# Analyses statistiques multivariées

Béatrice de Tilière

23 novembre 2009

# Table des matières

# Chapitre 1

# La Statistique

## 1.1 Généralités

*"La Statistique"* : méthode scientifique qui consiste à observer et à étudier une/plusieurs particularité(s) commune(s) chez un groupe de personnes ou de choses.

"La statistique" est à différencier d'"une statistique", qui est un nombre calculé à propos d'une population.

## 1.2 Un peu de vocabulaire

◦ *Population :* collection d'objets à étudier ayant des propriétés communes. Terme hérité des premières applications de la statistique qui concernait la démographie.

Exemple : ensemble de parcelles sur lesquelles on mesure un rendement, un groupe d'insectes...

◦ *Individu :* élément de la population étudiée.

Exemple : une des parcelles, un des insectes...

◦ *Variable :* propriété commune aux individus de la population, que l'on souhaite étudier. Elle peut être

- *qualitative :* couleur de pétales,
- *quantitative :* (numérique). Par exemple la taille, le poids, le volume. On distingue encore les variables
  - *continues* : toutes les valeurs d'un intervalle de $\mathbb{R}$ sont acceptables. Par exemple : le périmètre d'une coquille de moule.

- *discrètes* : seul un nombre discret de valeurs sont possibles. Par exemple : le nombre d'espèces ressencées sur une parcelle.

Les valeurs observées pour les variables s'appellent les *données*.

◦ *Echantillon* : partie étudiée de la population.

## 1.3 Collecte de données

La collecte de données (obtention de l'échantillon) est une étape clé, et délicate. Nous ne traitons pas ici des méthodes possibles, mais attirons l'attention sur le fait suivant.

Hypothèse sous-jacente en statistique : l'échantillon d'individus étudié est choisi au hasard parmi tous les individus qui auraient pu être choisi.

⇒ Tout mettre en oeuvre pour que ceci soit vérifié.

## 1.4 Deux directions en statistique

1. **Statistique descriptive**

   Elle a pour but de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses. Questions typiques :

   (a) Représentation graphique.

   (b) Paramètres de position, de dispersion, de relation.

   (c) Questions liées à des grands jeux de données.

2. **Statistique inférentielle**

   Les données ne sont pas considérées comme une information complète, mais une information partielle d'une population infinie. Il est alors naturel de supposer que les données sont des réalisations de variables aléatoires, qui ont une certaine loi de probabilité.

   Nécessite outils mathématiques plus pointus (théorie des probabilités).

   Questions typiques :

   (a) Estimation de paramètres.

   (b) Intervalles de confiance.

   (c) Tests d'hypothèse.

   (d) Modélisation : exemple (régression linéaire).

## 1.5 Statistique univariée / multivariée

Lorsque l'on observe une seule variable pour les individus de la population, on parle de *statistique univariée*, et de *statistique multivariée* lorsqu'on en observe au moins deux. Pour chacune des catégories, on retrouve les deux directions ci-dessus.

Exemple :
Univarié. Population : iris. Variable : longueur des pétales.
Multivarié. Population : iris. Variable 1 : longueur des pétales. Variable 2 : largeur des pétales.

## 1.6 Statistique descriptive multivariée et ce cours

Ce cours a pour thème la statistique descriptive dans le cas multivarié.

La statistique descriptive multivariée en général est un domaine très vaste. La première étape consiste à étudier la représentation graphique, et la description des paramètres de position, de dispersion et de relation. Ensuite, les méthodes principales se séparent en deux groupes.

1. Les méthodes factorielles (méthodes $R$, en anglais) : cherchent à réduire le nombre de variables en les résumant par un petit nombre de variables synthétiques. Selon que l'on travaille avec des variables quantitatives ou qualitatives, on utilisera l'*analyse en composantes principales*, ou l'*analyse de correspondance*. Les liens entre deux groupes de variables peuvent être traités grâce à l'*analyse canonique*.
2. Les méthodes de classification (méthodes $Q$, en anglais) : vise à réduire le nombre d'individus en formant des groupes homogènes.

Etant donné que ce cours ne dure que 5 semaines, nous ne traitons qu'un échantillon représentatif de méthodes. Nous avons choisi :

1. Paramètres de position, de dispersion, de relation.
2. Analyse en composantes principales (ACP).
3. Méthodes de classification.

# Chapitre 2

# Algèbre linéaire et représentation des vecteurs

Un des outils mathématique de base pour la statistique descriptive multivariée est l'algèbre linéaire. Ce chapitre consiste en quelques rappels qui seront utilisés dans la suite.

## 2.1 Matrices et vecteurs

• Une *matrice* $X$ est un tableau rectangulaire de nombres. On dit que $X$ est de *taille* $n \times p$, si $X$ a $n$ lignes et $p$ colonnes. Une telle matrice est représentée de la manière suivante :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

$x_{ij}$ est l'élément de la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne. On le note aussi $(X)_{ij}$

Exemple. $n = 3$, $p = 4$.

$$X = \begin{pmatrix} 3 & 4 & 8 & 2 \\ 2 & 6 & 1 & 1 \\ 1 & 1 & 0 & 5 \end{pmatrix}.$$

$$(X)_{22} = x_{22} = 6,$$
$$(X)_{13} = x_{13} = 8.$$

• La *transposée* de la matrice $X$, notée $X^t$, est obtenue à partir de $X$ en interchangeant les lignes et les colonnes. C'est à dire, $(X^t)_{ij} = x_{ji}$. Remarquer que si $X$ est une matrice de taille $n \times p$, alors $X^t$ est une matrice de taille $p \times n$.

EXEMPLE. Reprenant l'exemple ci-dessus, on a :

$$X^t = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \\ 8 & 1 & 0 \\ 2 & 1 & 5 \end{pmatrix}.$$

• Un *vecteur colonne* $\mathbf{x}$ est une matrice avec une seule colonne.

EXEMPLE.

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1.5 \\ 1 \end{pmatrix}.$$

Un *vecteur ligne* $\mathbf{x}^t$ est une matrice avec une seule ligne. Remarquer que la notation souligne le fait que c'est la transposée d'un vecteur colonne.

EXEMPLE.

$$\mathbf{x}^t = (1 \;\; 1.5 \;\; 1).$$

## 2.2 Opérations sur les matrices

Voici les opérations élémentaires définies sur les matrices.

• *Addition :* Soient $X$ et $Y$ deux matrices de même taille, disons $n \times p$. Alors la matrice $X + Y$ est de taille $n \times p$, et a pour coefficients :

$$(X + Y)_{ij} = x_{ij} + y_{ij}.$$

EXEMPLE.

$$X = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \\ 8 & 1 & 0 \\ 2 & 1 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 2 & 1 & 0 \\ 3 & 1 & 2 \\ 4 & 5 & 4 \\ 1 & 2 & 3 \end{pmatrix}, \quad X + Y = \begin{pmatrix} 5 & 3 & 1 \\ 7 & 7 & 3 \\ 12 & 6 & 4 \\ 3 & 3 & 8 \end{pmatrix}.$$

• *Multiplication par un scalaire :* Soit $X$ une matrice de taille $n \times p$, et $\lambda$ un nombre réel (aussi appelé *scalaire*), alors la matrice $\lambda X$ est de taille $n \times p$, et a pour coefficients :

$$(\lambda X)_{ij} = \lambda x_{ij}.$$

EXEMPLE.

$$X = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \\ 8 & 1 & 0 \\ 2 & 1 & 5 \end{pmatrix}, \quad \lambda = 2, \quad 2X = \begin{pmatrix} 6 & 4 & 2 \\ 8 & 12 & 2 \\ 16 & 2 & 0 \\ 4 & 2 & 10 \end{pmatrix}.$$

• *Multiplication de matrices :* Soit $X$ une matrice de taille $n \times p$, et $Y$ une matrice de taille $p \times q$, alors la matrice $XY$ est de taille $n \times q$, et a pour coefficients :

$$(XY)_{ij} = \sum_{k=1}^{p} x_{ik} y_{kj}.$$

EXEMPLE.

$$X = \begin{pmatrix} 3 & 2 & 1 \\ 4 & 6 & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 3 & 1 \\ 4 & 1 \\ 2 & 1 \end{pmatrix}, \quad XY = \begin{pmatrix} 19 & 6 \\ 38 & 11 \end{pmatrix}.$$

• La transposée vérifie les propriétés suivantes :

1. $(X + Y)^t = X^t + Y^t$
2. $(XY)^t = Y^t X^t$.

## 2.3 Interprétation géométrique des vecteurs

A priori, les matrices sont des tableaux de nombres. Il existe cependant une interprétation géométrique, qui va nous servir pour les statistiques multivariées. Les dessins correspondants sont faits au tableau pendant le cours.

• Interprétation géométrique des vecteurs

Un vecteur ligne de taille $1 \times n$, ou un vecteur colonne de taille $n \times 1$ représente un point de $\mathbb{R}^n$. La visualisation n'est possible que pour $n = 1, 2, 3$.

EXEMPLE. Le vecteur ligne $\mathbf{x}^t = (1 \;\; 2 \;\; 1)$ ou le vecteur colonne $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$, représente un point de $\mathbb{R}^3$.

• Projection orthogonale d'un point sur une droite

Pour étudier la projection orthogonale, on a besoin des définitions suivantes.

∘ Le *produit scalaire* de deux vecteurs $\mathbf{x}$, $\mathbf{y}$ de $\mathbb{R}^n$, noté $< \mathbf{x}, \mathbf{y} >$, est par définition :

$$< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i = (x_1 \cdots x_n) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{x}^t \mathbf{y}.$$

Deux vecteurs $\mathbf{x}$, $\mathbf{y}$ sont dits *orthogonaux*, si $< \mathbf{x}, \mathbf{y} >= 0$.

EXEMPLES.
Soient $\mathbf{x}^t = (1, 2, 3)$, $\mathbf{y}^t = (2, 3, 4)$, $\mathbf{z}^t = (1, 1, -1)$. Alors $< \mathbf{x}, \mathbf{y} >= 20$, et $< \mathbf{x}, \mathbf{z} >= 0$, donc $\mathbf{x}$ et $\mathbf{z}$ sont des vecteurs orthogonaux.

∘ La *norme* d'un vecteur $\mathbf{x}$, notée $\|\mathbf{x}\|$ est par définition :

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{< \mathbf{x}, \mathbf{x} >}.$$

Un vecteur de norme 1 est dit *unitaire*.

EXEMPLES.
Soient $\mathbf{x}^t = (1, 4, 5, 2)$, $\mathbf{y}^t = (1, 0, 0)$. Alors $\|\mathbf{x}\| = \sqrt{1 + 16 + 25 + 4} = \sqrt{46}$, et $\|\mathbf{y}\| = \sqrt{1 + 0 + 0} = 1$, donc $\mathbf{y}$ est un vecteur unitaire.

**Remarque 2.1**
– En dimension $n = 2$, on retrouve le théorème de Pythagore :

$$\|\mathbf{x}\| = \left\| \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\| = \sqrt{x_1^2 + x_2^2}.$$

– On peut montrer que : $< \mathbf{x}, \mathbf{y} >= \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cos \angle(\mathbf{x}, \mathbf{y})$.

∘ *Projection orthogonale d'un point sur une droite*

Soit $D$ une droite de $\mathbb{R}^n$ qui passe par l'origine, et soit $\mathbf{y}$ un vecteur directeur de cette droite. Alors la projection orthogonale $\tilde{\mathbf{x}}$ de $\mathbf{x}$ sur $D$ est donnée par :

$$\tilde{\mathbf{x}} = \frac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{y}\|} \frac{\mathbf{y}}{\|\mathbf{y}\|}.$$

La *coordonnée* de la projection $\tilde{\mathbf{x}}$ sur la droite $D$ est : $\dfrac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{y}\|}$. Remarquer que si $\mathbf{y}$ est un vecteur unitaire, alors la projection $\tilde{\mathbf{x}}$ est simplement donnée par : $\tilde{\mathbf{x}} =< \mathbf{x}, \mathbf{y} > \mathbf{y}$, et la coordonnée est alors le produit scalaire $< \mathbf{x}, \mathbf{y} >$.

