Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté de Science
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Licence en Informatique

Thème

Développement d'une Application basée sur l'Analyse en Composantes Principales

Réalisé par :

- **BENALLAL Zyneb**
- **TAHRAOUI Hayet**

Présenté le 10 Juin 2014 devant la commission d'examination composée de MM.

- *CHAOUCHE RAMADANE Lamia* *(Encadreur)*
- *LAHASAINI Mohammed* *(Examinateur)*
- *BELHOUCINE Amine* *(Examinateur)*

Année universitaire : 2013-2014

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience et le courage d'accomplir ce Modeste travail.

Et c'est avec une immense reconnaissance que nous tenons à remercier notre encadreur Mme « CHAOUCHE RAMDANE Lamia » pour son précieux conseil et son aide durant toute la période du travail.

On lui présente donc nos sentiments de gratitude.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Zyneb Et Hayet

Dédicaces

Je dédie ce modeste travail à :

Mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, que dieu te garde dans son vaste paradis, à toi mon père.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

A mes frères et mes sœurs, qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés avec ses précieux conseils.

A toute ma famille, et mes amis.

C'était avec toi mon binôme « Hayet » que j'ai partagé les bons moments pour que nous puissions achever notre mémoire dans les meilleures conditions.

A tous qui me connaisse, et m'aime, je vous dis merci.

A toute personne qui a contribué de prés ou de loin à la réalisation de ce mémoire.

Zyneb

Dédicaces

C’est avec mon énorme plaisir, un cœur ouvert et une joie immense, que je dédie ce modeste travail tout d’abord à mes parents pour leurs amour, leurs sacrifices et leurs encouragements qui ont fait de moi ce que je suis aujourd’hui.

A ma sœur Fatima Zohra qui n’a jamais cessé de m’encourager pour qui je souhaite la réussite pour sa soutenance.

A mes frères Sofiane, Salim, et Djawed merci pour tout ce que vous avez fait pour moi.

A mon binôme Zyneb avec qui j’ai partagé les joies et les difficultés durant ce projet merci pour tous.

A toute personne qui me connaisse et me considère comme amie.

Aux personnes qui m'ont encouragé et motivé, qui n'ont cessé d'œuvré pour ma réussite et pour mon bonheur.

Hayet

Table Des Matières

Introduction Générale

Il n'y a pas très longtemps, on ne pouvait pas traiter un tableau de 3000 lignes et 300 colonnes. L'apparition et le développement des ordinateurs a du coup levé cet obstacle de calcul, et a permis la conservation et l'exploitation des grandes masses de données. Cette amélioration continue de l'outil informatique a fortement contribué au développement et à la vulgarisation de nombreuses méthodes statistiques, devenues maintenant d'usage assez courant.

Aujourd'hui, des vastes données d'enquêtes sont dépouillées et, fournissent de grands tableaux qui se prêtent aisément à l'interprétation. Des données issues d'investigations spécifiques sont rassemblées et constituent une masse importante et apparemment indéchiffrable d'informations mais, qu'on peut désormais traiter sans difficultés.

Dans ce document, on s'est intéressé à l'analyse de données, et surtout à l'une des méthodes multidimensionnelles la plus employée connue sous le nom d'Analyse en Composantes Principales « ACP ».

L'opération de l'Analyse en Composantes Principales consiste à passer d'un tableau des données brutes contenant toute l'information recueillie sur le phénomène que nous souhaitons étudier à certaines représentations visuelles des données. Cette opération entraînera une certaine perte "d'information" que l'on essaie de minimiser. En échange, on obtient un gain en "signification", en particulier grâce aux représentations graphiques. Autrement dit, on passe du « magma » des données d'origine à des graphiques interprétables par l'utilisateur.

Nous allons essayer dans ce travail de présenter une application qui s'intéresse à réaliser une ACP sur un tableau de données.

Au long de ce mémoire nous allons présenter trois chapitres :

- Le premier : donnera un aperçu global sur l'Analyse des Données.
- Le deuxième : à travers ce dernier nous allons voir l'Analyse en Composantes Principales d'une façon détaillée.
- Le dernier chapitre sera consacré à définir notre application.
- Et l'annexe où nous allons interpréter un exemple avec les détails de calcul.

