RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

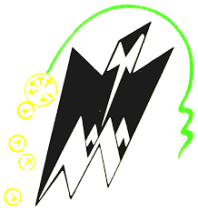
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI DE TIZI-OUZOU

FACULTÉ DES SCIENCES

DÉPARTEMENT MATHÉMATIQUE



En vue de l'obtention du Diplôme de Master en
Mathématiques appliquées à la gestion des entreprises

Thème :

Application de l'Analyse en Composantes Principales pour les données 2015 de NAFTAL

Présenté par :

AIT OUARAB Essaid

YAMOUTENE Hamza

Encadré par :

Dr. TALEB Youcef

Devant le membre de juré :
Pr.AIDENE Mouhamed
Pr.HAMADOUCHE Djamel

Représentant NAFTAL :
Mr.HARBANE Slimane

Le 30 septembre 2019

Table des matières

"Il faut avoir beaucoup étudié pour savoir peu."
— Montesquieu

"Chaque science, chaque étude, a son jargon inintelligible,
qui semble n'être inventé que pour en défendre les approches."
— Voltaire

" Il faut douter de toute chose "
— R.Descartes (1596-1650)

Liste des tableaux

Table des figures

Remerciement

Avant toute chose nous remercions Allah le tout puissant de nous avoir accordé la force et les moyens pour pouvoir réaliser ce travail.

Nous remercions chaleureusement et spécialement notre Promoteur **Dr Y.TALEB** qui a accepté de nous encadrer pour la réalisation de ce modeste travail, et qui nous a orienté dans notre travail. On exprime nos profonds remerciements pour son aide précieuse, sa disponibilité, son écoute ses conseils, sa compréhension tout au long de notre stage et la bonne ambiance de travail qu'il a su créer.

Nous tenons à exprimer nos sincères remerciements à **Mr S.HARBANE** notre encadreur au sein de l'entreprise NAFTAL, qui a manifesté un intérêt particulier à ce sujet.

Nous exprimons également notre gratitude à **Pr DJ.Hamadouche**,**Pr M.Aidene** , qui ont accepté d'examiner ce travail.

Nous ne saurons oublier le grand mérite des enseignants qui ont contribué à notre cursus particulièrement ceux du département "recherche opérationnelle" et qu'ils trouvent ici le témoignage de notre profonde reconnaissance.

Enfin, que toute personne qui, d'une façon ou d'une autre, a contribué à la réalisation de cette étude, trouve ici le témoignage de nos plus vives gratitudes.

Dédicace

Je dédie ce travail à :

. Mes chers parents, ma mère et mon père pour leurs sacrifices et leur soutiens tout au long mes études ;

. À mes grands mères ;

. À mes sœurs et frères ;

. À mes cousins et cousines .

. À mes oncles et tantes .

. À mon binôme Mr YAMOUTENE Hamza .

. À mes amis (es) avec qui j'ai vécu des beaux moments :
AS.Rafik,BM.Massinissa,A.Mohand, BAE.Mohand,B.Younes,H.Kamel,B.Kader,
B.Azzedine,M.ABDENOUR,S.Louiza,M.mohammed amokrane ,
AI.Juba,Hamza,F.Boukais,R.Nourdine.

. Sans oublier mes collègues de la ligue d'athlétisme de Tizi Ouzou
les membres de club ISWI,mes athlètes .

AIT OUARAB Essaid.

Dédicace

<u>Je dédie ce travail à :</u>

. Mes chers parents, ma mère Zina et mon père Mohammed pour leurs sacrifices et leur soutiens tout au long mes études ;

. À ma très chère fiancée $S.M$ et sa famille ;

. À mes sœurs et frères ;

. À mes belles sœurs et beaux frères.

. À mon binôme Mr AIT OUARAB Essaid (Brahim).

. À mes amis avec qui j'ai vécu des beaux moments au cours de mon cursus à l'université : Z.Farhat, Z.mohamed, Massinissa, L.Laid, JM.Abdelkader, L.Amirouche, K.AlaaEdin, M.AbdRaouf, G.Mohamed, S.Rabah, Z.Mohamed, B.hassan.

. Sans oublier mon chardonneret "Nitcha".

YAMOUTENE Hamza.

Notations

Notations

n représente le nombre d'individus.
p représente le nombre de variables quantitatives.
$\mathbb{R}^p$ L'espace individus.
$\mathbb{R}^n$ L'espace des variables.
Les tableaux :
X tableaux de données brutes. X^j la variable j.
Y tableaux de données centrés. Y^j la variable centrée j.
Z tableaux de données centrés réduit. Z^j la variable centrée réduite$j$.
e_i l'individu i.
e_{i_c} L'individu i de valeurs centrées.
$e_{i_{cr}}$ L'individu i de valeurs centrées réduites.
F_q le sous espace vectoriel sur lequel on projette les données $q < p$.
x_i^j La valeur de l'individu i pour la variable j.
D_p La matrice diagonale des poids (n, n). Si les poids sont égaux : $D_p = \frac{1}{n} I_n$.
g Le centre de gravité (le point moyen).
$\overline{x^j}$ Représente la moyenne arithmétique de la j^{eme} variable.
s_j L'écart type de la variable j.
s_j^2 La variance de la variable j.
$\mathbb{1} = (1, \ldots, 1)$ Désigne un vecteur de $\mathbb{R}^n$.
Les métriques :

M Désigne une métrique quelconque I_n, $D_{\frac{1}{s}}$ ou $D_{\frac{1}{s^2}}$.
I_n La matrice identité (diagonale) d'ordre n.
$D_{\frac{1}{s}}$ Désigne une matrice diagonale des inverses des écarts types d'ordre n.
$D_{\frac{1}{s^2}}$ Désigne une matrice diagonale des inverses des variances d'ordre n.
$d_M^2(e_i, e_{i'})$ La M distance entre les 2 individus $(e_i, e_{i'})$.
T Matrice carré inversible d'ordre p.
V Matrice des variances covariances (carrée symétrique d'ordre p).
$V(X^j)$ La variance de la variable X^j.
MVM est appelée matrice d'inertie du nuage.
$Cov(X^j, X^{j'})$ La covariance entre les deux variables $(X^j, X^{j'})$.
R Matrice des corrélations (carrée symétrique d'ordre p).

$r_{p,p-1}$ Le coefficient de corrélation entre les variables $(p, p-1)$.
I_g Indique l'inertie au centre de gravité g.
I_a Indique l'inertie en un point a.
I_Δ Indique l'inertie par rapport à l'axe Δ passant par g.
I_F Indique l'inertie par rapport au sous-espace F.
w_i est la coordonnée de l'individu e_i sur l'axe Δ_k.
W_i est le vecteur des coordonnées de l'unité e_i sur l'axe Δ_k.
N Le nuage des points.
u_{qj} est la $j$ème coordonnées du vecteur directeur unitaire u_q de Δ_q.
Λ est la matrice diagonale des valeurs propres de VM.
D_k vecteur de coordonnées des variables sur le k^{eme} axe factoriel du nuage de points variables.
d_{kj} La j^{eme} coordonnée de la k^{eme} variable sur le k^{eme} axe factoriel du nuage de points variables.
M^{-1} est la matrice inverse de M.
$cos^2(\theta_{ik})$ la qualité de la projection d'un individu e_i sur l'axe Δ_k.

Introduction

Introduction

L'analyse des données est une des branches les plus vivantes de la statistique. Ses principales méthodes se séparent en deux groupes :

- Les méthodes de classification,
- Les méthodes factorielles.

Les méthodes de classification visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes.

Les méthodes factorielles cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l'algèbre linéaire et donnant lieu à des représentations graphiques dans lesquelles les objets à décrire se transforment en des points sur des axes et des plans.

Les principales techniques factorielles sont :

L'analyse en composantes principales (Hotelling, 1933) qui analyse un ensemble de données (observations) faites sur un ensemble de variables quantitatives (numériques).

L'analyse des correspondances (Benzekri, 1964) qui est une technique de base pour analyser des tables de contingence qui peut être utilisé pour des variables qualitatives ou quantitatives positives de nature très divers.

L'analyse canonique.(Hotelling) qui contient à la Régression multiple et l'analyse discriminante comme des cas particulier.

Les techniques factorielles de l'analyse des données ont une partie de fondement générale commune à toutes : c'est celle qui s'appelle " L'analyse générale ", qui est basée sur les idées développées jadis par Eckart et Young (1936), qu'aujourd'hui elles sont développées encore plus théoriquement, surtout de point du vue informatique dans les dernières années et elles construisent ce qu'on appelle " Approximation d'une matrice par d'autres de rang inférieur ", qui est basée sur la théorie générale de décomposition singulières d'une matrice (Singular Value Des composition (SVD)).

Les méthodes d'analyse de données ont commencées à être développées dans les années 50 poussées par le développement de l'informatique et du stockage des données qui depuis n'a cessé de croître. L'analyse de données a surtout été développée en France par J.P. Benzécri qui a su par l'analyse des correspondances représenter les données de manière simple et interprétable. Il décrit l'analyse de données selon cinq principes, un peu désuets aujourd'hui :

1^{er} - principe : Statistique n'est pas probabilité.

2^{eme} - principe : Le modèle doit suivre les données et non l'inverse.

3^{eme} - principe : Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions.

4^{eme} - principe : Pour l'analyse des faits complexes et notamment des faits sociaux, l'ordinateur est indispensable.

5^{eme} - principe : Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique.

Ces cinq principes montrent bien l'approche d'une part de la statistique à la différence des probabilités - les modèles doivent coller aux données - et d'autre part de l'analyse de données - il faut traiter le plus grand nombre de données simultanément ce qui implique l'utilisation de l'ordinateur et ainsi l'utilisation de nouvelles techniques adaptées. L'analyse de données fait toujours l'objet de recherche pour s'adapter à tout type de données et faire face à des considérations de traitements en temps réel en dépit de la quantité de données toujours plus importante.

Les méthodes développées (l'analyse de données) sont maintenant souvent intégrées avec des méthodes issues de l'informatique et de l'intelligence artificielle (apprentissage numérique et symbolique) dans le data mining traduit en français par (fouille de donnée) ou encore extraction de connaissance à partir de données

Aujourd'hui les méthodes d'analyse de données sont employées dans un grand nombre de domaines qu'il est impossible d'énumérer. Actuellement ces méthodes sont beaucoup utilisées en marketing par exemple pour la gestion de la clientèle (pour proposer de nouvelles offres ciblées par exemple). Elles permettent également l'analyse d'enquêtes par exemple par l'interprétation de sondages (où de nombreuses données qualitatives doivent être prises en compte). Nous pouvons également citer la recherche documentaire qui est de plus en plus utile notamment avec internet (la difficulté porte ici sur le type de données textuelles ou autres). Le grand nombre de données en météorologie a été une des premières motivations pour le développement des méthodes d'analyse de données. En fait, tout domaine scientifique qui doit gérer de grande quantité de données de type varié ont recours à ces approches (écologie, linguistique, économie, etc) ainsi que tout domaine industriel (assurance, banque, téléphonie, etc). Ces approches ont également été mises à profit en traitement du signal et des images, où elles sont souvent employées comme pré traitements (qui peuvent être vus comme des filtres). En ingénierie mécanique, elles peuvent aussi permettre d'extraire des informations intéressantes sans avoir recours à des modèles parfois alourdis pour tenir compte de toutes les données.

Dans ce mémoire, On a élaboré un plan de travail qui se compose de deux parties, une partie théorique et une partie pratique. Dans un premier lieu on a cité les différents rappels algébriques et principes d'analyse en composantes principales (ACP), en seconde partie l'application de l'ACP pour les nos données récupérée auprès de l'entreprise NAFTAL 2015.

Chapitre 1 : Présentation de l'entreprise

Chapitre 1

Présentation de l'entreprise

1.1 Introduction

[1] NAFTAL est une société par actions (SPA) au capital social de 40 000 000 000 DA. Fondée en 1982 et filiale à 100% du Groupe Sonatrach, elle est rattachée à l'activité commercialisation. Elle a pour mission principale, la distribution et la commercialisation des produits pétroliers et dérivés sur le marché national. A l'ère de la mondialisation, NAFTAL a jugé indispensable la mise en place d'une nouvelle organisation par ligne de produit (bitumes, lubrifiants, réseau, logistique, GPL, pneumatique, Aviation, Marine).

1.2 NAFTAL filiale de SONATRACH

NAFTAL, société nationale de commercialisation et de distribution des produits pétroliers, filiale de SONATRACH a été créée en 1987 Décret n°87-190 . Sa mission essentielle consiste à distribuer et à commercialiser des produits pétroliers sur le marché national.

Elle intervient en qualité d'intermédiaire entre les fournisseurs nationaux et étrangers (raffineurs, manufacturiers et autres producteurs) et les utilisateurs de produits pétroliers implantés essentiellement en Algérie bien que depuis l'année 2002, elle cherche à s'internationaliser en essayant de pénétrer les marchés de certains pays limitrophes.

La distribution consiste à s'approvisionner, stocker, vendre et acheminer le produit vers le client en vue de son utilisation. C'est ainsi que NAFTAL assume deux grandes fonctions :

- La fonction logistique : qui comprend la circulation de tous les flux physiques du producteur à l'utilisateur : transport, livraison, stockage et manutention.

- La fonction commerciale : La fonction commerciale qui englobe la gestion du réseau, la vente, les actions promotionnelles et la gestion de la force de vente.

L'entreprise Entreprise nationale de raffinage et de distribution de produits pétroliers ERDP/NAFTAL Créée le 6 avril 1981 par décret N°80-101 a été constituée par le transfert des structures, moyens et

1. Site internet www.naftal.dz

biens, activités et personnel détenus et gérés auparavant par SONATRACH. L'ERDP/NAFTAL est entrée en activité le 1^{er} janvier 1982.

Sa mission consistait à prendre en charge le raffinage et la distribution des produits pétroliers en Algérie. En 1987, elle a connu une autre restructuration instituée par le décret N°87-189 du 27 août 1987 et qui s'est concrétisée par la création de deux entreprises :

NAFTEC : chargée du raffinage du pétrole.

NAFTAL : chargée de la distribution et de la commercialisation des produits pétroliers sur le marché national.

NAFTAL a bénéficié du monopole de la distribution des produits pétroliers de la date de sa création jusqu'à la fin des années 90 bien que la libéralisation de la distribution des produits pétroliers a débuté de manière effective quelques années auparavant avec le lancement des unités de fabrication des bitumes et par l'importation de pneumatiques par des privés nationaux dès 1991.

Avec la promulgation du décret N°97-435 du 17 novembre 1997 qui dans son article N°04 énonce que toutes personnes physiques ou morales peuvent exercer les activités de stockage, de distribution des produits pétroliers, de conditionnement des GPL et de transformation des bitumes, un nouveau cadre juridique a été tracé par les pouvoirs publics mettant fin à toute monopolisation du marché.

Cette libéralisation a été ensuite élargie par le décret N°04-89 du 22 mars 2004 permettant à toute personne physique ou morale d'exercer l'activité de fabrication des lubrifiants.

Depuis 1999, de multiples intervenants nationaux et étrangers se sont impliqués dans la distribution et la commercialisation des carburants, des GPL, des lubrifiants, des bitumes et des pneumatiques c'est-à-dire en exerçant une fonction identique à celle de NAFTAL.

C'est ainsi que NAFTAL se retrouve aujourd'hui, dans un nouveau contexte de libre concurrence marqué de surcroît, par les nouvelles dispositions de la loi N°05-07 du 28 avril 2005 relative aux hydrocarbures, par l'application des mesures énoncées par les accords d'association avec l'Union Européenne et par les préparations pour l'adhésion de l'Algérie à l'Organisation Mondiale du Commerce (OMC).

Désormais, la survie de NAFTAL dépend de sa capacité d'adaptation aux tendances d'un environnement dans lequel la mondialisation des marchés, la globalisation, la difficulté accrue de maîtrise des besoins de la clientèle, l'essor des NTIC et l'économie fondée sur le savoir. Ils constituent de plus, des phénomènes interdépendants entraînant dans leur sillage de nouveaux enjeux et de nouveaux défis.

1.3 L'offre de NAFTAL

NAFTAL pratique une politique de distribution dite extensive c'est-à-dire qu'elle s'attache à couvrir l'ensemble du territoire national. Son offre est très diversifiée. Elle est composée de

plusieurs gammes de produits et services.

Les carburants « terre » :

Il existe cinq types de carburants « terre » :

- Essence normale.
- Essence super.
- Essence Sans Plomb.
- Gas-oil.
- GPL/Carburant.

Les carburants Aviation :

- Le carburéacteur Jet A1.
- Le Kérosène (Jet déclassé).
- L'essence AVGAS 100LL.

Les carburants Marine :

- Le fuel-oil Bunker C.
- Le fuel-oil BTS.
- Le gas-oil.

Les gaz Pétrole Liquéfiés - GPL :

- Le butane conditionné.
- Le butane vrac.
- Le propane conditionné.
- Le propane en vrac.

Les lubrifiants :

- Les huiles-moteurs diesel.
- Les huiles-moteurs essences.
- Les huiles de transmission.
- Les huiles industrielles.
- Les huiles spéciales automobiles.

Les graisses :

- Les lubrifiants et produits spéciaux synthétiques pour moteurs d'avions.
- Les lubrifiants marins.

Les produits spéciaux :

- La paraffine.
- Les huiles aromatiques.
- Les essences spéciales.
- Le white spirit petroleum.
- Le toluène.
- Le xylène.

- Le methmix (aviation).

Les bitumes :

- Les bitumes purs.
- Les bitumes oxydés.
- Les bitumes fluidifiés.
- Les émulsions de bitumes.

Les pneumatiques :

- Le pneumatique « tourisme ».
- Les pneumatiques « poids lourds ».
- Les pneumatiques « véhicules utilitaires ».
- Les pneumatiques « moyens de manutention ».
- Les pneumatiques « tracteurs agricoles ».
- Le pneumatique « génie civil ».

Prestations de service :

- Services de vidange - lavage – graissage.
- Services de maintenance des équipements et installations (volucompteurs, cuves, citernes...).
- Installations d'équipements de distribution.

1.4 Le marché national des produits pétroliers

Le marché étant défini comme l'ensemble des clients actuels et potentiels capables et désireux de procéder à l'échange des produits et services. Pour NAFTAL, il s'agit de l'ensemble des utilisateurs nationaux voire étrangers des produits pétroliers et des services qui leur sont liés.

Le marché national peut être segmenté en plusieurs secteurs et entités :

Produits pétroliers	marchés (utilisateurs)
Carburants « terre »	Usagers de la route (automobilistes, transporteurs) Producteurs d'électricité (fuel)
Carburants aviation	Compagnies aériennes Ministère de la Défense Nationale Sûreté Nationale Protection civil
Carburants marine	Compagnies de navigation (armateurs) Ministère de la Défense Nationale Entreprises de pêche, artisans pêcheurs...
GPL	Ménages, commerçants, entreprises industrielles, Hôtels, collectivités locales, établissements hospitaliers, établissements scolaires et universitaires, institutions militaires, de sécurité, protection civile, agriculteurs, apiculteurs, restaurants, aviculteurs...
Lubrifiants	Usagers de la route, entreprises industrielles, Compagnies aériennes et navigation, entreprises de pêche...
Bitumes	Entreprise de travaux publics et de construction de routes Collectivités locales Fabricants de produits d'étanchéité
Produits spéciaux	Entreprises industrielles Cie aériennes Cie de Navigation
Pneumatiques	Usagers de la route Manutentionnaires Agriculteurs Entreprises de BTP

TABLE 1.1 – Les secteurs des produits

1.4.1 Le réseau national de distribution

Le réseau national de distribution des produits pétroliers comprend trois étapes qui sont :

- L'approvisionnement : C'est une relation entre la source et le centre de stockage primaire (entrepôts).C'est l'action d'acheminer des produits pétroliers d'une raffinerie vers un centre primaire soit par pipe ou par capotage (bateau).

- Le ravitaillement : C'est le transfert du stock entre l'entrepôt et les centres de stockages secondaires (dépôts). C'est l'action d'acheminer des produits pétroliers d'un centre primaire vers un centre secondaire soit par rail (train) ou par des camions (wagon citerne). Les dépôts n'ont aucune liaison avec les raffineries et chaque entrepôt couvre un ensemble de dépôts.

- La livraison : C'est une phase finale qui intervient au niveau du réseau de distribution ; elle a pour rôle d'assurer la disponibilité des produits dans les zones de consommations (stations-services). Le transport de carburants vers les stations se fait entièrement par des camions citernes.

- Livraison directe : Consiste à livrer les produits carburants d'un centre secondaire vers les clients en utilisant des camions propre à NAFTAL. Elle se fait généralement pour les stations les plus proches de dépôts d'OUED-AISSI.

- Livraison en droiture : Consiste à livrer les produits d'un centre primaire directement vers le client. C'est-à-dire du l'entrepôt d'Alger vers le client avec des camions privés que l'entreprise à alloué. D'une manière générale les raffineries approvisionnent tous les gros.

- Consommateurs : clients industriels, aéroports, entreprises de travaux routiers, etc. Ainsi pratiquement toutes les ventes de fuel lourd et de bitumes font l'objet de livraison en droiture.

1.5 L'organisation de NAFTAL

Il faut rappeler que la mission de NAFTAL consiste à acheminer son offre composée de produits et services diversifiées telle que définie ci-dessus, des lieux de raffinage (Arzew, Skikda..) ou des ports pour certains produits en provenance de raffineries algériennes en utilisant le cabotage ou encore d'installations de raffinage et de manufactures (pneumatiques) étrangères aux nombreux utilisateurs éparpillés à travers le territoire national.

Pour accomplir ses activités, le groupe NAFTAL dispose comme tout distributeur d'un réseau de distribution assez dense, organisé de manière à satisfaire toutes les exigences de la clientèle et gérée par une Direction Générale implantée à Alger et d'Unités administratives décentralisées appelées Districts intervenant chacune dans deux à trois Wilayas de façon à couvrir l'ensemble du territoire national.

