Université Abdelmalek Essaadi
Ecole Nationale des Sciences Appliquées
Al Hoceima, Maroc

Statistique en Grande Dimension

–Cours–
Analyse des données et Modélisation

Mohamed ADDAM
Professeur de Mathématiques

École Nationale des Sciences Appliquées d'Al Hoceima
–ENSAH–
Année Universitaire 2021/2022
addam.mohamed@gmail.com
m.addam@uae.ac.ma

28 février 2022

Table des matières

Notations

- $\mathbb{N} := \{0, 1, 2, \ldots\}$ l'ensemble des naturels,
- $(-\mathbb{N}) := \{\ldots, -2, -1, 0\}$ l'ensemble des opposés des naturels,
- $\mathbb{N}^* = \mathbb{N}\backslash\{0\} := \{n \in \mathbb{N} \ / \ n \neq 0\}$,
- $\mathbb{Z} := \mathbb{N} \cup (-\mathbb{N})$ l'ensemble des entiers,
- $\mathbb{D}$ l'ensemble des décimaux,
- $\mathbb{Q} := \{\frac{p}{q} \ / \ p \in \mathbb{Z},\ q \in \mathbb{N}^*\}$ l'ensemble des rationnels,
- $\mathbb{R}$ l'ensemble des nombres réels,
- $\mathbb{C}$ l'ensemble des nombres complexes.

On suppose connues les propriétés élémentaires de ces ensembles.

Chapitre 1

Statistique : Analyse univariée et multivariée

1.1 Statistique

1.1.1 Généralités

”La statistique” est une 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.

1.1.2 Vocabulaire

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

♢ Individus : éléments de la population étudiée.

♢ Variable : propriété commune aux individus de la population, que l’on souhaite étudier. Elle peut être

1. qualitative : couleur de pétales, sexe,...
2. quantitative (numérique) : comme la taille, le poids, le volume. On distingue encore les variables
   – continues : toutes les valeurs d’un intervalle de $\mathbb{R}$ sont acceptables.
   – discrètes : seul un nombre discret de valeurs sont possibles. Par exemple : le nombre d’espèce ressencées sur une parcelle.

   Les valeurs observées pour les variables s’appellent les **donnée**.

♢ Echantillon : partie étudiée de la population.

1.1.3 Collecte de données

La collecte de données (observation 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. C’est-à-dire **Tout mettre en oeuvre pour que ceci soit vérifié**.

1.1.4 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) Représntation 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 réalisations de variables aléatoires, qui ont une certaine loi de probabilité.
   Nécessite des outils mathématiques plus pointus et variés (Théorie des probabilités).
   Questions typiques :
   (a) Estimation de paramètres.
   (b) Intervalles de confiance.
   (c) Tests d'hypothèse.
   (d) Modélisation : exemple (régression linéaire).

1.1.5 Statistique univarié/ multivarié

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

Exemple 1.1.1 – *Univarié. Population : iris. Variable : longueur des pétales.*
– *Multivarié. Population : iris. Variable 1 : longueur des pétales. Variable 2 : largeur des pétales.*

1.1.6 Statistique descriptive

Ce cours a pour thème tous les types de statistiques, mais une grande partie du volume horaire sera consacrer à la statistique descriptive dans ces deux cas consécutifs : univarié et 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** dites méthodes **R** en anglais : ces méthodes cherchent à réduire le nombre de variables en les résumant par 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** dites méthodes **Q** en anglais : ces méthodes visent à réduire le nombre d'individus en formant des groupes homogènes.

1.2 Statistique descriptive élémentaire

Cette section est illustrée au tableau au moyen d'un jeu de données.

1.2.1 La matrice des données

Avant de pouvoir analyser les données, il faut un moyen pour les répertorier et stoker. L'outil naturel est d'utiliser une matrice A , appelée matrices des données. Nous nous restreignons au cas où les données sont de type quantitatif, ce qui est fréquent en médecine, biologie et au laboratoires d'analyse.
On suppose que l'on a une population constituée de n individus, et que pour chacun de ces individus, on observe p variables. Alors, les données sont répertoriées de la manière suivante :

$$A = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix}$$

L'élément a_{ij} de la matrice A représente l'observation de la j^{eme} variable pour l'individu i.
On va noter i^{me} ligne de A, représentant les données de toutes les variables pour le i^{eme} individu, par A_i^T.
On va noter j^{eme} colonne de A, représentant les données de la j^{eme} variable pour tous les individus, par $A_{(j)}$. Ainsi,

$$A_i^T = (a_{i1}, \dots, a_{ip}) \quad \text{et} \quad A_{(j)} = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

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.

Example 1.2.1 *Voici des données représentant les résultats de 6 individus à un test de statstique (variable 1) et de géologie (variable 2).*

$$A = \begin{pmatrix} 11 & 13,5 \\ 12 & 13,5 \\ 13 & 13,5 \\ 14 & 13,5 \\ 15 & 13,5 \\ 16 & 13,5 \end{pmatrix}$$

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.

1.2.2 Paramètres de position

Les quantités ci-dessous sont des généralisations naturelles du cas uni-dimensionnel. Soit $A_{(j)}$ les données de la j^{eme} variable pour les n individus.

Moyenne arithmétique

La moyenne arithmétique des données $A_{(j)}$ de la j^{eme} variable, notée $\overline{A_{(j)}}$, est :

$$\overline{A_{(j)}} = \frac{1}{n} \sum_{i=1}^{n} a_{ij}$$

on peut alors représenter les p moyennes arithmétiques des données des p variables sous la forme du vecteur ligne des moyennes arithmétiques, noté $\bar{A}^T$:

$$\bar{A}^T = (\overline{A_{(1)}}, \dots, \overline{A_{(p)}})$$

Exemple 1.2.1 *Le vecteur ligne des moyennes arithmétiques pour l'exemple des notes est*

$$\bar{A}^T = (\frac{11 + \ldots + 16}{6}, \frac{13,5 + \ldots + 13,5}{6}) = (13,5; 13,5)$$

Médiane

On supose que les vecteurs des données $A_{(j)}$ de la j^{eme} 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 :

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

Si n est pair, on prendra par convention

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

On peut aussi mettre les p médianes dans un vecteur ligne, noté m^T, et appelé le vecteur ligne des médianes

$$m^T = (m_{(1)}, \ldots, m_{(p)})$$

Exemple 1.2.2 *Le vecteur ligne des médianes pour l'exemple des notes est*

$$m^T = (\frac{13 + 14}{2}, \frac{13,5 + 13,5}{2}) = (13,5; 13,5).$$

1.3 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 sur 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.

1.3.1 Etendue

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

$$A^{max}_{(j)} = \max_{i\in\{1,\ldots,n\}} a_{i,j} \quad \text{et} \quad A^{min}_{(j)} = \min_{i\in\{1,\ldots,n\}} a_{i,j}$$

alors

$$\omega_{(j)} = A^{max}_{(j)} - A^{min}_{(j)}.$$

On peut représenter les p étendues sous la forme d'un vecteur ligne, appelé vecteur ligne des étendues, et noté w^T :

$$\mathrm{w}^T = (\omega_{(1)}, \ldots, \omega_{(p)})$$

Exemple 1.3.1 *Le vecteur des étendues de l'exemple des notes est :*

$$\mathrm{w}^T = (5, 0)$$

Remarque 1.3.1 *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 différentes, mais cela ne représente pas bien la réalité des données.*

1.3.2 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 $A_{(j)}$ de la j^{eme} variable, l'idée est de calculer la somme, pour chacune des données de cette variable, des distances à la moyenne, et de diviser par le nombre de données. Une première idée serait de calculer :

$$\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{A_{(j)}}) = \frac{1}{n}\left[(a_{1,j}-\overline{A_{(j)}})+\ldots+(a_{n,j}-\overline{A_{(j)}})\right],$$

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 additionne de même signe, disons positif. Une idée est de prendre la valeur absolue, et on obtient alors l'**écart de la moyenne** c'est-à-dire de calculer

$$E_m = \frac{1}{n}\sum_{i=1}^{n}|a_{i,j}-\overline{A_{(j)}}| = \frac{1}{n}\left[|a_{1,j}-\overline{A_{(j)}}|+\ldots+|a_{n,j}-\overline{A_{(j)}}|\right],$$

Une autre manière de procéder est de prendre les carrés, on obtient alors la **variance** :

$$\sigma^2(A_{(j)}) := \mathrm{Var}(A_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{A_{(j)}})^2 = \frac{1}{n}\left[(a_{1,j}-\overline{A_{(j)}})^2+\ldots+(a_{n,j}-\overline{A_{(j)}})^2\right].$$

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

$$\sigma(A_{(j)}) := \sqrt{\mathrm{Var}(A_{(j)})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{A_{(j)}})^2}.$$

Exemple 1.3.2 *Voici le calcul des variances et des écart-types pour l'exmple des notes*

$$\begin{aligned}\sigma^2(A_{(1)}) &= \frac{1}{6}\left[(11-13.5)^2+(12-13.5)^2+(13-13.5)^2+(14-13.5)^2+(15-13.5)^2+(16-13.5)^2\right]\\ &= 2.917\end{aligned}$$

$$\sigma(A_{(1)}) = \sqrt{2.917} = 1.708$$

$$\sigma^2(A_{(2)}) = \frac{1}{6}\left[6(13.5-13.5)^2\right] = 0$$

$$\sigma(A_{(2)}) = \sqrt{0} = 0$$

⊗Notation matricielle
La variance s'écrit naturellement comme la norme d'un vecteur. cette interprétation géométrique est utile

pour la suite de cette analyse statistqie.
On définit alors la matrice des moyennes arithmétiques, notée $\overline{A}$, par

$$\overline{A} = \begin{pmatrix} \overline{A_{(1)}} & \dots & \overline{A_{(p)}} \\ \vdots & \ddots & \vdots \\ \overline{A_{(1)}} & \dots & \overline{A_{(p)}} \end{pmatrix}$$

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

$$A - \overline{A} = \begin{pmatrix} a_{1,1} - \overline{A_{(1)}} & \dots & a_{1,p} - \overline{A_{(p)}} \\ \vdots & \ddots & \vdots \\ a_{n,1} - \overline{A_{(1)}} & \dots & a_{n,p} - \overline{A_{(p)}} \end{pmatrix}$$

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

$$\sigma^2(A_{(j)}) = \frac{1}{n}\langle (A - \overline{A})_{(j)}, (A - \overline{A})_{(j)} \rangle = \frac{1}{n}(A - \overline{A})_{(j)}^T (A - \overline{A})_{(j)} = \frac{1}{n}\|(A - \overline{A})_{(j)}\|^2.$$

De manière analogue, l'écart-type s'écrit sous la forme :

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

Exemple 1.3.3 *Réécrivons la variance pour l'exemple des notes en notation matricielle*

$$\overline{A} = \begin{pmatrix} 13,5 & 13,5 \\ 13,5 & 13,5 \\ 13,5 & 13,5 \\ 13,5 & 13,5 \\ 13,5 & 13,5 \\ 13,5 & 13,5 \end{pmatrix}, \quad \textit{et} \quad A - \overline{A} = \begin{pmatrix} -2,5 & 0 \\ -1,5 & 0 \\ -0,5 & 0 \\ 0,5 & 0 \\ 1,5 & 0 \\ 2,5 & 0 \end{pmatrix}$$

Ainsi

$$\sigma^2(A_{(1)}) = \frac{1}{6}\langle \begin{pmatrix} -2,5 \\ -1,5 \\ -0,5 \\ 0,5 \\ 1,5 \\ 2,5 \end{pmatrix}, \begin{pmatrix} -2,5 \\ -1,5 \\ -0,5 \\ 0,5 \\ 1,5 \\ 2,5 \end{pmatrix} \rangle = \frac{1}{6}\left\| \begin{pmatrix} -2,5 \\ -1,5 \\ -0,5 \\ 0,5 \\ 1,5 \\ 2,5 \end{pmatrix} \right\|^2 = 2,917.$$

De la même manière, on trouvera $\sigma^2(A_{(2)}) = 0$.

1.3.3 Variables centrées-réduites

Les données d'une variable sont dites centrées si 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 A est la matrice des données, on notera Z la matrice des données centrées réduites. Par définition, on a : $Z = (z_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq p}}$ avec

$$z_{i,j} = \frac{a_{i,j} - \overline{A_{(j)}}}{\sigma(\overline{A_{(j)}})}$$

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

Exemple 1.3.4 *Voici la matrice des données centrées réduites de l'exemple des notes. On se souvient que*

$$\sigma(A_{(1)}) = 1.708, \quad \sigma(A_{(2)}) = 0, \quad \overline{A_{(1)}} = 13.5, \quad \overline{A_{(2)}} = 13.5$$

Ainsi

$$Z = \begin{pmatrix} -1.464 & 0 \\ -0.878 & 0 \\ -0.293 & 0 \\ 0.293 & 0 \\ 0.878 & 0 \\ 1.464 & 0 \end{pmatrix}$$

1.4 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 $A_{(1)}, \ldots, A_{(p)}$ de p variables observées sur n individus.

1.4.1 Covariance

Pour tout $1 \leq i, j \leq p$, on définit la **covariance** entre les données $A_{(i)}$ et $A_{(j)}$ des i^{eme} et j^{eme} variables, notée $\mathrm{Cov}(A_{(i)}, A_{(j)})$, par :

$$\mathrm{Cov}(A_{(i)}, A_{(j)}) = \frac{1}{n}\langle (A - \overline{A})_{(i)}, (A - \overline{A})_{(j)} \rangle = \frac{1}{n}(A - \overline{A})_{(i)}^T (A - \overline{A})_{(j)}$$

Théorème 1.4.1 *(Köning-Huygens) La covariance est égale à :*

$$\mathrm{Cov}(A_{(i)}, A_{(j)}) = \left(\frac{1}{n}\langle A_{(i)}, A_{(j)} \rangle\right) - \overline{A_{(i)}}\ \overline{A_{(j)}}$$

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

$$\begin{aligned}
\mathrm{Cov}(A_{(i)}, A_{(j)}) &= \frac{1}{n}\langle (A_{(i)} - \overline{A}_{(i)}, A_{(j)} - \overline{A}_{(j)} \rangle \\
&= \frac{1}{n}\langle A_{(i)} - \overline{A_{(i)}}\ \mathbf{1}, A_{(j)} - \overline{A_{(j)}}\ \mathbf{1} \rangle \\
&= \frac{1}{n}\left[\langle A_{(i)}, A_{(j)} \rangle - \overline{A_{(i)}}\langle \mathbf{1}, A_{(j)} \rangle - \overline{A_{(j)}}\langle A_{(i)}, \mathbf{1} \rangle + \overline{A_{(i)}}\ \overline{A_{(j)}}\langle \mathbf{1}, \mathbf{1} \rangle\right] \\
&= \frac{1}{n}\left[\langle A_{(i)}, A_{(j)} \rangle - n\overline{A_{(i)}}\ \overline{A_{(j)}} - n\overline{A_{(j)}}\ \overline{A_{(i)}} + n\overline{A_{(i)}}\ \overline{A_{(j)}}\right] \\
&= \left(\frac{1}{n}\langle A_{(i)}, A_{(j)} \rangle\right) - \overline{A_{(i)}}\ \overline{A_{(j)}}
\end{aligned}$$

car $\langle A_{(i)}, \mathbf{1}\rangle = n\ \overline{A_{(i)}}$, $\langle \mathbf{1}, A_{(j)}\rangle = n\ \overline{A_{(j)}}$ et $\langle \mathbf{1}, \mathbf{1}\rangle = n$. □

Remarque 1.4.1 *1.* $\mathrm{Cov}(A_{(i)}, A_{(j)}) = \frac{1}{n}(A - \overline{A})^T_{(i)}(A - \overline{A})_{(j)}$*, c'est-à-dire* $\mathrm{Cov}(A_{(i)}, A_{(j)})$ *est le coefficient* (i, j) *de la matrice* $X = \frac{1}{n}(A - \overline{A})^T(A - \overline{A})$.

2. $\mathrm{Cov}(A_{(i)}, A_{(i)}) = \sigma^2(A_{(i)})$.

3. La matrice de covariance est symétrique. c'est-à-dire $\mathrm{Cov}(A_{(i)}, A_{(j)}) = \mathrm{Cov}(A_{(j)}, A_{(i)})$.

4. Dans ce cas de la variance, le théorème de köning-Huygens s'écrit :

$$\sigma^2(A_{(i)}) = \frac{1}{n}\|A_{(i)}\|^2 - \overline{A_{(i)}}^2.$$

Exemple 1.4.1 *Calculons la covariance entre les données des première et deuxième variables de l'exemple des notes, en utilisant le théorème de Köning-Huygens :*

$$\mathrm{Cov}(A_{(1)}, A_{(2)}) = \frac{1}{6}\left(11.13,5 + 12.13,5 + 13.13,5 + 14.13,5 + 15.13,5 + 16.13,5\right) - 13,5^2 = 0$$

Matrice de Covariance

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

$$C(A) = \frac{1}{n}(A - \overline{A})^T(A - \overline{A})$$

de sorte que l'on a

$$\mathrm{Cov}(A_{(i)}, A_{(j)}) = (C(A))_{i,j}$$

Remarquer que les coefficients sur la diagonale de la matrice $C(A)$ donnent les variances.

Exemple 1.4.2 *Calculons la matrice de covariance pour l'exemple des notes.*

$$C(A) = \frac{1}{6}(A - \overline{A})^T(A - \overline{A}) = \frac{1}{6}\begin{pmatrix} -2,5 & -1,5 & -0,5 & 0,5 & 1,5 & 2,5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -2,5 & 0 \\ -1,5 & 0 \\ -0,5 & 0 \\ 0,5 & 0 \\ 1,5 & 0 \\ 2,5 & 0 \end{pmatrix}$$

d'où la matrice de covariance

$$C(A) = \begin{pmatrix} 2,91667 & 0 \\ 0 & 0 \end{pmatrix}$$

Ainsi, on retrouve $\sigma^2(A_{(1)}) = (C(A))_{1,1} = 2.917$, $\sigma^2(A_{(2)}) = (C(A))_{2,2} = 0$ *et*

$$\mathrm{Cov}(A_{(1)}, A_{(2)}) = (C(A))_{1,2} = (C(A))_{2,1} = \mathrm{Cov}(A_{(2)}, A_{(1)}) = 0.$$

Variabilité totale de la matrice des données A

La variabilité totale de la matrice des données A est la trace de la matrice de covariance, c'est-à-dire

$$\mathrm{Tr}(C(A)) = \sum_{i=1}^{p} \sigma^2(A_{(i)}).$$

Cette quantité est importante car elle donne en quelque sorte la quantité d'information qui est contenue dans la matrice des données A. Elle joue un rôle clé dans l'analyse par composante principale.

1.4.2 Corrélation de Bravais-Pearson

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

$$\mathrm{r}(A_{(i)}, A_{(j)}) = \frac{\mathrm{Cov}(A_{(i)}, A_{(j)})}{\sigma(A_{(i)})\sigma(A_{(j)})} = \frac{\langle (A-\overline{A})_{(i)}, (A-\overline{A})_{(j)} \rangle}{\|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(j)}\|} = \cos\left((A-\overline{A})_{(i)}, (A-\overline{A})_{(j)}\right)$$

Propriété 1.4.1 *La corrélation de Bravais-Pearson satisfait les propriétés suivantes :*

1. $\mathrm{r}(A_{(i)}, A_{(i)}) = 1$ *pour tout* $1 \leq i \leq p$.
2. $|\mathrm{r}(A_{(i)}, A_{(j)})| \leq 1$ *pour tout* $1 \leq i, j \leq p$
3. $|\mathrm{r}(A_{(i)}, A_{(j)})| = 1$*, si et seulement si il existe un nombre* $\alpha \in \mathbb{R}$ *tel que*

$$(A-\overline{A})_{(j)} = \alpha (A-\overline{A})_{(i)}$$

Démonstration.

1. Pour ce point il suffit de prendre $j = i$ dans l'expression de $\mathrm{r}(A_{(i)}, A_{(j)})$ et on obtient

$$\mathrm{r}(A_{(i)}, A_{(i)}) = \frac{\mathrm{Cov}(A_{(i)}, A_{(i)})}{\sigma(A_{(i)})\sigma(A_{(i)})} = \frac{\langle (A-\overline{A})_{(i)}, (A-\overline{A})_{(i)} \rangle}{\|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(i)}\|} = \frac{\|(A-\overline{A})_{(i)}\|^2}{\|(A-\overline{A})_{(i)}\|^2} = 1$$

2. Pour le deuxième point, on utilisera l'inégalité de Cauchy-Schwarz :

$$|\langle (A-\overline{A})_{(i)}, (A-\overline{A})_{(j)} \rangle| \leq \|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(j)}\|$$

   alors

$$|\mathrm{r}(A_{(i)}, A_{(j)})| = \frac{|\langle (A-\overline{A})_{(i)}, (A-\overline{A})_{(j)} \rangle|}{\|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(j)}\|} \leq \frac{\|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(j)}\|}{\|(A-\overline{A})_{(i)}\| \|(A-\overline{A})_{(j)}\|} = 1.$$

3. $|\mathrm{r}(A_{(i)}, A_{(j)})| = 1$ si et seulement si $\mathrm{r}(A_{(i)}, A_{(j)}) = \pm 1$.
   si et seulement si $\cos\left((A-\overline{A})_{(i)}, (A-\overline{A})_{(j)}\right) = \pm 1$.
   si et seulement si l'angle $\left((A-\overline{A})_{(i)}, (A-\overline{A})_{(j)}\right) = k\pi$ avec $k \in \mathbb{Z}$.
   c'est-à-dire que $(A-\overline{A})_{(i)}$ et $(A-\overline{A})_{(j)}$ sont colinéaires.

□

Matrice de Corrélation

De manière analogue à la matrice de covariance, on définit la matrice de corrélation , de taille $(p \times p)$, notée $\mathbf{R}(A)$, par :

$$\mathbf{R}(A) = \left[\mathrm{r}(A_{(i)}, A_{(j)})\right]_{\substack{1 \leq i \leq p \\ 1 \leq j \leq p}}$$

Via les propriété du coefficient de corrélation, on remarque que les éléments diagonaux de la matrice de corrélation sont tous égaux à 1.

Exemple 1.4.3 *La matrice de corrélation de l'exemple des notes est*

$$\mathbf{R}(A) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Chapitre 2

Méthode des moindres carrées : Régression simple et multiple

2.1 Régression simple

Dans les sciences expérimentales, il est souvent nécessaire de ”résoudre” des systèmes qui n'ont pas de solution où bien qui ont une infinité de solution (la solution n'est pas unique). Supposons par exemple que les mesures x et y de deux grandeurs soient, d'après une loi connue, liées par une relation :

1. affine de type : $y = mx + p$ de paramètre m et p inconnus.
2. quadratique de type parabole : $y = ax^2 + bx + c$ de paramètres a, b et c inconnus.

Définition 2.1.1 *1. On appelle* **régression linéaire simpe**, *toute droite affine d'équation*

$$y = mx + p$$

de paramètres inconnus m et p, approchant le nuage de points (x_M, y_M) d'un domaine Ω de $\mathbb{R}^2$.

2. On appelle **régression quadratique simpe**, *toute parabole d'équation*

$$y = ax^2 + bx + c$$

de paramètres inconnus a, b et c, approchant le nuage de points (x_M, y_M) d'un domaine Ω de $\mathbb{R}^2$.

2.1.1 Régression linéaire simple

Soit $M_i = (x_i, y_i)$ avec $1 \leq i \leq N$ le nuage de points d'un domaine Ω de $\mathbb{R}^2$. On suppose que les points $(x_i, y_i)_v$ sont le résultat de N expériences indépendantes. On cherche la droite affine passant par un nombre maximale de point M_i et qui approche tous les autres points qui restent. En général, le système

$$y_i = mx_i + p, \quad 1 \leq i \leq N \tag{0.1}$$

n'a pas de solution. On cherche alors une **bonne approximation** des valeurs inconnues m et p. C'est-à-dire qu'il s'agit de trouver une solution fiable (m, p) au système de N équations et à deux incnnus (0.1). La méthode des moindres carrés est un des procédés permettant de résoudre ce genre de problème.

Coût d'énergie relative à la régression linéaire

Lors de N expériences indépendantes efféctuées, on obtient un écart d'erreur ε entre la vraie valeur de y et son approché par la méthode des moindres carrés, noté $\hat{y} = mx + p$. On a pour tout $1 \leq i \leq N$, $\varepsilon_i = y_i - \hat{y}_i$. Ainsi, on obtient un coût d'énergie qu'on note $\mathcal{J}(m,p)$ défini comme la somme des carrés des erreurs ε_i pour tout $1 \leq i \leq N$. Soit

$$\mathcal{J}(m,p) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - mx_i - p)^2.$$

L'application $\mathcal{J}$ est une fonction à deux variables allant de $\mathbb{R}^2$ à valeurs dans $[0, +\infty[$ et qu'il s'agit d'une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^2$.