Et on termine par une conclusion et une perspective de notre travail.

CHAPITRE I

Analyse Des Données

*I-1 - **Introduction :***

L'Analyse des données est une technique essentiellement descriptive, a pour but de décrire, de réduire, de classer et de clarifier les données.

Cette approche descriptive et multidimensionnelle permet de dire que l'Analyse des Données, c'est de la ''statistique descriptive perfectionnée''.

Cette analyse est un sous domaine des statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données ainsi qu'à en tirer une information statistique qui sert à décrire de façons plus succincte les principales informations contenues dans ces données. On peut aussi chercher à classer les données en différents sous groupes Plus homogènes [1].

Par exemple l'âge, le sexe et la catégorie socioprofessionnelle des joueurs de golf peuvent être étudiés simultanément [1].

*I-2- **Définition :***

- Pour J-P. Fénelon ''l'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement [2].
- Dans l'acception française, la terminologie « analyse des données » désigne un sous-ensemble de ce qui est appelé plus généralement la « statistique multi variée ». L'analyse des données est un ensemble de techniques descriptives, dont l'outil mathématique majeur est l'algèbre matricielle, et qui s'exprime sans supposer a priori un modèle probabiliste [3].

I-3- Variables et données:

I-3-1 Définition :

Une variable est une caractéristique étudiée pour une population donnée. Le sexe, la couleur préférée, le nombre de téléviseurs de votre foyer ou encore l'âge sont des variables.

Des milliers de variables peuvent être sujet aux études.

Il existe 2 types de variables [4]:

I-3-2 Les variables qualitatives:

Sont des variables représentées par des qualités, telles que le sexe, le programme d'études ou encore l'état civil. Les variables qualitatives s'expriment en modalités.

Les modalités sont comme des choix de réponses aux variables étudiées.

Pour les variables qualitatives, il y a encore 2 types de variables différentes [4]:

I-3-2-1 Les variables qualitatives nominales:

Sont des variables qui correspondent à des noms, il n'y a aucun ordre précis. Ce sont seulement des mots dans le désordre. Par exemple, le sexe a 2 modalités possible : féminin ou masculin. Ce sont des noms et peu importe l'ordre dans lequel on le présente. C'est exactement la même chose pour la profession ou encore le mets préféré, ce sont uniquement des noms ou l'ordre n'a pas d'importance [4].

I-3-2-2 Les variables qualitatives ordinales:

Sont des variables qui contiennent un ordre. Par exemple, le degré de satisfaction par rapport à votre fournisseur cellulaire. Les différentes modalités seraient : très satisfait, satisfait, insatisfait, très insatisfait. Les variables qualitatives ordinales sont très souvent des degrés de satisfaction, d'approbation, etc. [4]

I-3-3 Les variables quantitatives:

Sont quant à elles des variables représentées par des quantités telles que l'âge, le poids et la taille. Elles s'expriment en valeurs. Les valeurs représentent les choix de réponses aux variables quantitatives.

Il en va de même pour les variables quantitatives, 2 types différents, les variables quantitatives discrètes et continues [4].

I-3-3-1 Les variables quantitatives discrètes:

Sont des valeurs que l'on peut énumérer, il est inutile d'utiliser des classes pour les exprimer.

Par exemple, le nombre de personnes dans le ménage ou bien le nombre de présence au centre commercial par mois sont autant de possibilités pour des variables quantitatives discrètes [4].

I-3-3-2 Les variables quantitatives continues:

Sont des valeurs très nombreuses dont l'énumération serait fastidieuse. Il est donc préférable de les exprimer en classe de largeur égale. Par exemple, le poids est une variable quantitative continue.

La figure 1 résume bien le tout [4].