Le réseau de distribution est composé d'infrastructures et de Centres de stockage et de distribution de lubrifiants, de bitumes, de produits marine (pour les Districts situés dans les zones côtières), de produits aviation (pour chaque Aéroport civil implanté au niveau du territoire national), d'entrepôts et dépôts pour le stockage des carburants, de stations-service et points de vente (magasins).

L'entreprise est structurée en plusieurs niveaux :

- Assemblée Générale composée d'un seul actionnaire SONATRACH qui est propriétaire à 100% des actions de NAFTAL.
- Conseil d'Administration comprenant un Président (PDG de l'entreprise), des membres issus de la société mère SONATRACH et d'un représentant syndical).
- Président Directeur Général et son staff composé de Conseillers Principaux et de Conseillers. Branches (Commercialisation, Carburants, GPL, Activités Internationales).

- Directions Exécutives (Finances et Comptabilité, Ressources Humaines, Stratégie, Planification, Economie SPE).

- Directions Centrales (Audit, Procédures et Contrôle de gestion, Hygiène, Sécurité, Environnement, Qualité HSEQ).

- Directions (Administration Générale, Affaires Sociales et Culturelles).

- Les Branches sont considérées comme des structures opérationnelles et organisées elles-mêmes en plusieurs niveaux :

Niveau central : Directions d'activités et de Produits, Départements et services.

Niveau décentralisé : Districts (Unités administratives), Centres et Antennes administratives au niveau de chaque Wilaya.

Il s'agit d'une organisation fortement hiérarchisée, conçue selon les principes dictés par le taylorisme.

Il existe 21 Districts rattachés à la Branche Commercialisation et 19 Districts relevant de la Branche GPL. Pour conditionner le GPL en bouteilles de 13 Kg et 3 Kg et le propane en Bouteilles de 35 Kg, NAFTAL possède des Centres d'enfûtage (carrousel de conditionnement).

L'organisation de NAFTAL est établie pour remplir trois fonctions essentielles :

- Approvisionnement.
- Stockage des produits.
- Ventes des produits et prestations de services.

NAFTAL est dotée pour cela de moyens considérables :

- Un effectif de plus de 29900 personnes dont 3000 environ exercent à titre de temporaires. A signaler que presque 8% de l'effectif est de niveau cadre. Le taux des cadres supérieurs par rapport à l'effectif total est actuellement de 1,03%.

- Une flotte importante composée de plus de 3500 camions. Le transport des produits pétroliers est l'une des tâches essentielles de NAFTAL. Il exige souvent un certain savoir-faire pour se réaliser de manière efficace. A souligner que NAFTAL a commercialisé en 2004, 10 millions de tonnes de carburants et elle en a transporté 24 millions de tonnes. Cela montre clairement que la fonction logistique est essentielle dans les activités de l'entreprise. L'approvisionnement en carburant s'effectue en partie par voie ferroviaire. Il existe une entreprise mixte de transport de produits pétroliers.

- Société de Transport des Produits Energétiques (STPE) - dont le capital appartient pour 50% à NAFTAL et pour 50% à SNTF (Société Nationale de Transport Ferroviaire).

- Des hangars de stockage des lubrifiants et produits spéciaux.

- Des bacs de stockage des bitumes.

- Des entrepôts et dépôts de stockage des carburants.

Il faut noter que NAFTAL est propriétaire de plus de 660 stations-service dont 335 sont gérées par des tiers à titre de location (gestion libre).

Elle accomplit des tâches de grossiste de produits pétroliers autrement dit elle fournit des produits pétroliers à plus de 1250 stations-service privées et à 335 en gestion libre.

Elle joue le rôle de détaillant en commercialisant directement aux clients (gestion directe) à partir de ses propres stations- service au nombre de 329 à la date de janvier 2006.

1.6 Présentation du district

Le District commercialisation de TIZI OUZOU, se situe dans la zone industrielle d'OUED AISSI, il couvre une importante zone d'influence regroupant la wilaya de TIZI OUZOU, une partie de la wilaya de BOUMERDES et l'agence commerciale de BEJAIA.

Le district a comme mission de stockage, distribution et de commercialisé des produits pétroliers, il assure la bonne exploitation et la maintenance des infrastructures qui lui sont alliées, ainsi que le suivi et le contrôle des activités des antennes qui lui sont affiliées.

Notre stage s'est déroulé dans le département commercial de district CLP 215G.

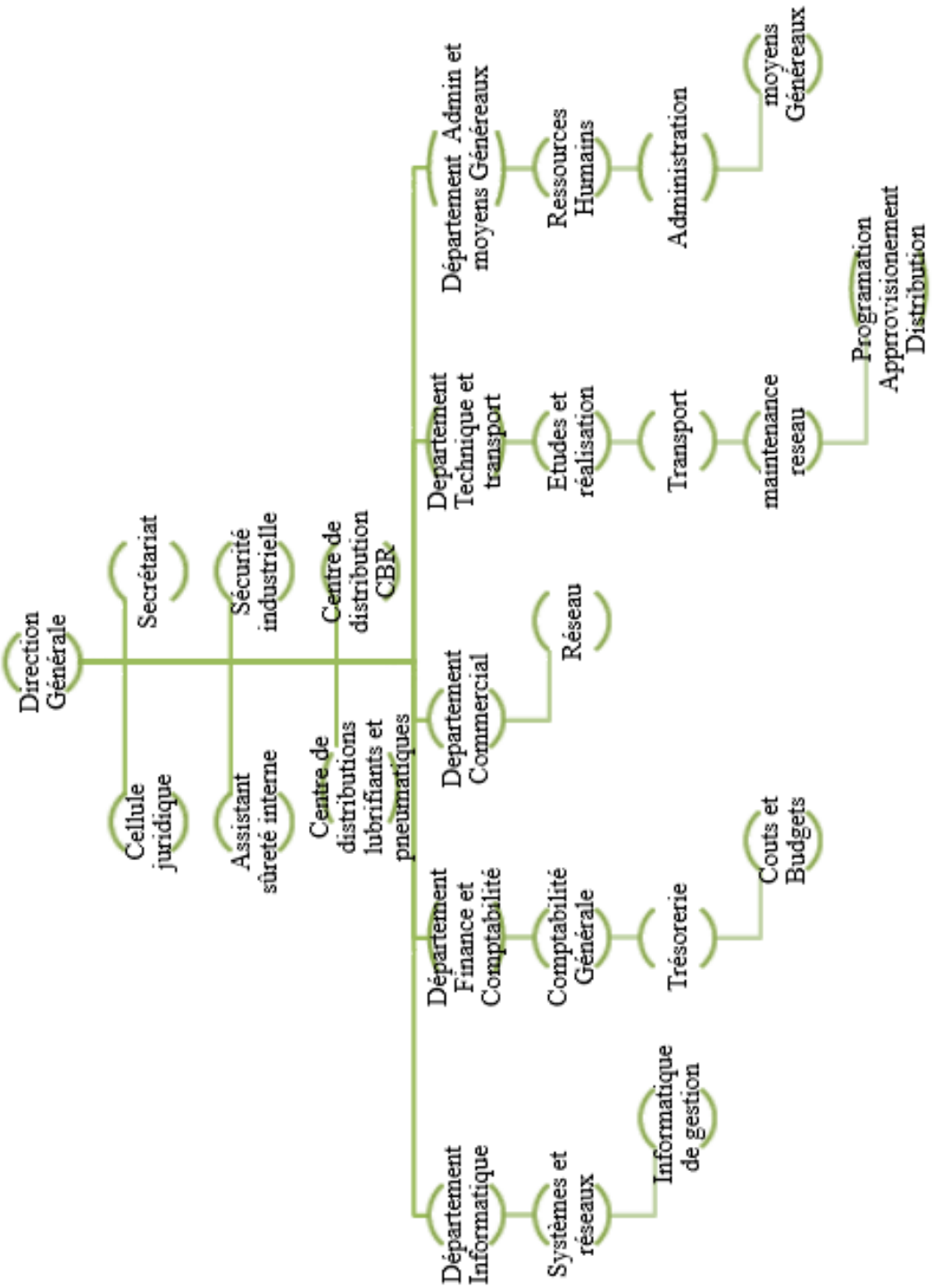


FIGURE 1.1 – L'organigramme de NAFTAL Tizi-Ouzou

1.6.1 Circuits et réseaux utilisés par NAFTAL "Oued-Aissi"

Le circuits : Le circuit de distribution du District commercialisation est comme suit :

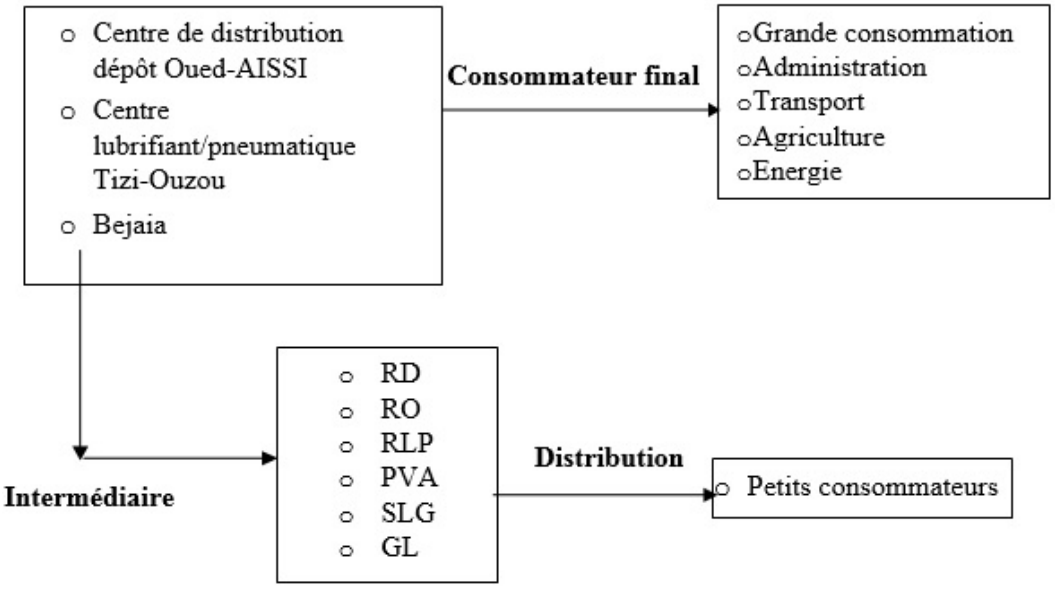


FIGURE 1.2 – Le circuit de distribution des produits d'entretiens automobile Tizi Ouzou.

La distribution directe se fait à partir des trois centres vers le consommateur final ou les clients présentés sont les secteurs suivants : Administration, transport, agriculture, énergie,(RD, RO, RLP, PVA, SLG, GL).

La distribution indirecte se fait par l'inter médiation d'un distributeur revendeur de NAFTAL comme les GD (gestion direct) et les GL (gestion libre) ou privé comme les RO (revendeur ordinaire), les RD (revendeur direct) ou les PVA (point de vente agrée), SLG (station lavage graissage), qui font le lien entre NAFTAL et les petits consommateurs.

Le réseau : L'entreprise dispose de :
- Centre multi-produit (CMP) TIZI-OUZOU
- Centre lubrifiant /pneumatique à Béjaia.
- Gérance libre.
- Gérance directe.

1.7 Identification de champs d'étude

Notre travail se base sur le service lubrifiant qui fait partie du département commercial, plus exactement le centre Lubrifiants et Pneumatique de Tizi-Ouzou.

1.7.1 Fiche technique du centre Lubrifiants et pneumatiques $215G$

Le Centre Lubrifiants et Pneumatiques 215 G est situé sur la route d'ALGER à la sortie Ouest de la ville de TIZI OUZOU vers ALGER. Il est composé de :

Plan de situation actuel du centre :
- Superficie du terrain : $10000M^2$ pour les Lubrifiants en Futs de $200L$.
- Un hangar de $3170M^2$ pour l'activité Pneumatiques.
- Bloc administratif : $100M^2$.
- poste de garde et sanitaires : $81M^2$.

Capacité de stockage :
- Lubrifiants : 800 Tonnes.
- Pneumatiques : 5000 Unités.

Effectif :
40 Agents.

CADRES	MAITRISE	EXECUTION	Effectif actuel du centre LP $215G$
06	**15**	**19**	

TABLE 1.2 – Effectifs du centre LP $215G$

Parc roulant : Le faible parc existant est insuffisant en égards aux objectifs assignés pour le centre.

En matière d'approvisionnement l'unité dispose de trois semi-remorques de capacité de 15 Tonnes chacun, chargés de couvrir les ravitaillements pneumatiques et lubrifiants.

Quant aux livraisons sont assurées par quatre camions de capacité 7 Tonnes chacun, d'un véhicule (léger) de service et deux camions pour la collecte des huiles usagées de capacité respectives $6M^3$ et $12M^3$).

L'acquisition d'un fourgon tollé pour la livraison des stations-service situées au centre-ville s'avère indispensable.

1.8 Conclusion

Étant l'unique distributeur de produits pétroliers sur le marché national, NAFTAL a pour mission la satisfaction de la demande nationale en tous produits pétroliers et vu la sensibilité de cette mission, l'objectif de NAFTAL dépasse le contexte économique pour prendre une dimension politique et sociale qui consiste à éviter toute perturbation dans sa chaine de distribution afin de bien servir ses clients dans les temps et gagner davantage de confiance de ces derniers à travers la bonne qualité de ses services.

Chapitre 2 : Rappels et compléments d'algèbre linéaire

Chapitre 2

Rappels et compléments d'algèbre linéaire

2.1 Notations

.Dans tout ce qui suit, E et F sont deux espaces vectoriels réels munis respectivement des bases canoniques $\varepsilon = e_j; j = 1, \ldots, p$ et $F = f_i; i = 1, \ldots, n$. On note indifféremment soit un vecteur de E ou de F, un endomorphisme de E, ou une application linéaire de E dans F, soit leurs représentations matricielles dans les bases définies ci-dessus.

2.2 Matrices

2.2.1 Notations

La matrice d'ordre $(n \times p)$ associées à une application linéaire de E dans F est décrite par un tableau :

$$A = \begin{pmatrix} a_1^1 & \ldots & a_1^j & \ldots & a_1^p \\ \vdots & & \vdots & & \vdots \\ a_i^1 & \ldots & a_i^j & \ldots & a_i^p \\ \vdots & & \vdots & & \vdots \\ a_n^1 & \ldots & a_n^j & \ldots & a_n^p \end{pmatrix} \tag{2.1}$$

On note par la suite :
$a_i^j = [A]_i^j$ le terme général de la matrice,
$a_i = [a_i^1, \ldots, a_i^p]'$ un vecteur-ligne mis en colonne,
$a^j = [a_1^j, \ldots, a_n^j]'$ un vecteur-colonne.

2.2.2 Types de matrices

Une matrice est dite :
- vecteur-ligne (colonne) si $n = 1 \quad (p = 1)$,
- vecteur-unité d'ordre p si elle vaut $1_p = [1, \ldots, 1]'$,
- scalaire si $\begin{cases} \text{n} = 1 \\ et \\ \text{p} = 1, \end{cases}$
- carrée si $n = p$.

Une matrice carrée est dite :
- Identité (I_p) si : $a_i^j = \sigma^j = \begin{cases} 0 \ si \ i \neq j; \\ 1 \ si \ i = j. \end{cases}$
- Diagonale si : $a_i^j = 0$ lorsque $i \neq j$,
- Symétrique si : $a_i^j = a_i^j; \forall (i; j)$,
- Triangulaire supérieure (inférieure) si : $a_i^j = 0$ lorsque $i > j \ \ (i < j)$.

2.2.3 Opérations sur les matrices

- Somme : $[A + B]_j^i = a_i^j + b_i^j$ pour A et B de même ordre $(n \times p)$.
- Multiplication par un scalaire : $[\alpha A]_i^j = \alpha a_i^j$ pour $\alpha \in \mathbb{R}$:
- Transposition : $[A']_i^j = a_i^j$; A' est d'ordre $(p \times n)$: $\begin{cases} (\text{A}')' = \text{A}\,; \\ (\text{A} + \text{B})' = \text{A}' + \text{B}'\,; \\ (\text{AB})' = \text{B}'\text{A}'\,; \end{cases}$

$$\begin{pmatrix} A_1^1 & A_2^1 \\ A_1^2 & A_2^2 \end{pmatrix}' = \begin{pmatrix} A_1^{1'} & A_1^{2'} \\ A_2^{1'} & A_2^{2'} \end{pmatrix}$$

- Produit scalaire élémentaire : $a'b = \sum_{i=1}^n a_i b_i$ où a et b sont des vecteurs-colonnes.
- Produit $[AB]_i^j = a_i' b^j$ avec $A_{(n\times p)}$; $B_{(p\times q)}$ et $AB_{(n\times q)}$, et pour des matrices par blocs, sous réserve de comptabilité des dimensions.

Soient :

$A = (a_{i,j})_{(i,j)\in[1,m]\times[1,n]} \in \text{M}_{m,n(K)}$ et $B = (b_{i,j})_{(i,j)\in[1,n]\times[1,p]} \in \text{M}_{n,p(K)}$ deux matrices telles que la largeur de la première soit égale à la hauteur de la seconde. Le produit de A par B est la matrice suivante : $AB = \sum_{k=1}^n a_{i,k} b_{k,j\,(i,j)\in[1,m]\times[1,p]} \in \text{M}_{m,p(K)}$.

Exemple 2.1. $\begin{pmatrix} A_1^1 & A_2^1 \\ A_1^2 & A_2^2 \end{pmatrix} \begin{pmatrix} B_1^1 & B_1^2 \\ B_2^1 & B_2^2 \end{pmatrix} = \begin{pmatrix} A_1^1 B_1^1 + A_2^1 B_1^2 & A_1^1 B_2^1 + A_2^1 B_2^2 \\ A_1^2 B_1^1 + A_2^2 B_1^2 & A_1^2 B_2^1 + A_2^2 B_2^2 \end{pmatrix}$

2.2.4 Propriétés des matrices carrées

La trace et le déterminant sont des notions intrinsèques, qui ne dépendent pas des bases de représentation choisies, mais uniquement de l'application linéaire sous-jacente.

Trace

Par définition, si A est une matrice $(p \times p)$: $\quad tr(A) = \sum_{j=1}^{p} a_{jj}$

Propriétés

$tr(\alpha) = \alpha$;

$tr(\alpha A) = \alpha tr(A)$;

$tr(A + B) = tr(A) + tr(B)$;

$tr(AB) = tr(BA)$; reste vrai si A est $(n \times p)$ et si B est $(p \times n)$;

$tr(CC)' = tr(C'C) = \sum_{i=1}^{n} \sum_{i=1}^{p} (c_i^j)^2$ et dans ce cas, C est $(n \times p)$.

Déterminant

Le déterminant d'une matrice carrée $A = (a_{i,j})$ d'ordre n est le nombre noté $\det(A)$ égal à :

$$\det(A) = \sum_{\sigma \in P_n} \varepsilon(\sigma) \prod_{j=1}^{n} a_{\sigma(j),j}$$

où :

P_n est l'ensemble des permutations de $1, 2, \cdots, n$;

$\varepsilon(\sigma)$ désigne la signature d'une permutation σ (égale à 1 si la permutation est paire et -1 si la permutation est impaire).

Propriétés

$|A| = \prod_{j=1}^{p} a_{jj}$; si A est triangulaire ou diagonale ;

$|\alpha A| = \alpha^p |A|$;

$|AB| = |A|\ |B|$;

$\begin{vmatrix} A & B \\ 0 & C \end{vmatrix} = |A|\,|C|$;

Exemple 2.2. Matrice 2×2 : $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$

Comatrice

$Com(A) \in \mathrm{M}_n(K)$ — ou matrice des cofacteurs — d'une matrice carrée $A \in \mathrm{M}_n(K)$ est définie par :

$$(ComA)_{i,j} = (-1)^{i+j} \det(A_{i,j}),$$

où $A_{i,j} \in \mathrm{M}_{n-1}(K)$ se déduit de A en supprimant la i-ème ligne et la j-ème colonne.

Inverse

L'inverse de A, lorsqu'elle existe, est la matrice unique notée A^{-1} telle que :

$A(A^{-1}) = (A^{-1})A = I$;

Elle existe si et seulement si $|A| \neq 0$.

Propriétés
$(A^{-1})' = (A')^{-1}$;
$(AB)^{-1} = (B^{-1})(A^{-1})$;
$|A^{-1}| = \dfrac{1}{|A|}$;

Définitions
Une matrice carrée A est dite :
- Symétrique si $A' = A$,
- Singulière si $|A| = 0$,
- Régulière si $|A| \neq 0$,
- Idempotente si $AA = A$,
- Définie-positive si, $\forall x \in \mathbb{R}^p; x'Ax \geq 0$, et si $x'Ax = 0 \rightarrow x = 0$,
- Positive, ou semi-définie-positive, si, $\forall x \in \mathbb{R}^p; x'Ax \geq 0$,
- Orthogonale si $AA' = A'A = I(A' = A^{-1})$.

2.3 Espaces euclidiens

E est un espace vectoriel réel de dimension p isomorphe à $\mathbb{R}^p$.

2.3.1 Sous-espaces

- Un sous-ensemble E_q de E est un sous-espace vectoriel (s.e.v.) de E s'il est non vide et stable : $\forall (x, y) \in E_q^2, \forall \alpha \in \mathbb{R}, \alpha(x + y) \in E_q$.

- Le q-uple $x_1, \ldots, x_q$ de E constitue un système linéairement indépendant si et seulement si :

$$\sum_{i=1}^{q} \alpha_i x_i = 0 \Rightarrow \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0.$$

- Un système linéairement indépendant $\varepsilon_q = e_1, \ldots, e_q$ qui engendre dans E un s.e.v. $E_q = e_1, \cdots, e_q$ on constitue une base et $\dim(E_q) = card(E_q) = q$.

