كلية العلوم و التقنيات فاس
ⵜⴰⵖⵚⵍⵓⵏⵜ ⵏ ⵜⵇⴰⵙⵙⵉⵢⵏ ⴷ ⵜⵙⵏⵙⴰⵙⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

Licence Mathématiques et Applications

(MA)

MEMOIRE DE FIN D'ETUDES

Pour l'obtention du Diplôme de Licence Sciences et Techniques

(LST)

L'analyse en composantes principales et Plans d'expériences

(Analyse et interprétation des résultats)

- ***Réalisé par : BOUZID El houssine***
- ***Encadré par: Pr. AMMOR Ouafae***

Soutenu le 06 juin 2018

Devant le jury composé de :

- ✓ *Pr .AMMOR Ouafae*
- ✓ *Pr. RAHMOUNI HASSANI Aziza*
- ✓ *Pr. EZZAKI Fatima*

Année Universitaire 2017-2018

Remerciement

Mes plus vifs remerciements vont aux personnes qui ont contribué au bon déroulement et à l'aboutissement de ce projet de fin d'étude.

En premier lieu je tiens à exprimer ma profonde gratitude à Madame Ouafae AMMOR qui m'a donné la chance de réaliser ce travail. Je tiens ses conseils avisés et la disponibilité constante dont il a toujours fait preuve.

Je suis très honoré que Madame RAHMOUNI HASSANI Aziza et Madame EZZAKI Fatima aient accepté de lire mon travail et d'être un jury de celui-là.

J'exprime ainsi toute ma reconnaissance aux membres de département de mathématiques de la Faculté des Sciences et Techniques de Fès, surtout les professeurs que j'ai rencontré durant ces trois années.

SOMMAIRE

Introduction :

Dans les dernières décennies, et avec les progrès informatiques, les statistiques ont évolué dans différentes domaines : L'analyse des données et les plans d'expériences sont les plus influencées par ce développement informatique.

Mon rapport de stage est partitionné en deux volets :

Le premier concerne l'une des principales méthodes d'analyse des données : L'ACP et surtout l'interprétation des résultats trouvés.

Le second traite l'interprétation des résultats issus à partir de l'application des plans d'expériences, en planifiant au mieux les essais afin de gagner en temps et en argent.

Ces deux démarches expérimentales : -L'ACP et les Plans d'expérience - vont aider l'expérimentateur à structurer sa recherche de manière plus efficace, à confronter et à valider ses propres hypothèses et à mieux comprendre les phénomènes étudiés afin de trouver des solutions pour les problèmes posés et faire ainsi des modèles et des prévisions futurs.

L'objectif de ce travail est de présenter, la mise en œuvre et l'intérêt de ces 2 méthodes, en présentant leurs aspects théoriques et l'interprétation de leurs résultats à travers des exemples pratiques.

A- L'Analyse en Composantes Principales

Le principe général de L'ACP :

Définition : L'ACP (Analyse en Composantes Principales) est une méthode de traitement des données quantitatives multidimensionnelles qui poursuit les deux objectifs suivants :

- Visualiser les données.
- Réduire la dimension effective des données.

Pourquoi ?

- Problème : Lorsque on étudie simultanément un nombre important de variables quantitatives comment en faire un graphique globale ?

La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan ou bien dans un espace à 3 dimension, mais dans un espace de dimension plus importante, l'objectif de L'ACP est de revenir à un espace de dimension réduite en déformant le moins possible la réalité, il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales, afin de représenter les données sur des graphique pour atteindre une finalité importante : l'interprétation des résultats.

Les principes théoriques de L'ACP :

Les données à analyser

Les données sont les mesures effectuées sur n unités $\{u_1, u_2, \ldots\ldots\ldots u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $v_1, v_2, \ldots\ldots\ldots, v_p$, d'où l'écriture matricielle suivante du tableau de données :

$$X = \begin{array}{c} \\ u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_p \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & \ldots & v_j & \ldots & v_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix} \end{array}$$

La ligne i décrit la valeur prise par l'individu u_i pour p valeurs, alors que la colonne j décrit la valeur de la variable v_j pour n individus.

Le problème à traiter

On cherche à extraire l'information pertinente contenue dans le tableau des données. Pour cela, on va le résumer en extrayant l'essentiel de sa structure en vue de faire des représentations graphiques à la fois fidèles aux données initiales et commodes à interpréter. Ces représentations devront se faire en dimension réduite.

Comment ?

Les supports théoriques :

On cherche des combinaisons linéaires des variables initiales, appelées composantes principales, s'écrivant sous la forme suivante :

$$C^1 = a_1^1 X^1 + a_1^2 X^2 + \cdots + a_1^p X^p$$
$$C^2 = a_2^1 X^1 + a_2^2 X^2 + \cdots + a_2^p X^p$$
$$\vdots$$

Telles que

☛☛ C^1 doit contenir un maximum d'information, c'est-à-dire disperser le plus possible les individus.

☛☛R $(C^1, C^2) = 0$ la condition de perpendicularité des axes 1 et 2, c'est-à-dire que la deuxième composante principale doit contenir l'information complémentaire de la première.

☛☛La variance de C^2 doit être, à son tour, la plus grande possible. Ainsi, cette deuxième composante principale fournit la plus grande information possible complémentaire à la première.

☛☛Le processus se déroule jusqu'à l'obtention de la dernière composante principale (la $p^{ème}$), les parts d'informations expliquées par chacune d'elles devenant de plus en plus faibles.

En conclusion, la phase essentielle de L'ACP consiste à transformer ces p variables quantitatives initiales, toutes plus ou moins corrélées entre elles, en p nouvelles variables quantitatives, non corrélées, appelées composantes principales.

Remarque :

Les données sont soit considérées en tant qu'individus décrits par leurs p variables, soit en tant que variables décrites par les n individus, d'où l'importance de la considération des deux nuages de points. Nous obtenons ainsi n points dans l'espace R^p, espace des variables et p points dans l'espace R^n celui des individus.

Mais le problème est de visualiser la forme des nuages, pour ce faire l'idée est d'étudier les projections sur des droites, des plans ou plus généralement sur des sous espace de dimension réduite s < p. Il faut donc chercher le sous-espace qui ajuste au mieux le nuage de points i.e. chercher à minimiser les déformations provoquées par la projection.

Nous allons donc chercher à ajuster au mieux le nuage des individus dans l'espace des variables puis le nuage des variables dans l'espace des individus

1) Ajustement du nuage des individus dans l'espace des variables :

L'objectif est de fournir des images approchées du nuage des individus - que nous noterons Ni dans R^p l'espace des variables.

Supposons que le nuage Ni est reconstitué de manière satisfaisante dans un sous-espace de dimension s < p.

a) Dans un premier temps, cherchons un sous-espace vectoriel à une dimension, i.e. une droite d1 passant par l'origine, qui ajuste au mieux le nuage des individus Ni. Nous considérons donc le cas où s= 1. La projection sur la droite d1 qui ajuste au mieux le nuage Ni

donne la dispersion ou inertie maximale le long de la droite d1. (La notion de variance se généralise en inertie).

À la recherche de d1 :

Proposition 1 : Maximiser la dispersion le long de la droite d1 revient à minimiser les distances des points du nuage Ni à la droite d1, c'est-à-dire que la droite d1 passe au plus près de tous les points du nuage Ni.

Soit u_1 le vecteur unitaire de la droite d1 on a la proposition suivante :

Proposition 2 : maximiser la dispersion le long de la droite d1 revient à maximiser une forme quadratique définie par $u_1^t X^t X u_1$

Le problème revient donc à trouver u_1 qui maximise cette forme quadratique avec la contrainte $u_1 u_1^{\ t} = 1$. Le sous-espace à une dimension optimal au sens de l'inertie maximale est donc l'axe d1 défini par le vecteur u_1 solution de ce problème.

b) Cherchons maintenant à déterminer le sous-espace à deux dimensions s'ajustant au mieux au nuage Ni, nous considérons donc le cas où s = 2.

Proposition 3 : Le sous-espace à deux dimensions qui ajuste au mieux le nuage Ni contient u_1.

Le sous-espace à deux dimensions est donc caractérisé par l'axe d1 et l'axe d2 défini par le vecteur u_2 orthogonal à u_1 vérifiant donc :

$$\begin{cases} u_2^t X^t X u_2 \text{ est maximal} \\ u_2^t u_2 = 1 \text{ (contraite de normalité)} \\ u_2^t u_2 = 0 \text{ (contrainte d'orthogonalité)} \end{cases}$$

Par récurrence, le sous-espace à s dimensions s'ajustant au mieux au nuage Ni contient les vecteurs $u_1, u_2, \ldots\ldots\ldots, u_{s-1}$ .Ce sous-espace est engendré par le sous-espace $\{u_1, u_2, \ldots\ldots\ldots, u_{s-1}\}$ de dimension $s-1$ et le vecteur u_s orthogonal à ce sous-espace, et vérifiant :

$$\begin{cases} u_s^t X^t X u_s \text{ est maximal} \\ u_s^t u_s = 1 \end{cases}$$

Proposition 4 : Une base orthonormée du sous-espace vectoriel de dimension s, s'ajustant au mieux au nuage Ni dans l'espace des variables, est constituée par les s vecteurs propres $u_1, u_2, \ldots\ldots\ldots, u_s$ correspondant aux s plus grandes valeurs propres de la matrice $X^t X$.

C'est la proposition fondamentale de l'ACP!

Définition : les s vecteur ainsi obtenu déterminent des axes qui s'appellent les axes factoriels ou les facteurs principaux.

2) Ajustement du nuage des variables dans l'espace des individus :

De la même façon que pour le nuage des individus Ni, nous cherchons une image du nuage des variables - que nous noterons Nv dans l'espace des individus. L'approche est identique à celle du nuage des individus, il suffit simplement de considérer la matrice XX^t au lieu de $X^t X$. Ainsi comme dans le premier cas, l'axe factoriel (ou axe d'inertie) est déterminé par v_s vérifiant :

$$\begin{cases} v_s^t X X^t v_s \text{ est maximal} \\ v_s^t v_s = 1 \text{ (contrainte de normalité)} \\ v_s^t v_q = 0 \text{ pour tout } q = \{1, 2, \ldots, s-1\} \text{ (orthogonalité)} \end{cases}$$

Le sous-espace d'ajustement est obtenu de la même manière que dans le cas des individus, par la proposition suivante :

Proposition 5 : Une base orthonormée du sous-espace vectoriel de dimension s, s'ajustant au mieux au nuage Nv dans l'espace des individus est constituée par les s vecteurs propres $(v_1, v_2, \ldots\ldots\ldots, v_S)$ correspondant aux s plus grandes valeurs propres de la matrice XX'.

Définition : Les s nouvelles variables $v_1, v_2, \ldots\ldots\ldots, v_S$ sont appelées composantes principales, c'est celles qui résument donc l'ensemble des variables initiales du tableau X.

Remarque :

Mathématiquement, on peut mettre en évidence des relations, dites relations de transition, entre les ajustements dans les deux espaces, ces relations montrent que les deux nuages doivent s'analyser et s'interpréter simultanément.

Mais après avoir visualiser les données brutes dans des espaces de dimension réduite par projection, comment peut-on théoriquement retrouver les données initiales ?

Bien entendue, il est possible de reconstruire de manière exacte le tableau de données X par une décomposition en valeurs singulières de la matrice X. En effet, puisque u_s est le sème vecteur propre de norme 1 de la matrice X'X, correspondant à la sème valeur propre, et v_s est le sème vecteur propre de norme 1 de la matrice XX' correspondant à la même valeur propre, nous avons :

$$X = \sum_{i=1}^{p} \sqrt{\lambda_i} v_i u_i^T$$

S'ajuster au mieux signifie donc reconstituer au mieux les positions des points des nuages par un nouvel ensemble de coordonnées.

Problème :

La difficulté majeur réside dans le choix de s, c'est-à-dire à partir de quelle valeur a-t-on une bonne reconstruction, ou encore une bonne proportion de la trace de $X^t X$?

Indice de qualité de la reconstruction :

La qualité globale de la reconstruction peut être mesurée par :

$$\tau = \frac{\sum_{i=1}^{s} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Ce coefficient est appelé taux d'inertie ou pourcentage de la variance relatif aux s premiers facteurs.

Un repère formé par les s premiers axes factoriels permet de reconstituer les positions de départ avec une bonne précision, si la somme des s valeurs propres associées représente une bonne proportion de la trace de la matrice X'X.

Nous obtenons ainsi une reconstruction approchée du tableau X en se limitant aux s premiers axes factoriels.

Récapitulatif

Etant donnée un tableau de données de taille (p, q) que l'on représente par une matrice X.

