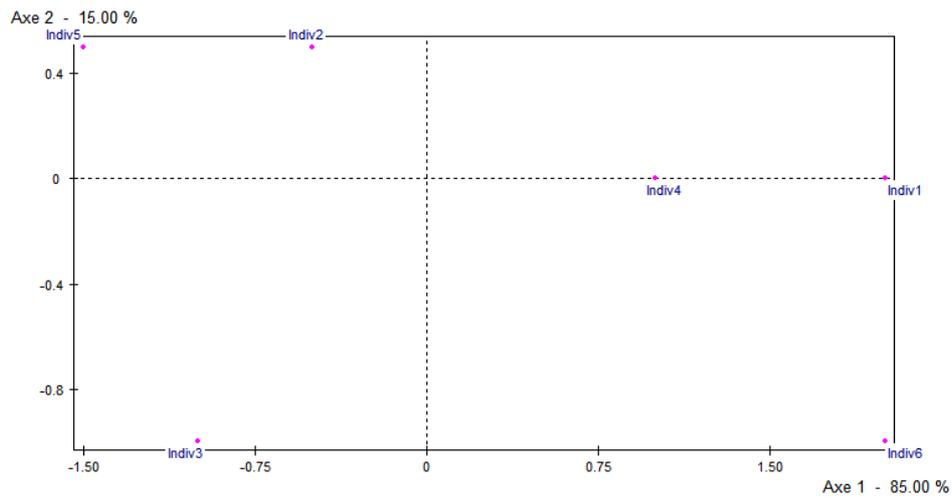


FIGURE 4.5 – Projection des individus sur le plan principal obtenu avec SPAD. Il explique 100% de l'inertie (de la variabilité) des données.

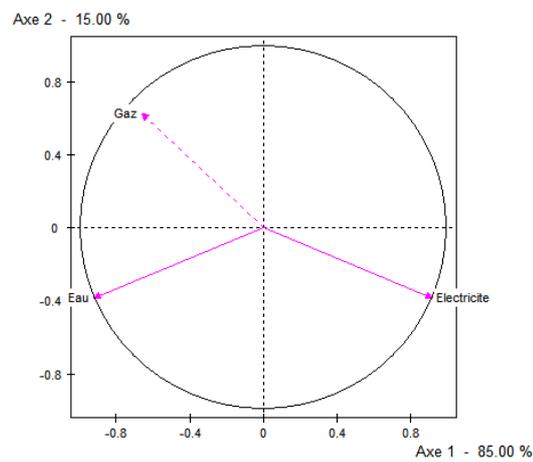


FIGURE 4.6 – Cercle des corrélations obtenu avec SPAD.

CHAPITRE 5

ANALYSE FACTORIELLE DE CORRESPONDANCES

On considère deux variables X et Y , qualitatives ou quantitatives décrivant un ensemble de N individus. L'Analyse Factorielle de Correspondances (AFC), introduite par Benzécri dans les années 1960, a pour but de décrire et d'étudier le tableau de contingence des variables X et Y lorsqu'elles ne sont pas indépendantes.

5.1 DISTRIBUTIONS MARGINALES ET CONDITIONNELLES

On suppose que la variable X a n modalités $\{x_1, \dots, x_n\}$, et que la variable Y a m modalités $\{y_1, \dots, y_m\}$.

5.1.1 DISTRIBUTIONS MARGINALES, TABLEAUX DE CONTINGENCE

Afin de définir le tableau de contingence des effectifs ou des fréquences des variables X et Y , on introduit les notions suivantes.

- Pour tout $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$, l'*effectif* du couple (x_i, y_j) , noté k_{ij} , est le nombre d'individus tels que $X = x_i$ et $Y = y_j$; sa *fréquence*, noté f_{ij} , est le quotient de son effectif par la taille de la population :

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

- Pour tout $i \in \{1, \dots, n\}$ (respectivement $j \in \{1, \dots, m\}$); l'*effectif marginal* de x_i (respectivement y_j), noté $k_{i\cdot}$ (respectivement $k_{\cdot j}$), est l'effectif de x_i (respectivement y_j) pour la série simple de X (respectivement Y) :

$$k_{i\cdot} = \sum_{j=1}^m k_{ij} \quad \text{et} \quad k_{\cdot j} = \sum_{i=1}^n k_{ij}.$$

- La *fréquence marginale* de x_i (respectivement y_j), notée $f_{i\cdot}$ (respectivement $f_{\cdot j}$), est le quotient de l'effectif marginal de x_i (respectivement y_j) par la taille de la population :

$$f_{i\cdot} = \frac{k_{i\cdot}}{N} = \sum_{j=1}^m f_{ij} \quad \text{et} \quad f_{\cdot j} = \frac{k_{\cdot j}}{N} = \sum_{i=1}^n f_{ij}.$$

Remarque. On a,

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

$$\sum_{i=1}^n k_{i\cdot} = \sum_{j=1}^m k_{\cdot j} = N \quad \text{et} \quad \sum_{i=1}^n f_{i\cdot} = \sum_{j=1}^m f_{\cdot j} = 1.$$

Notation. On note $F = (f_{ij})$ la matrice de taille $n \times m$ dont les coefficients sont les fréquences. Les données sont représentées sous la forme de tableaux croisés formés de n lignes et m colonnes, augmentés de leurs marges. Selon que l'on mette les effectifs ou les fréquences, on obtient le *tableau de contingence des effectifs* ou des *fréquences* :

$X \backslash Y$	y_1	\cdots	y_j	\cdots	y_m	Total
x_1	k_{11}	\cdots	k_{1j}	\cdots	k_{1m}	$k_{1\cdot}$
\vdots	\vdots		\vdots			\vdots
x_i	k_{i1}	\cdots	k_{ij}	\cdots	k_{im}	$k_{i\cdot}$
\vdots	\vdots		\vdots			\vdots
x_p	k_{p1}	\cdots	k_{pj}	\cdots	k_{pm}	$k_{p\cdot}$
Total	$k_{\cdot 1}$	\cdots	$k_{\cdot j}$	\cdots	$k_{\cdot m}$	N

FIGURE 5.1 – Tableau de contingence des effectifs.

$X \backslash Y$	y_1	\cdots	y_j	\cdots	y_m	Total
x_1	f_{11}	\cdots	f_{1j}	\cdots	f_{1m}	$f_{1\cdot}$
\vdots	\vdots		\vdots			\vdots
x_i	f_{i1}	\cdots	f_{ij}	\cdots	f_{im}	$f_{i\cdot}$
\vdots	\vdots		\vdots			\vdots
x_n	f_{n1}	\cdots	f_{nj}	\cdots	f_{nm}	$f_{n\cdot}$
Total	$f_{\cdot 1}$	\cdots	$f_{\cdot j}$	\cdots	$f_{\cdot m}$	1

FIGURE 5.2 – Tableau de contingence des fréquences.

EXEMPLE. Voici un exemple tiré de l'élection présidentielle de 1981 ; X désigne la variable "les départements", Y la variable "les candidats", k_{ij} est le nombre de voix (en milliers) obtenues par le candidat j dans le département i . Nous nous limitons à 4 candidats et 6 départements (on a $n = 6$, $m = 4$) :

Modalités pour X : {Ain, Hautes-Alpes, Ariège, Bouche du Rhône, Charente-Maritime, Drôme}

Modalités pour Y : {Mitterrand, Giscard d'Estaing, Chirac, Marchais}

Le tableau de contingence des effectifs est :

$X \backslash Y$	Mitterrand	Giscard	Chirac	Marchais	Total
Ain	51	64	36	23	174
Hautes-Alpes	14	17	9	9	49
Ariège	27	18	13	17	75
Bouche du Rhône	191	204	119	205	719
Charente-Maritime	71	76	47	37	231
Drôme	57	55	31	30	173
Total	411	434	255	321	1421

Dans cet exemple, $k_{23} = 9$, $k_{2.} = 49$, $k_{.3} = 255$, $N = 1421$.

Le tableau de contingence des fréquences est :

$X \backslash Y$	Mitterrand	Giscard	Chirac	Marchais	Total ($f_{i.}$)
Ain	51/1421	64/1421	36/1421	23/1421	174/1421
Hautes-Alpes	14/1421	17/1421	9/1421	9/1421	49/1421
Ariège	27/1421	18/1421	13/1421	17/1421	75/1421
Bouche du Rhône	191/1421	204/1421	119/1421	205/1421	719/1421
Charente-Maritime	71/1421	76/1421	47/1421	37/1421	231/1421
Drôme	57/1421	55/1421	31/1421	30/1421	173/1421
Total ($f_{.j}$)	411/1421	434/1421	255/1421	321/1421	1421/1421

5.1.2 DISTRIBUTIONS CONDITIONNELLES, TABLEAUX DES PROFILS

Deux lectures du tableau de contingence sont possibles, selon que l'on privilégie l'une ou l'autre des deux variables : lecture en ligne ou lecture en colonne.

En extrayant une colonne (respectivement une ligne) du tableau de contingence, on obtient une *distribution conditionnelle* de X (respectivement de Y). Précisément, l'extraction de la colonne j (respectivement de la ligne i) permet de répondre à la question « Sachant que Y vaut y_j (respectivement X vaut x_i), comment se comporte X (respectivement Y)? » Ainsi, une distribution conditionnelle est l'étude d'un critère lorsque l'autre reste fixe à une valeur ou modalité connue. On note $X|_{Y=y_j}$ (respectivement $Y|_{X=x_i}$) la distribution conditionnelle de X sachant $Y = y_j$ (respectivement de Y sachant $X = x_i$).

Pour tout $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$, on définit la *fréquence conditionnelle* de x_i sachant $Y = y_j$ (respectivement y_j sachant $X = x_i$), notée $f_{i|Y=y_j}$ ou plus simplement $f_{i|j}$ (respectivement $f_{j|X=x_i}$ ou $f_{j|i}$), comme étant la fréquence d'occurrence de x_i dans la colonne j (respectivement y_j dans la ligne i) du tableau de contingence :

$$f_{i|j} = \frac{k_{ij}}{k_{.j}} = \frac{f_{ij}}{f_{.j}} \quad \text{et} \quad f_{j|i} = \frac{k_{ij}}{k_{i.}} = \frac{f_{ij}}{f_{i.}}$$

Remarque. Les fréquences conditionnelles vérifient les égalités suivantes :

$$\forall j \in \{1, \dots, m\}, \sum_{i=1}^n f_{i|j} = 1 \quad \text{et} \quad \forall i \in \{1, \dots, n\}, \sum_{j=1}^m f_{j|i} = 1.$$

PROFILS LIGNES

Pour tout $i \in \{1, \dots, n\}$, le *profil-ligne* i , noté $\mathbf{f}_{|i}$, est le point de \mathbb{R}^m dont les coordonnées sont les fréquences conditionnelles sachant $X = x_i$:

$$\mathbf{f}_{|i} = (f_{1|i}, \dots, f_{m|i})' = \left(\frac{f_{i1}}{f_{i\cdot}}, \dots, \frac{f_{im}}{f_{i\cdot}} \right)'.$$

On construit alors le *tableau des profils-lignes* :

$\begin{array}{c} Y X \\ X \end{array}$	Mitterrand	Giscard	Chirac	Marchais	Total
Ain	51/174	64/174	36/174	23/174	1
Hautes-Alpes	14/49	17/49	9/49	9/49	1
Ariège	27/75	18/75	13/75	17/75	1
Bouche du Rhône	191/719	204/719	119/719	205/719	1
Charente-Maritime	71/231	76/231	47/231	37/231	1
Drôme	57/173	55/173	31/173	30/173	1

FIGURE 5.3 – Tableau des profils-lignes

On note $N(I)$ le nuage des profils-lignes représentés comme n points de \mathbb{R}^m :

$$N(I) = \{\mathbf{f}_{|i} \mid i \in \{1, \dots, n\}\}.$$

Chacun des points $\mathbf{f}_{|i}$ est affecté d'une masse $f_{i\cdot}$ qui est sa fréquence relative.

Le *centre de gravité* ou *barycentre* du nuage des profils-lignes, noté $\bar{\mathbf{f}}^1$, est la moyenne des profils-lignes affectés de leur masse. La $j^{\text{ème}}$ composante du vecteur $\bar{\mathbf{f}}^1$ de \mathbb{R}^m est, \bar{f}_j^1 :

$$\bar{f}_j^1 = \sum_{i=1}^n f_{i\cdot} \frac{f_{ij}}{f_{i\cdot}} = \sum_{i=1}^n f_{ij} = f_{\cdot j}.$$

Il s'agit de la fréquence marginale de la $j^{\text{ème}}$ colonne. Ainsi, on a :

$$\bar{\mathbf{f}}^1 = (f_{\cdot 1}, \dots, f_{\cdot m})'.$$

PROFILS-COLONNES

Pour tout $j \in \{1, \dots, m\}$, le *profil-colonne* j , noté $\mathbf{f}_{|j}$, est le point de \mathbb{R}^n dont les coordonnées sont les fréquences conditionnelles sachant $Y = y_j$:

$$\mathbf{f}_{|j} = (f_{1|j}, \dots, f_{n|j})' = \left(\frac{f_{1j}}{f_{\cdot j}}, \dots, \frac{f_{nj}}{f_{\cdot j}} \right)'.$$

On construit alors le *tableau des profils-colonnes* :

$X Y \backslash Y$	Mitterrand	Giscard	Chirac	Marchais
Ain	51/411	64/434	36/255	23/321
Hautes-Alpes	14/411	17/434	9/255	9/321
Ariège	27/411	18/434	13/255	17/321
Bouche du Rhône	191/411	204/434	119/255	205/321
Charente-Maritime	71/411	76/434	47/255	37/321
Drôme	57/411	55/434	31/255	30/321
marges	1	1	1	1

De manière symétrique, on note $N(J)$ le nuage des profils-colonnes représentés comme m points de \mathbb{R}^n :

$$N(J) = \{\mathbf{f}_{|j} \mid j \in \{1, \dots, m\}\}.$$

Chacun des points $\mathbf{f}_{|j}$ est affecté d'une masse $f_{\cdot j}$ qui est sa fréquence relative. Le *centre de gravité* ou *barycentre*, noté $\bar{\mathbf{f}}^c$, est la moyenne des profils-colonnes affectés de leur masse. Pour tout $i \in \{1, \dots, n\}$, la $i^{\text{ème}}$ composante du vecteur $\bar{\mathbf{f}}^c$ de \mathbb{R}^n est,

$$\bar{f}_i^c = \sum_{j=1}^m f_{\cdot j} \frac{f_{ij}}{f_{\cdot j}} = \sum_{j=1}^m f_{ij} = f_{i\cdot}.$$

Il s'agit de la fréquence marginale de la $i^{\text{ème}}$ ligne. Ainsi, on a :

$$\bar{\mathbf{f}}^c = (f_{1\cdot}, \dots, f_{n\cdot})'.$$

NOTATIONS

On appelle $P^l = (f_{j|i})$ (resp. $P^c = (f_{i|j})$) la matrice de taille $n \times m$ dont les coefficients sont les fréquences conditionnelles sachant $\{X = x_i\}$ (resp. sachant $\{Y = y_j\}$). Soit D_n (resp. D_m) la matrice diagonale de taille $n \times n$ (resp. $m \times m$) dont la diagonale est formée des fréquences marginales $\{f_{1\cdot}, \dots, f_{n\cdot}\}$ (resp. $\{f_{\cdot 1}, \dots, f_{\cdot m}\}$), et soit D_n^{-1} (resp. D_m^{-1}) la matrice diagonale dont la diagonale est formée des inverses. Alors, on a :

$$P^l = D_n^{-1}F, \quad P^c = FD_m^{-1}.$$

Remarque 5.1. La somme des coordonnées des points du nuage des profils-lignes (resp. profils-colonnes) est égale à 1. Ceci signifie que ces points se trouvent dans un hyper-plan de \mathbb{R}^m (resp. \mathbb{R}^n).

5.1.3 INDÉPENDANCE

Les variables X et Y sont dites *indépendantes* lorsque, $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$,

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

Proposition 5.1. *Les assertions suivantes sont équivalents :*

1. Les variables X et Y sont indépendantes,
2. $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}, f_{i|j} = f_{i\cdot}$.
3. $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}, f_{j|i} = f_{\cdot j}$.

5.2 AJUSTEMENT DU NUAGE DES PROFILS-LIGNES ET COLONNES

5.2.1 DISTANCE DU χ^2

La distance euclidienne usuelle de \mathbb{R}^m entre deux profils-lignes traduit la différence ou ressemblance entre les lois conditionnelles sachant $X = x_i$ et sachant $X = x_{i'}$, sans tenir compte des effectifs totaux de chacune des modalités :

$$d^2(\mathbf{f}_{|i}, \mathbf{f}_{|i'}) = \sum_{j=1}^m (f_{j|i} - f_{j|i'})^2 = \|\mathbf{f}_{|i} - \mathbf{f}_{|i'}\|^2.$$

Cependant, cette distance favorise les colonnes qui ont une masse $f_{.j}$ importante. Pour remédier à cela, on pondère chaque colonne par l'inverse de la masse de la colonne et l'on obtient une nouvelle distance appelée *distance du χ^2* :

$$d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \mathbf{f}_{|i'}) = \|\mathbf{f}_{|i} - \mathbf{f}_{|i'}\|_{D_m^{-1}}^2 = (\mathbf{f}_{|i} - \mathbf{f}_{|i'})' D_m^{-1} (\mathbf{f}_{|i} - \mathbf{f}_{|i'}) = \sum_{j=1}^m \frac{1}{f_{.j}} (f_{j|i} - f_{j|i'})^2.$$

Remarque. Deux profils sont proches si les fréquences conditionnelles sachant $X = i$ et $X = i'$ sont proches. Ceci est le cas si les variables X et Y sont indépendantes.