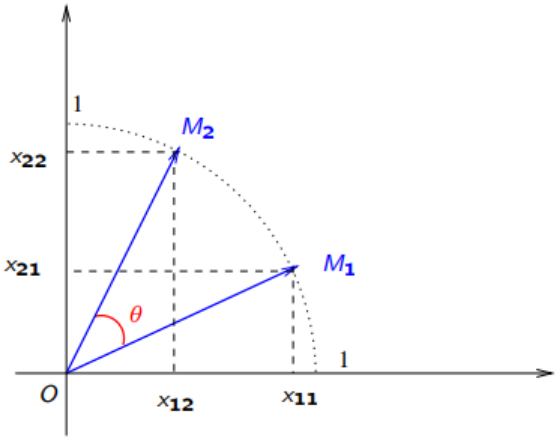
2.3.2 Produit scalaire

Un espace vectoriel de dimension p est dit euclidien, s'il est muni d'un produit scalaire qui est définit par :

$$f(x, y) =< x, y >, \forall x, y \in E.$$

où f est une forme bilinéaire, symétrique et définie positive, c'est à dire :
$\forall \alpha, \beta \in \mathbb{R}, \ \forall x, y, z \in E : < \alpha x + \beta y, z > \ \geq \ \alpha < x, z > + \ \beta < y, z >$
$< x, y >=< y, x >, \forall x, y \in E.$
$< x, x >=\| x \|^2 \geq 0, \forall x \in E - \{0\}$

$$\langle \overrightarrow{OM}_1, \overrightarrow{OM}_2 \rangle = \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$



$$\overrightarrow{OM_1} = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\overrightarrow{OM_2} = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

Norme : $\|\overrightarrow{OM_1}\| = \|\overrightarrow{OM_2}\| = 1$

FIGURE 2.1 – produit scalaire

Représentation matricielle du produit scalaire

Soit $e_1, \ldots, e_p$ une base de E, les vecteurs $x = (\alpha_1, \ldots, \alpha_p)$ et $y = (\beta_1, \ldots, \beta_p)$ s'écrivent alors : $x = \sum_{i=1}^{p} \alpha_i e_i$, $y = \sum_{i=1}^{p} \beta_i e_i$.

Grâce à la bilinéarité de f on écrira :

$$< x, y >= \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \beta_j e_i e_j.$$

La matrice M de terme général $< e_i, e_j >$ est appelée métrique. Le produit scalaire s'écrira donc :

$$< x, y >_M = x' M y$$

M est symétrique définie positive.

On dit que x et y sont $M-$ orthogonaux si $< x, y >_M = 0$

Dans ce cas :

$$\| x + y \|_M^2 = \| x \|_M^2 + \| y \|_M^2$$

2.3.3 Métrique euclidienne

Soit M une matrice carrée $(p \times p)$, symétrique, définie-positive ; M définit sur l'espace E :

- Un produit scalaire : $< x, y >_M = x' M y$,
- Une norme : $\| x \|_M = < x, x >_M^{\frac{1}{2}}$,
- Des angles : $\cos \theta_M(x, y) = \frac{< x, y >_M}{\|x\|_M \|y\|_M}$.
- Une distance : $d_M(x, y) = \| x - y \|_M$,

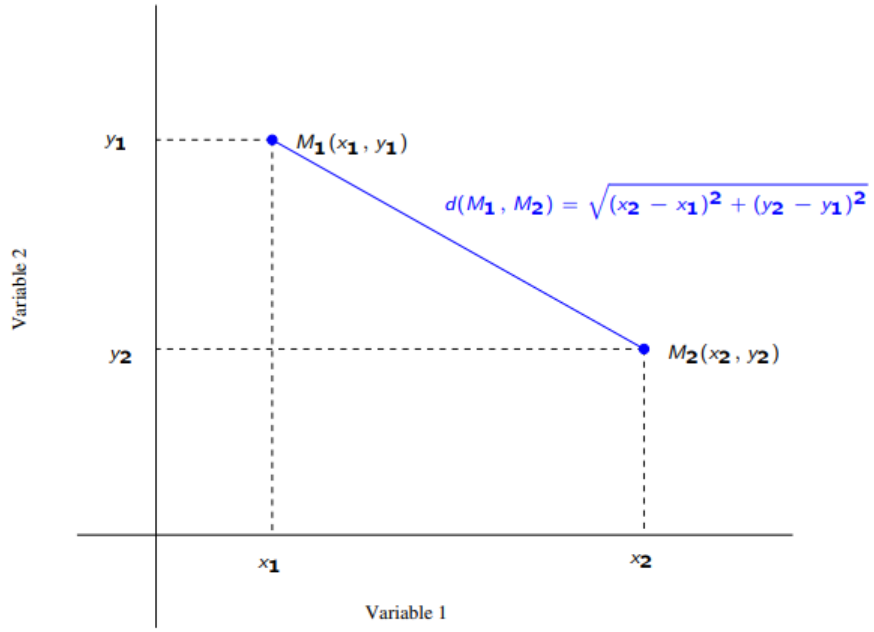


FIGURE 2.2 – distance euclidienne

La matrice M étant donnée, on dit que :
- Une matrice A est $M-$symétrique si $(MA)' = MA$,
- Deux vecteurs x et y sont $M-$orthogonaux si $< x, y >_M = 0$,
- Un vecteur x est $M-$normé si $\| x \|_M = 1$,
- Une base $\varepsilon_q = e_1, \cdots, e_q$ est $M-$ orthonormée si :

$$\forall (i,j), < e_i, e_j >_M = \delta_i^j.$$

2.3.4 Projection

Soit W un sous-espace de E et $\beta = b^1, \ldots, b^q$ une base de W; $P(p \times p)$ est une matrice de projection $M-$orthogonale sur W si et seulement si :

$$\begin{cases} \forall y \in E, Py \in W \\ et < Py, y - Py >_M = 0. \end{cases}$$

Toute matrice idempotente ($P^2 = P$) et $M-$ symétrique ($P'M = MP$) est une matrice de projection $M-$orthogonale et réciproquement.

Propriétés
- Les valeurs propres de P sont 0 ou 1 :
$u \in W$, $Pu = u, \lambda = 1$, de multiplicité $dim(W)$,
$v \perp W$, $Pv = 0, \lambda = 0$, de multiplicité $dim(W^\perp)$, (on note $(v \in W^\perp)$.)

$tr(P) = dim(W)$.
$P = B(B'MB)^{-1}B'M$, où $B = [b^1, \ldots, b^q]$.

- Dans le cas particulier où les b^j sont $M-$orthonormés : $P = BB'M = \sum_{i=1}^{i=q} b^j b^{j'} M$.

- Dans le cas particulier où $q = 1$ alors :

$$P = \frac{bb'}{b'Mb}M = \frac{1}{\| b \|_M} bb'M.$$

- Si $P_1, \ldots, P_q$ sont des matrices de projection $M-$orthogonales alors la somme $P_1 + \ldots + P_q$ est une matrice de projection $M-$orthogonale si et seulement si : $P_k P_j = \delta_k^j P_j$.

- La matrice $I - P$ est la matrice de projection $M-$orthogonale sur $W^\perp$.

2.4 Valeurs et vecteurs propres

Soit A une matrice carrée $(p \times p)$.

Définitions

- Par définition, un vecteur v définit une direction propre associée à une valeur propre λ si l'on a :

$$Av = \lambda v.$$

- Si λ est une valeur propre de A, le noyau $Ker(A - \lambda I)$ est un s.e.v. de E, appelé sous-espace propre, dont la dimension est majoré par l'ordre de multiplicité de λ. Comme cas particulier, $Ker(A)$ est le sous-espace propre associé, si elle existe, à la valeur propre nulle.

- Les valeurs propres d'une matrice A sont les racines, avec leur multiplicité, du polynôme caractéristique :

$$\mid A - \lambda I \mid = 0.$$

Théorème 2.1. *Soit deux matrices $A(n \times p)$ et $B(p \times n)$; les valeurs propres non nulles de AB et BA sont identiques avec le même degré de multiplicité.*

Si u est vecteur propre de BA associé à la valeur propre λ différente de zéro, alors $v = Au$ est vecteur propre de la matrice AB associé à la même valeur propre.

Les applications statistiques envisagées dans ce cours ne s'intéressent qu'à des types particuliers de matrices.

Théorème 2.2. *Une matrice A réelle symétrique admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base orthonormée de E ; A se décompose en :*

$$A = V \Lambda V' = \sum_{k=1}^{p} \lambda_k v^k v^{k'}$$

où V est une matrice orthogonale $[v^1, \ldots, v^p]$ des vecteurs propres orthonormés associés aux valeurs propres λ_k, rangées par ordre décroissant dans la matrice diagonale Λ.

Théorème 2.3. *Une matrice A réelle $M-$symétrique admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base $M-$orthonormée de E ; A se décompose en :*

$$A = V \Lambda V' M = \sum_{k=1}^{p} \lambda_k v^k v^{k'} M$$

où $V = [v^1, \ldots, v^p]$ est une matrice $M-$orthogonale ($V'MV = I_p$ et $VV' = M^{-1}$) des vecteurs propres associés aux valeurs propres λ_k, rangées par ordre décroissant dans la matrice diagonale Λ.

Les décompositions ne sont pas uniques : pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que pour une valeur propre multiple, une infinité de bases $M-$orthonormées peuvent être extraites du sous-espace propre unique associé.

Le rang de A est aussi le rang de la matrice Λ associée et donc le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.

Par définition, si A est positive, on note la racine carrée de A :

$$A^{\frac{1}{2}} = \sum_{k=1}^{p} \sqrt{\lambda_k} v^k v^{k'} M = V \Lambda^{\frac{1}{2}} V' M.$$

Propriétés

- Si $\lambda_k \neq \lambda_j$, $v^k \perp M^{v^j}$;
- $tr(A) = \sum_{k=1}^{p} \lambda_{k;}$ $\mid A \mid = \prod_{k=1}^{p} \lambda_k$;
- si A est régulière, $\forall k, \lambda_k \neq 0$;
- si A est positive, $\lambda_p \geq 0$;
- si A est définie positive, $\lambda_p > 0$.

Chapitre 3 : Analyse en Composantes Principales

Chapitre 3

Analyse en Composantes Principales

3.1 Introduction

L'analyse en composante principale est une technique d'analyse statistique, principalement descriptive qui travaille discrètement sur les mesures récoltées pour chaque individu. Elle consiste à représenter sous forme graphique le plus d'information possible contenue dans un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n unités. Elle permet ainsi de visualiser un espace à p dimensions à l'aide d'un espace de dimension plus petit.

Les résultats de l'ACP sont produits sous deux formes : feuilles de données (Résultats numériques) et graphiques. Alors que les feuilles de données peuvent être utilisées dans l'interprétation des résultats, les graphiques associés permettent une aide visuelle pour classer les variables et observations.

Résultats numériques : Le module ACP produit une grande variété de résultats, telles que les coordonnées factorielles des variables et observations, contributions des variables et observations, résultats factoriels, coefficients des résultats factoriels, cosinus carrés, valeurs propres, et statistiques descriptives.

Résultats graphiques : Le but principal de l'ACP est de récupérer un espace factoriel de plus petite dimension sur lequel les points originaux (variables ou observations) peuvent être projetés. Afin de la faciliter, des tracés en 2D des coordonnées factorielles peuvent être produits dans cette méthode. Cette option est disponible pour les variables et les observations. L'ACP représente également les tracés des valeurs propres de la matrice de corrélation ou covariance pour les variables actives, c'est-à-dire, le tracé des valeurs propres. Divers graphiques en $2D$ et $3D$ sont aussi disponibles pour les statistiques descriptives. De nombreuses options de représentation sont disponibles pour chaque graphique.

3.2 principe de l'ACP, Tableaux de données, et espaces associés

3.2.1 Principe de l'ACP

L'ACP est une méthode descriptive permettant de traiter des tableaux de données quantitatives multidimensionnelle X_n^p (de grandes dimension) ;

Où n représente le nombre d'individus e_i, $[e_1, e_2, \ldots, e_i, \ldots, e_n]$

Et p le nombre de variables quantitatives X^j, $[X^1, X^2, \ldots, X^j, \ldots, X^p]$.

Le but de l'ACP est de résumer la grande quantité d'informations contenues dans le tableau X_n^p, et cela dans un tableau de plus petite dimension C_n^q $(q < p)$.
Et ainsi fournir une représentation visuelle tels que :

- C^j est une combinaison linéaire des p variables quantitatives, X^j, $j = 1, \ldots, p$.
- Les variables C^j, $j = 1, \ldots, p$ sont **non corrélées** entre elles (les axes sont orthogonaux).
- Le tableau X peut être reconstitué à partir du nouveau tableau C.
- C contient le maximum d'informations sur X.

Autrement dit, on cherche à définir q nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible.

. Ces variables seront appelées : **« composantes principales »**,

. Les axes qu'elles déterminent : **« axes principaux »**,

. Les formes linéaires associées : **« facteurs principaux »**.

Remarque 3.1. *Perdre moins d'information possible veut dire :*

1. Le sous espace F_q souhaité, sur lequel on va projeter les points du nuage, devra être "ajusté" le mieux possible au nuage des individus c'est à dire la somme des carrés des distances entre les individus à F_q doit être minimale. 2.3.3,page 20

2. F_q est le sous-espace tel que le nuage projeté ait une inertie (dispersion) maximale. 2.3.4,page 21

3.2.2 Tableaux de données

Les données sont les mesures effectuées sur n individus $[e_1, e_2, \ldots, e_i, \ldots, e_n]$.
Les p variables quantitatives qui représentent ces mesures sont $[X^1, X^2, \ldots, X^j, \ldots, X^p]$.
On possède donc un tableau rectangulaire de mesure noté X dont les colonnes sont des variables et dont les lignes représentent des individus statistiques :

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^j & \ldots & x_1^p \\ x_2^1 & x_2^2 & \ldots & x_2^j & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^1 & x_i^2 & \ldots & x_i^j & \ldots & x_i^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \ldots & x_n^j & \ldots & x_n^p \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix} = \left(X^1, X^2, \ldots, X^j, \ldots, X^p\right)$$

e_i : L'individu i.
X^j : La variable j du tableau de données brutes.
x_i^j : La valeur de l'individu i pour la variable j.
On peut représenter chaque individus par le vecteur de ses mesures sur les p variables :

$$e_i' = [x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^p] \text{ ce qui donne : } e_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix}$$

Alors e_i est un vecteur de $\mathbb{R}^p$.
De façon analogue, on peut représenter chaque variable par un vecteur de $\mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n individus :

$$X^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{pmatrix}$$

3.2.3 Nuage de points

Nuage de points-individus

L'ensemble des points qui représentent les individus est appelé " nuage des individus ".

Le nuage des individus résume les coordonnées des n vecteurs individus e_i dans le repère de $\mathbb{R}^p$ dont les axes sont les p variables du tableau. Il permet de visualiser les ressemblances/dissemblances entre individus contenus dans le tableau de données X.

Nuages de points-variables

En faisant de même dans $\mathbb{R}^n$, chaque variable pourra être représenté par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé " nuage des variables ".

Le nuage des variables résume les coordonnées des p vecteurs variables dans le repère de $\mathbb{R}^n$ dont les axes sont déterminés par les n individus. Il permet de visualiser les liens entre les variables contenus dans le tableau de données X.

Remarque 3.2. *On ne peut pas visualiser ces représentations, puisque $\mathbb{R}^n$ et $\mathbb{R}^p$ sont de dimensions élevés, en générale supérieure à 3, donc on essai de trouver un espace sur lequel on projette ces données en perdant le moins possible d'information.*

3.2.4 Centre de gravité

le vecteur g le centre de gravité du nuage de point est un individu fictif, tel que :

$$g = \begin{pmatrix} \overline{x^1} \\ \vdots \\ \overline{x^p} \end{pmatrix}$$

Avec : $\overline{x^j} = \sum_{i=1}^{n} p_i x_i^j, j = 1, \cdots, p.$
Et : $(x^j)' D_p \uparrow_n$,
Ainsi :

$$g = \begin{pmatrix} \overline{x^1} \\ \vdots \\ \overline{x^p} \end{pmatrix} = \begin{pmatrix} (x^1)' D_p \uparrow_n \\ \vdots \\ (x^p)' D_p \uparrow_n \end{pmatrix}$$

D'où :

$$g = X' D_p \uparrow_n .$$

$\overline{x^j}$: représente la moyenne arithmétique de la variable j.
$\uparrow_n = (1, \ldots, 1)$ désigne un vecteur de $\mathbb{R}^n$.

Prendre g comme **origine** du nuage des points, revient alors à travailler sur le tableau de données centrées Y associé à X :

$$Y = \begin{pmatrix} x_1^1 - \overline{x^1} & \dots & x_1^j - \overline{x^j} & \dots & x_1^p - \overline{x^p} \\ x_2^1 - \overline{x^1} & \dots & x_2^j - \overline{x^j} & \dots & x_2^p - \overline{x^p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^1 - \overline{x^1} & \dots & x_i^j - \overline{x^j} & \dots & x_i^p - \overline{x^p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 - \overline{x^1} & \dots & x_n^j - \overline{x^j} & \dots & x_n^p - \overline{x^p} \end{pmatrix} = \left(Y^1, \dots, Y^j, \dots, Y^p\right)$$

Écriture matricielle :

$$\begin{aligned} Y &= X - \mathbb{1}_n \, g' \\ &= X - \mathbb{1}_n \mathbb{1}_n' \, D_p X \\ &= (I_n - \mathbb{1}_n \mathbb{1}_n' \, D_p) X. \end{aligned}$$

Avec $g' = \mathbb{1}_n' \, D_p X$.
Le vecteur des coordonnées centrées de l'individu i est :

$$e_{i_c} = \begin{pmatrix} x_i^1 - \overline{x^1} \\ \vdots \\ x_i^j - \overline{x^j} \\ \vdots \\ x_i^p - \overline{x^p} \end{pmatrix} \in \mathbb{R}^p$$

Et celui des coordonnées centrées de la variable j est :

$$Y^j = \begin{pmatrix} x_1^j - \overline{x^j} \\ \vdots \\ x_i^j - \overline{x^j} \\ \vdots \\ x_n^j - \overline{x^j} \end{pmatrix} = \begin{pmatrix} y_1^j \\ \vdots \\ y_i^j \\ \vdots \\ y_n^j \end{pmatrix} \in \mathbb{R}^n$$

Remarque 3.3. *Souvent, les données brutes x_i^j sont remplacées par les données de la forme $\frac{x_i^j - \overline{x^j}}{s_j}$ (dites centrées réduites) où $\overline{x^j}$ est la moyenne de la variable X^j et s_j est l'écart-type de la variable X^j. Le centrage permet de comparer les dispersions par rapport à un point de référence unique (la moyenne, qui vaut zéro pour la variable après centrage). En réduisant les variables, on les exprime toutes en unités d'écart-type, et on leur donne une variance égale à 1.*

$$Z = \begin{pmatrix} \frac{x_1^1 - \overline{x^1}}{s_1} & \dots & \frac{x_1^p - \overline{x^p}}{s_p} \\ \vdots & \vdots & \vdots \\ \frac{x_i^1 - \overline{x^1}}{s_1} & \dots & \frac{x_i^p - \overline{x^p}}{s_p} \\ \vdots & \vdots & \vdots \\ \frac{x_n^1 - \overline{x^1}}{s_1} & \dots & \frac{x_n^p - \overline{x^p}}{s_p} \end{pmatrix} = \left(Z^1, \dots, Z^j, \dots, Z^p \right)$$

Le vecteur des coordonnées centrées réduites de la variable j est :

$$Z^j = \begin{pmatrix} \frac{x_1^j - \overline{x^j}}{s_j} \\ \vdots \\ \frac{x_i^j - \overline{x^j}}{s_j} \\ \vdots \\ \frac{x_n^j - \overline{x^j}}{s_j} \end{pmatrix} = \begin{pmatrix} z_1^j \\ \vdots \\ z_i^j \\ \vdots \\ z_n^j \end{pmatrix}$$

3.3 Matrice de variance covariance et corrélations

3.3.1 Matrice de variance covariance

On appelle matrice de covariance empirique de p variables quantitatives $X^1, \ldots, X^j, \ldots, X^p$ mesurées sur un ensemble de n individus, la matrice symétrique à p lignes et p colonnes contenant sur sa diagonale principale les variances empiriques des p variables, et ailleurs, les covariances empiriques de ces variables deux à deux :

$$V = \begin{pmatrix} Var(X^1) & Cov(X^1,X^2) & \ldots & Cov(X^1,X^j) & \ldots & Cov(X^1,X^p) \\ Cov(X^2,X^1) & Var(X^2) & \ldots & Cov(x^2,X^j) & \ldots & Cov(X^2,X^p) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ Cov(X^j,X^1) & Cov(X^j,X^2) & \ldots & Var(X^j) & \ldots & Cov(X^j,X^p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X^p,X^1) & Cov(X^p,X^2) & \ldots & Cov(X^p,X^j) & \ldots & Var(X^p) \end{pmatrix}$$

avec

$$Var(X^j) = \sum_{i=1}^{n} p_i (x_i^j - \overline{x^j})^2$$

Et

$$s_j = \sqrt{\| Y^j \|^2_{D_p}} = \| Y^j \|_{D_p}$$

$$\begin{aligned} Cov(X^j, X^k) &= \sum_{i=1}^{n} p_i (x_i^j - \overline{x^j})(x_i^k - \overline{x^k}) \\ &= < X^j \restriction_n - \overline{X^j}, X^k - \overline{X^k} \restriction_n >_{D_p}, \\ &= < Y^j, Y^k >_{D_p} . \end{aligned}$$

avec :

$$\overline{x^j} = \sum_{i=1}^{n} p_i x_i^j$$

Écriture matricielle :
Le carré de la norme d'une variable centrée est sa variance :

$$\| Y^j \|^2_{D_p} = (Y^j)' D_p Y^j = {s^2}_j$$

Le produit scalaire entre deux variables centrées est leur covariance :

$$< Y^j, Y^k >_{D_p} = (Y^j)' D_p Y^k = Cov(Y^j, Y^k)$$

Le produit scalaire entre deux variables centrées réduites est leur coefficient de corrélation :

$$< Z^j, Z^k >_{D_p} = Cov(Z^j, Z^k) = r(Z^j, Z^k).$$

La variance de Z^j égale 1.
Écriture matricielle :

$$\begin{aligned} Z &= Y D_{\frac{1}{s}}; \\ &= (I_n - 1_n 1_n' \, D_p) X D_{\frac{1}{s}}. \end{aligned}$$

Si :

$$e_{i_c} = \begin{pmatrix} x_i^1 - \overline{x^1} \\ x_i^2 - \overline{x^2} \\ \vdots \\ x_i^j - \overline{x^j} \\ \vdots \\ x_i^p - \overline{x^p} \end{pmatrix} = \begin{pmatrix} y_i^1 \\ y_i^2 \\ \vdots \\ y_i^j \\ \vdots \\ y_i^p \end{pmatrix}$$

est le vecteur centré des p variables mesurées sur l'individus i, on peut voir que :

$$V = \sum_{i=1}^{n} p_i y_i y_i^{'} = \begin{pmatrix} \sum_{i=1}^n p_i (y_i^1)^2 & \cdots & \sum_{i=1}^n p_i y_i^1 y_i^j & \cdots & \sum_{i=1}^n p_i y_i^1 y_i^p \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n p_i y_i^j y_i^1 & \cdots & \sum_{i=1}^n p_i (y_i^j)^2 & \cdots & \sum_{i=1}^n p_i y_i^j y_i^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n p_i y_i^p y_i^1 & \cdots & \sum_{i=1}^n p_i y_i^p y_i^j & \cdots & \sum_{i=1}^n p_i (y_i^p)^2 \end{pmatrix}$$

Où $y_i^j = x_{ij} - \overline{x^j}$
On retrouve bien la matrice de covariance empirique :

$$V = \sum_{i=1}^{n} p_i y_i y_i^{'} = \sum_{i=1}^{n} p_i \begin{pmatrix} y_i^1 \\ \vdots \\ y_i^j \\ \vdots \\ y_i^p \end{pmatrix} \left(y_i^1, \cdots, y_i^j, \cdots, y_i^p \right)$$

$$V = X' D_p X - g g' = \; Y' D_p Y$$

Cette matrice est une matrice symétrique. Elle est définie positive si les p variables ne sont pas liées linéairement. On peut remarquer que sa trace est égale à la somme des variances empiriques des p variables.