EXEMPLE. Soit $D$ la droite de $\mathbb{R}^2$ passant par l'origine $(0, 0)$, et de vecteur directeur $\mathbf{y}^t = (1, 2)$. Soit $\mathbf{x}^t = (-1, 0)$. Alors la projection orthogonale $\tilde{\mathbf{x}}$ de $\mathbf{x}$ sur $D$ est :

$$\tilde{\mathbf{x}}^t = -\frac{1}{5}(1, 2) = -\left(\frac{1}{5}, \frac{2}{5}\right).$$

La coordonnée de $\tilde{\mathbf{x}}$ sur la droite $D$ est : $-1/\sqrt{5}$.

Preuve:

Le point $\tilde{\mathbf{x}}$ satisfait deux conditions :

1) Il appartient à la droite $D$, ce qui se traduit mathématiquement par, $\tilde{\mathbf{x}} = \lambda \mathbf{y}$, pour un $\lambda \in \mathbb{R}$.

2) Le vecteur $\mathbf{x} - \tilde{\mathbf{x}}$ est orthogonal au vecteur $\mathbf{y}$, ce qui se traduit par, $< \mathbf{x} - \tilde{\mathbf{x}}, \mathbf{y} >= 0$.

De 2) on obtient, $< \mathbf{x}, \mathbf{y} >=< \tilde{\mathbf{x}}, \mathbf{y} >$ . Utilisons maintenant 1) et remplaçons $\tilde{\mathbf{x}}$ par $\lambda \mathbf{y}$. On a alors :

$$< \mathbf{x}, \mathbf{y} >=< \lambda \mathbf{y}, \mathbf{y} >= \lambda \|\mathbf{y}\|^2,$$

d'où $\lambda = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{y}\|^2}$. On conclut en utilisant 1), et en remplaçant $\lambda$ par la valeur trouvée. □

# Chapitre 3

# Statistique descriptive élémentaire

Ce chapitre est illustré au tableau au moyen d'un petit jeu de données. Un jeu de données plus grand est traité dans l'Annexe A.

## 3.1 La matrice des données

Avant de pouvoir analyser les données, il faut un moyen pour les répertorier. L'outil naturel est d'utiliser une *matrice* $X$, appelée *matrice des données*. Nous nous restreignons au cas où les données sont de type *quantitatif*, ce qui est fréquent en biologie.

On suppose que l'on a $n$ individus, et que pour chacun de ces individus, on observe $p$ variables. Alors, les données sont répertoriées de la manière suivante :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

L'élément $x_{ij}$ de la matrice $X$ représente l'observation de la $j^{\text{ème}}$ variable pour l'individu $i$.

On va noter la $i^{\text{ème}}$ ligne de $X$, représentant les données de toutes les variables pour le $i^{\text{ème}}$ individu, par $X_i^t$. On va noter la $j^{\text{ème}}$ colonne de $X$, représentant les données de la $j^{\text{ème}}$ variable pour tous les individus, par $X_{(j)}$. Ainsi,

$$X_i^t = (x_{i1}, \cdots, x_{ip}).$$

$$X_{(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

On peut considérer cette matrice de deux points de vue différents : si l'on compare deux colonnes, alors on étudie la relation entre les deux variables correspondantes. Si par contre, on compare deux lignes, on étudie la relation entre deux individus.

EXEMPLE. Voici des données représentant les résultats de 6 individus à un test de statistique (variable 1) et de géologie (variable 2).

$$X = \begin{pmatrix} 11 & 13.5 \\ 12 & 13.5 \\ 13 & 13.5 \\ 14 & 13.5 \\ 15 & 13.5 \\ 16 & 13.5 \end{pmatrix}$$

Remarquer que lorsque $n$ et $p$ deviennent grands, ou moyennement grand, le nombre de données $np$ est grand, de sorte que l'on a besoin de techniques pour résumer et analyser ces données.

## 3.2 Paramètres de position

Les quantité ci-dessous sont des généralisations naturelles du cas uni-dimensionnel. Soit $X_{(j)}$ les données de la $j^{\text{ème}}$ variable pour les $n$ individus.

### 3.2.1 Moyenne arithmétique

La *moyenne arithmétique* des données $X_{(j)}$ de la $j^{\text{ème}}$ variable, notée $\overline{X_{(j)}}$, est :

$$\overline{X_{(j)}} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

On peut représenter les $p$ moyennes arithmétiques des données des $p$ variables sous la forme du *vecteur ligne des moyennes arithmétiques*, noté $\overline{\mathbf{x}}^t$ :

$$\overline{\mathbf{x}}^t = (\overline{X_{(1)}}, \cdots, \overline{X_{(p)}}).$$

EXEMPLE. Le vecteur ligne des moyennes arithmétiques pour l'exemple des notes est :

$$\overline{\mathbf{x}}^t = \left( \frac{11 + \cdots + 16}{6}, \frac{13.5 + \cdots + 13.5}{6} \right) = (13.5, 13.5).$$

### 3.2.2 Médiane

On suppose que les valeurs des données $X_{(j)}$ de la $j^{\text{ème}}$ variable sont classées en ordre croissant. Alors, lorsque $n$ est impair, la *médiane*, notée $m_{(j)}$, est l'"élément du milieu", c'est à dire :

$$m_{(j)} = x_{\frac{n+1}{2}j}.$$

Si $n$ est pair, on prendra par convention :

$$m_{(j)} = \frac{x_{\frac{n}{2},j} + x_{\frac{n}{2}+1,j}}{2}.$$

On peut aussi mettre les $p$ médianes dans un vecteur ligne, noté $\mathbf{m}^t$, et appelé le *vecteur ligne des médianes* :

$$\mathbf{m}^t = (m_{(1)}, \cdots, m_{(p)}).$$

EXEMPLE. Le vecteur ligne des médianes pour l'exemple des notes est :

$$\mathbf{m}^t = \left(\frac{13+14}{2}, \frac{13.5+13.5}{2}\right) = (13.5, 13.5).$$

## 3.3 Paramètres de dispersion

La moyenne ne donne qu'une information partielle. En effet, il est aussi important de pouvoir mesurer combien ces données sont dispersées autour de la moyenne. Revenons à l'exemple des notes, les données des deux variables ont la même moyenne, mais vous sentez bien qu'elles sont de nature différente. Il existe plusieurs manières de mesurer la dispersion des données.

### 3.3.1 Etendue

Soit $X_{(j)}$ les données de la $j^{\text{ème}}$ variable, alors l'*étendue*, notée $w_{(j)}$, est la différence entre la donnée la plus grande pour cette variable, et la plus petite. Mathématiquement, on définit :

$$X_{(j)}^{\max} = \max_{i\in\{1,\cdots,n\}} x_{ij}.$$
$$X_{(j)}^{\min} = \min_{i\in\{1,\cdots,n\}} x_{ij}.$$

Alors

$$w_{(j)} = X_{(j)}^{\max} - X_{(j)}^{\min}.$$

On peut représenter les $p$ étendues sous la forme d'un vecteur ligne, appelé *vecteur ligne des étendues*, et noté $\mathbf{w}^t$ :

$$\mathbf{w}^t = (w_{(1)}, \cdots, w_{(p)}).$$

EXEMPLE. Le vecteur ligne des étendues de l'exemple des notes est :

$$\mathbf{w}^t = (5, 0).$$

**Remarque 3.1** C'est un indicateur instable étant donné qu'il ne dépend que des valeurs extrêmes. En effet, vous pouvez avoir un grand nombre de données qui sont similaires, mais qui ont une plus grande et plus petite valeur qui sont très différentes, elles auront alors une étendue très différente, mais cela ne représente pas bien la réalité des données.

### 3.3.2 Variance et écart-type

Une autre manière de procéder qui tient compte de toutes les données, et non pas seulement des valeurs extrêmes, est la suivante.

On considère les données $X_{(j)}$ de la $j^{\text{ème}}$ variable, l'idée est de calculer la somme, pour chacune des données de cette variable, des distance à la moyenne, et de diviser par le nombre de données. Une première idée serait de calculer :

$$\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_{(j)}}) = \frac{1}{n}\left[(x_{1j} - \overline{X_{(j)}}) + \cdots + (x_{nj} - \overline{X_{(j)}})\right],$$

mais dans ce cas là, il y a des signes $+$ et $-$ qui se compensent et faussent l'information. En effet, reprenons l'exemple de la variable 1 ci-dessus. Alors la quantité ci-dessus est :

$$\frac{1}{6}\left[(11 - 13.5) + (12 - 13.5) + (13 - 13.5) + (14 - 13.5) + (15 - 13.5) + (16 - 13.5)\right] = 0,$$

alors qu'il y a une certaine dispersion autour de la moyenne. Pour palier à la compensation des signes, il faut rendre toutes les quantités que l'on somme de même signe, disons positif. Une idée est de prendre la valeur absolue, et on obtient alors l'*écart à la moyenne*. Une autre manière de procéder est de prendre les carrés, on obtient alors la *variance* :

$$\mathrm{Var}(X_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_{(j)}})^2 = \frac{1}{n}\left[(x_{1j} - \overline{X_{(j)}})^2 + \cdots + (x_{nj} - \overline{X_{(j)}})^2\right].$$

Pour compenser le fait que l'on prenne des carrés, on peut reprendre la racine, et on obtient alors l'*écart-type* :

$$\sigma(X_{(j)}) = \sqrt{\mathrm{Var}(X_{(j)})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X_{(j)}})^2}.$$

EXEMPLE. Voici le calcul des variances et des écart-types pour l'exemple des notes.

$$\begin{aligned}\mathrm{Var}(X_{(1)}) &= \frac{1}{6}\left((11-13.5)^2+(12-13.5)^2+\cdots+(16-13.5)^2\right) = 2.917\\ \sigma(X_{(1)}) &= 1.708\\ \mathrm{Var}(X_{(2)}) &= \frac{1}{6}\left(6(13.5-13.5)^2\right) = 0\\ \sigma(X_{(2)}) &= 0.\end{aligned}$$

• Notation matricielle

La variance s'écrit naturellement comme la norme d'un vecteur. Cette interprétation géométrique est utile pour la suite.

On définit la matrice des moyennes arithmétiques, notée $\overline{X}$, par :

$$\overline{X} = \begin{pmatrix} \overline{X_{(1)}} & \cdots & \overline{X_{(p)}} \\ \vdots & \vdots & \vdots \\ \overline{X_{(1)}} & \cdots & \overline{X_{(p)}} \end{pmatrix},$$

alors la matrice $X - \overline{X}$ est :

$$X - \overline{X} = \begin{pmatrix} x_{11} - \overline{X_{(1)}} & \cdots & x_{1p} - \overline{X_{(p)}} \\ \vdots & \vdots & \vdots \\ x_{n1} - \overline{X_{(1)}} & \cdots & x_{np} - \overline{X_{(p)}} \end{pmatrix}.$$

Et donc la variance des données $X_{(j)}$ de la $j^{\text{ème}}$ variable est égale à $1/n$ fois le produit scalaire de la $j^{\text{ème}}$ colonne avec elle-même ; autrement dit à $1/n$ fois la norme au carré du vecteur donné par la $j^{\text{ème}}$ colonne. Mathématiquement, on écrit ceci ainsi :

$$\mathrm{Var}(X_{(j)}) = \frac{1}{n} < (X-\overline{X})_{(j)}, (X-\overline{X})_{(j)} >= \frac{1}{n}((X-\overline{X})_{(j)})^t(X-\overline{X})_{(j)} = \frac{1}{n}\|(X-\overline{X})_{(j)}\|^2.$$

De manière analogue, l'écart type s'écrit :

$$\sigma(X_{(j)}) = \frac{1}{\sqrt{n}}\|(X-\overline{X})_{(j)}\|.$$

EXEMPLE. Réécrivons la variance pour l'exemple des notes en notation matricielle.

$$\overline{X} = \begin{pmatrix} 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \\ 13.5 & 13.5 \end{pmatrix}, \text{ et } \qquad X - \overline{X} = \begin{pmatrix} -2.5 & 0 \\ -1.5 & 0 \\ -0.5 & 0 \\ 0.5 & 0 \\ 1.5 & 0 \\ 2.5 & 0 \end{pmatrix}.$$

Ainsi :

$$\mathrm{Var}(X_{(1)}) = \frac{1}{6}\left\langle \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix}, \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix} \right\rangle = \frac{1}{6}\left\| \begin{pmatrix} -2.5 \\ -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \\ 2.5 \end{pmatrix} \right\|^2 = 2.917.$$

De manière analogue, on trouve $\mathrm{Var}(X_{(2)}) = 0$.