De manière symétrique, on introduit la distance du χ^2 sur les profils-colonnes.

$$d_{D_n^{-1}}^2(\mathbf{f}_{|j}, \mathbf{f}_{|j'}) = \|\mathbf{f}_{|j} - \mathbf{f}_{|j'}\|_{D_n^{-1}}^2 = (\mathbf{f}_{|j} - \mathbf{f}_{|j'})' D_n^{-1} (\mathbf{f}_{|j} - \mathbf{f}_{|j'}) = \sum_{i=1}^n \frac{1}{f_{i.}} (f_{i|j} - f_{i|j'})^2.$$

Un des avantages de la distance du χ^2 est de satisfaire au principe d'équivalence distributionnelle qui dit que si deux profils-lignes (colonnes) sont identiques, alors on peut regrouper dans le tableau de contingence les deux lignes (colonnes), en une ligne (colonne) formée de la somme des effectifs, sans changer la distance entre profils-lignes ou profils-colonnes.

Propriété 5.1. *La distance du χ^2 vérifie le principe d'équivalence distributionnelle.*

Démonstration. Nous montrons le résultat pour les profils-lignes, la démonstration étant symétrique pour les profils-colonnes.

Supposons que les profils lignes $\mathbf{f}_{|i_1}$ et $\mathbf{f}_{|i_2}$ soient identiques, c'est-à-dire que pour tout $j \in \{1, \dots, m\}$, $\frac{f_{i_1 j}}{f_{i_1.}} = \frac{f_{i_2 j}}{f_{i_2.}}$. Ceci implique que :

$$\frac{f_{i_1 j} + f_{i_2 j}}{f_{i_1.} + f_{i_2.}} = \frac{f_{i_1 j}}{f_{i_1.}} = \frac{f_{i_2 j}}{f_{i_2.}}.$$

Remplaçons dans le tableau de contingence les lignes i_1 et i_2 par une ligne i_0 formée de la somme des effectifs. Alors, la $j^{\text{ème}}$ coordonnée du profil-ligne $\mathbf{f}_{|i_0}$ de i_0 est :

$$\begin{aligned} \frac{k_{i_0 j}}{k_{i_0.}} &= \frac{k_{i_1 j} + k_{i_2 j}}{k_{i_1.} + k_{i_2.}}, \text{ par définition} \\ &= \frac{f_{i_0 j}}{f_{i_0.}} = \frac{f_{i_1 j} + f_{i_2 j}}{f_{i_1.} + f_{i_2.}}, \text{ en divisant le numérateur et dénominateur par } k.. \\ &= \frac{f_{i_1 j}}{f_{i_1.}} = \frac{f_{i_2 j}}{f_{i_2.}}, \end{aligned}$$

car les profils-lignes $\mathbf{f}_{|i_1}$ et $\mathbf{f}_{|i_2}$ soient identiques. Ainsi, le nouveau profil-ligne i_0 est le même que i_1 ou i_2 et les distances entre profils-lignes restent inchangées.

Montrons qu'il en est de même pour les distances entre les profils-colonnes j et j' . Il suffit de montrer que la différence des distances entre j et j' , avant et après intégration, est nulle. Retournant à la définition de la distance entre profils-colonnes, nous déduisons que cette différence est égale à,

$$\begin{aligned} & \frac{1}{f_{i_1 \cdot}} \left(\frac{f_{i_1 j}}{f_{\cdot j}} - \frac{f_{i_1 j'}}{f_{\cdot j'}} \right)^2 + \frac{1}{f_{i_2 \cdot}} \left(\frac{f_{i_2 j}}{f_{\cdot j}} - \frac{f_{i_2 j'}}{f_{\cdot j'}} \right)^2 - \frac{1}{f_{i_0 \cdot}} \left(\frac{f_{i_0 j}}{f_{\cdot j}} - \frac{f_{i_0 j'}}{f_{\cdot j'}} \right)^2 \\ &= f_{i_1 \cdot} \left(\frac{f_{i_1 j}}{f_{i_1 \cdot} f_{\cdot j}} - \frac{f_{i_1 j'}}{f_{i_1 \cdot} f_{\cdot j'}} \right)^2 + f_{i_2 \cdot} \left(\frac{f_{i_2 j}}{f_{i_2 \cdot} f_{\cdot j}} - \frac{f_{i_2 j'}}{f_{i_2 \cdot} f_{\cdot j'}} \right)^2 - f_{i_0 \cdot} \left(\frac{f_{i_0 j}}{f_{i_0 \cdot} f_{\cdot j}} - \frac{f_{i_0 j'}}{f_{i_0 \cdot} f_{\cdot j'}} \right)^2 \end{aligned}$$

Remarquons que les trois termes entre parenthèses sont les mêmes. Par convention, notons-le avec l'indice i_0 . Ainsi, cette différence s'écrit,

$$(f_{i_1 \cdot} + f_{i_2 \cdot} - f_{i_0 \cdot}) \left(\frac{f_{i_0 j}}{f_{i_0 \cdot} f_{\cdot j}} - \frac{f_{i_0 j'}}{f_{i_0 \cdot} f_{\cdot j'}} \right)^2.$$

Or par définition, $f_{i_0 \cdot} = \sum_{j=1}^m \frac{k_{i_0 j}}{k_{\cdot \cdot}} = \frac{k_{i_1 j} + k_{i_2 j}}{k_{\cdot \cdot}} = f_{i_1 \cdot} + f_{i_2 \cdot}$, ce qui termine la preuve. \square

Remarque. Dans la représentation graphique du nuage $N(I)$, si deux profils-lignes i_1 et i_2 sont identiques, ils sont représentés par le même point $\mathbf{f}_{|i_1} = \mathbf{f}_{|i_2}$ dans \mathbb{R}^m . Nous venons de voir que le point $\mathbf{f}_{|i_0}$ du profil-ligne i_0 obtenu par agrégation de i_1 et i_2 est aussi le même point de \mathbb{R}^m .

5.2.2 INERTIE DES NUAGES PROFILS-LIGNES ET COLONNES

L'inertie totale du nuage des profils-lignes $N(I)$, notée $\mathcal{J}_{\text{tot}}^{D_m^{-1}}$, est la moyenne pondérée des distances au centre de gravité $\bar{\mathbf{f}}^1$:

$$\mathcal{J}_{\text{tot}}^{D_m^{-1}} = \sum_{i=1}^n f_i \cdot d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \bar{\mathbf{f}}^1).$$

L'inertie totale du nuage de profils-lignes $N(J)$, notée $\mathcal{J}_{\text{tot}}^{D_n^{-1}}$, est défini de manière analogue.

Rappelons la définition de la statistique du χ^2 observée :

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

On a la proposition suivante.

Proposition 5.2. *L'inertie totale du nuage des profils-lignes $N(I)$ et celle des profils-colonnes $N(J)$ sont égales à $\frac{1}{N}$ fois le χ^2 :*

$$\mathcal{J}_{\text{tot}}^{D_m^{-1}} = \mathcal{J}_{\text{tot}}^{D_n^{-1}} = \frac{1}{N} \cdot \chi^2.$$

Démonstration. Démontrons cela pour l'inertie totale du nuage des profils-lignes. On a,

$$\begin{aligned}
 \mathcal{J}_{\text{tot}}^{D_m^{-1}} &= \sum_{i=1}^n f_{i\cdot} d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \bar{\mathbf{f}}^1) \\
 &= \sum_{i=1}^n f_{i\cdot} \sum_{j=1}^m \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 \\
 &= \sum_{i=1}^n \sum_{j=1}^m f_{i\cdot} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot}} \right)^2, \text{ en réduisant au même dénominateur} \\
 &= \sum_{i=1}^n \sum_{j=1}^m \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \frac{1}{N} \chi^2.
 \end{aligned}$$

La preuve est symétrique pour l'inertie du nuages des profils-colonnes. \square

Ainsi l'inertie totale mesure l'écart à l'indépendance des données. On pourra rejeter l'hypothèse nulle d'indépendance des variables X et Y si la valeur du χ^2 observée dépasse la valeur χ_0^2 qui a une probabilité d'être dépassée inférieure à un seuil fixé au préalable.

5.2.3 AJUSTEMENT DU NUAGE DES PROFILS-LIGNES ET COLONNES

L'ajustement du nuage se fait selon le même principe qu'en Analyse en Composantes Principales : on souhaite garder le maximum de dispersion entre les projections de toutes les paires de profils-lignes, sauf que la métrique choisie n'est pas l'identité mais la métrique D_m^{-1} .

D'après le théorème 3.2, il est équivalent de maximiser l'inertie totale de la projection du nuage des profils-lignes, et ceci revient à faire l'analyse par rapport à l'origine du nuage centré. Cependant, une des particularités de l'analyse de correspondances (due au fait que les points du nuage se trouvent dans un hyper-plan, voir remarque 5.1) est qu'il revient au même de procéder à l'analyse par rapport à l'origine si l'on ignore le premier axe factoriel correspondant à la valeur propre 1 (valeur propre 0 lorsqu'on effectue l'analyse par rapport au nuage centré), voir Figure 5.4.

Soit donc \mathcal{D} une droite qui passe par l'origine de vecteur directeur D_m^{-1} -unitaire \mathbf{a} . Notons π la projection D_m^{-1} -orthogonale sur cette droite. On cherche à maximiser :

$$\sum_{i=1}^n f_{i\cdot} d_{D_m^{-1}}^2(\mathbf{0}, \pi(\mathbf{f}_{|i})).$$

Lemme 5.1. *On a les égalités suivantes :*

$$\begin{aligned}
 \sum_{i=1}^n f_{i\cdot} d_{D_m^{-1}}^2(\mathbf{0}, \pi(\mathbf{f}_{|i})) &= \mathbf{a}' D_m^{-1} (P^1)' D_n P^1 D_m^{-1} \mathbf{a} = \mathbf{u}' (P^1)' D_n P^1 \mathbf{u} \\
 &= \mathbf{a}' D_m^{-1} F' D_n^{-1} F D_m^{-1} \mathbf{a} = \mathbf{u}' F' D_n^{-1} F \mathbf{u},
 \end{aligned}$$

où $\mathbf{u} = D_m^{-1} \mathbf{a}$.

Démonstration. La première égalité s'obtient de manière analogue au lemme 2.2 du chapitre 2. La deuxième égalité découle du fait que $P^1 = D_n^{-1} F$ (voir section 5.1.2). \square

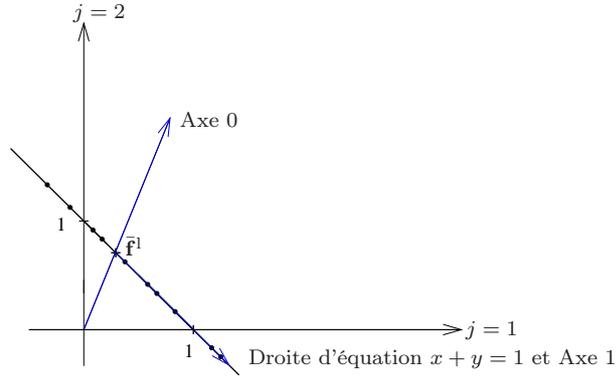


FIGURE 5.4 – Représentation lorsque $m = 2$. Le nuage des profils-lignes se trouvent sur la droite d'équation $x + y = 1$. Dans le cas où $m = 2$ il s'agit aussi d'un des axe principaux. L'axe 0 est l'autre axe principal, correspondant à la valeur propre 0 lorsque les données sont centrées, et à la valeur propre 1 lorsqu'elles ne le sont pas.

Remarque 5.2. On a,

$$\|\mathbf{a}\|_{D_m^{-1}}^2 = \mathbf{a}' D_m^{-1} \mathbf{a} = \mathbf{u}' D_m D_m^{-1} D_m \mathbf{u} = \|\mathbf{u}\|_{D_m}^2.$$

D'où, $\|\mathbf{a}\|_{D_m^{-1}}^2 = 1$ si et seulement si $\|\mathbf{u}\|_{D_m}^2 = 1$. Le fait de travailler avec \mathbf{u} plutôt que \mathbf{a} permet de simplifier certaines écritures dans la suite.

Ainsi, nous cherchons \mathbf{u}_1 solution du problème d'optimisation suivant :

$$\begin{cases} \mathbf{u}' F' D_n^{-1} F \mathbf{u} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\|_{D_m} = 1 \text{ est satisfaite.} \end{cases}$$

Ensuite, nous cherchons \mathbf{u}_2 , D_m -unitaire et D_m -orthogonal à \mathbf{u}_1 , tel que $\mathbf{u}_2' F' D_n^{-1} F \mathbf{u}_2$ soit maximal. En itérant les étapes, le problème devient :

Trouver un repère D_m -orthonormal $(\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_m)$ tel que $\mathbf{u}_1' F' D_n^{-1} F \mathbf{u}_1$ soit maximale, puis $\mathbf{u}_2' F' D_n^{-1} F \mathbf{u}_2$, etc.

(★)

On se retrouve dans le cadre de la section 1.8, avec $M = D_m$, et $A = F' D_n^{-1} F$. Mais cette fois-ci, on ne tient pas compte de la valeur propre 1 qui, on le rappelle, correspond à la valeur propre 0 du nuage centré. En utilisant le théorème de Perron-Frobenius, on peut montrer que la valeur propre 1 est la plus grande valeur propre de la matrice à diagonaliser. Ainsi, on ne va retenir que les valeurs propres non-nulles et *strictement inférieures* à 1. Notons q le nombre de telles valeurs propres, alors $q \leq \min\{m, n\} - 1$. Le théorème 1.4 nous dit :

Théorème 5.1. *La solution au problème d'optimisation sous contrainte (★) est le repère D_m -orthonormé $(\bar{\mathbf{f}}^1, \mathbf{u}_1, \dots, \mathbf{u}_q)$ formé des vecteurs propres de $D_m^{-1} F' D_n^{-1} F$, tel que \mathbf{u}_1 correspond à la plus grande valeur propre $\lambda_1 < 1$ de $D_m^{-1} F' D_n^{-1} F$, puis \mathbf{u}_2 à la deuxième plus grande valeur propre $\lambda_2 < 1$ de $D_m^{-1} F' D_n^{-1} F$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall k \in \{1, \dots, q\}, \quad \sum_{i=1}^n f_i d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \pi_k(\mathbf{f}_{|i})) = \mathbf{u}_k' F' D_n^{-1} F \mathbf{u}_k = \lambda_k.$$

Remarque 5.3. Le vecteur propre associé à la valeur propre $\lambda_0 = 1$ est $\mathbf{u}_0 = (1, \dots, 1)'$. L'axe correspondant est $\mathbf{a}_0 = D_m \mathbf{u}_0 = (f_{\cdot 1}, \dots, f_{\cdot m})' = \bar{\mathbf{f}}^1$, voir Figure 5.4.

Définition 5.1. Les vecteurs D_m -unitaires $\mathbf{u}_1, \dots, \mathbf{u}_q$ correspondant à des valeurs propres non nulles et différentes de 1 de $D_m^{-1} F' D_n^{-1} F$ sont les *axes factoriels*.

Les vecteurs directeurs $\mathbf{a}_1, \dots, \mathbf{a}_q$ des droites sur lesquelles on projette, associés au facteurs $\mathbf{u}_1, \dots, \mathbf{u}_q$ s'appellent les *axes principaux*, où on rappelle que, pour tout $k \in \{1, \dots, q\}$,

$$\mathbf{u}_k = D_m^{-1} \mathbf{a}_k.$$

Définition 5.2. Pour tout $i \in \{1, \dots, n\}$, pour tout $k \in \{1, \dots, q\}$, la coordonnée ψ_{ik} du $i^{\text{ème}}$ profil-ligne sur le $k^{\text{ème}}$ axe principal \mathbf{a}_k est :

$$\Psi_{ik} = \langle \mathbf{f}_{|i}, \mathbf{a}_k \rangle_{D_m^{-1}} = \mathbf{f}'_{|i} D_m^{-1} \mathbf{a}_k = \mathbf{f}'_{|i} \mathbf{u}_k = \sum_{j=1}^m \frac{f_{ij}}{f_{i\cdot}} u_{jk}.$$

Définissons $U = (\mathbf{u}_1, \dots, \mathbf{u}_q)$ la matrice dont les colonnes sont les facteurs. Alors, on peut écrire les coordonnées des profils-lignes dans le nouveau repère $(\mathbf{0}, \mathbf{a}_1, \dots, \mathbf{a}_q)$ dans la matrice Ψ :

$$\Psi = P^1 U = D_n^{-1} F U.$$

La $i^{\text{ème}}$ ligne $\Psi'_i = \mathbf{f}'_{|i} U$ représente les coordonnées du $i^{\text{ème}}$ profil-ligne sur les q axes principaux. La $k^{\text{ème}}$ colonne $\Psi^k = P^1 \mathbf{u}_k$ représente les coordonnées de tous les profil-lignes sur le $k^{\text{ème}}$ axe principal \mathbf{a}_k . On l'appelle la $k^{\text{ème}}$ *composante principale* du nuage des profils-lignes.

Pour le nuage des profils-colonnes, tout se passe de manière symétrique. On peut résumer les deux ajustements dans le tableau suivant :

Profils-lignes	← Nuage →	Profils-colonnes
Dans \mathbb{R}^m	Éléments de construction	Dans \mathbb{R}^n
$D_m^{-1} F' D_n^{-1} F$	Matrice à diagonaliser	$D_n^{-1} F D_m^{-1} F'$
\mathbf{u}_k, D_m -unitaire	Axes factoriels (vecteurs propres de la matrice)	\mathbf{v}_k, D_n -unitaire
$\mathbf{a}_k = D_m \mathbf{u}_k$	Axes principaux	$\mathbf{b}_k = D_n \mathbf{v}_k$
$\Psi^k = D_n^{-1} F \mathbf{u}_k$	Composantes principales	$\Phi^k = D_m^{-1} F' \mathbf{v}_k$

5.3 LIEN ENTRE LES AJUSTEMENTS DANS LES DEUX ESPACES : REPRÉSENTATION SIMULTANÉE

5.3.1 RELATION ENTRE LES AJUSTEMENTS DANS LES DEUX ESPACES

De manière analogue à ce que nous avons fait aux chapitres 2 et 4, nous pouvons relier les vecteurs propres et valeurs propres des matrices $D_m^{-1} F' D_n^{-1} F$ et $D_n^{-1} F D_m^{-1} F'$.