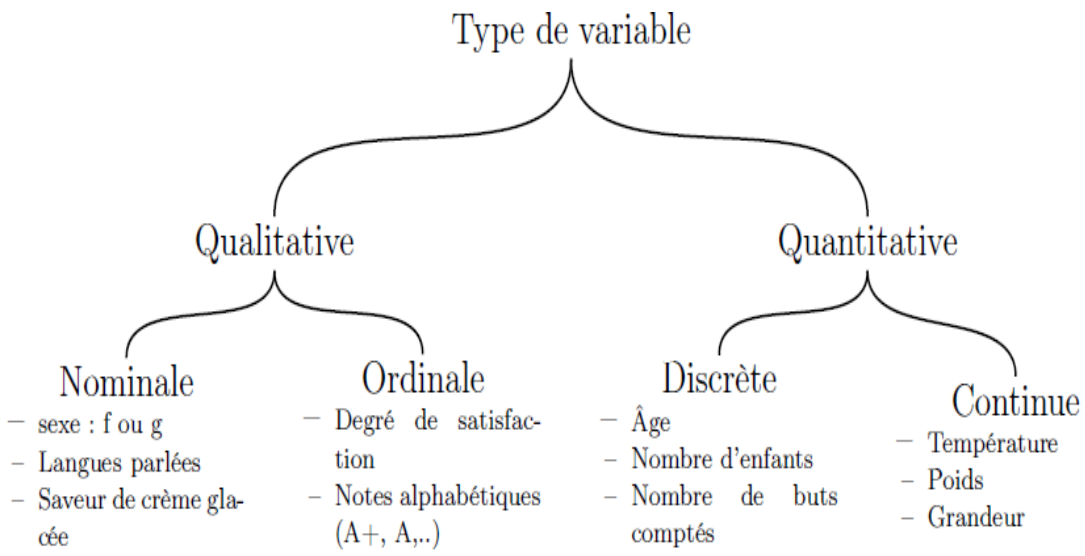
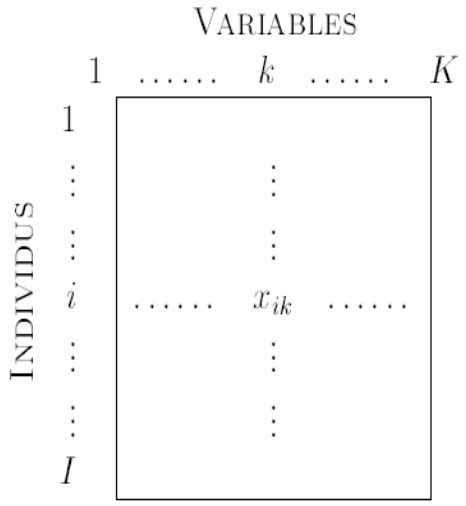


Figure I.1. Diagramme des différents types de variables [4].

I-4- Tableau de données :

Les méthodes d'analyse de données supposent souvent une organisation des données particulière, naturelle, mais parfois difficile à réaliser selon l'application et les données. Le choix d'un tableau permet une organisation dans le plan de toutes les données et ainsi de traiter simultanément toute l'information. Ainsi la plupart des méthodes nécessitent une organisation des données présentée par le schéma suivant selon les données ce tableau est quelque peu modifié, mais l'idée de tableau reste présente dans toutes les méthodes d'analyse de données [5].



Représentations des données [5]

Ainsi les observations ou individus ou encore unités statistiques sont représentés en ligne et sont chacun décrits par des variables ou caractères.

D'après le schéma ci-dessus X_{ik} est donc la valeur de la variable k pour l'individu i avec k = 1;… ; K et i = 1;…; I. Par abus de notations, pour des considérations de simplification de celles-ci, I représente à la fois le nombre d'individus et l'ensemble des individus {1; ..; i;.. ; I}, de même K représente le nombre de variables et l'ensemble des variables {1;…; k;…; K}.

Cette représentation des données peut faciliter la lecture de petits tableau, i.e. lorsqu'il y a peu de données. Cependant, dès lors que la taille du tableau est grand, ou que nous recherchons des relations entre plus de deux individus ou plus de deux variables, cette représentation et les techniques simples de la statistique descriptive ne suffisent plus [5].