Il est souhaitable de centrer et réduire les variables i.e. normer la matrice X en colonne, soit Z la matrice centrée réduite obtenue :

$$Z = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{1q}-\bar{x}_q}{\sigma_q} \\ \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{2q}-\bar{x}_q}{\sigma_q} \\ \vdots & & \ddots & \vdots \\ \frac{x_{p1}-\bar{x}_1}{\sigma_1} & \frac{x_{p2}-\bar{x}_2}{\sigma_2} & \cdots & \frac{x_{pq}-\bar{x}_q}{\sigma_q} \end{pmatrix}$$

A partir de cette matrice, on définit la matrice R des corrélations entre les q variables prises deux à deux :

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1q} \\ \rho_{21} & 1 & \cdots & \rho_{2q} \\ \vdots & & \ddots & \vdots \\ \rho_{q1} & \cdots & \cdots & 1 \end{pmatrix}$$

R résume la structure des dépendances linéaires entre les q variables et on a

$$R = \frac{1}{p} Z^T Z$$

On extrait les valeurs propres les plus grandes de la matrice R des corrélations.

Les valeurs propres représentent les variances des individus sur les axes correspondants.

En pratique, on arrête l'extraction des valeurs propres lorsque la somme des s valeurs propres que l'on a déterminés représente un pourcentage satisfaisant de la variance (pourcentage d'inertie).

On détermine les vecteurs propres associés aux valeurs propres, ce sont les axes factoriels.

Les vecteurs propres permettent le calcul des composantes principales, et donc le calcul des coordonnées des variables et des individus sur les nouveaux axes principaux.

Nombre d'axes à retenir :

Les critères les plus utilisés sont les suivants :

a- Critère de Kaiser (variables centrées et réduites) : On ne retient que les axes associés aux valeurs propres supérieurs à 1, c'est-à-dire dont la variance est supérieure à celle des variables d'origine.

• Et plus généralement, on ne garde que les valeurs propres supérieures à leur moyenne.

b- Éboulis des valeurs propres : on cherche un «coude », une cassure dans le graphe des valeurs propres et on ne conserve que les valeurs jusqu'à ce « coude ».

5) Nouvelles coordonnées des individus

On cherche à calculer les coordonnées des individus sur les nouveaux axes, les vecteurs propres obtenus précédemment permettent de calculer les nouvelles coordonnées des variables dans le repère des composantes principales. Ces nouvelles coordonnées sont obtenues en multipliant la matrice des données centrées réduites Z avec la matrice des vecteurs propres P

6) Nouvelles coordonnées des variables

Pour calculer les nouvelles coordonnées des variables, il faut multiplier chaque vecteur propre par la racine carrée de la valeur propre associée.

7) On projet les points de nuage sur le plan ou l'espace formé par les premiers axes factoriels

8) Interprétation des projections

Cette projection entraîne une déformation du nuage de points, d'où l'importance d'étudier la qualité de représentation des points individus.

Pour connaître la qualité de représentation d'un individu, on postule les règles énoncées dans le tableau ci-contre :

Valeur de $\cos^2(\alpha)$	Signification
≈ 1	Très bonne représentation
≥ 0.5	Représentation acceptable
<0.5	Mauvaise représentation

Avec α l'angle entre le vecteur individu, et la composante principale

Le nuage des individus dans l'espace des variables est formé de p points.

Les points représentés dans l'espace factoriel définie par les premiers axes sont les projections des individus.

Une proximité entre les projections de deux points individus s'interprète comme un comportement analogue des q variables associées

On représente les q points variables sur le même graphique, les coordonnées des points variables s'interprète en terme de corrélation.

Le coefficient de corrélation entre une variable u et un axe F étant le cosinus de l'angle formé par u et F.

Une proximité entre deux points variables signifie que les deux variables correspondantes sont corrélées, la corrélation est encore plus significative lorsque les points représentatifs de ces deux variables sont éloignés de l'origine.

Une proximité entre un point variable et un point individu signifie que la variable joue un rôle important pour l'individu considéré.

On cherche si c'était possible une interprétation des axes principaux.

Exemple théorique illustratif :

Faire une ACP normée de la matrice $X = \begin{pmatrix} 3 & 4 \\ 7 & 2 \\ 5 & 6 \\ 5 & 4 \end{pmatrix}$

On a $X = \overbrace{\begin{pmatrix} 3 & 4 \\ 7 & 2 \\ 5 & 6 \\ 5 & 4 \end{pmatrix}}^{les\ variables}$ $\left.\begin{matrix}\\\\\\\\\end{matrix}\right\}$ *les individus*

Centrons et réduisons alors cette matrice en colonne :

Pour se faire posons : $C_1 = \begin{pmatrix} 3 \\ 7 \\ 5 \\ 5 \end{pmatrix}$ *et* $C_2 = \begin{pmatrix} 4 \\ 2 \\ 6 \\ 4 \end{pmatrix}$

Calculons la moyenne de chaque colonne : $\bar{x}_j = \frac{1}{4}\sum_{i=1}^{4} x_{ij}$

On obtient alors $\bar{x}_1 = 5$ *et* $\bar{x}_2 = 4$ la matrice centrée est alors : $\bar{X} = \begin{pmatrix} -2 & 0 \\ 2 & -2 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}$

Calculons la variance de chaque colonne : on a $\text{var}(C_j) = \frac{1}{4}\sum_{i=1}^{4}(x_{ij} - \bar{x}_j)^2 = \frac{1}{4}\|\overline{C}_j\|$

Avec les $\overline{C}_j$ sont les colonnes de la matrice centrée $\overline{X}$, on a alors :

$$\sigma(C_1) = \sqrt{\frac{1}{4}\|\overline{C}_1\|} = \sqrt{\frac{1}{4}\left[(-2)^2 + 2^2 + 0^2 + 0^2\right]} = \sqrt{2}$$

De la même façon on obtient $\sigma(C_2) = \sqrt{\frac{1}{4}\|\overline{C}_2\|} = \sqrt{2}$

Ainsi on obtient la matrice centrée réduite $Z = \begin{pmatrix} -\sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}$

La matrice des corrélations R s'obtient par la formule suivant : $R = \frac{1}{4}Z^t Z$

D'où $R = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$

Cherchons maintenant les valeurs et les vecteurs propres de cette matrice :

L'équation caractéristique es donnée par : $\det(R - xI_2) = \begin{vmatrix} 1-x & -\frac{1}{2} \\ -\frac{1}{2} & 1-x \end{vmatrix} = (x - \frac{3}{2})(x - \frac{1}{2})$

Les deux valeurs propres sont alors $\lambda_1 = \frac{3}{2}$ et $\lambda_2 = \frac{1}{2}$ (on les classe de plus grande au plus petite)

Les sous espaces propres :

$$E_{\lambda_1} = E_{\frac{3}{2}} = \ker(R - \frac{3}{2}I_2) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in IR^2 \ tq \ \ R\begin{pmatrix} x \\ y \end{pmatrix} = \frac{3}{2}\begin{pmatrix} x \\ y \end{pmatrix} \right\}$$

Soit $u = \begin{pmatrix} x \\ y \end{pmatrix} \in IR^2$ on a :

$$u \in \ker(R - \frac{3}{2} I_2) \Leftrightarrow \left[\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} - \frac{3}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Leftrightarrow -\frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Leftrightarrow y = -x$$

Donc $E_{\frac{3}{2}} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in IR^2 \ tq \quad y = -x \right\} = \left\langle \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\rangle$

Un vecteur unitaire de ce sous espace est $\begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$

De façon similaire on obtient : $E_{\frac{1}{2}} == vect\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$

Un vecteur unitaire de ce sous espace est $\begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$

Les deux vecteurs $\begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$ *et* $\begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$ sont les composantes principales, elles forment alors le nouveau repère dans lequel on va visualiser les données.

Les nouvelles coordonnées des individus :

S'obtiennent en multipliant la matrice des données centrées réduites par la matrice des vecteurs propres, donc

$$M_{\substack{nouvelles coordonnées \\ des individus}} = \begin{pmatrix} -\sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{matrix} \text{les abscisses} & \text{les ordonnées} \\ \begin{pmatrix} -1 & -1 \\ 2 & 0 \\ -1 & 1 \\ 0 & 0 \end{pmatrix} \end{matrix}$$

Les nouvelles coordonnées des variables :

S'obtiennent en multipliant chaque vecteur propre par la racine carrée de la valeur propre associée

Les nouvelles coordonnées de la première variable sont $\begin{pmatrix} x \\ y \end{pmatrix} = \sqrt{\frac{3}{2}} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}$

Les nouvelles coordonnées de la deuxième variable sont $\begin{pmatrix} x \\ y \end{pmatrix} = \sqrt{\frac{1}{2}} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$

Toutes ces coordonnées nous permettront de représenter les résultats graphiquement, de les analyser et de les interpréter.

Qualité de représentation des individus :

Se calcule par : $qlt_{axe\,k}(individu\ \ i) = \cos^2(\alpha_{ik}) = \frac{c_{ik}^2}{\sum_{j=1}^{p} c_{ij}^2}$

Avec p le nombre de colonnes de la matrice X c'est aussi le nombre de composantes principales et les c_{ij} sont les nouvelles coordonnées des individus dans les composantes principales

Ici on a deux composantes principales, les nouvelles coordonnées sont

$$M_{\substack{nouvelles coordonnées \\ des individus}} = \begin{array}{c} \begin{matrix} \text{les abscisses} & \text{les ordonnées} \end{matrix} \\ \begin{pmatrix} -1 & -1 \\ 2 & 0 \\ -1 & 1 \\ 0 & 0 \end{pmatrix} \end{array}$$

Donc

$$qlt_{axe1}(individu\ 1) = \cos^2(\alpha_{11}) = \frac{c_{11}^2}{\sum_{j=1}^{2} c_{1j}^2} = \frac{(-1)^2}{(-1)^2 + (-1)^2} = \frac{1}{2}$$

$$qlt_{axe2}(individu\ 3) = \cos^2(\alpha_{32}) = \frac{c_{32}^2}{\sum_{j=1}^{2} c_{3j}^2} = \frac{1^2}{(-1)^2 + 1^2} = \frac{1}{2}$$

Toutes les autres qualités se calculent de la même manière

Contribution des individus à la formation des axes :

La contribution d'un individu i à la formation d'un axe k est donnée par : $CTR_k(i) = \frac{c_{ik}^2}{n\lambda_k}$

$$CTR_1(1) = \frac{(-1)^2}{4 \times \frac{3}{2}} = \frac{1}{6} \qquad CTR_2(1) = \frac{(-1)^2}{4 \times \frac{1}{2}} = \frac{1}{2}$$

$$CTR_1(2) = \frac{2^2}{4 \times \frac{3}{2}} = \frac{2}{3} \qquad CTR_2(2) = \frac{0^2}{4 \times \frac{3}{2}} = 0$$

$$CTR_1(3) = \frac{(-1)^2}{4 \times \frac{3}{2}} = \frac{1}{6} \qquad CTR_2(3) = \frac{1^2}{4 \times \frac{1}{2}} = \frac{1}{2}$$

$$CTR_1(4) = \frac{0^2}{4 \times \frac{3}{2}} = 0 \qquad CTR_2(4) = \frac{0^2}{4 \times \frac{3}{2}} = 0$$

La somme des contributions vaut 1

Problème :

en réalité les études expérimentales recensent des données de grandes taille qu'on ne peut pas les traiter manuellement, il' est dur d'effectuer à la main le calcul des valeurs propres d'une matrice d'ordre élevé !

Le développement informatique prend alors la relève !

De la théorie à la pratique :

L'ACP en pratique

En pratique pour réaliser une ACP on suit une démarche en plusieurs étapes :

1) Préparation des données

- S'assurer que les données sont quantitatives.

- Données manquantes : L'ACP ne sait pas traiter les données manquantes. Certains logiciels proposent de supprimer les individus possédant des données manquantes, alors que d'autres vont remplacer la donnée manquante par un zéro.

2) Réaliser les calculs

On a vu les fondements théoriques de calcul. Vu la taille du tableau de données que l'on traite, c'est le logiciel qui réalisera cette étape.

Le logiciel produit alors différents tableaux et graphiques qu'il faudra interpréter.

3) Interpréter les résultats :

a) Déterminer le nombre d'axes de l'analyse

Pour répondre à cette question, il faut consulter le tableau des valeurs propres qui accompagne L'ACP. Les valeurs propres sont classées de façon décroissante.

Il y a deux manières célèbres pour déterminer le nombre d'axes à prendre en compte : l'éboulis des valeurs propres et le critère de Kaiser.