Définition 2.1.2 *On appelle solution obtenue par la méthode des moindres carrés, la solution* (m_0, p_0) *telle que*

$$\mathcal{J}(m_0, p_0) = \min_{(m,p)\in\mathbb{R}^2} \mathcal{J}(m,p).$$

Solution par la méthode des moindres carrés

La fonctionnelle $\mathcal{J}$ étant de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^2$, alors nous pouvons trouver less points critiques de $\mathcal{J}$ par résoudre l'équation vectorielle à deux inconnus (m,p) donnée par :

$$\nabla \mathcal{J}(m,p) = 0_{\mathbb{R}^2}$$

où

$$\nabla \mathcal{J}(m,p) = \begin{pmatrix} \frac{\partial \mathcal{J}}{\partial m}(m,p) \\ \\ \frac{\partial \mathcal{J}}{\partial p}(m,p) \end{pmatrix}.$$

Un peu de calcul des dérivées partielles nous permet de trouver l'ensemble $\mathcal{E}_c$ des points critiques de la fonctionnelle $\mathcal{J}(m,p)$.

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial m}(m,p) &= -2\sum_{i=1}^{N} x_i(y_i - mx_i - p) = 0, \\ \frac{\partial \mathcal{J}}{\partial p}(m,p) &= -2\sum_{i=1}^{N} (y_i - mx_i - p) = 0, \end{aligned}$$

ce qui conduit vers le système suivant

$$\begin{cases} \displaystyle\sum_{i=1}^{N} x_i y_i &= m\left(\sum_{i=1}^{N} x_i^2\right) + p\left(\sum_{i=1}^{N} x_i\right), \\ \displaystyle\sum_{i=1}^{N} y_i &= m\left(\sum_{i=1}^{N} x_i\right) + Np \end{cases}$$

d'où le système matricielle suivant :

$$\begin{pmatrix} \displaystyle\sum_{i=1}^{N} x_i^2 & \displaystyle\sum_{i=1}^{N} x_i \\ \displaystyle\sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} m \\ \\ p \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N} x_i y_i \\ \displaystyle\sum_{i=1}^{N} y_i \end{pmatrix} \tag{0.2}$$

Le pproblème d'approximation par régression linéaire admet une unique solution sit et seulement si le système matricielle (0.2) admet une unique solution. C'est-à-dire que la matrice

$$A = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix}$$

est inversible. Dans ce cas, on obtient

$$\begin{cases} m = \dfrac{N\left(\sum_{i=1}^{N} x_i y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}, \\ p = \dfrac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}, \end{cases}$$

Remarque 2.1.1 *Si la matrice A n'était pas inversible, alors on procède à la résolution par la méthode du pseudo-inverse de Moore-Penrose.Ainsi de trouver le meilleur couple $(m^\dagger, p^\dagger)$ de parmi tous les points critiques de la fonctionnelle $\mathcal{J}(m,p)$.*

Exemple 2.1.1 *1. Soit $N = 3$ et on considère les points $(x_1, y_1) = (1, 2)$, $(x_2, y_2) = (-1, 0)$ et $(x_3, y_3) = (2, -1)$.*
Le système

$$\begin{cases} 2 = m + p, \\ 0 = -m + p, \\ -1 = 2m + p \end{cases}$$

n'admet pas de solution, en effet, la solution des deux premières équations du système est $(m, p) = (1, 1)$ alors que le couple $(1, 1)$ ne satisfait pas la troisième équation. La somme

$$\mathcal{J}(m,p) = (2 - m - p)^2 + (m - p)^2 + (-1 - 2m - p)^2 = 6m^2 + 4pm + 3p^2 - 2p + 5$$

est, pour toute valeur de p, un trinôme en m dont le minimum obtenu pour $m = -\frac{p}{3}$ est égal à :

$$\mathcal{J}(-\frac{p}{3}, p) = \frac{7}{3}p^2 - 2p + 5.$$

Le minimum de $\mathcal{J}(m,p)$ atteint si

$$p = \frac{3}{7}, \quad m = -\frac{p}{3} = -\frac{1}{7}$$

est égal à $\mathcal{J}(-\frac{1}{7}, \frac{3}{7}) = \frac{32}{7}$.

2. *On pourra pour ce même exemple calculer la matrice* $A = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$ *et le vecteur* $b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. *Alors*

$$\begin{cases} m &= \frac{0-2}{3(1+1+4)-2^2} = -\frac{2}{14} = -\frac{1}{7}, \\ p &= \frac{1.6-2.0}{3(1+1+4)-2^2} = \frac{6}{14} = \frac{3}{7}, \end{cases}$$

D'où la droite de régression approchant y *est* $\hat{y} = -\frac{1}{7}x + \frac{3}{7}$.

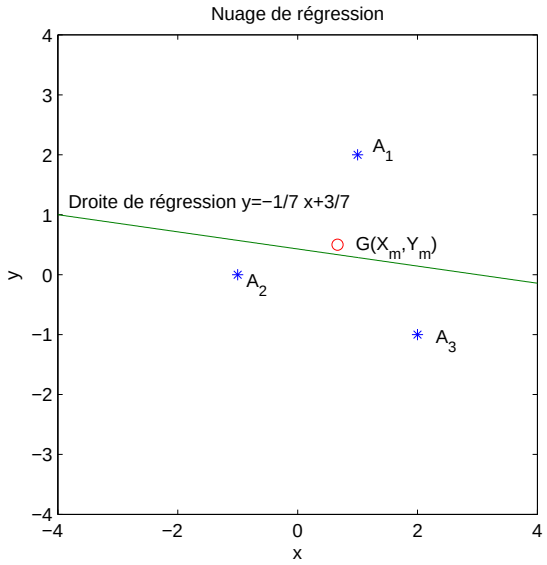

FIGURE 2.1 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$, $Y_m = \bar{Y}$ et $G(X_m, Y_m)$ est le centre de gravité de A_1, A_2 et A_3.

2.1.2 Régression quadratique simple

Soit $M_i = (x_i, y_i)$ avec $1 \leq i \leq N$ le nuage de points d'un domaine Ω de $\mathbb{R}^2$. On suppose que les points $(x_i, y_i)_{1 \leq i \leq N}$ sont le résultat de N expériences indépendantes. On cherche la droite affine passant par un nombre maximale de point M_i et qui approche tous les autres points qui restent. En général, le système

$$y_i = ax_i^2 + bx_i + c, \quad 1 \leq i \leq N \tag{0.3}$$

n'a pas de solution. On cherche alors une **bonne approximation** des valeurs inconnues a, b et c. C'est-à-dire qu'il s'agit de trouver une bonne solution (a, b, c) au système de N équations et à trois incnnus (0.3). La méthode des moindres carrés est un des procédés permettant de résoudre ce genre de problème.

Coût d'énergie relative à la régression quadratique

Lors de N expériences indépendantes effectuées, on obtient un écart d'erreur ε entre la vraie valeur de y et son approché par la méthode des moindres carrés, noté $\hat{y} = ax^2 + bx + c$. On a pour tout $1 \leq i \leq N$,

$\varepsilon_i = y_i - \hat{y}_i$. Ainsi, on obtient un coût d'énergie qu'on note $\mathcal{J}(a,b,c)$ défini comme la somme des carrés des erreurs ε_i pour tout $1 \leq i \leq N$. Soit

$$\mathcal{J}(a,b,c) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - ax_i^2 - bx_i - c)^2.$$

L'application $\mathcal{J}$ est une fonction à deux variables allant de $\mathbb{R}^3$ à valeurs dans $[0, +\infty[$ et qu'il s'agit d'une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^3$.

Définition 2.1.3 *On appelle solution obtenue par la méthode des moindres carrés, la solution* $(\hat{a}, \hat{b}, \hat{c})$ *telle que*

$$\mathcal{J}(\hat{a}, \hat{b}, \hat{c}) = \min_{(a,b,c)\in\mathbb{R}^3} \mathcal{J}(a,b,c).$$

Solution par la méthode des moindres carrés

La fonctionnelle $\mathcal{J}$ étant de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^3$, alors nous pouvons trouver less points critiques de $\mathcal{J}$ par résoudre l'équation vectorielle à deux inconnus (m,p) donnée par :

$$\nabla\mathcal{J}(a,b,c) = 0_{\mathbb{R}^3}$$

où

$$\nabla\mathcal{J}(a,b,c) = \begin{pmatrix} \frac{\partial\mathcal{J}}{\partial a}(a,b,c) \\ \\ \frac{\partial\mathcal{J}}{\partial b}(a,b,c) \\ \\ \frac{\partial\mathcal{J}}{\partial c}(a,b,c) \end{pmatrix}.$$

Un peu de calcul des dérivées partielles nous permet de trouver l'ensemble $\mathcal{E}_c$ des points critiques de la fonctionnelle $\mathcal{J}(a,b,c)$.

$$\begin{aligned} \frac{\partial\mathcal{J}}{\partial a}(a,b,c) &= -2\sum_{i=1}^{N} x_i^2(y_i - ax_i^2 - bx_i - c) = 0, \\ \frac{\partial\mathcal{J}}{\partial b}(a,b,c) &= -2\sum_{i=1}^{N} x_i(y_i - ax_i^2 - bx_i - c) = 0, \\ \frac{\partial\mathcal{J}}{\partial b}(a,b,c) &= -2\sum_{i=1}^{N} (y_i - ax_i^2 - bx_i - c) = 0, \end{aligned}$$

ce qui conduit vers le système suivant

$$\left\{ \begin{array}{rcl} \displaystyle\sum_{i=1}^{N} x_i^2 y_i &=& \displaystyle a\left(\sum_{i=1}^{N} x_i^4\right) + b\left(\sum_{i=1}^{N} x_i^3\right) + c\left(\sum_{i=1}^{N} x_i^2\right), \\ \displaystyle\sum_{i=1}^{N} x_i y_i &=& \displaystyle a\left(\sum_{i=1}^{N} x_i^3\right) + b\left(\sum_{i=1}^{N} x_i^2\right) + c\left(\sum_{i=1}^{N} x_i\right), \\ \displaystyle\sum_{i=1}^{N} y_i &=& \displaystyle a\left(\sum_{i=1}^{N} x_i^2\right) + b\left(\sum_{i=1}^{N} x_i\right) + cN \end{array} \right.$$

d'où le système matricielle suivant :

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 y_i \\ \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \tag{0.4}$$

Le pproblème d'approximation par régression linéaire admet une unique solution sit et seulement si le système matricielle (0.4) admet une unique solution. C'est-à-dire que la matrice

$$A = \begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i & N \end{pmatrix}$$

est inversible.

Remarque 2.1.2 *Si la matrice A n'était pas inversible, alors on procède à la résolution par la méthode du pseudo-inverse de Moore-Penrose.Ainsi de trouver le meilleur couple $(a^\dagger, b^\dagger, c^\dagger)$ de parmi tous les points critiques de la fonctionnelle $\mathcal{J}(a, b, c)$.*

Exemple 2.1.2 *Dans un plan vectoriel rapporté à un repère donné, on considère l'ensemble Ω de cinq points de coordonnées*

$$(-1, 2.5), \quad (0, 1.25), \quad (1, 1), \quad (2, -1.5) \quad \textit{et} \quad (3, -4).$$

Il s'agit de $N = 5$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 2$ et $x_5 = 3$, puis $y_1 = 2.5$, $y_2 = 1.25$, $y_3 = 1$, $y_4 = -1.5$ et $y_5 = -4$. Alors, on peut calculer la matrice A et le vecteur b :

$$A = \begin{pmatrix} 99 & 35 & 15 \\ 35 & 15 & 5 \\ 15 & 5 & 5 \end{pmatrix} \quad \textit{et} \quad b = \begin{pmatrix} -38.5 \\ -16.5 \\ -0.75 \end{pmatrix}$$

D'où le système linéaire suivant :

$$\begin{pmatrix} 99 & 35 & 15 \\ 35 & 15 & 5 \\ 15 & 5 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -43.5 \\ -11.5 \\ -5.75 \end{pmatrix} \tag{0.5}$$

D'où $a = -1.0536$, $b = 1.5321$ et $c = 0.4786$. Ainsi la parabole de régression

$$\hat{y} = -1.0536\, x^2 + 1.5321\, x + 0.4786$$

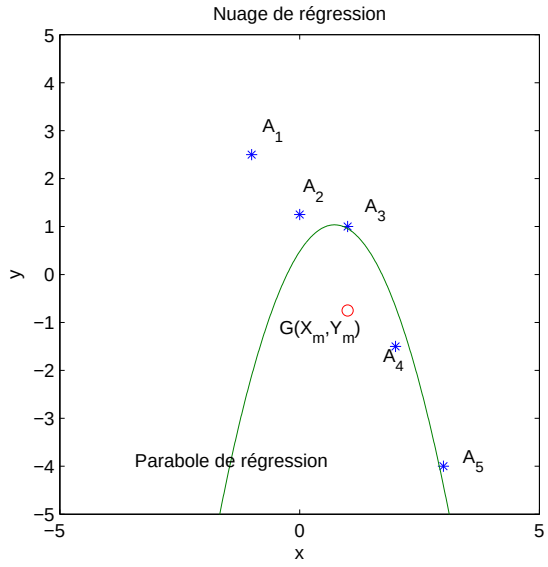

FIGURE 2.2 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$, $Y_m = \bar{Y}$ et $G(X_m, Y_m)$ est le centre de gravité de $A_1, \ldots, A_5$.

Définition 2.1.4 *Soit y une variable explicative par N tests expérimentals. On suppose que est définie par une des régressions simples linéaire où bien quadratique. Soit $\hat{y}$ l'approché de y selon la méthode des moindres carrés.*

1. *On appelle la Somme des Carrées des Ecarts (SCE) des résidus $y - \hat{y}$ la quantité définie par*

$$\mathrm{SCE}_{residu} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

2. *On appelle la Somme des Carrées des Ecarts (SCE) totale de y, la quantité définie par*

$$\mathrm{SCE}_y = \sum_{i=1}^{N} (y_i - \bar{y})^2,$$

   avec $\bar{y}$ est la moyenne arithmétique de y sur N tests expérimentals.

3. *On appelle la Somme des Carrées des Ecarts (SCE) expliquée par la régression la quantité définie par*

$$\mathrm{SCE}_{regr} = \mathrm{SCE}_y - \mathrm{SCE}_{residu}$$

4. *On appelle le* **Carré Moyen** *dû à la régression :*

$$\mathrm{CM}_{regr} = \frac{\mathrm{SCE}_{regr}}{p}$$

   avec p désigne le nombre des variables explicatives.

5. *On appelle le* **Carré Moyen** *des résidus :*

$$\mathrm{CM}_{residu} = \frac{\mathrm{SCE}_{residu}}{n - p - 1}$$

   avec n est le nombre d'individus.

6. *On appelle F calculé de Fisher Snedecor, notée $\mathrm{F}_{calcule}$, la valeur de F correspondant à p et $n-p-1$ degrés de liberté :*
$$\mathrm{F}_{calcule} = \frac{\mathrm{CM}_{regr}}{\mathrm{CM}_{residu}} = \frac{n-p-1}{p}\frac{\mathrm{SCE}_{regr}}{\mathrm{SCE}_{residu}}$$
7. *On appelle* **intensité dû à la régression** *où bien le* **coefficient de détermination**, *notée $I = R^2$, la quantité définie par :*
$$\mathrm{I} = \mathrm{R}^2 = \frac{\mathrm{SCE}_{regr}}{\mathrm{SCE}_{y}}$$
8. *On appelle le* **coefficient de détermination ajusté**, *notée $R^2_{ajusté}$, la quantité définie par :*
$$R^2_{ajusté} = 1 - \frac{n-1}{n-p}(1-R^2).$$

2.1.3 Coefficients de régression standardisés

L'équation de régression simple à une seule variable explicative X est donnée par :
$$\hat{Y} = aX + b$$
où a est le coefficient de régression et b est l'ordonnée à l'origine(valeur de $\hat{Y}$ pour $X = 0$).

Définition 2.1.5 *Soit X une variable explicative.*
1. *On dit que X est centrée si et seulement si $E(X) = \bar{X} = 0$.*
2. *On dit que X est réduite si et seulement si $\sigma^2(X) = 1$.*

Définition 2.1.6 *Soit X une variable explicative de moyenne $\bar{X}$ et de variance $\sigma^2(X)$.*
On appelle **variable explicative standardisée** *relative à X, notée Z, la variable explicative centré-réduite exprimée par :*
$$Z = \frac{X - \bar{X}}{\sigma(X)}$$

Le coefficient b disparaît si Y et X sont standarisés (centrés-réduites), puisque la standardisation (centrage et réduction des variables) opère un changement d'origine.
L'équation de régression standardisée est :
$$\hat{Y} = \alpha Z$$
ainsi, on peut obtenir le coefficient de régression α sans passer par la standardisation des variables grâce à la relation :
$$\alpha = a\ \frac{\sigma_X}{\sigma_Y}.$$

Définition 2.1.7 *On appelle coefficient de* **corrélation de Bravais-Pearson** *entre y et X, noté r_{YX}, la quantité calculée par l'expression :*
$$\mathrm{r}_{YX} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$
Le coefficient de détermination R^2 sera calculer encore par $R^2 = \mathrm{R}^2_{YX}$

Individus/ variables	Y	X
1	20	10
2	82	40
3	44	20
4	65	30
5	25	15
Somme	236	115

Propriété 2.1.1 *1. Le facteur $R^2.100$ désigne le pourcentage du nuage de régression expliqué par la droite de régression*

$$Y = aX + b$$

2. Si $R^2.100$ est proche de 100%, alors la droite de régression explique bien le nuage de points où bien la droite passe très procche de tous les points.

3. Il est donc possible d'utiliser cette droite pour résumer le nuage de régression.

Exemple 2.1.3 *On considère le tableau de valeurs suivant :*

Le tableau des caalculs permettant de trouver les éléments d'une régression sur des données centrées (x, y) et non centrée (X, Y) est le suivant :

Ind	Y	X	X^2	XY	Y^2	$y = Y - \bar{Y}$	$x = X - \bar{X}$	y^2	x^2	xy	$\hat{y}$
1	20	10	100	200	400	-27.20	-13.00	739.84	169.00	353.60	18.99
2	82	40	1600	3280	6724	34.80	17.00	1211.04	289.00	591.60	84.09
3	44	20	400	880	1936	-3.20	-3.00	10.24	9.00	9.60	40.69
4	65	30	900	1950	4225	17.80	7.00	316.84	49.00	124.60	62.39
5	25	15	225	375	625	-22.20	-8.00	492.84	64.00	177.60	29.84
S	236	115	3225	6685	13910	0.00	0.00	2770.80	580.00	1257.00	236.00

Il permet de calculer les caractéristiques qui conduisent aux paramètres de la régression :

$$\begin{aligned}
\bar{Y} &= \frac{1}{5}\sum_{k=1}^{5} Y_k = \frac{236}{5} = 47.2 \\
\bar{X} &= \frac{1}{5}\sum_{k=1}^{5} X_k = \frac{115}{5} = 23 \\
\sigma^2(Y) &= \frac{1}{5}\sum_{k=1}^{5} Y_k^2 - \bar{Y}^2 = \frac{13910}{5} - (47.2)^2 = 554.16 \\
\sigma(Y) &= \sqrt{554.16} = 23.54 \\
\sigma^2(X) &= \frac{1}{5}\sum_{k=1}^{5} X_k^2 - \bar{X}^2 = \frac{3225}{5} - (23)^2 = 116 \\
\sigma(X) &= \sqrt{116} = 10.77
\end{aligned}$$

$$\begin{aligned} \mathrm{r}_{YX} &= \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = 0.9916 \\ \mathrm{r}_{XY} &= \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{1}{n}\frac{\sum_{k=1}^{n} X_k Y_k - n\bar{X}.\bar{Y}}{\sigma(X)\sigma(Y)} = \frac{1}{5}\frac{6685 - 5 * 47.2 * 23}{10.77 * 23.54} = 0.9916 \\ R^2 &= \mathrm{r}_{XY}^2 = (0.9916)^2 = 0.9833 \end{aligned}$$

Le coefficient de détermination R^2, nous indique que 98.33% du nuage de régression est expliqué par la droite de régression $Y = aX + b$.
La méthode de calcul des paramètres a et b de la droite de régression consiste à minimiser la somme des carrés des résidus entre les valeurs observées Y_k et les valeurs calculées $\hat{Y}_k$.
On démontre que

$$\hat{a} = R\frac{\sigma(Y)}{\sigma(X)} = 0.9916\ \frac{23.54}{10.77} = 2.17$$

ainsi

$$\hat{b} = \bar{Y} - \hat{a}\bar{X} = 47.2 - 2.17 * 23 = -2.71$$

La droite de régression passe par le point $G(\bar{X}, \bar{Y})$ qui est le centre de gravité du nuage des points des individus.

$$\hat{Y} = 2.17X - 2.71$$

Le nuage de régression permet de connaître l'information concernant les individus du tableau. Par exemple, on visualise le point A_1 proche de A_5 et le point A_5 et le point A_1 sont loins du point A_2. Il est possible aussi de quantifier cette information en calculant toutes les distances au carré (théorème de Pythagore) entre les paires de points et de les classer par ordre croissant.
Le graphe de régression montre que le nuage de points est inséré dans une ellipse aux bords aplatis, ce qui signifie que ce nuage peut être résumé au moyen d'une droite de régression. Cette observation est confirmée par le calcul du coefficient de corrélation $R = 0.9916$, ce qui signifie qu'il existe une relation étroite et positive entre X et Y. Il est donc possible de substituer au nuage de régression, la droite $\hat{Y} = 2.17X - 2.71$ ou encore la droite sur variables centrées $\hat{y} = 2.17x$ qui a pour origine le point $G(\bar{X}, \bar{Y})$. (Voir les détails sur le tableau)

On peut donc calculer les projections au sens des moindres carrés (parallèlement à l'axe des ordonnées) des 5 points sur la droite de régression.
Ces projections sont données pour les variables non centrées par les calculs $\hat{Y}_1, \ldots, \hat{Y}_5$. On constate alors que si on calcule la distance, par exemple, entre $\hat{Y}_1$ et $\hat{Y}_5$ au carré, on trouve environ celle du nuage de régression entre le point A_1 et le point A_5.
*Par conséquent, l'information concernant les 5 points sur l'axe $\hat{Y}$ est conservée par rapport à celle du nuage de régression. On peut donc dire que l'analyse de données a eu lieu puisqu'e l'***information est pratiquement identique sur l'axe que dans le plan***.*
On peut aussi résumer l'information contenue dans le nuage de points en utilisant non pas les projections sur la droite de régression des points au sens des méthodes de moindres carrés, mais leurs projections orthogonales sur cette même droite, en concervant pour origine de l'axe, le point G et en construisant un vecteur unitaire dont on connaît les coordonnées dans l'espace $\mathbb{R}^2$; les projections orthogonales des 5 points sur cette droite dont données par le **produit scalaire** *entre le vecteur unitaire et un vecteur qui a pour origine le point G et pour extrémité le point à projeter. On pourrait constater que , dans ce cas aussi, la distance au carré par exemple entre le point A_1 et le point A_5 projetés est approximativement identique*

à celle du plan entre les mêmes points . L'analyse de données est donc encore réalisable en procédant da la sorte.