Soit \mathbf{u} un vecteur propre non nul de la matrice $D_m^{-1} F' D_n^{-1} F$ pour la valeur propre λ :

$$D_m^{-1} F' D_n^{-1} F \mathbf{u} = \lambda \mathbf{u}.$$

En pré-multipliant cette égalité par $D_n^{-1}F$, nous obtenons :

$$D_n^{-1}FD_m^{-1}F'(D_n^{-1}F\mathbf{u}) = \lambda(D_n^{-1}F\mathbf{u}).$$

En faisant l'argument symétrique pour l'autre matrice, nous déduisons que ces deux matrices ont les mêmes valeurs propres non-nulles et que \mathbf{u} est vecteur propre de $D_m^{-1}F'D_n^{-1}F$ si et seulement si $\mathbf{v} = D_n^{-1}F\mathbf{u}$ est vecteur propre de $D_n^{-1}FD_m^{-1}F'$.

Pour \mathbf{u}_k , D_m -normé à 1, on veut que \mathbf{v}_k soit D_n -normé à 1. Ainsi,

$$\mathbf{v}_k = \frac{D_n^{-1}F\mathbf{u}_k}{\|D_n^{-1}F\mathbf{u}_k\|_{D_n}}.$$

Or, $\|D_n^{-1}F\mathbf{u}_k\|_{D_n}^2 = \mathbf{u}_k'F'D_n^{-1}D_nD_n^{-1}F\mathbf{u}_k = \mathbf{u}_k'F'D_n^{-1}F\mathbf{u}_k = \lambda_k$, d'après le théorème 5.1, d'où $\mathbf{v}_k = \frac{D_n^{-1}F\mathbf{u}_k}{\sqrt{\lambda_k}}$. Par un argument symétrique, on a $\mathbf{u}_k = \frac{D_m^{-1}F'\mathbf{v}_k}{\sqrt{\lambda_k}}$. Nous avons démontré le résultat suivant :

Théorème 5.2. *Soient $\lambda_1 \geq \dots \geq \lambda_q > 0$ les valeurs propres non-nulles de $D_m^{-1}F'D_n^{-1}F$. Pour $k \in \{1, \dots, q\}$, soit \mathbf{u}_k (resp. \mathbf{v}_k) le vecteur propre D_m -unitaire (resp. D_n -unitaire) correspondant à la valeur propre λ_k de $D_m^{-1}F'D_n^{-1}F$, respectivement $D_n^{-1}FD_m^{-1}F'$. Alors, ces vecteurs propres sont liés par les relations :*

$$\begin{cases} \mathbf{v}_k = \frac{D_n^{-1}F\mathbf{u}_k}{\sqrt{\lambda_k}} = \frac{P^l\mathbf{u}_k}{\sqrt{\lambda_k}} = \frac{\Psi^k}{\sqrt{\lambda_k}} \\ \mathbf{u}_k = \frac{D_m^{-1}F'\mathbf{v}_k}{\sqrt{\lambda_k}} = \frac{(P^c)'\mathbf{v}_k}{\sqrt{\lambda_k}} = \frac{\Phi^k}{\sqrt{\lambda_k}}, \end{cases}$$

où on rappelle que P^l est la matrice des profils-lignes et P^c , celle des profils-colonnes.

On déduit les relations suivantes pour les coordonnées des projections :

$$\Psi^k = P^l\mathbf{u}_k = \sqrt{\lambda_k}\mathbf{v}_k, \quad (5.1)$$

$$\Phi^k = (P^c)'\mathbf{v}_k = \sqrt{\lambda_k}\mathbf{u}_k. \quad (5.2)$$

5.3.2 RELATIONS QUASI-BARYCENTRIQUES, REPRÉSENTATION SIMULTANÉE

En exprimant \mathbf{u}_k de (5.1) par l'expression du théorème et de manière analogue pour \mathbf{v}_k , on obtient :

$$\Psi^k = P^l \frac{\Phi^k}{\sqrt{\lambda_k}}, \quad \Phi^k = (P^c)' \frac{\Psi^k}{\sqrt{\lambda_k}}$$

Ce qui, en composante donne :

$$\begin{aligned} \psi_{ik} &= \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^m \frac{f_{ij}}{f_{i\cdot}} \phi_{jk} \\ \phi_{jk} &= \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^n \frac{f_{ij}}{f_{\cdot j}} \psi_{ik}. \end{aligned}$$

Ainsi, au coefficient de dilatation $\frac{1}{\sqrt{\lambda_k}}$ près, les projections des points d'un nuage sont, sur un axe, les barycentres des projections des points de l'autre nuage. Ceci justifie la représentation simultanée des deux nuages dans un même repère.

5.4 AIDE À L'INTERPRÉTATION

• D'après la proposition 5.2, on sait que N fois l'inertie totale du nuage des profils-lignes (ou profils-colonnes) est la statistique du χ^2 observée qui, sous l'hypothèse d'indépendance des données, suit asymptotiquement une loi du χ^2 à $(m-1)(n-1)$ degrés de liberté. L'inertie s'exprime aussi comme la somme des valeurs propres différentes de 1. Ainsi, la statistique observée :

$$\chi^2 = N \left(\sum_{k=1}^q \lambda_k \right)$$

permet d'effectuer un test du χ^2 pour évaluer l'indépendance des données. On pourra rejeter l'hypothèse d'indépendance des données si la valeur observée du χ^2 dépasse la valeur χ_0^2 qui a une probabilité d'être dépassée inférieure à un seuil fixé au préalable. L'AFC des données fait sens si l'hypothèse d'indépendance des données peut être rejetée.

L'inertie total du nuage de points mesure l'écart à l'indépendance des données, et chacun des axes principaux explique une part de l'écart à l'indépendance des données.

• Pour tout $k \in \{1, \dots, q\}$, le *pourcentage d'inertie expliquée* par les k premiers axes est :

$$F_k = \frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_q}.$$

Les nombres $(F_k)_{k=1}^q$ servent à déterminer le nombre d'axes à garder.

Le logiciel SPAD donne, pour chacun des profils-lignes i , $i \in \{1, \dots, n\}$:

- **Coordonnées dans le nouveau repère** : $(\psi_{ik} : k \in \{1, \dots, q\})$.
- **Distance au carré du profil-ligne i au centre de gravité** : $d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \bar{\mathbf{f}}^1)$.

$$d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \bar{\mathbf{f}}^1) = \sum_{j=1}^m \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i \cdot}} - f_{\cdot j} \right)^2.$$

On remarque que la distance s'annule lorsque le profil-ligne i est égal au profil moyen.

- **Contribution du profil-ligne i à l'inertie sur l'axe principal k** : $\text{CRT}_k(i)$.

$$\text{CRT}_k(i) = \frac{f_{i \cdot} d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \pi_k(\mathbf{f}_{|i}))}{\lambda_k} = \frac{f_{i \cdot} \psi_{ik}^2}{\lambda_k}.$$

On a $\sum_{i=1}^n \text{CRT}_k(i) = 1$ car $\sum_{i=1}^n f_{i \cdot} d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \pi_k(\mathbf{f}_{|i})) = \lambda_k$. Cette quantité permet de savoir dans quelle mesure le profil-ligne i contribue à l'inertie λ_k sur l'axe k .

- **Cosinus carré entre le vecteur $(\bar{\mathbf{f}}^1 \mathbf{f}_{|i})$ et la droite \mathcal{D}_k** : $\cos_k^2(i)$

$$\cos_k^2(i) = \frac{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \pi_k(\mathbf{f}_{|i}))}{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \mathbf{f}_{|i})} = \frac{\psi_{ik}^2}{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \mathbf{f}_{|i})}.$$

Un cosinus carré proche de 1 signifie que le profil-ligne i est proche de l'axe factoriel \mathcal{D}_k et par conséquent bien représenté par cet axe. Un cosinus carré proche de 0 signifie que le profil-ligne i est loin de l'axe seulement si la distance à l'origine est grande.

On obtient les quantités analogues pour le nuage des profils-colonnes.

5.5 CALCULS POUR LE THÉORÈME 5.1

La matrice $D_m^{-1}F'D_n^{-1}F$ n'est pas symétrique. On peut toutefois construire une matrice symétrique S qui a les mêmes valeurs propres. Pour tout $k \in \{1, \dots, q\}$, le vecteur \mathbf{u}_k est solution de :

$$D_m^{-1}F'D_n^{-1}F\mathbf{u}_k = \lambda_k\mathbf{u}_k.$$

En pré-multipliant cette équation par $(D_m)^{\frac{1}{2}}$, on obtient :

$$(D_m)^{-\frac{1}{2}}F'D_n^{-1}F(D_m)^{-\frac{1}{2}}[(D_m)^{\frac{1}{2}}\mathbf{u}_k] = \lambda_k[(D_m)^{\frac{1}{2}}\mathbf{u}_k],$$

La matrice $S = (D_m)^{-\frac{1}{2}}F'D_n^{-1}F(D_m)^{-\frac{1}{2}}$, de taille $m \times m$, est symétrique et a les mêmes valeurs propres non-nulles que $D_m^{-1}F'D_n^{-1}F$. Posons $W = D_n^{-\frac{1}{2}}F(D_m)^{-\frac{1}{2}}$ la matrice de travail. Alors,

$$W'W = S.$$

D'autre part \mathbf{u}_k est vecteur propre de $D_m^{-1}F'D_n^{-1}F$ si et seulement si $\tilde{\mathbf{u}}_k := (D_m)^{\frac{1}{2}}\mathbf{u}_k$ est vecteur propre de S . De plus,

$$\|\mathbf{u}_k\|_{D_m}^2 = \mathbf{u}_k'D_m\mathbf{u}_k = ((D_m)^{\frac{1}{2}}\mathbf{u}_k)'((D_m)^{\frac{1}{2}}\mathbf{u}_k) = \|(D_m)^{\frac{1}{2}}\mathbf{u}_k\|^2 = \|\tilde{\mathbf{u}}_k\|^2.$$

Ainsi, le vecteur \mathbf{u}_k est D_m -normé à 1 si et seulement si le vecteur $\tilde{\mathbf{u}}_k$ est normé à 1 pour la métrique usuelle.

En pratique, on procède donc comme suit.

1. On détermine la matrice de travail $W = D_n^{-\frac{1}{2}}F(D_m)^{-\frac{1}{2}}$. C'est une matrice de taille $n \times m$ dont les coefficients sont :

$$w_{ij} = \frac{f_{ij}}{\sqrt{f_{i\cdot}f_{\cdot j}}}.$$

2. On calcule la matrice symétrique $S = W'W$.
3. On détermine les valeurs propres $\lambda_0, \dots, \lambda_q$ non-nulles de S et les vecteurs propres unitaires $\tilde{\mathbf{u}}_0, \dots, \tilde{\mathbf{u}}_q$ correspondants. Les valeurs propres de S sont celles de la matrice $D_m^{-1}F'D_n^{-1}F$. La plus grande valeur propre sera $\lambda_0 = 1$, que l'on exclut. En utilisant la remarque 5.3, on déduit que le vecteur propre $\tilde{\mathbf{u}}_0$ associé à la valeur propre λ_0 est $\tilde{\mathbf{u}}_0 = (\sqrt{f_{\cdot 1}}, \dots, \sqrt{f_{\cdot m}})'$.

4. On retrouve les axes factoriels en posant $\mathbf{u}_k = D_m^{-\frac{1}{2}}\tilde{\mathbf{u}}_k$.

5.6 EXERCICE

Exercice 5.1. On considère l'exemple déjà évoqué de l'élection présidentielle de 1981.

$X \backslash Y$	Mitterrand	Giscard	Chirac	Marchais
Ain	51	64	36	23
Hautes-Alpes	14	17	9	9
Ariège	27	18	13	17
Bouche du Rhône	191	204	119	205
Charente-Maritime	71	76	47	37
Drôme	57	55	31	30

Le candidat Marchais et le département Drôme sont supposés illustratifs. Soit donc le tableau de contingence des effectifs des éléments actifs.

$X \backslash Y$	Mitterrand	Giscard	Chirac	Total
Ain	51	64	36	151
Hautes-Alpes	14	17	9	40
Ariège	27	18	13	58
Bouche du Rhône	191	204	119	514
Charente-Maritime	71	76	47	194
Total	354	379	224	957

1. Donner le tableau de contingence des fréquences.
2. Construire le tableau des profils-lignes et profils-colonnes.
3. Quel est le point moyen du nuage des profils-ligne $N(I)$ et du nuage des profils colonnes $N(J)$?
4. Calculer la distance du χ^2 entre les profils-ligne de l'Ariège et de la Charente-Maritime.
5. Déterminer la matrice de travail W et la matrice à diagonaliser S .
6. Quelle est la valeur propre triviale de S et le vecteur propre associé.
7. Soit $\lambda_1 = 0.0035269$, la première valeur propre non-triviale de S et

$$\tilde{\mathbf{u}}_1 = (0.750930, -0.653700, -0.093711)'$$

le vecteur propre associé. Soit,

$$\tilde{\mathbf{u}}_2 = (-0.25728, -0.42030, 0.87015)'$$

le vecteur propre associé à la troisième valeur λ_2 . En déduire λ_2 .

8. Calculer les axes factoriels \mathbf{u}_1 et \mathbf{u}_2 ; puis les coordonnées des départements sur les axes principaux \mathbf{a}_1 et \mathbf{a}_2 , ainsi que les coordonnées de l'élément illustratif "la Drôme".
9. Calculer les coordonnées des candidats sur les axes principaux \mathbf{b}_1 et \mathbf{b}_2 , ainsi que les coordonnées de l'élément illustratif "Marchais".
10. Effectuer l'AFC de ces données avec SPAD. Expliquer toutes les issues non encore expliquées.

Solution de l'exercice 5.1.

1. Tableau de contingence des fréquences.

$X \backslash Y$	Mitterrand	Giscard	Chirac	Total
Ain	51/957	64/957	36/957	151/957
Hautes-Alpes	14/957	17/957	9/957	40/957
Ariège	27/957	18/957	13/957	58/957
Bouche du Rhône	191/957	204/957	119/957	514/957
Charente-Maritime	71/957	76/957	47/957	194/957
Total	354/957	379/957	224/957	957/957

Spad. Les fréquences marginales ($f_{i.}$) et ($f_{.j}$), qui se trouvent dans la dernière colonne et la dernière ligne du tableau de contingence des fréquences, sont les poids relatifs des variables et des individus de Corbi-2 et Corbi-4 du fichier .xls de spad.

2. Tableau des profils-ligne

$X \backslash Y_{ X}$	Mitterrand	Giscard	Chirac	Total
Ain	$\frac{51}{151}$	$\frac{64}{151}$	$\frac{36}{151}$	1
Hautes-Alpes	$\frac{14}{40}$	$\frac{17}{40}$	$\frac{9}{40}$	1
Ariège	$\frac{27}{58}$	$\frac{18}{58}$	$\frac{13}{58}$	1
Bouche du Rhone	$\frac{191}{514}$	$\frac{204}{514}$	$\frac{119}{514}$	1
Charente Maritime	$\frac{71}{194}$	$\frac{76}{194}$	$\frac{47}{194}$	1

Tableau des profils-colonnes

$X_{ Y} \backslash Y$	Mitterrand	Giscard	Chirac
Ain	51/354	64/379	36/224
Hautes-Alpes	14/354	17/379	9/224
Ariège	27/354	18/379	13/224
Bouche du Rhone	191/354	204/379	119/224
Charente Maritime	71/354	76/379	47/224
Total	1	1	1

3. D'après le cours, le centre de gravité du nuage des profils-ligne est le vecteur formé des fréquences marginales ($f_{.j}$).

$$\bar{\mathbf{f}}^l = \left(\frac{354}{957}, \frac{379}{957}, \frac{224}{957} \right)'$$

D'après le cours, le centre de gravité du nuage des profils-colonne est le vecteur formé des fréquences marginales ($f_{i.}$).

$$\bar{\mathbf{f}}^c = \left(\frac{151}{957}, \frac{40}{957}, \frac{58}{957}, \frac{514}{957}, \frac{194}{957} \right)'$$

4. D'après le cours,

$$d_{D_m^{-1}}^2(\mathbf{f}_{|i}, \mathbf{f}_{|i'}) = \sum_{j=1}^m \frac{1}{f_{\cdot j}} (f_{j|i} - f_{j|i'})^2.$$

Pour le profil de l'Ariège : $(\frac{27}{58}, \frac{18}{58}, \frac{13}{58})$, et de la Charente-Maritime : $(\frac{71}{194}, \frac{76}{194}, \frac{47}{194})$, on obtient :

$$d_{D_m^{-1}}^2(\text{Ar}, \text{Ch-M}) = \frac{957}{354} \left(\frac{27}{58} - \frac{71}{194} \right)^2 + \frac{957}{379} \left(\frac{18}{58} - \frac{76}{194} \right)^2 + \frac{957}{224} \left(\frac{13}{58} - \frac{47}{194} \right)^2 \simeq 0,0449.$$