### 3.3.3 Variables centrées-réduites

Les données d'une variable sont dites *centrées* si on leur soustrait leur moyenne. Elles sont dites *centrées réduites* si elles sont centrées et divisées par leur écart-type. Les données d'une variable centrées réduites sont utiles car elles n'ont plus d'unité, et des données de variables différentes deviennent ainsi comparables.

Si $X$ est la matrice des données, on notera $Z$ la matrice des données centrées réduites. Par définition, on a :

$$(Z)_{ij} = z_{ij} = \frac{x_{ij} - \overline{X_{(j)}}}{\sigma(X_{(j)})}.$$

Remarquer que si $\sigma(X_{(j)})$ est nul la quantité ci-dessus n'est pas bien définie. Mais dans ce cas, on a aussi $x_{ij} - \overline{X_{(j)}} = 0$ pour tout $i$, de sorte que l'on pose $z_{ij} = 0$.

**Exemple 3.1** Voici la matrice des données centrées réduites de l'exemple des notes. On se souvient que

$$\begin{aligned} \sigma(X_{(1)}) = 1.708 \quad \sigma(X_{(2)}) = 0 \\ \overline{X_{(1)}} = 13.5 \quad \overline{X_{(2)}} = 13.5. \end{aligned}$$

Ainsi,

$$Z = \begin{pmatrix} \frac{11-13.5}{1.708} & 0 \\ \frac{12-13.5}{1.708} & 0 \\ \frac{13-13.5}{1.708} & 0 \\ \frac{14-13.5}{1.708} & 0 \\ \frac{15-13.5}{1.708} & 0 \\ \frac{16-13.5}{1.708} & 0 \end{pmatrix} = \begin{pmatrix} -1.464 & 0 \\ -0.878 & 0 \\ -0.293 & 0 \\ 0.293 & 0 \\ 0.878 & 0 \\ 1.464 & 0 \end{pmatrix}.$$

## 3.4 Paramètres de relation entre deux variables

Après la description uni-dimensionnelle de la matrice des données, on s'intéresse à la liaison qu'il existe entre les données des différentes variables. Nous les comparons deux à deux.

Rappelons le contexte général. Nous avons les données $X_{(1)}, \cdots, X_{(p)}$ de $p$ variables observées sur $n$ individus.

### 3.4.1 Covariance

Pour tout $i$ et $j$ compris entre 1 et $p$, on définit la *covariance* entre les données $X_{(i)}$ et $X_{(j)}$ des $i^{\text{ème}}$ et $j^{\text{ème}}$ variables, notée $\text{Cov}(X_{(i)}, X_{(j)})$, par :

$$\text{Cov}(X_{(i)}, X_{(j)}) = \frac{1}{n} < (X - \overline{X})_{(i)}, (X - \overline{X})_{(j)} >= \frac{1}{n}((X - \overline{X})_{(i)})^t (X - \overline{X})_{(j)}.$$

**Théorème 1 (Köning-Huygens)** *La covariance est égale à :*

$$\text{Cov}(X_{(i)}, X_{(j)}) = \left(\frac{1}{n} < X_{(i)}, X_{(j)} >\right) - \overline{X_{(i)}}\,\overline{X_{(j)}}.$$

Preuve:
Par définition de la matrice $\overline{X}$, nous avons $\overline{X}_{(i)} = \overline{X_{(i)}}\,\mathbf{1}$, où $\overline{X_{(i)}}$ est la moyenne des données de la $i^{\text{ème}}$ variable, et $\mathbf{1}$ est le vecteur de taille $n \times 1$, formé de 1. Utilisant la bilinéarité du produit scalaire, nous obtenons :

$$\begin{aligned}
\text{Cov}(X_{(i)}, X_{(j)}) &= \frac{1}{n} < X_{(i)} - \overline{X}_{(i)}, X_{(j)} - \overline{X}_{(j)} > \\
&= \frac{1}{n} < X_{(i)} - \overline{X_{(i)}}\,\mathbf{1}, X_{(j)} - \overline{X_{(j)}}\,\mathbf{1} > \\
&= \frac{1}{n}\left[< X_{(i)}, X_{(j)} > - \overline{X_{(i)}} < \mathbf{1}, X_{(j)} > - \overline{X_{(j)}} < X_{(i)}, \mathbf{1} > + \overline{X_{(i)}}\,\overline{X_{(j)}} < \mathbf{1}, \mathbf{1} >\right] \\
&= \frac{1}{n}\left[< X_{(i)}, X_{(j)} > - n\overline{X_{(i)}}\,\overline{X_{(j)}} - n\overline{X_{(j)}}\,\overline{X_{(i)}} + n\overline{X_{(i)}}\,\overline{X_{(j)}}\right] \\
&\qquad (\text{car} < \mathbf{1}, X_{(j)} >= n\overline{X_{(j)}}), < X_{(i)}, \mathbf{1} >= n\overline{X_{(i)}}, \text{ et } < \mathbf{1}, \mathbf{1} >= n) \\
&= \left(\frac{1}{n} < X_{(i)}, X_{(j)} >\right) - \overline{X_{(i)}}\,\overline{X_{(j)}}.
\end{aligned}$$

□

**Remarque 3.2**

1. $\mathrm{Cov}(X_{(i)}, X_{(j)}) = \frac{1}{n}((X - \overline{X})^t(X - \overline{X}))_{ij}$, c'est à dire $\mathrm{Cov}(X_{(i)}, X_{(j)})$ est le coefficient $(i, j)$ de la matrice $\frac{1}{n}(X - \overline{X})^t(X - \overline{X})$.
2. $\mathrm{Cov}(X_{(i)}, X_{(i)}) = \mathrm{Var}(X_{(i)})$.
3. La covariance est symétrique, i.e. : $\mathrm{Cov}(X_{(i)}, X_{(j)}) = \mathrm{Cov}(X_{(j)}, X_{(i)})$.
4. Dans le cas de la variance, le Théorème de Köning-Huygens s'écrit :

$$\mathrm{Var}(X_{(j)}) = \left(\frac{1}{n}\|X_{(j)}\|^2\right) - \overline{X_{(j)}}^2.$$

Exemple. Calculons la covariance entre les données des première et deuxième variables de l'exemple des notes, en utilisant le Théorème de König-Huygens :

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) = \frac{1}{6}[11 \cdot 13.5 + 12 \cdot 13.5 + 13 \cdot 13.5 + \cdots + 16 \cdot 13.5] - 13.5^2 = 0.$$

• Matrice de covariance

Les variances et covariances sont naturellement répertoriées dans la *matrice de covariance des données* $X$, de taille $p \times p$, notée $V(X)$, définie par :

$$V(X) = \frac{1}{n}(X - \overline{X})^t(X - \overline{X}).$$

De sorte que l'on a

$$\mathrm{Cov}(X_{(i)}, X_{(j)}) = (V(X))_{ij}.$$

Remarquer que les coefficients sur la diagonale de la matrice $V(X)$ donnent les variances.

Exemple. Calculons la matrice de covariance pour l'exemple des notes.

$$\begin{aligned} V(X) &= \frac{1}{6}(X - \overline{X})^t(X - \overline{X}) \\ &= \frac{1}{6}\begin{pmatrix} -2.5 & -1.5 & -0.5 & 0.5 & 1.5 & 2.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -2.5 & 0 \\ -1.5 & 0 \\ -0.5 & 0 \\ 0.5 & 0 \\ 1.5 & 0 \\ 2.5 & 0 \end{pmatrix} = \begin{pmatrix} 2.91667 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Ainsi, on retrouve :

$$\begin{aligned} \mathrm{Var}(X_{(1)}) = (V(X))_{11} = 2.917, \quad \mathrm{Var}(X_{(2)}) = (V(X))_{22} = 0 \\ \mathrm{Cov}(X_{(1)}, X_{(2)}) = \mathrm{Cov}(X_{(2)}, X_{(1)}) = (V(X))_{12} = (V(X))_{21} = 0. \end{aligned}$$

• La *variabilité totale* de la matrice des données $X$, est par définition :

$$\mathrm{Tr}(V(X)) = \sum_{i=1}^{p} \mathrm{Var}(X_{(i)}).$$

Cette quantité est importante car elle donne en quelque sorte la quantité d'information qui est contenue dans la matrice $X$. Elle joue un rôle clé dans l'ACP.

### 3.4.2 Corrélation de Bravais-Pearson

La *corrélation de Bravais-Pearson* entre les données $X_{(i)}$ et $X_{(j)}$ des $i^{\text{ème}}$ et $j^{\text{ème}}$ variables, notée $r(X_{(i)}, X_{(j)})$, est par définition :

$$\begin{aligned} r(X_{(i)}, X_{(j)}) &= \frac{\mathrm{Cov}(X_{(i)}, X_{(j)})}{\sigma(X_{(i)})\sigma(X_{(j)})} = \frac{< (X-\overline{X})_{(i)}, (X-\overline{X})_{(j)} >}{\|(X-\overline{X})_{(i)}\| \cdot \|(X-\overline{X})_{(j)}\|} \\ &= \cos\angle\left((X-\overline{X})_{(i)}, (X-\overline{X})_{(j)}\right). \end{aligned}$$

**Proposition 2** *La corrélation de Bravais-Pearson satisfait les propriétés :*

1. $r(X_{(i)}, X_{(i)}) = 1$,
2. $|r(X_{(i)}, X_{(j)})| \leq 1$,
3. $|r(X_{(i)}, X_{(j)})| = 1$, *si et seulement si il existe un nombre* $a \in \mathbb{R}$, *tel que*

$$(X-\overline{X})_{(j)} = a(X-\overline{X})_{(i)}.$$

Preuve:

Pour le point 1, il suffit d'écrire :

$$r(X_{(i)}, X_{(i)}) = \frac{< (X-\overline{X})_{(i)}, (X-\overline{X})_{(i)} >}{\|(X-\overline{X})_{(i)}\| \cdot \|(X-\overline{X})_{(i)}\|} = \frac{\|(X-\overline{X})_{(i)}\|^2}{\|(X-\overline{X})_{(i)}\|^2} = 1.$$

Le point 2 est une conséquence du fait que la corrélation est un cosinus. De plus le cosinus de l'angle formé par deux vecteurs vaut 1 en valeur absolue, ssi ces deux vecteurs sont colinéaires, ce qui est exactement la condition donnée au point 3. □

• Conséquences

1. Le Point 3 s'écrit en composantes :

$$x_{1j} - \overline{X_{(j)}} = a(x_{1i} - \overline{X_{(i)}}), \cdots, x_{nj} - \overline{X_{(j)}} = a(x_{ni} - \overline{X_{(i)}}),$$

ainsi $|r(X_{(i)}, X_{(j)})| = 1$, si et seulement si il y a une dépendance linéaire entre les données $X_{(i)}$ et $X_{(j)}$ des $i^{\text{ème}}$ et $j^{\text{ème}}$ variables. Voir le dessin fait au tableau.

2. Si la corrélation est proche de 1, cela implique une relation linéaire entre les données, mais pas forcément une *causalité*. Ces deux phénomènes peuvent être reliés entre eux par une troisième variable, non mesurée qui est la cause des deux. Par exemple, le nombre de coups de soleil observés dans une station balnéaire peut être fortement corrélé au nombre de lunettes de soleil vendues ; mais aucun des deux phénomènes n'est la cause de l'autre.
3. Si la corrélation est proche de 0, cela ne signifie pas qu'il n'y a pas de relation entre les données des variables, cela veut seulement dire qu'il n'y a pas de relation *linéaire*. Elle pourrait par exemple être quadratique, ou autre.