Il est important que les valeurs propres des axes retenus restituent une bonne proportion de la variance.

b) Sélectionner les individus et variables à interpréter

Les graphiques de L'ACP sont les projections des variables et des individus sur un plan factoriel déterminé. On commencera par interpréter le premier plan factoriel (celui formé par les facteurs C^1 *et* C^2) car c'est celui qui concentre la plus grande partie de l'information du nuage.

Sur un plan factoriel, on n'interprète que les variables et les individus qui sont bien représentés. Pour les individus, on utilisera les contributions absolues et relatives alors que pour les variables, on n'interprètera que celles qui sont proches du cercle de corrélation.

c) les sorties graphiques

Deux graphiques sont données par les logiciels : celui des variables et celui des individus

La représentation des variables.

Ce graphique se distingue par la présence d'un cercle de corrélation. On interprète deux types de positions :

1-Les positions des variables par rapport aux axes afin de déterminer quelles sont les variables qui font les axes.

2-Les positions des variables les unes par rapport aux autres. Le coefficient de corrélation entre deux variables étant le cosinus de l'angle formé par les vecteurs correspondants on en déduit que :

- deux variables qui sont proches ou confondus sont corrélées positivement (coefficient de corrélation proche de 1),

- deux variables opposées (formant un angle de π) sont corrélées négativement (coefficient de corrélation proche de -1)

- deux variables positionnées à angle droit (angle de $\frac{\pi}{2}$) ne sont pas du tout corrélées (coefficient de corrélation égal à 0)

La représentation des individus

Deux cas se présentent :

L'ACP est réalisé sur un tableau comportant beaucoup d'individus. Dans ce cas, on ne pourra pas interpréter les positions relatives de tous les individus car le nuage sera tellement dense.

Toutefois, si un individu est atypique, il va ressortir du nuage et on pourra alors l'identifier pour éventuellement le supprimer et effectuer un nouveau passage sans cet individu. Dans ce cas, on a souvent recours à une méthode classification automatique afin de regrouper les individus qui sont proches les uns des autres et ainsi de constituer des type d'individus ayant un comportement similaire.

Sous réserve d'une bonne représentation, la proximité de deux individus sur un plan factoriel est synonyme d'individus ayant un comportement similaire. Si deux individus ont exactement les mêmes valeurs aux différentes variables, ils seront superposés sur les différents plans factoriels. De même, des individus ou des groupes d'individus s'opposant par rapport à un axe factoriel, s'opposeront par rapport aux variables qui font cet axe.

Exemples pratiques :

Exemple 1 :

Dans cet exemple on va traiter les notes de 15 étudiants d'une classe en tronc commun en huit matières : l'arabe, le français, l'anglais, la philosophie, les maths, la physique et la chimie, les sciences de la vie et de la terre et l'éducation sportive

Les résultats sont regroupés dans le tableau ci-dessous :

N	Arabe	Français	Anglais	Philosophie	Math	PC	SVT	ES
1	16	14.5	14.75	13.5	10	11.75	11	16.5
2	11.5	14	13	10	16	17.5	15.75	15.5
3	12	11.75	14	11.25	14.75	16	16.5	16
4	13.5	12	15	12.5	15	14.5	14.75	15
5	15.75	16.5	15	14.25	9.75	11	11.25	14.75
6	11.75	10.75	13	12.75	15	14.5	16	15.5
7	14.5	13.75	16	15.5	15	14	13.75	13
8	12.5	10	11	13	13.5	12.75	15	14
9	13.5	15	12.75	10	12.5	11.75	10	15
10	17	14.25	16	15.75	11.75	13	12.5	16.75
11	15.5	16	14.75	13.25	12	12.5	12.75	13.75
12	13.75	16	17.5	13.5	16.75	17	16.5	15
13	14	11.75	14.5	12.5	13	11.75	14	16.25
14	10.5	9.5	11.75	13	11	12.5	10	17
15	15.5	13.25	14	14	13.75	15	14.5	14

Ces données sont toutes quantitatifs on les traite alors par une ACP à l'aide de R

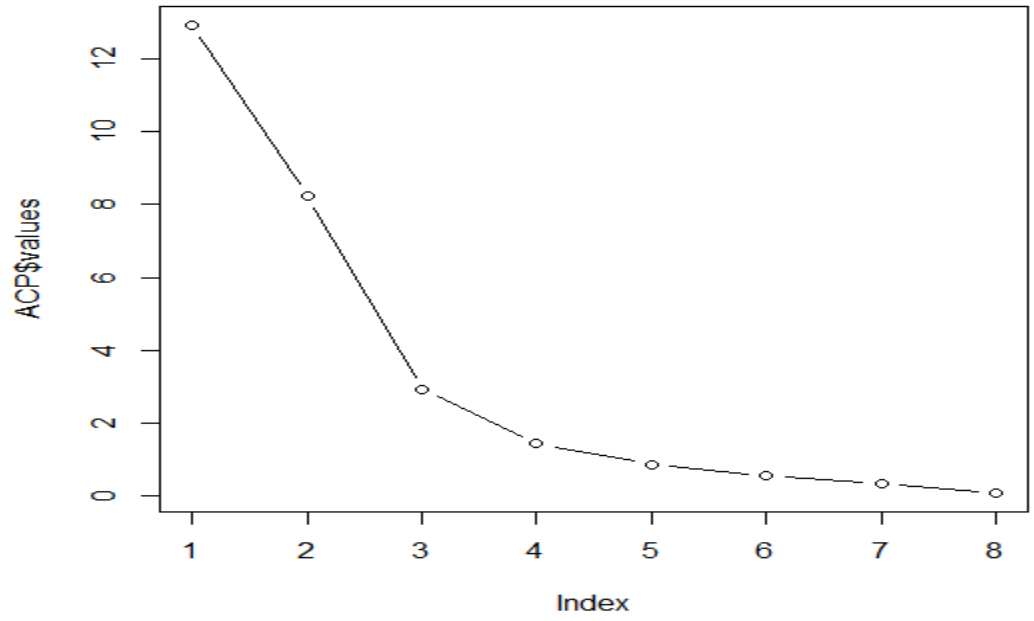
Interprétation des résultats :

Détermination de nombre d'axes à retenir :

On utilise pour cela l'éboulis des valeurs propres représenté ci-contre :

Dans ce graphique, on cherche un coude et on ne conserve que les valeurs propres avant ce coude.

Les axes factoriels principaux sont alors les vecteurs propres associés.



On constate alors que la cassure (le coude) se commence à partir de la troisième valeur propre ; on ne retient alors que les deux premiers axes factoriels correspondants.

Le tableau suivant résume le pourcentage de la variance expliquée par chacun des deux axes ainsi que la variance expliquée par le plan issu de ces deux axes :

	Axe 1	Axe 2
Variance	3.240	2.455
% de la variance	40.503	30.687
% cummulatif de la variance variance	40.503	71.190

Les deux axes retenus expliquent 71.19% de la variance totale.

Résultats sur les individus :

Puisque on n'interprète que les points individus bien représentés, alors on doit se méfier de la qualité de représentation de chaque individu dans le plan principal.

Cette qualité de représentation se calcule en sommant les cos2 de chaque individu pour les deux axes. Le tableau ci-dessous regroupe les la qualité de représentation de chaque individus ainsi que leur contribution à la formation des deux axes :

individus	Axe 1	Contribution	Cos2	Axe 2	Contribution	Cos2
1	-2.571	13.604	0.798	-0.713	1.380	0.061
2	3.056	19.220	0.763	0.108	0.032	0.001
3	2.395	11.798	0.846	-0.146	0.058	0.003
4	1.037	2.211	0.584	0.506	0.695	0.139
5	-3.098	19.750	0.915	0.179	0.087	0.003
6	2.028	8.461	0.778	-0.581	0.917	0.064
7	-0.411	0.348	0.020	2.141	12.464	0.553
8	1.206	2.994	0.174	-1.425	5.515	0.243
9	-0.734	1.109	0.061	-1.608	7.021	0.292
10	-2.392	11.774	0.561	0.800	1.737	0.063
11	-1.593	5.223	0.491	0.889	2.146	0.153
12	1.284	3.394	0.125	3.161	27.142	0.759
13	-0.235	0.113	0.022	-0.824	1.843	0.266
14	0.031	0.002	0.000	-3.617	35.522	0.847
15	-0.002	0.000	0.000	1.129	3.458	0.421

En sommant les cos2 et en adoptant les règles postulées dans le tableau page 17 on conclut que tous les individus, à l'exception des individus 8, 9, 13 et 15, sont bien représentés dans le plan principal.

Les individus bien représentés dans l'axe 1	Les individus bien représentés dans l'axe 2	Les individus bien représentés dans le plan
1 2 3 4 5 6 10	7 12 14	1 2 3 4 5 6 7 10 11 12 14

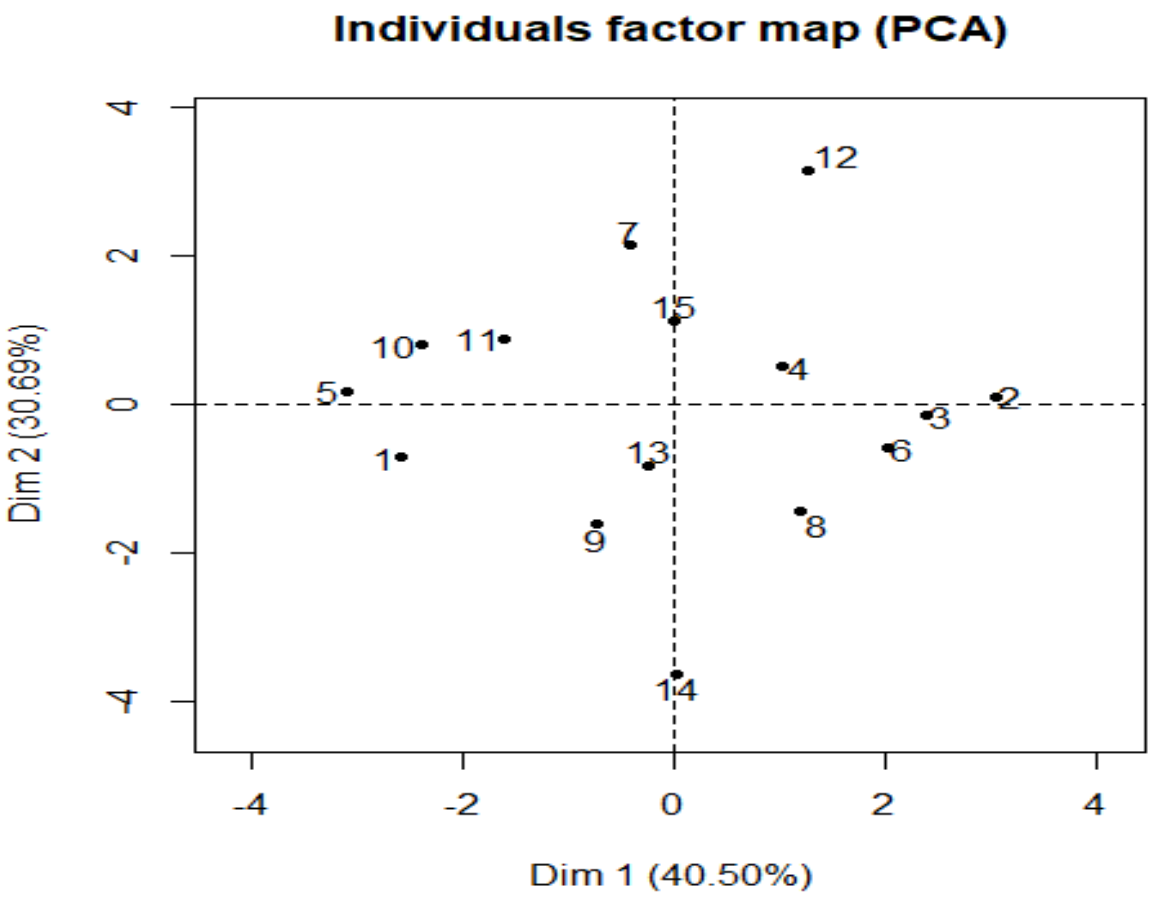
Les individus qui participent le plus à la formation du premier axe sont ceux qui ont une contribution supérieure à la moyenne c.-à-d. supérieure à $\frac{100\%}{15}$ =6.667.

Les individus contribuent à la formation de l'axe 1	Les individus contribuent à la formation de l'axe 2
1 2 3 5 6 10	7 14 12

Les individus qui contribuent le plus à la formation de l'axe 1 se caractérisent par des résultats meilleurs dans l'une des deux disciplines et moyens dans l'autre.

Or les individus qui contribuent à la formation de deuxième axe se caractérisent par des résultats homogènes dans toutes les matières, bonnes ou moyennes.