5. La matrice de travail $W = (w_{ij})$ est la matrice de taille $n \times m$, dont les coefficients sont :

$$w_{ij} = \frac{f_{ij}}{\sqrt{f_{i\cdot} \cdot f_{\cdot j}}} = \frac{k_{ij}}{\sqrt{k_{i\cdot} \cdot k_{\cdot j}}}.$$

$$W = \begin{pmatrix} \frac{51}{\sqrt{151 \times 354}} & \frac{64}{\sqrt{151 \times 379}} & \frac{36}{\sqrt{151 \times 224}} \\ \frac{14}{\sqrt{40 \times 354}} & \frac{17}{\sqrt{40 \times 379}} & \frac{9}{\sqrt{40 \times 244}} \\ \frac{27}{\sqrt{58 \times 354}} & \frac{18}{\sqrt{58 \times 379}} & \frac{13}{\sqrt{58 \times 244}} \\ \frac{191}{\sqrt{514 \times 354}} & \frac{204}{\sqrt{514 \times 379}} & \frac{119}{\sqrt{514 \times 244}} \\ \frac{71}{\sqrt{194 \times 354}} & \frac{76}{\sqrt{194 \times 379}} & \frac{47}{\sqrt{194 \times 244}} \end{pmatrix}.$$

La matrice à diagonaliser S est

$$S = W'W = \begin{pmatrix} 0,37 & 0,38 & 0,29 \\ 0,38 & 0,40 & 0,30 \\ 0,29 & 0,30 & 0,23 \end{pmatrix}.$$

6. D'après le cours, la valeur propre triviale de S est la valeur propre $\lambda_0 = 1$ et le vecteur propre associé est

$$\tilde{\mathbf{u}}_0 = (\sqrt{f_{\cdot 1}}, \sqrt{f_{\cdot 2}}, \sqrt{f_{\cdot 3}})' = \left(\sqrt{\frac{354}{957}}, \sqrt{\frac{379}{957}}, \sqrt{\frac{224}{957}} \right)'$$

7. D'après un résultat du cours, $\text{Tr}(S) = \lambda_0 + \lambda_1 + \lambda_2$; d'où

$$\lambda_2 = \text{Tr}(S) - (\lambda_0 + \lambda_1) \simeq 1,0036530 - (1 + 0,0035269) \simeq 0,0001261$$

8. Remarquons que les vecteurs propres $\tilde{\mathbf{u}}_1$ et $\tilde{\mathbf{u}}_2$ sont normés à 1 pour la métrique usuelle. Ainsi, pour $k = 1, 2$, les facteurs principaux \mathbf{u}_k sont donnés par : $\mathbf{u}_k = (D_m)^{-\frac{1}{2}} \tilde{\mathbf{u}}_k$, avec

$$(D_m)^{-\frac{1}{2}} = \begin{pmatrix} \sqrt{\frac{954}{354}} & 0 & 0 \\ 0 & \sqrt{\frac{954}{379}} & 0 \\ 0 & 0 & \sqrt{\frac{954}{224}} \end{pmatrix}.$$

$$\mathbf{u}_1 = \begin{pmatrix} 1,235 \\ -1,039 \\ -0,194 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0,423 \\ -0,668 \\ 1,799 \end{pmatrix}.$$

D'après le cours, les coordonnées des départements (profils-lignes) sur les axes principaux $\mathbf{a}_1, \mathbf{a}_2$ sont enregistrées dans la matrice $\Psi = P^l U$, où $U = (\mathbf{u}_1, \mathbf{u}_2)$ est la matrice dont les colonnes sont les facteurs $\mathbf{u}_1, \mathbf{u}_2$, et $P^l = D_n^{-1} F$ est la matrice des profils lignes (le tableau des profils-lignes sans la dernière colonne).

$$P^l = D_n^{-1} F = \begin{pmatrix} \frac{51}{151} & \frac{64}{151} & \frac{36}{151} \\ \frac{14}{40} & \frac{17}{40} & \frac{9}{40} \\ \frac{27}{58} & \frac{18}{58} & \frac{13}{58} \\ \frac{191}{514} & \frac{204}{514} & \frac{119}{514} \\ \frac{71}{194} & \frac{76}{194} & \frac{47}{194} \end{pmatrix} = \begin{pmatrix} 0,34 & 0,42 & 0,24 \\ 0,35 & 0,43 & 0,23 \\ 0,47 & 0,31 & 0,22 \\ 0,37 & 0,40 & 0,23 \\ 0,37 & 0,39 & 0,24 \end{pmatrix}.$$

$$U = \begin{pmatrix} 1,235 & -0,423 \\ -1,039 & -0,668 \\ -0,194 & 1,799 \end{pmatrix}.$$

$$\Psi = P^l U = \begin{pmatrix} 0,34 & 0,42 & 0,24 \\ 0,35 & 0,43 & 0,23 \\ 0,47 & 0,31 & 0,22 \\ 0,37 & 0,40 & 0,23 \\ 0,37 & 0,39 & 0,24 \end{pmatrix} \begin{pmatrix} 1,235 & -0,423 \\ -1,039 & -0,668 \\ -0,194 & 1,799 \end{pmatrix} = \begin{pmatrix} -0,069 & 0,003 \\ -0,053 & -0,027 \\ 0,209 & -0,001 \\ 0,002 & -0,006 \\ -0,002 & 0,019 \end{pmatrix}.$$

Spad. La matrice Ψ se retrouve dans les deux dernières colonnes du premier tableau de l'onglet Corbi-4 du fichier .xls de Spad.

Les coordonnées Ψ'_6 de l'élément illustratif la Drôme sont données par :

$$\Psi'_6 = \begin{pmatrix} \frac{57}{143} & \frac{55}{143} & \frac{31}{143} \end{pmatrix} U = \begin{pmatrix} 0,051 & -0,036 \end{pmatrix}.$$

Spad. Les coordonnées de l'élément illustratif la Drôme se trouvent dans les deux dernières colonnes du premier tableau de l'onglet Corbi-5 du fichier .xls de Spad.

9. Les coordonnées des candidats (profils-colonnes) sur les axes principaux $\mathbf{b}_1, \mathbf{b}_2$ sont enregistrées dans la matrice $\Phi = P^c V$, où $V = (\mathbf{v}_1, \mathbf{v}_2)$ est la matrice dont les colonnes sont les facteurs $\mathbf{v}_1, \mathbf{v}_2$. Pour calculer ces coordonnées, on utilise plutôt les formules de transition : Pour $k = 1, 2$, $\Phi^k = \sqrt{\lambda_k} \mathbf{u}_k$. D'où

$$\Phi = \left(\sqrt{\lambda_1} \mathbf{u}_1 \quad \sqrt{\lambda_2} \mathbf{u}_2 \right) = \begin{pmatrix} 0,073 & -0,005 \\ -0,062 & -0,007 \\ -0,012 & 0,020 \end{pmatrix}.$$

Spad. La matrice Φ se retrouve dans les deux dernières colonnes du premier tableau de l'onglet Corbi-2 du fichier .xls de Spad.

Pour calculer les coordonnées de l'individu illustratif "Marchais" sur les axes principaux,

on utilise la formule de transition :

$$\begin{aligned} \Phi'_4 &= \begin{pmatrix} \frac{23}{143} & \frac{9}{291} & \frac{17}{291} & \frac{205}{291} & \frac{37}{291} \end{pmatrix} V = \begin{pmatrix} \frac{23}{143} & \frac{9}{291} & \frac{17}{291} & \frac{205}{291} & \frac{37}{349} \end{pmatrix} \begin{pmatrix} \frac{\Psi^1}{\sqrt{\lambda_1}} & \frac{\Psi^2}{\sqrt{\lambda_2}} \end{pmatrix} \\ &= \begin{pmatrix} 0,0790 & 0,0309 & 0,0584 & 0,7045 & 0,1271 \end{pmatrix} \begin{pmatrix} -1,1692 & 0,2538 \\ -0,8910 & -2,4246 \\ 3,5188 & -0,0951 \\ 0,0284 & -0,5222 \\ -0,0336 & 1,7166 \end{pmatrix} \\ &= \begin{pmatrix} 0,1013 & -0,2101 \end{pmatrix}. \end{aligned}$$

10. Avec SPAD, nous obtenons le plan factoriel des données.

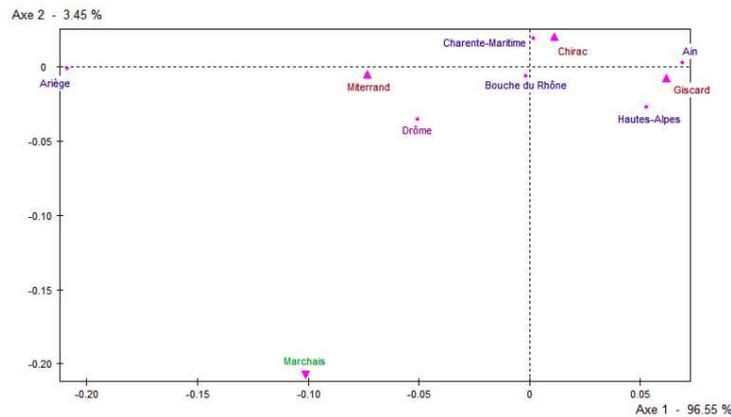


FIGURE 5.5 – Plan factoriel des données obtenu avec SPAD.

Après l'analyse, les seules données qu'il nous reste à expliquer sont les distances carrées au centre de gravité, les contributions des profils à l'inertie des axes, les cosinus carrés. Nous résumons ces données dans le cas des profils-lignes dans le tableau ci-dessous, et les calculons pour le premier profil-ligne.

Individus	Distance carré centre gravité	CRT		cos ²	
		1	2	1	2
Ain	0,005	21,570	1,073	0,998	0,002
Hautes-Alpes	0,004	3,321	24,447	0,791	0,208
Ariège	0,044	75,042	0,067	1,000	0,000
Bouche du Rhône	0,000	0,043	14,495	0,076	0,914
Charente Maritime	0,000	0,023	59,931	0,010	0,993

- Distance au carré du profil-ligne 1 au centre de gravité : $d_{D_m^{-1}}^2(\mathbf{f}_1, \bar{\mathbf{f}}^1)$.

On peut montrer que cette distance est égale à la distance *usuelle* de l'origine au point défini par les coordonnées dans le nouveau repère du profil-ligne :

$$d_{D_m^{-1}}^2(\mathbf{f}_{|1}, \bar{\mathbf{f}}^1) = \psi_{11}^2 + \psi_{12}^2 \simeq (-0,069)^2 + (0,003)^2 \simeq 0,005.$$

- Contribution du profil-ligne 1 à l'inertie sur les axes principaux $(\text{CRT}_1(1), \text{CRT}_2(1))$.

$$\begin{aligned} (\text{CRT}_1(1), \text{CRT}_2(1)) &= \left(\frac{f_1 \cdot \psi_{11}^2}{\lambda_1}, \frac{f_1 \cdot \psi_{12}^2}{\lambda_2} \right) \\ &\simeq \left(\frac{15,779(0,069)^2}{0,0035269}, \frac{15,779(0,003)^2}{0,0001261} \right) \simeq (21,570, 1,073). \end{aligned}$$

- Cosinus carré entre le vecteur $(\bar{\mathbf{f}}^1 \mathbf{f}_{|1})$ et le premier axe principal.

$$\cos_1^2(1) = \frac{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \pi_k(\mathbf{f}_{|i}))}{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \mathbf{f}_{|i})} = \frac{\psi_{ik}^2}{d_{D_m^{-1}}^2(\bar{\mathbf{f}}^1, \mathbf{f}_{|i})} \simeq \frac{0,069^2}{0,005} \simeq 0,9983$$

Analyse.

En regardant les contributions relatives et les \cos^2 , on déduit que la construction de l'axe 1 provient essentiellement des départements Ariège (3) et Ain (1) qu'il oppose, et la construction de l'axe 2 provient des trois autres départements. Il faut cependant faire attention à l'interprétation des Hautes-Alpes dont le \cos^2 montre que la projection n'est pas représentative.

CHAPITRE 6

MÉTHODES DE CLASSIFICATION

Ce chapitre concerne les *méthodes de classification* ou *cluster analysis* en anglais. Le but est de regrouper les individus en classes qui sont le plus homogène possible.

La première idée pour trouver la meilleure partition serait de fixer un critère d'optimalité, puis de parcourir toutes les partitions possibles, de calculer le critère et de déterminer laquelle des partitions est la meilleure. Ceci n'est pas réaliste car le nombre de partitions devient vite gigantesque. Par exemple, pour 20 individus, le nombre de partition est 51 724 158 235 372! On a donc recourt à d'autres procédés.

Il y a deux grands types de méthodes de classification :

- *Classifications hiérarchiques*. A chaque instant, on a une décomposition de l'espace des individus en classes disjointes. Ces méthodes peuvent être ascendantes ou descendantes. Nous traiterons les premières (CAH). Dans ce cas, au début, chaque individu forme une classe à lui tout seul. Puis, à chaque étape, les deux classes les plus "proches" sont fusionnées. A la dernière étape, il ne reste plus qu'une seule classe regroupant tous les individus.
- *Classifications non-hiérarchiques*. Dans ce cas, le nombre de classes est fixé à l'avance. Il s'agit essentiellement des techniques d'agrégation autour des centres mobiles. Ces méthodes sont particulièrement intéressantes dans le cas de grands tableaux car elles sont peu coûteuses en temps de calcul et en espace de mémoire.

Dans ces méthodes, les individus sont regroupés dans des classes *homogènes*. Ceci signifie que les individus d'une même classe sont *proches*. Afin de mesurer la proximité entre individus, nous avons besoin d'un choix de distance, objet de la section suivante.

On considère un échantillon formé de n individus sur lequel on étudie p variables. Les données sont représentées dans une matrice X de taille $n \times p$.

6.1 DISTANCES

Voici les distances les plus utilisées. Soit $\mathbf{x}_i, \mathbf{x}_{i'}$ une paire d'individus.

1. La distance usuelle : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\| = \sqrt{\sum_{j=1}^p (x_{ij} - x_{i'j})^2}$.
2. La distance de Manhattan : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^p |x_{ij} - x_{i'j}|$.
3. La distance de Mahalanobis : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_{V^{-1}}$, où V est la matrice des variances.

Remarque. Si les unités des variables ne sont pas comparables, on pourra aussi considérer les données centrées-réduites. Une autre alternative est de prendre le tableau issu de l'ACP.

Par définition, une distance satisfait aux conditions suivantes :

1. Symétrie : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = d(\mathbf{x}_{i'}, \mathbf{x}_i)$.
2. Positivité : $d(\mathbf{x}_i, \mathbf{x}_{i'}) \geq 0$.
3. $d(\mathbf{x}_i, \mathbf{x}_{i'}) = 0 \iff \mathbf{x}_i = \mathbf{x}_{i'}$.
4. Inégalité triangulaire : $d(\mathbf{x}_i, \mathbf{x}_{i'}) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_{i'})$.

Dans certain cas, il est utile de relaxer la dernière hypothèse, on parlera alors de *dissimilarité*. Par exemple, une question importante en biologie est celle de la classification des espèces. Les n individus sont alors décrits par la présence (1) ou absence (0) de p caractéristiques. Il s'agit d'un *tableau disjonctif complet*. Les distances ci-dessus ne sont pas adéquates, et on préférera les dissimilarités suivantes.

On enregistre les quantités :

- $a_{ii'}$ = nombre de caractéristiques communes aux individus \mathbf{x}_i et $\mathbf{x}_{i'}$.
- $b_{ii'}$ = nombre de caractéristiques possédées par \mathbf{x}_i , mais pas par $\mathbf{x}_{i'}$.
- $c_{ii'}$ = nombre de caractéristiques possédées par $\mathbf{x}_{i'}$, mais pas par \mathbf{x}_i .
- $d_{ii'}$ = nombre de caractéristiques possédées ni par \mathbf{x}_i , ni par $\mathbf{x}_{i'}$.

On a alors les définitions de dissimilarités suivantes :

1. Jaccard : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = 1 - \frac{a_{ii'}}{a_{ii'} + b_{ii'} + c_{ii'}}$.
2. Russel et Rao : $d(\mathbf{x}_i, \mathbf{x}_{i'}) = 1 - \frac{a_{ii'}}{a_{ii'} + b_{ii'} + c_{ii'} + d_{ii'}}$.

EXEMPLE. On a 5 individus, et les variables sont :

1. Var 1 : Présence / absence d'ailes.
2. Var 2 : Présence / absence de pattes.
3. Var 3 : Présence / absence de bec.

Les données sont : $X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. Ainsi : $a_{12} = 1$, $b_{12} = 1$, $c_{12} = 1$, $d_{12} = 0$, et la distance

de Jaccard est : $d(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{1}{3} = \frac{2}{3}$.

Les distances ou dissimilarités entre toutes les paires d'individus sont répertoriées dans une matrice $D = (d_{ii'})$ de taille $n \times n$, appelée *matrice des distances* :

$$d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'}).$$

Remarquer que seuls $\frac{n(n-1)}{2}$ termes sont significatifs car la matrice est symétrique et les termes de la diagonale sont nuls.

6.2 MÉTHODE DE CLASSIFICATION HIÉRARCHIQUE (CAH)

Le principe consiste à créer, à chaque étape, une partition obtenue en agréant deux-à-deux les éléments les plus proches. L'algorithme fournit une hiérarchie de partitions, se présentant sous la forme d'un *arbre/dendogramme*. L'intérêt de cet arbre est qu'ils peut donner une idée du nombre de classes existant effectivement dans la population.

6.2.1 ALGORITHME

Pour l'instant, on n'a qu'une notion de distance/dissimilarité entre individus. On suppose que l'on a aussi une notion de distance entre *classes*, que l'on précisera ultérieurement. Voici l'algorithme de classification hiérarchique.