• Matrice de corrélation

De manière analogue à la matrice de covariance, on définit la *matrice de corrélation*, de taille $p \times p$, notée $R(X)$, par :

$$(R(X))_{ij} = r(X_{(i)}, X_{(j)}).$$

Remarquer que les éléments diagonaux de cette matrice sont tous égaux à 1.

Exemple. La matrice de corrélation de l'exemple des notes est :

$$R(X) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

# Chapitre 4

# Analyse en Composantes Principales (ACP)

Méthode factorielle, ou de type $R$ (en anglais). A pour but de réduire le nombre de variables en perdant le moins d'information possible, c'est à dire en gardant le maximum de la variabilité totale.

Pratiquement, cela revient à projeter les données des variables pour les individus sur un espace de dimension inférieure en maximisant la variabilité totale des nouvelles variables. On impose que l'espace sur lequel on projète soit orthogonal (pour ne pas avoir une vision déformée des données).

## 4.1 Etape 1 : Changement de repère

Soit $X$ la matrice des données. Pour plus de visibilité, on considère la matrice des données centrées $X-\overline{X}$. Le $i^{\text{ème}}$ vecteur ligne $(X-\overline{X})_i^t$ représente les données de toutes les variables pour le $i^{\text{ème}}$ individu. Pour simplifier les notations, on écrit $\mathbf{x}^t = (X-\overline{X})_i^t$.

• Représentation graphique du $i^{\text{ème}}$ individu

On peut représenter $\mathbf{x}^t$ par un point de $\mathbb{R}^p$. Alors,

- chacun des axes de $\mathbb{R}^p$ représente une des $p$ variables,
- les coordonnées de $\mathbf{x}^t$ sont les données des $p$ variables pour le $i^{\text{ème}}$ individu.

• Nouveau repère

Soient $\mathbf{q}_1, \cdots, \mathbf{q}_p$, $p$ vecteurs de $\mathbb{R}^p$, unitaires et deux à deux orthogonaux. On considère les $p$ droites passant par l'origine, de vecteur directeur $\mathbf{q}_1, \cdots, \mathbf{q}_p$ respectivement. Alors

ces droites définissent un nouveau repère. Chacun des axes représente une nouvelle variable, qui est combinaison linéaire des anciennes variables.

• Changement de repère pour le $i^{\text{ème}}$ individu

On souhaite exprimer les données du $i^{\text{ème}}$ individu dans ce nouveau repère. Autrement dit, on cherche à déterminer les nouvelles coordonnées du $i^{\text{ème}}$ individu. Pour $j = 1, \cdots, p$, la coordonnée sur l'axe $\mathbf{q}_j$ est la coordonnée de la projection orthogonale de $\mathbf{x}$ sur la droite passant par l'origine et de vecteur directeur $\mathbf{q}_j$. Elle est donnée par (voir Chapitre 2) :

$$< \mathbf{x}, \mathbf{q}_j >= \mathbf{x}^t \mathbf{q}_j.$$

Ainsi les coordonnées des données du $i^{\text{ème}}$ individu dans ce nouveau repère sont répertoriées dans le vecteur ligne :

$$(\mathbf{x}^t \mathbf{q}_1 \ \cdots \ \mathbf{x}^t \mathbf{q}_p) = \mathbf{x}^t Q = (X - \overline{X})_i^t Q,$$

où $Q$ est la matrice de taille $p \times p$, dont les colonnes sont les vecteurs $\mathbf{q}_1, \cdots, \mathbf{q}_p$. Cette matrice est *orthonormale*, *i.e.* ses vecteurs colonnes sont unitaires et deux à deux orthogonaux.

• Changement de repère pour tous les individus

On souhaite faire ceci pour les données de tous les individus $(X - \overline{X})_1^t, \cdots, (X - \overline{X})_n^t$. Les coordonnées dans le nouveau repère sont répertoriées dans la matrice :

$$Y = (X - \overline{X})Q. \tag{4.1}$$

En effet, la $i^{\text{ème}}$ ligne de $Y$ est $(X - \bar{X})_i^t Q$, qui représente les coordonnées dans le nouveau repère des données du $i^{\text{ème}}$ individu.

## 4.2 Etape 2 : Choix du nouveau repère

Le but est de trouver un nouveau repère $\mathbf{q}_1, \cdots, \mathbf{q}_p$, tel que la quantité d'information expliquée par $\mathbf{q}_1$ soit maximale, puis celle expliquée par $\mathbf{q}_2$, etc... On peut ainsi se limiter à ne garder que les 2-3 premiers axes. Afin de réaliser ce programme, il faut d'abord choisir une mesure de la quantité d'information expliquée par un axe, puis déterminer le repère qui optimise ces critères.

### 4.2.1 Mesure de la quantité d'information

La variance des données centrées $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ variable représente la dispersion des données autour de leur moyenne. Plus la variance est grande, plus les données de cette variable sont dispersées, et plus la quantité d'information apportée est importante.

La quantité d'information contenue dans les données $(X - \overline{X})$ est donc la somme des variances des données de toutes les variables, c'est à dire la *variabilité totale* des données $(X - \overline{X})$, définie à la Section 3.4.1 :

$$\sum_{j=1}^{p} \mathrm{Var}((X - \overline{X})_{(j)}) = \mathrm{Tr}(V(X - \overline{X})) = \mathrm{Tr}(V(X)).$$

La dernière égalité vient du fait que $V(X - \overline{X}) = V(X)$. Etudions maintenant la variabilité totale des données $Y$, qui sont la projection des données $X - \overline{X}$ dans le nouveau repère défini par la matrice orthonormale $Q$. Soit $V(Y)$ la matrice de covariance correspondante, alors :

**Lemme 3**

1. $V(Y) = Q^t V(X) Q$

2. *La variabilité totale des données $Y$ est la même que celle des données $X - \overline{X}$.*

Preuve:

$$\begin{aligned} V(Y) &= \frac{1}{n}(Y - \overline{Y})^t (Y - \overline{Y}) \\ &= \frac{1}{n} Y^t Y, \text{ (car } \overline{Y} \text{ est la matrice nulle)} \\ &= \frac{1}{n}((X - \overline{X})Q)^t (X - \overline{X})Q \text{ (par (4.1))} \\ &= \frac{1}{n} Q^t (X - \overline{X})^t (X - \overline{X}) Q \text{ (propriété de la transposée)} \\ &= Q^t V(X) Q. \end{aligned}$$

Ainsi, la variabilité totale des nouvelles données $Y$ est :

$$\begin{aligned} \mathrm{Tr}(V(Y)) &= \mathrm{Tr}(Q^t V(X) Q) = \mathrm{Tr}(Q^t Q V(X)), \text{ (propriété de la trace)} \\ &= \mathrm{Tr}(V(X)) \text{ (car } Q^t Q = \mathrm{Id}, \text{ étant donné que la matrice } Q \text{ est orthonormale).} \end{aligned}$$

□

### 4.2.2 Choix du nouveau repère

Etant donné que la variabilité totale des données projetées dans le nouveau repère est la même que celle des données d'origine $X - \overline{X}$, on souhaite déterminer $Q$ de sorte que

la part de la variabilité totale expliquée par les données $Y_{(1)}$ de la nouvelle variable $\mathbf{q}_1$ soit maximale, puis celle expliquée par les données $Y_{(2)}$ de la nouvelle variable $\mathbf{q}_2$, etc... Autrement dit, on souhaite résoudre le problème d'optimisation suivant :

$$\boxed{\begin{array}{c}\text{Trouver une matrice orthonormale } Q \text{ telle que} \\ \text{Var}(Y_{(1)}) \text{ soit maximale, puis Var}(Y_{(2)}), \text{ etc...}\end{array}} \tag{4.2}$$

Avant d'énoncer le Théorème donnant la matrice $Q$ optimale, nous avons besoin de nouvelles notions d'algèbre linéaire.

• Théorème spectral pour les matrices symétriques

Soit $A$ une matrice de taille $p \times p$. Un vecteur $\mathbf{x}$ de $\mathbb{R}^p$ s'appelle un *vecteur propre* de la matrice $A$, s'il existe un nombre $\lambda$ tel que :

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Le nombre $\lambda$ s'appelle la *valeur propre* associée au vecteur propre $\mathbf{x}$.

Une matrice carrée $A = (a_{ij})$ est dite *symétrique*, ssi $a_{ij} = a_{ji}$ pour tout $i, j$.

**Théorème 4 (spectral pour les matrices symétriques)** *Si $A$ est 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 exite une matrice orthonormale $Q$ telle que :*

$$Q^t A Q = D$$

*et $D$ est la matrice diagonale formée des valeurs propres de $A$.*

• Théorème fondamental de l'ACP

Soit $X - \overline{X}$ la matrice des données centrées, et soit $V(X)$ la matrice de covariance associée (qui est symétrique par définition). On note $\lambda_1 \geq \cdots \geq \lambda_p$ les valeurs propres de la matrice $V(X)$. Soit $Q$ la matrice orthonormale correspondant à la matrice $V(X)$, donnée par le Théorème 4, telle que le premier vecteur corresponde à la plus grande valeur propre, etc...Alors, le théorème fondamental de l'ACP est :

**Théorème 5** *La matrice orthonormale qui résoud le problème d'optimisation* (4.2) *est la matrice $Q$ décrite ci-dessus. De plus, on a :*

1. $\text{Var}(Y_{(j)}) = \lambda_j$,

2. $\mathrm{Cov}(Y_{(i)}, Y_{(j)}) = 0$, *quand* $i \neq j$,
3. $\mathrm{Var}(Y_{(1)}) \geq \cdots \geq \mathrm{Var}(Y_{(p)})$,

Les colonnes $\mathbf{q}_1, \cdots, \mathbf{q}_p$ de la matrice $Q$ décrivent les nouvelles variables, appelées les *composantes principales*.

Preuve:
On a :

$$\begin{aligned} V(Y) &= Q^t V(X) Q, \quad \text{(par le Lemme 3)} \\ &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & & \lambda_{p-1} & 0 \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}, \quad \text{(par le Théorème 4)} \end{aligned}$$

Ainsi,

$$\begin{aligned} \mathrm{Var}(Y_{(j)}) &= (V(Y))_{jj} = (Q^t V Q)_{jj} = \lambda_j \\ \mathrm{Cov}(Y_{(i)}, Y_{(j)}) &= (V(Y))_{ij} = (Q^t V Q)_{ij} = 0. \end{aligned}$$

Ceci démontre 1 et 2. Le point 3 découle du fait que l'on a ordonné les valeurs propres en ordre décroissant.

Le dernier point non-trivial à vérifier est l'optimalité. C'est à dire que pour toute autre matrice orthonormale choisie, la variance des données de la première variable serait plus petite que $\lambda_1$, etc... Même si ce n'est pas très difficile, nous choisissons de ne pas démontrer cette partie ici. □

## 4.3 Conséquences

Voici deux conséquences importantes du résultat que nous avons établi dans la section précédente.

• Restriction du nombre de variables

Le but de l'ACP est de restreindre le nombre de variables. Nous avons déterminé ci-dessus des nouvelles variables $\mathbf{q}_1, \cdots, \mathbf{q}_p$, les *composantes principales*, qui sont optimales. La part de la variabilité totale expliquée par les données $Y_{(1)}, \cdots, Y_{(k)}$ des $k$

premières nouvelles variables ($k \leq p$), est :

$$\frac{\mathrm{Var}(Y_{(1)}) + \cdots + \mathrm{Var}(Y_{(k)})}{\mathrm{Var}(Y_{(1)}) + \cdots + \mathrm{Var}(Y_{(p)})} = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}.$$

Dans la pratique, on calcule cette quantité pour $k = 2$ ou 3. En multipliant par 100, ceci donne le pourcentage de la variabilité totale expliquée par les données des 2 ou 3 premières nouvelles variables. Si ce pourcentage est raisonable, on choisira de se restreindre aux 2 ou 3 premiers axes. La notion de raisonable est discutable. Lors du TP, vous choisirez 30%, ce qui est faible (vous perdez 70% de l'information), il faut donc être vigilant lors de l'analyse des résultats.