Ces tableaux d'analyse s'accompagnent par le graphique des individus qui rend les résultats clairs et visibles :



Pour l'axe 1 on constate que les points individus 1, 2, 3, 5 et 10 sont éloignés de l'origine ce qui justifie leur bonne qualité de représentation, ainsi ils sont très proches de cet axe ce qui justifie leur importante contribution à la formation de cet axe.

De même pour l'axe 2 les individus 14, 7 et 12 sont très éloignés de l'origine donc ils ont une bonne qualité de représentation dans cet axe, ils sont ainsi proches de cet axe chose qui signifie leur forte contribution à la formation de cet axe.

Résultats sur les variables :

De même que pour les individus on n'interprète que les variables qui sont bien représentées pour cela on étudie la qualité de représentation de chaque variable ainsi que leur contribution à la formation des deux axes principaux, on a alors le tableau suivant fournie par L'ACP :

variables	Axe 1	contribution	Cos2	Axe 2	contribution	Cos2
arabe	-0.800	19.767	0.640	0.492	9.872	0.242
Français	-0.507	7.928	0.257	0.604	14.838	0.364
Anglais	-0.321	3.180	0.103	0.833	28.297	0.695
Philosophie	-0.568	9.956	0.323	0.409	6.811	0.167
Math	0.840	21.756	0.705	0.493	9.900	0.243
PC	0.788	19.173	0.621	0.488	9.689	0.238
SVT	0.768	18.209	0.590	0.502	10.271	0.252
ES	-0.031	0.030	0.001	-0.503	10.324	0.253

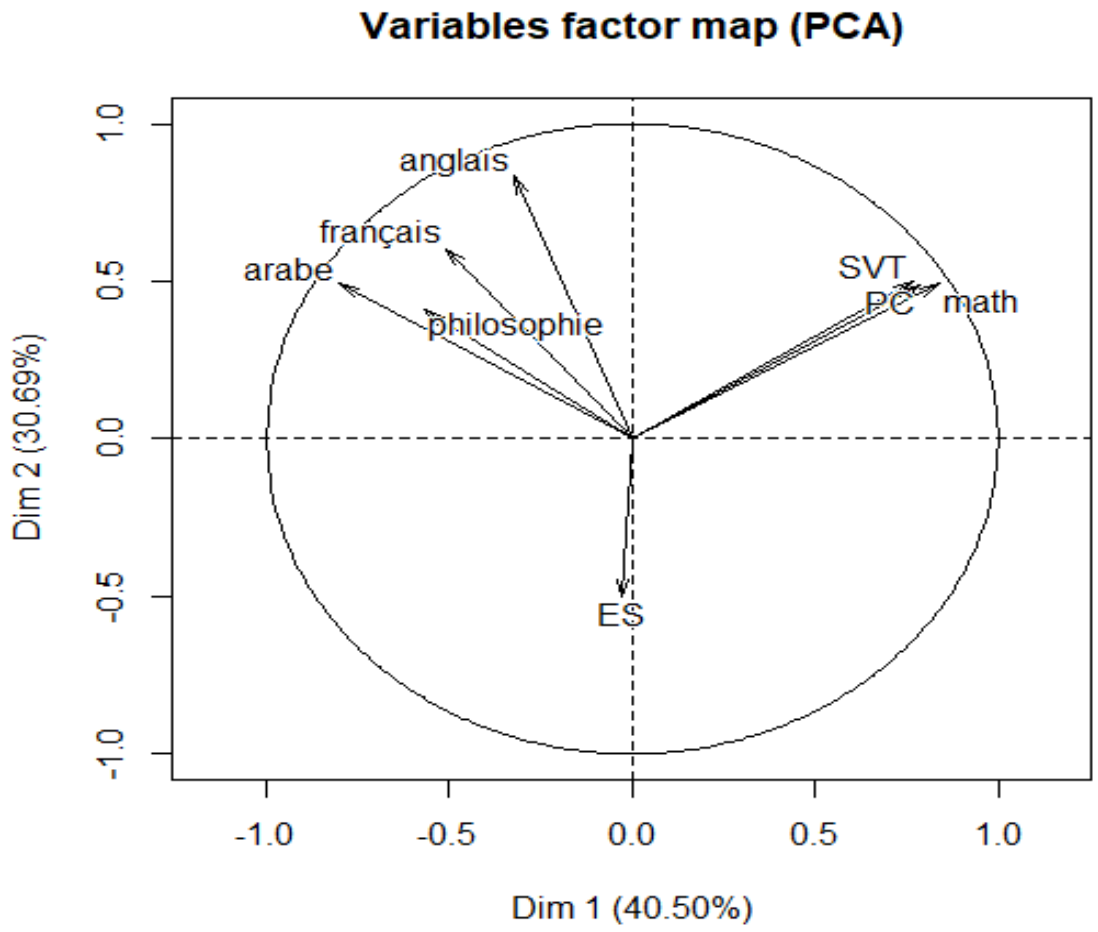
On résume ces résultats dans les deux sous tableaux suivants :

Les variables bien représentés dans l'axe 1	Les variables bien représentées dans l'axe 2	Dans le plans
Arabe, math, PC et SVT	L'anglais	Arabe, français, anglais math, PC et SVT

On choisit les variables ayants une contribution supérieure à la moyenne ($\frac{100\%}{8} = 12.5$)

Les variables qui contribuent le plus à la formation de l'axe 1	Les variables qui contribuent le plus à la formation de l'axe 2
Arabe, math , PC et SVT	Anglais

Le cercle de corrélation résume clairement les résultats décrits par les tableaux



Le cercle des corrélations montre que les deux variables philosophie et ES (éducation sportive) sont mal représentées dans le plans principales car ils sont éloignées du cercle et que toutes les autres variables sont bien représentées dans ce plan.

Interprétation des axes :

On cherche à donner un sens aux axes retenus, pour cela on a étudié les contributions des individus et des variables à la formation de ces axes que l'on résume dans le tableau suivant :

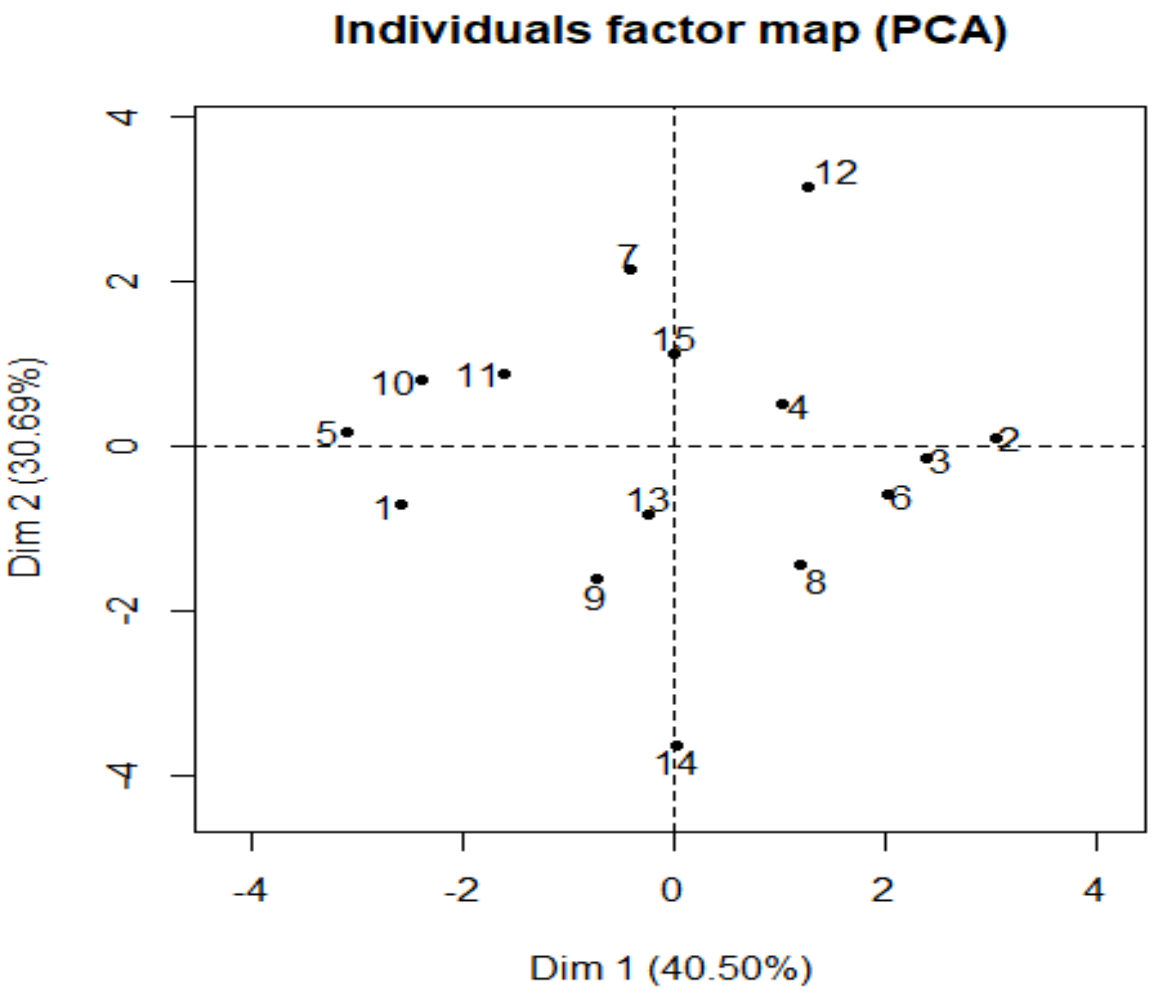
	Contribution pour l'axe 1	Contribution pour l'axe 2
Les individus	1 2 3 5 10	7 12 14
Les variables	Arabe, math , PC et SVT	Anglais

Chaque individus de l'ensemble { 1, 2, 3, 5, 10} a soit des résultats meilleures en arabe et moyennes en math , PC et SVT soit l'inverse, et ils ont en générale des bons résultats en une discipline et moyenne à l'autre, par contre, les individus 7 12 14 ont des résultats homogènes dans toutes les matières et la note d'anglais est la plus fortes pour le 7 et le 12 (resp. 16 et 17.5)

De ces constatation on peut dire que l'axe 1 correspond aux étudiants qui sont soit littéraires soit scientifique et l'axe 2 correspond à ceux qui ont des scores presque identique dans tous les modules.

Interprétation des individus (suite) :

Maintenant on va s'intéresser aux ressemblances entre les individus, et pour cela on doit consulter le graphique représentatif de ces derniers :

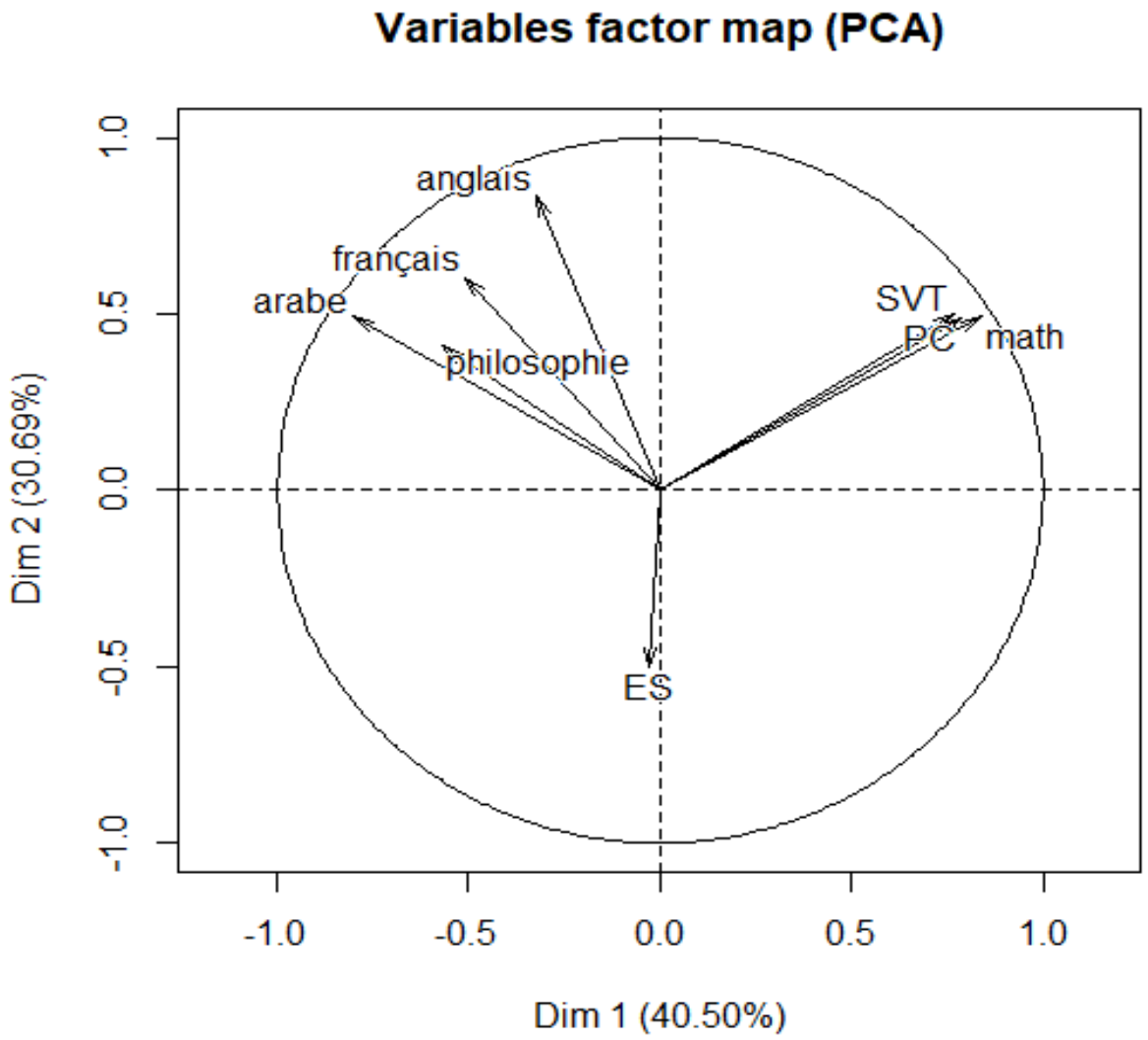


Les deux individus 2 et 5 s'opposent par rapport au deuxième axe ce qui traduit que leurs résultats des différentes matières le sont.