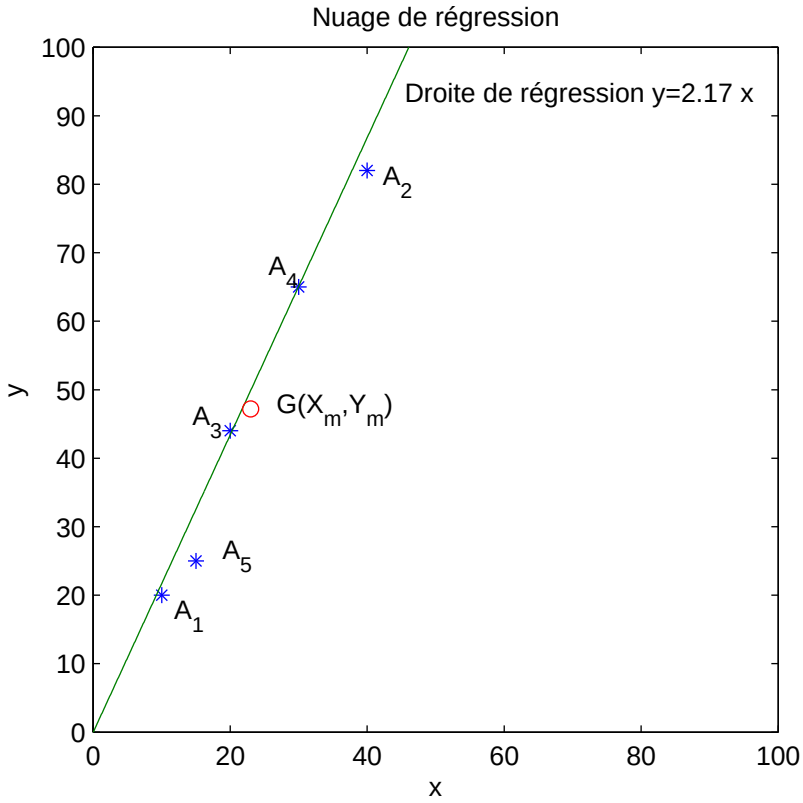

FIGURE 2.3 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$ et $Y_m = \bar{Y}$

Remarque 2.1.3 *Lorsque l'on travaille sur les variables centrées, on a les coordonnées suivantes des vecteurs $\vec{x}$ et $\vec{y}$:*

$$X - \bar{X} = \vec{x} = \begin{pmatrix} -13 \\ 17 \\ -3 \\ 7 \\ -8 \end{pmatrix} \quad \textit{et} \quad Y - \bar{Y} = \vec{y} = \begin{pmatrix} -27.2 \\ 34.8 \\ -3.2 \\ 17.8 \\ -22.2 \end{pmatrix}$$

Le produit scalaire entre les vecteurs $\vec{x}$ et $\vec{y}$ s'écrit :

$$(\vec{x}, \vec{y}) = \sum_{k=1}^{5} x_k y_k = 1257.$$

De ce fait :

$$\frac{1}{n}(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{k=1}^{n} x_k y_k = \textit{Cov}(x, y).$$

D'où :

$$\sigma^2(x) = \textit{Cov}(x, x) = \frac{1}{n} \sum_{k=1}^{n} x_k^2 = (\vec{x}, \vec{x}) = \frac{\|\vec{x}\|^2}{n}$$

et

$$\sigma(x) = \frac{\|\vec{x}\|}{\sqrt{n}}$$

De plus :

$$R = \frac{Cov(x,y)}{\sigma(x)\sigma(y)} = \frac{\frac{(\vec{x},\vec{y})}{n}}{\frac{\|\vec{x}\|}{\sqrt{n}} \cdot \frac{\|\vec{y}\|}{\sqrt{n}}} = \frac{(\vec{x},\vec{y})}{\|\vec{x}\|\|\vec{y}\|}.$$

Par ailleurs on sait que

$$(\vec{x},\vec{y}) = \|\vec{x}\|\|\vec{y}\|\cos(\theta)$$

avec θ est l'angle formé par les deux vecteurs $\vec{x}$ et $\vec{y}$. D'où

$$R = \cos(\theta)$$

Ainsi, lorsque les variables sont centrées, le coefficient de corrélation entre les 2 variables est égal au cosinus de l'angle formé par les vecteurs représentant ces variables. *Quand on centre et on réduit des variable sous leurs formes centrée-réduite*

$$y_k = \frac{Y_k - \bar{Y}}{\sigma(Y)}$$

on forme des vecteurs qui ont tous la même dimension ($\sigma^2(y) = 1$). De ce fait, la variance est la distance commune à tous les vecteurs (ils se situent sur un cercle de rayon 1) et ils se positionnent les uns par rapport aux autres par le coefficient de corrélation linéaire que l'on déduit à partir de l'angle formé par les deux vecteurs.

2.2 Régression multiple

La science thématiquement combinatoire sur une base spatiale, elle offre beaucoup plus d'exemples où une répartition est "explicable" par la conjonction de plusieurs facteurs : il faut, par conséquent, passer d'un modèle de régression simple à un modèle de régression multiple, où plusieurs variables "explicatives" notées $X_1, \ldots, X_p$ rendent compte de la variabilité de Y, variable "à expliquer" (Y et les X_j étant des variables quantitatives continues connues par individu).

Définition 2.2.1 *Soient $X_1, \ldots, X_p$ des variables et Y une autre variable*

1. *Les $(X_j)_{1\leq j\leq p}$ sont appelées des variables explicatives et elles sont quantitatives où bien qualitatives.*
2. *La variable Y est appelée une variable expliquée par les $X_1, \ldots, X_p$ si Y dépende des.*
3. *p est le nombre de variables explicatives.*

La régression multiple est une extension du modèle de régression simple, à une différence près : alors que la régression simple est symétrique (on peut permuter les rôles de Y et X, tour à tour variables à expliquer et explicative), la régression multiple est, elle, dissymétrique : c'est bien la distribution de Y qu'il s'agit d'expliquer par celles des (X_j)

2.2.1 Equation de la régression multiple

La régression multiple consiste à projeter les points d'un nuage multidimensionnel sur un hyperplan (généralisation d'un plan à plus de 2 dimensions). Comme régression simple, l'ajustement des projections est réalisé par les moindres carrés, tels que soit minimale la somme des carrés des projections de Y_i sur l'hyperplan (parallèlement à l'axe de Y).
L'équation de régression multiple (avec p variables explicatives) est :

$$\hat{Y} = a_1 X_1 + \ldots + a_p X_p + b$$

où les (a_j) sont les coefficients de régression multiple et b est la valeur de Y à l'origine $0_{\mathbb{R}^p} = (0, \ldots, 0)$.
L'expression $\hat{Y}$ signifie la valeur approchée de la variable exacte Y.

2.2.2 Coefficient de régression standardisés

Le coefficient b disparaît si y et les X_j sont **standardisé**, puisque la standardisation (centrage et réduction des variables) opère un changement d'origine (le nouvel origine devient $0_{\mathbb{R}^p} = (0, \ldots, 0)$) et d'échelle (la nouvelle unité =$(1, \ldots, 1)$).
L'équation de régression devient alors

$$\hat{Y} = \alpha_1 Z_1 + \ldots + \alpha_p Z_p$$

où Y et les Z_j sont des variables standardisées (centrés-réduites) et les α_j sont les coefficients de régression standardisés, comparables entre eux car débarrassés des effets de différences de moyenne, d'écart-type et d'unité de mesure.

On peut obtenir les coefficients de régression α_j sans passer par la standardisation des variables grâce à la relation

$$\alpha_j = a_j \, \frac{\sigma(X_j)}{\sigma(Y)}, \quad j = 1, \ldots, p$$

avec $\sigma(X_j)$ est l'écart-type de la variable X_j et $\sigma(Y)$ est l'écart-type de la variable Y.

2.2.3 Indépendance des variables explicatives

Pour qu'on puisse additionner les effets des variables explicatives et, donc, connaître la part d'explication de Y par chacune des variables explicatives X_j, il faut qu'elles soient **indépendantes** les unes des autre. Ce qui est souhaitable, c'est donc que :

- les variables explicatives X_j soient **très peu corrélées entre elles**,
- les variables explicatives X_j soient **bien corrélées** avec la variable à expliquer Y.

Ce sont des conditions à vérifier avant de poursuivre.
Et, si l'indépendance des X_j est vérifiée, alors les coefficients :

- a_j s'interprètent comme en régression simple (quand X_j augmente de 1 alors Y augmente de a_j)
- $\alpha_j = a_j \, \frac{\sigma(X_j)}{\sigma(Y)}$, $(j = 1, \ldots, p)$ indiquent la part de variance de Y due à chacun des X_j.

Et, si l'indépendance des X_j n'est pas vérifiée, il faudra se débarrasser de l'effet de leurs redondances. c'est-à-dire que si X_j et X_k ne sont pas indépendantes alors $X_k = \beta X_j$ alors

$$a_j X_j + a_k X_k = (a_j + \beta a_k) X_j = a'_j X_j$$

avec $a'_j = a_j + \beta a_k$ est le nouveau coefficient de X_j en régression multiple après avoir débarrassé de la redondance.

2.2.4 Résidus de la régression

Les résidus de la régression ($\mathcal{E}_i = Y_i - \hat{Y}_i$) doivent être considérés comme en régression simple et, comme en régression simple, il y a intérêt à étudier leur distribution (histogramme de $Y - \hat{Y}$) et à les cartographier, par exemple avec une légende en 3 classes :

1. ($Y_i - \hat{Y}_i$) très inférieur à 0, le modèle sous estime la valeur Y_i observée,
2. ($Y_i - \hat{Y}_i$) voisin de 0, le modèle estime la valeur Y_i observée,
3. ($Y_i - \hat{Y}_i$) très supérieur à 0, le modèle sur estime la valeur Y_i observée.

Les résidus, s'ils sont assez importants, peuvent traduire :

1. la nécessité d'ajouter une variable explicative oubliée,
2. l'existence d'individus "hors norme", situés loin de l'hyperplan,
3. des particularités locales,
4. l'effet d'une erreur aléatoire (d'échantillonnage ou sur les mesures).

2.2.5 Conditions de validité d'une régression multiple

Soient $(X_j)_{1 \leq j \leq p}$ des variables explicatives et Y une variable à expliquer.

Définition 2.2.2 *Deux variables explicatives ou plus satisfont la condition d'***homoscédasticité** *s'elles ont la même variance où bien elles ont à peu près la même variance.*

Pour procéder à une régression multiple pour expliquer une variable Y par des variables explicatives (X_j), il faut satisfaire les conditions suivantes :

1. la relation entre chaque variable explicative X_j et la variable à expliquer doit être **linéaire** ; si ce n'est pas le cas, il faut pratiquer une transformation des variables en relation non linéaire avec Y (carrés, log, exp,...) ou utiliser d'autres techniques (réseaux de neurons, par exemple)
2. il ne doit pas y avoir des variables **colinéaires**, c'est à dire de variables dont la somme des valeurs est égale à une constante ; par exemple, dans une régression entre revenu moyen par habitant en Y et pourcentages d'emploi dans les 3 secteurs primaire, secondaire et tertiaire, l'une de ces 3 variables explicatives doit être enlevée (car son $\%$ se déduit de 100% moins la somme des 2 autres) et les résultats n'en seront pas changés.
3. les variables explicatives doivent être **indépendantes** (avoir de très faibles corrélations entre elles) ; dans le cas contraire, il peut aussi être fait appel aux réseaux de neuron.
4. il est par contre souhaitable que chacune ait une bonne corrélation avec Y.

En cas d'**erreur aléatoire**, d'échantillonnage ou de mesure, sur Y (mais pas sur les X_j, considérés cmme dénués d'erreur aléatoire), on pourra procéder à des tests supposant, comme en régression simple,

1. la **normalité** des résidus $Y - \hat{Y}$.
2. leur **homoscédasticité** (varaince à peu près égale quelque soit l'intervalle de valeurs de $\hat{Y}$).

2.2.6 Régression multiple à trois variables explicatives

L'équation de régression multiple à trois variables explicatives est :

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + b, \tag{0.6}$$

On cherche alors une **bonne approximation** des valeurs inconnues a_1, a_2, a_3 et b. C'est-à-dire qu'il s'agit de trouver une bonne solution (a_1, a_2, a_3, b) au système à trois incnnus (0.6). Comme dans le cas d'une régression simple, la méthode des moindres carrés reste un des procédés permettant de résoudre ce type de problème.

Coût d'énergie relative à la régression quadratique

Lors de N expériences indépendantes efféctuées, on obtient un écart d'erreur ε entre la vraie valeur de Y et son approché par la méthode des moindres carrés, noté $\hat{Y} = a_1X_1 + a_2X_2 + a_3X_3 + b$. On a pour tout $1 \leq i \leq N$, $\varepsilon_i = Y_i - \hat{Y}_i$. Ainsi, on obtient un coût d'énergie qu'on note $\mathcal{J}(a_1, a_2, a_3, b)$ défini comme la somme des carrés des erreurs ε_i pour tout $1 \leq i \leq N$. Soit

$$\mathcal{J}(a_1, a_2, a_3, b) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - a_1X_1^{(i)} + a_2X_2^{(i)} + a_3X_3^{(i)} + b)^2.$$

L'application $\mathcal{J}$ est une fonction à deux variables allant de $\mathbb{R}^3$ à valeurs dans $[0, +\infty[$ et qu'il s'agit d'une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^3$.

Définition 2.2.3 *On appelle solution obtenue par la méthode des moindres carrés, la solution* $(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{b})$ *telle que*

$$\mathcal{J}(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{b}) = \min_{(a_1,a_2,a_3,b)\in\mathbb{R}^4} \mathcal{J}(a_1, a_2, a_3, b).$$

Solution par la méthode des moindres carrés

La fonctionnelle $\mathcal{J}$ étant de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^4$, alors nous pouvons trouver less points critiques de $\mathcal{J}$ par résoudre l'équation vectorielle à 4 inconnus (a_1, a_2, a_3, b) donnée par :

$$\nabla\mathcal{J}(a_1, a_2, a_3, b) = 0_{\mathbb{R}^4}$$

où

$$\nabla\mathcal{J}(a_1, a_2, a_3, b) = \begin{pmatrix} \frac{\partial\mathcal{J}}{\partial a_1}(a_1, a_2, a_3, b) \\ \frac{\partial\mathcal{J}}{\partial a_2}(a_1, a_2, a_3, b) \\ \frac{\partial\mathcal{J}}{\partial a_3}(a_1, a_2, a_3, b) \\ \frac{\partial\mathcal{J}}{\partial b}(a_1, a_2, a_3, b) \end{pmatrix}.$$

Un peu de calcul des dérivées partielles nous permet de trouver l'ensemble $\mathcal{E}_c$ des points critiques de la fonctionnelle $\mathcal{J}(a_1, a_2, a_3, b)$.

$$\begin{aligned}
\frac{\partial\mathcal{J}}{\partial a_1}(a_1, a_2, a_3, b) &= -2\sum_{i=1}^{N} X_1^{(i)}(Y_i - a_1X_1^{(i)} + a_2X_2^{(i)} + a_3X_3^{(i)} + b) = 0, \\
\frac{\partial\mathcal{J}}{\partial a_2}(a_1, a_2, a_3, b) &= -2\sum_{i=1}^{N} X_2^{(i)}(Y_i - a_1X_1^{(i)} + a_2X_2^{(i)} + a_3X_3^{(i)} + b) = 0, \\
\frac{\partial\mathcal{J}}{\partial a_3}(a_1, a_2, a_3, b) &= -2\sum_{i=1}^{N} X_3^{(i)}(Y_i - a_1X_1^{(i)} + a_2X_2^{(i)} + a_3X_3^{(i)} + b) = 0, \\
\frac{\partial\mathcal{J}}{\partial b}(a_1, a_2, a_3, b) &= -2\sum_{i=1}^{N} (Y_i - a_1X_1^{(i)} + a_2X_2^{(i)} + a_3X_3^{(i)} + b) = 0,
\end{aligned}$$

ce qui conduit vers le système suivant

$$\left\{\begin{array}{rcl}
\displaystyle\sum_{i=1}^{N} X_1^{(i)}Y_i &=& a_1\left(\displaystyle\sum_{i=1}^{N}(X_1^{(i)})^2\right)+a_2\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)}\right)+a_3\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)}\right)+b\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}\right),\\
\displaystyle\sum_{i=1}^{N} X_2^{(i)}Y_i &=& a_1\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)}\right)+a_2\left(\displaystyle\sum_{i=1}^{N}(X_1^{(i)})^2\right)+a_3\left(\displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)}\right)+b\left(\displaystyle\sum_{i=1}^{N}X_2^{(i)}\right),\\
\displaystyle\sum_{i=1}^{N} X_3^{(i)}Y_i &=& a_1\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)}\right)+a_2\left(\displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)}\right)+a_3\left(\displaystyle\sum_{i=1}^{N}(X_3^{(i)})^2\right)+b\left(\displaystyle\sum_{i=1}^{N}X_3^{(i)}\right),\\
\displaystyle\sum_{i=1}^{N} Y_i &=& a_1\left(\displaystyle\sum_{i=1}^{N}X_1^{(i)}\right)+a_2\left(\displaystyle\sum_{i=1}^{N}X_2^{(i)}\right)+a_3\left(\displaystyle\sum_{i=1}^{N}X_3^{(i)}\right)+bN
\end{array}\right.$$

d'où le système matricielle suivant :

$$\begin{pmatrix}
\displaystyle\sum_{i=1}^{N}(X_1^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_1^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}(X_2^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}(X_3^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_3^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}X_3^{(i)} & N
\end{pmatrix}\begin{pmatrix} a_1\\ a_2\\ a_3\\ b\end{pmatrix}=\begin{pmatrix}
\displaystyle\sum_{i=1}^{N}X_1^{(i)}Y_i\\
\displaystyle\sum_{i=1}^{N}X_2^{(i)}Y_i\\
\displaystyle\sum_{i=1}^{N}X_3^{(i)}Y_i\\
\displaystyle\sum_{i=1}^{N}Y_i
\end{pmatrix}\tag{0.7}$$

Le pproblème d'approximation par régression linéaire multiple admet une unique solution sit et seulement si le système matricielle (0.7) admet une unique solution. C'est-à-dire que la matrice

$$A=\begin{pmatrix}
\displaystyle\sum_{i=1}^{N}(X_1^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_1^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}(X_2^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)}X_3^{(i)} & \displaystyle\sum_{i=1}^{N}(X_3^{(i)})^2 & \displaystyle\sum_{i=1}^{N}X_3^{(i)}\\
\displaystyle\sum_{i=1}^{N}X_1^{(i)} & \displaystyle\sum_{i=1}^{N}X_2^{(i)} & \displaystyle\sum_{i=1}^{N}X_3^{(i)} & N
\end{pmatrix}$$

est inversible.

Remarque 2.2.1 *Si la matrice A n'était pas inversible, alors on procède à la résolution par la méthode du pseudo-inverse de Moore-Penrose.Ainsi de trouver le meilleur couple $(a_1^\dagger, a_2^\dagger, a_3^\dagger, b^\dagger)$ de parmi tous les points critiques de la fonctionnelle $\mathcal{J}(a_1, a_2, a_3, b)$.*

Exemple 2.2.1 *Il est fournit dans le tableau 2.2, expliquant Y, la température moyenne annuelle de 6 villes du nord ouest du Maroc par leurs latitude X_1 et longitude X_2. L'exemple n'a d'autre utilité que calculatoire et montrer la façon dont on calcule les coefficients de régression et d'autres éléments lors de*

TABLE 2.1 – Variables explicatives des températures moyennes annuelles

Ind	X_1	X_2	Y	$\hat{Y}$	$\varepsilon = Y - \hat{Y}$
Tanger	48.55	7.6	9.6	10.197	-0.597
Tétouan	47.60	7.5	10.6	10.573	0.027
Al-Hoceima	48.00	7.8	11.3	10.572	0.728
Lârache	48.70	6.2	9.5	9.272	0.228
Chéfchaoun	47.63	6.8	9.5	10.131	-0.631
Asilah	47.78	6.3	10	9.756	0.244

ce type d'expériences.
L'équation de régression est

$$Y = a_1 X_1 + a_2 X_2 + b$$

avec Y est la température annuelle. Il s'agit d'un modèle à deux variables explicatives $p = 2$ et à six individus $N = 6$.

1. **Calcul de la matrice** A

$$A = \begin{pmatrix} \sum_{i=1}^{6} (X_1^{(i)})^2 & \sum_{i=1}^{6} X_1^{(i)} X_2^{(i)} & \sum_{i=1}^{6} X_1^{(i)} \\ \sum_{i=1}^{6} X_1^{(i)} X_2^{(i)} & \sum_{i=1}^{6} (X_2^{(i)})^2 & \sum_{i=1}^{6} X_2^{(i)} \\ \sum_{i=1}^{6} X_1^{(i)} & \sum_{i=1}^{6} X_2^{(i)} & 6 \end{pmatrix} = \begin{pmatrix} 13850.0978 & 2027.218 & 288.26 \\ 2027.218 & 299.22 & 42.20 \\ 288.26 & 42.20 & 6 \end{pmatrix}$$

2. **Calcul du vecteur du second membre** B

$$B = \begin{pmatrix} \sum_{i=1}^{6} X_1^{(i)} Y_i \\ \sum_{i=1}^{6} X_2^{(i)} Y_i \\ \sum_{i=1}^{6} Y_i \end{pmatrix} = \begin{pmatrix} 2905.975 \\ 427.09 \\ 60.50 \end{pmatrix}$$

3. **Résolution du système**

$$\begin{pmatrix} 13850.0978 & 2027.218 & 288.26 \\ 2027.218 & 299.22 & 42.20 \\ 288.26 & 42.20 & 6 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ b \end{pmatrix} = \begin{pmatrix} 2905.975 \\ 427.09 \\ 60.50 \end{pmatrix}$$

D'où

$$\begin{pmatrix} a_1 \\ a_2 \\ b \end{pmatrix} = \begin{pmatrix} -0.45962 \\ 0.61181 \\ 27.862 \end{pmatrix}$$

Finalement, l'équation de régression multiple

$$\hat{Y} = -0.45962 X_1 + 0.61181 X_2 + 27.862.$$

4. **Calcul de moyennes, variances et écart-type**

$$\begin{aligned}
\bar{X}_1 &= \frac{1}{6}(48.55 + 47.60 + 48.00 + 48.70 + 47.63 + 47.78) = 48.043,\\
\bar{X}_2 &= \frac{1}{6}(7.6 + 7.5 + 7.8 + 6.2 + 6.8 + 6.3) = 7.033,\\
\bar{Y} &= \frac{1}{6}(9.6 + 10.6 + 11.3 + 9.5 + 9.5 + 10) = 10.083,\\
\bar{\hat{Y}} &= \frac{1}{6}(10.197 + 10.573 + 10.572 + 9.272 + 10.131 + 9.756) = 10.0835,\\
\sigma^2(X_1) &= \frac{1}{6}(48.55^2 + 47.60^2 + 48.00^2 + 48.70^2 + 47.63^2 + 47.78^2) - 48.043^2 = 0.2198,\\
\sigma^2(X_2) &= \frac{1}{6}(7.6^2 + 7.5^2 + 7.8^2 + 6.2^2 + 6.8^2 + 6.3^2) - 7.033^2 = 0.407,\\
\sigma^2(Y) &= \frac{1}{6}(9.6^2 + 10.6^2 + 11.3^2 + 9.5^2 + 9.5^2 + 10^2) - 10.083^2 = 0.4514,\\
\sigma^2(\hat{Y}) &= \frac{1}{6}(10.197^2 + 10.573^2 + 10.572^2 + 9.272^2 + 10.131^2 + 9.756^2) - 10.0835^2 = 0.2099
\end{aligned}$$

5. **Coefficients de régression standardisés**

   (a) $\alpha_1 = a_1 * \frac{\sigma(X_1)}{\sigma(Y)} = -0.45962 * \sqrt{\frac{0.2198}{0.4514}} = -0.321$

   (b) $\alpha_2 = a_2 * \frac{\sigma(X_2)}{\sigma(Y)} = 0.61181 * \sqrt{\frac{0.407}{0.4514}} = 0.581$
   D'où l'équation de régression standardisée

$$\hat{Y} = -0.321Z_1 + 0.581Z_2$$

   avec $Z_1 = \frac{X_1 - 48.043}{0.469}$ *et* $Z_2 = \frac{X_2 - 7.033}{0.638}$.

   Un coefficient de régression standardisé exprime l'augmentation moyenne de Y quand une variable explicative augmente d'un écart-type et que les autres variables explicatives sont **maintenues constantes**.
   Ici, les coefficients de régression standardisés indiquent, pour les 6 villes considérées, l'influence sur leurs températures moyennes annuelles :

   (a) de la latitude à longitude constante,

   (b) de la longitude à latitude constante.

6. **Résidus de la régression** *: Les deux dernières colonnes du tableau 2.1 indiquent les températures prédites par le modèle de régression linéaire multiple ($\hat{Y}$) et les résidus de la régression (différence entre températures réelles Y et celles prédites par l'équation de régression $\hat{Y}$).*
   Par exemple pour Tanger, $\hat{Y} = 10.197$. Et le résidu est $9.6 - 10.197 = -0.597$ ce qui veut dire que le modèle **surestime** *donc la température de Tanger.*
   Il est clair que l'on a ici un exercice d'école et que l'étude thermique de la région Nord-Ouest du Maroc et de ses abords nécessiterait bien d'autres stations et variable (altitude, par exemple). Le but, ici, n'est que d'illustrer les principales aides à l'explication des résultats.
   On vérifie sur le tableau 2.1 que la moyenne des résidus est nulle (aux arrondis de calcul près). L'importance des écarts $Y - \hat{Y}$ est un premier indicateur de la qualité de l'ajustement par moindres carrés d'une régression multiple. Il faut donc regarder de près, cartographier et interpréter les résidus les plus forts (< 0 et > 0).
   Les résidus du tableau 2.1 (dernière colonne) semblent forts, notamment 3 d'entres eux :

   (a) La température moyennee annuelle est nettement surestimée par l'équation de régression à Chéfchaoun (altitude plus élevée : 1000 métres) et à Tanger.