Si on doit travailler avec des variables centrées et réduites, on passe du tableau des valeurs centrées au tableau des valeurs centrées et réduites : $Z = Y D_{\frac{1}{s}}$.

Avec $D_{\frac{1}{s}}$ la matrice diagonale des inverses des écarts-type empiriques des variables :

$$D_{\frac{1}{s}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \frac{1}{s_j} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \frac{1}{s_p} \end{pmatrix}$$

3.3.2 Matrice des corrélation

Si on calcule la matrice de covariance à partir d'un tableau de données centrées et réduites, on obtient la matrice des corrélations empiriques qui résume la structure des dépendances linéaires entre les p variables prises deux à deux, notée R :

$$\begin{aligned} Z'D_pZ &= D_{\frac{1}{s}}Y'D_pYD_{\frac{1}{s}}, \\ &= D_{\frac{1}{s}}VD_{\frac{1}{s}}, \\ &= R. \end{aligned}$$

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & \cdots & r_{1p} \\ r_{21} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \ddots \\ \vdots & & \cdots & \ddots & r_{p-1,p} \\ r_{p,1} & \cdots & \cdots & r_{p,p-1} & 1 \end{pmatrix}$$

Z est le tableau centré réduit.

D est la matrice des poids.

Coefficient de corrélation

Le coefficient de corrélation entre deux variables, égal au rapport entre la covariance entre ces deux variables et le produit de leurs écart-types, permet de mesurer le sens et l'intensité de la relation entre celles-ci. Si sa valeur est négative, cela signifie que d'une manière générale quand la valeur de la première variable augmente, celle de la deuxième diminue, et réciproquement.
À l'inverse si la valeur est positive, cela signifie que les deux variables augmentent et diminuent ensemble.

Tel que :

$$R = (r_{jk})j, k \in 1, \cdots, p.$$

$$r_jk = \frac{< Y^j, Y^k >_{D_p}}{\| Y^j \|_{D_p} \| Y^k \|_{D_p}} = \frac{cov(y^j, y^k)}{s^j s^k}.$$

Indépendamment du sens de la relation, la valeur absolue du coefficient permet de mesurer l'intensité de la relation entre les deux variables : plus elle est proche de 1, plus la relation est forte, plus elle s'approche de 0 moins elle n'a de signification lorsque l'on a que quelques dizaines d'individus statistiques, on ne considère en général comme significatifs que les coefficients ayant une valeur absolue supérieure à 0,5 c'est à dire compris entre -1 et $-0,5$ ou entre $+0,5$ et $+1$.

3.4 Espace des variables

Objectif : Trouver le plan de projection du nuage de points variables $\{\mathbb{N}(J) = X^j \in \mathbb{R}^n, j = 1, \ldots, p.\}$ tel que les angles entre les variables (et donc les corrélations) soient les moins déformés possible. L'espace $\mathbb{R}^n$ est muni de la métrique des des poids D_p.

3.4.1 Matrice des poids

Afin de calculer les distances entre deux variables, il est nécessaire d'attribuer des poids p_i, $i = 1, \ldots, n$ aux n individus selon l'importance que l'on souhaite leur donner.

Ces poids qui sont des nombres positifs de somme 1 comparables à des fréquen ces, sont regroupés dans une matrice diagonale D de taille n : 2.2.2 page 17

$$D_p = \begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix}$$

Dans le cas le plus usuel de poids égaux, on a bien sur $D_p = \frac{1}{n} I_n$ i.e : $p_i = \dfrac{1}{n}, \forall i = 1, \ldots, n.$ *où* I_n est la matrice identité d'ordre n.

Généralement, on définit le produit scalaire entre deux variables par :

$$< X^j, X^k >_{D_p} = (X^j)' D_p X^k = \sum_{i=1}^{n} p_i x_i^j x_i^k.$$

Le cosinus de l'angle θ_{jk} entre deux variables centrées est donné par :

$$\cos\theta_{jk} = \frac{< Y^j, Y^k >_{D_p}}{\| Y^j \|_{D_p} \| Y^k \|_{D_p}} = \frac{s_{jk}}{s_j s_k}.$$

Dans le cas variables centrées réduites, ce produit scalaire est la $Cov(Z^j, Z^k)$ car :

$$\left\langle \frac{X^j - \overline{x^j}}{s_j}, \frac{X^k - \overline{x^k}}{s_k} \right\rangle_D = \sum_{i=1}^{n} p_i \frac{x_i^j - \overline{x^j}}{s_j} \frac{x_i^k - \overline{x^k}}{s_k} = Cov(Z^j, Z^k)$$

Donc :

$$Var\left(\frac{X^j - \overline{x^j}}{s_j}\right) = \sum_{i=1}^{n} p_i \frac{x_i^j - \overline{x^j}}{s_j} \frac{x_i^j - \overline{x^j}}{s_j} = \left\langle \frac{X^j - \overline{x^j}}{s_j}, \frac{X^j - \overline{x^j}}{s_j} \right\rangle = \left\| \frac{X^j - \overline{x^j}}{s_j} \right\|^2.$$

De plus :

$$\left|\left|\frac{X^j - \overline{x^j}}{s_j}\right|\right|^2 = \frac{1}{s_j^2}\left(\sum_{i=1}^{n} p_i(x_i^j - \overline{x^j})(x_i^j - \overline{x^j})\right) = 1$$

Donc le nuage des variables est situé sur une sphère de rayon 1. De plus le cosinus de l'angle de ces deux variables n'est autre que leur coefficient de corrélation linéaire :

$$\cos\theta_{jk} = \frac{\left\langle X^j - \overline{x^j}, X^k - \overline{x^k}\right\rangle}{\| X^j - \overline{x^j} \| \| X^k - \overline{x^k} \|} = \sum_{i=1}^{n} p_i \left(\frac{x_i^j - \overline{x^j}}{s_j}\right)\left(\frac{x_i^k - \overline{x^k}}{s_k}\right)$$

L'interprétation d'un coefficient de corrélation comme un cosinus est une propriété très importante puisque elle donne un support géométrique, donc visuel, au coefficient de corrélation.

3.4.2 Graphiques associés aux variables

Cercle de corrélation

Le cercle des corrélations est la projection du nuage des variables sur le plan des composantes principales. Ce sont donc les projections des colonnes de Z sur les plans formés par ces axes.

Les axes factoriels sont :

- Des combinaisons linéaires des colonnes de Z ;
- Des vecteurs de $\mathbb{R}^n$;
- Orthogonaux 2 à 2.

Z¹ et Z² ont une corrélation proche de 1.

Z¹ et Z³ ont une corrélation proche de 0.

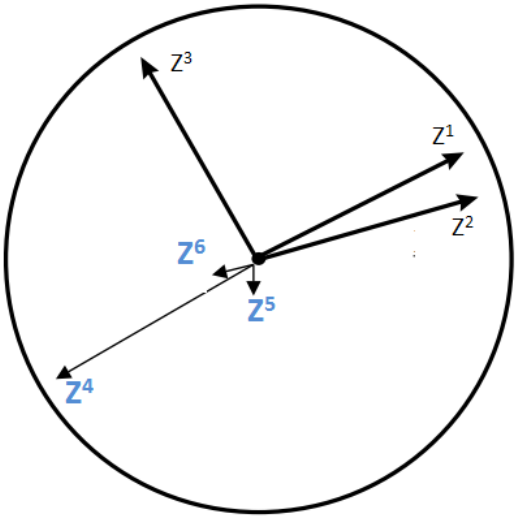


FIGURE 3.1 – cercle des correlations

Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

Si les valeurs prises par deux variables particulières sont très voisines pour tous les individus, ces variables seront représentées par deux points très proches dans $\mathbb{R}^n$. Cela peut vouloir dire que ces variables mesurent une même chose ou encore qu'elles sont liées par une relation particulière.

3.5 Espace des individus

Objectif : Trouver le plan de projection du nuage de points individus $\{\mathbb{N}(I) = e_i \in \mathbb{R}^p, i = 1, \ldots, n.\}$ tel que les distances entre les individus soient les mieux conservées possible.

Chaque individu i sera considéré comme un élément d'un espace vectoriel $\mathbb{R}^p$ (espace des individus). L'ensemble des n individus est un nuage de points de $\mathbb{R}^p$ dont le barycentre est le point g.

3.5.1 Le rôle de la métrique " le choix de la distance "

Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance utilisé par l'ACP dans l'espace où sont représentées les individus, est la distance euclidienne classique. La distance entre deux individus e_i et e_j est définie par : šection 2.3.3,page 20

$$d^2_M(e_i, e_j) = \| e_i - e_j \|^2_M = (e_i - e_j)' M (e_i - e_j)$$

Où M est une matrice symétrique de taille p définie positive ;

Et $(e_i - e_j)'$ est la transposée du vecteur $(e_i - e_j)$.

Dans le cas d'une ACP centré réduite, les métriques les plus utilisées sont les métriques diagonales des inverses des variances qui reviennent à diviser chaque caractère par son écart type (donner à chaque caractère la même importance). La métrique est :

$$M = D_{\frac{1}{s^2}} = \begin{pmatrix} \frac{1}{s_1^2} & \ldots & 0 & 0 \\ \vdots & \ddots \frac{1}{s_2^2} & & 0 \\ 0 & & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{s_p^2} \end{pmatrix}$$

Qui rend la distance entre individus invariante par transformation linéaire. Cette métrique est intéressante quand les variables sont hétérogènes. La distance entre deux individus ne dépend plus des unités de mesure car $\frac{x_i^j}{s_j}$ est sans dimension. De plus, elle donne à chaque caractère la même importance quel que soit leur dispersion. On dit que la métrique $M = D_{\frac{1}{s^2}}$ rétablit l'équilibre entre les variables. On peut décomposer cette matrice sous la forme classique $M = T'T$, où T est inversible puisque M est supposée définie positive, alors le produit scalaire est :

$$< e_i, e_j >_M = e_i' M e_j$$

peut s'écrire :

$$e_i' T' T e_j = (Te_i)'(T' e_j) = < Te_i, Te_j >_{I_p} .$$

2.2.4, 2.2.4, 2.3.2 page 18, 19, 19

Tout se passe comme si on avait transformé les données initiales du tableau X par la matrice T et utilisé ensuite le produit scalaire ordinaire. Dans le cas d'une ACP centrée non réduite,

la métrique $M = I$ elle revient à utiliser le produit scalaire usuel, elle conduit à privilégier les variables les plus dispersées pour lesquels les différences entre individus sont les plus fortes et à négliger les différences entre les autres variables.

3.5.2 Représentation des individus dans les nouveaux axes

Pour faire la représentation des individus dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir w_{ik}, coordonnée de l'individu e_i sur l'axe Δ_k, on projette orthogonalement le vecteur $\overrightarrow{ge_i}$, sur cet axe et on obtient :

$$\begin{cases} & w_{i_k} = \langle \overrightarrow{ge_i}, \overrightarrow{u_k} \rangle \\ Et & W_i = e'_{i_c} M u. \end{cases}$$

Où W_i est le vecteur des coordonnées de l'individu e_i.

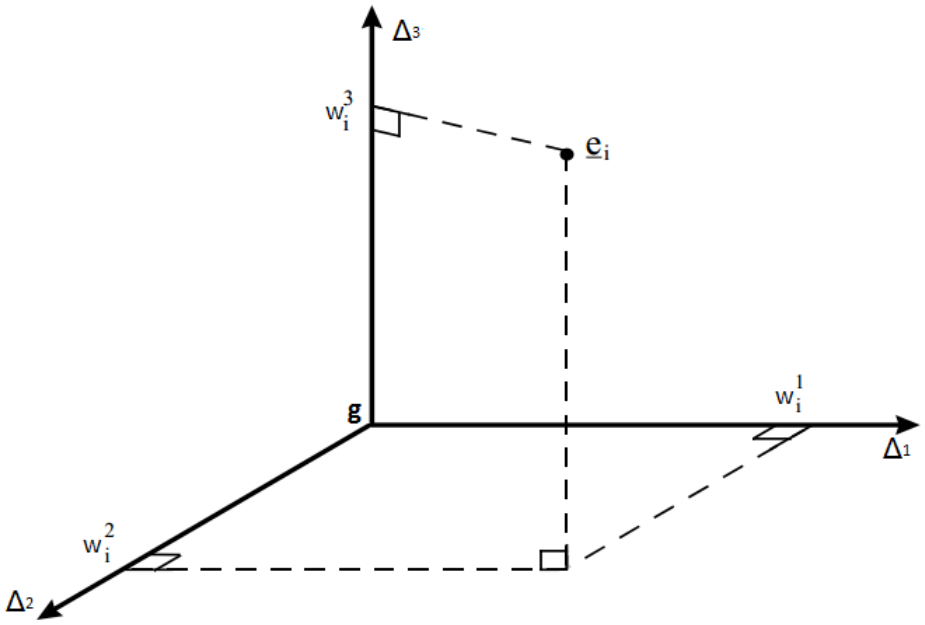


FIGURE 3.2 – projection orthogonale

Remarque 3.4. *L'orientation des axes est complètement arbitraire et peut différer d'un logiciel à l'autre. Le signe des coordonnées des individus sur un axe n'a donc pas de signification. En revanche, la comparaison des signes peut s'interpréter. Si deux individus e_i et e_j ont sur un axe Δ, le premier une coordonnées positive et le second une coordonnée négative, cela signifie qu'ils s'opposent sur cet axe.*

3.5.3 Projection des individus sur un sous espace

Le module ACP recherche les droites ajustant au mieux les nuages de points dans l'espace vectoriel à p dimensions, $\mathbb{R}^p$, des individus, au sens des moindres carrés. Mathématiquement parlant, l'objectif est d'obtenir un ensemble de vecteurs orthogonaux. Chaque vecteur de cet ensemble est proportionnel aux axes factoriels de l'espace, $\mathbb{R}^n$, des variables, et peut générer une droite dans $\mathbb{R}^p$ avec la propriété des moindres carrés. Ces vecteurs sont appelés les axes factoriels et sont davantage utilisés dans le calcul des coordonnées factorielles des points (individus) dans l'espace $\mathbb{R}^p$. La projection des individus sur l'espace vectoriel F_q, généré par l'ensemble des facteurs, peut révéler la structure cachée des données.

On ne peut pas visualiser directement le nuage $N(I)$ des individus du fait de la dimension importante de l'espace $\mathbb{R}^p$ ($p > 3$). Le principe de l'ACP consiste à projeter orthogonalement le nuage $N(I)$ sur un plan (plus généralement sur un sous espace de l'espace $\mathbb{R}^p$).

Pour donner un sens mathématique à cela, nous introduisons la notion de projection $M-$orthogonale associée à la métrique choisie M.

Définition 3.1. Soit F un sous-espace vectoriel de $\mathbb{R}^p$. La matrice de projection M-orthogonale sur F est l'unique matrice $P \in M_p$ (M_p matrice carrée de taille p x p) vérifiant :

pour tout $e_i \in \mathbb{R}^p$:

$$\begin{cases} p_{e_i} \in F \\ et \ < p_{e_i}, e_i - p_{e_i} >_M = 0. \end{cases}$$

P est tel que :

$$\begin{cases} P^2 = P \\ et \quad {}^tPM = MP. \end{cases}$$

Notons F^* le complémentaire $M-$orthogonal de F :

$$F^* = [e_i \in \mathbb{R}^p, \forall e_{i'} \in F \Rightarrow < e_i, e_{i'} >_M = 0].$$

Rappelons que tout vecteur e_i de $\mathbb{R}^p$ peut s'écrire :

$$e_i = P_{F_{e_i}} + P_{F^*_{e_i}}$$

Cette relation dit aussi que

$$P_{F^*_{e_i}} = e_i - P_{F_{e_i}}$$

Revenons maintenant au nuage des n individus, de centre de gravité g, où le plan (ou le sous espace) est choisi de façon à ce que la projection orthogonale déforme le moins possible le nuage. En terme de distance entre individus, le sous espace cherché est tel que :

$$I_F = \sum_{i=1}^{n} p_i \parallel e_i - f_i \parallel^2$$

soit minimal.

Cette écriture n'est autre que la forme classique du critère des moindres carrées ; par conséquent le sous espace passera par le point fictif g barycentre du nuage $N(I)$ des individus.
Or d'après le théorème de Pythagore, on a :

$$\| e_i - g \|^2 = \| e_i - f_i \|^2 + \| f_i - g \|^2$$

Donc :

$$\sum_{i=1}^{n} p_i \| e_i - g \|^2 = \sum_{i=1}^{n} p_i \| e_i - f_i \|^2 + \sum_{i=1}^{n} p_i \| f_i - g \|^2$$

$$I_g \qquad = \qquad I_F \qquad + \qquad I_F*$$

Par conséquent, minimiser l'expression I_F ci-dessus, revient à maximiser I_F*, puisque I_g est constante.

Si on note f_i^* la projection orthogonale de e_i sur F^* qui est le complémentaire orthogonal de F, on peut écrire :

$$d^2(f_i, e_i) + d^2(f_i^*, e_i) = d^2(g, e_i) = d^2(g, f_i) + d^2(g, f_i^*).$$

On en déduit (théorème de Huygens), que :

$$I_F + I_F^* = I_g = I_{F \oplus F*}$$

Où $\oplus$ est la somme directe, il suffit donc de remarquer que le projecteur associé à la somme directe de deux sous espaces $M-$orthogonaux est la somme des projecteurs associés à chacun des espaces.

Dans le cas particulier où le sous espace est de dimension 1, c'est-à-dire est un axe, I_{F^*} est une mesure de l'allongement du nuage selon cet axe. On emploie pour I_{F^*} les expressions " d'inertie portée par l'axe " ou bien " d'inertie expliquée par l'axe". En projetant le nuage des individus sur un sous espace F, on perd l'inertie mesurée par I_F, on ne conserve que celle mesurée par I_F^*.

De plus, si on décompose l'espace $\mathbb{R}^p$ comme la somme de sous espaces de dimension 1 et orthogonaux entre eux :

$$\Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_p$$

On peut écrire :

$$I_g = I_{\Delta_1^*} + I_{\Delta_2^*} + \cdots + I_{\Delta_p^*}.$$

Remarque 3.5. *Si deux individus sont bien projetés, alors leur distance en projection est proche de leur distance dans* $\mathbb{R}^p$

3.5.4 Coordonnées factorielles des individus

Les coordonnées factorielles des individus ne sont pas des corrélations comme c'est le cas pour les variables (lorsque nous analysons une matrice de corrélations). Il s'agit simplement des points qui sont projetés sur les droites traversant le nuage de points (au sens des moindres carrés) multidimensionnel de l'espace produit par les vecteurs des individus. Dans ce sens, c'est leur amplitude relative qui est importante. Les individus contribuant le plus à un facteur particulier (par opposition à ceux qui ont une contribution moyenne) sont les plus représentatifs du concept représentant le facteur construit. Par exemple, si nous pouvions clairement qualifier un facteur dans une analyse comme un facteur "corpulence des individus", les individus contribuant le plus à ce facteur seraient ceux dont la corpulence est la plus forte et ceux dont la corpulence est la plus faible par rapport aux individus de corpulence moyenne.