Les trois individus 1, 5 et 10 sont proches, ils ont alors une ressemblance réelle de point de vue des variables, ces trois individus se caractérisent par des bons résultats aux disciplines littéraires et moyens aux disciplines scientifiques, il en est de même pour le groupe {2, 3, 6} mais celui-là a par contre des résultats meilleurs en math, PC et SVT et moyens en arabe, français et anglais.

Interprétation des variables (suite) :

Le graphique des variables se caractérise par le cercle des corrélations ci-dessous :



Les trois variables math, PC et SVT sont très corrélées entre elles, la corrélation étant positive ce qui justifie que les trois variables agissent dans le même sens pour les individus c.-à-d. qu'une bonne note dans l'un de ces trois modules s'accompagne par des bonnes notes aux deux autres et inversement.

L'examen de la matrice des corrélations ci-dessous nous indique que les deux disciplines littéraire et scientifique sont corrélées négativement et les modules de même discipline sont corrélés positivement, ce qui signifie que les deux groupes de modules s'opposent pour les individus qui les décrivent.

	Arabe	Français	Anglais	Philo	Math	PC	SVT	ES
Arabe	1.000	0.666	0.623	0.627	-0.437	-0.391	-0.299	-0.188
Français	0.666	1.000	0.653	0.153	-0.150	-0.040	-0.205	-0.278
Anglais	0.623	0.653	1.000	0.514	0.158	0.185	0.157	-0.108
Philo	0.627	0.153	0.514	1.000	-0.274	-0.258	-0.130	-0.183
Math	-0.437	-0.150	0.158	-0.274	1.000	0.868	0.854	-0.301
PC	-0.391	-0.040	0.185	-0.258	0.868	1.000	0.799	-0.095
SVT	-0.299	-0.205	0.157	-0.130	0.854	0.799	1.000	-0.224
ES	-0.188	-0.278	-0.108	-0.183	-0.301	-0.095	-0.224	1.000

Ainsi on a déterminé l'intérêt des étudiants pour chaque discipline, ce qui facilitera leur orientation soit vers des disciplines littéraires, soit vers des disciplines scientifiques.

Exemple 2 :

Sur un échantillon de 20 familles à Sefrou on a distribué un questionnaire pour recueillir des réponses aux questions suivantes

-Le nombre d'enfants : enf

-Le salaire : slr

-la quantité de farine en Kg : frn

-la quantité de poisson en Kg : poi

-la quantité de lait en *l*

-la viande en Kg : vnd

-l'huile d'olive en *l* : HO

-les fruits en Kg : frui

-le nombre œufs : efs

-les légumes en Kg : lgm

-le thé en *l*

-le poulet en Kg : plt

-l’huile végétal en *l :* HV

On va étudier la relation entre le nombre d’enfants, les revenus et la consommation mensuelle de ces 11 aliments.

On veut réaliser une ACP du tableau ci-contre pour connaitre les structures et les modes de consommation des habitants de ce quartier.

Les données issues de cet enquête sont représentées dans le tableau ci-dessous :

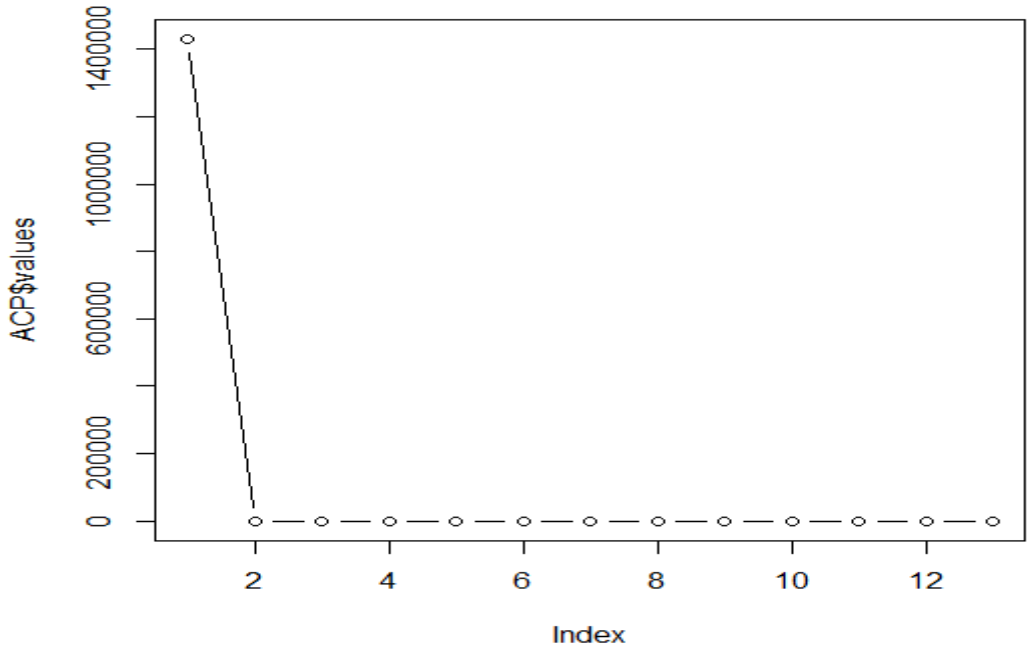
famille	enf	slr	frn	poi	lait	vnd	HV	HO	frui	efs	lgm	Thé	plt
1	4	3000	40	5	15	4	5	1	15	8	30	20	4
2	3	2000	40	1	4	0.5	5	0	6	60	40	50	6
3	1	5000	30	5	30	6	4	2	15	10	20	8	3
4	5	3500	50	4	10	4	7	1	8	10	40	50	2
5	2	2500	40	2	6	1	5	0.5	10	45	30	40	5
6	1	2300	25	1	4	1	5	1	5	15	30	30	4
7	0	2200	15	1	6	0	2	0.5	8	40	30	20	4
8	3	6000	50	4	30	8	7	2	10	20	30	10	3
9	3	3400	40	3	40	4	5	1	15	15	40	15	5
10	3	2700	45	1	6	1	5	0	8	60	50	30	4
11	5	2250	50	0	5	0	9	0.5	4	40	55	50	3
12	2	4500	35	2	10	2	5	1	6	40	30	20	2
13	2	3300	30	3	15	6	4	1	20	20	40	35	4
14	0	1700	10	1	4	0.5	2	0	8	60	30	20	3
15	1	2200	20	0.5	0	1	3	0.5	4	30	35	30	2
16	2	1900	30	1	4	0.5	5	0	4	30	35	40	4
17	2	2300	40	1	6	1	5	1	6	30	40	30	4
18	2	5200	40	2	20	8	6	1	20	20	25	20	2
19	4	4000	50	2	18	6	7	2	15	15	30	20	5
20	2	2500	30	1	5	1	4	0	6	20	40	40	4

Analyse et interprétation des résultats :

Ces données sont toutes quantitatives on les traite alors par une ACP.

Le nombre d'axe à retenir :

Pour répondre à cette question on consulte l'éboulis des valeurs propres :



En utilisant ce graphique, on peut dire que les deux premiers axes restituent une bonne proportion de la variance totale, en effet ces deux axes expliquent environ 71.26% ; 46.21% pour l'axe 1 et 25.05% pour le deuxième axe comme l'indique le tableau suivant :

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	6.008	3.256	1.176	0.707	0.572	0.416	0.370
% of var.	46.213	25.049	9.043	5.442	4.399	3.204	2.843
Cumulative % of var.	46.213	71.262	80.306	85.747	90.146	93.350	96.193

La chute est importante dès la troisième valeur propre (de 25.049% à 9.043%) d'où le choix des deux premiers axes.

Qualité de représentation des individus :

Le tableau suivant rassemble les qualités de représentation des individus dans les deux axes retenus :

Individus	Qualité dans l'axe 1 Cos 2	Qualité dans l'axe 2 Cos2	Qualité dans le plan ∑Cos2
1	0.495	0.005	0.500
2	0.550	0.134	0.684
3	0.714	0.250	0.964
4	0.093	0.514	0.607
5	0.303	0.008	0.311
6	0.210	0.170	0.380
7	0.309	0.663	0.972
8	0.845	0.013	0.858
9	0.395	0.005	0.400
10	0.416	0.193	0.609
11	0.178	0.723	0.901
12	0.020	0.088	0.108
13	0.206	0.020	0.226
14	0.389	0.527	0.916
15	0.434	0.249	0.683
16	0.823	0.001	0.824
17	0.463	0.063	0.526
18	0.617	0.026	0.643
19	0.579	0.122	0.701
20	0.665	0.000	0.665

Sur le premier axe on peut considérer que les individus 3, 8, 16, 18, et 20 sont bien représentés, les deux individus 2 et 19 ont une qualité acceptable et tous les autres individus sont mal représentés dans ce premier axe.

De même pour l'axe 2 les individus bien représentés dans cet axe sont 7, 11, et 14, le 4 à une qualité acceptable et tout le reste est mal représenté dans cet axe.

Mais dans le plan formé de ces deux axes on peut considérer que les individus 1, 2, 3, 4, 7, 8, 10, et 11 ainsi que les individus de 14 à 20 sont bien représentés dans ce plan.

Les individus bien représentés sont déterminés, on étudie maintenant leur participation à la formation des axes, le tableau ci-contre résume la contribution des individus, nous ne conservons que celles supérieures à la moyenne (supérieures à 100/20=5)

La contribution la plus importante à la formation de l'axe 1 est celle des individus 8, 3, 18, 2, 19 et 14

Contribution +	Contribution -
3 8 18 19	2 14

De même les individus les plus influents sur la formation du deuxième axe sont 11, 14, 7, 4 et 3

Contribution +	Contribution -
4 11	3 7 14

Individus	Contribution Pour l'axe 1	Contribution pour l'axe 2
1	4.081	0.075
2	7.402	3.325
3	14.959	9.680
4	1.233	12.529
5	1.260	0.063
6	1.041	1.555
7	3.505	13.887
8	17.309	0.496
9	4.371	0.108
10	3.772	3.233
11	3.891	29.223
12	0.095	0.757
13	1.374	0.248
14	6.690	16.741
15	4.121	4.369
16	4.672	0.012
17	1.181	0.297
18	8.700	0.669
19	7.041	2.731
20	3.302	0.000

Analyse des variables :

Qualité de représentation des variables :

Cette qualité se mesure comme pour les individus :

Variables	Qualité dans l'axe 1	Qualité dans l'axe 2	Qualité dans le plan
Nombre d'enfants	0.070	0.856	0.926
Le salaire	0.811	0.000	0.811
La farine	0.183	0.738	0.921
Le pisson	0.682	0.002	0.684
Le lait	0.743	0.001	0.744
La viand	0.888	0.000	0.888
L'huile végétal	0.117	0.754	0.871
L'huile d'olive	0.787	0.000	0.787
Les fruits	0.555	0.021	0.576
Les œufs	0.512	0.001	0.513
Les légumes	0.236	0.496	0.732
Le thé	0.384	0.364	0.748
Le poulet	0.040	0.023	0.063

Conclusion : les variables bien représentées dans le premier axe sont : le salaire, le poisson, le lait, la viande, l'huile d'olive, les fruits et les œufs.