(b) Elle est nettement sous estimée à Al-Hoceima.

7. **Corrélations de Bravais-Pearson entre variable (Corrélations partielles)** *: Il s'agit de calculer les coefficients*

$$\begin{aligned} \mathrm{r}_{YX_1} &= \frac{Cov(X_1,Y)}{\sigma(X_1)\sigma(Y)} = -0.281, \\ \mathrm{r}_{YX_2} &= \frac{Cov(X_2,Y)}{\sigma(X_2)\sigma(Y)} = 0.629 \\ \mathrm{r}_{X_2X_1} &= \frac{Cov(X_1,X_2)}{\sigma(X_1)\sigma(X_2)} = -0.056 \end{aligned}$$

Le tableau 2.2 fournit les coefficient de détermination entre les variables : Elles doivent être mini-

	Y	X_2
X_1	0.079	0.00314
X_2	0.396	1

TABLE 2.2 – r^2 entre variables du tableau 2.1

males entre les X_j, variables explicatives (indépendance) et bonnes entre variables explicatives X_j et variables à expliquer Y.
Les contraintes d'indépendance entre variables explicatives (latitude et longitude des 6 villes) est ici respectée puisque leur coefficient de détermination r^2 (variance commune) est de **0.00314***.*
Le coefficient de détermination r^2 (voir Tableau 2.2) entre :

(a) Température (Y) et latitude (X_1) est de 0.079 ($\mathrm{r}_{YX_1} = -0.281$) : les températures moyennes tendent légèrement à être plus chaudes au Nord, où les villes sont d'altitude plus basse).

(b) Température (Y) et longitude (X_2) est de 0.396 ($\mathrm{r}_{YX_2} = 0.629$) : les températures moyennes tendent légèrement à être plus chaudes à l'ouest, où les villes sont situé sur l'atlantique).

8. **Plan de régression** *: Une régression linéaire multiple avec comme variables indépendantes la latitude (X_1) et la longitude (X_2) nous donne ici un plan de régression. Connaissant la latitude et la longitude on peut extrapoler la variable Y à tout l'espace-domaine de notre étude- découpé en un maillage plus ou moins fin. On obtient alors une surface de tendance d'ordre 1 comme l'illustre le schéma suivant :*

2.3 Tests sur données d'échantillon

Si les données proviennent d'un échantillon **représentatif** dont on veut généraliser les résultats à toute la population mère (toute la zone dans l'exemple), on procédera à des tests de significativité des résultats de la régression (dont les éléments sont furnis par la plupart des logiciels statistiques).

2.3.1 Résidus comme erreur aléatoire du modèle de régression

Comme en régression simple, la distribution des résidus ($\mathcal{E}_i = Y_i - \hat{Y}_i$), exprimés dans l'unité de mesure de Y (en degrés Celsius dans l'exemple), doit alors donner lieu à examen :

1. la distribution des $\mathcal{E}_i$ doit être normale (Variable aléatoire de loi Gaussienne),
2. le nuage de points de E (en ordonnées)-$\hat{Y}$ (en abscisses) ne doit pas montrer de nettes croissance ou décroissance des valeurs de E en fonction de celles de $\hat{Y}$,

si la distribution de l'erreur aléatoire est gaussienne $\mathcal{N}(\mu, \sigma^2)$, on peut donc utiliser la distribution de probabilités de la loi de Gauss pour extrapoler les résultats.

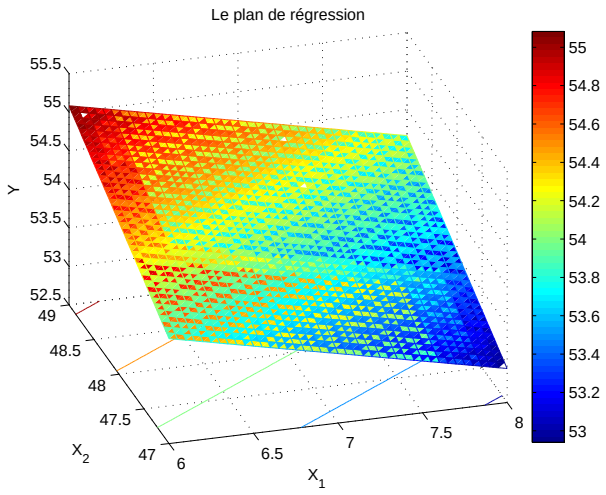

FIGURE 2.4 – Surface de tendance d'odre 1-Plan de régression.

2.3.2 Significativité de l'ensemble des variables explicatives

On effectue une analyse de variance et un test F de Fisher-Snedecor : on calculera les quantités suivantes SCE_Y, $SCE_{\text{résidu}}$, SCE_{regr}, CM_{regr}, $CM_{\text{résidu}}$ et puis $F_{\text{calculé}}$.
On lit dans la table du F de Fissher-Snedecor (pour un risque d'erreu choisi) la valeur de F correspondant à p et $n-p-1$ degrés de liberté. Si $F_{\text{calculé}} > F_{\text{lu}}$, on accepte (au resque d'erreur choisi) l'hypothèse que la régression est généralisable à la population mère (toute la zone Nord-Ouest du Maroc). Le tableau 2.3 fournit les valeurs pour cette analyse de variance.

TABLE 2.3 – Analyse de variance relative à la régression du tableau 2.1

SCE due à	SCE	Degrés de liberté	Carré Moyen
régression	1.2715	$p=2$	0.6357
Résidus	1.3968	$n-p-1=3$	0.4656
totale	2.6683	$n-1=5$	

$$F_{\text{calculé}} = \frac{CM_{\text{regr}}}{CM_{\text{résidu}}} = \frac{0.6357}{0.4656} = 1.3653.$$

Au risque d'erreur de 5%, F_{lu} dans la table pour 2 et 3 degrés de liberté vaut 19.16 (voir la table de Fisher-Snedecor fournie par un des logiciels statistiques).
$F_{\text{calculé}} < F_{\text{lu}}$: **on ne peut généraliser la régression à toute la zone**.
On vérifie par ailleurs que, sur l'échantillon de 6 villes, l'intensité de la relation est faible :

$$I = \frac{CM_{\text{regr}}}{CM_Y} = \frac{1.2715}{2.6683} = 0.4765.$$

Latitude et longitude n'expliquent, dans l'échantillon, que $I \times 100 = 47.65\%$ des variations inter-cités de températures annuelles moyennes.

2.4 Corrélation multiple

Le coefficient de corrélation multiple est le coefficient de Bravais Pearson entre Y et $\hat{Y}$, c'est-à-dire entre valeurs observées et prédites par le modèle de régression. Comme en régression simple, c'est le carré

du coefficient de corrélation (R^2 : coefficient de détermination) qui exprime le **pourcentage de variance pris en compte par le modèle** et qui mesure donc la quaalité de l'ajustement linéaire.
Si les variables explicatives X_j sont parfaitement indépendantes les unes des autres (aucune redondance entre elles), R^2 multiple est la somme des r^2 entre chaque X_j et Y :

$$\mathrm{R}^2 = \mathrm{r}^2_{YX_1} + \mathrm{r}^2_{YX_2}.$$

Dans l'exemple du tableau 2.1, le coefficient de corrélation multiple (corrélation simple $\mathrm{r}_{Y\hat{Y}}$ entre les variables Y et $\hat{Y}$) est

$$\mathrm{r}_{Y\hat{Y}} = \frac{Cov(Y,\hat{Y})}{\sigma(Y)\sigma(\hat{Y})} = 0.6962$$

et le coefficient de détermination est de 0.4765. R^2 mesure la variance expliquée par la régression

$$\mathrm{R}^2 = I = \frac{CM_{\text{regr}}}{CM_Y} = \frac{1.2715}{2.6683} = 0.4765 \simeq 0.4746 = \mathrm{r}^2_{YX_1} + \mathrm{r}^2_{YX_2}.$$

Comme l'analyse de variance l'avait déjà révélé, l'équation de régression multiple n'explique que 47.65% des différences de température moyenne annuelle entre les 6 ville de l'échantillon tandis que 52.35% de celle-ci reste inexpliquée (et due à d'autres facteurs joueant sur la variation de la température annuelle). L'analyse de variance et R^2 fournissent donc la même information (variance de Y explicable par l'ensemble des X_j).

Chapitre 3

L'analyse en composantes principales

3.1 Introduction

La 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 variabilité totale.
Pratiquement, cela revient à projeter les données 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).

3.2 Etape 1 : Changement de repère

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

- **Représentation graphique du $i^{\text{ème}}$ individu**
  On peut représenter 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 x^T sont les données des p variables pour $i^{\text{ème}}$ individu.
- **Nouveau repère**
  Soient $v_1, \ldots, v_p$, p vecteurs de $\mathbb{R}^p$, unitaires et deux à deux orthogonaux. On considère les p droitess passant par l'origine, de vecteurs directeurs $v_1, \ldots, v_p$ respectivement. Alors ces droites définissent un nouveau repère. Chacun des axes représente une nouvelle variables, 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 $j = 1, \ldots, p$, la coordonnée sur l'axe v_j est la coordonnée de la projection orthogonale de x sur la droite passant par l'origine et de vecteur directeur v_j. Elle est donnée par (voir le chapitre 1) :

  $$(x, v_j) = x^T v_j.$$

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

  $$(x^T v_1, \ldots, x^T v_p) = x^T Q = (A - \bar{A})_i^T Q$$

  où Q est la matrice de taille $(p \times p)$, dont les colonnes sont les vecteurs $v_1, \ldots, v_p$. Cette matrice est **orthonormale**, c'est-à-dire 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 $(A-\bar{A})_1^T, \ldots, (A-\bar{A})_n^T$. Les coordonnées dans le nouveau repère sont répertoriées dans la matrice :

  $$B = (A - \bar{A})Q$$

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

3.3 Etape 2 : Choix du nouveau repère

Le but est de trouver un nouveau repère $v_1, \ldots, v_p$, tel que la quantité d'information expliquée par v_1 soit maximale, puis celle expliquée par v_2, etc... On peut ainsi se limiter à ne garder que les 2 où 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.

3.3.1 Mesure de la quantité d'information

La variance dess données centrées $(A - \bar{A})_{(j)}$ 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 $(A - \bar{A})$ est donc des variances des données de toutes les variables, c'est-à-dire la **variabilité totale** des données $(A - \bar{A})$, définie précédement

$$\sum_{j=1}^{p} \sigma^2((A - \bar{A})_{(j)}) = \text{Tr}(C(A - \bar{A})) = \text{Tr}(C(A)).$$

La dernière égalité vient du fait que $C(A-\bar{A}) = C(A)$ (les matrices de covariances soient égales). Etudions maintenant la variabilité totale des données B, qui sont la projection des données $C(A-\bar{A}))$ dans le nouveau repère défini par la matrice orthonormale Q. Soit $C(B)$ la matrice de covariance correspondante, alors :

Propriété 3.3.1 *1. $C(B) = Q^T C(A) Q$,*

2. La variabilité totale des données B est la même que celle des données $(A - \bar{A})$.

Démonstration.

1. On a

$$\begin{aligned} C(B) &= \frac{1}{n}(B - \bar{B})^T(B - \bar{B}) \\ &= \frac{1}{n}B^T B \quad (\text{car } \bar{B} \text{ est la matrice nulle}) \\ &= \frac{1}{n}((A - \bar{A})Q)^T(A - \bar{A})Q \\ &= \frac{1}{n}Q^T(A - \bar{A})^T(A - \bar{A})Q \\ &= Q^T C(A) Q \end{aligned}$$

2. Ainsi, la variabilité totale des nouvelles données B est

$$\begin{aligned} \text{Tr}(C(B)) &= \text{Tr}(Q^T C(A) Q) = \text{Tr}(Q^T Q C(A)), \quad (\text{propriéte de la trace}) \\ &= \text{Tr}(C(A)) \end{aligned}$$

car $Q^T Q = Id$, étant donné que la matrice Q est orthonormale.

□

3.3.2 Choix du nouveaur repère

Etant donné que la variabilité totale des données projetées dans le nouveau repère est la même que celle des données d'origine $(A - \bar{A})$, on souhaite déterminer Q de sorte que la part de la variabilité totale expliquée par les donnée $B_{(1)}$ de la nouvelle variable v_1 soit maximale, puis celle expliquée par les données $B_{(2)}$ de la nouvelle variable v_2, etc...
Autrement dit, on souhaite résoudre le problème d'optimisation suivant :
"**Trouver une matrice orthonormale** Q **telle que** $\sigma^2(B_{(1)})$ **soit maximale, puis** $\sigma^2(B_{(2)})$**, etc...**"
Avant d'énoncer le théorème donnant la matrice Q optimale, nous avons besoin de nouvelles notions d'algère linéaire.

- **Théorème spectral pour les matrices symétriques**
  Soit A une matrice de taille $(p \times p)$. Un vecteur x de $\mathbb{R}^p$ s'appelle un **vecteur propre** de la matrice A, s'il existe un nombre λ tel que :
  $$Ax = \lambda x.$$
  Le nombre λ s'appelle la valeur propre associée au vecteur propre x.
  Une matrice carrée $A = (a_{ij})$ est dite symétrique si et seulement si $a_{ij} = a_{ji}$, pour tout i, j.
  Théorème 3.3.1 *Si A est une matrice symétrique de taille $(p \times p)$, alors il existe une base orthonormale de $\mathbb{R}^p$ formée de vecteur propres de A. De plus, chacune des valeurs propres associée est réelle. Autrement dit, il existe une matrice orthonormale Q telle que*
  $$Q^T AQ = D$$
  avec D est la matrice diagonale formée des valeurs propres de A
- **Théorème fondamentale de l'ACP**
  Soit $(A - \bar{A})$ la matrice des données centrées, et soit $C(A)$ la matrice de covariance associée (qui est symétrique par définition). On note $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ les valeurs propres de la matrice $C(A)$. Soit Q la matrice orthonormale correspondant à la matrice $C(A)$, donnée par le Théorème(3.3.1), telle que le premier vecteur corresponde à la plus grande valeur propres, etc... Alors, le théorème fondamentale de l'ACP est :
  Théorème 3.3.2 *La matrice orthonormale qui résout le problème d'optimisation est la matrice Q décrite ci-dessus. De plus, on a :*
  1. $\sigma^2(B_{(j)}) = \lambda_j$,
  2. $\mathrm{Cov}(B_{(i)}, B_{(j)}) = 0$, *quand* $i \neq j$,
  3. $\sigma^2(B_{(1)}) \geq \sigma^2(B_{(2)}) \geq \ldots \geq \sigma^2(B_{(p)})$.

  Les colonnes $v_1, \ldots, v_p$ de la matrice Q décrivent les nouvelles variables, appelées les **composantes principales**
  Démonstration. On a
  $$\begin{aligned} C(B) &= Q^T C(A) Q \quad (\text{d'après la propriété}(3.3.1)) \\ &= \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix} \end{aligned}$$
  Ainsi,
  $$\sigma^2(B_{(j)}) = (C(B))_{jj} = (Q^T C(A) Q)_{jj} = \lambda_j$$
  $$\mathrm{Cov}(B_{(i)}, B_{(j)}) = (C(B))_{ij} = (Q^T C(A) Q)_{ij} = 0$$

ceci démontre les deux premières assertions du théorème. Le troisième point 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 λ_1, etc... Même si ce n'est pas très difficile, nous choisissons de ne pas traiter cette partie ici. □

3.4 Conséquences de l'ACP

Voici deux conséquences importantes du résultats que nous avons établi dans 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 $v_1, \ldots, v_p$, les **composantes principales**, qui sont optimales. La part de la variabilité totale expliquée par les données $B_{(1)}, \ldots, B_{(k)}$ des k premières nouvelles variables ($k \leq p$), est :

$$\frac{\sigma^2(B_{(1)}) + \ldots + \sigma^2(B_{(k)})}{\sigma^2(B_{(1)}) + \ldots + \sigma^2(B_{(p)})} = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p}$$

  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 purcentage est raisonnable, on choisira de se restreindre aux 2 ou 3 premiers axes. La notion de raisonnable est discutable. Lors des travaux pratiques, 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 $(A-\bar{A})_{(j)}$ de la $j^{\text{ème}}$ ancienne variable et celle $B_{(k)}$ de la la $k^{\text{ème}}$ nouvelle variable. La matrice de covariance $C(A, B)$ de $A - \bar{A}$ et B est donnée par :

$$\begin{aligned} C(A,B) &= \frac{1}{n}(A-\bar{A})^T(B-\bar{B}) \\ &= \frac{1}{n}(A-\bar{A})^T B \quad (\text{car } \bar{B} \text{ est la matrice nulle}) \\ &= \frac{1}{n}(A-\bar{A})^T(A-\bar{A})Q, \quad (\text{par définition de la matrice } B \\ &= Q(Q^T C(A) Q), \quad (\text{car } QQ^T = Id), \\ &= QD, \end{aligned}$$

  car par le théorème spectral, D est la matrice diagonale des valeurs propres.
  Ainsi :

$$\text{Cov}(A_{(j)}, B_{(k)}) = (C(A,B))_{jk} = q_{jk}\lambda_k.$$

  De plus, $\sigma^2(A_{(j)}) = (C(A))_{jj} = \mu_{jj}$ et $\sigma^2(B_{(k)}) = \lambda_k$. Ainsi la correlation entre $A_{(j)}$ et $B_{(k)}$ est donnée par :

$$\text{r}(A_{(j)}, B_{(k)}) = \frac{\lambda_k q_{jk}}{\sqrt{\lambda_k \mu_{jj}}} = \sqrt{\frac{\lambda_k}{\mu_{jj}}}\, q_{jk}.$$

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

Remarque 3.4.1 *Le raisonnement ci-dessus n'est pas valable que si la dépendance entre les données des variables est linéaire. En effet, dire qu'une corrélation forte (resp. faible) est équivalente à une dépendance forte (resp. 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 (*$\log$, $\exp$*, racines,....).*

3.5 Dans la 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 $R(A)$. En effet :

$$\begin{aligned}
\text{Cov}(Z_{(i)}, Z_{(j)}) &= \text{Cov}\left(\frac{A_{(i)} - \overline{A_{(i)}}}{\sigma_{(i)}}, \frac{A_{(j)} - \overline{A_{(j)}}}{\sigma_{(j)}}\right) \\
&= \frac{1}{\sigma_{(i)}\sigma_{(j)}}\text{Cov}\left(A_{(i)} - \overline{A_{(i)}}, A_{(j)} - \overline{A_{(j)}}\right) \\
&= \frac{1}{\sigma_{(i)}\sigma_{(j)}}\text{Cov}\left(A_{(i)}, A_{(j)}\right) \\
&= \text{r}(A_{(i)}, A_{(j)}).
\end{aligned}$$

3. La matrice Q est la matrice orthogonale correspondant à la matrice $R(A)$, donnée par le Théorème spectral pour les matrices symétriques.
4. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ sont les valeurs propres de la matrice de corrélation $R(A)$.
5. La corrélation entre $Z_{(j)}$ et $Z_{(k)}$ est :

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

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

3.6 Exemple d'application

Soit le tableau de données suivant :

TABLE 3.1 – Le tableau est représenté sous la forme $A_{(3,2)}$

	ind var	x_1	x_2
	A_1	4	5
	A_2	6	7
	A_3	8	0

- **Représentation graphique** du nuage des 3 points individus dans l'espace $\mathbb{R}^2$ des variables (x_1 en abscisse et x_2 en ordonnée). Le système d'axes est orthonormé : une base $\{\vec{i}, \vec{j}\}$ telle que $\|\vec{i}\| = \|\vec{j}\| = 1$ et $(\vec{i}, \vec{j}) = 0$.

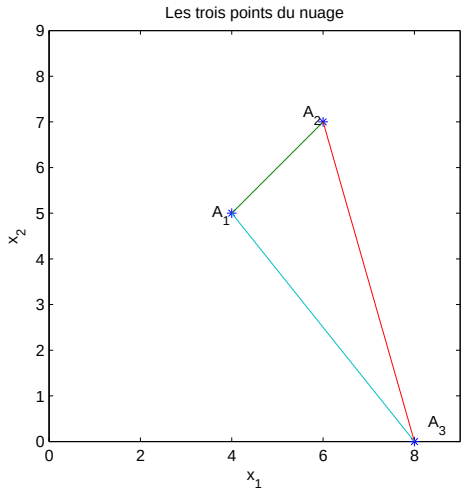

FIGURE 3.1 – Les trois points du nuage constituent l'information des lignes du Tableau (3.1). Les positions relatives de ces 3 points peuvent être calculées en utilisant la distance euclidienne.

- **Calcul des caractéristiques des colonnes du tableau**
  Calcul de la moyenne et de l'écart-type de x_1 et x_2 :

$$\begin{aligned} \bar{x}_1 &= \frac{18}{3} = 6 \quad \text{et} \quad x_2 = \frac{12}{3} = 4 \\ \sigma^2(x_1) &= \frac{116}{3} - 6^2 = 2.67 \quad \text{et} \quad \sigma(x_1) = 1.633 \\ \sigma^2(x_2) &= \frac{74}{3} - 4^2 = 8.67 \quad \text{et} \quad \sigma(x_2) = 2.944 \end{aligned}$$

  Calcul de la moyenne et de l'écart-type de A_1, A_2 et A_3 :

$$\begin{aligned} \bar{A_1} &= \frac{9}{2} = 4.5, \quad \bar{A_2} = \frac{13}{2} = 6.5 \quad \text{et} \quad \bar{A_3} = \frac{8}{2} = 4 \\ \sigma^2(A_1) &= \frac{41}{2} - 4.5^2 = 0.25 \quad \text{et} \quad \sigma(A_1) = 0.5 \\ \sigma^2(A_2) &= \frac{85}{2} - 6.5^2 = 0.25 \quad \text{et} \quad \sigma(A_2) = 0.5 \\ \sigma^2(A_3) &= \frac{64}{2} - 4^2 = 16 \quad \text{et} \quad \sigma(A_3) = 4 \end{aligned}$$

- **Construction du tableau des variables centrées et réduites**

TABLE 3.2 – Le tableau est représenté sous la forme $Z_{(3,2)}$

	ind var	x_1	x_2	$x_1 - \bar{x}_1$	$x_2 - \bar{x}_2$	$z_1 = \frac{x_1 - \bar{x}_1}{\sigma(x_1)}$	$z_2 = \frac{x_2 - \bar{x}_2}{\sigma(x_2)}$
	Z_1	4	5	-2	1	$-1.225 = -\sqrt{\frac{3}{2}}$	$0.34 = \sqrt{\frac{3}{26}}$
	Z_2	6	7	0	3	0	$1.02 = \frac{3\sqrt{13}}{13}$
	Z_3	8	0	2	-4	$1.225 = \sqrt{\frac{3}{2}}$	$-1.36 = -4\sqrt{\frac{3}{26}}$
	Somme	18	12	0	0	0	0

On vérifie que : $\bar{z}_1 = \bar{z}_2 = 0$, $\sigma^2(z_1) = \sigma^2(z_2) = 1$ et $\text{Cov}(z_1, z_2) = \text{r}_{z_1,z_2}$.

- **Représentation graphique** du nuage des 3 points individus dans l'espace $\mathbb{R}^2$ des variables centrées réduites (z_1 en abscisse et z_2 en ordonnée). Le système des axes est orthonormé : une base $\{\vec{i}, \vec{j}\}$ telle que $\|\vec{i}\| = \|\vec{j}\| = 1$ et $(\vec{i}, \vec{j}) = 0$. Dans cet espace, l'origine des axes (point 0) est confondu avec le centre de gravité du triangle (point $G(\bar{z}_1 = 0, \bar{z}_2 = 0)$)

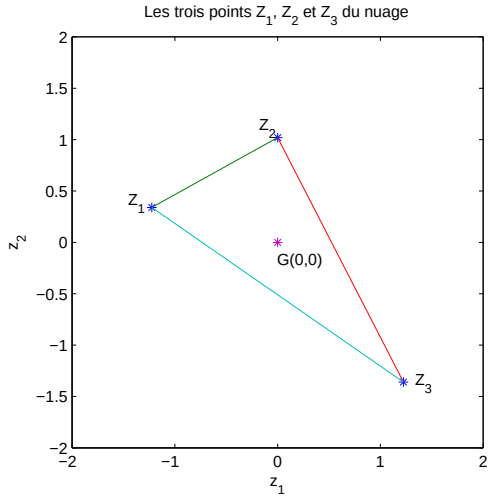

FIGURE 3.2 – Les trois points du nuage constituent l'information des lignes du Tableau (3.2). Les positions relatives de ces 3 points peuvent être calculées en utilisant la distance euclidienne.