3.5.5 Graphiques associés aux individus

Pour les graphiques associés aux individus, une paire d'axes factoriels est sélectionnée parmi l'ensemble des axes factoriels. Les points de l'espace factoriel générés par les individus sont alors projetés sur le plan factoriel généré par la paire d'axes sélectionnée. Ces graphiques peuvent être utilisés pour classer les observations individuelles (individus) dans des catégories. Les individus sont classés en fonction de leurs coordonnées correspondantes sur les axes factoriels. En considérant différents couples d'axes parmi les facteurs calculés, davantage de classes d'individus peuvent être mises en évidence.

Deux points sont très voisins si leurs p coordonnées sont très proches. Les deux individus concernés sont alors caractérisés par des valeurs presque égales pour chaque variable.

3.6 L'analyse

3.6.1 Moments d'inertie

Définitions :

L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité g. Elle mesure la dispersion totale du nuage de points. Elle est donc aussi égale à la somme des variances des variables étudiées. Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1. L'inertie totale est alors égale à p (nombre de variables).

Ce moment d'inertie totale est important car c'est une mesure de la "dispersion" du nuage des individus par rapport à son centre de gravité. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est concentré sur son centre de gravité.

Inertie du nuage des individus autour d'un point a

$a \in \mathbb{R}^n$; On appelle inertie totale du nuage autour du point $\mathbb{R}^n$ la quantité :

$$I_a = \sum_{i=1}^{n} p_i \parallel e_i - a \parallel_M^2$$

Proposition : Relation de Huyglens :

$$I_a = I_g + \parallel g - a \parallel_M^2$$

Preuve :

$I_a = \sum_{i=1}^n p_i \parallel e_i - a \parallel_M^2$,
$= \sum_{i=1}^n p_i \parallel e_i - g + g - a \parallel_M^2$,
$= \sum_{i=1}^n p_i < e_i - g + g - a, e_i - g + g - a >_M$,
$= \sum_{i=1}^n p_i < e_i - g, e_i - g >_M + \sum_{i=1}^n p_i < g - a, g - a >_M + \sum_{i=1}^n p_i < e_i - g, g - a >_M$
$+ \sum_{i=1}^n p_i < g - a, e_i - g >_M$,
$= \sum_{i=1}^n p_i \parallel e_i - g \parallel_M^2 + \sum_{i=1}^n p_i \parallel g - a \parallel_M^2 + 2 \sum_{i=1}^n p_i < e_i - g, g - a >_M$,

On doit montrer que $2 \sum_{i=1}^n p_i < e_i - g, g - a >_M$ est nul :
$2 \sum_{i=1}^n p_i < e_i - g, g - a >_M = \sum_{i=1}^n p_i (g - a)' M (e_i - g)$,
$= (g - a)' M (\sum_{i=1}^n p_i (e_i - g))$,
Comme $\sum_{i=1}^n p_i (e_i - g) = 0_{\mathbb{R}^p}$.

$$\sum_{i=1}^{n} p_i e_i - g = \begin{pmatrix} \sum_{i=1}^n p_i e_i^1 \\ \vdots \\ \sum_{i=1}^n p_i e_i^p \end{pmatrix} - g,$$

$$= \begin{pmatrix} \overline{e}_1 \\ \vdots \\ \overline{e}_p \end{pmatrix} - g,$$

$$= g - g = 0_{\mathbb{R}^p}.$$

Inertie du nuage des individus par rapport à un axe passant par g

L'inertie du nuage des individus par rapport à un axe Δ passant par g est égale, par définition, à :

$$I_\Delta = \sum_{i=1}^{n} d^2 p_i^j(w_{i\Delta}, e_i) = \sum_{i=1}^{n} p_i^t(w_{i\Delta} - e_i)M(w_{i\Delta} - e_i)$$

où

$w_{i\Delta}$ est la projection orthogonale de e_i sur l'axe Δ. Cette inertie mesure la proximité à l'axe Δ du nuage des individus.

L'inertie du nuage des individus par rapport à un sous-espace vectoriel F passant par g

Cette inertie est, par définition, égale à :

$$I_F = p_i^t(w_{iF} - e_i)M(w_{iF} - e_i)$$

où w_{iF} est la projection orthogonale de e_i sur le sous-espace F.

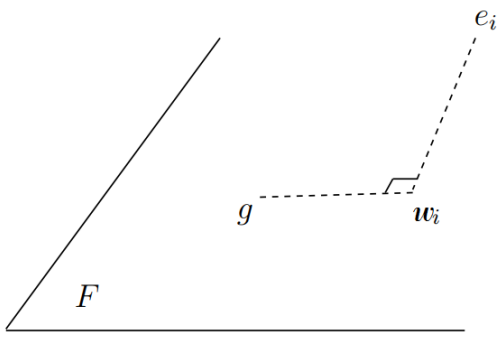


FIGURE 3.3 – Projection orthogale des individus sur F

Propriétés de l'inertie

1. Relation de Huyghens : $\forall e_i \in \mathbb{R}^p, I_{e_i} = I_g + \| g - e_i \|^2_M$.

En particulier, g est le point (unique) en lequel l'inertie est minimale.

2. $I_g = trace(MV) = trace(VM)$, où la trace d'une matrice est la somme de ses termes diagonaux.

3. Si $M = I$:

$$I_g = trace(MV) = trace(V).$$

4. Si $M = D_{\frac{1}{s^2}}$: $trace(MV) = trace(D_{\frac{1}{s^2}}V) = trace(D_{\frac{1}{s}}VD_{\frac{1}{s}})$, ce qui est égale à : $trace(R) = p$.

L'inertie est donc égale au nombre de variables et ne dépend pas de leurs valeurs.

Remarque 3.6. *L'inertie totale est invariante, si on translate tous les points du nuage d'un même segment. Ainsi, on a $I_g(X) = I_O(Y)$, inertie totale du nuage associé au tableau centré Y, de centre de gravité $O \in \mathbb{R}^p$.*

Inertie totale du nuage des individus

On note I_g le moment d'inertie du nuage des individus par rapport au centre de gravité g :

$$I_g = \sum_{i=1}^{n} p_i^t(e_i - g)M(e_i - g) = \sum_{i=1}^{n} p_i \| e_i - g \|^2_M = \sum_{i=1}^{n}\sum_{j=1}^{p} p_i^j(x_i^j - \overline{x^j})^2.$$

Si $M = I$:

$$I_g = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_i^j - \overline{x^j})^2.$$

On peut voir, en inversant l'ordre des signes sommes, que I_g peut aussi s'écrire sous la forme suivante :

$$I_g = \sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x^j})^2\right] = \sum_{j=1}^{p} var(Y^j).$$

où $Var(Y^j)$ est la variance empirique de la variable Y^j. Sous cette forme, on constate que l'inertie totale est égale à la trace de la matrice de covariance V.

$$I_g = Trace(V)$$

3.6.2 Contribution des axes à l'inertie totale

La contribution d'une variable est en fait la contribution relative d'une variable à la variance d'un axe. Les valeurs de cette statistique permettent de sélectionner les variables à interpréter par rapport à leurs coordonnées factorielles, c'est-à-dire, leurs corrélations avec les axes factoriels. Naturellement, il faut étudier les variables qui expliquent relativement le plus de variance sur l'axe factoriel.

Comme dans le cas des variables, la contribution d'un individu est également la contribution relative de cet individu à la variance d'un axe factoriel. Ainsi, d'une certaine manière, la contribution d'un individu est une mesure de l'importance d'un individu sur un axe factoriel. Plus la contribution d'un individu sera importante et plus il aura de poids sur ce facteur. Par conséquent, lorsque vous interprétez les composantes principales, vous devez commencer par les individus dont les contributions sont les plus importantes. Rigoureusement parlant, cette statistique ne doit pas être interprétée pour les individus supplémentaires dans la mesure où seuls les individus actifs contribuent aux axes.

En utilisant le théorème de Huygens, on peut décomposer l'inertie totale du nuage des individus :

$$I_g = I_{\Delta_1} + I_{\Delta_2} + \cdots + I_{\Delta_p} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Cette inertie vaut p en ACP normée et $s_1^2 + \cdots + s_p^2$ en ACP non normée.

La contribution **absolue** de l'axe Δ_k à l'inertie totale du nuage des individus est égale à la valeur propre qui lui est associée :

$$CTA(\Delta_k/I_g) = \lambda_k$$

Sa contribution **relative** ou " pourcentage d'inertie expliquée par Δ_k " est égale à :

$$CTR(\Delta_k/I_g) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à :

$$CTR(\Delta_1 \oplus \Delta_2/I_g) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Et la Part d'inertie expliquée par les q premières composantes principales :

$$CTR(\Delta_1 \oplus \Delta_2, \cdots, \oplus\Delta_q/I_g) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres sont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

3.6.3 Valeurs et vecteurs propres

Les valeurs propres

Les valeurs propres de la matrice de corrélation, ou de la matrice de covariance des variables actives jouent un rôle important dans le calcul des composantes principales. En plus de déterminer les coordonnées factorielles des variables et individus, elles donnent une assez bonne idée de la variance expliquée par le nombre de facteurs donné. Cette information peut de plus être utilisée

pour déterminer l'ordre dans lequel vous pouvez proposer de réduire les dimensions de l'espace original des variables ou individus, sans perdre beaucoup d'information. Sur la base des valeurs propres, beaucoup de critères peuvent être utilisés pour décider du nombre idéal de facteurs dans une situation donnée. Puisque la somme des valeurs propre est égale au nombre de variables "actives", la moyenne des valeurs propres est de 1, et l'approche générale est de commencer tout d'abord avec les valeurs propres supérieures à 1.

Les vecteurs propres

Les vecteurs propres sont les coefficients à affecter aux variables initiales pour obtenir les composantes principales.

3.6.4 Choix du nombre d'axes

Pour avoir une déformation minimale du nuage des points il faut que l'axe sur lequel on projette permette la dispersion maximale. Le principal intérêt de l'ACP consiste à réduire la dimension de l'espace de la représentation, le choix du nombre d'axes à retenir est un point essentiel qui n'a pas de solution rigoureuse. Il existe plusieurs types de procédures pour guider le choix de ce nombre d'axes :

1. On peut choisir le nombre q d'axes à retenir en fonction d'un pourcentage d'inertie fixé a priori.

2. On peut choisir de retenir les q axes apportant une inertie λ_q supérieure à l'inertie moyenne par variable. En ACP normée, l'inertie moyenne par variable vaut 1, et on choisit q tel que $\lambda_q > 1$ et $\lambda_{q+1} < 1$. C'est la règle de Kaiser.

3. Visualiser l'histogramme des valeurs propres (qui n'est pas un histogramme) et chercher une "cassure".

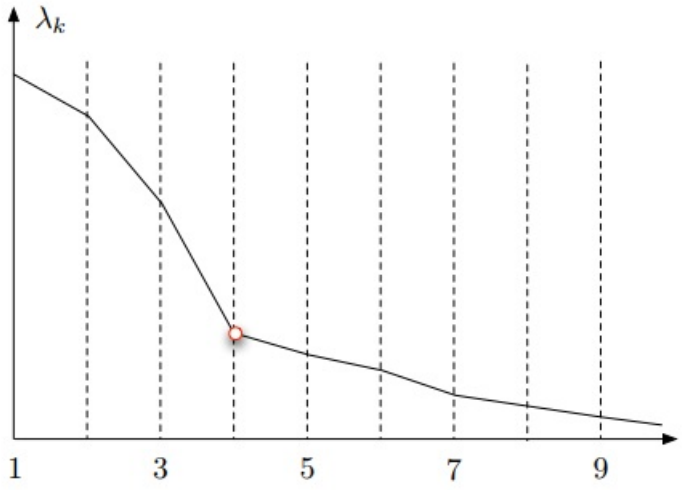


FIGURE 3.4 – Eboulis des valeurs propres

Pour quantifier cette cassure, on peut utiliser la règle du coude :
i. Calculer les différences premières :

$$\lambda_1 - \lambda_2 = \epsilon_1 \; et \; \lambda_2 - \lambda_3 = \epsilon_2 \cdots$$

ii. Calculer les différences secondes :

$$\epsilon_1 - \epsilon_2 = \delta_1 \; et \; \epsilon_2 - \epsilon_3 = \delta_2 \cdots$$

iii. Retenir les q axes tel que $\delta_1, \ldots, \delta_{q-1}$ soient toutes positives et que δ_q soit négative.

4. intervalle de confiance

Remarque 3.7. *Les intervalles de confiance d'Anderson ne sont licites que si le nuage de points est gaussien. On ne l'affiche donc qu'à titre indicatif.*

3.6.5 Axes principaux

Objectif :
Définir l'espace principal revient à :
. Définir p nouvelles variables comme axes du repère du nuage de points individus : les composantes principales. Les p axes principaux sont définis séquentiellement :
- On détermine l'axe (premier axe principal) sur lequel le nuage se déforme le moins possible en projection, c'est celui associé à la plus grande valeur propre. On le note u^1.
- On cherche un second axe, sur lequel le nuage se déforme le moins en projection, après le premier axe, tout en étant orthogonal au premier, Le deuxième axe est celui associé à la deuxième valeur propre. On le note $u^2, \ldots$.
- On réitère jusqu'à l'obtention de p axes.

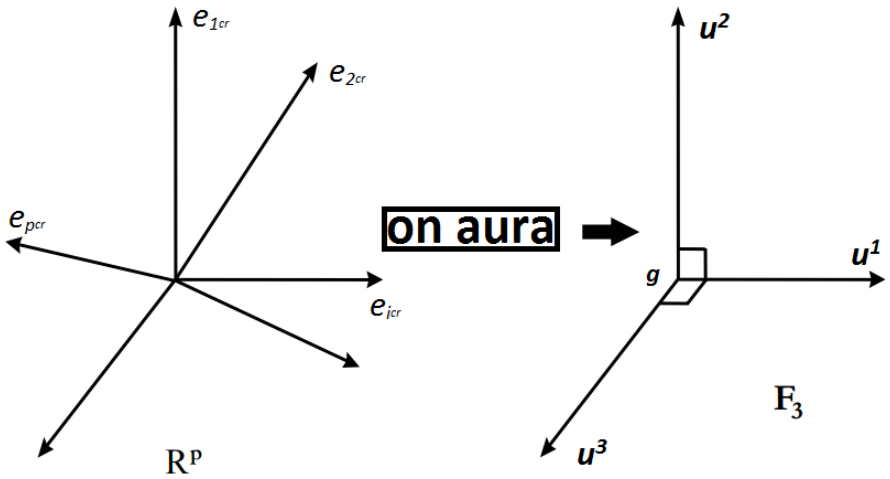


FIGURE 3.5 – Axes principaux

Remarque 3.8. *. Dans le second repère, les axes ne véhiculent pas la même information selon leurs rangs : leurs capacités à « résumer » le nuage se détériore au fur et à mesure que l'on observe des axes de rang élevé.*

. le meilleur axe (le premier axe principal) sera celui sur lequel le nuage de points projeté est de dispersion, c'est à dire tel que le nuage projeté est d'inertie maximale.

. Le second axe sera celui qui, après le premier, est tel que le nuage projeté est d'inertie maximale, tout en étant orthogonal au premier ....

. Idem pour le nuage de points variables.

Définition 3.2. On appelle axe principale la droite de $\mathbb{R}^n$ passant par g qui maximise l'inertie du nuage projeté sur cette droite, on le trouve facilement par la méthode de Lagrange.

Sur l'axe défini par le vecteur unitaire u, on associée la forme linéaire a. a est un élément de $\mathbb{R}^{p^*}$ (dual de l'espace des individus).

A l'axe principal u $M-$normé à I est associé le facteur principal $a = Mu$ puisque u était vecteur propre de VM

$$VM = \lambda \Rightarrow MVMu = \lambda Mu$$

Les facteurs principaux sont les vecteurs propres M^{-1} (M^{-1} est la matrice inverse de M), normés de MV, et comme $\mathbb{R}^p$ est muni de la métrique M son dual doit être muni d'une métrique M^{-1} , donc ${}^tuM^{-1}u = 1$. Les facteurs principaux sont $M-$orthogonaux.

Théorème 3.1. *La matrice VM admet p valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ de vecteurs propres, $M-$unitaires, associés à $u_1, \cdots, u_p$ deux à deux $M-$orthogonaux.*

Pour $k = 1, \cdots, p$, le sous-espace vectoriel F_k de dimension k portant l'inertie maximale est engendré par les vecteurs $u_1, \cdots, u_k$. De plus, $I_{F_k} = \sum_{i=1}^{k} \lambda_i$

Remarque 3.9. *Si u est le vecteur propre associé à la valeur propre λ de la matrice M, alors automatiquement $-u$ est également un vecteur propre :*

$$Mu = \lambda u \Leftrightarrow M(-u) = \lambda(-u)$$

Recherche de l'axe Δ_1 passant par g d'inertie minimale

On cherche un axe Δ_1 passant par g d'inertie I_{Δ_1} minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du nuage, rechercher Δ_1 tel que I_{Δ_1} est minimale, est équivalent à chercher Δ_1^* tel que $I_{\Delta_1 *}$ soit maximale.

$$I_{\Delta_1} \text{ est minimale} \Longleftrightarrow I\Delta_1 * \text{ est maximale.}$$

On définit l'axe Δ_1 par son vecteur directeur unitaire $\overrightarrow{u_1}$. Il faut donc trouver $\overrightarrow{u_1}$ tel que $I_{\Delta_1 *}$ est maximum sous la contrainte que $\| \overrightarrow{u_1} \|^2 = 1$.

Expressions algébriques de I_{Δ_1} et de $\| \vec{u_1} \|^2$

L'inertie du nuage projeté sur un axe Δ de vecteur directeur $M-$ unitaire u par rapport au centre de gravité g est :

$$I_\Delta = \sum_{i=1}^{n} p_i ||w_i - g||_M^2$$

Où $w_i - g$ est le projeté $M-$ orthogonal de e_i sur Δ.

D'où

$$I_\Delta = \sum_{i=1}^{n} p_i ({}^t u M e_{i_c})({}^t e_i M u)$$

$$= {}^t u M \left[\sum_{i=1}^{n} p_i e_{i_c}^t e_{i_c} \right] M u.$$

On reconnaît la matrice de variance covariance empirique V des p variables entre crochets, la matrice MVM est appelée matrice d'inertie du nuage ; elle définit la forme quadratique d'inertie qui, à tout vecteur u $M-$normé à 1, associe l'inertie projetée sur l'axe défini par u. La matrice d'inertie ne se confond avec la matrice de variance covariance que si $M = I$. On a :

$$I_\Delta = {}^t u_1 M V M u_1$$

Et

$$\| \overrightarrow{u_1} \|^2 = {}^t u M u = 1$$

Recherche du maximum

Le problème à résoudre : trouver u_1 tel que ${}^t u_1 M V M u_1$ soit maximum avec la contrainte ${}^t u_1 M u_1 = 1$ est le problème de la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte (les inconnues sont les composantes de u_1). La méthode des multiplicateurs de Lagrange peut alors être utilisée.

Dans le cas de la recherche de u_1, il faut calculer les dérivées partielles de :

$$g(u_1) = g(u_{11}, u_{12}, \cdots, u_{1p}) = {}^t u_1 M V M u_1 - \lambda_1 ({}^t u_1 M u_1 - 1)$$

En utilisant la dérivée matricielle, on obtient :

$$\frac{\partial g(u_1)}{\partial u_1} = 2MVMu_1 - 2\lambda_1 M u_1 = 0.$$

Le système à résoudre est :

$$\begin{cases} MVMu_1 - \lambda_1 &= 0 \quad (1) \\ {}^t u_1 M u_1 - 1 &= 0 \quad (2) \end{cases}$$

De l'équation matricielle (1) de ce système on déduit que u_1 est vecteur propre de la matrice VM associé à la valeur propre λ_1

En multipliant à gauche par ${}^t u_1$ les deux membres de l'équation(1) on obtient :

$$ {}^t u_1 MVMu_1 - \lambda_1^t u_1 Mu_1 = 0$$

Et en utilisant l'équation (2) on trouve que :

$$ {}^t u_1 MVMu_1 = \lambda_1$$

On reconnait que le premier membre de l'équation précédente est égal à l'inertie I_{Δ_1} qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice VM et que cette valeur propre est égale à l'inertie portée par l'axe λ_1.

L'axe Δ_1 pour lequel le nuage des individus a l'inertie minimum a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice VM.

Recherche des axes suivants

On recherche ensuite un deuxième axe Δ_2 orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 passant par g par son vecteur directeur unitaire u_2. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à :

$$ I_{\Delta_2} = {}^t u_2 MVu_2$$

Elle doit être maximum avec les deux contraintes suivantes :

$$ {}^t u_2 Mu_2 = 1 \; et \; {}^t u_1 Mu_1 = 0$$

La deuxième contrainte exprime que le deuxième axe doit être orthogonal au premier et que le produit scalaire des deux vecteurs directeurs est nul. En appliquant la méthode des multiplicateurs de Lagrange, cette fois avec deux contrainte, on trouve que u_2 est le vecteur propre de VM correspondant à la deuxième plus grande valeur propre.

On peut chercher de nouveaux axes en suivant la même procédure. Les nouveaux axes tous vecteurs propres de VM correspondant aux valeurs propres données. La matrice VM étant une matrice symétrique réelle, elle possède p vecteurs propres réels, formant une base orthogonale de $\mathbb{R}^p$:

$$\begin{cases} \Delta_1 \perp \Delta_2 \perp \cdots \Delta_p; \\ u_1 \perp u_2 \perp \cdots \perp u_p; \\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p; \\ I_{\Delta_1} \geq I_{\Delta_2} \geq \cdots \geq I_{\Delta_p}; \end{cases}$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale

des vecteurs propres de VM. On appelle les nouveaux axes, axes principaux.