1. *Initialisation*. Partition en n classes C_1, \dots, C_n , où chaque individu représente une classe. On suppose donné la matrice des distances entre individus.
2. *Etape k* . ($k = 0, \dots, n - 1$). Les données sont $n - k$ classes C_1, \dots, C_{n-k} , et la matrice des distances entre les différentes classes. Pour passer de k à $k + 1$:
 - Trouver dans la matrice des distances la plus petite distance entre deux classes. Regrouper les deux classes correspondantes. Obtenir ainsi $n - k - 1$ nouvelles classes, C_1, \dots, C_{n-k-1} .
 - Recalculer la matrice des distances qui donnent les $\frac{(n-k)(n-k-1)}{2}$ distances entre les nouvelles classes.
 - Poser $k := k + 1$.

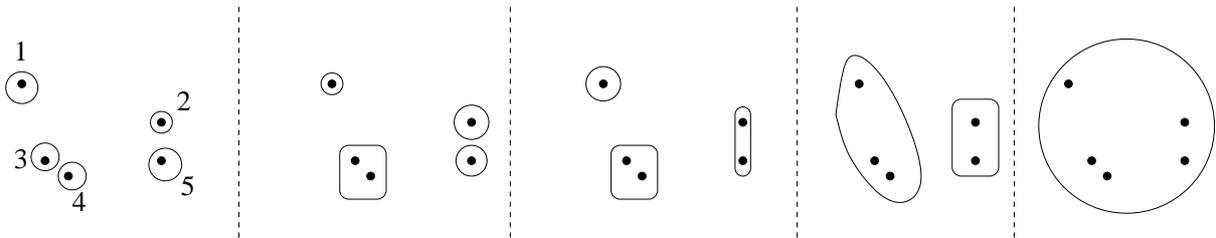


FIGURE 6.1 – Illustration de la CAH.

Le résultat de l'algorithme est représenté sous forme d'un *arbre*, aussi appelé *dendogramme*. La hauteur des branches représente la distance entre les deux éléments regroupés.

Les avantages et inconvénients de l'algorithme sont :

- *Avantages*. Algorithme simple, permettant une très bonne lecture des données.
- *Inconvénients*.
 1. Selon la définition de distance entre les classes, on trouve des résultats très différents. Une idée est donc d'appliquer la méthode avec différentes distances, et de trouver les groupes stables.

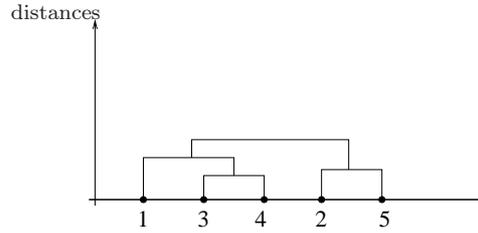


FIGURE 6.2 – Exemple d’arbre/dendrogramme issu d’une CAH.

2. La complexité de l’algorithme est en $O(n^3)$, ainsi même sur un nombre de données petit, on arrive rapidement à saturation de la puissance d’un ordinateur. En effet, l’algorithme est constitué de n étapes, et à chaque fois il faut parcourir la matrice des distances qui est de taille $\frac{(n-k)(n-k-1)}{2}$.

Voici deux définitions naturelles de distances entre classes. Soient C et C' deux classes.

La *distance du saut minimum* entre les classes C et C' , notée $d_{\min}(C, C')$, est par définition :

$$d_{\min}(C, C') = \min_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'} d(\mathbf{x}_i, \mathbf{x}_{i'}).$$

C’est la plus petite distance entre éléments des deux classes.

La *distance du saut maximum* entre les classes C et C' , notée $d_{\max}(C, C')$, est par définition :

$$d_{\max}(C, C') = \max_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'} d(\mathbf{x}_i, \mathbf{x}_{i'}).$$

C’est la plus grande distance entre éléments des deux classes.

La troisième distance entre classes est celle de *Ward*. Elle utilise la notion d’inertie vue au chapitre 3 et nous lui consacrons la section suivante.

6.2.2 DISTANCE DE WARD ENTRE CLASSES

On suppose que les individus sont attribués des poids relatifs $(p_i)_{i=1}^n$, avec $\sum_{i=1}^n p_i = 1$, et que la distance entre individus est la distance euclidienne usuelle.

RAPPELS DU CHAPITRE 3

L’inertie totale du nuage de points est :

$$J_{\text{tot}} = \sum_{i=1}^n p_i d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^n p_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2.$$

Supposons de plus que les individus soient regroupés en k classes C_1, \dots, C_k . Pour tout $\ell \in \{1, \dots, k\}$:

- $p_{(\ell)}$ est le poids relatif de la classe C_ℓ :

$$p_{(\ell)} = \sum_{i \in I_\ell} p_i. \quad \text{Ainsi,} \quad \sum_{\ell=1}^k p_{(\ell)} = 1.$$

- $\bar{\mathbf{x}}_{(\ell)}$ est le centre de gravité ou barycentre de la classe C_ℓ :

$$\bar{\mathbf{x}}_{(\ell)} = (\bar{x}_{(\ell)}^1, \dots, \bar{x}_{(\ell)}^p)',$$

où pour tout $j \in \{1, \dots, p\}$,

$$\bar{x}_{(\ell)}^j = \frac{1}{p_{(\ell)}} \sum_{i \in I_\ell} p_i x_{ij}.$$

L'inertie inter-classes du nuage de points est :

$$J_{\text{inter}} = \sum_{\ell=1}^k p_{(\ell)} d^2(\bar{\mathbf{x}}_{(\ell)}, \bar{\mathbf{x}}) = \sum_{\ell=1}^k p_{(\ell)} \|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}\|^2.$$

Nous avons aussi vu la notion d'inertie intra-classes et le théorème de Huygens, qui dit que :

$$J_{\text{tot}} = J_{\text{inter}} + J_{\text{intra}}.$$

DISTANCE DE WARD ENTRE CLASSES

Une classe est homogène si son inertie est faible, ainsi pour l'algorithme de classification, nous souhaitons avoir une inertie intra-classe qui soit faible, ou de manière équivalente, une inertie inter-classe qui soit élevée (vu que l'inertie totale est constante d'après le théorème de Huygens).

Quand on fusionne deux classes C_ℓ et $C_{\ell'}$ pour obtenir une nouvelle classe $C_\ell \cup C_{\ell'}$, l'inertie inter-classe diminue. On note $\bar{\mathbf{x}}_{(\ell \cup \ell')}$ le centre de gravité de la classe $C_\ell \cup C_{\ell'}$:

$$\bar{\mathbf{x}}_{(\ell \cup \ell')} = \frac{p_{(\ell)} \bar{\mathbf{x}}_{(\ell)} + p_{(\ell')} \bar{\mathbf{x}}_{(\ell')}}{p_{(\ell)} + p_{(\ell')}}.$$

Ce nouveau centre de gravité est naturellement affecté du poids relatif $p_{(\ell)} + p_{(\ell')}$.

On définit la *distance de Ward*, notée $d_{\text{ward}}(C_\ell, C_{\ell'})$ entre les classes C_ℓ et $C_{\ell'}$, comme étant la perte de l'inertie inter-classe :

$$d_{\text{ward}}(C_\ell, C_{\ell'}) = p_{(\ell)} \|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}\|^2 + p_{(\ell')} \|\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}\|^2 - (p_{(\ell)} + p_{(\ell')}) \|\bar{\mathbf{x}}_{(\ell \cup \ell')} - \bar{\mathbf{x}}\|^2.$$

Ainsi, à chaque étape de l'algorithme, on fusionne les deux classes telles que la perte d'inertie inter-classe soit minimale.

Remarque 6.1.

- Au début de l'algorithme, on a une partition en n classes, où chaque classe représente un seul individu. Chacun des individus est le centre de gravité de sa propre classe et l'inertie inter-classe est l'inertie totale du nuage de points.
- A la fin de l'algorithme, on a une seule classe regroupant tous les individus. Le centre de gravité de la classe est le centre de gravité du nuage et l'inertie inter-classe est donc nulle.
- En conséquence, la somme des niveaux d'agrégation des différents niveaux de l'arbre, qui est la somme des pertes d'inerties inter-classes, est égale à l'inertie totale du nuage de points.

Il est plus aisé de calculer la distance de Ward entre classes à l'aide de la proposition suivante.

Proposition 6.1. *La distance de Ward est égale à :*

$$d_{\text{ward}}(C_\ell, C_{\ell'}) = \frac{p(\ell)p(\ell')}{p(\ell) + p(\ell')} \|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}\|^2.$$

Démonstration. D'après le théorème de Huygens on a :

$$p(\ell)\|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}\|^2 + p(\ell')\|\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}\|^2 = p(\ell)\|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2 + p(\ell')\|\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2 + (p(\ell) + p(\ell'))\|\bar{\mathbf{x}}_{(\ell \cup \ell')} - \bar{\mathbf{x}}\|^2.$$

Ainsi, en utilisant la définition de la distance de Ward on obtient,

$$d_{\text{ward}}(C_\ell, C_{\ell'}) = p(\ell)\|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2 + p(\ell')\|\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2.$$

Or,

$$\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell \cup \ell')} = \bar{\mathbf{x}}_{(\ell)} - \frac{p(\ell)\bar{\mathbf{x}}_{(\ell)} + p(\ell')\bar{\mathbf{x}}_{(\ell')}}{p(\ell) + p(\ell')} = \frac{p(\ell')[\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}]}{p(\ell) + p(\ell')},$$

et de manière symétrique :

$$\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}_{(\ell \cup \ell')} = \frac{p(\ell)[\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}_{(\ell)}]}{p(\ell) + p(\ell')}.$$

On conclut que :

$$d_{\text{ward}}(C_\ell, C_{\ell'}) = p(\ell)\|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2 + p(\ell')\|\bar{\mathbf{x}}_{(\ell')} - \bar{\mathbf{x}}_{(\ell \cup \ell')}\|^2 = \frac{p(\ell)p(\ell')}{p(\ell) + p(\ell')} \|\bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}\|^2.$$

□

CHOIX DE LA BONNE PARTITION

Nous nous limitons ici à une approche empirique du choix de nombre de classes. L'idée est de repérer un saut (si possible) entre les agrégations courtes distances (branches courtes de l'arbre) et les longues distances (branches longues de l'arbre). On peut ensuite vérifier la robustesse de cette répartition en classes avec la méthode des centres mobiles par exemple (voir la prochaine section). Parfois le nombre adéquat de classes est donné par le type de données. Il existe aussi des tests statistiques pour déterminer le bon nombre de classes.

Exercice 6.1.

Effectuer la CAH du tableau suivant selon la méthode de Ward.

	\mathbf{x}^1	\mathbf{x}^2	poids
\mathbf{x}'_1	1	1	$\frac{3}{10}$
\mathbf{x}'_2	2	2	$\frac{3}{10}$
\mathbf{x}'_3	4	3	$\frac{1}{10}$
\mathbf{x}'_4	4	4	$\frac{2}{10}$
\mathbf{x}'_5	3	5	$\frac{1}{10}$

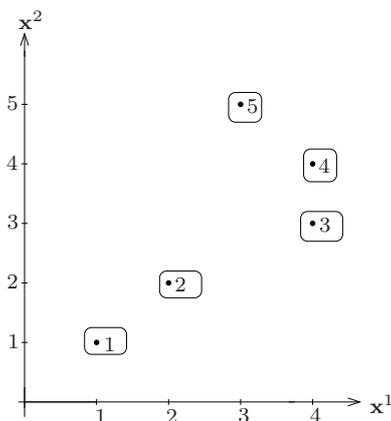


FIGURE 6.3 – Initialisation de l'algorithme.

Solution de l'exercice 6.1.

INITIALISATION

Au départ on a 5 classes C_1, \dots, C_5 , chacune contenant un seul individu. Le centre de gravité de chacune des classes est l'individu lui-même.

- Calcul des distances usuelles au carré entre les centres de gravité des classes.

$\ \bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}\ ^2$	1	2	3	4	5
1	0	2	13	18	20
2		0	5	8	10
3			0	1	5
4				0	2
5					0

Par exemple, $\|\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}\|^2 = \|\mathbf{x}_1 - \mathbf{x}_2\|^2 = (1 - 2)^2 + (1 - 2)^2 = 2$.

- Calcul des pseudo poids $\frac{p_{(\ell)}p_{(\ell')}}{p_{(\ell)}+p_{(\ell'')}}$ multipliés par 10 (afin d'avoir de quantités plus simples à écrire, mais cela ne change pas les relations d'ordre entre les coefficients) :

10.pseudo-poids	1	2	3	4	5
1	X	$\frac{3}{2}$	$\frac{3}{4}$	$\frac{6}{5}$	$\frac{3}{4}$
2		X	$\frac{3}{4}$	$\frac{6}{5}$	$\frac{3}{4}$
3			X	$\frac{2}{3}$	$\frac{1}{2}$
4				X	$\frac{2}{3}$
5					X

Par exemple, $10 \frac{p_1 p_2}{p_1 + p_2} = 10 \frac{\frac{3}{10} \frac{3}{10}}{\frac{3}{10} + \frac{3}{10}} = \frac{3}{2}$.

- Calcul des distances de Ward entre les classes (qui sont les individus à cette étape), multipliées par 10 :

$$d_{\text{ward}}(C_\ell, C_{\ell'}) = \frac{p(\ell)p(\ell')}{p(\ell) + p(\ell')} \|\bar{\mathbf{x}}^{(\ell)} - \bar{\mathbf{x}}^{(\ell')}\|^2.$$

$10 \cdot d_{\text{ward}}(C_\ell, C_{\ell'})$	1	2	3	4	5
1	0	3	9,75	21,6	15
2		0	3,75	9,6	7,5
3			0	0,67	2,5
4				0	1,33
5					0

Par exemple, $10 \cdot d_{\text{ward}}(C_1, C_2) = \frac{3}{2} \cdot 2 = 3$.

ÉTAPE 0

Au vu de la matrice des distances, on décide d'agréer les classes 3 et 4 en une nouvelle classe notée C_6 .

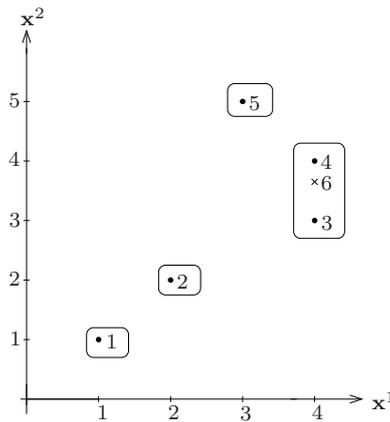


FIGURE 6.4 – Étape 0 de l'algorithme.

- On calcule le centre de gravité $\bar{\mathbf{x}}_{(6)}$ de cette nouvelle classe :

$$\bar{\mathbf{x}}_{(6)} = \frac{1}{\frac{1}{10} + \frac{2}{10}} \left[\frac{1}{10} \begin{pmatrix} 4 \\ 3 \end{pmatrix} + \frac{2}{10} \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right] = \frac{1}{3} \begin{pmatrix} 12 \\ 11 \end{pmatrix} = \begin{pmatrix} 4 \\ \frac{11}{3} \end{pmatrix} \simeq \begin{pmatrix} 4 \\ 3,667 \end{pmatrix}.$$

Le centre de gravité $\bar{\mathbf{x}}_{(6)}$ est affecté du poids $\frac{1}{10} + \frac{2}{10} = \frac{3}{10}$. La matrice des données correspondantes est :

	\mathbf{x}^1	\mathbf{x}^2	poids
\mathbf{x}'_1	1	1	$\frac{3}{10}$
\mathbf{x}'_2	2	2	$\frac{3}{10}$
\mathbf{x}'_5	3	5	$\frac{1}{10}$
$\bar{\mathbf{x}}'_{(6)}$	4	$\frac{11}{3}$	$\frac{3}{10}$

- Calcul des distances usuelles au carré et des pseudo-poids multipliés par 10.

$\ \bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}\ ^2$	1	2	5	6
1	0	2	20	16, 11
2		0	10	6, 78
5			0	2, 78
6				0

10.pseudo-poids	1	2	5	6
1	0	$\frac{3}{2}$	$\frac{3}{4}$	$\frac{3}{2}$
2		0	$\frac{3}{4}$	$\frac{3}{2}$
5			0	$\frac{3}{4}$
6				0

- Calcul des distances de Ward multipliées par 10

$10 \cdot d_{\text{ward}}(C_\ell, C_{\ell'})$	1	2	5	6
1	0	3	15	24, 16
2		0	7,5	10, 17
5			0	2, 09
6				0

ÉTAPE 1

Au vu de la matrice des distances de Ward, on décide d'agréer les classes 5 et 6 en une nouvelle classe notée C_7 .

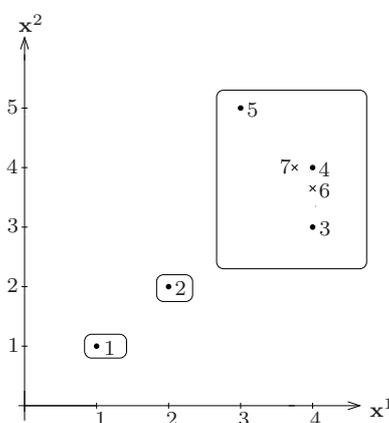


FIGURE 6.5 – Étape 1 de l'algorithme.

On calcule le centre de gravité $\bar{\mathbf{x}}_{(7)}$ de cette nouvelle classe :

$$\bar{\mathbf{x}}_{(7)} = \frac{1}{\frac{1}{10} + \frac{3}{10}} \left[\frac{1}{10} \begin{pmatrix} 3 \\ 5 \end{pmatrix} + \frac{3}{10} \begin{pmatrix} 4 \\ 3,667 \end{pmatrix} \right] = \begin{pmatrix} 3,75 \\ 4 \end{pmatrix}.$$

Le nouveau centre de gravité est affecté du poids $\frac{3}{10} + \frac{1}{10} = \frac{4}{10}$.