Dans l'espace $\mathbb{R}^3$ des individus, se situent les deux variables centrées réduites. On a

$$z_1\left(-\sqrt{\frac{3}{2}}, 0, \sqrt{\frac{3}{2}}\right) \quad \text{et} \quad z_2\left(\sqrt{\frac{3}{26}}, \frac{3\sqrt{13}}{13}, -4\sqrt{\frac{3}{26}}\right)$$

Avec un système d'axes orthonormé, en utilisant les variables centrées réduites dans l'espace à trois dimensions des individus avec un système orthonormé on peut calculer :

$$d^2(0, z_1) = \left(-\sqrt{\frac{3}{2}}\right)^2 + 0^2 + \left(\sqrt{\frac{3}{2}}\right)^2 = 3$$

D'où $\frac{1}{3}d^2(0, z_1) = 1$ la variance de z_1.

$$d^2(0, z_2) = \left(\sqrt{\frac{3}{26}}\right)^2 + \left(\frac{3\sqrt{13}}{13}\right)^2 + \left(-4\sqrt{\frac{3}{26}}\right)^2 = 3$$

D'où $\frac{1}{3}d^2(0, z_2) = 1$ la variance de z_2.
Dans cet espace, la distance au carré entre l'origine et une variable est, à $n = 3$ près, la variance de la variable. Quand les variables sont centrées et réduites, toutes les variables sont **équidistantes** de l'origine. cette distance est, au nombre d'observations près, la variance des variables.

Présentation des calculs

$$X_{(3,2)} = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{bmatrix} \end{array}$$

$$\bar{x}_j = \begin{bmatrix} 6 & 4 \end{bmatrix} \qquad Z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)}$$

$$\sigma(x_j) = \begin{bmatrix} 2\sqrt{\frac{2}{3}} & \sqrt{\frac{26}{3}} \end{bmatrix}$$

Tableau des variables centrées réduites :

$$Z_{(3,2)} = \begin{array}{c} \\ Z_1 \\ Z_2 \\ Z_3 \end{array} \begin{array}{c} \begin{array}{cc} z_1 & z_2 \end{array} \\ \begin{bmatrix} -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{26}} \\ 0 & 3\sqrt{\frac{3}{26}} \\ \sqrt{\frac{3}{2}} & -4\sqrt{\frac{3}{26}} \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{cc} z_1 & z_2 \end{array} \\ \begin{bmatrix} -1.225 & 0.34 \\ 0 & 1.02 \\ 1.225 & -1.36 \end{bmatrix} \end{array}$$

$\bar{Z}_j = \begin{bmatrix} 0 & 0 \end{bmatrix}$ La moyenne des variables centrées et réduites est égale à 0

$\sigma(z_j) = \begin{bmatrix} 1 & 1 \end{bmatrix}$ L'écart-type des variables centrées et réduites est égale à 1

De plus,

$$\mathrm{r}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)} = \mathrm{Cov}(x,y)$$

puisque $\sigma(x) = \sigma(y) = 1$. Donc, le coefficient de corrélation linéaire r entre deux variables est égal à la covariance.

Remarque 3.6.1 *On peut aussi traiter l'information contenue dans le tableau de départ en utilisant me tableau des individus centrés réduits.*

$$X_{(3,2)} = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \bar{x}_i & \sigma(x_i) \end{array} \\ \begin{bmatrix} 4 & 5 \end{bmatrix} \quad \begin{bmatrix} 4.5 \\ 6.5 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 0.5 \\ 0.5 \\ 4 \end{bmatrix} \\ \begin{bmatrix} 6 & 7 \end{bmatrix} \\ \begin{bmatrix} 8 & 0 \end{bmatrix} \end{array}$$

$$Q_{(3,2)} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix} \quad \textit{avec} \quad q_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)}$$

Il est possible de représenter l'information contenue dans ce nouveau tableau comme précédemment et d'en tirer des conclusions.

- **Calcul du produit matriciel** $\frac{1}{n} Z^T Z$

$$\frac{1}{3} Z^T Z = \frac{1}{3} \begin{bmatrix} -\sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{26}} & 3\sqrt{\frac{3}{26}} & -4\sqrt{\frac{3}{26}} \end{bmatrix} \begin{bmatrix} -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{26}} \\ 0 & 3\sqrt{\frac{3}{26}} \\ \sqrt{\frac{3}{2}} & -4\sqrt{\frac{3}{26}} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 3 & -\frac{15}{2\sqrt{3}} \\ -\frac{15}{2\sqrt{3}} & 3 \end{bmatrix}$$

$$\frac{1}{3}Z^TZ = \begin{bmatrix} 1 & -0.69 \\ -0.69 & 1 \end{bmatrix}$$

Le résultat de ce calcul est une matrice carrée, de taille (2×2), notée R, contenant les coefficients de corrélation linéaires des variables.
Cette matrice carrée R a pour dimension le nombre de variables. Elles est possède les propriétés suivantes :

(a) Elle est symétrique.

(b) Elle a des 1 sur la diagonale principale (les variances des variables)

(c) Elle a des valeurs inférieures ou égales à 1 en valeur absolue.

Dans cette matrice R, on a sur la diagonale les variances des variables, or dans une exercice du chapitre précédent on a vu que cette variance était, au nombre d'observations près, la distance de la variable à l'origine. Elle contient de part et d'autre de la diagonale le coefficient de corrélation linéaire entre les deux variables. Or dans un exercice (chapitre précédent), on a vu que ce coefficient de corrélation était le cosinus de l'angle formé par les deux variables. L'angle formé par les deux variables peut donc en être déduit.
Avec la matrice R, il est donc possible de représenter dans l'espace les positions relatives des variables entre elles. Cette matrice R nous donne donc l'information recherchée concernant les variables. C'est la raison pour laquelle elle porte le nom de matrice d'information des variables.

- **Calcul du produit matriciel** ZZ^T

$$ZZ^T = \begin{bmatrix} -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{26}} \\ 0 & 3\sqrt{\frac{3}{26}} \\ \sqrt{\frac{3}{2}} & -4\sqrt{\frac{3}{26}} \end{bmatrix} \begin{bmatrix} -\sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{26}} & 3\sqrt{\frac{3}{26}} & -4\sqrt{\frac{3}{26}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{21}{13} & \frac{9}{26} & -\frac{51}{26} \\ \frac{9}{26} & \frac{27}{26} & -\frac{36}{26} \\ -\frac{51}{26} & -\frac{36}{26} & \frac{87}{26} \end{bmatrix} = V.$$

Cette matrice V n'est pas une matrice de corrélation, mais elle y ressemble. On lui donne le nom de **matrice d'information des individus**. Elle est symétrique ; sa diagonale est la somme des carrés des individus ligne du tableau et de part et d'autre on trouve la somme des produits lignes deux à deux des individus.

- **Caractéristiques de la matrice** $\mathrm{R} = \frac{1}{n}Z^TZ$
Les caractéristiques d'une matrice sont données par les vecteurs propres associés aux valeurs propres de la matrice.
On appelle vecteur propre associé à la valeur propre λ de la matrice R toute solution du système homogène

$$\mathrm{R}V = \lambda V \Leftrightarrow V \in \ker(\mathrm{R} - \lambda I).$$

On sait que si dans ce système d'équaations le déterminant de la matrice $(\mathrm{R} - \lambda I)$ est différent de 0, alors ce système possède une et une seule solution qui est $V = 0$ et que l'on appelle la solution triviale. C'est la raison pour laquelle pour que ce système ait des solutions autres que celle-ci, il faut que

$$\det(\mathrm{R} - \lambda I) = 0.$$

Or ce déterminant conduit à une équation (équation caractéristique de la matrice R) qui a pour variable λ et pour degré la dimension de la matrice R.

Les racines de cette équation donnent les différentes valeurs de λ et portent le nom de valeurs propres de la matrice R. Pour chacune des valeurs propres, on pourra calculer à partir du système de départ, une infinité de vecteurs V qu'on appelle les vecteurs propres de R. Parmi cette infinité de vecteurs propres, on recherche par la suite le vecteur propre de norme 1 (c'est-à-dire le vecteur propre unitaire). Dans ce cas on a :

$$\mathrm{R}V = \lambda V \Leftrightarrow V \in \ker(\mathrm{R} - \lambda I) \Leftrightarrow (\mathrm{R} - \lambda I)V = 0$$

On note $V = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ le vecteur propre de R associé à la valeur propre λ.

$$(\mathrm{R} - \lambda I)V = 0 \Leftrightarrow \begin{pmatrix} 1-\lambda & -0.69 \\ -0.69 & 1-\lambda \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Leftrightarrow \begin{cases} (1-\lambda)v_1 - 0.69v_2 = 0 \\ -0.69v_1 + (1-\lambda)v_2 = 0 \end{cases}$$

***Calcul de valeurs propres :**

$$\det(\mathrm{R} - \lambda I) = \begin{vmatrix} 1-\lambda & -0.69 \\ -0.69 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - (0.69)^2$$

$$= (1 - \lambda - 0.69)(1 - \lambda + 0.69) = 0.$$

d'où la matrice R admet deux valeurs propres distinctes $\lambda_1 = 1.69$ et $\lambda_2 = 0.31$. Si on additionne $\lambda_1 + \lambda_2 = 1.69 + 0.31 = 2$, on obtient la dimension de la matrice R (le nombre de variables du tableau).

***Calcul de vecteurs propres associés :**

– Pour $\lambda_1 = 1.69$

$$\begin{cases} (1-1.69)v_1 - 0.69v_2 = 0 \\ -0.69v_1 + (1-1.69)v_2 = 0 \end{cases} \Leftrightarrow \begin{cases} -0.69v_1 - 0.69v_2 = 0 \\ -0.69v_1 - 0.69v_2 = 0 \end{cases} \Leftrightarrow v_1 + v_2 = 0$$

d'où $V = \begin{pmatrix} k \\ -k \end{pmatrix}$ avec $k \in \mathbb{R}$. On a une infinité de vecteurs propres portés par la seconde bissectrice du plan $\{\vec{v}_1, \vec{v}_2\}$.

Pour touver un vecteur propre normé il faut que

$$\|V\|^2 = 1 \Leftrightarrow k^2 + k^2 = 2k^2 = 1 \Leftrightarrow k = \pm\frac{\sqrt{2}}{2}.$$

En retenant pour k la valeur positive, on définit : $b_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$ le vecteur propre normé de R.

– Pour $\lambda_2 = 0.31$

$$\begin{cases} (1-0.31)v_1 - 0.69v_2 = 0 \\ -0.69v_1 + (1-0.31)v_2 = 0 \end{cases} \Leftrightarrow \begin{cases} 0.69v_1 - 0.69v_2 = 0 \\ -0.69v_1 + 0.69v_2 = 0 \end{cases} \Leftrightarrow v_1 - v_2 = 0$$

d'où $W = \begin{pmatrix} k \\ k \end{pmatrix}$ avec $k \in \mathbb{R}$. On a une infinité de vecteurs propres portés par la première bissectrice du plan $\{\vec{v}_1, \vec{v}_2\}$.

Pour touver un vecteur propre normé il faut que

$$\|V\|^2 = 1 \Leftrightarrow k^2 + k^2 = 2k^2 = 1 \Leftrightarrow k = \pm\frac{\sqrt{2}}{2}.$$

En retenant pour k la valeur positive, on définit : $b_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$ le vecteur propre normé de R.

Ces vecteurs propres normés constiruent une nouvelle base orthonormée dans laquelle la norme de chaque vecteur égale à 1 et leur produit sclaire est nul :

$$(b_1, b_2) = \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2} = 0$$

on peut alors placer les coordonnées (dans l'ancienne base) de ces vecteurs dans une matrice Q, dans l'ordre décroissant de leurs valeurs propres.

$$Q = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

cette matrice est une matrice orthogonale et vérifie donc : $Q^{-1} = Q$ et $Q^T Q = I$.

- **Caractéristiques de la matrice** $V = ZZ^T$
  Si on calcule comme précédemment les valeurs propres de la matrice V :

$$\det(V - \lambda I) = \begin{vmatrix} \frac{21}{13} - \lambda & \frac{9}{26} & -\frac{51}{26} \\ \frac{9}{26} & \frac{27}{26} - \lambda & -\frac{36}{26} \\ -\frac{51}{26} & -\frac{36}{26} & \frac{87}{26} - \lambda \end{vmatrix} = 0$$

On trouve $\lambda_1 = 5.07$, $\lambda_2 = 0.93$ et $\lambda_3 = 0$.
Si on porte dans un tableau les valeurs propres de V et R on a : On remarque que si on multiplie les

TABLE 3.3 – Le tableau présente un bilan entre V et R

V	R
$\lambda_1 = 5.07$	$\lambda_1 = 1.69$
$\lambda_2 = 0.93$	$\lambda_2 = 0.31$
$\lambda_3 = 0$	$----$
$\sum_{i=1}^{3} \lambda_i = 6$	$\sum_{i=1}^{2} \lambda_i = 2 = m$

valeurs propres de la matrice R par $n = 3$, on obtient les deux premières valeurs propres de la matrice V et que la dernière valeur propre de V est nulle. C'est-à-dire que

$$\lambda_i(V) = n\lambda_i(\mathrm{R}), \quad \text{pour } i \in \{1; 2\}.$$

Propriété 3.6.1 *Soit $n \geq m$ deux entriers naturels et soient Z une matrice de taille $(n \times m)$, $V = ZZ^T$ et $\mathrm{R} = \frac{1}{n}Z^T Z$ deux matrices qui généralisent le cas traité précédemment. On désigne par $\lambda_i(V)$ les valeurs propres de V et par $\lambda_i(\mathrm{R})$ les valeurs propres de R. Alors on a les propriérés suivantes :*

(a) $m = \sum_{i=1}^{m} \lambda_i(\mathrm{R})$

(b) $\lambda_i(V) = n\lambda_i(\mathrm{R})$ *pour tout* $1 \leq i \leq m$

(c) La matrice V admet $(n - p)$ valeurs propres nulles.

Remarque 3.6.2 *On peut aussi démontrer qu'il est possible de calculer les vecteurs propres de V connaissant ceux de R. Et donc, qu'en définitive, les caractéristiques de R permettent de calculer celles de V et réciproquement.*

Chapitre 4

Méthodes de classification

4.1 Introduction

Ce chapitre est dédié aux 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 la classe 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-hiérarchique(partitionnement).** Décomposition de l'espace des individus en classes disjointes.
2. **Classifications hiérarchique.** 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. Plus, 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 4.1.1 *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.*

4.2 Moyenne et barycentre, variance et inertie

4.2.1 Cas d'une variable

Une variable x définie sur un ensemble K d'individus se représente par un nuage de points sur un axe, c'est à dire que $x : K \to \mathbb{R}$ est une variable aléatoire réelle. L'individu i est représenté par le point d'abscisse $x(i) = x_i$ la valeur prise par la variable x pour l'individu i.

(a) **Moyenne et barycentre** : si l'importance des individus est la même pour tous, la **moyenne** de la variable x, notée $\bar{x}$, est égale à :

$$\bar{x} = \frac{1}{n} \sum_{i \in K} x_i$$

plus généralement, si l'individu a un poids p_i(par exemple si les individus représentent des populations d'effectifs inégaux), la moyenne $\bar{x}$ s'écrit :

$$\bar{x} = \frac{\sum_{i \in K} p_i x_i}{\sum_{i \in K} p_i}.$$

Souvent les poids sont tels que $\sum_{i\in K} p_i = 1$ ce qui allège l'écriture : $\bar{x} = \sum_{i\in K} p_i x_i$.

Définition 4.2.1 *Sur l'axe de représentation du nuage, le points d'abscisse $\bar{x}$ est le barycentre des points x_i muni des poids p_i. Ce barycentre est la traduction géométrique de la notion statistique de moyenne.*
En retirant à chaque x_i la moyenne $\bar{x}$, on obtient une variable centrée. En passant de x à $x - \bar{x}$ on effectue une translation du nuage sur l'axe (ou une translation de l'origine de l'axe) qui fait coïncider son barycentre avec l'origine.

(b) **Variance et inertie** : Si l'importance des individus est la même pour tous, la **variance** d'une variable x, notée $\sigma^2(x)$, est égale à :

$$\sigma^2(x) = \frac{1}{n}\sum_{i\in K}(x_i - \bar{x})^2.$$

Si l'individu i a un poids p_i elle s'écrit :

$$\sigma^2(x) = \frac{\sum_{i\in K} p_i(x_i - \bar{x})^2}{\sum_{i\in K} p_i}.$$

Lorsque les poids sont tels que $\sum_{i\in K} p_i = 1$ ce qui allège l'écriture :

$$\sigma^2(x) = \sum_{i\in K} p_i(x_i - \bar{x})^2$$

La variance mesure la dispersion des valeurs autour de la moyenne. Le fait de considérer les carrés des écarts et non les valeurs absolues des écarts facilite les calculs et permet des décompositions suivant le théorème de Pythagore et celui de Huygens. L'écart-type $\sigma(x)$ est la racine carrée de la variance.

Définition 4.2.2 *(a) L'inertie I_a d'un point i de poids p_i par rapport à un point A de coordonnée x_a est, par définition le produit du poids de i par le carré de sa distance à A soit :*

$$I_i = p_i(x_i - x_a)^2.$$

(b) L'inertie d'un nuage de pints est la somme des inerties des points du nuage. L'inertie I d'un nuage de points représenté sur un axe, par rapport au point G d'abscisse $\bar{x}$, est égale à

$$\sum_{i\in K} I_i = \sum_{i\in K} p_i(x_i - x_a)^2;$$

on retrouve la variance lorsque $\sum_{i\in K} p_i = 1$.

(c) La notion statistique de variance correspond à la notion mécanique d'inertie d'un nuage de points par rapport à son barycentre.

lorsqu'on divise chaque valeur $x_i - \bar{x}$ de la variable centrée par son écart-type $\sigma(x)$, on obtient une variable de moyenne 0 et de variance 1, appelée **variable centrée-réduite**.
La transformation géométrique qui permet de passer de $x - \bar{x}$ à $\frac{x-\bar{x}}{\sigma(x)}$ est une homothétie de centre G et de rapport $\frac{1}{\sigma(x)}$.

(c) **Théorème de Huygens** : La forme la plus simple du théorème de Huygens est la relation entre l'inertie d'un nuage par rapport à un point quelconque Z d'abscisse z et son inertie par rapport à G. La première est égale à la seconde augmentée de l'inertie, par rapport à Z, de G affecté du poids total du nuage :

$$\sum_{i\in K} p_i(x_i - z)^2 = \sum_{i\in K} p_i(x_i - \bar{x})^2 + \left(\sum_{i\in K} p_i\right)(\bar{x} - z)^2.$$

Remarque 4.2.1 *En appliquant cette relation à J sous-nuage de K, on obtient la forme décrite ci-après sous laquelle le théorème de Huygens est rencontré le plus souvent en statistique.*
L'inertie d'un nuage de points dans lequel on distingue J sous-nuages est la somme des inerties de ces sous-nuages par rapport à leur barycentre (inertie intra) augmentée de l'inertie du nuage des J barycentres chacun affecté du poids total du sous-nuage qu'il représente (inertie inter). Ceci s'écrit, en notant K_j le $j^{ième}$ sous-nuage, $\bar{x}_j$ son barycentre et p_j son poids ($p_j = \sum_{i\in K_j} p_i$) :

$$\sum_{i\in K} p_i(x_i - z)^2 = \sum_{j=1}^{k}\sum_{i\in K_j} p_i(x_i - \bar{x}_j)^2 + \sum_{j=1}^{k} p_j(\bar{x}_j - z)^2.$$

4.2.2 Cas de deux variables

ces propriétés se généralisent à un tableau de données comportant 2 variables x et y. L'ensemble des valeurs des 2 variables se représente par un nuage dans un plan rapporté à deux axes orthogonaux correspondant respectivement aux deux variables. Un individu i est représenté par un point dont les 2 coordonnées sont ses valeurs x_i et y_i.

(a) **Centrage et réduction** : Le point G de coordonnées $(\bar{x}, \bar{y})$ est le barycentre des points munis des poids p_i. Quand on retire à chaque valeur x_i la moyenne $\bar{x}$ et à chaque valeur y_i la moyenne $\bar{y}$, on obtient un tableau centré. La transformation géométrique qui permet de passer du nuage associé au tableau initial au nuage associé au tableau centré est une translation qui fait coïncider l'origine O et le barycentre $G(\bar{x}, \bar{y})$.
Quand on divise les valeurs $x_i - \bar{x}$ par $\sigma(x)$ et $y_i - \bar{y}$ par $\sigma(y)$, on obtient un tableau $z_i^{(x)} = \frac{x_i - \bar{x}}{\sigma(x)}$ et $z_i^{(y)} = \frac{y_i - \bar{y}}{\sigma(y)}$ centré-réduit. La transformation géométrique qui permet de passer du nuage centré au nuage centré-réduit est la composition de deux homothéties de centre G(la première, de rapport $\frac{1}{\sigma(x)}$ dans la direction de abscisse x, la seconde, de rapport $\frac{1}{\sigma(y)}$ dans la direction de abscisse y).

$$\begin{pmatrix} z_i^{(x)} \\ z_i^{(y)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma(x)} & 0 \\ 0 & \frac{1}{\sigma(y)} \end{pmatrix} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix}$$

Un nuage centré-réduit possède, en projection sur chaque axe, une inertie égale à 1.

(b) **Théorème de Huygens** : Le carré de la distance d'un point i à l'origine vaut

$$d(O, i) = \|\overrightarrow{Oi}\|^2 = x_i^2 + y_i^2.$$

On déduit que l'inertie I_i de i est

$$I_i = p_i d(O, i) = p_i\|\overrightarrow{Oi}\|^2 = p_i x_i^2 + p_i y_i^2.$$

D'où, l'inertie totale du nuage des points i est :

$$I = \sum_{i\in K} p_i\|\overrightarrow{Oi}\|^2 = \sum_{i\in K} p_i x_i^2 + \sum_{i\in K} p_i y_i^2.$$

L'inertie du nuage se décompose donc suivant les deux axes : elle est la somme des inerties de ses deux propjections suivant les deux directions orthogonales. Si les variables sont centrées-réduites, l'inertie du nuage vaut 1 dans chaque direction et vaux donc 2 dans le plan.

Le théorème de Huygens se généralise sans difficulté au cas de deux variables puisque l'inertie d'un nuage se décompose sur chaque axe suivant le théorème de Pythagore :

$$\sum_{i\in K} p_i \|\overrightarrow{Oi}\|^2 = \sum_{j=1}^{k}\sum_{i\in K_j} p_i \|\overrightarrow{iG_j}\|^2 + \sum_{j=1}^{k} p_j \|\overrightarrow{G_jZ}\|^2,$$

avec $\|\overrightarrow{iG_j}\|^2 = (x_i - x_{G_j})^2 + (y_i - y_{G_j})^2$ et $\|\overrightarrow{G_jZ}\|^2 = (x_{G_j} - x_Z)^2 + (y_{G_j} - y_Z)^2$.

Dans ce qui suit, nous procédons à la génaralisation de l'étude au cas de plusieurs variables :

4.3 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 $i^{\text{ième}}$ individu sont représentées par la $i^{\text{ième}}$ 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 $\mathbb{R}^p$.

Dans la suite, on écrira aussi la $i^{\text{ième}}$ ligne X_i^T de la matrice X sous la 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 $i^{\text{ième}}$ colonne (de longueur n) de la matrice X.

Définition 4.3.1 *Une distance entre les données X_i et X_j des $i^{\text{ième}}$ et $j^{\text{ième}}$ individus, est un nombre noté $d(X_i, X_j)$, qui satisfait les propriétés suivantes :*

(a) $d(X_i, X_j) = d(X_j, X_i)$ *pour tout* i *et* j.

(b) $d(X_i, X_j) \geq 0$ *pour tout* i *et* j.

(c) $d(X_i, X_j) = 0$, *si et seulement si* $i = j$.

(d) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ *pour tout* i, j *et* k.

Ainsi une distance représente une dissimilarité entre individus. cependant, on parlera de **dissimilarité**, au sens strict du termes, seulement lorsque les propriétés $(a) - (b) - (c)$ 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 : $d(X_i, X_j) = \sqrt{(X_i - X_j)^T C(X)(X_i - X_j)}$ où $C(X)$ est la matrice de covariance de X.

