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

Analyse des données statistiques pour l'Ingénieur

–Cours Pour Génie Civil 1–
Analyse des données statistiques et Modélisation

Mohamed ADDAM
Professeur de Mathématiques

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

16 Mars 2020

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* $\hat{y} = -\frac{1}{7}x + \frac{3}{7}$.

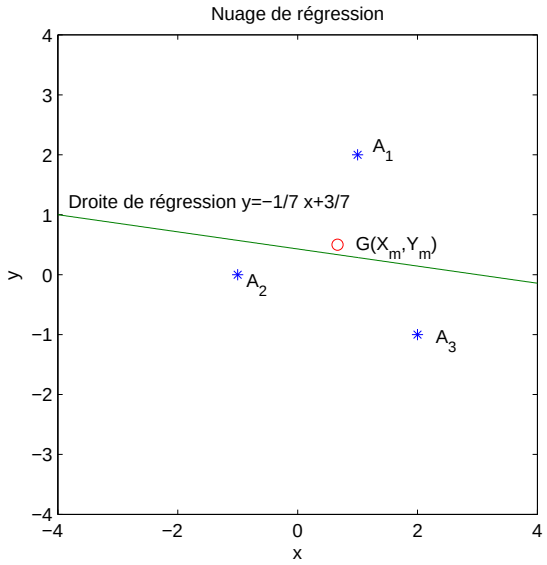

FIGURE 2.1 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$, $Y_m = \bar{Y}$ et $G(X_m, Y_m)$ est le centre de gravité de A_1, A_2 et A_3.

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

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

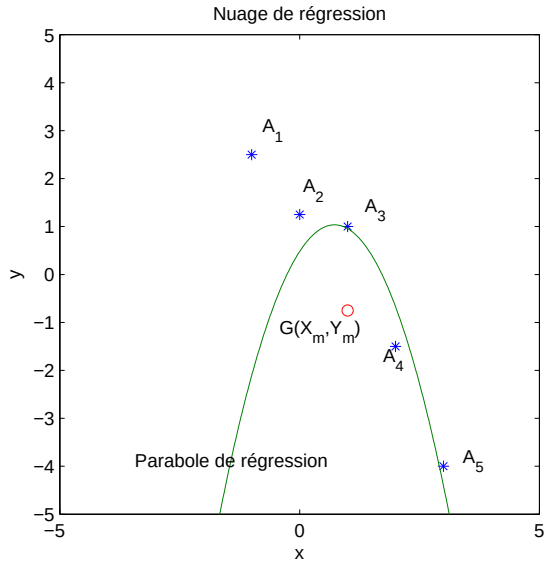

FIGURE 2.2 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$, $Y_m = \bar{Y}$ et $G(X_m, Y_m)$ est le centre de gravité de $A_1, \ldots, A_5$.

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.