La nouvelle matrice des données est :

	\mathbf{x}^1	\mathbf{x}^2	poids
\mathbf{x}'_1	1	1	$\frac{3}{10}$
\mathbf{x}'_2	2	2	$\frac{3}{10}$
$\bar{\mathbf{x}}'_{(7)}$	3,75	4	$\frac{4}{10}$

- Calcul des distances usuelles au carré et des pseudo-poids multipliés par 10.

$\ \bar{\mathbf{x}}_{(\ell)} - \bar{\mathbf{x}}_{(\ell')}\ ^2$	1	2	7
1	0	2	16,56
2		0	7,06
7			0

10.pseudo-poids	1	2	7
1	0	$\frac{3}{2}$	$\frac{12}{7}$
2		0	$\frac{12}{7}$
7			0

- Calcul des distances de Ward multipliées par 10

$10 \cdot d_{\text{ward}}(C_\ell, C_{\ell'})$	1	2	7
1	0	3	28,4
2		0	12,1
7			0

ÉTAPE 2

Au vu de la matrice des distances de Ward, on décide d'agréer les classes 1 et 2 en une nouvelle classe notée C_8 .

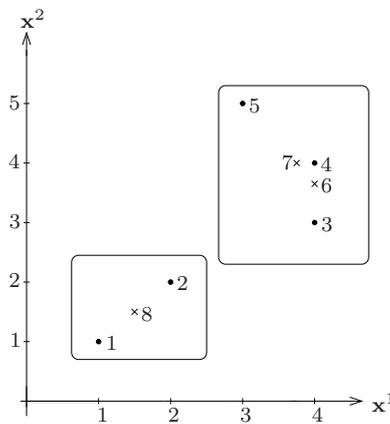


FIGURE 6.6 – Étape 2 de l'algorithme.

On calcule le centre de gravité $\bar{\mathbf{x}}_{(8)}$ de cette nouvelle classe :

$$\bar{\mathbf{x}}_{(8)} = \frac{1}{\frac{3}{10} + \frac{3}{10}} \left[\frac{3}{10} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{3}{10} \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right] = \begin{pmatrix} 1,5 \\ 1,5 \end{pmatrix}.$$

Le nouveau centre de gravité est affecté du poids $\frac{3}{10} + \frac{3}{10} = \frac{6}{10}$.

La nouvelle matrice des données est :

	x^1	x^2	poids
$\bar{x}'_{(7)}$	3,75	4	$\frac{4}{10}$
$\bar{x}'_{(8)}$	1,5	1,5	$\frac{6}{10}$

- Calcul des distances usuelles au carré et des pseudo-poids multipliés par 10.

$\ \bar{x}_{(\ell)} - \bar{x}_{(\ell')}\ ^2$	7	8
7	0	11,31
8		0

10.pseudo-poids	7	8
7	0	2,4
8		0

- Calcul des distances de Ward multipliées par 10

$10 \cdot d_{\text{ward}}(C_\ell, C_{\ell'})$	7	8
7	0	27,15
8		0

ÉTAPE 3

Au vu de la matrice des distances de Ward, on décide d'agréer les classes 7 et 8 en une nouvelle classe notée C_9 .

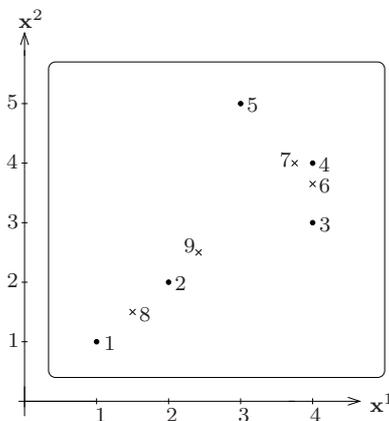


FIGURE 6.7 – Étape 3 de l'algorithme.

On calcule le centre de gravité $\bar{x}_{(9)}$ de cette nouvelle classe :

$$\bar{x}_{(9)} = \frac{1}{\frac{4}{10} + \frac{6}{10}} \left[\frac{4}{10} \begin{pmatrix} 3,75 \\ 4 \end{pmatrix} + \frac{6}{10} \begin{pmatrix} 1,5 \\ 1,5 \end{pmatrix} \right] = \begin{pmatrix} 2,4 \\ 2,5 \end{pmatrix}.$$

Le nouveau centre de gravité est affecté du poids $\frac{4}{10} + \frac{6}{10} = 1$.

FIN DE L'ALGORITHME

D'après la remarque 6.1 :

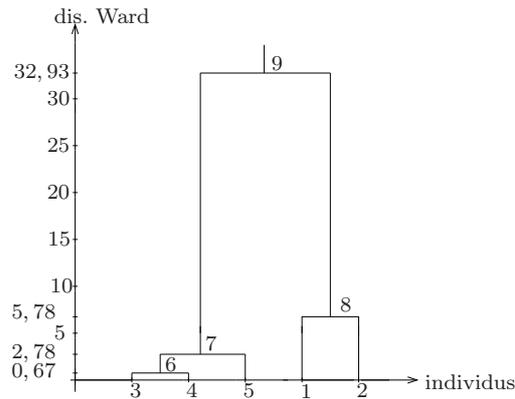


FIGURE 6.8 – Dendrogramme de la CAH.

- le centre de gravité du nuage de points est, $\bar{\mathbf{x}}_{(9)} = (2.4, 2.5)'$.
- L'inertie totale du nuage de points est la somme des pertes d'inertie inter-classes :

$$J_{\text{tot}} = 0,67 + 2,09 + 3 + 27,15 \simeq 32,9.$$

Au vu du dendrogramme, il semble raisonnable de garder trois classes dans cet exemple.

6.3 MÉTHODES NON HIÉRARCHIQUES : MÉTHODE DES CENTRES MOBILES

Les auteurs à l'origine de cette méthode et de ses extensions sont Forgy, Mac Queen, Diday.

On fixe le nombre de classes k à l'avance. Cette méthode permet alors de partager n individus en k classes de manière rapide.

-*Algorithme.*

1. *Initialisation.* Choisir k centres provisoires dans \mathbb{R}^p (tirés au hasard).
2. *Pas de l'algorithme.*
 - Chacun des individus est associé à la classe dont le centre est le plus proche. On obtient ainsi une partition des individus en k classes.
 - Remplacer les k centres par les centres de gravité des nouvelles classes.
 - Recommencer jusqu'à stabilisation.

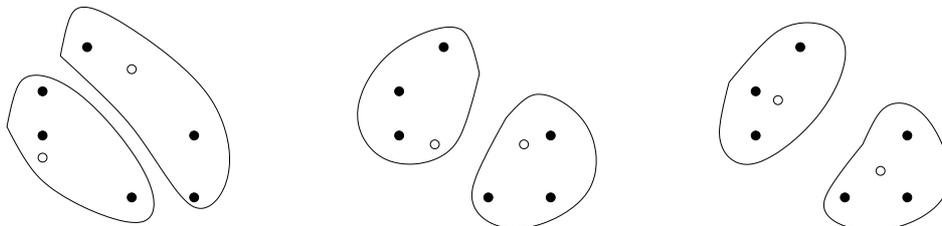


FIGURE 6.9 – Illustration de la méthode des centres mobiles, $k = 2$.

- *Avantages.*

1. On peut montrer qu'à chaque étape l'inertie intra-classe diminue (bonne notion d'homogénéité).
2. Algorithme rapide, qui permet de traiter un grand nombre de données.

- *Inconvénients.*

1. On doit fixer le nombre de classes à l'avance. Donc on ne peut déterminer le nombre idéal de groupes. Mais cela peut être utile pour consolider une CAH.
2. Le résultat dépend de la condition initiale. Ainsi, on n'est pas sûr d'atteindre la partition en k classes, telle que J_{intra} est minimum (on a minimum "local" et non "global").

CHAPITRE 7

ANALYSE DISCRIMINANTE

7.1 INTRODUCTION

L'analyse discriminante désigne un ensemble de techniques qui permettent de décrire, expliquer et prédire l'appartenance d'un individu à une classe pré-définie, les individus étant caractérisés par un certain nombre de variables quantitatives (numériques).

7.1.1 CONTEXTE

On considère un échantillon I formé de n individus, $\text{card}(I) = n$.

On dispose de la donnée de p variables numériques, indexées par un ensemble J , $\text{card}(J) = p$. Les données sont représentées dans une matrice $X = (x_{ij})$ de taille $n \times p$.

- Pour $i \in \{1, \dots, n\}$, la $i^{\text{ème}}$ ligne, notée \mathbf{x}'_i , représente les données de toutes les variables pour le $i^{\text{ème}}$ individu.
- Pour $j \in \{1, \dots, p\}$, la $j^{\text{ème}}$ colonne, notée \mathbf{x}^j , représente les données de la $j^{\text{ème}}$ variable pour tous les individus.
- Le coefficient x_{ij} est la donnée de la $j^{\text{ème}}$ variable pour le $i^{\text{ème}}$ individu.

De plus, on suppose que les individus sont regroupés en k classes C_1, \dots, C_k au moyen d'une variable qualitative nominale \mathbf{y} ayant k modalités. Pour chaque $\ell \in \{1, \dots, k\}$, on note I_ℓ l'ensemble des individus de la classe C_ℓ , avec $\text{card}(I_\ell) = n_\ell$. Ainsi, $\sum_{\ell=1}^k n_\ell = n$.

On dispose donc du tableau suivant.

	\mathbf{x}^1	\dots	\mathbf{x}^j	\dots	\mathbf{x}^p	\mathbf{y}
\mathbf{x}_1	x_{11}	\dots	x_{1j}	\dots	x_{1n}	1
\vdots	\vdots		\vdots		\vdots	\vdots
\mathbf{x}_{n_1}	$x_{n_1 1}$	\dots	$x_{n_1 j}$	\dots	$x_{n_1 n}$	1
\vdots	\vdots		\vdots		\vdots	\vdots
$\mathbf{x}_{n_1 + \dots + n_{k-1} + 1}$		\dots		\dots		k
\vdots	\vdots		\vdots		\vdots	\vdots
$\mathbf{x}_{n_1 + \dots + n_k}$		\dots		\dots		k

L'analyse discriminante a deux buts principaux.

- *Approche descriptive.* Confirmer la distinction en k groupes à partir des observations des données X . Restituer géométriquement cette séparation en groupes. Pour cela, il faut déterminer les combinaisons des variables qui séparent le mieux possible les k groupes.
- *Approche décisionnelle.* Étant donné un ou plusieurs nouveaux individus sur lequel on a mesuré les p variables, établir une règle de classement qui permette d'affecter l'individu à une des k classes.

Les domaines d'applications de l'analyse discriminante sont nombreux.

- Les premiers travaux sur l'analyse discriminante sont consacrés aux mesures de boîtes crâniennes humaines et avaient pour but, soit de déterminer le sexe des individus, soit de mettre en évidence les différences entre groupes humains (travaux de Pearson et Mahalanobis, 1920).
- *Données économiques et financières.* Travaux de Fisher (1936), Welch et Ward (1939-1944), Rao (1947-1949), Anderson (1951).
- *Médecine.* Aide au diagnostic (les classes représentent des types de maladie), et au pronostic (les classes représentent les issues probables d'une maladie).
- *Météorologie.* Préviation d'avalanches, tempêtes, etc.
- *Domaine bancaire.* Préviation de risques en bourses, risques de faillites de clients débiteurs.
- *Fiscalité.* Détection de fraude (Deux groupes : fraudeurs/non-fraudeurs).

7.1.2 NOTATIONS

Si les données ont été récoltées dans les mêmes conditions pour les n individus, on leur attribue un même poids relatif $\frac{1}{n}$. Il se peut que ce ne soit pas le cas, auquel cas on leur attribue des poids relatifs $(p_i)_{i=1}^n$, tels que $\sum_{i=1}^n p_i = 1$.

Les notations sont celles du chapitre 3. Rappelons-en quelques unes. Pour tout $\ell \in \{1, \dots, k\}$:

- $p_{(\ell)}$ est le poids relatif de la classe C_ℓ :

$$p_{(\ell)} = \sum_{i \in I_\ell} p_i. \quad \text{Ainsi,} \quad \sum_{\ell=1}^k p_{(\ell)} = 1.$$

- $\bar{\mathbf{x}}_{(\ell)}$ est le centre de gravité ou barycentre de la classe C_ℓ :

$$\bar{\mathbf{x}}_{(\ell)} = (\bar{x}_{(\ell)}^1, \dots, \bar{x}_{(\ell)}^p)',$$

où pour tout $j \in \{1, \dots, p\}$,

$$\bar{x}_{(\ell)}^j = \frac{1}{p_{(\ell)}} \sum_{i \in I_\ell} p_i x_{ij}.$$

- La matrice des variances-covariances est notée V , celle des variances intra-classes est W et celle des variances inter-classes est B .

Un des outils clé pour ce chapitre est la notion d'inertie totale J_{tot} du nuage de points $N(I)$, d'inertie intra-classe J_{intra} et d'inertie inter-classe J_{inter} de la section 3.5.

7.2 DISCRIMINATION LINÉAIRE

L'analyse factorielle discriminante cherche à déterminer des nouvelles variables, combinaisons linéaires des variables d'origine, qui permettent de visualiser le mieux possible en projection la séparation en k classes. Une combinaison linéaire des variables revient à considérer un nouvel axe. Le premier axe, de vecteur directeur \mathbf{a}_1 , sera choisi de sorte à ce que l'inertie intra-classe des données projetées sur cet axe soit minimale et l'inertie inter-classe des données projetées soit maximale. Ensuite, on cherchera un deuxième axe de vecteur directeur \mathbf{a}_2 , orthogonal au premier, qui vérifie ce même principe. On continue jusqu'à ce que l'on retrouve toute l'information de départ. Cependant, comme les nouveaux axes sont construits de sorte à ce que l'information soit maximale, il suffira d'en considérer 1, 2 ou 3, ce qui permet une bonne visualisation de la répartition en classes des données (dans \mathbb{R} , \mathbb{R}^2 , \mathbb{R}^3).

7.2.1 PRINCIPE DE DISCRIMINATION LINÉAIRE

On considère le nuage des individus $N(I)$ dans \mathbb{R}^p . On suppose que \mathbb{R}^p est muni d'une métrique d_M , que l'on fixera par la suite.

INERTIE POUR UNE PROJECTION DES DONNÉES

On rappelle, voir section 3.4.5, que l'inertie totale du nuage de points $\mathcal{J}_{\text{tot}}^M$ mesure la dispersion des données, et est définie par :

$$\mathcal{J}_{\text{tot}}^M(X) = \sum_{i=1}^n p_i d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}).$$

Soit \mathcal{D} une droite de \mathbb{R}^p passant par le centre de gravité du nuage $\bar{\mathbf{x}}$, de vecteur directeur \mathbf{a} que l'on suppose M -normé à 1 : $\|\mathbf{a}\|_M = 1$. Notons π la projection M -orthogonale sur cette droite et $\pi(X)$ le nuage projeté.

Calculons l'inertie totale du nuage de points projeté sur la droite \mathcal{D} :

$$\begin{aligned} \mathcal{J}_{\text{tot}}^M(\pi(X)) &= \sum_{i=1}^n p_i d_M^2(\pi(\mathbf{x}_i), \pi(\bar{\mathbf{x}})) = \sum_{i=1}^n p_i \|\pi(\mathbf{x}_i - \bar{\mathbf{x}})\|_M^2 \\ &= \sum_{i=1}^n p_i \|\pi(\mathbf{y}_i)\|_M^2, \text{ par définition des données centrées, section 3.4.1} \\ &= \sum_{i=1}^n p_i \langle \mathbf{y}_i, \mathbf{a} \rangle_M^2, \text{ par définition de la projection } M\text{-orthogonale, section 1.5.4} \\ &= \sum_{i=1}^n p_i \mathbf{a}' M \mathbf{y}_i \mathbf{y}_i' M \mathbf{a} \\ &= \mathbf{a}' M \left(\sum_{i=1}^n p_i \mathbf{y}_i \mathbf{y}_i' \right) M \mathbf{a} = \mathbf{a}' M V M \mathbf{a}, \text{ d'après la propriété 3.6.} \end{aligned}$$

Notons \mathbf{u} le vecteur $\mathbf{u} = M \mathbf{a}$, appelé le *facteur associé à l'axe* \mathbf{a} . Alors, l'inertie totale du nuage de points projeté sur la droite \mathcal{D} est égale à :

$$\mathcal{J}_{\text{tot}}^M(\pi(X)) = \mathbf{a}' M V M \mathbf{a} = \mathbf{u}' V \mathbf{u}.$$

On parle aussi de l'*inertie totale du nuage de point sur le facteur \mathbf{u}* .

De manière analogue l'*inertie intra-classe sur le facteur \mathbf{u}* est :

$$\mathcal{J}_{\text{intra}}^M(\pi(X)) = \mathbf{u}'W\mathbf{u},$$

et l'*inertie inter-classe sur le facteur \mathbf{u}* est :

$$\mathcal{J}_{\text{inter}}^M(\pi(X)) = \mathbf{u}'B\mathbf{u}.$$

Du théorème 3.3 qui dit que $V = W + B$, on déduit le corollaire suivant :

Corollaire 7.1. *L'inertie totale sur le facteur \mathbf{u} est égale à l'inertie intra-classe sur le facteur \mathbf{u} plus l'inertie inter-classe sur le facteur \mathbf{u} :*

$$\mathcal{J}_{\text{tot}}^M(\pi(X)) = \mathcal{J}_{\text{intra}}^M(\pi(X)) + \mathcal{J}_{\text{inter}}^M(\pi(X)) \Leftrightarrow \mathbf{u}'V\mathbf{u} = \mathbf{u}'W\mathbf{u} + \mathbf{u}'B\mathbf{u}.$$

PRINCIPE DE DISCRIMINATION LINÉAIRE

Dans un premier temps, nous cherchons une droite \mathcal{D}_1 de vecteur directeur \mathbf{a}_1 , passant par $\bar{\mathbf{x}}$, telle que l'inertie intra-classe du nuage de points projeté sur \mathcal{D}_1 soit minimale et l'inertie inter-classe du nuage de points projeté sur \mathcal{D}_1 soit maximale.