3.6.6 Composantes principales

On projette les individus sur le sous-espace Δ_{u_1} de dimension 1, on obtient un vecteur appelé la première composante principale :

$$C^1 = \begin{pmatrix} c_1^1 \\ c_2^1 \\ \vdots \\ c_n^1 \end{pmatrix} \in \mathbb{R}^n;$$

Avec $c_i^1 =< y_i, u_1 >_M$ projection de l'individu i sur le premier sous espace : $c_i^1 = y_i' M u_1$. On peut écrire aussi :

$$C^1 = \begin{pmatrix} < y_1, u_1 >_M \\ < y_2, u_1 >_M \\ \vdots \\ < y_n, u_1 >_M \end{pmatrix};$$

Ainsi, on projette les individus de même sur le deuxième sous espace Δ_{u_2} de dimension 1, on obtient la deuxième composante principale :

$$C^1 = \begin{pmatrix} c_1^2 \\ c_2^2 \\ \vdots \\ c_n^2 \end{pmatrix} \in \mathbb{R}^n = y_i' M u_2;$$

Ainsi de suite pour les autres composantes.

On note $C^1, \ldots, C^q, \ldots, C^p$ les composantes principales ; C^q étant la nouvelle variable correspondant à l'axe Δ_{u_q} :

$$C^q = \sum_{j=1}^{p} u_{qj} Y^j = Y u_q = Y M u_q$$

Et de façon générale :

$$C = YU.$$

Avec $U = u_1, u_2, \ldots, u_q$.

Pour obtenir ces coordonnées, on écrit que chaque composante principale est une combinaison linéaire des variables initiales.

La k^{eme} composante principale s'écrit :

$$C^q = u_{q1} Y^1 + \ldots + u_{qp} Y^p$$

Propriétés des Composantes Principales

moyenne nulle : La moyenne des composantes principales "les nouvelles variables" doivent être toutes nulles.

$$\overline{C^k} = 0, \forall \alpha = 1, \ldots, p$$

Corrélation nulle : La corrélation entre les composante principales est nulle.

$$cor(C^k, C^l) = 0$$

On a :

$$cov(C^k, C^l) =< C^k, C^l >_{D_p}$$

Or $C^k = Xu_k$,
Donc

$$cov(C^k, C^l) = u'_k \, {}^tY D_p Y u_l = u'_k V u_l$$

Car $Y' D_p Y = V$.
Rappelons que u_l est vecteur propre de VM associé à la valeur propre λ_l,
Ainsi :

$$cov(C^k, C^l) = \lambda u'_k u_l = 0$$

Car deux vecteurs propres associés à deux valeurs propres distinctes sont orthogonaux.
D'où :

$$cor(C^k, C^l) = \frac{cov(C^k, C^l)}{s_k s_l} = 0$$

Avec s_k est l'écart type de la variable C^k.

Variance égale la valeur propre : la variance de chaque composante principale égale la valeur propre correspondante.

$$\| C^k \|^2 = var\, C^k = \lambda_k$$

En effet

$$Var(C) = C'DC = u'\, Y'DYu = u'Vu$$

Or :

$$Vu = \lambda M^{-1} u$$

Donc :

$$Var(C) = \lambda\, u' M^{-1} u = \lambda$$

Nombre : bien que l'objectif soit en général de n'utiliser qu'un petit nombre de Composantes Principales, l'ACP en construit initialement p, autant que de variables originales.

Ce n'est que par la suite que l'analyste décidera du nombre de Composantes à retenir.

"Retenir q Composantes Principales " veut dire Remplacer les observations originales par leur projections orthogonales dans le sous-espace à q dimensions défini par les q premières Composantes

Principales.

Orthogonalité : les Composantes Principales définissent des directions de l'espace des observations qui sont deux à deux orthogonales. Autrement dit, l'ACP procède à un changement de repère orthogonal, les directions originales étant remplacées par les Composantes Principales.

Dé-corrélation : les Composantes Principales sont des variables qui s'avèrent être deux à deux dé-corrélées.

Ordre et sous espaces optimaux : la propriété fondamentale des Composantes Principales est de pouvoir être classées par ordre décroissant d'importance dans le sens suivant :

∗ Si l'analyste décide de décrire ses données avec seulement $q(q < p)$ combinaisons linéaires de ses variables originales tout en perdant le moins possible d'information, alors ces k combinaisons linéaires sont justement les q premières Composantes Principales.

Ainsi, le meilleur sous-espace à q dimensions dans lequel projeter les observations est celui engendré par les q premières Composantes Principales. Autrement dit, les sous espaces de projection optimale sont emboités, ce qui est une propriété forte et très utile.

3.6.7 Qualité de Représentation

Une des questions importantes qui doit être traitée dans l'ACP est le nombre de composantes principales qui pourraient idéalement représenter l'ensemble complet des points (variables ou individus). Comme chaque valeur propre de la matrice de corrélation ou de covariance est représentative de la variance expliquée par une composante principale, un pourcentage de variance cumulée (expliquée) peut être attribué à un nombre donné de facteurs. Ceci représente la qualité de représentation et est une mesure importante de la variance comptant pour un ensemble de composantes principales donné.

Qualité de la représentation des variables

On mesure la qualité de la projection d'une variable Z^j sur l'axe Δ_α par le carré du cosinus de l'angle $\theta_{j\alpha}$ entre le vecteur z^j et l'axe Δ_α :

$$cos^2(\theta_{kj}) = \frac{v_{kj}^2}{\parallel Z^k \parallel^2} = v_{kj}^2$$

On mesure la qualité de la projection d'une variable Z^k sur le plan $(\Delta_k, \Delta_{j'})$ par le carré du cosinus de l'angle $\theta_{k(j,j')}$ entre le vecteur Z^k et sa projection orthogonale sur (j, j') :

$$cos^2(\theta_{j,j'}) = u_{kj}^2 + u_{kj'}^2$$

$\sqrt{cos^2(\theta_{k(j,j')})}$ est donc " la longueur de la flèche ".
Plus la flèche est proche du cercle, meilleure est la qualité de la représentation de la variable.

Une variable sera d'autant mieux représentée sur un axe, un plan, ou un sous-espace que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. En

effet, le coefficient de corrélation empirique entre une ancienne variable Z^j et une nouvelle variable C_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point v_j représentant la variable sur l'axe avec cet axe.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1.

Qualité de la représentation des individus

On mesure la qualité de la projection d'un individu e_i sur l'axe Δ_k par le carré du cosinus de l'angle θ_{ik} entre le vecteur e_i et l'axe Δ_k :

$$cos^2(\theta_{ik}) = \frac{w_{ik}^2}{\| e_i \|^2} = \begin{cases} \simeq 1 \ tres\ bonne\ representation; \\ \geq 0.5 \ representation\ acceptable; \\ \leq 0.5 \ mauvaise\ representation. \end{cases}$$

On mesure la qualité de la projection d'un individu i sur le plan $(\Delta_\alpha, \Delta'_\alpha)$ par le carré du cosinus de l'angle $\theta_i(\alpha, \alpha')$ entre le vecteur z_i et sa projection orthogonale sur $(\Delta_\alpha, \Delta'_\alpha)$:

$$cos^2(\theta i_{k,k'}) = \frac{w_{ik}^2 + w_{ik'}^2}{\| e_i \|^2}$$

Plus la valeur du cos^2 est proche de 1, meilleure est la qualité de la représentation de l'individu. Si deux individus sont bien projetés, alors leur distance en projection est proche de leur distance dans $\mathbb{R}^p$. Lorsque des points projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentants ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches sur un axe peuvent ne pas être proches dans l'espace.

Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut s'assurer que ces individus sont bien représentés dans le plan. Pour que l'individu e_i soit bien représenté sur un plan, il faut que l'angle entre le vecteur $\overrightarrow{e_{i_c}}$ et le plan soit petit. On calcule donc le carré de cosinus de cet angle. En utilisant le théorème de Pythagore, on peut montrer que le carré de cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égale à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension k quelconque. Si le carré du cosinus de l'angle entre $\overrightarrow{e_{i_c}}$ et le plan est proche de 1, on pourra dire que l'individu e_{i_c} est bien représenté par sa projection sur le plan. Et si deux individus sont bien représenté en projection sur un plan et on a des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace. Le carré du cosinus de l'angle θ_{ik} entre $\overrightarrow{e_{i_c}}$ et un axe Δ_k de vecteur directeur unitaire u_k est égale à :

$$\cos^2(\theta_{ik}) = \frac{< \overrightarrow{e_{ic}}, \overrightarrow{u_k} >^2_M}{\| \overrightarrow{e_{ic}} \|^2} = \frac{(w_{ik})^2}{\| e_{i_c} \|^2_M}$$

En utilisant le théorème de Pythagore on peut calculer le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{e_{ic}}$ et le plan engendré par les deux axes $\Delta_k \oplus \Delta_{k'}$:

$$\cos^2(\theta_{ikk'}) = \cos^2(\theta_{ik}) + \cos^2(\theta_{ik'})$$

Si, après l'étude des pourcentages d'inertie expliqués par les sous-espaces successifs engendrés par les nouveaux axes, on a décidé de ne retenir qu'un sous-espace de dimension $k < p$, on pourra calculer la qualité de la représentation d'un individu e_{i_c} en calculant le carré du cosinus de l'angle de $\overrightarrow{e_{ic}}$ avec ce sous-espace.

Remarque 3.10. *Si un individu est très proche du centre de gravité dans l'espace, c'est- à-dire si* $\| \overrightarrow{e_{ic}} \|^2_M$ *est très petit, le point représentant cet individu sur un plan sera bien représenté.*

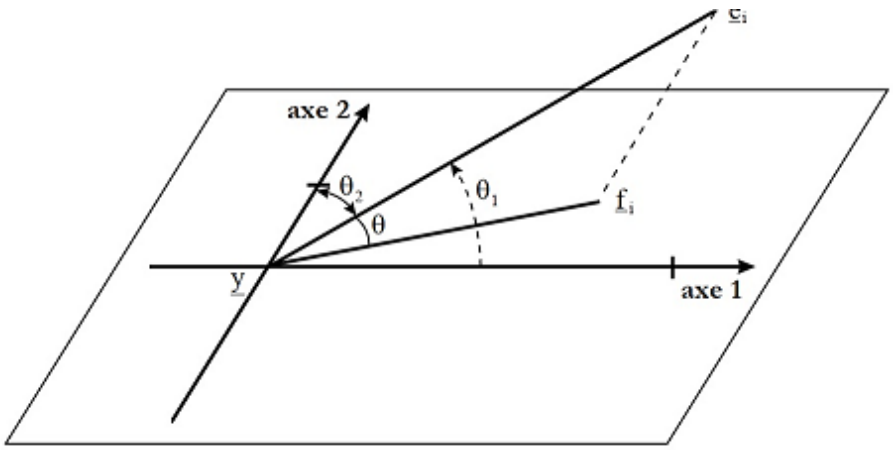


FIGURE 3.6 – Projection sur un plan de l'angle des carrés de cosinus

3.6.8 Interprétation des nouveaux axes en fonction des anciennes variables

Interprétation des coordonnées factorielles des variables. Comme mentionné plus tôt, les coordonnées factorielles ne sont rien d'autre que les corrélations entre une variable et les axes factoriels. Plus la valeur absolue du poids factoriel d'une variable sur un facteur particulier est élevée, plus la variable est corrélée à ce facteur. En d'autres termes, plus la coordonnée factorielle d'une variable sera importante, plus la variable contribuera au concept représenté par ce facteur. Par exemple, un facteur avec des poids élevés pour trois mesures de la corpulence d'une personne, comme le poids en kilos, la taille en centimètres, et le tour de poitrine en centimètres, pourraient sans

doute être interprétés comme représentatifs de la "corpulence" (c'est-à-dire, que ces trois variables contribuent le plus fortement à cet axe).

On peut interpréter les axes principaux en fonction des anciennes variables. Une ancienne variable X^j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondante à cet axe.

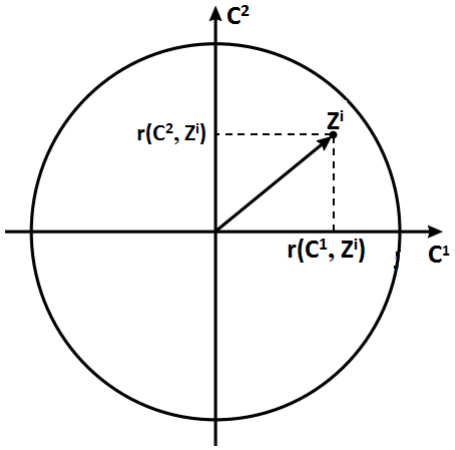


FIGURE 3.7 – La relation entre anciennes et nouvelles variables

Interprétation de la position des variables

Sur le graphique du cercle des corrélations, on peut interpréter les positions des variables initiales les unes par rapport aux autres en termes de corrélations. Deux points très proches du cercle des corrélations, donc bien représentés dans le plan, seront très corrélés positivement entre eux s'ils sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, ils sont très corrélées négativement. Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, seront anti corrélées entre elles.

Si deux variables sont bien projetées, alors leur angle en projection est proche de leur angle dans $\mathbb{R}^n$.

Sachant que la corrélation entre deux variables est le cosinus de l'angle entre les variables centrées-réduites :

- un angle de 90° correspond à une corrélation nulle.
- un angle nul correspond à une corrélation de 1.
- un angle de 180° correspond à une corrélation de -1.

Corrélations entre "variables initiales" et "composantes principales"

La méthode la plus naturelle pour donner une signification à une composante principale C^q est de la relier aux variables initiales Z^j en calculant les coefficients de corrélation linéaire $Cor(C^q, Z^j)$. On obtient alors, ce que l'on appelle communément le "cercle de corrélation", dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et +1.

Les représentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1 si la représentation est faite sur un plan (2 dimensions).

Dans le cas centré ou $M = I$ on peut montrer que les variances-covariances et les coefficients de corrélation empiriques des composantes principales avec les variables initiales sont :

$$cov(C^q, Y^j) = u_q' Y' D Y^j = u_q' Y' D Y \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = u_q' V \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \lambda_q u_q' \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \lambda_q u_{qj}$$

Enfin :

$$Cor(C^q, Y^j) = \sqrt{\lambda_q} \frac{u_{qj}}{\sqrt{Var(X^j)}}$$

Où u_{qj} est la $j$ème coordonnées du vecteur directeur unitaire u_q de Δ_q.

Lorsqu'on choisit la métrique $M = D_{\frac{1}{s^2}}$ ce qui revient à travailler sur données centrée réduites, le calcule de $Cor(C^q, Z^j)$ est particulièrement simple :

En effet :

$$Cor(C, X^j) = Cor(C, Z^j) = \frac{{}^t C D Z^j}{s_z}$$

comme $Var(C) = \lambda$:

$$Cor(C, X^j) = \frac{{}^t C D Z^j}{\sqrt{\lambda}}$$

Or $C = Yu$ où u, facteur principal associé à C est vecteur propre de R associé à la valeur propre λ :

$$Cor(C, X^j) = {}^t U {}^t Z D Z^j = \frac{{}^t Z^j D Z U}{\sqrt{\lambda}}$$

${}^t Z^j D Z$ est la $j$ème ligne de ${}^t Z^j D Z = R$, donc ${}^t Z^j D Z U$ est la $j$ème composante de RU.

comme $RU = \lambda U$, il vient :

$$Cor(C^q, Z^j) = \sqrt{\lambda_q} u_{qj}$$

Ces calculs s'effectuent pour chaque composante principale.

De façons générale, la matrice de covariance des composantes principales est égale à V_c :

$$V_c = {}^t U {}^t Y D Y U = {}^t U V U = \Lambda$$

Où Λ est la matrice diagonale des valeurs propres de VM :

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

Et la matrice des covariances entre les composantes principales et les anciennes variables Vaut :

$$Cov(C, Y^j) = {}^t YDYU = VU = U\Lambda$$

Remarque 3.11. *Lorsqu'on ne travaille pas sur des données centrées réduites, il vaut mieux éviter d'interpréter les proximités entre points variables si ceux-ci ne sont pas proches de la circonférence. Par contre dans le cas de l'ACP centrée-réduite, le cercle des corrélations est la projection exacte de l'ensemble des variables centrées réduites sur le sous espace engendré par z_1 et z_2.*

3.6.9 Interprétation des nouveaux axes en fonction des individus

Chaque axe principal Δ_k, de vecteur directeur u_k, représente une nouvelle variable C_k de dimension n, construite comme combinaison linéaire des variables (axes) de départ, appelée composante principale. La coordonnée c_{i_k} d'un individu i donné sur cet axe correspond à la valeur de la composante principale prise par cet individu.

Les composantes principales sont construites de manière à restituer la majeure partie de l'information du tableau. Elles déforment le moins possible l'information. La première composantes principale sera une combinaison linéaire des variables de départ de dispersion (de variance) maximale.

Les composantes principales sont non corrélées (les axes sont orthogonaux).

L'interprétation des coordonnées factorielles s'effectue par rapport à leur contribution à la variance. Dans une première étape, nous recherchons les individus qui possèdent les contributions les plus élevées pour un facteur sélectionné. Nous pouvons alors sélectionner un sous-ensemble de ces individus et rechercher les contributions supérieures à la contribution moyenne. Le sous-ensemble de ces points est alors divisé en deux ensembles : le premier avec les coordonnées négatives, et le second avec les coordonnées positives. Ce partitionnement permet de mettre en évidence les différences qui existent parmi les individus, révélant ainsi la structure des données cachée dans les individus.

Lorsque on calcule l'inertie I_{Δ_k} portée par l'axe Δ_k on peut voir quelle est la part de cette inertie due à un individu e_i particulier.

Dire que C_k est très corrélé avec une variable X^j signifie que les individus ayant une forte coordonnée positive sur l'axe 1 sont caractérisés par une valeur de X^j nettement supérieure à la moyenne (rappelons que l'origine des axes représente le centre de gravité du nuage).

Inversement si les individus ne sont pas anonymes, ils aident à l'interprétation des axes principaux et des composantes principales : on cherchera par exemple les individus opposés le long

d'un axe.

Il est très utile aussi de calculer pour chaque axe la contribution apporté par les divers individus à cet axe.

3.6.10 Interprétation similaire des graphiques

En observant le cercle de corrélations des variables et le premier plan factoriel des projections des individus en similaire, on va avoir une idée sur les valeurs des variables, et ça nous aides dans l'interprétation des résultats.

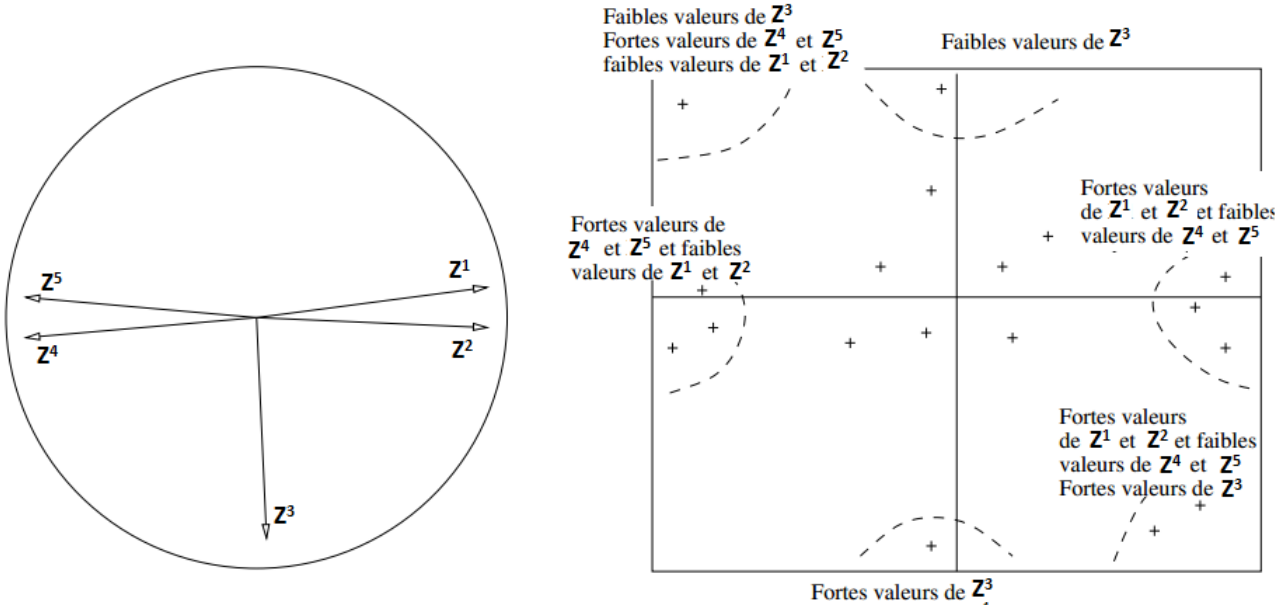


FIGURE 3.8 – Représentations des variables et des individus

On a $Yu_j = C^j$; en post-multipliant les deux membres de cette relation par $u_{j'}$ et en sommant sur j on trouve :

$$Y \sum_{j=1}^{p} u_j u_{j'} = \sum_{j=1}^{p} C^j u_{j'}$$

Comme les u_j sont orthogonaux , donc $\sum_{j=1}^{p} u_j u_{j'} = I$.

D'où :

$$Y = \sum_{j=1}^{p} C^j u_{j'}$$

On peut donc ainsi reconstituer le tableau de données (centré) à partir des facteurs et composantes principales. Si l'on se contente de sommer sur les q premiers termes correspondant aux q plus grandes valeurs propres, on obtient alors la meilleure approximation de Y par une matrice de rang q au sens des moindres carrés.

Récapitulation

Algorithme ACP :

1. Calculer les moyennes des variables $X_j,\ j = 1, \ldots, p$.

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. Centrer le tableau X pour que g soit situé à l'origine [obligatoire] :

$$y_j = x_j - \overline{x_j}$$

Et réduire si les données sont hétérogènes et cela pour éliminer l'effet des unités :

$$z = \frac{y_j}{s_j}$$

3. Calculer la matrice de variance covariance $V = X^t D_p X = \frac{1}{n} X^t X$.
4. Calculer les valeurs propres et les vecteurs propres de V.
5. Calculer les projections des individus et des variables sur les axes factoriels :

$$\begin{cases} C^j = Xu, \\ W_j = \sqrt{\lambda_j} u \end{cases}$$

6. Représenter graphiquement les individus et les variables.
7. Interpréter les résultats de l'analyse.

Pratique de l'ACP :

1. Choisir les variables actives.
2. Choisir de réduire ou non les variables.
3. Réaliser l'ACP.
4. Choisir le nombre de dimensions à interpréter.
5. Interpréter simultanément le graphe des individus et celui des variables.
6. Utiliser les indicateurs pour enrichir l'interprétation.
7. Revenir aux données brutes pour interpréter.