Exemple 4.3.1 *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

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

$$d(X_1, X_3) = \sqrt{(1.5-8.2)^2 + (2-2.7)^2 + (3-9)^2 + (2.8-1.2)^2} = 9.161,$$

et

$$d(X_2, X_3) = \sqrt{(1-8.2)^2 + (3.1-2.7)^2 + (6.2-9)^2 + (5.3-1.2)^2} = 8.755.$$

- **Exemple 2 : Similarité entre objets décrits par les individus binaires**
  Une question importante en biologie est la classification des espèces (penser aux classification 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 4.3.2 *On a 5 individus, et les variables sont*

  1. *Var*1 *: Présence / abssence d'ailes.*
  2. *Var*2 *: Présence / abssence de pattes.*
  3. *Var*3 *: Présence / abssence 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** : $d(X_i, X_j) = 1 - \frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}}$
  2. **Russel et Rao** : $d(X_i, X_j) = 1 - \frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}+d_{ij}}$

  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 la différence avec 1. c'est -à-dire $1 - \ldots$.
  Exercice 4.3.1 *Calculer les distances de Jaccard dans l'exemple ci-dessus.*
- **Exemple 3 : Abondance d'espèces en écologie**
  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 donnés 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 $i^{\text{ième}}$ et $j^{\text{ième}}$ individus de sorte à ce qu'ils soient sur la sphère de rayon 1 dans $\mathbb{R}^p$:

$$\tilde{X}_i = \frac{X_i}{\|X_i\|} \quad \text{et} \quad \tilde{X}_j = \frac{X_j}{\|X_j\|}.$$

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 représentées sont répertoriées dans une matrice notée $D = (d_{ij})$, de taille $n \times n$, telle que :

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

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

4.4 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.

Propriété 4.4.1 *Soit $S(n, k)$ le nombre de partitions de n éléments en k parties. Alors, $S(n, k)$ satisfait la relation de récrrences suivante :*

$$\begin{cases} S(n,k) &= kS(n-1,k) + S(n-1,k-1), \quad \textit{pour} \quad k = 2, \ldots, n-1. \\ S(n,n) &= S(1,1) = 1. \end{cases}$$

Démonstration.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 :

(a) 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 manière de la rajouter.

(b) Soit on partitionne les $(n-1)$ premiers objets en $(k-1)$ groupes, et le dernier élément forme ungroupe 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 :

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\mathrm{C}_k^j j^n.$$

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 premières valeurs de $S(n)$ sont : $S(1) = 1$, $S(2) = 2$, $S(3) = 5$, $S(4) = 15$, $S(5) = 52$, $S(6) = 203$, $S(7) = 877$, $S(8)4140$, $S(9) = 21147$,. . ., $S(20) = 51724158235372$.

Définition 4.4.1 *Le nombre $S(n,k)$ de partitions de n éléments en k paries 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.*

4.5 Inertie d'un nuage de points

Cette section introduit une notion dont le ,ien 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.

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

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 moyenne $\bar{\mathrm{x}}^T$ 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 $\bar{\mathrm{x}}^T$ écrit sous forme d'une colonne est noté $\bar{\mathrm{x}}$. Alors :

Définition 4.5.1
1. *L'inertie du nuage de points par rapport à son centre de gravité est la somme pondérée des éloignements au centre de gravité.*
2. *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, c'est-à-dire*

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

c'est une sorte de variance pour le $i^{ième}$ individu.

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

$$I = \frac{1}{n}\sum_{i=1}^{n} I_i = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \bar{\mathrm{x}}\|^2.$$

Propriété 4.5.1 *1. I caractérise la* **dispersion** *ou la* **forme** *du nuage par rapport à son centre : au plus I est élevée, au plus le nuage est dispersé autour de son centre de gravité.*

2. Une inertie nulle signifie que tous les individus sont identiques.

3. Lorsque les variables sont centrées et réduites alors $I = 1$.

4. L'inertie mesure la quantité d'information contenue dans le tableau de données X.

4.5.2 Inertie inter-classe, inertie intra-classe

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

Définition 4.5.2 *1. L'inertie de l'individu i dans la classe C_ℓ est :*

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

*et l'***inertie de la classe** *C_ℓ est la somme des inerties des individus dans cette classe, c'est-à-dire*

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

2. *L'***inertie intra-classe**, *notée I_{intra}, est par définition :*

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

3. *L'***inertie inter-classe**, *notée I_{inter}, est la somme pondérée des inerties des centres de gravité des différentes classes, c'est-à-dire :*

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

Remarque 4.5.1 *Si une classe est "bien regroupée" autour de son centre de gravité, son inertie est* **faible**. *Ainsi, un bon critère pour avoir des classes homogène est d'avoir une inertie intra-classe qui soit* **petite que possible**.

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

Comme conséquence du théorème de Huygens, on a

$$I = I_{inter} + I_{intra}$$

où bien

$$I = \sum_{i \in K} p_i \|\overrightarrow{Oi}\|^2 = \sum_{j=1}^{k} \sum_{i \in K_j} p_i \|\overrightarrow{iG_j}\|^2 + \sum_{j=1}^{k} p_j \|\overrightarrow{G_j Z}\|^2 = \sum_{i=1}^{k} \sigma^2(X_i),$$

avec $\|\overrightarrow{iG_j}\|^2 = (x_i - x_{G_j})^2 + (y_i - y_{G_j})^2$ et $\|\overrightarrow{G_j Z}\|^2 = (x_{G_j} - x_Z)^2 + (y_{G_j} - y_Z)^2$.

On peut encore définir une distance ou éloignement entre individus par :

$$d^2(e_i, e_k) = \|e_i - e_k\|^2 = \sum_{j=1}^{p} (e_{ij} - e_{kj})^2 = (e_i - e_k)^T (e_i - e_k)$$

donc l'éloignement d'un point e_i du nuage par rapport au centre de gravité est :

$$d^2(e_i, G) = \sum_{j=1}^{p} (e_{ij} - \bar{x}_j)^2.$$

4.6 Méthodes non hiérarchiques : méthode de centres mobiles

Origine et extension

L'origine de la méthode provient de Forgy, Mac Queen et 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 provisoirs 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.

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_{intra} est minimum (on a minimum "local" et non "global").

Remarque 4.6.1 *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-Khawarizmi ($\sim 780 - -850$), puis le savant arabe Averoès (12ième siècle) évoque une méthode similaire. C'est un moine nommé Adelard de Barth (12ième 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'***algorithme** *est actuellement une science à part entière.*

4.7 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 $C_1, \ldots, C_n$, où chaque individu représente une classe. On suppose donné la matrice des distance entre individus.
2. **Etape** k($k = 0, \ldots, n-1$). Les données sont $n-k$ classes $C_1, \ldots, C_{n-k}$, 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 correspondante. Obtenir ainsi $(n-k-1)$ nouvelles classes, $C_1, \ldots, C_{n-k-1}$.
   – recalculer la matrice des distances qui donnent les $\frac{(n-k)(n_k-1)}{2}$ distances entre les nouvelles classes.
   – Poser $k := k+1$.

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.

Avantages

Algorithme simple, permettant une très bonne lecture des données.

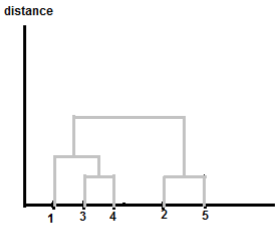

FIGURE 4.1 – Arbre/ dendogramme correspondant.

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 staables.
2. Choix de la bonne partition : repérer un saut (si possible) entre les agrégations courtes distances (branches courtes de l'arbre) et les groupes 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 ordre de $\mathcal{O}(n^3)$, 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 et 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
$$d(C, C') = \min_{\substack{X_i \in C \\ X_j \in C'}} d(X_i, X_j)$$
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
$$d(C, C') = \max_{\substack{X_i \in C \\ X_j \in C'}} d(X_i, X_j)$$
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 distance entre tous les individus des deux classes.

4. **Méthode de Ward pour les 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 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)

$$\delta(C, C') = \frac{|C|}{n}\|\bar{x}(C) - \bar{x}\|^2 + \frac{|C'|}{n}\|\bar{x}(C') - \bar{x}\|^2 - \frac{|C| + |C'|}{n}\|\bar{x}(C \cup C') - \bar{x}\|^2$$

$$\delta(C, C') = \frac{|C|.|C'|}{|C| + |C'|}\|\bar{x}(C) - \bar{x}(C')\|^2$$

   où $\bar{x}(C \cup C')$ 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.

4.8 Algorithme de Ward

1. Au pas k, en agrégeant deux éléments (individus et / ou groupes d'individus), on passe d'une partition en $n - k + 1$ classes à une partition en $n - k$ classes.
2. La nouvelle partition (en $n - k$ classes), présente une inertie intra plus grande (éventuellement égale) que celle de la précédente (en $n - k + 1$ classes) : en agrégeant deux classes, on ne peut qu'augmenter l'inertie intra. Cela découle d'une autre forme du théorème Huygens selon laquelle l'inertie d'un nuage par rapport à un point est minimum lorsque ce point est le centre de gravité du nuage (ce qui fait apparaître aussi que l'inertie intra n'augmente pas dans le seul cas très particulier où less deux classes agrégées ont le même centre de gravité).

L'idée de Ward consiste à choisir à chaque pas le regroupement de classes tel que l'**augmentation de l'inertie intra soit minimum**. Cet algorithme ne fournit évidement pas de partitions globalement optimales (sauf au premier pas ce qui est sans intérêt pratique) : il faudrait pour cela remettre en cause à chaque pas les regroupements du pas précédent mais cela ferait perdre l'emboîtement des partitions et donc l'arbre hiérarchique.
Si l'on note :

1. G_i (resp. G_j) le centre de gravité de la classe C_i (resp. C_j),
2. p_i (resp. p_j) la somme des poids des éléments de la classe C_i (resp. C_j),

on montre que l'augmentation de l'inertie intra due au regroupement des classess C_i et C_j s'écrit :

$$\delta(C_i, C_j) = \frac{p_i p_j}{p_i + p_j}\|G_i - G_j\|^2.$$

Tel que le critère minimisé à chaque pas et qui définit l'indice de niveau des nœuds de la hiérarchie. Cette écriture fait apparaître que, à chaque pas, on rgroupe des classes :

1. proches, c'est-à-dire que $\|G_i - G_j\|^2$ soit petit ;
2. de faibles poids, c'est-à-dire telles que $p_{ij} = \frac{p_i p_j}{p_i + p_j}$ soit petit.

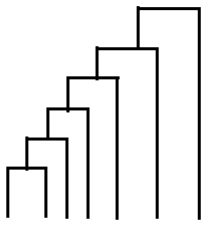

FIGURE 4.2 – Arbe hiérarchique présentant un effet de chaîne. Les individus s'agrègent un par un au groupe déjà constitué. Les partitions obtenues par coupure d'un tel arbe, mettant toutes en évidence un seul groupe et des individus isolés, sont généralement sans intérêt pratique.

ce dernier point montre bien pourquoi l'algorithme de Ward est peu sensible à l'effet de chaîne, fréquent par exemple lorsque l'on utilise l'algorithme du saut minimum, qui conduit à des arbres difficilement exploitable (voir Figure.4.2) : l'algorithme de Ward favorise l'agrégation entre eux des éléments isolés.

On peut montrer que, lorsque l'algorithme de Ward agrège la classe k à la classe (constituée à une étape antérieure de l'algorithme) réunissant les classes i et j, l'augmentation de l'inertie intra est plus grande que celle consécutive à l'agrégation des classes i et j. Soit $\delta(k, \{i, j\}) \geq \delta(i, j)$. L'augmentation d'inertie intra étant utilisée comme indice de niveau, cette propriété assure que l'arbre hiérarchique ne présente par d'inversion (voir Figure.4.3).

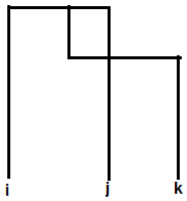

FIGURE 4.3 – Inversion dans un arbre hiérarchique. k s'agrège au groupe $\{i, j\}$ à niveau inférieur à celui de l'agrégation entre i et j. Ce phénomène est impossible avec les algorithmes usuels.

Chapitre 5

L'analyse factorielle des correspondances

5.1 Données, Notations, Hypothèse d'indépendance

A l'origine, l'Analyse Factorielle des Correspondances (AFC) a été conçue pour étudier des tableaux appelés couramment tableaux de contingence (ou tableaux croisés). Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même popultation de n individus. La Figure.5.1 résume les principales notations.

Exemple 5.1.1 *Dans cet exemple, la population est constituée par l'ensemble des individus qui ont quitté le système scolaire français en 1792 et qui occupent un emploi en 1973 ; pour chaque individu, on connaaît son niveau de diplôme et sa catégorie d'emploi.*

TABLE 5.1 – (1)=Sans Diplôme, (2)=BEPC, (3)=BEP/CAP, (4)=BAC Général, (5)=BAC Technique, (6)=DEUG/ENT, (7)=DUT/BTS, (8)=SUP. Elèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi : sexe masculin.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Total
Agriculteur	15068	2701	5709	297	1242	-	322	-	25339
Ingénieur	-	337	309	917	-	308	-	4383	6254
Technicien	302	1697	2242	1969	1399	357	1943	381	10290
Ouvrier qualifié	10143	3702	30926	314	1861	-	-	337	47283
Ouvrier non qualifié	59394	8087	17862	2887	1696	-	-	323	90249
Cadre supérieur	596	298	892	1227	298	2362	318	6781	12772
Cadre moyen	2142	2801	672	6495	924	2807	2301	4030	22172
Employé qualifié	5445	7348	4719	4353	1280	614	982	-	24741
Employé non qualifié	4879	4987	1514	3478	886	1326	-	661	17731
Total	97969	31958	64845	21937	9586	7774	5866	16896	256831

TABLE 5.2 – (1)=Sans Diplôme, (2)=BEPC, (3)=BEP/CAP, (4)=BAC Général, (5)=BAC Technique, (6)=DEUG/ENT, (7)=DUT/BTS, (8)=SUP. Elèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi : sexe féminin.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Total
Agriculteur	5089	1212	1166	-	-	-	-	-	7467
Ingénieur	-	-	-	316	-	-	304	1033	1653
Technicien	281	-	320	320	283	-	683	-	1887
Ouvrier qualifié	7470	1859	4017	1752	657	-	285	-	16040
Ouvrier non qualifié	299974	4334	4538	1882	-	-	-	-	40751
Cadre supérieur	-	-	-	2236	595	911	569	6788	11099
Cadre moyen	1577	1806	4549	17063	875	4152	15731	3991	49744
Employé qualifié	21616	19915	32452	16137	5865	1256	3332	1286	101859
Employé non qualifié	19849	7325	6484	5111	898	294	635	-	40596
Total	85879	36451	53526	44817	9173	6613	21539	13098	271096

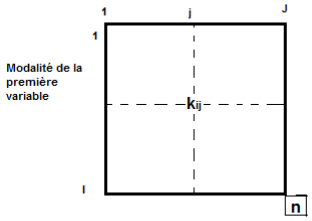

FIGURE 5.1 – Tableau des données brutes. I : ensemble des lignes et nombre de lignes (8 niveaux de diplôme). J : ensemble des colonnes et nombre de colonnes (9 catégories d'emploi). k_{ij} : nombres d'individus possédant à la fois la modalité i de la première variable et la modalité j de la seconde (i.e. qui ont le niveau de diplôme i et qui occupent un emploi de la catégorie j). $\sum_i \sum_j k_{ij} = n$ le nombre tatal d'individus.

On parle indifféremment de la modalité i (par exemple le baccalauréat) ou de la classe i, c'et-à-dire de la classe des individus qui possèdent la modalité i (par exemple les bacheliers).
Dans ce chapitre, nous nous limitons à l'étude d'un tableau de contingence. Cependant, la plupart des notations introduites et des résultats présentés peuvent être généralisés à des tableaux qui ne sont pas strictement de ce type. Le cas très important du tableau disjonctif complet fait l'objet d'une thématique "Analyse des Correspondances Multiples". La conclusion du présent chapitre donne quelques points de repère sur l'application de l'AFC à d'autres tableaux que les tableaux de contingence.
On considère souvent le tableau des fréquences relatives F, obtenu en divisant chaque effectif k_{ij} par l'effectif total n. Ce nouveau tableau définit une mesure de probablité sur l'ensemble produit $I \times J$. Ses marges, ou probabilités marginales, ont pour termes général $f_{i.}$ pour la marge-colonne et $f_{.j}$ pour la marge-ligne (voir Figure.5.2)

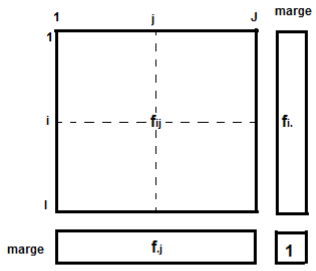

FIGURE 5.2 – Tableau F des fréquences relatives et ses marges

$$
\begin{aligned}
f_{ij} &= \frac{1}{n} k_{ij}, \\
f_{i.} &= \sum_j f_{ij}, \\
f_{.j} &= \sum_i f_{ij}, \\
\sum_i f_{i.} &= \sum_j f_{.j} = \sum_i \sum_j f_{ij} = 1.
\end{aligned}
$$

Un tableau de contingence exprime la liaison entre deux variables qualitatives. Classiquement, pour une mesure de probabilité, on dit qu'il y a **indépendance** entre les deux variables lorsque, pour tout i et pour tout j, on a l'égalité

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

Il y a **liaison** entre les deux variables dès que certaines cases du tableau f_{ij} différent du produit $f_{i.} f_{.j}$. Si f_{ij} est supérieur à ce produit, les modalités i et j s'associent plus qu'elles ne le font dans l'hypothèse d'indépendance : On dit que i et j s'attirent.
Au cntraire, si f_{ij} est inférieur au produit des marges, i et j s'associent au moins que dans l'hypothèse d'indépendance : on dit qu'il y a répulsion entre ces deux modalités.

L'**indépendance** s'exprime aussi en considérant le tableau comme un ensemble de lignes. En effet, l'égalité ci-dessus est équivalente à l'égalité

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

La quantité $f_{.j}$ représente le pourcentage de la population totale qui possède la modalité j tandis que $\frac{f_{ij}}{f_{i.}}$ représente ce même pourcentage dans la sous-population possédant la modalité i. Lorsqu'il y a indépendance, les I sous-populations caractérisées par les modalités i de la première variable se répartissent selon les J modalités j de la deuxièmevariable avec les mêmes pourcentages. Toutes les lignes sont alors proportionnelles. La réciproque est vraie : lorsque toutes les lignes sont proportionnelles, elles sont proportionnelles à la marge $f_{.j}$ et les deux variables sont indépendantes. Il y a **liaison** dès lors que les lignes ne sont pas toutes proportionnelles à la marge, c'est-à-dire lorsqu'elles ne sont pas identiques du point de vue de leur association avec l'ensemble des colonnes.

5.2 Objectifs

Bien que le tableau étudié soit de nature très différente de celui étudié en ACP, les objectifs de l'AFC peuvent s'exprimer de manière analogue à ceux de l'ACP : on cherche à obtenir une typologie des lignes, une typologie de colonnes et à relier ces deux typologies entre elles ; mais la notion de ressemblance entre deux lignes, ou entre deux colonnes, est différente de celle de l'ACP.
Dans un tableau de contingence, la ressemblance, entre deux lignes d'une part et entre deux colonnes d'autre part, s'exprime de manière totalement symétrique. Deux lignes sont considérées comme proches si elles s'associent de la même façon à l'ensemble des colonnes ; les termes «trop» et «trop peu» sont pris en référnce à la situation d'indépendance. Symétriquement, deux colonnes sont proches si elles s'associent de la même façon à l'ensemble des lignes.
Schématiquement, l'étude de l'ensemble des lignes revient à mettre en évidence une typologie dans laquelle on cherche les lignes dont la répartition s'écarte le plus de celle de l'ensemble de la population, celles qui se ressemblent entre elles (dans le sens précisé ci-dessus) et celles qui s'opposent. Pour mettre en relation la typologie des lignes avec l'ensemble des colonnes, on caractérise chaque groupe de lignes par les colonnes auxquelles ce groupe s'associe trop ou trop peu.
L'étude de l'ensemble des colonnes est absolument analogue. Cette approche, grâce à la notion de ressemblance utilisée, permet d'étudier la liaison entre les deux variables, c'est-à-dire l'écart du tableau à l'hypothèse d'indépendance. L'analyse de cette liaison est l'objectif fondamental de l'AFC.
Enfin, bien qu'il y soit fait peu référence par la suite, il faut signaler que l'AFC, comme toute Analyse Factorielle, est utilisée aussi dans le but de réduire la dimension des données en conservant le plus d'information possible.

5.3 Transformations des données en profils

En AFC, le tableau brut n'est pas analysé directement. Dans l'étude des lignes, le tableau des données est transformé en divisant chaque terme f_{ij} de la ligne i par la marge $f_{i.}$ de cette ligne i. La nouvelle ligne est appelée profil-ligne (voir Figure). Cette trasformation découle de l'objectif qui vise à étudier la liaison entre les deux variables au travers de l'écart entre les pourcentage en lignes. Elle se justifie aussi de façon directe puisque la comparaison de deux lignes du tableau brut risque d'être influencée principalement par leurs effectifs marginaux. Ainsi, dans le tableau croisant emplois et diplômes, la différence entre les lignes brutes Bac technique et Bac général traduit essentiellement une différence entre les effectifs globaux de ces deux diplômes.

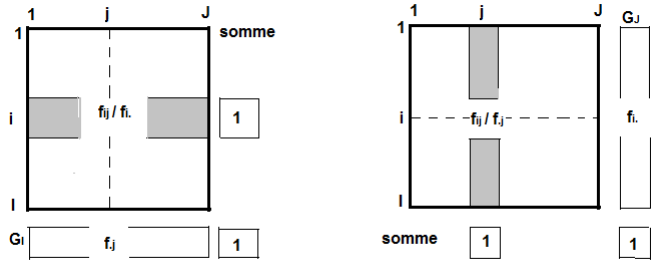

FIGURE 5.3 – Profil-ligne (à gauche) et profil-colonne (à droite). G_I et G_J : profils marginaux

Le nombre $\frac{f_{ij}}{f_{i.}}$ représente, dans notre exemple, la probabilité d'occuper un emploi de la catégorie j sachant que l'on détient le niveau de diplôme i. Le profil ligne i n'est rien d'autre que la loi de probabilité conditionnelle définie par i sur l'ensemble des colonnes. Pour analyser l'écart à l'indépendance, on confronte ces profils au profil ligne marginal (=établi sur l'ensemble de la population) de terme général $f_{.j}$ et noté G_J.

Du fait du rôle symétrique joué par les lignes et les colonnes, un raisonnement analogue peut être mené à propos des colonnes. Il conduit à la notion de profil-colonne (voir Figure.5.3).

Ainsi, en AFC, selon que l'on s'intéresse aux lignes ou aux colonnes, on ne considère pas le même tableau transformé. Toutefois, les deux transformations en profils possèdent la même signification vis-à-vis des objets qu'elles concernent. Ces transformations sont intéressantes en elles-mêmes indépendamment de tout contexte d'analyse factorielle. Lorsqu'un tableau croisé est commenté, il est presque toujours présenté sous la forme de pourcentages, par rapport aux lignes ou aux colonnes selon les aspects que l'on cherche à mettre en évidence.