Dans la quantité $\mathbf{u}'V\mathbf{u} = \mathbf{u}'W\mathbf{u} + \mathbf{u}'B\mathbf{u}$, nous cherchons donc à minimiser $\mathbf{u}'W\mathbf{u}$ et maximiser $\mathbf{u}'B\mathbf{u}$, sur tous les vecteurs \mathbf{u} de la forme $\mathbf{u} = M\mathbf{a}$.

Il y a deux manières (équivalentes) de résoudre ce problème. Remarquons que lorsque l'on multiplie \mathbf{u} par une constante c , l'équation devient $c^2\mathbf{u}'V\mathbf{u} = c^2\mathbf{u}'W\mathbf{u} + c^2\mathbf{u}'B\mathbf{u}$, et le problème d'optimisation ne change pas de nature.

1. Ainsi, on peut choisir c de sorte à ce que le membre de gauche soit égal à 1. Autrement dit, on peut choisir le vecteur \mathbf{u} de sorte à ce que $\mathbf{u}'V\mathbf{u} = \|\mathbf{u}\|_V^2 = 1$, et maximiser $\mathbf{u}'B\mathbf{u}$. Ceci nous permet de préciser le choix de métrique. On a supposé le vecteur \mathbf{a} M -normé à 1, *i.e.*, $\mathbf{a}'M\mathbf{a} = 1$. De la relation, $\mathbf{u} = M\mathbf{a}$, on déduit $\mathbf{a} = M^{-1}\mathbf{u}$; d'où

$$\mathbf{a}'M\mathbf{a} = \mathbf{u}'M^{-1}MM^{-1}\mathbf{u} = \mathbf{u}'M^{-1}\mathbf{u}.$$

Pour que cette quantité soit égale à 1, il suffit de choisir $M^{-1} = V$, c'est-à-dire $M = V^{-1}$. La distance considérée sur \mathbb{R}^p est donc $d_{V^{-1}}$.

2. On peut aussi choisir c de sorte à ce que $c^2\mathbf{u}'W\mathbf{u} = 1$. Autrement dit, on peut choisir le vecteur \mathbf{u} de sorte à ce que $\mathbf{u}'W\mathbf{u} = \|\mathbf{u}\|_W^2 = 1$, et maximiser $\mathbf{u}'B\mathbf{u}$.

Par un raisonnement analogue, ceci revient à choisir la métrique $d_{W^{-1}}$ sur \mathbb{R}^p .

Les deux méthodes étant équivalentes, il suffit de choisir l'une ou l'autre. Les deux approches sont présentes dans la littérature; nous choisissons de détailler la première, et mentionnerons brièvement la deuxième. Dans la suite, nous supposons que les matrices V et W sont régulières, donc inversibles.

7.2.2 DISCRIMINATION DES DONNÉES : APPROCHE 1

Choisissons d'abord l'approche 1 au problème d'optimisation. La métrique considérée sur \mathbb{R}^p est $d_{V^{-1}}$ et on cherche \mathbf{u}_1 solution du problème suivant :

$$\left\{ \begin{array}{l} \mathbf{u}'B\mathbf{u} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\|_V = 1 \text{ est satisfaite.} \end{array} \right.$$

D'après la proposition 1.1, on sait que la solution \mathbf{u}_1 est un vecteur propre V -unitaire correspondant à la plus grande valeur propre λ_1 de la matrice $V^{-1}B$. On a alors, $\mathbf{u}'_1B\mathbf{u}_1 = \lambda_1$.

On cherche ensuite \mathbf{u}_2 , V -unitaire et V -orthogonal à \mathbf{u}_1 , tel que $\mathbf{u}'_2B\mathbf{u}_2$ soit maximal. En itérant ces étapes, le problème devient :

Trouver un repère V -orthonormé $(\bar{\mathbf{x}}, \mathbf{u}_1, \dots, \mathbf{u}_p)$ de \mathbb{R}^p tel que $\mathbf{u}'_1B\mathbf{u}_1$ soit maximal, puis $\mathbf{u}'_2B\mathbf{u}_2$, etc.

Le théorème 1.4 énonce dans ce contexte nous dit :

Théorème 7.2. *La solution au problème d'optimisation sous contrainte est la base¹ V -orthonormée $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ formée des vecteurs propres de $V^{-1}B$, telle que \mathbf{u}_1 correspond à la plus grande valeur propre λ_1 de $V^{-1}B$, puis \mathbf{u}_2 à la deuxième valeur propre λ_2 de $V^{-1}B$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall \alpha \in \{1, \dots, p\}, \quad \mathbf{u}'_\alpha B \mathbf{u}_\alpha = \lambda_\alpha.$$

Remarque 7.1. Comme V est inversible, le rang de la matrice $V^{-1}B$ est égal au rang de B . Or, le rang de la matrice B est au plus de $k - 1$, où k est le nombre de classes. Ainsi, il y a au plus $k - 1$ valeurs propres non-nulles. Notons r le rang de la matrice B , alors :

$$\text{rg}(B) = r \leq k - 1.$$

Définition 7.1. Les vecteurs unitaires $\mathbf{u}_1, \dots, \mathbf{u}_r$, correspondant à des valeurs propres non-nulles de $V^{-1}B$ sont les *facteurs discriminants*. Chaque vecteur \mathbf{u}_α est combinaison linéaire des vecteurs $\mathbf{x}^1, \dots, \mathbf{x}^p$, ce qui signifie que \mathbf{u}_α représente une nouvelle variable "abstraite", combinaison des variables d'origine.

Les vecteurs directeurs $\mathbf{a}_1, \dots, \mathbf{a}_r$ des droites sur lesquelles on projette, associés au facteurs $\mathbf{u}_1, \dots, \mathbf{u}_r$ s'appellent les *axes discriminants*, où on rappelle que, pour tout $\alpha \in \{1, \dots, r\}$, $\mathbf{u}_\alpha = V^{-1}\mathbf{a}_\alpha$.

Définition 7.2. Pour tout $\alpha \in \{1, \dots, r\}$, on appelle $\alpha^{\text{ème}}$ *variable discriminante*, le vecteur

$$\Psi^\alpha = YV^{-1}\mathbf{a}_\alpha = Y\mathbf{u}_\alpha,$$

qui représente les coordonnées de tous les individus (centrés) sur le $\alpha^{\text{ème}}$ axe discriminant \mathbf{a}_α .

1. Cette base existe d'après le théorème 1.3 qui est la forme généralisée du théorème spectral pour les matrices symétriques.

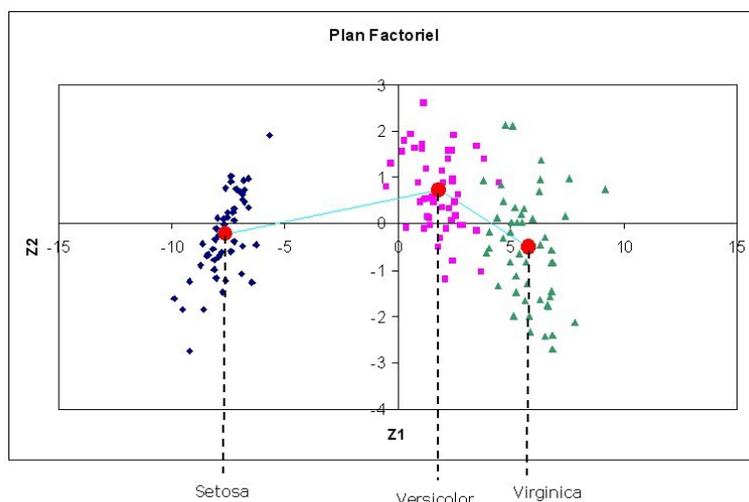


FIGURE 7.1 – Analyse discriminante linéaire sur les données des Iris de Fisher. Source : Dodabe [Public domain], via Wikimedia Commons.

Définition 7.3. Pour tout $\alpha \in \{1, \dots, r\}$, on appelle *rapport de corrélations²* associé au facteur \mathbf{u}_α , la fonction :

$$\eta^2(\Psi^\alpha) = \frac{\mathbf{u}'_\alpha B \mathbf{u}_\alpha}{\mathbf{u}'_\alpha V \mathbf{u}_\alpha}.$$

On parle aussi du *pouvoir discriminant* du facteur \mathbf{u}_α .

Propriété 7.1. Pour tout $\alpha \in \{1, \dots, r\}$, le rapport de corrélations associé au facteur \mathbf{u}_α satisfait aux propriétés suivantes :

- $0 \leq \eta^2(\Psi^\alpha) \leq 1$
- $\eta^2(\Psi^\alpha) = \lambda_\alpha$.

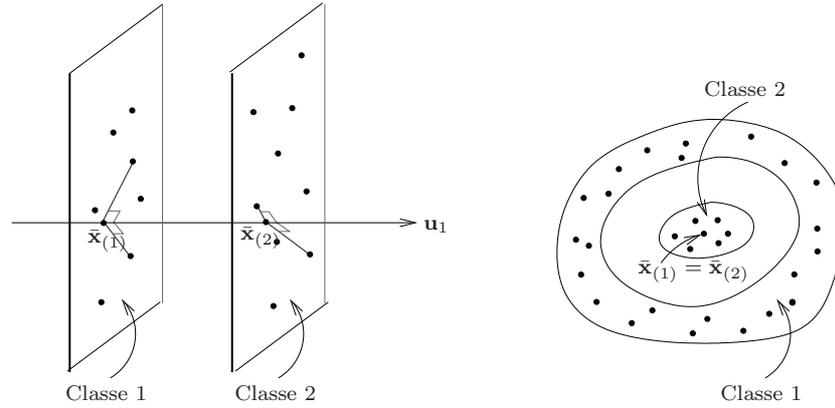
Démonstration. La première propriété découle du fait que les matrices B et V sont définies positives. La deuxième propriété est une conséquence du théorème 7.2 et du fait que le vecteur \mathbf{u}_α est V -normé à 1. \square

Remarque 7.2. Regardons les cas extrêmes pour la plus grande valeur propre λ_1 .

- Si $\lambda_1 = 1$, en projection l'inertie intra-classe est nulle. Les k nuages de points sont donc chacun dans un hyper-plan orthogonal à \mathbf{u}_1 , voir Figure 7.2 (gauche).
- Si $\lambda_1 = 0$, en projection l'inertie inter-classe est nulle, ce qui signifie que les centres de gravité des classes sont confondus. Attention, cela ne veut pas dire qu'il n'y a pas de discrimination possible, mais qu'il n'y a pas de discrimination *linéaire* possible, elle pourrait être quadratique ou autre, voir Figure 7.2 (droite).

Les valeurs propres $\lambda_1, \dots, \lambda_r$ sont une mesure pessimiste du pouvoir discriminant des facteurs $\mathbf{u}_1, \dots, \mathbf{u}_r$. En effet, elles mesurent une discrimination linéaire entre les individus, mais il se pourrait qu'il existe une discrimination d'une autre sorte.

2. Dans la littérature anglophone, la racine de ce rapport est appelé *corrélation canonique*



CALCULS POUR LE THÉORÈME 7.2

Pour calculer les vecteurs propres et valeurs propres de la matrice $V^{-1}B$, on se réfère aux conséquences de la deuxième démonstration du théorème 1.3.

Soit la matrice symétrique $V^{-\frac{1}{2}}BV^{-\frac{1}{2}}$, de taille $p \times p$; la matrice étant symétrique, elle est donc diagonalisable dans une base orthonormée. Soit $\lambda_1, \dots, \lambda_r$ ses valeurs propres non nulles en ordre décroissant, et $\mathbf{p}^1, \dots, \mathbf{p}^r$ les vecteurs propres correspondants. Alors, pour tout $\alpha \in \{1, \dots, r\}$, λ_α est valeur propre de la matrice $V^{-1}B$ et $\mathbf{u}_\alpha = V^{-\frac{1}{2}}\mathbf{p}^\alpha$ est le vecteur propre, V -normé à 1, associé.

De plus, par définition (voir équation (3.3)), les coefficients de la matrice B sont :

$$b_{jj'} = \sum_{\ell=1}^k p^{(\ell)} (\bar{x}_{(\ell)}^j - \bar{x}^j) (\bar{x}_{(\ell)}^{j'} - \bar{x}^{j'}).$$

Autrement dit, la matrice B est égale au produit CC' , où $C = (c_{j\ell})$ est la matrice rectangulaire de taille $p \times k$, dont les coefficients sont définis par :

$$c_{j\ell} = \sum_{\ell=1}^k \sqrt{p^{(\ell)}} (\bar{x}_{(\ell)}^j - \bar{x}^j).$$

Ainsi, la matrice $V^{-\frac{1}{2}}BV^{-\frac{1}{2}}$ est de la forme XX' :

$$V^{-\frac{1}{2}}BV^{-\frac{1}{2}} = V^{-\frac{1}{2}}CC'V^{-\frac{1}{2}} = (V^{-\frac{1}{2}}C)(V^{-\frac{1}{2}}C)'$$

Se rappelant de l'exercice 1.3 qui montre que les matrices XX' et $X'X$ ont les mêmes valeurs propres non-nulles et qui explicite un lien entre les vecteurs propres correspondants, nous pouvons plutôt considérer la matrice symétrique $X'X = C'V^{-1}C$, de taille $k \times k$, ce qui évite de calculer la puissance $\frac{1}{2}$ de l'inverse de V . Nous obtenons donc le résultat suivant.

Proposition 7.1. *Soit $C'V^{-1}C$ la matrice de taille $k \times k$. Soient $\lambda_1, \dots, \lambda_r$ ses valeurs propres non-nulles, en ordre décroissant et $\mathbf{q}^1, \dots, \mathbf{q}^r$ les vecteurs propres correspondants. Alors, pour tout $\alpha \in \{1, \dots, r\}$, λ_α est valeur propre de la matrice $V^{-1}B$ et $\hat{\mathbf{u}}_\alpha = V^{-1}C\mathbf{q}^\alpha$ est un vecteur propre correspondant. Pour obtenir \mathbf{u}_α , il suffit de V -normer à 1 le vecteur $\hat{\mathbf{u}}_\alpha$: $\mathbf{u}_\alpha = \hat{\mathbf{u}}_\alpha / \|\hat{\mathbf{u}}_\alpha\|_V$.*

7.2.3 ÉVALUATION DES FACTEURS DISCRIMINANTS

Pour tout $\alpha \in \{1, \dots, r\}$, on se rappelle que la valeur propre λ_α mesure le pouvoir discriminant (linéaire) du facteur \mathbf{u}_α , avec $1 \geq \lambda_1 \geq \dots \geq \lambda_r \geq 0$. En utilisant une statistique, il est possible d'évaluer de manière plus précise ce pouvoir discriminant.

Pour tout $\alpha \in \{0, \dots, r-1\}$, on définit le *lambda de Wilks* :

$$\Lambda_{\alpha+1} = \prod_{\beta=\alpha+1}^r \frac{1}{1 + \lambda_\beta}.$$

Si les valeurs propres $\{\lambda_1, \dots, \lambda_r\}$ sont toutes "trop petites", alors $\Lambda_{\alpha+1}$ sera "trop grand". Plus précisément, en supposant que les données sont une réalisation d'un modèle probabiliste (du type de ce qui est fait à la section 7.3.2) et en effectuant certaines hypothèses sur les paramètres de la loi du modèle, la loi de $\Lambda_{\alpha+1}$ est connue et on a à disposition des tables (ou un ordinateur) pour la calculer explicitement. On peut alors comparer la valeur obtenue avec la valeur attendue et conclure (avec une certaine marge d'erreur) sur le pouvoir discriminant du facteur. On procède par itération sur α jusqu'à ce que $\Lambda_{\alpha+1}$ dépasse la valeur seuil choisie.

On peut aussi étudier certaines transformations de cette statistique, comme celle de Bartlett et Rao, qui suivent aussi des lois connues.

7.2.4 DISCRIMINATION DES DONNÉES : APPROCHE 2

Avec la deuxième approche, la métrique considérée sur \mathbb{R}^p est $d_{W^{-1}}$ et on cherche \mathbf{v}_1 solution du problème suivant :

On cherche donc \mathbf{v}_1 solution du problème suivant :

$$\begin{cases} \mathbf{v}'B\mathbf{v} \text{ est maximale} \\ \text{la contrainte } \|\mathbf{u}\|_W = 1 \text{ est satisfaite.} \end{cases}$$

En itérant les étapes, le problème devient :

Trouver une base W -orthonormée $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ de \mathbb{R}^p telle que $\mathbf{v}'_1 B \mathbf{v}_1$ soit maximale, puis $\mathbf{v}'_2 B \mathbf{v}_2$, etc.

Le théorème 1.4 énonce dans ce contexte nous dit :

Théorème 7.3. *La solution au problème d'optimisation sous contrainte est la base³ W -orthonormée $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ formée des vecteurs propres de $W^{-1}B$, telle que \mathbf{v}_1 correspond à la plus grande valeur propre μ_1 de $W^{-1}B$, puis \mathbf{v}_2 à la deuxième valeur propre μ_2 de $W^{-1}B$, etc. où les valeurs propres sont comptées avec multiplicité. On a alors,*

$$\forall \alpha \in \{1, \dots, p\}, \quad \mathbf{v}'_\alpha W^{-1} B \mathbf{v}_\alpha = \mu_\alpha.$$

3. Cette base existe d'après le théorème 1.3 qui est la forme généralisée du théorème spectral pour les matrices symétriques.