Les variables bien représentées dans le deuxième axe sont : le nombre d'enfants, la farine, l'huile végétal et on peut aussi ajouter la variable « légumes » car sa qualité de représentation dans cet axe est proche de 0.5.

Mais sur le plan toutes les variables sont bien représentées à l'exception de la variable poulet qui a une qualité largement inférieure à 0.5.

Contribution des variables à la formation des axes :

On ne conserve que les contributions supérieures à la contribution moyenne (supérieure à 100/13≈7.692)

Conclusion :

Les variables contribuent le plus à la formation du premier axe sont : viande, salaire, huile d'olive, lait, poisson, fruits et œufs.

Celles qui contribuent le plus à la formation du deuxième axe sont : nb enfants, farine, huile végétal, légumes et la variable thé.

Les variables	CTR Axe1	CTR Axe 2
Nbenfants	1.168	26.272
Salair	13.494	0.000
Farine	3.047	22.660
Poisson	11.350	0.070
lait	12.365	0.034
Viand	14.780	0.000
Huile végétal	1.945	23.165
Huile d'olive	13.105	0.004
fruits	9.236	0.650
oeufs	8.517	0.026
légumes	3.924	15.238
thé	6.396	11.186
poulet	0.673	0.694

Interprétation des axes :

L'axe 1 :on regroupe dans le tableau suivant les contributions des individus :

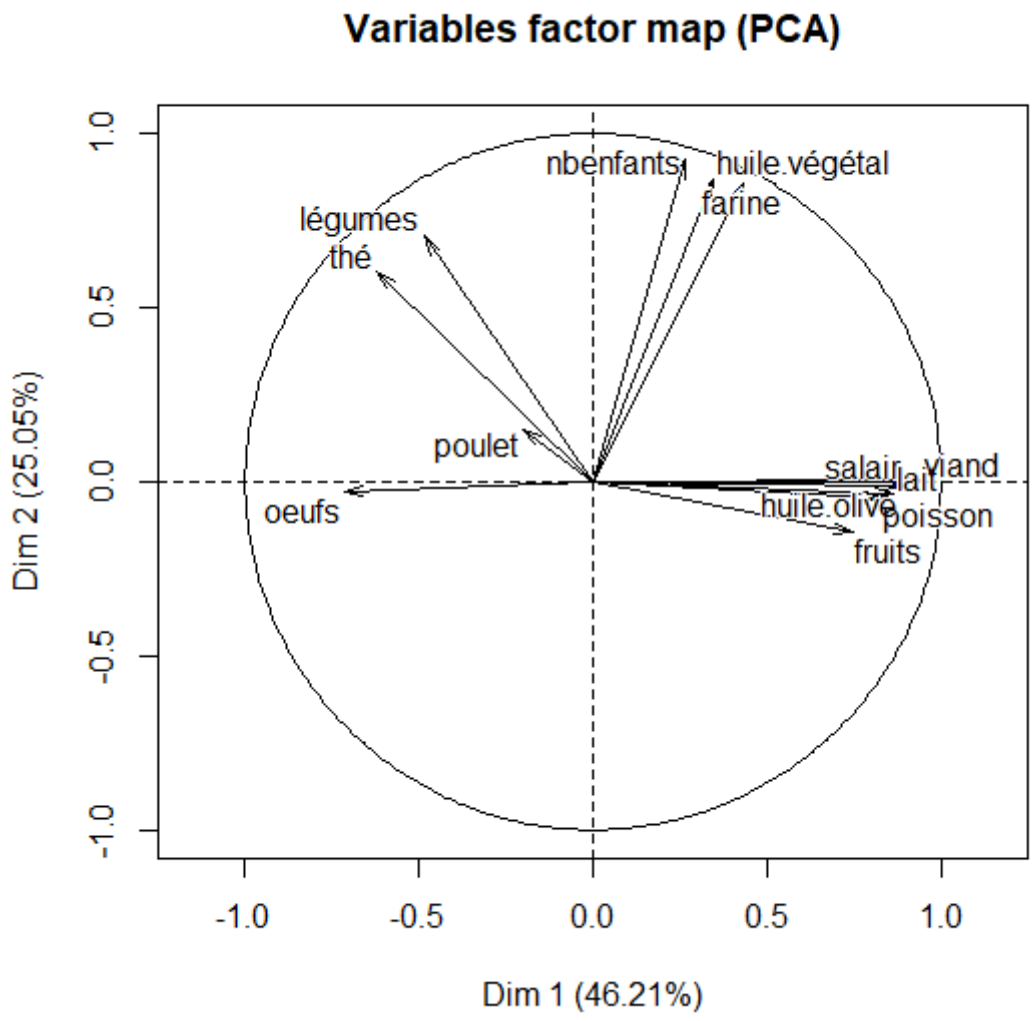
	Contribution +	Contribution -
Les individus	3 8 18 19	2 14
Les variables	viand, salair, huile d'olive, lait, poisson et fruits.	Œufs

L'axe 1 oppose alors les familles dont le salaire est important et dont la consommation de la viande, d'huile d'olive, du lait, de poisson et des fruits est importante, contre les familles dont

le salaire est faible ainsi que la consommation de ces derniers aliments et dont la consommation des œufs, des légumes et du thé est remarquable.

l'axe 2 : les individus qui contribuent positivement à la formation de cet axe sont le 11 et le 4 dont le nombre d'enfants est important ainsi que la consommation de la farine, d'huile d'olive, des légumes et du thé, par contre les individus contribuent négativement à la formation de cet axe n'ont pas d'enfants et ils ont une consommation faibles de ces aliment.

Ces résultats sont clairs grâce au cercle des corrélations :



Ce cercle nous indique que toutes les variables sont bien représentées car elles sont proches du cercle sauf pour la variable poulet qui est loin du cercle et par suite elle est mal représentée dans le plan principal.

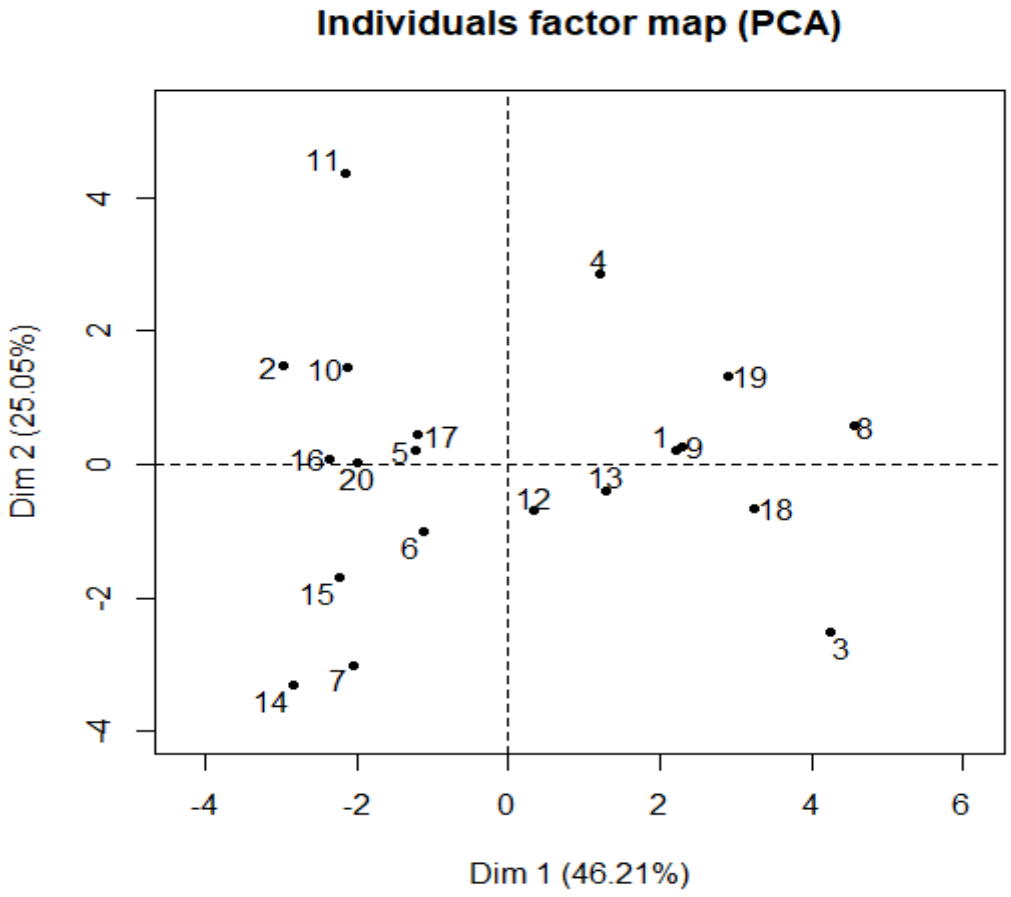
Les variables salaire, viande, fruits, poisson, lait et huile d'olive sont très corrélées entre elles, la corrélation étant positive ce qui traduit alors qu'une forte valeur de l'une de ces variables entraine une forte valeur des autres, elles sont aussi corrélées avec le premier axe ce qui justifie leur forte contribution à la formation de cet axe.

De même les variables nb enfants, farine et huile végétal sont corrélées entre elles et avec le deuxième axe ce qui signifie leur forte influence sur la formation de cet axe.

De toutes ces constatations on peut conclure que le premier axe principal mesure l'état économique des familles, et l'axe 2 mesure leur état social.

Etude des proximités entre les individus :

Pour cela on doit consulter la représentation des individus dans le plan principal



On rappelle que les individus bien représentés dans ce plan sont : 1, 2, 3, 4, 7, 8, 10, et 11 ainsi que les individus de 14 à 20.

Les deux individus 16 et 20 sont très proches, ces deux familles présentent le même comportement de point de vu des variables associées, en effet leur consommation est pratiquement identique comme le montre le tableau des données brutes.

Il en est de même pour les familles 7 et 14 et pour les deux familles 2 et 10.

Les deux individus 1 et 16 s'opposent par rapport au deuxième axe, ils s'opposeront alors aux variables qui font cet axe.

Bien entendu la famille 16 a deux enfants mais avec une consommation importante des légumes, du thé et des œufs par contre la famille 1 a 4 enfants mais avec une consommation faibles de ces aliment et forte dans ce qui concerne le poisson, le lait, la viande et les fruits cette différence est dû à l'écart entre les salaires des deux (le 1 a un salaire de 3000 dh mais le 16 n'a que 1900 dh).

La symétrie est aussi claire entre les deux individus 10 et 15 mais cette fois-ci par rapport au premier axe, ces deux s'opposent alors par rapport aux variables qui contribuent à la formation de cet axe. En effet la famille 10 consomme les poissons, le lait les fruits et les œufs par contre la famille 15 a une consommation faibles de ces aliments, l'écart de salaire entre les deux familles est de 500 dh, dans ce cas on peut dire que la consommation des deux familles est liées au nombre d'enfants seulement.

Conclusion : en analysant les résultats obtenus par cette ACP et en revenant au tableau des données brutes on peut conclure que les familles salariées plus de 3000 dh se dirigent vers une consommation des aliments chères (les poissons, le lait, la viande et les fruits), par contre les familles salariées moins de 3000 dh se dirigent vers une consommation moins chère (la farine, les œufs et les légumes).

Cette différence est liées aux salaires et au nombre d'enfants et donc au pouvoir d'achat de chaque catégorie socioéconomique.

B- LES PLANS D'EXPERIENCES :

I- Définition et objectifs

Définition :

Les plans d'expériences comme J'ai énoncé à l'introduction constituent essentiellement une stratégie de planification d'expériences afin d'obtenir des conclusions solides et adéquates de manière efficace et économique

a) Formalisme mathématique et vocabulaire :

a-1) vocabulaire :

Les plans d'expériences permettent d'organiser au mieux les essais qui accompagnent une recherche scientifique, cette recherche consiste à s'intéresser à une grandeur particulière qui dépend d'un grand nombre de variables contrôlables, cette dépendance peut se traduire sous forme d'une relation mathématique reliant la grandeur y aux différents variables associées notons les $x_1, x_2, \ldots\ldots\ldots\ldots x_k$.

$$y = f(x_1, x_2, \ldots\ldots\ldots\ldots x_k)$$

La grandeur y est appelée réponse et les variables associées sont appelées facteurs.

L'étude de cette grandeur se ramène alors à déterminer la fonction f qui lie la réponse y aux différents facteurs $x_1, x_2, \ldots\ldots\ldots\ldots x_k$.