Résumé

Soit un tableau de données,X_{np} contenant les observations de n individus statistiques sur p variables quantitatives continues.

L'espace des colonnes $\mathbb{R}^n$ est muni d'une métrique $D = diag(\ldots, p_i, \ldots)$ des poids des individus.

L'espace des lignes $\mathbb{R}^p$ est muni d'une métrique M.

En ACP on peut analyser :

la matrice des données centrées Y,

la matrice des données centrées-réduites Z.

L'ACP consiste alors à analyser deux nuages de points :

les n points individus de $\mathbb{R}^p$ (les lignes) avec la métrique $M = I_p$,

les p points variables de $\mathbb{R}^n$ (les colonnes) avec la métrique $M = \frac{1}{n} I_n$

On distingue alors deux type d'ACP :

l'ACP non normée (sur matrice des covariances) qui analyse Y, l'ACP normée (sur matrice des corrélations) qui analyse Z.

Chapitre 4 : Application de l'ACP

Chapitre 4

Application de l'ACP

4.1 Introduction

Dans ce travail que nous achevons avec l'application de l'analyse en composante principale nous étudions des données récupérés auprès de l'entreprise NAFTAL,ces données sont sous forme de tableau de donnée sur les lubrifiants de l'année 2015 notre application se fera sur les 14 individus(produits lubrifiants) et 05 variables quantitatives,l'objectifs de cette analyse est de détecter les différents liens et relations entre nos différentes variables , et cela avec une représentation de données graphiquement sur les plans d'axes principaux qui constituent le meilleur résumé possible de l'information et établir une similarité entre les individus et les variables,notre études se fera avec le logiciel de statistique R.

4.2 Présentation du logiciel R

4.2.1 Origines

Le logiciel **R** est un logiciel de statistique créé par Ross Ihaka et Robert Gentleman. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre.

C'est un clone du logiciel S-plus qui est fondé sur le langage de programmation orienté objet, développé par **AT et T** Bell Laboratories en 1988 . Ce logiciel sert à manipuler des données,à tracer des graphiques et à faire des analyses statistiques sur ces données.

4.2.2 l'utilité de l'utilisation du logiciel

Tout d'abord R est un logiciel gratuit et à code source ouvert (open source). Il fonctionne sous UNIX (et Linux), Windows et Macintosh. C'est donc un logiciel multi-plates-formes. Il est

développé dans la mouvance des logiciels libres par une communauté sans cesse plus vaste de bénévoles motivés.

Tout le monde peut d'ailleurs contribuer à son amélioration en y intégrant de nouvelles fonctionnalités ou méthodes d'analyse non encore implémentées. Cela en fait donc un logiciel en rapide et constante évolution.

C'est aussi un outil très puissant et très complet, particulièrement bien adapté pour la mise en œuvre informatique de méthodes statistiques.L'avantage en est toutefois double :
– l'approche est pédagogique puisqu'il faut maîtriser les méthodes statistiques pour parvenir à les mettre en œuvre ;
– l'outil est très efficace lorsque l'on domine le langage R puisque l'on devient alors capable de créer ses propres outils, ce qui permet ainsi d'opérer des analyses très sophistiquées sur les données

Le logiciel R est particulièrement performant pour la manipulation de données, le calcul et l'affichage de graphiques. Il possède, entre autres choses :
– un système de documentation intégré très bien conçu (en englais) ;
– des procédures efficaces de traitement des données et des capacités de stockage de ces données ;
– une suite d'opérateurs pour des calculs sur des tableaux et en particulier sur des matrices ;
– une vaste et cohérente collection de procédures statistiques pour l'analyse de données ;
– des capacités graphiques évoluées ;
– un langage de programmation simple et efficace possibilités d'entrée-sortie.

4.2.3 Les différents packages R utlisées

Un package R est un ensemble cohérent de fonctions, de jeux de données et de documentation permettant de compléter les fonctionnalités du système de base ou d'en ajouter de nouvelles. Les packages sont installés depuis le site Comprehensive R Archive Network (CRAN) ;

FactoMineR

est un package R dédié à l'analyse exploratoire multidimensionnelle de données (à la Française). Il a été développé et il est maintenu par François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, et J. Mazet.

- Il permet de mettre en oeuvre des méthodes analyses de données telles que l'analyse en composantes principales (ACP), l'analyse des correspondances (AC), l'analyse des correspondances multiples (ACM) ainsi que des analyses plus avancées.

- Il permet l'ajout d'information supplémentaire telle que des individus et/ou des variables supplémentaires.

- Il fournit un point de vue géométrique et de nombreuses sorties graphiques.

- Il fournit de nombreuses aides à l'interprétation (description automatique des axes, nombreux indicateurs, ...).

- Il peut prendre en compte diverses structures sur les données (structure sur les variables, hiérarchie sur les variables, structure sur les individus).

- Beaucoup de matériels pédagogique (MOOC, livres, etc.) est disponible pour expliquer aussi bien les méthodes que la façon de les mettre en oeuvre avec FactoMineR.

- Il gère les données manquantes avec missMDA

- Il a une interface Shiny qui permet de construire des graphes de façon interactive avec Factoshiny.

- Il propose une interprétation automatique des résultats obtenus avec FactoMineR grâce à FactoInvestigate .

Factoshiny

C'est une interface graphique qui permet de paramétrer les méthodes et de modifier les graphes de façon interactive Il n'est pas nécessaire de savoir programmer .L'objet résultat de Factoshiny peut être réutilisé pour modifier les graphes. L'interface est rouverte avec le dernier paramétrage et les dernières options graphiques qui peuvent être modifiés. permet d'améliorer facilement et de façon interactive les graphiques pour les rendre beaucoup plus lisibles.

missMDA

Le package missMDA est complémentaire de FactoMineR. Il permet de gérer les données manquantes pour les méthodes d'analyses. Il permet de faire de l'imputation simple et multiple.

L'imputation simple consiste à remplacer les valeurs manquantes par des valeurs plausibles. Cela revient à compléter le jeu de données qui peut ensuite être analysé par n'importe quelle méthode d'analyse factorielle.

missMDA impute les valeurs manquantes de sorte que les valeurs imputées n'ont aucune influence sur les résultats de l'analyse factorielle,pas d'influence dans le sens où les valeurs imputées n'ont aucun poids, et donc les résultats de l'analyse factorielle sont obtenues uniquement avec les valeurs observées

4.3 Présentation de données

Nous allons, à partir données extraites dans des bilans et archives comptables et fiches techniques des produits (lubrifiants) NAFTAL de l'année 2015, étudier les divers facteurs liés aux lubrifiants (huiles) Nous allons traiter un tableau de 14 individus représentant les lubrifiants commercialisés par l'entreprise NAFTAL et de 5 variables que nous allons décrire ci-après :

Variables

1. Quantités totales vendues en lubrifiant pendant un an en (Tonnes) : notée **QV** :
2. Prix de vente unitaire pour 01 (Dinars) : noté **PVU** .
3. La distance de vidange requise en (Kilomètres) : notée **VID** .

4. La viscosité est une mesure de résistance à l'écoulement d'un fluide. La viscosité d'une huile moteur s'exprime par 2 grades. Un grade à froid et un grade à chaud.En (mm^2 / S, Centistoke) :noté **VISCO** .

5. Le point d'écoulement se réfère à la température la plus basse à laquelle un lubrifiant continue de s'écouler. En dessous de ce point, l'huile tend à s'épaissir et à cesser de s'écouler librement. (C°) : noté **P.ECOU**.

Individus

Les 14 individus (lubrifiants) que nous allons décrire ci-après : [1]

Particuliers
Les lubrifiants pour moteurs à essence et diesel
-NAFTILIA SUPER
-CHELIA SUPER TD 15W40
-NAFTILIA SYNTH 10W40 P
-NAFTILIA SYNTH ECO 5W30
-NAFTILIA SYNTHETIQUE PLUS
-NAFTILIA VP SUPER 15W40 P

Professionels
-Huiles moteurs
TISKA
TASFALOUT
TILIA B
-Graisses
TASSADIT A2
TESSALA graisse
-Huiles hydraulique
TORADA
-Turbine
TORBA
-Engrenages
FODDA

1. https ://www.naftal.dz/fr/index.php/produits

4.4 L'utilisation de R pour L'analyse des données

4.4.1 Code d'application sous R

```
rm(list=ls())
ls()
require(graphics) ;require(stats)
# setwd("C :/Utilisateurs/brahim/Bureau/Données naftal")
getwd()

# Library des packages
library(FactoMineR)
library(Rcmdr)
library(Factoshiny)
library(missMDA)
library(FactoInvestigate)
library(rgl)
library(MASS)
library(factoextra)
library("plot3D")
library("plot3Drgl")

# Importation des données 2015
annee2015<- read.table(file="2015.csv",header=TRUE,sep=";",dec=",")
```

Voir figure 4.8 page 78

```
X <- annee2015[2 :6] Voir figure 4.9 page 78
# Estimation des données manquantes Voir figure 4.10 page 79
nbx <- estim_ncpPCA(X)
nbx$ ncp
nbx$ criterion

# Package factoMineR
res.imputex <- imputePCA(X,ncp=0)
res.imputex$ completeObs

# Application de L'ACP avec le nouveau tableau de données estimer. voir figures 4.3, 4.4
```

pages 73 ,74

```
res.acpx <- PCA(res.imputex$ completeObs)

# Simulation de l'interafce graphique
resshiny = PCAshiny(res.acpx)
```

```
# Matrice des corrélation Voir figure 4.11 page 79
round(cor(res.imputex$ completeObs),2)
round(det(cor(res.imputex$ completeObs)),2)
summary(res.imputex$ completeObs)

# Calcules de valeurs propres manuel

acpx<-princomp(res.imputex$ completeObs,cor=T,scores=F)
valeurs_propres <- get_eigenvalue(res.acpx)
round(valeurs_propres,3)

inertie <- acpx$ sdevˆ 2/sum(acpx$ sdevˆ 2)∗ 100
round(inertie,3)

# Eboulis des valeurs propres voir figure 4.1 page 70
plot(1 :5,acpx$ sdevˆ 2,type="b",ylab="valeurs propres ",xlab="composante",main="scree plot")
# Diagramme de l'inertie et eboulis des valeurs propres voir figure 4.2 page 71
fviz_eig(res.acpx)
barplot(inertie,ylab="% d'inertie",names.arg=round(acpx$ sdevˆ 2,2))

# Les différents résultats des L'ACP
print(attributes(res.acpx))
res.acpx$ svd : Décomposition en valeurs singulières définies par la matrice identité.
res.acpx$ eig : calcule les valeurs propres et les vecteurs propres d'une matrice
res.acpx$ var :Résultats des variables
res.acpx$ ind : Résultats des variables
res.acpx$ call

# graphe des individus par les cos2 et les contribution
fviz_pca_ind(res.acpx,
        col.ind = "cos2", # Colorer par le cos2
        gradient.cols = c("# 00FF00", "# E7B800", "# 0000FF"),
        repel = TRUE )

fviz_pca_ind(res.acpx,
        col.ind = "contrib", # Colorer par le contrib
        gradient.cols = c("# 00FF00", "# E7B800", "# 0000FF"),
        repel = TRUE     )

# Graphe des variables par les cos2 et les contribution
```

```
fviz_pca_var(res.acpx,
             col.var = "contrib",
             gradient.cols = c("# 00FF00", "# E7B800", "# 0000FF"),
             repel = TRUE )

fviz_pca_var(res.acpx,
             col.var = "cos2", # Colorer par le cos2
             gradient.cols = c("# 00FF00", "# E7B800", "# 0000FF"),
             repel = TRUE)

# Graphe assemblé. voir figure 4.5 page 74
fviz_pca_biplot(res.acpx, repel = TRUE,
                col.var = "# 00FF00",
                col.ind = "# 0000FF", )
# Résultats des variables

res.var <- get_pca_var(res.acpx)
res.var$ coord          # Coordonnées
res.var$ contrib        # Contributions aux axes
res.var$ cos2           # Qualité de représentation

# Résultats des individus
res.ind <- get_pca_ind(res.acpx)
res.ind$ coord          # Coordonnées
res.ind$ contrib        # Contributions aux axes
res.ind$ cos2           # Qualité de représentation
```

4.4.2 Le Tableau de données

Nos différents données de 05 variables et 14 individus (produits) sont stockées dans un tableau excel sous forme .csv sous le nom (annee2015). Ensuite on procèdera à l'extraction de nouveau tableau avec les données quantitatives d'ou on va appliqué l'ACP.

4.4.3 Etudes de tableau de données

Vu qu'on a deux éléments manquants de l'année 2015 on doit estimer les valeurs manquantes à l'aide de la fonction **estim_ncpPCA** Voir figure 4.10 page 79 .

Après estimation de ce dernier tableau On procédera a l'utilisation du package **FactoMineR**,avec les commandes suivantes :
res.imputex <- imputePCA(X,ncp=0) ,**res.imputex$completeObs**

Le package **FactoMineR** permet de mettre on oeuvre la méthode d'analyse en composante principale qui nous fournis un point de vue géométrique et de nombreuses sorties graphique et différents indicateurs qui nous permet de faire des interprétations à nos résultats.

L'ACP centrée réduite

L'ACP se fera avec la fonction **princomp()** cette dernière permet de centré et réduire le tableau initiale (cor=T,scores=T).

Les différents résultats de l'ACP

print(attributes(res.acpx))
- sdev : Les écarts types des composants principaux.
- eig : Valeurs propres
- call : L'appel correspondant.
- center : Les moyens qui ont été soustraits.
- scale : Les mises à l'échelle appliquées à chaque variable,(pour que l'attribut soir réduit)
- n.obs :Le nombre d'observation
- loading :La matrice de rotation,dont chacune des colonnes est un vecteur propre.

Choix de nombre d'axes

Les valeurs propres nous renseignent sur la fraction de l'inertie total prise en compte par chaque axes.

- Eboulis des valeurs propres :

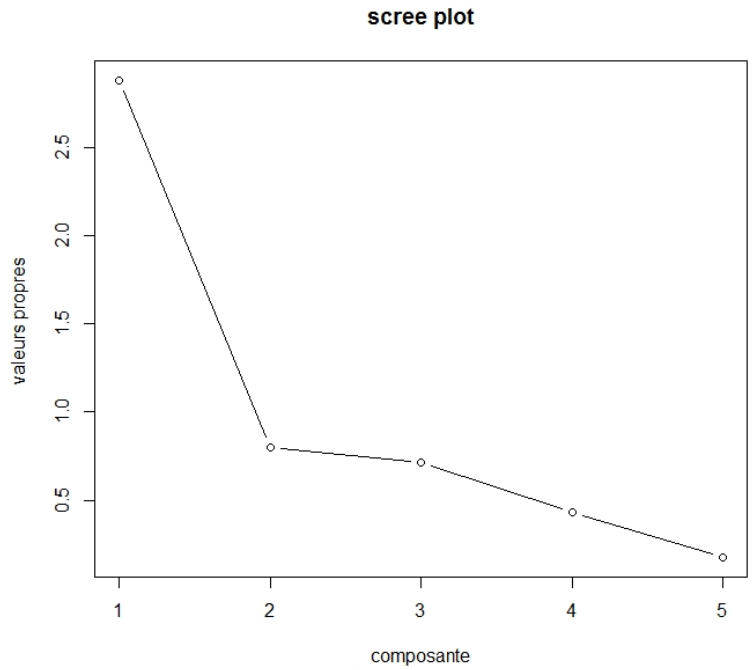


FIGURE 4.1 – Eboulis des Valeurs Propres

Dans ce qui suit on va appliquer la règle suivante :

1ère règle : Le "scree test" ou test du coude. On observe le graphique des valeurs propres et on ne retient que les valeurs qui se retrouvent à gauche du point d'inflexion,dans notre étude un changement de pente est décelé après la troisième valeur propre inclus,le "Scree-test " de Cattell, nous permet de garder deux axes factoriels portant 73,64% de l'inertie totale. Figure4.2 Page 71.

- Graphe d'inertie :

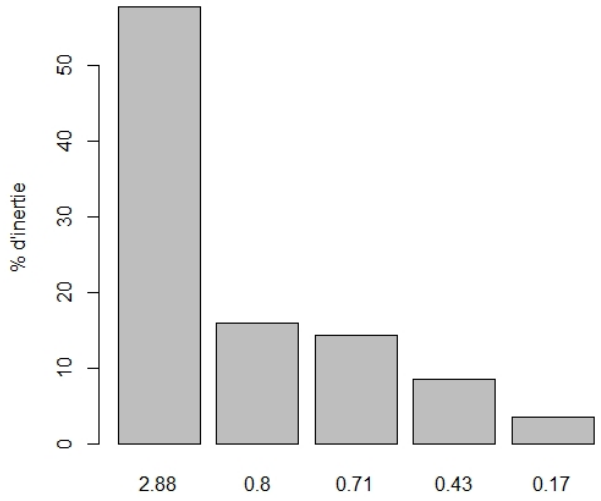


FIGURE 4.2 – Graphe d'inertie

On voit clairement que la première valeur propre 2.88 présente une inertie de 56 % sur le premier axe factoriel puis la deuxième valeur qu'est de 0. 79 avec une inertie de 15% sur le deuxième axe factoriel avec un total de 73,64% de la suite des inerties se trouve dans la figure 4.12 page 79

Résumé des Résultats sur les données selon le choix des valeurs propres

-COORD : est la corrélation entre les variables d'origine et les nouvelles variables synthétiques (axes principaux). On interprète ce coefficient comme n'importe quelle corrélation linéaire.

-COS^2 : représente la qualité de la représentation et la répartition de la variables sur les différents facteurs. La somme horizontale sera égale à 100%.

-CTR : représente la contribution de chaque variable à la construction du facteur(axes). La somme verticale est de 100%.

Les différents résultats sont résumés dans les tableaux suivants :

- Résultats sur les Individus : Le graphe des individus voir la figure 4.3 page 73

Individus

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
1	2.078	0.006	0.000	0.000	-0.217	0.423	0.011	-1.983	39.316	0.911
2	2.199	-0.885	1.941	0.162	1.665	24.800	0.573	0.154	0.239	0.005
3	2.702	1.365	4.617	0.255	-1.568	22.002	0.337	-0.147	0.217	0.003
4	1.819	-0.665	1.095	0.134	0.691	4.266	0.144	-1.201	14.419	0.436
5	3.946	3.745	34.750	0.901	1.225	13.413	0.096	0.004	0.000	0.000
6	2.323	2.103	10.955	0.820	0.699	4.365	0.090	0.219	0.481	0.009
7	1.430	-0.960	2.281	0.450	-0.653	3.819	0.209	0.265	0.700	0.034
8	1.162	1.064	2.805	0.839	-0.256	0.586	0.049	0.288	0.828	0.061
9	1.255	-0.564	0.789	0.202	-0.901	7.266	0.516	0.307	0.942	0.060
10	0.584	0.372	0.342	0.405	-0.210	0.393	0.129	0.161	0.258	0.076
11	2.074	-0.420	0.437	0.041	0.136	0.165	0.004	1.980	39.202	0.912
12	3.964	-3.809	35.945	0.923	0.866	6.706	0.048	0.165	0.273	0.002
13	1.366	-0.078	0.015	0.003	-1.078	10.387	0.622	0.275	0.756	0.041
14	1.443	-1.275	4.027	0.781	-0.397	1.410	0.076	-0.487	2.369	0.114

- Résultats sur les variables : Le cercle de corrélation des variables voir la figure 4.4 page 74

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
QV	-0.727	18.313	0.528	0.477	28.442	0.227	-0.354	17.524	0.125
PVU	0.747	19.375	0.559	0.542	36.791	0.294	0.128	2.284	0.016
VID	0.836	24.218	0.698	0.379	17.988	0.144	0.042	0.251	0.002
VISCO	0.839	24.390	0.703	-0.332	13.769	0.110	0.099	1.370	0.010
P.ECOU	-0.629	13.703	0.395	0.155	3.010	0.024	0.749	78.571	0.561

4.4.4 Interprétations

Figures selon l'axe 1 et 2

- Graphe des individus :

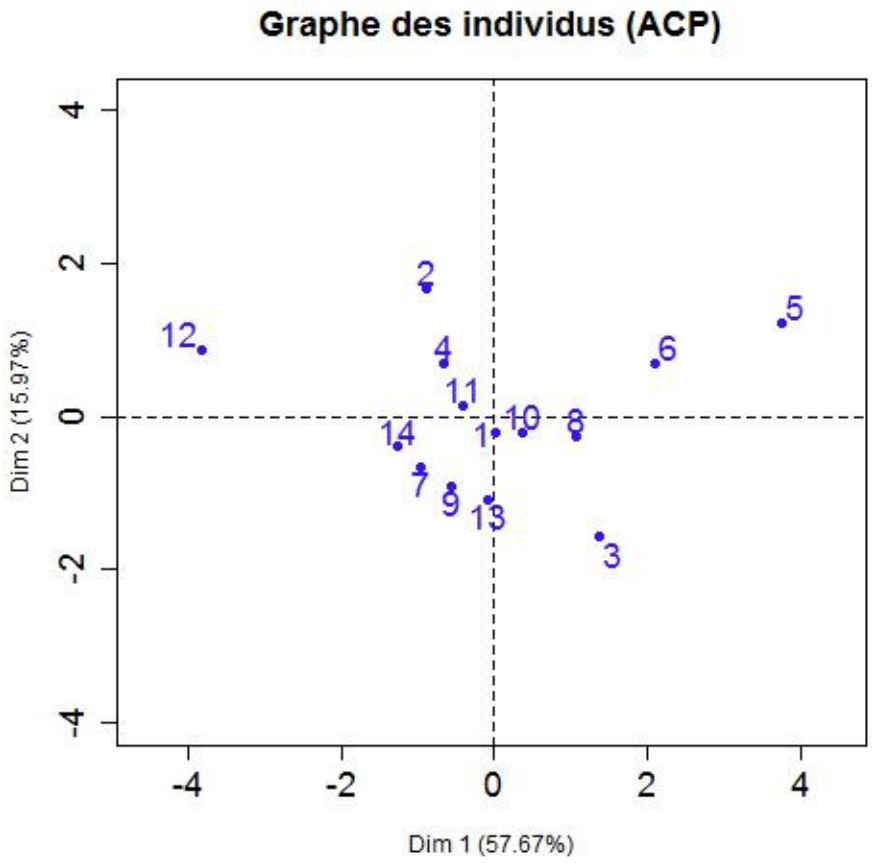


FIGURE 4.3 – Graphe des Individus selon Les Cos^2 et contributions

- Cercle de corrélation des variables :