• Corrélation entre les données des anciennes et des nouvelles variables

Etant donné que les nouvelles variables sont dans un sens "artificielles", on souhaite comprendre la corrélation entre les données $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ ancienne variable et celle $Y_{(k)}$ de la $k^{\text{ème}}$ nouvelle variable. La matrice de covariance $V(X, Y)$ de $X - \overline{X}$ et $Y$ est donnée par :

$$\begin{aligned} V(X,Y) &= \frac{1}{n}(X - \overline{X})^t(Y - \overline{Y}) \\ &= \frac{1}{n}(X - \overline{X})^t(Y - \overline{Y}), \text{ (car } \overline{Y} \text{ est la matrice nulle)} \\ &= \frac{1}{n}(X - \overline{X})^t(X - \overline{X})Q, \text{ (par définition de la matrice } Y) \\ &= Q(Q^t V(X) Q) \text{ (car } Q^t Q = \mathrm{Id}) \\ &= QD, \text{ (par le Théorème spectral, où } D \text{ est la matrice des valeurs propres).} \end{aligned}$$

Ainsi :

$$\mathrm{Cov}(X_{(j)}, Y_{(k)}) = V(X,Y)_{jk} = q_{jk}\lambda_k.$$

De plus, $\mathrm{Var}(X_{(j)}) = (V(X))_{jj} = v_{jj}$, et $\mathrm{Var}(Y_{(k)}) = \lambda_k$. Ainsi la correlation entre $X_{(j)}$ et $Y_{(k)}$ est donnée par :

$$r(X_{(j)}, Y_{(k)}) = \frac{\lambda_k q_{jk}}{\sqrt{\lambda_k v_{jj}}} = \frac{\sqrt{\lambda_k} q_{jk}}{\sqrt{v_{jj}}}.$$

C'est la quantité des données $(X - \overline{X})_{(j)}$ de la $j^{\text{ème}}$ ancienne variable "expliquée" par les données $Y_{(k)}$ de la $k^{\text{ème}}$ nouvelle variable.

**Attention :** Le raisonnement ci-dessus n'est valable que si la dépendance entre les données des variables est linéaire (voir la Section 3.4.2 sur les corrélations). En effet, dire qu'une corrélation forte (faible) est équivalente à une dépendance forte (faible) entre les données, n'est vrai que si on sait à priori que la dépendance entre les données est linéaire. Ceci est donc à tester sur les données avant d'effectuer une ACP. Si la dépendance entre les données n'est pas linéaire, on peut effectuer une transformation des données de sorte que ce soit vrai (log, exponentielle, racine, ...).

## 4.4 En pratique

En pratique, on utilise souvent les données centrées réduites. Ainsi,

1. La matrice des données est la matrice $Z$.
2. La matrice de covariance est la matrice de corrélation $R(X)$. En effet :

$$\begin{aligned}\mathrm{Cov}(Z_{(i)}, Z_{(j)}) &= \mathrm{Cov}\left(\frac{X_{(i)} - \overline{X_{(i)}}}{\sigma_{(i)}}, \frac{X_{(j)} - \overline{X_{(j)}}}{\sigma_{(j)}}\right), \\ &= \frac{\mathrm{Cov}(X_{(i)} - \overline{X_{(i)}}, X_{(j)} - \overline{X_{(j)}})}{\sigma_{(i)}\sigma_{(j)}}, \\ &= \frac{\mathrm{Cov}(X_{(i)}, X_{(j)})}{\sigma_{(i)}\sigma_{(j)}}, \\ &= r(X_{(i)}, X_{(j)}).\end{aligned}$$

3. La matrice $Q$ est la matrice orthogonale correspondant à la matrice $R(X)$, donnée par le Théorème spectral pour les matrices symétriques.
4. $\lambda_1 \geq \cdots \geq \lambda_p$ sont les valeurs propres de la matrice de corrélation $R(X)$.
5. La corrélation entre $Z_{(j)}$ et $Y_{(k)}$ est :

$$r(Z_{(j)}, Y_{(k)}) = \sqrt{\lambda_k} q_{jk},$$

car les coefficients diagonaux de la matrice de covariance (qui est la matrice de corrélation) sont égaux à 1.

# Chapitre 5

# Méthodes de classification

Ce chapitre concerne les méthodes de *classification*, ou de type $Q$ (en anglais). En anglais on parle aussi de "cluster analysis". Le but est de regrouper les individus dans des classes qui sont le plus "homogène" possible. On "réduit" maintenant le nombre d'individus, et non plus le nombre de variables comme lors de l'ACP. Il y a deux grands types de méthodes de classification :

1. *Classifications non-hiérarchiques (partitionnement).* Décomposition de l'espace des individus en classes disjointes.
2. *Classifications hiérarchiques.* A chaque instant, on a une décomposition de l'espace des individus en classes disjointes. 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.

**Remarque 5.1** On retrouve les méthodes de classification en statistique descriptive et inférentielle. Dans le premier cas, on se base sur les données uniquement ; dans le deuxième, il y a un modèle probabiliste sous-jacent. On traitera ici le cas descriptif uniquement.

## 5.1 Distance entre individus

Dans les méthodes de classification, les individus sont regroupés dans des classes *homogènes*. Ceci signifie que les individus d'une même classe sont *proches*. On a donc besoin d'une notion de proximité entre individus. Il existe un concept mathématique adéquat, à la base de toute méthode de classification, qui est celui de *distance*.

Soit $X$ la matrice des données, de taille $n \times p$. Ainsi il y a $n$ individus, et $p$ variables. Les données de toutes les $p$ variables pour le $i^{\text{ème}}$ individu sont représentées par la $i^{\text{ème}}$ ligne de la matrice $X$, notée $X_i^t$, qu'il faut imaginé comme étant un vecteur de $\mathbb{R}^p$, de sorte que l'on a en tout $n$ points de $\mathbb{R}^p$.

*Attention !*
Dans la suite, on écrira aussi la $i^{\text{ème}}$ ligne $X_i^t$ de la matrice $X$ sous forme d'un vecteur colonne, noté $X_i$, en accord avec les conventions introduites. Il ne faut cependant pas confondre $X_i$ avec $X_{(i)}$ qui est la $i^{\text{ème}}$ colonne (de longueur $n$) de la matrice $X$.

Une *distance* entre les données $X_i$ et $X_j$ des $i^{\text{ème}}$ et $j^{\text{ème}}$ individus, est un nombre, noté $d(X_i, X_j)$, qui satisfait les propriétés suivantes :

1. $d(X_i, X_j) = d(X_j, X_i)$.
2. $d(X_i, X_j) \geq 0$.
3. $d(X_i, X_j) = 0$, si et seulement si $i = j$.
4. $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, (inégalité triangulaire).

Ainsi une distance représente une dissimilarité entre individus. Cependant, on parlera de *dissimilarité*, au sens strict du terme, seulement lorsque les propriétés 1 à 3 sont satisfaites.

Nous présentons maintenant plusieurs exemples de distances et dissimilarités. Si les variables ont des unités qui ne sont pas comparables, on peut aussi considérer les données centrées réduites. Une autre alternative est de prendre la matrice donnée par l'ACP. Quelque soit la matrice choisie, nous gardons la notation $X = (x_{ij})$.

• Exemple 1 : Données numériques

Les distances usuellement utilisées sont :

1. Distance euclidienne : $d(X_i, X_j) = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$.
2. Distance de Manhattan : $d(X_i, X_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$.
3. Distance de Mahalanobis : $d(X_i, X_j) = \sqrt{(X_i - X_j)^t V(X)^{-1} (X_i - X_j)}$, où $V(X)$ est la matrice de covariance de $X$.

Exemple. On considère la matrice des données suivantes :

$$X = \begin{pmatrix} 1.5 & 2 & 3 & 2.8 \\ 1 & 3.1 & 6.2 & 5.3 \\ 8.2 & 2.7 & 9 & 1.2 \end{pmatrix}.$$

Alors les distances euclidiennes sont :

$$d(X_1, X_2) = \sqrt{(1.5-1)^2 + (2-3.1)^2 + (3-6.2)^2 + (2.8-5.3)^2} = 4.236,$$
$$d(X_1, X_3) = 9.161, \quad d(X_2, X_3) = 8.755.$$

• Exemple 2 : Similarité entre objets décrits par des variables binaires

Une question importante en biologie est la classification des espèces (penser aux classifications de Darwin). C'est le cas traité par cet exemple. Les $n$ individus sont alors décrits par la présence (1) ou l'absence (0) de $p$ caractéristiques, on parle de données binaires. Dans ce cas, les distances ci-dessus ne sont pas adaptées. Il existe d'autres définitions plus adéquates.

On enregistre les quantités suivantes :

$a_{ij}$ = nombre de caractéristiques communes aux individus $X_i$ et $X_j$.

$b_{ij}$ = nombre de caractéristiques possédées par $X_i$, mais pas par $X_j$.

$c_{ij}$ = nombre de caractéristiques possédées par $X_j$, mais pas par $X_i$.

$d_{ij}$ = nombre de caractéristiques possédées ni par $X_i$, ni par $X_j$.

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 : $a_{12} = 1,\ b_{12} = 1,\ c_{12} = 1,\ d_{12} = 0.$

On a alors les définitions de dissimilarités suivantes :

1. Jaccard : $d(X_i, X_j) = 1 - \dfrac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$.
2. Russel et Rao : $d(X_i, X_j) = 1 - \dfrac{a_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}}$.

Remarquer que dans les définitions ci-dessus, le terme de droite représente une similarité entre les individus, et c'est un nombre compris entre 0 et 1. Ainsi, afin d'avoir une dissimilarité, on prend $1 - \cdots$.

Exercice : Calculer les distances de Jaccard dans l'exemple ci-dessus.

• Exemple 3 : Abondance d'espèces en écologie

Cet exemple concerne le TP de la semaine 5. L'*écologie* a pour but d'étudier les interactions d'organismes entre eux, et avec leur environnement. Un exemple classique est l'étude de l'abondance de certaines espèces en différents sites. Les individus de la matrice des données $X$ sont dans ce cas des "sites", et les variables sont des "espèces". Le coefficient $x_{ij}$ de la matrice des données donne l'abondance de l'espèce $j$ sur le site $i$.

Dans ce cas, on peut utiliser les distances de l'Exemple 1, mais il existe aussi d'autres notions plus appropriées à l'écologie. En particulier :

1. Dissimilarité de Bray-Curtis : $d(X_i, X_j) = \dfrac{\sum_{k=1}^{p} |X_{ik} - X_{jk}|}{\sum_{k=1}^{p} (X_{ik} + X_{jk})}$
2. Distance de corde : On normalise les données des $i^{\text{ème}}$ et $j^{\text{ème}}$ individus de sorte à ce qu'ils soient sur la sphère de rayon 1 dans $\mathbb{R}^p$ : $\tilde{X}_i = \frac{X_i}{\|X_i\|}, \tilde{X}_j = \frac{X_j}{\|X_j\|}$. Alors la distance $d(X_i, X_j)$ de $X_i$ à $X_j$ est la distance euclidienne entre $\tilde{X}_i$ et $\tilde{X}_j$, c'est à dire $d(X_i, X_j) = \|\tilde{X}_i - \tilde{X}_j\|$.

• Matrice des distances

Les distances entre les données de tous les individus sont répertoriées dans une matrice, notée $D = (d_{ij})$, de taille $n \times n$, telle que :

$$d_{ij} = d(X_i, X_j).$$

Remarquer que seuls $\frac{n(n-1)}{2}$ termes sont significatifs, étant donné que la matrice est symétrique ($d_{ij} = d_{ji}$), et que les termes sur la diagonale sont nuls.

## 5.2 Le nombre de partitions

La première idée pour trouver la meilleure partition de $n$ individus, serait de fixer un critère d'optimalité, puis de parcourir toutes les partitions possibles, de calculer ce critère, et de déterminer laquelle des partitions est la meilleure. Ceci n'est cependant pas réaliste étant donné que le nombre de partitions devient vite gigantesque, comme nous allons le voir ci-dessous.