Pour ce faire on étudie l'influence de chaque facteur, en limitant ses variations entre deux niveaux. La borne inférieure est le niveau bas. La borne supérieure est le niveau haut, le facteur peut prendre les valeurs comprise entre ces deux niveaux c'est le domaine du facteur.

S'il y a plusieurs facteurs, chacun d'eux à son domaine de variation. Afin d'avoir une représentation commune pour tous les facteurs, on a la convention suivante : indiquer les niveaux bas par –1 et les niveaux hauts par +1

Chaque facteur est représenté par un axe perpendiculaire aux autres axes

Une expérience donnée est alors représentée par un point dans ce système d'axes. Ce point est le point expérimental

Le domaine d'étude est celui limité par ces axes.

a-2) Modélisation mathématique :

En l'absence de toute information sur la fonction qui lie la réponse aux facteurs, on se donne a priori une loi d'évolution dont la formulation la plus générale est la suivante :

$$y = f(x_1, x_2, \ldots\ldots\ldots x_k)$$

Un développement limité nous permet d'approcher cette fonction par un polynôme de degré plus ou moins élevé $y = a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + \sum a_{ii} x_i^2 + \cdots$

Tel que : – x_i représente un niveau du facteur *i*,

– x_j représente un niveau du facteur *j*,

– $a_0, a_i, a_{ij} \ldots\ldots$ sont les coefficients du polynôme et ce sont les inconnues.

Ce modèle est appelé le modèle postulé. Mais à ce modèle on doit apporter deux corrections, la première est le manque d'ajustement qui traduit le fait que le modèle choisi par l'expérimentateur avant les expériences est probablement différent du modèle réel qui régit le phénomène étudié. Il y a un écart entre ces deux modèles. Cet écart est le manque d'ajustement on le note Δ.

La seconde correction est la prise en compte de la nature aléatoire de la réponse, car si l'on mesure plusieurs fois une réponse en un même point expérimental, on n'obtiendra pas exactement le même résultat. Il y a une dispersion des résultats. Les dispersions ainsi constatées sont appelées erreurs aléatoires ou erreurs expérimentales et on les note ε.

La modélisation doit être modifiée ainsi :

$$y = f(x_1, x_2, \ldots\ldots\ldots x_k) + \Delta + \varepsilon$$

a-3) Détermination des coefficients :

Pour déterminer les coefficients du modèle on adopte l'écriture matricielle suivante :

$$y = Xa + e$$

- y est le vecteur des réponses
- X est la matrice du modèle dépendant des points expérimentaux choisis pour exécuter le plan et du modèle postulé,
- a est le vecteur des coefficients,
- e est le vecteur des erreurs.

Pour identifier les coefficients on doit réaliser les essais. Mais le problème se pose lorsque on étudie plusieurs facteurs car il s'agit d'effectuer tous les essais possible ce qui est déraisonnable et coût en temps et en finance. Les plans d'expérience ont pour but de simplifier cette difficulté.

Un plan d'expériences est alors une liste ordonnée d'essais à effectuer, permettant d'identifier les coefficients d'un modèle donné de manière efficace, et on peut les classer en deux catégories :

1 - Plans complets

Une première catégorie de plans d'expériences est destinée à fournir une information la plus complète possible sur des systèmes présentant relativement peu de facteurs Ces plans consistent à tester toutes les combinaisons possibles, en faisant varier tous les facteurs à tous leurs niveaux, par exemple si on veut étudier une grandeur dépendant de six paramètres on doit effectuer $2^6 = 64$ essais.

Pour ce faire, une technique simple a été proposée pour le cas où chaque facteur n'a que deux niveaux. Elle consiste à numéroter les facteurs et à faire varier successivement leurs niveaux de la façon suivante :

1) Au premier essai, tous les facteurs sont au niveau bas.

2) On change le niveau du premier facteur à chaque essai...

3) celui du deuxième facteur tous les 2 essais...

4) et, plus généralement, celui du k ème facteur tous les 2^{k-1} essais.

2 - Plans réduits

En pratique, les plans complets ne sont utilisables que sur des systèmes avec très peu de facteurs, par exemple si on veut étudier une grandeur qui dépend de 7 facteurs on doit effectuer $2^7 = 128$ essais ce qui est dur, les plans réduits ont donc été proposés. Ils permettent de réduire le nombre d'essais à effectuer.

Pour cela, on prend rarement en compte toutes les interactions possibles dans le modèle.

Les plans réduits sont basés alors sur l'hypothèse d'élimination des interactions d'ordre trois et toutes celles d'ordre plus élevé.

L'usage de ces plans demande tout d'abord d'écrire le modèle (c'est-à-dire de lister les facteurs et les interactions à prendre en compte) et de choisir le nombre de niveaux des facteurs, dans notre étude on associe à chaque facteur deux niveaux. Différentes techniques sont alors utilisables ; nous en présentons la plus fréquente.

Méthode de Box et Hunter

La méthode de Box et Hunter permet de construire soi-même des plans réduits à partir de plans complets. Elle s'adresse exclusivement aux modèles à deux niveaux par facteur et se base sur la

définition suivante : soient xi et x_j deux facteurs admettant chacun deux niveaux, notés +1 et -1. On appelle niveau de l'interaction entre x_i et x_j, et on note I_{ij}, le produit de leurs niveaux respectifs Ainsi, si x_i et x_j sont tous deux au niveau haut (+1) ou au niveau bas (-1), leur interaction est au niveau haut (+1) ; dans le cas contraire, elle est au niveau bas. Le niveau de l'interaction de deux facteurs exprime donc formellement si, lors d'un essai donné, les deux facteurs agissent dans le même sens ou non. Elle se généralise à plus de deux facteurs : ainsi, étant donnés trois facteurs x_i, x_j et x_k admettant chacun deux niveaux, on note I_{ijk} le produit de leurs trois niveaux respectifs, et ainsi de suite. La méthode de Box et Hunter consiste à négliger l'interaction d'ordre le plus élevé, et à ne conserver que les essais donnant un même signe (par exemple +1) à cette interaction.

Ces produits entre les niveaux des facteurs est connus sous le nom de calcul de Box :

Calcul de Box :

Supposons que notre travail se fait en dimension 4

On désigne par I le vecteur signe identité $I=\begin{bmatrix}+1\\+1\\+1\\+1\end{bmatrix}$

Et soit u et v deux vecteurs signes quelconques prenons par exemple $u=\begin{bmatrix}+1\\-1\\+1\\+1\end{bmatrix}$ et $v=\begin{bmatrix}-1\\+1\\-1\\-1\end{bmatrix}$

Définition : la multiplication de deux vecteurs signes ou plus se fait terme par terme.

Alors on a les propriétés suivantes :

Commutativité : cette multiplication est commutative : $uv=vu$

La multiplication d'un vecteur signe par lui-même donne le vecteur identité $uu=I$

La multiplication d'un vecteur signe par le vecteur signe identité est le vecteur lui-même $uI=u$

Ces plans sont aussi basés sur un théorème fondamentale celui des alias, qui permet de regrouper les inconnues deux à deux (ou plus), en s'arrangeant pour que les colonnes du tableau possèdent deux à deux des niveaux identiques Ces regroupements de coefficients s'appellent des alias ou des contrastes.

La théorie des alias est très utile pour comprendre et interpréter les résultats d'un plan fractionnaire.

Cette théorie s'accompagne par des hypothèses qui aident à résoudre le système et d'analyser les résultats :

Hypothèse 1 : les interactions d'ordre égal ou supérieur à 3 peuvent être négligées.

Hypothèse 2 : si un contraste est négligeable, tous les termes aliasés sont eux-mêmes négligeables ; une compensation des termes est très improbable.

Hypothèse 3 : si deux effets de facteurs sont négligeables, on supposera que leur interaction l'est aussi.

Hypothèse 4 : une interaction comportant deux facteurs dont l'un a un effet négligeable, est généralement une interaction négligeable.

Ces postulats ainsi définis présentent en collaboration avec la méthode de Box un aide pour l'analyse des résultats et pour la bonne construction des plans fractionnaires.

On ne va pas se retarder aux aspects théoriques calculatoires car les logiciels nous donne les résultats aisément, notre rôle est d'interpréter ces résultats, et d'extraire le maximum d'informations à travers les exemples suivants :

II. Exemples pratiques

Exemple 1 : Un plan complet !

On veut étudier le rendement d'une réaction chimique en ne s'intéressant qu'à trois facteurs la température, la pression, et la quantité de catalyseur, et tel que chacun de ces facteurs a deux niveaux on résume les données dans le tableau ci-dessous :

Facteurs	Niveau bas (-)	Niveau haut (+)
Température (C)	60	80
Pression (bar)	2	4
Masse catalyseur (g)	50	80

Le pH du milieu réactionnelle, la nature de catalyseur, et la masse des réactifs restent les mêmes pendant tous les essais.

On a trois facteurs à étudier chacun d'eux a deux niveaux bien déterminés, les facteurs à conserver constants pendant les essais ont été précisés et seront vérifiés avant chaque essai, et puisque le nombre d'essais à effectuer est raisonnable (8), il convient alors de choisir un plans d'expérience complet 2^3 dont le modèle a priori est polynomiale de premier degré (ici c'est le degré de chaque facteur) d'où on écrit :

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_1x_2x_3$$

- y est la réponse, c'est le rendement dans notre exemple.
- Les a_i sont les effets des facteurs, a_{ij} est l'interaction de deux facteurs et a_{123} l'interaction des trois facteurs.

Cette équation relie le rendement y aux facteurs sélectionnés par l'expérimentateur, en tenant compte de leurs interactions, et en utilisant la méthode de la numérotation des essais on obtient la matrice d'expérience suivante :

N° essai	Température	Pression	Masse catalyseur	Rendement
1	-1	-1	-1	80
2	+1	-1	-1	72
3	-1	+1	-1	95
4	+1	+1	-1	60
5	-1	-1	+1	88
6	+1	-1	+1	55
7	-1	+1	+1	120
8	+1	+1	+1	69

De ce tableau on constate que les forts rendements de la réaction sont du côté des faibles températures, ainsi le rendement le plus fort est obtenu lorsque la température est au niveau bas et les autres facteurs au niveau haut.

A partir du modèle postulé et en adoptant la règle de Box pour les interactions des facteurs, on obtient la matrice X des effets associée au modèle.

Cette matrice regroupe tous les essais possibles et toutes les interactions des facteurs.

essai	I	T	P	C	TP	TC	PC	TPC
1	+1	-1	-1	-1	+1	+1	+1	-1
2	+1	+1	-1	-1	-1	-1	+1	+1
3	+1	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	+1	-1	+1	-1	-1	-1
5	+1	-1	-1	+1	+1	-1	-1	+1
6	+1	+1	-1	+1	-1	+1	-1	-1
7	+1	-1	+1	+1	-1	-1	+1	-1
8	+1	+1	+1	+1	+1	+1	+1	+1

Cette matrice est d'Hadamard, elle vérifie $X^T X = nI_8$

Rappel : une matrice d'Hadamard est une matrice de taille $n \times p$ vérifiant :

1) le produit scalaire des colonnes prises deux à deux est nul (orthogonalité).

2) la norme euclidienne au carré d'une colonne est égale à n.

Cette matrice vérifie alors : $X^T X = nI_p$ et $XX^T = nI_n$

L'estimation des coefficients est donnée par $\hat{a} = (X^t X)^{-1} X^t Y$, or puisque il s'agit d'une matrice d'Hadamard alors on écrit $\hat{a} = \frac{1}{n} X^T Y$, l'inversion de la matrice se remplace alors par une transposition.

On obtient alors le vecteur des coefficients :

$$a^t = (79.875 \quad -15.875 \quad 6.125 \quad 3.125 \quad -5.625 \quad -5.125 \quad 5.375 \quad 1.125)$$

Le modèle s'écrit alors :

$$y = 79.875 - 15.875x_1 + 6.125x_2 + 3.125x_3 - 5.625x_1x_2 - 5.125x_1x_3 + 5.375x_2x_3 + 1.125x_1x_2x_3$$

Tel que les x_i sont les trois facteurs étudiés exprimés en unité codé (-1 et +1).

NB : comment peut-on passer des unités codées aux unités physiques réelles ?

Bien entendu le passage des variables d'origines V aux variables codées v et inversement est effectué par la formule suivante : $v = \frac{V - V_0}{pas}$

Avec $V_0 = \frac{niveau\ haut\ réel + niveau\ bas\ réel}{2}$ et $pas = \frac{niveau\ haut\ réel - niveau\ bas\ réel}{2}$

Analyse et interprétation :

C'est le facteur température qui est le plus influent, avec un effet négatif, ce qui signifie que le rendement de la réaction baisse quand le facteur température passe du niveau bas au niveau haut, ce qui justifie la première constatation, on fixe alors le niveau de la température à -1.