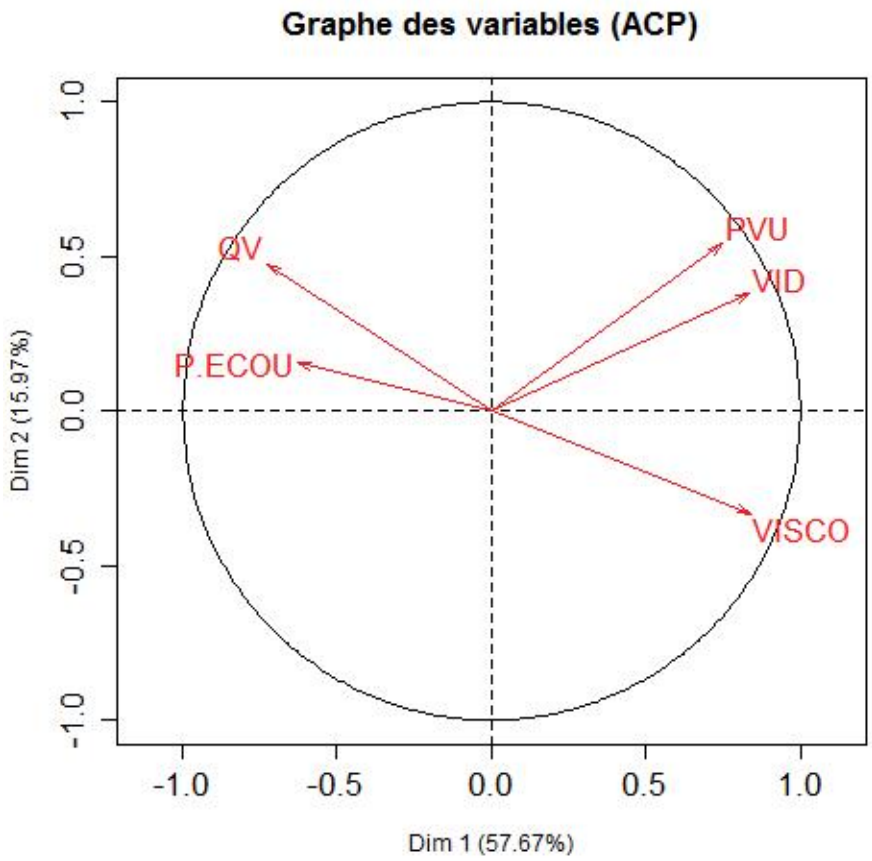


FIGURE 4.4 – Cercle de corrélation des variables

- Graphe des variables et individus :

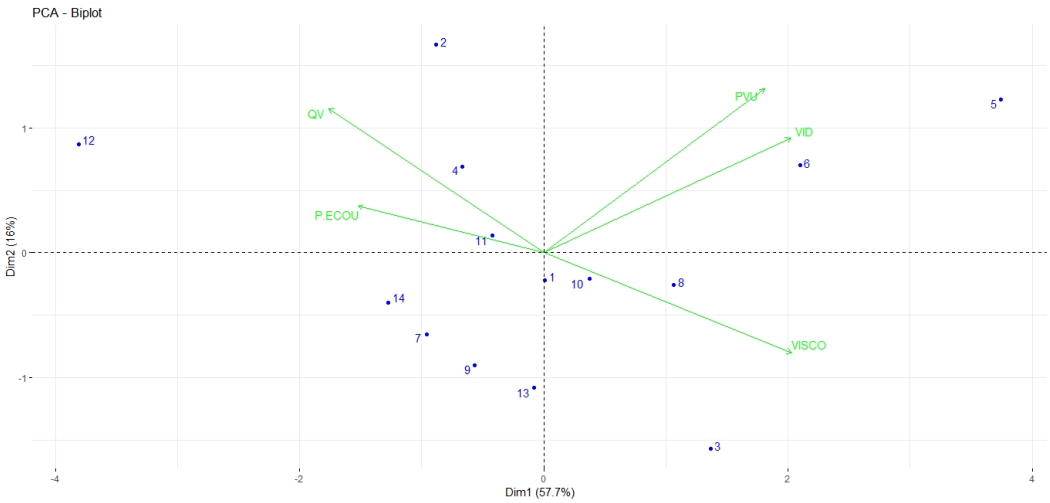


FIGURE 4.5 – Graphe variables et individus

intérprétation selon l'axe 1 et 2

Après la simulation des données avec le logiciel R où on a utilisé les différents pack-ages pour analyser ces données, suite à cette analyse on a pu avoir les résultats qui figure dans la fin du chapitre 4 et on procèdera à l'interprétation de ces dernières :

- L'axe 1 :

Comme les figures 4.3 4.4 4.5 pages 73 74 74 nous montre, on voit bien que les variables VID,VISCO et PVU ont des valeurs proches de 1 cela signifie que leurs valeurs sont liées aux coordonnées de l'axe1,autrement dit les huiles qui sont à gauche ont des faible coordonnées sur l'axe1 en revanche leurs valeurs de VID,VISCO et PVU sont faible. comparons maintenant aux huiles(individus) qui se trouvent aux proximités du centre (milieu) qu'ont des valeurs au alentours de la moyennes en PVU,VID et VISCO par contre les huiles(individus) qui sont à droite ont des valeurs élevées en qualité de VID et VISCO par rapport aux autres huiles au même aux autres variables.

On a toutes les variables liées positivement à l'axe 1 sont les huiles qui se trouvent à droite du graphe et qui ont des fortes valeurs en PVU,VID et VSICO (exp :5,6) contrairement aux variables liées négativement à l'axe 1 qui se trouve à droite du graphe et qui ont des faibles valeurs pour P.ECOU et Q.V à gauche du graphe on voit bien que les les huiles ont des fortes valeurs en P.ECOU et Q.V contrairement aux PVU,VID et VISCO.

- L'axe 2 :

Les corrélations sont moins fortes,ceci est normal car cet axe est moins important par rapport à l'axe 1 sauf PVU (0.54).

D'après les figures 4.6 4.7 page 76 76 on voit bien que les huiles qui se trouvent en hauts du graphe ont des valeurs élevés en PVU,les huiles qui ont des valeurs élevées en haut du graphe ont des valeurs faibles par rapport aux variables qui se trouvent en dessus de l'axe 2 contrairement aux huiles qui se trouvent en dessous de l'axe 2 qui ont des fortes valeurs en VISCO, en revanche des faibles valeurs par rapport aux autres variables. Voir l'exemple précédent :3.7 pages 56

Figures selon l'axe 1 et 3

- Graphe des individus :

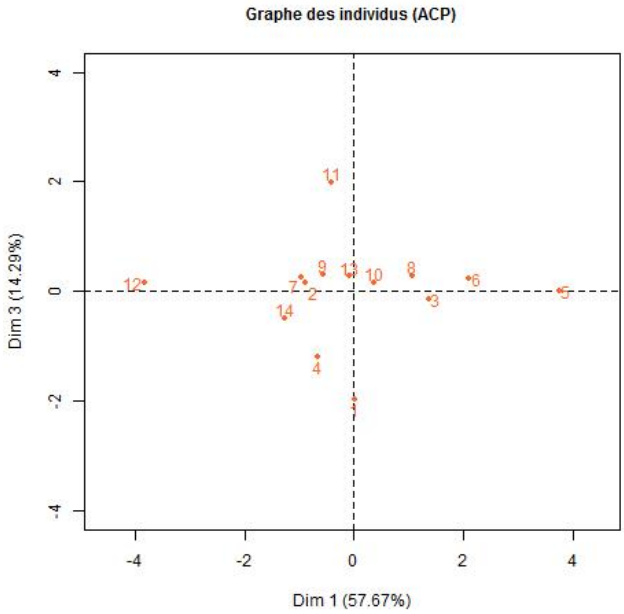


FIGURE 4.6 – Graphe des Individus selon la 1 ère et 3 ème valeures propres

- Cercle de corrélation des variables :

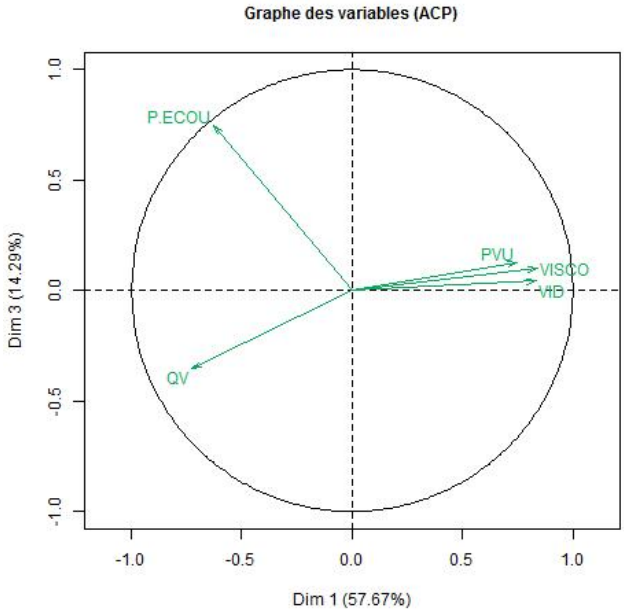


FIGURE 4.7 – Graphe des variables selon la 1 ère et 3 ème valeures propres

intérprétation selon l'axe 1 et 3

- L'axe 1 :

D'après le cercle des corrélations et le graphe des individus Figure 4.6 ,4.7 on voit bien que les variables PVU,VSCO et VID sont très liées positivement à l'axe 1 ce qui implique que les huiles qui se trouvent à droite du graphe des individus ont des fortes valeurs en ces variables et des valeurs faibles en P.ECOU et QV.

Par contre les deux variables QV,P.ECOU sont liées négativement à l'axe 1 en revanche les huiles qui se trouvent à gauche ont des fortes valeurs en QV et P.ECOU et des faibles valeurs par rapport aux autres variables.

Les individus qui contribuent le plus à la construction de l'axe 1 sont les proches de cet axe et qu'ils contiennent la maximum d'informations et parmi ces individus les huiles (3,5,6,12,14) contribuent avec un pourcentage de 90.4 % .

- L'axe 2 :

On voit clairement que les corrélations entre les variables PVU,VISCO,VID et QV sont très faibles avec une contribution totale de 22 % sauf la variable P.ECOU qui et liée avec 78 % à la construction de l'axe 2.

Les huiles qui se trouvent en dessus du graphe ont des fortes valeurs en P.ECOU et des valeurs faibles en QV et des valeurs moyenne en PVU,VISCO et VID qui sont faiblement et positivement liées à l'axe 2.

Les huiles qui se trouvent en dessous du graphe ont des fortes valeurs en QV et faibles valeurs en P.ECOU et moyenne pour le reste des variables.

Les individus huiles (1,4,11,14) parmi les autres individus contribuent avec 95,2 % à la construction de l'axe 2.

4.5 Annexes

- Tableau de données :

```
> annee2015
                    désignation.année.2015      QV      PVU   VID  VISCO P.ECOU
1     NAFTILIA SUPER  huiles moteurs essence   50.040 210058.2 15000 14.800    -34
2  CHELIA SUPER TD 15W40 huiles moteur diesel  93.114 261195.6 18000 14.367    -15
3                           FODDA Engrenages   20.700 179744.9 16000 24.000    -25
4                     NAFTILIA SYNTH 10W40 P   71.427 320636.6 12500 14.000    -27
5                    NAFTILIA SYNTH ECO 5W30    1.371 489675.5 22500 21.000    -29
6                  NAFTILIA SYNTHETIQUE PLUS    8.807 443765.6 17500 19.000    -25
7                  NAFTILIA VP SUPER 15W40 P   25.774 236713.2 11000 15.500    -19
8                                  TASFALOUT    2.160 308930.0 16000 18.000    -23
9                      TASSADIT A2 graisses    11.160 221320.0 12000 16.000    -20
10                          TESSALA graisse    13.986 267254.9 15500     NA    -22
11                                TILIA B       3.420 296147.7 14500 15.540     -9
12                          TISKA Hydraulique 125.640 141999.0 10500 11.400     -9
13                                 TORADA       0.360 163230.0 15000 16.500    -21
14                             TORBA turbine    55.800 161810.0 13000 14.900     NA
```

FIGURE 4.8 – Données de l'année 2015

- Tableau des données quantitatives avec les données manquante :

```
> X
        QV      PVU   VID  VISCO P.ECOU
1   50.040 210058.2 15000 14.800    -34
2   93.114 261195.6 18000 14.367    -15
3   20.700 179744.9 16000 24.000    -25
4   71.427 320636.6 12500 14.000    -27
5    1.371 489675.5 22500 21.000    -29
6    8.807 443765.6 17500 19.000    -25
7   25.774 236713.2 11000 15.500    -19
8    2.160 308930.0 16000 18.000    -23
9   11.160 221320.0 12000 16.000    -20
10  13.986 267254.9 15500     NA    -22
11   3.420 296147.7 14500 15.540     -9
12 125.640 141999.0 10500 11.400     -9
13   0.360 163230.0 15000 16.500    -21
14  55.800 161810.0 13000 14.900     NA
```

FIGURE 4.9 – Tableau des données quantitatives avec les données manquantes

- Données de l'année 2015 après estimation des valeurs manquantes :

```
> res.imputex$completeObs
           QV      PVU   VID  VISCO    P.ECOU
 [1,]  50.040 210058.2 15000 14.800 -34.00000
 [2,]  93.114 261195.6 18000 14.367 -15.00000
 [3,]  20.700 179744.9 16000 24.000 -25.00000
 [4,]  71.427 320636.6 12500 14.000 -27.00000
 [5,]   1.371 489675.5 22500 21.000 -29.00000
 [6,]   8.807 443765.6 17500 19.000 -25.00000
 [7,]  25.774 236713.2 11000 15.500 -19.00000
 [8,]   2.160 308930.0 16000 18.000 -23.00000
 [9,]  11.160 221320.0 12000 16.000 -20.00000
[10,]  13.986 267254.9 15500 16.539 -22.00000
[11,]   3.420 296147.7 14500 15.540  -9.00000
[12,] 125.640 141999.0 10500 11.400  -9.00000
[13,]   0.360 163230.0 15000 16.500 -21.00000
[14,]  55.800 161810.0 13000 14.900 -21.38462
```

FIGURE 4.10 – Données de l'année 2015 après estiamtion des valeurs manquantes

- Matrice des corrélation et son déterminant :

```
> round(cor(res.imputex$completeObs),2)
          QV   PVU   VID VISCO P.ECOU
QV      1.00 -0.40 -0.38 -0.66   0.31
PVU    -0.40  1.00  0.69  0.39  -0.33
VID    -0.38  0.69  1.00  0.63  -0.40
VISCO  -0.66  0.39  0.63  1.00  -0.45
P.ECOU  0.31 -0.33 -0.40 -0.45   1.00
> round(det(cor(res.imputex$completeObs)),2)
[1] 0.12
```

FIGURE 4.11 – Matrice des corrélations

- Les valeurs propres :

```
> eig.val <- get_eigenvalue(res.acpx)
> eig.val
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  2.8833361        57.666721                    57.66672
Dim.2  0.7986337        15.972674                    73.63940
Dim.3  0.7145957        14.291914                    87.93131
Dim.4  0.4284915         8.569829                    96.50114
Dim.5  0.1749431         3.498861                   100.00000
> inertie <- acpx$sdev^2/sum(acpx$sdev^2)*100
> round(inertie,3)
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
57.667 15.973 14.292  8.570  3.499
```

FIGURE 4.12 – Les valeurs propres et l'inertie

4.6 Conclusion

Durant l'année 2015 et après notre application on contacte bien qu'y ont a des produits qui influencent sur le chiffre d'affaires d'une part et sur la rentabilité de ces derniers quoique leurs caractéristiques techniques (variable) sont différentes et on constate que parmi ces individus y a parmi eux des huiles qu'ont enregistrées des fortes valeurs pertinentes comme les produits (2,3, 5,6, 12) et des variables aussi plus influentes durant l'année 2015 comme VID, VISCO, PVU.

D'après nos différentes données récupérer auprès de l'entreprise NAFTAL district oued aissi. L'huile hydraulique occupe la part du lion selon la quantité vendue qui est de 125 tonnes/an qui exprime une consommation trop élevée dans la zone de Tizi Ouzou dû à l'utilisation des pompes et vérins et des matériaux industrielles hydraulique. Et une quantité de 93 tonnes/an enregistrés des huiles à moteur diesel du à l'élargissement du parc automobile, qui à connue une augmentation donc le marché a plus de demande de ce type de lubrifiant.Contrairement aux huiles synthétique avec des faibles quantités en raison des prix de ces derniers et en raison de certaines industries nécessitant des performances plus élevés.

Conclusion générale

Conclusion Générale

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Il s'agit de résumer l'information contenue dans un fichier en un certain nombre de variables synthétiques, combinaisons linéaires des variables originelles. On les appelle « composantes principales », ou « axes factoriels », ou tout simplement « facteurs ». Nous devons les interpréter pour comprendre les principales idées forces que recèlent les données.

L'ACP est une méthode très efficace pour représenter des données corrélées entre elles. Elle est largement utilisée dans des études de marché, d'opinion et de plus en plus dans le domaine industriel.

Notre étude est basé sur les lubrifiants qui sont des produits consommables assurant diverses fonctions dans les mécanisme d'où ils interviennent. Ils peuvent être utiliser pour refroidir,nettoyer,étancher,lubrifier.la combinaison de ces rôles permet de réduire l'usure des pièces en mouvement et garantir la longévité des mécanisme.ainsi investir dans une lubrification adapter permet d'éviter les surcouts de maintenance.

Dans notre cas on peux dire que les lubrifiants (huiles) sont présent dans la majorité des secteurs d'activités à grande consommation on trouve essentiellement les activités industrielles pour lesquelles le rendement des chaines de production sont des critère de compétitivité.le rôle même des lubrifiants est de permettre d'atteindre ces rendements tous en protégeant les mécanismes des usures.Pour les raison semblable on retrouve également le secteur du transport,qu'il soit de tourisme ou commercial.

Bibliographie

Bibliographie

Ouvrages :
[1] Fiches technique des Lubrifiants de l'entreprise NAFTAL.
[2] Chiffres d'affaires CDS de l'année 2015 de l'entreprise NAFTAL.
[3] G. Saporta, « Probabilités, analyse des données et statistique »,
Dunod, 2006 ; pages 155 à 179.
[4] H.Leila « Analyse de données » polycopié 2015/2016 ; pages 03 à 12 .
[5] L.Lebart, A.Morineau,M.Piron « Statistique exploratoire multidimensionnelle »
Dunod,1995 ;pages 32 à 57 .
[6] F.Bertrand,M.Maumy-Bertrand ; « Initiation à la statistique avec R ».
[7]V.Richard « S'initier à l'analyse des données avec le logiciel R ».
[8] F. Husson, S.Lê, J.Pagès « Analyse de données avec R » 2éme édition 2016.
[9]Analyse en Composantes Principales (ACP) Principes et pratique,
R.RAKOTOMALALA,Lyon 2.

Sites internet :
http ://www.naftal.dz/
http ://wikistat.fr/
http ://tutoriels-data-mining.blogspot.com/
https ://cran.r-project.org

Résumé

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Il s'agit de résumer l'information contenue dans un fichier en un certain nombre de variables synthétiques, combinaisons linéaires des variables originelles. On les appelle « composantes principales », ou « axes factoriels », ou tout simplement « facteurs ». Nous devons les interpréter pour comprendre les principales idées forces que recèlent les données.

Dans ce mémoire, On a élaboré un plan de travail qui se compose de deux parties, une partie théorique et une partie pratique. Dans un premier lieu on a cité les différents rappels algébriques et principes d'analyse en composantes principales (ACP), en seconde partie l'application de l'ACP pour les nos données récupérée auprès de l'entreprise NAFTAL 2015.

Notre étude est basé sur les lubrifiants qui sont des produits consommables assurant diverses fonctions dans les mécanisme d'où ils interviennent. Ils peuvent être utiliser pour refroidir,nettoyer,étancher,lubrifier.la combinaison de ces rôles permet de réduire l'usure des pièces en mouvement et garantir la longévité des mécanisme.ainsi investir dans une lubrification adapter permet d'éviter les surcouts de maintenance.

Mots clés :

ACP : Analyse en composantes principales .
$\mathbf{X}(\mathbf{n}, \mathbf{p})$:Tableau des donnés brutes.
$\mathbf{Y}(\mathbf{n}, \mathbf{p})$:Tableau des donnés centrées .
$\mathbf{Z}(\mathbf{n}, \mathbf{p})$:Tableau des donnés centrées réduites.
QV : Quantités vendus (Tonnes).
PVU : Prix de vente unitaire (Dinars).
VID : Vidange (Kilométres).
VISCO : Indice de viscosité (centistoke).
P.ECOU : Indice point d'écoulement (degrés).