7.2.5 LIEN ENTRE LES DEUX APPROCHES

Soit λ valeur propre de $V^{-1}B$, de vecteur propre associé \mathbf{u} , telle que $0 < \lambda < 1$. Alors,

$$V^{-1}B\mathbf{u} = \lambda\mathbf{u} \Leftrightarrow B\mathbf{u} = \lambda(W + B)\mathbf{u} \Leftrightarrow B\mathbf{u} = \frac{\lambda}{1 - \lambda}W\mathbf{u} \Leftrightarrow W^{-1}B\mathbf{u} = \frac{\lambda}{1 - \lambda}\mathbf{u}.$$

Ainsi, les matrices $V^{-1}B$ et $W^{-1}B$ ont les mêmes vecteurs propres et les valeurs propres comprises entre 0 et 1 strictement, sont liées par la relation :

$$\forall \alpha \in \{1, \dots, r\}, \quad \frac{\lambda_\alpha}{1 - \lambda_\alpha} = \mu_\alpha, \quad \lambda_\alpha = \frac{\mu_\alpha}{1 + \mu_\alpha}. \quad (7.1)$$

Comme les vecteurs propres \mathbf{u}_α de $V^{-1}B$ sont V -normés à 1, et les vecteurs propres \mathbf{v}_α de $W^{-1}B$ sont W -normés à 1, on en déduit la relation :

$$\forall \alpha \in \{1, \dots, r\}, \quad \mathbf{u}_\alpha = \frac{\mathbf{v}_\alpha}{\|\mathbf{v}_\alpha\|_V}, \quad \mathbf{v}_\alpha = \frac{\mathbf{u}_\alpha}{\|\mathbf{u}_\alpha\|_W}.$$

7.3 ANALYSE FACTORIELLE DISCRIMINANTE : PRISE DE DÉCISION

Une fois trouvés les facteurs discriminants $\mathbf{u}_1, \dots, \mathbf{u}_r$ qui séparent au mieux les individus répartis en k classes, et après avoir évalué leur pouvoir discriminant, nous souhaitons trouver la classe d'affectation d'un nouvel individu \mathbf{x} décrit par un vecteur :

$$\mathbf{x} = (x^1, \dots, x^p)'$$

Nous allons étudier deux modèles d'affectation différents : le modèle géométrique et le modèle Bayésien.

7.3.1 RÈGLE GÉOMÉTRIQUE D'AFFECTION

Le modèle géométrique consiste à calculer la distance entre le nouvel individu \mathbf{x} et les centres de gravité $\bar{\mathbf{x}}_{(1)}, \dots, \bar{\mathbf{x}}_{(k)}$ de chacune des classes, et à affecter \mathbf{x} à la classe dont la distance est minimale.

Le choix de la distance dépend de si on suit l'approche 1 ou 2. Supposons que l'on suive l'approche 1 auquel cas on considère la distance $d_{V^{-1}}$ sur \mathbb{R}^p . Alors, pour tout $\ell \in \{1, \dots, k\}$,

$$\begin{aligned} d_{V^{-1}}^2(\mathbf{x}, \bar{\mathbf{x}}_{(\ell)}) &= (\mathbf{x} - \bar{\mathbf{x}}_{(\ell)})'V^{-1}(\mathbf{x} - \bar{\mathbf{x}}_{(\ell)}) \\ &= \mathbf{x}'V^{-1}\mathbf{x} - 2\mathbf{x}'V^{-1}\bar{\mathbf{x}}_{(\ell)} + \bar{\mathbf{x}}_{(\ell)}'V^{-1}\bar{\mathbf{x}}_{(\ell)}. \end{aligned}$$

Comme $\mathbf{x}'V^{-1}\mathbf{x}$ ne dépend pas de ℓ , la méthode géométrique revient à affecter \mathbf{x} à la classe ℓ telle que la quantité

$$\mathbf{x}'V^{-1}\bar{\mathbf{x}}_{(\ell)} - \frac{1}{2}\bar{\mathbf{x}}_{(\ell)}'V^{-1}\bar{\mathbf{x}}_{(\ell)}, \quad (7.2)$$

est maximale.

L'argument est similaire si on suit l'approche 2 avec pour choix de métrique $d_{W^{-1}}$.

Remarque 7.3. Cette méthode ne prend pas en compte le fait que le poids des différentes classes sont parfois très inégaux, ce qui peut être le cas si une des classes représente une maladie rare. La méthode suivante va permettre de résoudre ce problème.

7.3.2 MODÈLE BAYSIEN D'AFFECTION

Cette approche est probabiliste, elle a l'avantage de tenir compte des probabilités a priori des différentes classes, parfois très inégales (diagnostic d'une maladie rare par exemple).

L'ensemble des individus $I = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ est considéré comme étant un univers probabiliste, muni d'une tribu et d'une probabilité \mathbb{P} . Les individus sont répartis en classes C_1, \dots, C_k . Supposons que l'on ait, pour tout $\ell \in \{1, \dots, k\}$, une estimation de la probabilité $\mathbb{P}[C_\ell]$ qu'un individu appartienne à la classe C_ℓ et une estimation de la probabilité conditionnelle $\mathbb{P}[(x^1, \dots, x^p) | C_\ell]$ d'observer les données (x^1, \dots, x^p) sachant que l'individu appartient à la classe C_ℓ .

Étant donné ce modèle, on va affecter le nouvel individu observé $\mathbf{x} = (x^1, \dots, x^p)'$ à la classe C_ℓ , telle que la probabilité $\mathbb{P}[C_\ell | \mathbf{x}]$ (probabilité d'obtenir la classe C_ℓ sachant que \mathbf{x} est réalisé) soit maximale. D'après la formule de Bayes, cette probabilité est égale à :

$$\mathbb{P}[C_\ell | \mathbf{x}] = \frac{\mathbb{P}[\mathbf{x} | C_\ell] \mathbb{P}[C_\ell]}{\mathbb{P}[\mathbf{x}]}.$$

On souhaite maximiser cette quantité sur $\ell \in \{1, \dots, k\}$. Comme le dénominateur ne dépend pas de ℓ , on peut l'enlever du problème de maximisation.

En pratique, il faut faire des hypothèses supplémentaires pour calculer $\mathbb{P}[C_\ell | \mathbf{x}]$. Voici un exemple classique.

MODÈLE BAYSIEN D'AFFECTION DANS LE CAS NORMAL

On suppose que, pour tout $\ell \in \{1, \dots, k\}$, la loi conditionnelle d'un individu sachant qu'il appartient à la classe C_ℓ , est une loi normale $\mathcal{N}(\mathbf{m}_{(\ell)}, \Sigma)$, où $\mathbf{m}_{(\ell)}$ est le vecteur des moyennes et Σ est la matrice des variances. La densité d'une telle loi est f_ℓ définie sur \mathbb{R}^p par :

$$\forall \mathbf{x} \in \mathbb{R}^p, \quad f_\ell(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{(\ell)})' \Sigma^{-1} (\mathbf{x} - \mathbf{m}_{(\ell)})\right).$$

La quantité à maximiser devient alors $f_\ell(\mathbf{x})\mathbb{P}[C_\ell]$ ou, de manière équivalente, son logarithme.

En omettant tout ce qui ne dépend pas de ℓ , on conclut que le modèle Bayésien d'affectation dans le cas normal consiste à affecter l'individu \mathbf{x} à la classe ℓ pour laquelle la quantité,

$$\mathbf{x}' \Sigma^{-1} \mathbf{m}_{(\ell)} - \frac{1}{2} \mathbf{m}_{(\ell)}' \Sigma^{-1} \mathbf{m}_{(\ell)} + \ln \mathbb{P}[C_\ell] \tag{7.3}$$

est maximale.

Remarque 7.4.

- Le fait que la matrice des variances Σ est supposée être la même pour toutes les classes sous-entend que les k nuages de points ont la même dispersion, autrement dit la même forme. Si cette hypothèse n'est pas vérifiée, on pourra considérer des matrices des variances Σ_ℓ distinctes pour les différentes classes; la définition du score discriminant est adaptée et correspond à l'analyse discriminante quadratique.
- Les calculs ci-dessus supposent $\mathbb{P}[C_\ell]$, $\mathbf{m}_{(\ell)}$ et Σ connus. Dans la pratique, on les remplacera par des estimations.

7.3.3 LE CAS DE DEUX CLASSES

On suppose que $k = 2$. Dans le cas de deux classes, les résultats deviennent très explicites.

LA MATRICE B DES VARIANCES INTER-CLASSE

D'après la propriété 3.9, la matrice B est égale à :

$$B = p_{(1)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}})' + p_{(2)}(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}})'$$

Or, on a

$$p_{(1)} + p_{(2)} = 1 \text{ et } \bar{\mathbf{x}} = p_{(1)}\bar{\mathbf{x}}_{(1)} + p_{(2)}\bar{\mathbf{x}}_{(2)}.$$

Donc,

$$\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}} = \bar{\mathbf{x}}_{(1)} - p_{(1)}\bar{\mathbf{x}}_{(1)} - p_{(2)}\bar{\mathbf{x}}_{(2)} = (1 - p_{(1)})\bar{\mathbf{x}}_{(1)} - p_{(2)}\bar{\mathbf{x}}_{(2)} = p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}),$$

et par symétrie,

$$\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}} = p_{(1)}(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)}).$$

D'où

$$\begin{aligned} B &= p_{(1)}p_{(2)} \left[p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})' + p_{(1)}(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)})(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)})' \right] \\ &= p_{(1)}p_{(2)} \left[(1 - p_{(1)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})' + p_{(1)}(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)})(\bar{\mathbf{x}}_{(2)} - \bar{\mathbf{x}}_{(1)})' \right] \\ &= p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'. \end{aligned}$$

On voit bien que la matrice B est de rang 1. Il en est donc de même pour les matrices $V^{-1}B$ et $W^{-1}B$.

VALEURS PROPRES DE $V^{-1}B$ ET $W^{-1}B$

Comme les matrices $V^{-1}B$ et $W^{-1}B$ sont de rang 1, elles ont exactement une valeur propre non-nulle. Notons λ la valeur propre de $V^{-1}B$, et μ celle de $W^{-1}B$. Alors,

$$\begin{aligned} \lambda &= \text{Tr}(V^{-1}B) \\ &= \text{Tr}(V^{-1}p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})') \\ &= p_{(1)}p_{(2)} \text{Tr}(V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'), \text{ propriété de la trace (section 1.3.1)} \\ &= p_{(1)}p_{(2)} \text{Tr}((\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})), \text{ propriété de la trace (section 1.3.1)} \\ &= p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}), \text{ car la matrice est de dimension 1.} \end{aligned}$$

De manière analogue,

$$\mu = p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'W^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})$$

En conclusion,

$$\begin{aligned} \lambda &= p_{(1)}p_{(2)}d_{V^{-1}}^2(\bar{\mathbf{x}}_{(1)}, \bar{\mathbf{x}}_{(2)}) \\ \mu &= p_{(1)}p_{(2)}d_{W^{-1}}^2(\bar{\mathbf{x}}_{(1)}, \bar{\mathbf{x}}_{(2)}), \\ \lambda &= \frac{\mu}{1 + \mu}, \text{ d'après (7.1)} \end{aligned} \tag{7.4}$$

où l'on a la notation usuelle pour les distances, voir section 1.5.2 : $d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'M(\mathbf{x} - \mathbf{y})$.

VECTEURS PROPRES DE $V^{-1}B$ ET $W^{-1}B$

Nous avons l'intuition suivante. Étant donné qu'il n'y a que deux classes, l'*axe discriminant* est la droite de vecteur directeur $\hat{\mathbf{a}} = (\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})$ qui passe par les centres de gravité des deux classes et par $\bar{\mathbf{x}}$. Le facteur associé est $M\hat{\mathbf{a}}$, où M est la métrique choisie (V^{-1} ou W^{-1}). Remarquer que ce vecteur n'est pas forcément M^{-1} -normé à 1, mais sa norme ne joue pas de rôle lorsqu'on s'intéresse aux valeurs propres des matrices.

Les quantités, $\hat{\mathbf{u}} = V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})$, et $\hat{\mathbf{v}} = W^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})$ s'appellent *fonction de Fisher associée à la métrique V^{-1} , respectivement W^{-1}* .

Montrons que $\hat{\mathbf{u}}$ est effectivement vecteur directeur pour l'axe factoriel associé à la métrique V^{-1} . Autrement dit, montrons que $\hat{\mathbf{u}}$ est vecteur propre de la matrice $V^{-1}B$ pour la valeur propre λ . On a,

$$\begin{aligned} V^{-1}B\hat{\mathbf{u}} &= V^{-1}p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}), \text{ d'après la définition de } B \text{ et } \hat{\mathbf{u}} \\ &= p_{(1)}p_{(2)}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}), \text{ on a permuté les scalaires} \\ &= \lambda\hat{\mathbf{u}}, \text{ d'après la définition de } \lambda \text{ et } \hat{\mathbf{u}}. \end{aligned}$$

On démontre de manière analogue que $\hat{\mathbf{v}}$ est vecteur directeur pour l'axe factoriel associé à la métrique W^{-1} .

D'après la section 7.2.5, nous savons que les vecteurs $\hat{\mathbf{u}}$ et $\hat{\mathbf{v}}$ sont colinéaires ; en effet, les matrices $W^{-1}B$ et $V^{-1}B$ ont les mêmes vecteurs propres. Ainsi, il existe $a \in \mathbb{R}$, tel que $\hat{\mathbf{v}} = a\hat{\mathbf{u}}$:

$$W^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}) = aV^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}).$$

Pré-multipliant par $(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)})'$, et utilisant les équations (7.4), on obtient,

$$a = \frac{\mu}{\lambda} = (1 + \mu),$$

et donc :

$$W^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}) = (1 + \mu)V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}).$$

RÈGLE GÉOMÉTRIQUE D'AFFECTATION (RÈGLE DE FISHER)

Soit $\mathbf{x} = (x^1, \dots, x^p)'$ un nouvel individu. Alors, d'après le critère (7.2), on décide d'affecter \mathbf{x} à la classe 1, si

$$\begin{aligned} \mathbf{x}'V^{-1}\bar{\mathbf{x}}_{(1)} - \frac{1}{2}\bar{\mathbf{x}}_{(1)}'V^{-1}\bar{\mathbf{x}}_{(1)} &> \mathbf{x}'V^{-1}\bar{\mathbf{x}}_{(2)} - \frac{1}{2}\bar{\mathbf{x}}_{(2)}'V^{-1}\bar{\mathbf{x}}_{(2)} \\ \Leftrightarrow \mathbf{x}'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}) &> \frac{1}{2}(\bar{\mathbf{x}}_{(1)} + \bar{\mathbf{x}}_{(2)})'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}) \end{aligned}$$

Si les données sont centrées ($\bar{\mathbf{x}} = \mathbf{0}$) et les deux classes ont mêmes effectifs ($p_{(1)} = p_{(2)}$) :

$$\bar{\mathbf{x}} = p_{(1)}\bar{\mathbf{x}}_{(1)} + p_{(2)}\bar{\mathbf{x}}_{(2)} = p_{(1)}(\bar{\mathbf{x}}_{(1)} + \bar{\mathbf{x}}_{(2)}) = \mathbf{0}.$$

Ainsi, dans ce cas, on affecte \mathbf{x} à la classe 1 si $\mathbf{x}'V^{-1}(\bar{\mathbf{x}}_{(1)} - \bar{\mathbf{x}}_{(2)}) > 0$.

On obtient des critères analogues avec la matrice W^{-1} .

RÈGLE BAYSIENNE D'AFFECTION

Soit $\mathbf{x} = (x^1, \dots, x^p)'$ un nouvel individu. Alors, d'après le critère (7.3), on décide d'affecter \mathbf{x} à la classe 1, si

$$\begin{aligned} & \mathbf{x}'\Sigma^{-1}\mathbf{m}_{(1)} - \frac{1}{2}\mathbf{m}'_{(1)}\Sigma^{-1}\mathbf{m}_{(1)} + \ln \mathbb{P}[C_1] > \mathbf{x}'\Sigma^{-1}\mathbf{m}_{(2)} - \frac{1}{2}\mathbf{m}'_{(2)}\Sigma^{-1}\mathbf{m}_{(2)} + \ln \mathbb{P}[C_2] \\ \Leftrightarrow & \mathbf{x}'\Sigma^{-1}(\mathbf{m}_{(1)} - \mathbf{m}_{(2)}) - \frac{1}{2}(\mathbf{m}_{(1)} + \mathbf{m}_{(2)})'\Sigma^{-1}(\mathbf{m}_{(1)} - \mathbf{m}_{(2)}) > \ln \frac{\mathbb{P}[C_2]}{\mathbb{P}[C_1]} \end{aligned}$$

Lorsque $\mathbb{P}[C_2] = \mathbb{P}[C_1]$, que l'on estime $\mathbf{m}_{(\ell)}$ par $\bar{\mathbf{x}}_{(\ell)}$, Σ^{-1} par $\frac{n}{n-2}W^{-1}$, on retrouve le critère géométrique écrit avec la matrice W^{-1} .

BIBLIOGRAPHIE

- [Bar01] M. Bardos. *Analyse discriminante : application au risque et scoring financier*. Dunod, 2001.
- [EP08] B. Escofier and J. Pagès. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*. Dunod, 2008.
- [Ghe14] B. Ghermani. Notes de cours, 2014.
- [LPM06] L. Lebart, M. Piron, and A. Morineau. *Statistique exploratoire multidimensionnelle : Visualisations et inférences en fouille de données, 2006, 4eme édition*. Dunod, 2006.
- [Sap11] G. Saporta. *Probabilités, Analyse des données et Statistiques*,. 3 ème édition révisée, Paris, Editions Technip, 2011.
- [Ten07] M. Tenenhaus. *Statistique : Méthodes pour décrire, expliquer et prévoir*. Dunod, 2007.