Exemple 5.3.1 *Il s'agitd'une analyse factorielle pour deux variables quantitatives, sous forme d'un tableau de* **contingence** *(tableau tris croisés).*

- *Elle repose sur l'étude de la liaison (Khi-deux=χ^2) entre les deux variables.*
- *Les modalités des deux variables seront représentées dans un graphique à 2 ou 3 dimensions.*

TABLE 5.3 – Tableau des données"Tableau de contingence"

	Limousin	France	International
CAI	1	2	5
AGE	11	3	2
AGT	3	7	1

TABLE 5.4 – Tableau des données"Tableau à plat"

Mention	Lieu	Effectif
CAI	Limousin	1
CAI	France	2
CAI	International	5
AGE	Limousin	11
AGE	France	3
AGE	International	2
AGT	Limousin	3
AGT	France	7
AGT	International	1

L'analyse des correspondances réalise deux analyses factorielles, une sur chacun des nuages :
- *Nuage des* **profils-lignes** *: répartition en probabilité où bien en fréquence (%) des individus suivant les modalités de la deuxième variable ;*
- *Nuage des* **profils-colonnes** *: répartition en probabilité où bien en fréquence (%) des individus suivant les modalités de la première variable.*

TABLE 5.5 – Tableau profils-lignes

	Limousin	France	International	Somme
CAI	$\frac{1}{8} \times 100 = 12,5\%$	$\frac{2}{8} \times 100 = 25\%$	$\frac{5}{8} \times 100 = 62,5\%$	8
AGE	$\frac{11}{16} \times 100 = 68,8\%$	$\frac{3}{16} \times 100 = 18,8\%$	$\frac{2}{16} \times 100 = 12,4\%$	16
AGT	$\frac{3}{11} \times 100 = 27,3\%$	$\frac{7}{11} \times 100 = 63,6\%$	$\frac{1}{11} \times 100 = 9,1\%$	11

TABLE 5.6 – Tableau profils-colonnes

	Limousin	France	International
CAI	$\frac{1}{15} \times 100 = 6,7\%$	$\frac{2}{12} \times 100 = 16,7\%$	$\frac{5}{8} \times 100 = 62,5\%$
AGE	$\frac{11}{15} \times 100 = 73,3\%$	$\frac{3}{12} \times 100 = 25,0\%$	$\frac{2}{8} \times 100 = 25,0\%$
AGT	$\frac{3}{15} \times 100 = 20,0\%$	$\frac{7}{12} \times 100 = 58,3\%$	$\frac{1}{8} \times 100 = 12,5\%$
Somme	15	12	8

5.4 Ressemblance entre profils : Distance du χ^2

En AFC, la ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre leurs profils connue sous le nom de distance du χ^2. Elle est définie de façon symétrique pour les lignes et pour les colonnes. Soit :

$$d_{\chi^2}(\text{profil-ligne } i, \text{profil-ligne } \ell) = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{\ell j}}{f_{\ell .}} \right)^2,$$

$$d_{\chi^2}(\text{profil-colonne } j, \text{profil-colonne } k) = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2$$

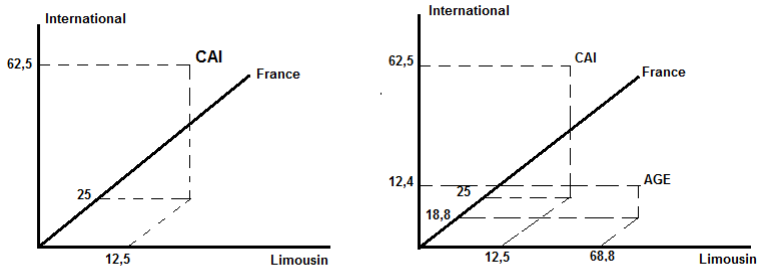

FIGURE 5.4 – Représentation graphique des Profils-lignes

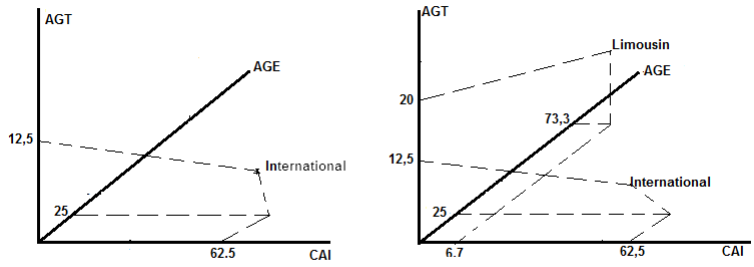

FIGURE 5.5 – Représentation graphique des Profils-colonnes

Dans ces relations, la distance entre deux lignes dépend essentiellement des différences terme à terme les deux profils dont elle fait une somme des carrés pondérés. La pondération $\frac{1}{f_{.j}}$ équilibre l'influence des colonnes sur la distance entre les lignes : elle augmente les termes, *a priori* plus faibles, concernant les modalités rares ; elle joue, jusqu'à un certain point, un rôle analogue à celui de la division par l'écart-type dans le cas des variables numériques.

La distance du χ^2 jouit d'une propriété fondamentale appelée **équivalence distributionnelle**. Selon cette propriété, si deux colonnes proportionnelles d'un tableau sont cumuléses en une seule, la distance entre les profils-lignes est inchangée. Le cas d'une proportionnalité parfaite entre deux colonnes ne se rencontre guère en pratique mais constitue une situation limite dont on peut être assez proche.

5.5 La dualité

Les deux nuages N_I et N_J constituent deux représentations d'une même tableau, l'une à travers ses profils-lignes, l'autre à travers ses profils-colonnes. Il s'ensuit que les analyses de ces deux nuages ne sont pas indépendantes : les relations entre ces deux analyses sont communément regroupées sous le terme de dualité.
Cette dualité est plus fondamentale et plus riche en AFC qu'en ACP car les lignes et les colonnes représentent des objets de même nature, ce qui n'est pas le cas en ACP.

5.5.1 Statistique χ^2 et inertie des deux nuages N_I et N_J

Lorsque l'on étudie un tableau de contingence, c'est-à-dire une population de n individus au travers de deux variables qualitatives, il est classique de mesurer la significativité de la liaison entre ces deux variables à l'aide de la statistique χ^2. Applliquée à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques que l'on obtiendrait en moyenne si les deux variables étaient indépendantes. Elle s'écrit :

$$\chi^2 = \sum_i \sum_j \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}} = n. \sum_{i,j} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

La statistique χ^2 est égale, au coefficient n près, à l'inertie totale par rapport à leur barycentre de l'un ou l'autre des nuages N_I et N_J. En effet, dans $\mathbb{R}^I$, l'inertie totale de N_I par rapport à G_I s'écrit

$$\text{Inertie}(N_I) = \sum_i \text{Inertie}(i) = \sum_i f_{i.}\, d^2(i, G_I) = \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

Soit :

$$\chi^2 = n\text{Inertie}(N_I) = n.\text{Inertie}(N_J) = n. \sum_i \sum_j \frac{1}{f_{i.}f_{.j}} (f_{ij} - f_{i.}f_{.j})^2$$

Cette double égalité montre que l'inertie totale de chacun des deux nuages N_I et N_J représente, sous deux formes différentes, la liaison entre les deux variables.
Finalement, l'inertie de chaque nuage est égale à l'écart à l'indépendance des deux variables :

$$\Phi^2 = \frac{\chi^2}{n} = \sum_i \sum_j \frac{1}{f_{i.}f_{.j}} (f_{ij} - f_{i.}f_{.j})^2 .$$

Cette quantité mesure l'intensité de la liaison entre deux variables qualitatives (cette liaison est d'autant plus intense que les modalités de l'une s'associent exclusivement aux modalités de l'autre) et non sa significativité (elle ne dépend pas de l'effectif total) ; l'indicateur χ^2, lui, mesure la significativité (une liaison forte peut ne pas être significative si elle est observée sur très peu d'individus ; une liaison faibles peut être significative si elle est observée sur beaucoup d'individus).

– L'inertie de chaque nuage est mesuré par la relation

$$\Phi^2 = \sum_i \sum_j \frac{1}{f_{i.}f_{.j}} (f_{ij} - f_{i.}f_{.j})^2 .$$

– Les nuages sont projetés dans l'espace à 2 ou 3 dimensions qui minimise la perte d'inertie.
– L'origine des axes correspond au **profil moyen**.

5.5.2 Dualité entre les facteurs sur I et les facteurs sur J

De même q'en ACP, on appelle **facteur** l'ensemble des coordonnées des projections des points d'un nuage sur l'un des axes factoriels ; les facteurs sur les lignes sont les projections de N_I et les facteurs sur les colonnes les projections de N_J. Le rang d'un facteur est le rang de l'axe factoriel correspondant. Outre leur inertie totale identique, les nuages N_I et N_J possèdent une propriété remarquable : leur ajustement conduit à deux suites de facteurs «duaux». Plus précisément, nous montrons que

1. les inerties associées aux axes de même rang dans chacun des nuages sont égales ;

2. les facteurs (de même rang) sur les lignes et ceux sur les colonnes sont liés par des relation dites de transition (elles permettent de transiter de $\mathbb{R}^I$ dans $\mathbb{R}^J$ et inversement).

Les deux paragraphes suivants détaillent cette dualité dont la conséquence essentielle est la suivante : les facteurs sur I et sur J de même rang doivent être interprétés conjointement car ils mettent en évidence la même part de liaison, exprimée pour l'un en termes dites **profils-lignes** et pour l'autre en termes de **profils-colonne**.

a) **Relations de transition** : Les formules de transition précisent les relations entre les points représentant d'un part les lignes et d'autre part les colonnes. Avec le notations suivantes :

(i) $F_s(i)$: projection de ligne i sur l'axe de rang s de N_I,

(ii) $G_s(j)$: projection de la colonne j sur l'axe de rang s de N_J,

(iii) λ_s : valeur commune de l'inertie associée à chacun de ces deux axes,

Les deux relations de transition s'écrivent :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{j.}} F_s(i)$$

Ces deux prpriétés, qui expriment les résultats de l'analyse d'un nuage en fonction des résultats de l'analyse de l'autre nuage, conduisent à une économie de calcul. Mais surtout, elles donnent un sens à une représentation simultanée des lignes et des colonnes.

b) **Représentation simultanée des lignes et des colonnes ; relations barycentriques** : La représentation simultanée s'obtient en superposant les projections de chacun des deux nuages N_I et N_J sur des plans engendrés par des axes de même rang pour les deux nuages. Sur les graphiques ainsi obtenus, les rapports entre la position des points lignes et les poins colonnes dus aux relations de transition peuvent être décrits ainsi :

" au coefficient $\frac{1}{\sqrt{\lambda_s}}$ près, la projection, notée $F_s(i)$, de la ligne i sur l'axe de rang s (dans $\mathbb{R}^J$) est la barycentre des projections, notées $G_s(j)$, des colonnes j sur l'axe de rang s (dans $\mathbb{R}^I$), chaque colonne j étant affectée du poids $\frac{f_{ij}}{f_{i.}}$ (cette expression d'une formule de transition est appelée propriété barycentrique)."

Les éléments "lourds" attirant le barycentre, une colonne j attire d'autant plus une ligne i que la valeur de $\frac{f_{ij}}{f_{i.}}$ est élevée.

Sur les plans factoriels, les points éloignés de l'origine retiennent particulièrement l'attention car ce sont les profils les plus différents du profil moyen. On trouve donc, pour un facteur, du même côté qu'une ligne i les colonnes j auxquelles elle s'associe le plus et, à l'opposé, celles auxquelles elle s'associe le moins. Il est ainsi possible d'interpréter la position d'**une ligne** par rapport à l'**ensemble des colonnes**, ce qui justifie l'intérêt pratique de la représentation simultanée.

La formulation symétrique vaut, en inversant les rôles joués par les lignes et les colonnes. D'où le nom de double propriété barycentrique donnée à ce qui est **la principale règle d'interprétation des graphiques de l'AFC**. Cette double propriété est non seulement spécifique de l'AFC en cherchant à construire des fonctions définies sur les lignes et les colonnes d'un tableau de contingence telles que la double propriété barycentrique soit vérifiée.

La représentation simultanée en AFC est universellement adoptée, ce qui n'est pas le cas de celle de l'ACP. On peut citer deux arguments importants en faveur de cette superposition.

– Alors qu'en ACP les lignes et les colonnes représentent des objets de nature bien différentes (individus et variables), les lignes et les colonnes, daans l'AFC d'un tableau de contingence, sont de même nature, à savoir des classes d'individus. Selon ce simple point de vue, cela ne pose aucun problème de figurer toutes ces classes sur un même graphique.

– Il existe d'autres présentations de l'AFC dans lesquelles les classes d'individus que constituent les lignes et les colonnes d'un tableau de contingence sont situées dans un même espace : leur représentation simultanée est alors naturelle.

En résumé, sur le graphique de la représentation simultanée des lignes et des colonnes, la position relative de deux points d'un même ensemble (lignes et colonnes) s'interprète en tant que distance tandis que la position d'**un** point d'un ensemble par rapport à celle de **tous** les points de l'autre ensemble s'interprète en tant que barycentre. Toute association entre **une** ligne et **une** colonne sugrérée par une proximité sur le graphique doit être contrôlée sur le tableau de données.

5.5.3 Interprétation de l'inertie des axes

L'inertie d'un point (ou d'un nuage de points) dans un espace euclidien se décompose sur toute base orthogonale : c'est la somme de ses inerties sur chacun des axes de cette baase.

L'ajustement des nuages N_I et N_J décompose leur inertie selon des directions privilégiées : du fait de l'orthogonalité des axes, la somme des inerties d'un nuage sur chacun des axes est égale à l'inertie totale du nuage.

Contrairement au cas de l'ACP, dans laquelle l'inertie des nuages est égale au nombre de variables, cette inertie en AFC traduit la structure du tableau : l'inertie de chacun des deux nuages, des profils-lignes et des profils-colonnes, est égale à la statistique Φ^2. L'AFC propose donc une décomposition de cette statistique et chaque facteur représente une part de la liaison entre les variables. L'inertie d'un facteur a donc une signification en absolu, et pas seulement en pourcentage de l'inertie totale du nuage : elle mesure en absolu l'importance de la part de liaison qu'il représente. Nous donnons l'interprétation des deux valeurs limites entre lesquelles elle se situe.

Lorsqu'un tableau vérifie les relations d'indépendance, les nuages sont concentrés en un point (leur barycentre) ; tous les profils-lignes sont identiques et égaux à la marge ligne $\{f_{.j};\quad j = 1, \ldots, J\}$ et tous les profils-colonnes sont identiques et égaux à la marge-colonne $\{f_{i.};\quad i = 1, \ldots, I\}$. L'inertie des nuages N_I et N_J relativement à leur centre de gravité est nulle et l'AFC ne donne aucun facteur (ou plutôt toute direction est associée à une inertie projetée nulle).

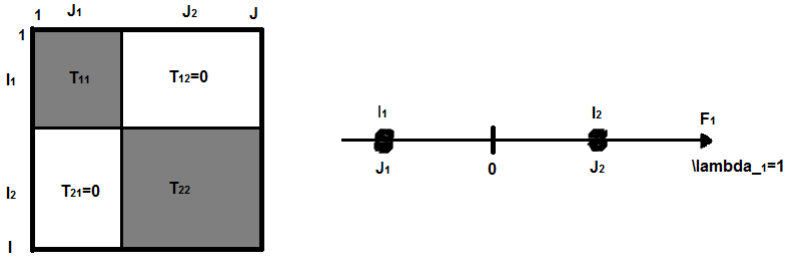

FIGURE 5.6 – Arbre/ dendogramme correspondant.

5.5.4 Formule de reconstitution des données

a la décomposition de l'inertie, on peut associer une décomposition du tableau lui-même. En effet, on peut montrer que

$$f_{ij} = f_{i.}f_{.j}\left(1 + \sum_s \frac{1}{\sqrt{\lambda_s}} F_s(i)G_s(j)\right)$$

Cette formule, appelée **formule de reconstitution des données**, permet de recalculer les valeurs du tableau initial en fonction des marges et des facteurs. Lorsque l'on dépouille les résultats d'une AFC, on limite généralement l'interprétation aux premiers facteurs. Cela revient à considérer non pas le tableau des données mais son approximation obtenue à l'aide des premiers termes de la somme ci-dessus.

Cette relation met en évidence une décomposition de l'écart du tableau relativement à l'hypothèse d'indépendance en une somme de tableaux dont chacun ne dépend que d'un cuple de facteurs (F_s, G_s) de même rang. Elle formalise l'aspect de l'objectif annoncé : décomposition de la liaison en éléments simple. En effet, chaque tableau de terme général $a_{ij} = f_{i.}f_{.j}F_s(i)G_s(j)$ exprime une liaison simple puisque le terme de la case (i, j) ne dépend que de la ligne i et de la colonne j.

Si le valeurs de $F_s(i)$ et de $G_s(j)$ sont de même signe, cette case exprime une attirance entre i et j ; dans le cas contraire, il exprime une répulsion d'autant plus importante que $F_s(i)$ et $G_s(j)$ sont grands en valeurs absolue.

5.6 Nombre d'axes et Inertie totale

Dans $\mathbb{R}^J$, le nuage N_I est contenu dans un sous-espace de dimension $J-1$; dans cet espace, on peut trouver au maximum $J-1$ dimensions orthogonales d'inertie non nulle. De même, dans l'espace $\mathbb{R}^I$, le nuage N_J est contenu dans un sous-espace de dimension $I-1$; dans cet espace, on peut trouver au maximum $I-1$ dimensions orthogonales d'inertie non nulle. Compte tenu de la dualité (même inertei sur les axes de même rang dans les deux espaces), en AFC on peut trouver au maximum $K = \min\{I-1, J-1\}$ axes d'inertie non nulle.

L'inertie associée à un axe étant au maximum égale à 1, l'inertie totale en AFC est donc comprise entre 0 (indépendance) et $K = \min\{I-1, J-1\}$ (liaison d'intensité maximum=association stricte entre les modalités des deux variables mises en correspondances).

Chapitre 6

Analyse factorielle : Calculs et dualité

6.1 Introduction

Les méthodes d'analyse factorielle sont fondées sur des principes communs : à partir d'un tableau de données, on construit deux nuages de points représentant respectivement les lignes et les colonnes ; ces deux nuages sont projetés chacun sur une suite d'axes orthogonaux maximisant l'inertie projetée ; sur chacun des axes , les deux nuages ont la même inertie projetée et les projections des points sont liées d'un nuage à l'autre par les relations dites de transition.

Dans ce chapitre, nous indiquons comment calculer ces facteurs, montrons la dualité des deux nuages et donnons des démonstrations des formules de transition. Le cadre dans lequel nous nous plaçons et assez général. Non seulement il recouvre l'ACP et l'AFC, mais il permet d'indroduire et de calculer les facteurs d'analyse factorielles fondées sur d'autres distances et d'autres poids.

6.2 Calcul des axes d'inertie et des facteurs d'un nuage de points

Le problème est posé en ces termes : étant donné un nuage de I points noté N_I dans un espace euclidien de dimension J, on cherche une suite d'axes orthonormés (pour laa métrique de l'espace $\mathbb{R}^J$) telle que l'inertie du nuage projeté sur ces axes soit maximum.

L'ensemble des coordonnées des I points du nuage sur un des axes définit une fonction numérique sur I, appelée facteur sur I. Dans les résultats d'une analyse, seuls les facteurs apparaissent, les axes n'étant que des intermédiaires de calcul. Pour obtenir les facteurs et leur inertie, nous utilisons des techniques simples de calcul matriciel.

6.2.1 Notations : les matrices X, M et D

Les coordonnées x_{ij} des I points du nuage N_I dans l'espace $\mathbb{R}^J$ forment un tableau , ou une matrice, de tailles $I \times J$, notée X. L'espace $\mathbb{R}^J$ est muni d'une métrique euclidienne qui peut être différente de la métrique canonique (ou usuelle). Cette métrique dérive d'un produit scalaire dont la matrice, de dimension $I \times J$, est notée M. Nous nous restreignons à des métriques associées à des matrices diagonales car elles seules sont facilement interprétables en termes de données initiales. En effet, lorsque M est diagonale, la distance d_M entre deux points i et ℓ de N_I s'écrit, en notant m_j les éléments diagonaux de M :

$$d^2_M(i,\ell) = \sum_j m_j (x_{ij} - x_{\ell j})^2.$$

Les coefficients m_j pondèrent l'influence de chaque colonne j dans les distances entre éléments ; cette propriété justifie leur nom de "poids des colonnes". Or, lorsque M n'est pas diagonale, ses termes apparaissent comme des poids associés à des couples de colonnes, ce qui n'a pas de résonance concrète.

Le produit scalaire (associé à d_M) entre deux vecteurs u et v s'écrit :

$$\langle u, v \rangle_M = u^T M v = v^T M u$$

avec u^T et v^T désignent les transposées de u et v, respectivement.

Les coordonnées des points de N_I et la métrique de l'espace $\mathbb{R}^J$ définissent entièrement la forme du nuage mais, dans le calcul des axes d'inertie, le poids des points de N_I intervient. Ces poids, notés p_i, sont rangés dans une matrice diagonale, de taille $I \times I$, notée D. Toute l'information nécessaire pour calculer les facteurs est contenue dans trois matrices X, M et D.

Dans ce qui suit nous ferons appel au cours d'algèbre linéaire, et sur tout les matrices et applications linéaires. Une matrice associée à une application linéaire et ou une matrice carrée associée à un endomorphisme.

6.2.2 Projection d'un nuage sur un axe

Soit u un vecteur unitaire, pour la métrique M, i.e. vérifiant $u^T M u = 1$, d'un axe quelconque de $\mathbb{R}^J$. L'ensemble des coordonnées des projections des I points du nuage N_I sur l'axe u constitue un vecteur de dimension I, que nous notons F_u. Pour tout point i du nuage N_I, $F_u(i) = x_i^T M u$ où x_i est le vecteur de $\mathbb{R}^J$ dont les coordonnées sont celles de i : x_i^T n'est autre que la ligne de la matrice X. De cette égalité, on déduit la relation matricielle

$$F_u = XMu.$$

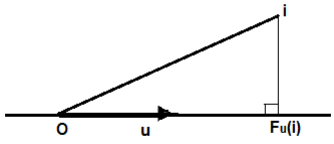

FIGURE 6.1 – Projection $F_u(i)$ du point i sur l'axe défini par le vecteur unitaire u

$$\begin{aligned} F_u(i) &= \langle x_i, u \rangle_M = x_i^T M u, \\ F_u &= XM_u, \\ \sum_i p_i \left(F_u(i)\right)^2 &= F_u^T D F_u. \end{aligned}$$

6.2.3 Inertie du nuage projeté

L'inertie du nuage projeté sur u est égale

$$\text{Inertie du nuage projeté sur } \ u = \sum_i p_i \left(F_u(i)\right)^2 .$$

Cette quantité s'écrit matriciellement en fonction de la matrice diagonale D et du vecteur F_u sous la forme $F_u^T D F_u$. Comme $F_u = XM_u$, l'inertie vaut $\mathcal{J}(u) = u^T M X^T D X M u$.

Chercher un axe de $\mathbb{R}^J$ tel que l'inertie du nuage projeté soit maximum revient donc à chercher un vecteur u, unitaire pour la métrique M (i.e. $u^T M u = 1$), rendant maximum la quantité $\mathcal{J}(u) = u^T M X^T D X M u$.

6.2.4 Calcul des axes d'inertie maximum ; cas de la métrique identité

Lorsque le produit scalaire sur $\mathbb{R}^J$ est le produit scalaire canonique, la matrice M est la matrice identité et l'écriture des expressions ci-dessus s'allège et devient plus facile : on cherche un vecteur u, vérifiant $u^T u = 1$ et rendant maximum $\mathcal{J}(u) = u^T X^T D X u$.

La matrice $X^T D X$ est symétrique, donc elle est diagnalisable, et ses vecteurs propres forment une base orthonormée de $\mathbb{R}^J$. Soient $\lambda_1, \ldots, \lambda_s, \ldots, \lambda_J$ les valeurs propres de la matrice $X^T D X$ rangées par ordre décroissant et $\{e_s;\ s = 1, \ldots, J\}$ une base orthonormée de vecteurs propres associés (i.e. $X^T D X e_s = \lambda_s e_s$). Décomposons le vecteur u sur cette base. On a

$$u = \sum_{s=1}^{J} u_s e_s \quad \text{avec} \quad \sum_{s=1}^{J} u_s^2 = 1,$$

$$X^T D X u = X^T D X \sum_{s=1}^{J} u_s e_s = \sum_{s=1}^{J} \lambda_s u_s e_s$$

L'inertie projetée sur u s'écrit donc :

$$u^T X^T D X u = \sum_{s=1}^{J} \lambda_s u_s^2 \leq \lambda_1 \sum_{s=1}^{J} u_s^2 = \lambda_1.$$

Ainsi, avec la contrainte $\sum_{s=1}^{J} u_s^2 = 1$, cette inertie est majorée par λ_1. Ce maximum est atteint lorsque la première composante u_1 de u vaut 1 ou -1 et que les autres sont nulles c'est-à-dire lorsque $u = \pm e_1$. l'inertie du nuage projeté sur un axe est donc maximum lorsque cet axe est colinéaire aux vecteurs propres de $X^T D X$ associés à sa plus grande valeur propre λ_1. Elle vaut alors λ_1.

Les vecteurs propres de la matrice symétrique $X^T D X$ étant orthogonaux deux à deux, le même raisonnement montre que la direction orthogonale à u_1 qui maximise l'inertie du nuage projeté est celle d'un vecteur propres associé à la deuxième valeur propre λ_2 de $X^T D X$; cette inertie vaut alors λ_2. La suite d'axes orthogonaux maximisant l'inertie projetée est donc définie par la suite de vecteurs propres de $X^T D X$ rangés par valeurs propres décroissantes(les valeurs propres snt supposées distinctes ce qui est toujours le cas en pratique).

6.2.5 Calcul des axes d'inertie maximum pour une métrique quelconque

Si la métrique M n'est pas la métrique identique, le raisonnement précédent s'applique sans changement majeur. En effet, la matrice $X^T DXM$ définit un endomorphisme de $\mathbb{R}^J$ symétrique pour la métrique M. Rappelons que la M-symétrie d'un endomorphisme A est définie par l'égalité, pour tout couple de vecteurs u et v, des deux expressions :

$$\langle u, Av\rangle_M = \langle Au, v\rangle_M.$$

Matriciellement : $A^T M = MA$; on trouve la notion usuelle de matrice symétrique si M est la matrice identité. Cette égalité est vérifiée pour $X^T DXM$:

$$\langle u, X^T DXMv\rangle_M = u^T MX^T DXMv = \langle X^T DXMu, v\rangle_M.$$

L'endomorphisme $X^T DXM$, étant M-symétrique, est diagonalisable et admet une base M-orthonormée de vecteurs propres.
Nus utiliserons le même principe qu'avant : Soient $\lambda_1, \ldots, \lambda_s, \ldots, \lambda_J$ les valeurs propres de la matrice $X^T DXM$ rangées par ordre décroissant et $\{e_s;\ s = 1, \ldots, J\}$ une base M-orthonormée de vecteurs propres associés (i.e. $X^T DXMe_s = \lambda_s e_s$). Décomposons le vecteur u sur cette base. On a

$$u = \sum_{s=1}^{J} u_s e_s \quad \text{avec} \quad \langle u, u\rangle_M = 1,$$

$$X^T DXMu = X^T DXM \sum_{s=1}^{J} u_s e_s = \sum_{s=1}^{J} \lambda_s u_s e_s$$

L'inertie projetée sur u s'écrit donc :

$$\langle X^T DXMu, u\rangle_M = u^T MX^T DXMu = \sum_{s=1}^{J} \lambda_s u_s^2 \langle e_s, e_s\rangle_M \leq \lambda_1 \sum_{s=1}^{J} u_s^2 = \lambda_1.$$

Ainsi, avec la contrainte $\sum_{s=1}^{J} u_s^2 = 1$, cette inertie est majorée par λ_1.