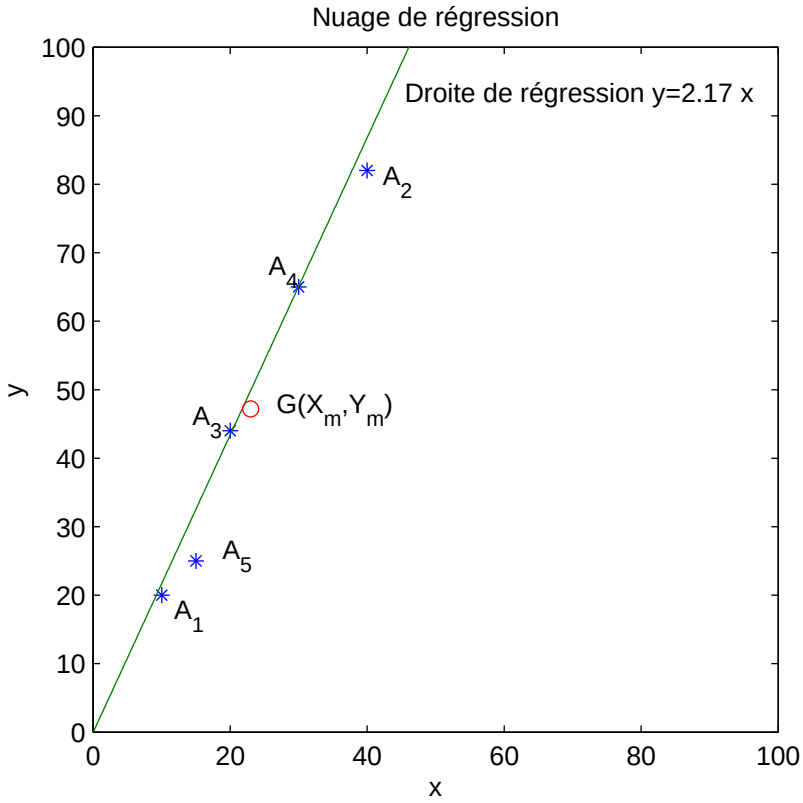

FIGURE 2.3 – Cette figure montre le nuage de régression et sa distribution dans un repère orthonormé avec $X_m = \bar{X}$ et $Y_m = \bar{Y}$

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 cottélations 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 $\mathcal{J}(a_1, a_2, a_3, b)$.

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

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.

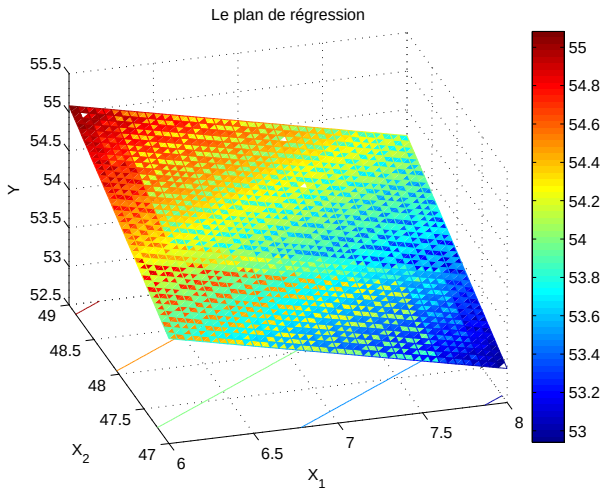

FIGURE 2.4 – Surface de tendance d'odre 1-Plan de régression.

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 $(\vec{i}, \vec{j}) = 0$.

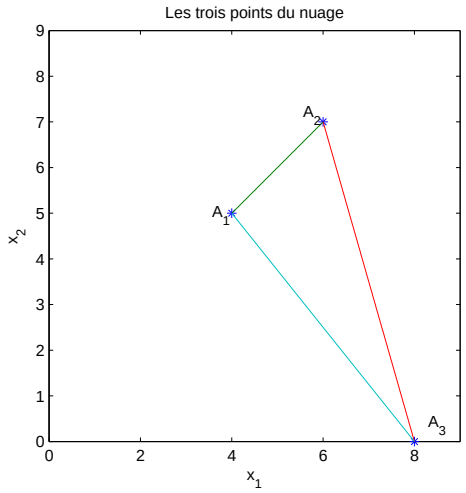

FIGURE 3.1 – Les trois points du nuage constituent l'information des lignes du Tableau (3.1). Les positions relatives de ces 3 points peuvent être calculées en utilisant la distance euclidienne.

- **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 $G(\bar{z}_1 = 0, \bar{z}_2 = 0)$)

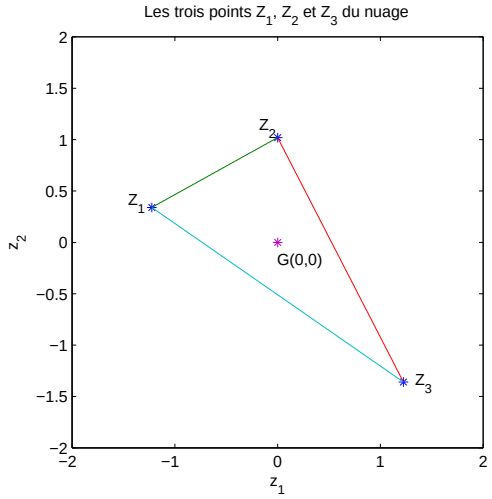

FIGURE 3.2 – Les trois points du nuage constituent l'information des lignes du Tableau (3.2). Les positions relatives de ces 3 points peuvent être calculées en utilisant la distance euclidienne.

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** $\frac{1}{n}Z^TZ$

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

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

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.*