Soit $S(n, k)$ le nombre de partitions de $n$ éléments en $k$ parties. Alors, $S(n, k)$ satisfait la relation de récurrence suivante :

$$S(n, k) = kS(n-1, k) + S(n-1, k-1), \quad k = 2, \cdots, n-1.$$
$$S(n, n) = S(n, 1) = 1.$$

Preuve:

Vérifions d'abord les conditions de bord. Il n'y a bien sûr qu'une seule manière de partitionner $n$ éléments en $n$ classes, ou en 1 classe.

Si l'on veut partitionner $n$ éléments en $k$ classes, alors il y a deux manières de le faire :

1. Soit on partitionne les $n-1$ premiers objets en $k$ groupes, et on rajoute le $n$-ième à un des groupes existants, de sorte qu'il y a $k$ manières de le rajouter.
2. Soit on partitionne les $n-1$ premiers objets en $k-1$ groupes, et le dernier élément forme un groupe à lui tout seul, de sorte qu'il n'y a qu'une seule manière de le faire.

□

On peut montrer que la solution de cette récurrence est donnée par :

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\left(\begin{array}{c} k \\ j \end{array}\right) j^n.$$

Soit $S(n)$ le nombre total de partitions de $n$ éléments. Alors,

$$S(n) = \sum_{k=1}^{n} S(n,k),$$

et on peut montrer que

$$S(n) = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}.$$

Les première valeurs pour $S(n)$ sont :

$$S(1) = 1,\ S(2) = 2,\ S(3) = 5,\ S(4) = 15,\ S(5) = 52,\ S(6) = 203,\ S(7) = 877,$$
$$S(8) = 4140,\ S(9) = 21\,147, \cdots, S(20) = 51\,724\,158\,235\,372\,!!!$$

**Terminologie** : Le nombre $S(n,k)$ de partitions de $n$ éléments en $k$ parties s'appelle *nombre de Stirling de deuxième espèce*. Le nombre $S(n)$ de partitions de $n$ éléments s'appelle *nombre de Bell*, il est habituellement noté $B_n$.

## 5.3 Inertie d'un nuage de points

Cette section introduit une notion dont le lien se fera plus tard avec le sujet. On la met ici parce qu'elle intervient dans les deux méthodes de classification que nous allons étudier.

On considère la matrice des données $X = (x_{ij})$ de taille $n \times p$, et on suppose que la distance entre les données des individus est la distance euclidienne.

• Inertie d'un individu, inertie d'un nuage de points

Souvenez-vous que les données de chaque individu sont interprétées comme un vecteur/point de $\mathbb{R}^p$, de sorte que l'on imagine la matrice des données $X$ comme étant un nuage de $n$ points dans $\mathbb{R}^p$. Dans ce contexte, on peut interpréter le vecteur des moyennes $\bar{\mathbf{x}}^t$ comme étant le *centre de gravité* du nuage de points.

Soit $X_i^t$ la donnée de toutes les variables pour l'individu $i$. Souvenez-vous que le vecteur $X_i^t$ écrit sous forme d'une colonne est noté $X_i$, et le vecteur $\bar{\mathbf{x}}^t$ écrit sous forme d'une colonne est noté $\bar{\mathbf{x}}$. Alors l'*inertie de l'individu* $i$, notée $I_i$, est par définition la distance au carré de cet individu au centre de gravité du nuage de points, i.e.

$$I_i = \|X_i - \bar{\mathbf{x}}\|^2.$$

C'est une sorte de variance pour le $i^{\text{ème}}$ individu.

L'*inertie du nuage de points*, notée $I$, est la moyenne arithmétique des inerties des individus :

$$I = \frac{1}{n}\sum_{i=1}^{n} I_i$$

• Inertie inter-classe, inertie intra-classe

On suppose que les individus sont regroupés en $k$ classes $C_1, \cdots, C_k$. Soit $n_\ell$ le nombre d'individus dans la classe $C_\ell$ (on a donc $\sum_{\ell=1}^{k} n_\ell = n$). Soit $\bar{\mathbf{x}}(\ell)$ le centre de gravité de la classe $C_\ell$, et $\bar{\mathbf{x}}$ le centre de gravité du nuage de points. Alors l'inertie de l'individu $i$ dans la classe $C_\ell$ est :

$$I_i = \|X_i - \bar{\mathbf{x}}(\ell)\|^2,$$

et l'*inertie de la classe* $C_\ell$ est la somme des inerties des individus dans cette classe, i.e.

$$\sum_{i \in C_\ell} \|X_i - \bar{\mathbf{x}}(\ell)\|^2.$$

L'*inertie intra-classe*, notée $I_{\text{intra}}$, est $1/n$ fois la somme des inerties des différentes classes, donc :

$$I_{\text{intra}} = \frac{1}{n}\sum_{\ell=1}^{k}\sum_{i \in C_\ell} \|X_i - \bar{\mathbf{x}}(\ell)\|^2.$$

L'*inertie inter-classe* est la somme pondérée des inerties des centres de gravité des différentes classes, c'est à dire :

$$I_{\text{inter}} = \frac{1}{n}\sum_{\ell=1}^{k} n_\ell \|\bar{\mathbf{x}}(\ell) - \bar{\mathbf{x}}\|^2.$$

**Remarque 5.2** Si une classe est "bien regroupée" autour de son centre de gravité, son inertie est faible. Ainsi, un bon critère pour avoir des classes homogènes est d'avoir une inertie intra-classe qui soit aussi petite que possible.

• Lien entre inertie du nuage de points, inertie intra/inter-classe

Comme conséquence du Théorème de König-Huygens, on a :

$$I = I_{\text{inter}} + I_{\text{intra}}.$$

**Remarque 5.3** On a vu ci-dessus que pour avoir des classes homogènes, il faut une inertie intra-classe qui soit aussi petite que possible. En utilisant le Théorème de König-Huygens, cela revient à dire qu'il faut une inertie inter-classe aussi grande que possible.

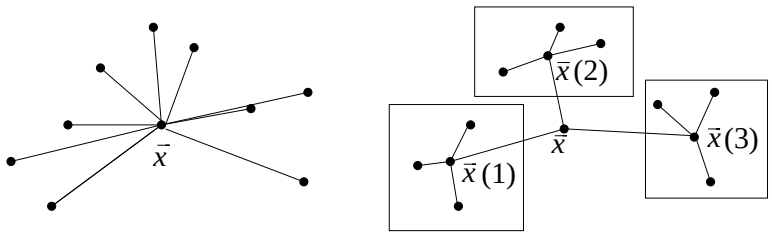



FIG. 5.1 – Illustration de l'inertie et du Théorème de König-Huygens ($p = 2, n = 9, k = 3$).

## 5.4 Méthodes non hiérarchiques : méthode des centres mobiles

• Origine et extensions : 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.

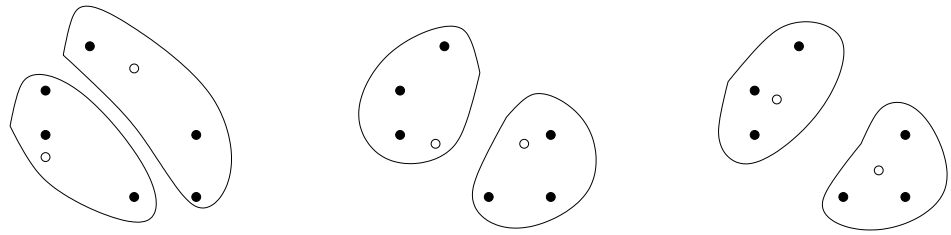


FIG. 5.2 – Illustration de la méthode des centre mobiles, $k = 2$.

• 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.
2. Le résultat dépend de la condition initiale. Ainsi, on n'est pas sûr d'atteindre la partition en $k$ classes, telle que $I_{\text{intra}}$ est minimum (on a minimum "local" et non "global").

**Remarque 5.4** *Un peu plus sur la notion d'algorithme ...* Il n'y a pas d'accord absolu sur la définition d'un algorithme, nous en donnons néanmoins une. Un *algorithme* est une liste d'instructions pour accomplir une tâche : étant donné un état initial, l'algorithme effectue une série de tâches successives, jusqu'à arriver à un état final.

Cette notion a été systématisée par le mathématicien perse Al Khuwarizmi ($\sim 780 - 850$), puis le savant arabe Averroès (12ème siècle) évoque une méthode similaire. C'est un moine nommé Adelard de Barth (12ème siècle) qui a introduit le mot latin "algorismus", devenu en français "algorithme". Le concept d'algorithme est intimement lié aux fonctionnement des ordinateurs, de sorte que l'*algorithmique* est actuellement une science à part entière.

## 5.5 Méthodes de classification hiérarchiques

Pour le moment, on n'a qu'une notion de dissimilarité entre *individus*. Afin de décrire la méthode, on suppose que l'on a aussi une notion de dissimilarité entre *classes*.

• Algorithme

1. *Initialisation.* Partition en $n$ classes $C_1, \cdots, C_n$, où chaque individu représente une classe. On suppose donné la matrice des distances entre individus.
2. *Etape $k$.* ($k = 0, \cdots, n-1$). Les données sont $n-k$ classes $C_1, \cdots, C_{n-k}$, 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, $C_1, \cdots, C_{n-k-1}$.
   - Recalculer la matrice des distances qui donnent les $\frac{(n-k)(n-k-1)}{2}$ distances entre les nouvelles classes.
   - Poser $k := k+1$.

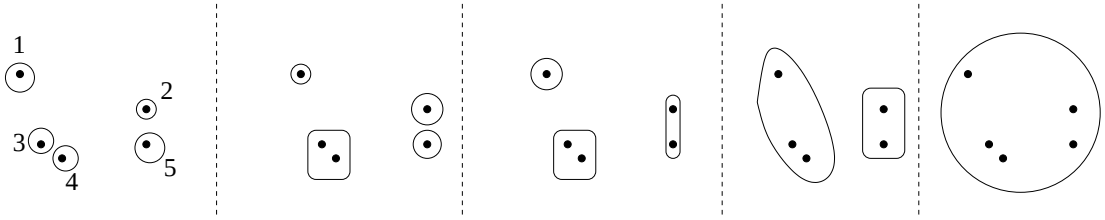



FIG. 5.3 – Illustration de la classification hiérarchique.

• Représentation

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.

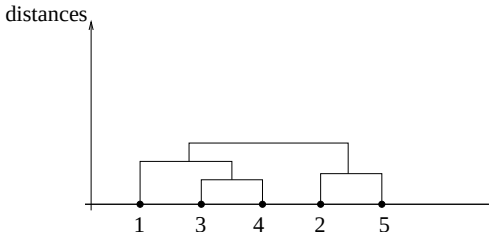



FIG. 5.4 – Arbre / dendogramme correspondant.

• 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.

2. Choix de la bonne partition : repérer un saut (si possible) entre les agrégations courtes distances (branches courtes de l'arbre) et les longues distances (branches longues de l'arbre). Parfois le nombre adéquat de classe est donné par le type de données. Il existe aussi des tests statistiques pour déterminer le bon nombre de classes.
3. 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 $\frac{(n-k)(n-k-1)}{2}$.

• Distance entre classes

Voici plusieurs définitions possibles de distances entre des classes formées de plusieurs individus. Soient $C$, $C'$ deux classes.