6.2.6 Calcul des facteurs et de leur inertie

Notons F_s le facteur de rang s défini par la projection du nuage sur le s^e-axe d(inertie. Pour calculer les facteurs F_s, on peut diagonaliser la matrice $X^T DXM$, calculer une suite de vecteurs propres M-normés u_s associés aux valeurs propres λ_s, et appliquer aux vecteurs u_s la matrice XM, soit $F_s = XMu_s$.

Il est possible aussi d'obtenir directement les facteurs F_s et leur inertie en diagonalisant la matrice $XMX^T D$ de taille $I \times I$. En effet, les égalités ci-dessous montrent que si u_s est un vecteur propre de $X^T DXM$ associé à λ_s, alors $F_s = XMu_s$ est un vecteur propre de $XMX^T D$ associé à la même valeur propre λ_s :

$$\begin{aligned} &X^T DXMu_s = \lambda_s, \\ &(XM)(X^T DXMu_s) = \lambda_s (XM)u_s, \\ &XMX^T DF_s = \lambda_s F_s. \end{aligned}$$

L'inertie du nuage N_I projetée sur u_s est la somme des carrés des termes de F_s pondérés par les poids des éléments i soit :

$$\sum_{i=1}^{I} p_i (F_s(i))^2 = F_s^T DF_s = \lambda_s.$$

6.2.7 Définition du nuage des colonnes de X

Le nuage des colonnes N_J comprend J points situés dans un espace de dimension I, noté $\mathbb{R}^I$. Les coordonnées x_{ij} de ces points sont contenues dans les colonnes de X (qui sont d'ailleurs les lignes de la transposée X^T).
Pour qu'il y ait dualité entre le nuage des lignes N_I et le nuage des colonnes N_J, il nécessaire que ces deux nuages représentent la même information et soient construits de façon symétrique.
Tout d'abord, il est logique d'affecter à chaque colonne j le poids m_j (terme général de la matrice M déjà interprété comme un poids de colonne ; rappelons que nous nous sommes limités aux métriques associées à une matrice diaagonale).
Ainsi, le choix des poids des éléments du nuage N_J et le choix de la métrique dans $\mathbb{R}^J$ sont liés.

En outre, la construction symétrique des deux nuages implique que le poids des individus du nuage N_I induise la métrique dont $\mathbb{R}^I$ est muni. De façon directe, on peut remarquer qu'il revient au même de dupliquer un élément i ou de doubler son poids.
Dans $\mathbb{R}^I$, la distance entre deux points est la même dans ces deux cas à condition d'adopter la métrique D.
Le tableau.6.1 résume les poids et les métriques mis en jeu. Les nuages N_I et N_J ainsi construits sont dits duaux en ce sens qu'ils représentent tous deux les mêmes données $\{X, M, D\}$.

TABLE 6.1 – Les deux nuages duaux

	Espace	Métrique	Poids	Coordonnées du point k
Nuage des lignes N_I	$\mathbb{R}^J$	M	D	k^e-ligne de X
Nuage des colonnes N_J	$\mathbb{R}^I$	D	M	k^e-ligne de X^T

6.3 Nuages des lignes et des colonnes en ACP et en AFC

Le cadre général choisi pour démontrer les principaux résultats d'analyse factorielle suppose que l'on peut définir de manière totalement symétrique, à partir du tiplet $\{X, M, D\}$, le nuage des lignes et celui des colonnes. En Analyse des Correspondances, comme en Analyse en Composantes Principales, il est possible de calculer des matrices $\{X, M, D\}$ permettant cette construction symétrique. Nous en faisons une description dans ce qui suit :

6.3.1 Matrices X, M, D en ACP

En Analyse en Composantes Principales, on a

- La matrice X est le tableau des données centrés et généralement réduites.
- Dans certains cas assez rares, on souhaite conserver l'échelle de chaque variable : la matrice X est alors la matrice des variables centrées non réduites.
- La matrice diagonale D contient les poids des individus. Dans la plupart des cas, tous les individus ont le même poids $= \frac{1}{I}$ mais il possible de leur affecter des poids différents. Notns que les poids p_i des individus doivent avoir pour somme 1 afin que le "cosinus" dans $\mathbb{R}^I$ traduise exactement la corrélation.

Les variables ont presque toujours un poids égal à 1, mais il est possible, là encore, de modifier ces poids pour moduler l'influence respective des variables.

Si l'on ne centre pas les variables, l'analyse factorielle est techniquement possible : ses résultats s'interprètent alors comme les projections duales du nuage des lignes et du nuage des colonnes mais il ne s'agit plus alors véritablement d'une ACP en ce sens qu'elle n'a pas les mêmes propriétés. Ainsi, c'est le centrage qui permet d'interpréter les axes factoriels de $\mathbb{R}^J$ comme les directions de plus grande variablité de N_I ; en l'absence de centrage, ces axes sont influencés non seulement par la forme du nuage N_I mais aussi par sa position par rapport à l'origine. Par ailleurs, le centrage parmet d'interpréter le cosinus de l'angle entre deux vecteurs représentant deux colonnesdans $\mathbb{R}^I$ comme un coefficient de corrélation.

Temarquons que la matrice $X^T DX$ est, dans le cas des données centrées-réduites, la matrice des corrélations (et la matrice des covariances lorsque les données sont seulement centrées).
Le calcul des axes factoriels ne dépendant des données X qu'au travers de cette matrice, il apparaît clairement ici que ces axes ne dépendent que des liaisons linéaires entre variables.

6.3.2 Matrices X, M, D en AFC

La représentation de l'AFC met l'accent sur l'analyse des nuages des profils des lignes et des colonnes du tableau des données. Ainsi, les deux matrices qui contiennent les coordonnées des profils des lignes et des colonnes du tableau de données correspondent à deux transformations différentes de ce tableau et, d'autre part, les métriques employées dans l'analyse d'un nuage ne sont pas les poids de l'autre nuage mais leur inverse. Cela laisserait à penser que l'AFC n'entre pas dans le carde général défini au début de ce chapitre. Nous intrduisons ici une autre définition de l'AFC avec des matrices X, M et D qui respectent toutes les conditions.

a) **Une autre définition de l'AFC** : Le terme générale de la matrice X s'écrit :

$$x_{ij} = \frac{f_{ij}}{f_{i.}f_{.j}} - 1 = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}.$$

Cette matrice contient les écartss (rapprtés au produit $f_{i.}f_{.j}$) entre le tableau des données f_{ij} et le tableau de terme général $f_{i.}f_{.j}$ qui correspond à l'hypothèse d'indépendance.
Les matrices M et D sont ddiagonales de coefficients $f_{.j}$ et $f_{i.}$ respectivement. Les poids des lignes sont donc égaux aux $f_{i.}$ et ceux des colonnes sont égaux aux $f_{.j}$.

b) **Equivalence entre les deux définitions** : Pour montrer qu'avec ces matrices on obtient les résultat de l'AFC présenté précédement, il afaut que les nuages de lignes et de colonnes obtenus par les deux approches sont isomorphes. Le nuage des lignes de X est, comme le nuage des profils-lignes du tableau de données, situé dans un espace de dimension J. Les coordonnées des points sont différentes et les deux espaces ne sont pas munis de la même métrique. L'un est muni de la métrique M et l'autre de la métrique de χ^2 qui n'est autre que l'inverse M^{-1} de M.

On peut véifier directement que les distances entre les couples de points homologues sont les même. Mais cette égalité découle d'un isomorphisme induit par M que l'on peut utiliser dans toute analyse factorielle et qui a une signification intéressante en AFC. En effet, la métrique M de l'espace $\mathbb{R}^J$ définit un isomorphisme de $\mathbb{R}^J$ dans son dual, noté $(\mathbb{R}^J)^*$. Si l'on munit $(\mathbb{R}^J)^*$ de la métrique M^{-1}, l'application M est un isomorphisme d'espaces euclidiens : les distance et les formes sont conservées grâce à :

$$\langle u, v\rangle_M = \langle Mu, Mv\rangle_{M^{-1}}.$$

Le dual, noté E^*, d'un espace vectoriel E est l'espace des formes linéaires : $f : E \to \mathbb{R}$. Projeter, dans E au sens de la métrique M, le vecteur v sur u revient à appliquer à v la forme linéaire Mu.

Cette forme linéaire est l'élément de E^* associe à chaque u de E par l'application M. Le schéma suivant résume les relations entre un espace euclidien et son dual

$$\begin{array}{ccc} (R^J, M) & \leftrightarrows_{M}^{M^{-1}} & ((R^J)^*, M^{-1}) \\ u & & Mu \end{array}$$

Or le nuage des profils-lignes dans $(\mathbb{R}^J)^*$, noté ici N_I^*, est l'image, par cet isomorphisme M, du nuage N_I défini dans la section précédente. En effet, si l'on applique M au point i de N_I, sa j^e coordonnée devient :

$$x_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \xrightarrow{M} x_{ij}^* = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}} = \frac{f_{ij}}{f_{i.}} - f_{.j}.$$

On retrouve la coordonnée, après centrage, du point i dans le nuage des lignes de l'AFC du chapitre précédent. L'AFC présenté dans ce chapitre est l'analys factorielle de N_I dans (R^J, M). L'AFC présenté dans le chapitre précédent est celle de N_I^* dans $((R^J)^*, M^{-1})$.

L'isomorphisme entre ces deux nuages assure la même décomposition sur les axes d'inertie, donc l'égalité des facteurs de rangs homologues. Notons que les axes d'inertie sont situés dans des espaces différents et, par conséquent, sont différents.

6.4 Dualité

6.4.1 Relations entre les axes d'inertie et les facteurs de deux nuages

Le calcul des axes d'inertie et les facteurs du nuage des colonnes est absolument identique à celui du nuage des lignes. Tous les résultats concernant le nuage des colonnes se déduisent de ceux obtenus pour le nuage des lignes, en remplaçant X par sa transposée X^T et enéchangeant les matrices M et D.

Dans l'espace $\mathbb{R}^I$, on cherche une suite de vecteurs $\{v_s,\ s = 1, \ldots, I\}$, chacun rendant maximum la quantité

$$\mathcal{J}(v_s) = v_s^T X M X^T D v_s$$

sous la double contrainte d'être unitaire (i.e. $v_s^T D v_s = 1$) et orthogonal aux vecteurs déjà trouvés (i.e. $v_s^T D v_t = 0$ pour tout $t < s$).
La solution est donnée par l'équation

$$X M X^T D v_s = \mu_s v_s$$

qui exprime que v_s est un vecteur propre unitaire de $X M X^T D$ associé à la valeur propres μ_s de rang s.
La comparaison de cette équation avec l'équation aux facteurs des sections précédentes (i.e. $X M X^T D F_s = \lambda_s F_s$) conduit aux deux résultats suivants :

1. $\mu_s = \lambda_s$: les inerties projetées des nuuages N_I et N_J sur leurs axes principaux de même rang sont identiques. Ces valeurs propres étant positives ou nulles, les inerties totales des deux nuages sont égales à

   $$\sum_s \lambda_s.$$

   Lorsque les matricess $X^T D X M$ et $X M X^T D$ ne sont pas de même taille et admettent des nombres différents de valeurs propres, les valeurs propres non communes aux deux matices sont nulles.

2. Les facteurs F_s et les axes v_s sont vecteurs propres, de la même matrice $X M X^T D$, associés à la même valeur propre. Or, les équations aux vecteurs propres caractérisent ces vecteurs à la norme près(sauf en cas d'égalité de plusieurs valeurs propres, cas particulier ne se produisant jamais avec des données réelles). Le facteur F_s et l'axe v_s sont donc deux vecteurs colinéaires de $\mathbb{R}^I$.

Le vecteur v_s étant unitaire et la norme de F_s étant donnée par :

$$\|F_s\|_D^2 = u_s^T M X^T D X M u_s = \lambda_s$$

il en résulte la relation très importante :

$$v_s = \frac{1}{\sqrt{\lambda_s}} F_s.$$

Un raisonnement analogue, compaarant l'équation aux facteurs G_s sur les colonnes et l'équation aux axes u_s, conduit à la relation symétrique de la précédente :

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s.$$

Ces deux dernières relations sont illustrées dans la figure 6.2.

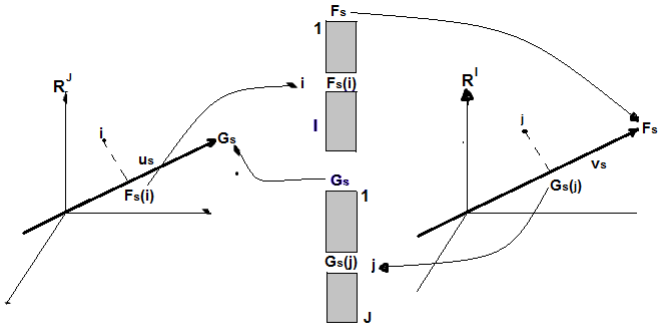

FIGURE 6.2 – Relations entre les axes d'inertie d'un nuage et les facteurs de l'autre nuage

L'ensemble des résultats est présenté schématiquement dans 6.2

TABLE 6.2 – Les deux nuages duaux, leurs axes d'inertie et leurs facteurs.

	Nuage N_I	Nuage N_J
Espace	$\mathbb{R}^J$	$\mathbb{R}^I$
Métrique	M	D
Coordonnées	X	X^T
Poids	D	M
Axe d'inertie	u_s	v_s
Equation	$X^T D X M u_s = \lambda_s u_s$	$X M X^T D v_s = \lambda_s v_s$
Norme	$\|u_s\|_M = 1$	$\|v_s\|_D = 1$
Orthogonalité	$\langle u_s, u_t \rangle_M = 0$ si $s \neq t$	$\langle v_s, v_t \rangle_M = 0$ si $s \neq t$
Facteur	$F_s = XMu_s$	$G_s = X^T D v_s$
Equation	$XMX^T D F_s = \lambda_s F_s$	$X^T D X M G_s = \lambda_s G_s$
Norme	$\|F_s\|_D = \sqrt{\lambda_s}$	$\|G_s\|_M = \sqrt{\lambda_s}$
Orthogonalité	$\sum_s p_i F_s(i) F_t(i) = 0$ si $s \neq t$	$\sum_s m_j F_s(j) F_t(j) = 0$ si $s \neq t$
Inertie sur l'axe s	λ_s	λ_s
Inertie totale	$\sum_s \lambda_s = \sum_i \sum_j p_i p_j x_{ij}^2$	$\sum_s \lambda_s = \sum_i \sum_j p_i p_j x_{ij}^2$

6.4.2 Le schéma de dualité

La méthode factorielle consiste à analyser simultanément d'une part dans $(\mathbb{R}^J, M)$ le nuage N_I affecté des poids contenus dans D et d'autre part dans $(\mathbb{R}^I, D)$ le nuage N_J affecté des poids contenus dans M.
Les matrices XM et $X^T D$ définissent des applications de $\mathbb{R}^I$ dans $\mathbb{R}^J$ et de $\mathbb{R}^J$ dans $\mathbb{R}^I$ qui lient les facteurs et les axes des deux nuages.
L'application M a déjà été considérée comme un isomorphisme de $\mathbb{R}^J$ dans son dual $\mathscr{L}(\mathbb{R}^J, \mathbb{R}) = \mathbb{R}^{J*}$. De même, D définit un isomorphisme de $\mathbb{R}^I$ dans sn dual $\mathscr{L}(\mathbb{R}^I, \mathbb{R}) = \mathbb{R}^{I*}$. L'analyse des nuages N_I et N_J est équivalente à celle de leurs images N_I^* et N_J^* par M et D.

Le matrice X définit donc une application X de $\mathbb{R}^{J*}$ dans $\mathbb{R}^I$. de façon analogue, la matrice X^T définit une application de $\mathbb{R}^{I*}$ dans $\mathbb{R}^J$. La figure.6.3, appelée **le schéma de dualité**, récapitule ces applications et les relations qui permettent de passer des axes (ou des facteurs) d'un nuage aux axes (et aux facteurs) des autres nuages.

Si, par exemple, on applique au vecteuur u_s de $\mathbb{R}^J$ successivement M, X, D et X^T, on obtient u_s au coefficient λ_s près. L'écriture de cette propriété pour n'importe quel axe principal ou n'importe quel facteur fournit l'équation qui le caractérise, c'est-à-dire la matrice dont il est vecteur propre. Ainsi, par exemple, les axes principaux u_s^* de N_I^* vérifient l'équaation :

$$MX^T D X u_s^* = \lambda_s u_s^*.$$

Appliqué à l'AFC, ce diagramme schématise bien les deux présentations de la méthode. Si l'on met en évidence les écarts à l'indépendance de chaque case du tableau de données, on analyse les nuages N_I et N_J, et la matrices des poids pour un nuage est la métrique pour l'autre. Si l'on met en évidence les profils, on analyse les nuages N_I^* et N_J^*, et la matrice des poids pour un nuage est l'inverse de la métrique pour l'autre.

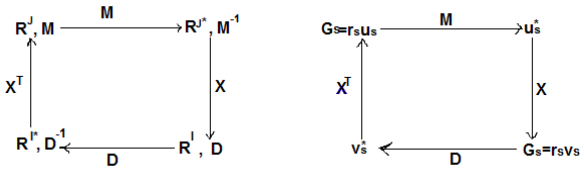

FIGURE 6.3 – Le schéma de dualité. M, D, X et X^T désignent ici les applications associées aux matrices de même noms. A gauche, les espaces en jeu et leur métriques, à droite, les résultats de l'analyse factorielle dans ces espaces. $r_s = \sqrt{\lambda_s}$.

6.4.3 Formules de transition

Unaspect de la liaison entre les analyses de chacun des deux nuages a été exprimé à l'aide des relations suivantes :

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s, \quad v_s = \frac{1}{\sqrt{\lambda_s}} F_s.$$

Elles indiquent que, dans l'espace $\mathbb{R}^I$, la représentation des colonnes (G_s) sert de base (u_s) à la représentation des lignes et réciproquement. La liaison entre les facteurs des deux nuages est donc une liaison fondamentale et il est nécessaire de les interpréter conjointement.

Les formules de transition permettent de calculer les projections de l'un des deux nuages en fonction des facteurs sur l'autre nuage. Elles décrivent directement des relations entre axes et facteurs et s'écrivent :

$$F_s = \frac{1}{\sqrt{\lambda_s}} XMG_s$$

$$G_s = \frac{1}{\sqrt{\lambda_s}} X^T DF_s.$$

Ce qui donne, point par point

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j m_j x_{ij} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i p_i x_{ij} F_s(i).$$

Ces formules montrent comment, de façon concrète, les facteurs des deux nuages doivent s'interpréter conjointement, c'est-à-dire comment chacun des ensembles peut servir de support et d'aide à l'interprétation des facteurs de l'autre ensemble.

Dans une représentation superposant les projections des lignes et des colonnes (pour les facteurs de même rang), la relation entre la position d'un élément d'un ensemble et celles de tous les éléments de l'autre ensemble peut s'exprimer ainsi :

- si x_{ij} est positif, il y a attirance entre i et j,
- si x_{ij} est négatif, il y a répulsion.

Les poids m_j et p_i pondèrent cette influence. Un élément i (resp. j) est donc situé du côté des éléments j (resp. i) pour lesquels les valeurs de x_{ij} sont les plus grandes.

Si l'on applique ces formules à l'AFC, on obtient :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}} f_{.j} G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j) - \frac{1}{\sqrt{\lambda_s}} \sum_j f_{.j} G_s(j).$$

6.5 Reconstruction des données et Approximation de X

6.5.1 Frmule d'approximation de x_{ij}

La projection d'un nuage sur ses axes d'inertie corespond à un changement de base orthonormée. En écrivant, par exemple, le vecteur x_i représentant la ligne i dans la base orthonormée des axes u_s, on obtient :

$$x_i^T = \sum_s F_s(i) u_s.$$

D'où, pour sa composante x_{ij} sur la base canonique :

$$x_{ij} = \sum_s F_s(i) u_s(j) = \sum_s \frac{F_s(i) G_s(j)}{\sqrt{\lambda_s}}.$$

Cette dernière expression, appelée formule de reconstitution des données, permet de calculer les valeurs x_{ij} en fonction des facteurs et des valeurs propres de l'analyse.

En limitant la somme à ses premiers termes, on obtient des valeurs approchées. La formule de reconstitution d'ordre S ne retient que les S premiers termes de la somme ; plus S est grand, plus l'approximation se rapproche des données initiales.

6.5.2 Interprétation dans l'espace des matrices

La formule de reconstitution des données s'écrit matriciellement :

$$X = \sum_s \frac{1}{\sqrt{\lambda_s}} F_s G_s^T = \sum_s \sqrt{\lambda_s} v_s u_s^T.$$

La matrice X est ainsi décomposée en une somme de matrices de rang 1 (le rang d'une matrice est la dimension de l'espace vectoriel engendré par ses colonnes ou par ses lignes).

Considérons l'espace des matrices de taille $I \times J$, noté $\mathbb{R}^{I \times J}$, muni de la métrique diagonale des produits $m_j p_i$. Dans cet espace, les matrices $\Lambda_s = v_s u_s^T$ (de rang 1) forment un système orthonormé et $\sigma_s = \sqrt{\lambda_s} v_s u_s^T$ est la projection de X sur $v_s u_s^T$

$$\begin{aligned}
\langle v_s u_s^T, v_t u_t^T \rangle_{m_j p_i} &= \sum_i \sum_j v_s(i) u_s(j) v_t(i) u_t(j) m_j p_i \\
&= \sum_i v_s(i) v_t(i) p_i \sum_j u_s(j) u_t(j) m_j \\
&= \begin{cases} 0 & , \text{si} \quad s \neq t, \\ 1 & , \text{si} \quad s = t. \end{cases}
\end{aligned}$$

$$\begin{aligned}
\langle X, v_t u_t^T \rangle_{m_j p_i} &= \sum_i \sum_j x_{ij} v_s(i) u_s(j) m_j p_i \\
&= \sum_i p_i v_s(i) F_s(i) \\
&= \sqrt{\lambda_s}.
\end{aligned}$$

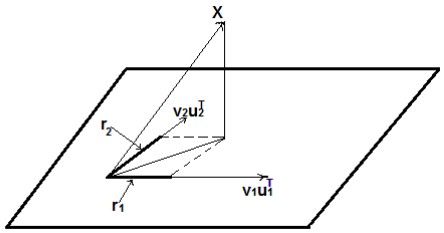

FIGURE 6.4 – Dans $\mathbb{R}^{I\times J}$, la reconstitution d'ordre 2 de X est une projection de X sur un plan.

L'analyse factorielle décompose la matrice X, en tant que vecteur da l'espace $\mathbb{R}^{I\times J}$, sur un système orthonormé de matrices de rang 1. La restriction de la formule de reconstitution des données, à ses S premiers vecteurs (voir la figure.6.4). Cette approximation est une matrice de rang S.

Le carré de la norme de la différence entre X est son approximation d'ordre S est égal à la somme des valeurs propres d'ordre supérieur à S.

On peut définir l'analyse factorielle par cette décomposition. L'objctif est alors d'approcher le tableau X avec un tableau de rang fixé S (S étant supérieur ou égal à 1 et inférieur à I et à J). On réalise l'ajustement avec le critère des moindres carrés pondérés, la case (i, j) ayant le poids $m_j p_i$.

On cherche alors une suite orthogonale de matrice de rang I, qui s'écrivent donc comme le produit d'un vecteur A_s de $\mathbb{R}^I$ et d'un vecteur B_s de $\mathbb{R}^J$, qui minimisent l'expression

$$\mathcal{J}(A_s, B_s) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{s=1}^{S} A_s(i)B_s(j)\right)^2 m_j p_i.$$

Quelques calculs, en procédant par itération sur s, permettent de vérifier que la solution unique est donnée par les premiers facteurs de l'analyse factorielle.