Effet	valeur
Constante	79.875
Température	-15.875
Pression	6.125
Masse catalyseur	3.125
Interaction TP	-5.625
Interaction TC	-5.125
Interaction PC	5.375
Interaction TPC	1.125

Lorsque on fixe le facteur température à-1 on constate que le rendement est plus fort lorsque le facteur pression est au niveau haut +1 quel que soit le niveau du troisième facteur.

Les facteurs pression et masse catalyseur ont des effets positifs, ce qui signifie que la réponse augmente lorsque ces deux effets passent du niveau bas au niveau haut. L'interaction TP est relativement importante, d'effet négatif ce qui montre que les deux facteurs température et pression agissent dans deux sens opposés, il en est de même pour l'interaction TC, et c'est le contraire pour l'interaction PC dont les facteurs pression et masse catalyseur agissent dans le même sens.

L'interaction d'ordre 3 est pratiquement nulle.

Donc pour maximiser le rendement on fixe le facteur température à son niveau bas et les autres à ses niveaux hauts.

Mais avant d'utiliser ce modèle pour prédire des résultats, on doit se méfier de sa validité, on doit alors étudier la signification de chaque coefficient du modèle.

En supposant que les coefficients appartiennent à une population normale, l'écart réduit des coefficients suit alors une loi de Student à n – p degré de liberté ; avec n le nombre d'essais utilisés, et p le nombre des coefficients du modèle.

Dans notre exemple on a utilisé 8 essais, pour déterminer 7 coefficients (le coefficient a_{123} est négligeable) donc on a un degré de liberté

Rendement mesuré	Rendement calculé	Résidu r_i
80	81.125	-1.125
72	70.875	1.125
95	93.875	1.125
60	61.125	-1 .125
88	86.875	1.125
55	56.125	-1.125
120	121.125	-1.125
69	67.875	1.125

La somme des carrés résiduels est SCR=10.125, ici on a un seul degré de liberté d'où :

$$\text{SCMR}=\frac{SCR}{1}$$

Les coefficients sont calculés avec 8 réponses, d'où l'estimation de la variance des coefficients par var (a_i)= $\frac{SCR}{8}$ $d'où \ \text{var}(a_i)=\frac{10.125}{8}=1.265625$.

L'écart type des coefficients est alors $\sigma(a_i)=\sqrt{1.265625}=1.125$.

Le tableau ci-contre nous permet de tester la signification des coefficients avec un test de Student :

Coefficient	Valeur	Ecart type	Le t de Student	La p-value
a_0 La constante	79.875	1.125	71	0.00896588
a_1 Température (T)	-15.875	1.125	-14.11	0.04504303
a_2 Pression (P)	6.125	1.125	5.44	0.11573365
a_3 Masse catal(C)	3.125	1.125	2.78	0.21982532
a_{12} Interaction (TP)	-5.625	1.125	-5	0.12566592
a_{13} Interaction (TC)	-5.125	1.125	-4.56	0.13743402
a_{23} Interaction (PC)	5.375	1.125	4.78	0.13129052

La « p-value » est la probabilité pour qu'un coefficient soit nul, cette valeur nous confirme que tous les coefficients sont significatifs.

Exemple 2 : un plan fractionnaire

On va reprendre l'exemple précédent, mais cette fois ci on va s'intéresser à 4 facteurs, la température, la pression, la masse du catalyseur, et le pH du milieu.

Puisque les deux réactifs coûtent chère, on ne va pas procéder avec un plan complet 2^4 mais avec un plan fractionnaire 2^{4-1}, ce plan est obtenu en conservant les essais dont l'interaction d'ordre 4 est positive.

La matrice des effets est alors celle du premier exercice, elle contient 8 termes, et ils nous restent 8 autres

essai	I	T	P	C	TP	TC	PC	TPC
1	1	-1	-1	-1	+1	+1	+1	-1
2	1	+1	-1	-1	-1	-1	+1	+1
3	1	-1	+1	-1	-1	+1	-1	+1
4	1	+1	+1	-1	+1	-1	-1	-1
5	1	-1	-1	+1	+1	-1	-1	+1
6	1	+1	-1	+1	-1	+1	-1	-1
7	1	-1	+1	+1	-1	-1	+1	-1
8	1	+1	+1	+1	+1	+1	+1	+1

Il nous reste qu'à placer les 8 autres termes !

Comment ? : lorsque on écrit la matrice des effets correspondant au plan complet 2^4, et on sélectionne les essais dont l'interaction d'ordre 4 est positive, on constate que l'interaction TPC et le facteur H ont le même signe pour chaque essai sélectionnés, d'où on choisit H=TPC comme alias ou contraste initiale.

Les règles de Box nous permet de conclure le générateur d'alias I=TPCH ce générateur nous permet de trouver tous les autre alias on a alors :

T=PCH P=TCH C=TPH TP=CH TC=PH PC=TH

Ce regroupement nous permet d'écrire :

$$y = (a_0 + a_{1234}) + (a_1 + a_{234})x_1 + (a_2 + a_{134})x_2 + (a_3 + a_{124})x_3 + (a_{12} + a_{34})x_1x_2 + (a_{13} + a_{24})x_1x_3 + (a_{23} + a_{14})x_2x_3 + (a_4 + a_{123})x_3x_2x_3$$

La matrice des effets s'écrit alors :

Essai	I=TPCH	T=PCH	P=TCH	C=TPH	TP=CH	TC=PH	PC=TH	TPC=H	y
1	1	-1	-1	-1	1	1	1	-1	89
2	1	1	-1	-1	-1	-1	1	1	68
3	1	-1	1	-1	-1	1	-1	1	73
4	1	1	1	-1	1	-1	-1	-1	85
5	1	-1	-1	1	1	-1	-1	1	55
6	1	1	-1	1	-1	1	-1	-1	115
7	1	-1	1	1	-1	-1	1	-1	82
8	1	1	1	1	1	1	1	1	70
Alias	L1	L2	L3	L4	L5	L6	L7	L8	

On écrit alors :

L1= $a_0 + a_{1234}$ L5= $a_{12} + a_{34}$

L2= $a_1 + a_{234}$ L6= $a_{13} + a_{24}$

L3= $a_2 + a_{134}$ L7= $a_{23} + a_{14}$

L4= $a_3 + a_{124}$ L8= $a_{123} + a_4$

Le calcul de ses contrastes nous donne :

L1=79.625 L3 =-2.125 L5=-4.875 L7=-2.375

L2=4.875 L4=0.875 L6=7.125 L8=-13.1

Le contraste L4 peut être considéré comme nul.

Pour déterminer les coefficients du modèle on fait appel aux hypothèses d'interprétation des alias :

Hypothèse 1 : les interactions d'ordre égal ou supérieur à trois peuvent être négligées.

D'où on a : a_0 =L1=79.625 ; a_1 =L2=4.875 ; a_2 =L3=-2.125 ; a_3 =L4≈0 ;

a_4 =L8=-13.125

Hypothèse 2 : si un contraste est faible, tous les termes aliasés dans ce contraste sont négligeables.

Hypothèse 3 : si deux effets de facteurs sont négligeables, leur interaction l'est aussi.

Hypothèse 4 : une interaction comportant deux facteurs dont l'un a un effet négligeable, est généralement négligeable

Puisque a_3 est pratiquement faible, alors d'après cette hypothèse on peut écrire :

$a_{13} = a_{23} = a_{34} = 0$ Il s'ensuit alors que :

a_{12} = L5 = -4.875 a_{14} = L7 = -2.375 et a_{24} = L6 = 7.125

Le modèle en unité codée s'écrit alors :

$$y = 79.625 + 4.875x_1 - 2.125x_2 - 13.125x_4 - 4.875x_1x_2 - 2.375x_1x_4 + 7.125x_2x_4$$

Sous réserve de la validité du modèle, on constate que le facteur 3 est nul ; la masse du catalyseur n'a alors aucune influence sur le rendement de la réaction étudiée, ainsi que sur les autres facteurs.

Dans cette interaction chimique le pH a l'effet le plus important (-13.125), cet effet est négatif ce qui signifie que le rendement baisse quand ce facteur passe de son niveau bas à son niveau haut, on fixe alors le pH à son niveau bas.

L'effet	valeurs
Température	4.875
Pression	-2.125
Catalyseur	0
Le pH	-13.125
Interaction TP	-4.875
Interaction TC	0
Interaction TH	-2.375
Interaction PC	0
Interaction PH	7.125
Interaction CH	0

Au contraire, la température agit positivement sur le rendement, ce dernier croitra lorsque celle-là passe de niveau bas au niveau haut, on fixe alors la température à son niveau haut.

L'interaction entre ces deux facteurs est négative (TH=-2.375) ce qui confirme cette influence opposée de la température et de la pression sur le rendement chimique ; c'est la même chose pour le facteur pression qui influe négativement sur le rendement, et qui agit positivement avec le pH (interaction PH=7.125) et négativement avec la température (interaction TP=-4.875), on fixe alors le facteur pression à son niveau bas.

Sous ces conditions le rendement vaut 115 mais on peut avoir d'autres valeurs soit grandes soit petites en cherchant les maximas de la fonction polynomiale :

$$y = 79.625 + 4.875x_1 - 2.125x_2 - 13.125x_4 - 4.875x_1x_2 - 2.375x_1x_4 + 7.125x_2x_4$$

Et les points où ils sont atteints, tout en conservant ceux appartiennent au domaine d'étude.

Conclusion :

Dans ce document nous avons présentés des généralités sur deux méthodes les plus célèbres dans le domaine de la recherche expérimental : l'analyse en composante principale et les plans d'expériences.

La première consiste à étudier un grand nombre de données non structurées afin d'en tirer un maximum d'information en un minimum de temps, ce qui semble merveille, mais ce procédé est imparfait dans la mesure que les variables étudiés doivent être quantitatives et de point de vue de la perte d'information dû à la déformation du nuage de points par la projection.

La deuxième permet d'étudier l'influence de plusieurs variables sur une grandeur physique tout en effectuant un minimum d'essais. Les plans mentionnés dans ce travail sont des plans où chaque facteur prend deux niveaux, ils sont basés sur la méthode de Box et Hunter qui consiste à diviser le nombre d'essais possibles sur deux, ce qui est déraisonnable lorsque le nombre des facteurs étudiés augmente.

Pour remédier à ce problème d'autres plans ont été mises en place que tel que les plans de TAGUCHI, de Placket et les carrés Gréco-Latin qui étudie plusieurs facteurs chacun d'eux prend plusieurs niveaux ; la compréhension de ces dernier plans nécessite des notions complexes dépassent le cadre de ce travail.

Bibliographie

[1] KARAM Sandrine, Application de la méthodologie des plans d'expériences et de l'analyse de données à l'optimisation des processus de dépôt, thèse doctorale 2004 université de Limoges.

[2] EL MEROUANI Mohamed, cour magistrale d'ACP et d'AFC 2014.

[3] Pierre-Louis GONZALEZ, Analyse en Composantes Principales.

[4]Alain MORINEAU, ACP-analyse en composantes principales, www.deenov.com.

[5] André Bouchier, l'analyse des données à l'usage des non mathématicien, deuxième partie : L'analyse en composantes principales, AGRO. M-INRA-formation permanente, janvier 2006.

[6] Alain BACCINI, statistique descriptive multidimensionnelle (pour les nuls) mai 2010, Institut de Mathématiques de Toulouse

[7] Thierry BLAYAC analyse de données, 2011-2012, université Montpellier 1.

[8] Gilbert Saporta, Probabilité Analyse des données et Statistique, mars 2006 édition THECNIP.

[9] Arnaud MARTIN, L'Analyse de données, polycopiés de cours, ENSEITA, 2004.

[10] Philippe TRIBOULET, notions de base sur les plans d'expériences, 2008

[11] Philippe ALEXIS, cours de plan d'expérience (sous forme littérale), décembre 2015

[12] R.BCHITOU, Chimiométrie ; cours de master chimie, université MOHAMMED V Agdal Rabat.

[13] Lionel GENDRE- Arnaud SAVARY –Bruno SOULIER, les plans d'expériences, ENS CACHAN décembre 2009

[14] Claude HOINARD, les plans factoriels complets, cours et énoncés d'exercices, mai 2009.

[15] Husson François, cours sur la planification expérimentale, les plans fractionnaires.

[16] Frédéric Bertrand, plans factoriels complets, plans fractionnaires : cas des facteurs ayant deux modalités ; université de Strasbourg, mars 2012.

[17] Jacques Goupy- Lee Creighton, Introduction aux plans d'expériences, $3^{ème}$ édition ; l'usine nouvelle, collection DUNOD.