1. *Le saut minimum / single linkage.* La distance du saut minimum entre les classes $C$ et $C'$, notée $d(C, C')$, est par définition :
$$d(C, C') = \min_{X_i \in C,\, X_j \in C'} d(X_i, X_j).$$
C'est la plus petite distance entre éléments des deux classes.
2. *Le saut maximum / complete linkage.* La distance du saut maximum entre les classes $C$ et $C'$, notée $d(C, C')$, est par définition :
$$d(C, C') = \max_{X_i \in C,\, X_j \in C'} d(X_i, X_j).$$
C'est la plus grande distance entre éléments des deux classes.
3. *Le saut moyen / average linkage.* La distance du saut moyen entre les classes $C$ et $C'$, notée $d(C, C')$, est par définition :
$$d(C, C') = \frac{1}{|C||C'|} \sum_{X_i \in C} \sum_{X_j \in C'} d(X_i, X_j).$$
C'est la moyenne des distances entre tous les individus des deux classes.
4. *Méthode de Ward pour distances euclidiennes.* Cette méthode utilise le concept d'*inertie*. On se souvient qu'une classe est homogène si son inertie est faible, ainsi on souhaite avoir une inertie intra-classe qui soit faible.
Quand on fusionne deux classes $C$ et $C'$, l'inertie intra-classe augmente. Par le Théorème de König-Huygens, cela revient à dire que l'inertie inter-classe diminue (car l'inertie totale du nuage de points est constante). On définit la *distance de Ward*, notée $d(C, C')$ entre les classes $C$ et $C'$, comme étant la perte de l'inertie inter-classe (ou gain d'inertie intra-classe)
$$\begin{aligned} d(C, C') &= \frac{|C|}{n}\|\bar{\mathbf{x}}(C) - \bar{\mathbf{x}}\|^2 + \frac{|C'|}{n}\|\bar{\mathbf{x}}(C') - \bar{\mathbf{x}}\|^2 - \frac{|C| + |C'|}{n}\|\bar{\mathbf{x}}(C \cup C') - \bar{\mathbf{x}}\|^2, \\ &= \frac{|C|.|C'|}{|C| + |C'|}\|\bar{\mathbf{x}}(C) - \bar{\mathbf{x}}(C')\|^2, \end{aligned}$$

où $\bar{\mathbf{x}}(C \cup C')$ est le centre de gravité de la classe $C \cup C'$, $|C|$ (resp. $|C'|$) est la taille de la classe $C$ (resp. $C'$).

Ainsi, on fusionne les deux classes telles que cette perte (ce gain) soit minimum.

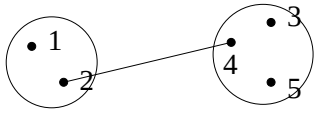
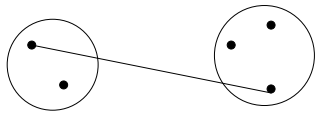
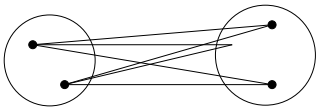
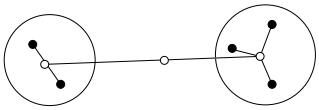
| classe C | classe C' | d(C,C') | Nom |
|---|---|---|---|
|   | | d(2,4) | saut minimum |
|  | | d(1,5) | saut maximum |
|  | | (1/6)(d(1,3)+d(1,4)+...+d(2,5)) | saut moyen |
|  | | (\|C\|.\|C'\|/(\|C\|+\|C'\|)) $\|\bar{x}(C)-\bar{x}(C')\|^2$ | Ward |

FIG. 5.5 – Exemple de calcul des distances.

# Annexe A

# Exercices et exemples

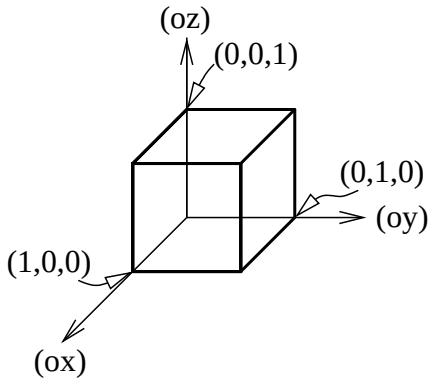
Exercice 1 sur le Chapitre 2

Soit $X$ la matrice : $X = \begin{pmatrix} 2 & 3 & -1 \\ 1 & 2 & -1 \end{pmatrix}$, et soit $\mathbf{y}$ le vecteur $\mathbf{y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. On note $X_{(1)}, X_{(2)}, X_{(3)}$, les 3 vecteurs colonnes de la matrice $X$.

1. Calculer $X^t\mathbf{y}$.
2. Représenter dans le plan ($\mathbb{R}^2$) :
   - $X_{(1)}, X_{(2)}, X_{(3)}$,
   - la droite $D$ passant par l'origine $(0, 0)$, de vecteur directeur $\mathbf{y}$.
3. Calculer le produit scalaire de $X_{(1)}$ avec $\mathbf{y}$. Même question avec $X_{(2)}, X_{(3)}$. Comparer avec le résultat obtenu en 1.
4. Calculer la norme au carré du vecteur $\mathbf{y}$.
5. En déduire la projection du point $X_{(1)}$ sur la droite $D$. Même question avec $X_{(2)}, X_{(3)}$. Faire une représentation graphique des projections.
6. Calculer la coordonnée de chacune de ces projections.

## Exercice 2 sur le Chapitre 2

On considère le cube ci-dessous.




1. Calculer le cosinus de l'angle formé par les vecteurs, $\mathbf{x} = (1, 1, 0)$ et $\mathbf{y} = (0, 1, 1)$.
   *Indication.* Utiliser la formule, $\cos \angle(\mathbf{x}, \mathbf{y}) = \dfrac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$.
2. En déduire l'angle formé par ces deux vecteurs.
3. Représenter cet angle graphiquement.

## Exemple pour le Chapitre 3

On étudie la quantité d'écorce (en centigrammes) sur 28 arbres dans les 4 directions cardinales N, E, S, O.

• Matrice des données

$X =$

| N | E | S | O |
|---|---|---|---|
| 72 | 66 | 76 | 77 |
| 60 | 53 | 66 | 63 |
| 56 | 57 | 64 | 58 |
| 41 | 29 | 36 | 38 |
| 32 | 32 | 35 | 36 |
| 30 | 35 | 34 | 26 |
| 39 | 39 | 31 | 27 |
| 42 | 43 | 31 | 25 |
| 37 | 40 | 31 | 25 |
| 33 | 29 | 27 | 36 |
| 32 | 30 | 34 | 28 |
| 63 | 45 | 74 | 63 |
| 54 | 46 | 60 | 52 |
| 47 | 51 | 52 | 43 |
| 91 | 79 | 100 | 75 |
| 56 | 68 | 47 | 50 |
| 79 | 65 | 70 | 61 |
| 81 | 80 | 68 | 58 |
| 78 | 55 | 67 | 60 |
| 46 | 38 | 37 | 38 |
| 39 | 35 | 34 | 37 |
| 32 | 30 | 30 | 32 |
| 60 | 50 | 67 | 54 |
| 35 | 37 | 48 | 39 |
| 39 | 36 | 39 | 31 |
| 50 | 34 | 37 | 40 |
| 43 | 37 | 39 | 50 |
| 48 | 54 | 57 | 43 |

• Interprétation de la matrice

- Le vecteur ligne $X_3^t = (56 \quad 57 \quad 64 \quad 58)$, représente la donnée de toutes les variables pour le $3^{\text{ème}}$ arbre.

- Le vecteur colonne

$$X_{(1)} = \begin{pmatrix} 72 \\ 60 \\ 56 \\ 41 \\ 32 \\ 30 \\ 39 \\ 42 \\ 37 \\ 33 \\ 32 \\ 63 \\ 54 \\ 47 \\ 91 \\ 56 \\ 79 \\ 81 \\ 78 \\ 46 \\ 39 \\ 32 \\ 60 \\ 35 \\ 39 \\ 50 \\ 43 \\ 48 \end{pmatrix}$$

représente les données de la variable "quantité d'écorce sur le côté nord" pour tous les individus.

• Moyenne arithmétique

Le vecteur ligne des moyennes arithmétiques est :

$$\bar{\mathbf{x}}^t = (50.536, 46.179, 49.679, 45.179).$$

où $\overline{X_{(1)}} = \frac{72+60+56+\cdots+43+48}{28} = 50.536$ est la moyenne arithmétique de la variable $N$, $\overline{X_{(2)}} = 46.179$ celle de la variable $E$ etc...

• Etendue

Le vecteur des maximums est :

$$\mathbf{x}^t_{\max} = (91, 79, 100, 77).$$

Le vecteur des minimums est :

$$\mathbf{x}^t_{\min} = (30, 30, 27, 25).$$

Donc, le vecteur des étendues est :

$$\mathbf{w}^t = (61, 49, 73, 52),$$

où $w_{(1)} =$ est l'étendue de la variable $N$, etc...

• Variances et écarts-types

$$\begin{aligned}
\mathrm{Var}(X_{(1)}) &= \frac{1}{28}[(72-50.536)^2 + (60-50.536)^2 + \cdots + (48-50.536)^2] = 280.034 \\
\sigma(X_{(1)}) &= \sqrt{280.034} = 16.734 \\
\mathrm{Var}(X_{(2)}) &= 212.075 \\
\sigma(X_{(2)}) &= \sqrt{212.075} = 14.563 \\
\mathrm{Var}(X_{(3)}) &= 337.504 \\
\sigma(X_{(3)}) &= \sqrt{337.504} = 18.371 \\
\mathrm{Var}(X_{(4)}) &= 217.932 \\
\sigma(X_{(4)}) &= \sqrt{217.932} = 14.763.
\end{aligned}$$

Par le Théorème de König-Huygens, on a aussi :

$$\mathrm{Var}(X_{(1)}) = \frac{1}{28}[72^2 + 60^2 + \cdots + 48^2] - 50.536^2,$$

et de manière similaire pour les autres variables.

• Matrice de covariance

$$\begin{aligned}
\mathrm{Cov}(X_{(1)}, X_{(2)}) =& \frac{1}{28}[(72-50.536)(66-46.179) + (60-50.536)(53-46.179) + \cdots + \\
&+ (48-50.536)(54-46.179)].
\end{aligned}$$

Et de manière similaire pour les autres. Ainsi, on peut calculer la matrice de covariance $V(X)$ :

$$V(X) = \begin{pmatrix} 280.034 & 215.761 & 278.136 & 218.190 \\ 215.761 & 212.075 & 220.879 & 165.254 \\ 278.136 & 220.879 & 337.504 & 250.272 \\ 218.190 & 165.254 & 250.272 & 217.932 \end{pmatrix}$$

• Matrice de corrélation

$$r(X_{(1)}, X_{(1)}) = 1$$
$$r(X_{(1)}, X_{(2)}) = \frac{\mathrm{Cov}(X_{(1)}, X_{(2)})}{\sigma_{(1)}\sigma_{(2)}} = \frac{215.761}{16.734 \times 14.563} = 0.885.$$

Et de manière similaire pour les autres. Ainsi, on peut calculer la matrice de corrélation $R(X)$ :

$$R(X) = \begin{pmatrix} 1 & 0.885 & 0.905 & 0.883 \\ 0.885 & 1 & 0.826 & 0.769 \\ 0.905 & 0.826 & 1 & 0.923 \\ 0.883 & 0.769 & 0.923 & 1 \end{pmatrix}.$$

## Exercice sur le Chapitre 3

On mesure en microlitre la quantité de trois types de substances émise par des roses qui subissent trois traitements différents. On obtient les données suivantes.

| | Subs. 1 | Subs. 2 | Subs. 3 |
|---|---|---|---|
| Rose 1 | 4 | 3 | 6 |
| Rose 2 | 2 | 5 | 8 |
| Rose 3 | 0 | 1 | 7 |

1. Calculer la matrice des moyennes arithmétiques pour ces données.
2. Calculer la matrice de covariance $V$. En déduire les variances et les covariances.
3. Recalculer la covariance entre les données des première et troisième variables en utilisant le Théorème de König-Huygens.
4. Calculer la matrice de corrélation $R$. En déduire les corrélations.
5. Représenter dans $\mathbb{R}^3$ les trois vecteurs $(X - \overline{X})_{(1)}, (X - \overline{X})_{(2)}, (X - \overline{X})_{(3)}$, ainsi que les angles dont le cosinus est calculé par les corrélations.

*Remarque.* La taille de ce jeu de données permet de faire les calculs à la main. Le nombre d'individus est néanmoins trop petit pour pouvoir tirer des conclusions à partir des corrélations.

Exemple pour le Chapitre 4 : une ACP à but pédagogique.

Source : G : Saporta "Probabilités, Analyse des données en statistiques". L'étude (1972) concerne la consommation annuelle en francs de 8 denrées alimentaires (variables), par 8 catégories socio-professionnelles (individus).

- Les variables
  1. Var 1 : Pain ordinaire.
  2. Var 2 : Autre pain.
  3. Var 3 : Vin ordinaire.
  4. Var 4 : Autre vin.
  5. Var 5 : Pommes de terre.
  6. Var 6 : Légumes secs.
  7. Var 7 : Raisins de table.
  8. Var 8 : Plats préparés.
- Les individus
  1. Exploitants agricoles.
  2. Salariés agricoles.
  3. Professions indépendantes.
  4. Cadres supérieurs.
  5. Cadres moyens.
  6. Employés.
  7. Ouvriers.
  8. Inactifs.
- Matrice des données

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

• Quelques statistiques élémentaires

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\overline{\mathbf{x}}^t$ | 127.125 | 4.875 | 96.125 | 58.875 | 38.625 | 6.625 | 10.125 | 24.375 |
| $\mathrm{Var}^t$ | 778.696 | 10.125 | 1504.7 | 980.696 | 61.9821 | 7.98214 | 19.8393 | 149.982 |
| $\sigma^t$ | 27.905 | 3.182 | 38.790 | 31.316 | 7.873 | 2.825 | 4.454 | 12.247 |

• Données centrées réduites

$$Z = \begin{pmatrix} 1.429 & -1.218 & 1.724 & -1.146 & 0.301 & 0.486 & -0.926 & -1.500 \\ 1.249 & -0.903 & 1.156 & -1.496 & 0.174 & 1.902 & -1.375 & -0.7655 \\ -0.291 & 0.353 & -0.699 & -0.091 & 0.047 & -0.575 & 0.645 & 1.357 \\ -1.437 & 1.924 & -0.853 & 1.664 & -1.476 & -1.283 & 1.768 & 1.194 \\ -0.864 & 0.039 & -0.725 & 0.578 & -0.841 & -0.929 & 0.196 & 0.459 \\ -0.577 & -0.274 & -0.621 & 0.227 & -0.587 & -0.221 & -0.028 & 0.295 \\ 0.103 & -0.589 & -0.518 & -0.219 & 0.555 & 0.132 & -0.701 & -0.683 \\ 0.389 & 0.667 & 0.538 & 0.482 & 1.825 & 0.486 & 0.420 & -0.357 \end{pmatrix}.$$

• Matrice des corrélations multipliées par 100

$$R(X) = \begin{pmatrix} 100 & -77.366 & 92.619 & -90.579 & 65.635 & 88.856 & -83.343 & -85.585 \\ -77.366 & 100 & -60.401 & 90.444 & -33.289 & -67.337 & 95.882 & 77.122 \\ 92.619 & -60.401 & 100 & -75.016 & 51.708 & 79.173 & -66.901 & -82.799 \\ -90.579 & 90.444 & -75.016 & 100 & -41.857 & -83.860 & 92.393 & 71.979 \\ 65.635 & -33.289 & 51.708 & -41.857 & 100 & 60.292 & -40.993 & -55.396 \\ 88.856 & -67.337 & 79.173 & -83.860 & 60.292 & 100 & -82.445 & -75.092 \\ -83.343 & 95.882 & -66.901 & 92.393 & -40.993 & -82.445 & 100 & 83.445 \\ -85.585 & 77.122 & -82.799 & 71.979 & -55.396 & -75.092 & 83.445 & 100 \end{pmatrix}.$$

• Valeurs propres de la matrice $R(X)$, et % de la variation totale expliquée

| Valeurs propres | % Variation totale |
|---|---|
| 6.21 | 77.6 |
| 0.880 | 88.6 |
| 0.416 | 93.8 |
| 0.306 | 97.6 |
| 0.168 | 99.7 |
| 0.0181 | 99.9 |
| 0.00345 | 100 |
| $7.36 \times 10^{-}12$ | 100 |

Exemple de calcul du % de la variation totale :

$$\% \text{ Variation totale } 1 = \frac{6.21}{6.21 + +0.880 + \cdots + 0.00345 + 7.36 \times 10^{-12}}.$$

On déduit que 2 composantes principales suffisent pour représenter 88.6% de la variation totale.

• Matrice des vecteurs propres $Q$ de $R(X)$

$$Q = \begin{pmatrix} -0.391 & -0.138 & 0.162 & 0.119 & 0.294 & 0.398 & -0.107 & 0.729 \\ 0.349 & -0.441 & 0.320 & 0.218 & -0.265 & 0.521 & 0.423 & -0.118 \\ -0.349 & -0.202 & 0.681 & -0.0289 & 0.246 & -0.465 & 0.254 & -0.180 \\ 0.374 & -0.260 & 0.0735 & -0.397 & -0.346 & -0.423 & 0.0333 & 0.575 \\ -0.246 & -0.744 & -0.558 & -0.0740 & 0.176 & -0.108 & 0.0934 & -0.135 \\ -0.365 & -0.128 & 0.0324 & 0.519 & -0.669 & -0.185 & -0.313 & 0.0127 \\ 0.373 & -0.326 & 0.254 & 0.0637 & 0.272 & 0.0163 & -0.766 & -0.159 \\ 0.362 & 0.0502 & -0.162 & 0.708 & 0.333 & -0.360 & 0.225 & 0.219 \end{pmatrix}.$$

• Plot des coordonnées des individus sur les deux premiers axes principaux
(2 premières colonnes de $Y = ZQ$)

$$Y = \begin{pmatrix} -3.153 & 0.229 & 0.785 & -0.581 & 0.539 & 0.020 & -0.020 & 3.302 * 10^{-9} \\ -3.294 & 0.418 & 0.328 & 0.857 & -0.461 & 0.004 & 0.029 & -4.575 * 10^{-9} \\ 1.376 & -0.054 & -0.517 & 0.799 & 0.700 & 0.054 & -0.004 & 8.482 * 10^{-10} \\ 4.077 & -0.164 & 0.962 & 0.014 & -0.242 & 0.118 & -0.014 & 3.170 * 10^{-9} \\ 1.607 & 0.801 & -0.163 & -0.385 & 0.037 & -0.130 & 0.109 & -1.829 * 10^{-8} \\ 0.754 & 0.756 & -0.322 & -0.064 & -0.192 & -0.183 & -0.102 & 1.556 * 10^{-8} \\ -0.841 & 0.171 & -0.914 & -0.515 & -0.274 & 0.218 & -0.005 & 1.760 * 10^{-9} \\ -0.526 & -2.157 & -0.159 & -0.123 & -0.106 & -0.102 & 0.009 & -1.782 * 10^{-9} \end{pmatrix}.$$

On remarque que la variabilité est la plus grande le long de l'axe 1.

Le premier axe met en évidence l'opposition (quant aux consommations étudiées) qui existe entre cadres supérieurs (individus $1, 2$) et agriculteurs (individus 4).

Le deuxième axe est caractéristique des inactifs (individus 8) qui sont opposés à presque toutes les autres catégories.

• Interprétation des composantes principales

Ceci se fait en regardant les corrélations avec les variables de départ.

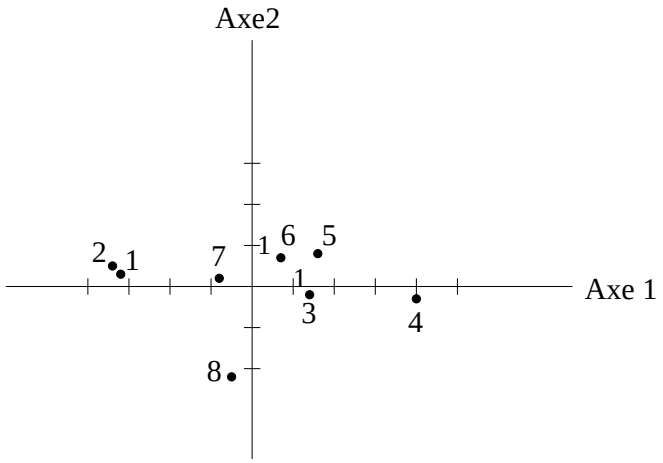



FIG. A.1 – Projection des individus sur les deux premiers axes principaux.

| Variables | Axe 1 : $Y_{(1)}$ | Axe 2 : $Y_{(2)}$ |
|---|---|---|
| Var 1 : $X_{(1)}$ | $\sqrt{\lambda_1}q_{11} = -0.97$ | $\sqrt{\lambda_2}q_{12} = -0.129$ |
| Var 2 : $X_{(2)}$ | $\sqrt{\lambda_1}q_{21} = 0.87$ | $\sqrt{\lambda_2}q_{22} = -0.413$ |
| Var 3 : $X_{(3)}$ | $\sqrt{\lambda_1}q_{31} = -0.87$ | $\sqrt{\lambda_2}q_{32} = -0.189$ |
| Var 4 : $X_{(4)}$ | $\sqrt{\lambda_1}q_{41} = 0.93$ | $\sqrt{\lambda_2}q_{42} = -0.244$ |
| Var 5 : $X_{(5)}$ | $\sqrt{\lambda_1}q_{51} = -0.614$ | $\sqrt{\lambda_2}q_{52} = -0.7$ |
| Var 6 : $X_{(6)}$ | $\sqrt{\lambda_1}q_{61} = -0.91$ | $\sqrt{\lambda_2}q_{62} = -0.12$ |
| Var 7 : $X_{(7)}$ | $\sqrt{\lambda_1}q_{71} = 0.93$ | $\sqrt{\lambda_2}q_{72} = -0.306$ |
| Var 8 : $X_{(8)}$ | $\sqrt{\lambda_1}q_{81} = 0.9$ | $\sqrt{\lambda_2}q_{82} = 0.0471$ |

La première composante principale mesure donc la répartition de la consommation entre aliments ordinaires bon marchés (Var 1 : Pain ordinaire, Var 3 : Vin ordinaire, Var 6 : Légumes secs) et aliments plus recherchés (Var 2 : Autre pain, Var 4 : Autre vin, Var 7 : Raisins de table, Var 8 : Plats préparés).

La deuxième composante principale est liée essentiellement à la consommation de pommes de terre, dont une consommation élevée caractérise les inactifs.

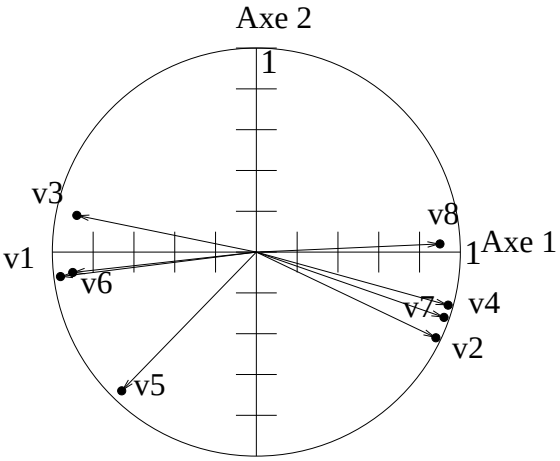



FIG. A.2 – Corrélations des données des anciennes variables avec les deux premiers axes principaux.

# Bibliographie


[1] Dytham C. *Choosing and Using Statistics : A Biologist's Guide.*

[2] Legendre L., Legendre P., Ecologie numérique. *Presses de l'Université du Québec, 1984.*

[3] Mardia K.V., Kent J.T., Bibby J.M., Multivariate analysis. *Probability and Mathematical Statistics. A series of Monographs anx Textbooks.*

[4] Saporta, G. Probabilités, analyse des données et statistique. *Editions Technip, 1990.*

[5] Venables W. N., Ripley B. D., Modern Applied Statistics with S. *Fourth Edition. Springer. ISBN 0-387-95457-0, 2002.*

[6] Verecchia, E. Notes de cours.

[7] Stahel, W. Notes de cours disponibles sur le web. http ://www.stat.math.ethz.ch/ stahel/