I-5- Méthodes d'analyses des données :

I-5-1 Analyse par réduction des dimensions :

La représentation des données multidimensionnelles dans un espace à dimension réduite est le domaine des analyses en composantes principales, analyse factorielle des correspondances, analyse des correspondances multiple, analyse canonique, positionnement multidimensionnel, analyse factorielle multiple. Ces méthodes permettent de représenter le nuage de points à analyser dans un plan ou dans un espace à trois dimensions sans trop de perte d'information, et sans hypothèse

statistique préalable. En mathématiques, elles exploitent le calcul matriciel et l'analyse des vecteurs et des valeurs propres [3].

I-5-1-1 Analyse en composantes principales :

L'analyse en composantes principales est une méthode descriptive utilisée pour réduire les K variables en des combinaisons linéaires des K' variables initiales, que leur variance soit maximale et que les nouvelles variables soient orthogonales entre elles suivant une distance particulière.. En ACP, les variables sont quantitatives.

Les composantes, les nouvelles variables, définissent un sous-espace à p dimensions sur lequel sont projetés les individus avec un minimum de pertes d'information. Dans cet espace le nuage de points est plus facilement représentable et l'analyse est plus aisée.

Son objectif est de représenter sous forme graphique l'essentiel .de l'information contenue dans un tableau de données quantitatif [3].

I-5-1-2 Analyse factorielle des correspondances :

Le but de l'AFC est de trouver des liens ou correspondances entre deux variables qualitatives (nominales). Cette technique traite les tableaux de contingence de ces deux variables. En fait, une AFC est une ACP sur ces tableaux dérivés du tableau initial munis de la métrique du Khi- deux χ^2. Le principe de l'AFC est identique à celui de l'ACP. Les axes explicatifs qui sous-tendent le tableau de fréquences de deux variables qualitatives sont recherchés et présentés dans un graphique.

I-5-1-3 Analyse des correspondances multiples :

L'ACM se propose d'analyser k ($k \geq 2$) variables qualitatives d'observations sur i individus. Comme il s'agit d'une analyse factorielle elle aboutit à la représentation des données dans un espace à dimensions réduites engendré par les facteurs. L'ACM est l'équivalent de l'ACP pour les variables qualitatives et elle se réduit à l'AFC lorsque le nombre de variables qualitatives est égal à 2.

Formellement, une ACM est une AFC appliquée sur le tableau disjonctif complet, ou bien une AFC appliquée sur le tableau de Burt, ces deux tableaux étant issus du tableau initial.

Un tableau disjonctif complet est un tableau où les variables sont remplacées par leurs modalités et les éléments par 1 si la modalité est remplie 0 sinon pour chaque individu. Un tableau de Burt est le tableau de contingence des k variables prises deux à deux [3].

I-5-1-4 Analyse canonique :

L'analyse canonique permet de comparer deux groupes de variables quantitatives appliqués tous deux sur les mêmes individus. Le but de l'analyse canonique est de comparer ces deux groupes de variables pour savoir s'ils décrivent un même phénomène, auquel cas l'analyste pourra se passer d'un des deux groupes de variables.

L'analyse canonique généralise des méthodes aussi diverses que la régression linéaire, l'analyse discriminante et l'analyse factorielle des correspondances.

Enfin l'analyse canonique généralisée étend l'analyse canonique ordinaire à l'étude de k groupes de variables ($k > 2$) appliquées sur le même espace des individus. Elle admet comme cas particuliers l'ACP, l'AFC et l'ACM, l'analyse canonique simple, mais aussi la régression simple, et multiple, l'analyse de la variance, l'analyse de la covariance et l'analyse discriminante [3].

I-5-1-5 Positionnement multidimensionnel :

Pour utiliser cette technique les tableaux ne doivent pas être des variables caractéristiques d'individus mais des « distances » entre les individus. L'analyste souhaite étudier les similarités et les dis similarités entre ces individus.

Le positionnement multidimensionnel (« *multidimensional scaling* » ou MDS) est donc une méthode factorielle applicable sur des matrices de distances entre individus .Cette méthode ne fait pas partie de ce qu'on nomme habituellement l'analyse des données. Mais elle a les mêmes caractéristiques que les méthodes précédentes : elle est fondée sur le calcul matriciel et ne demande pas d'hypothèse probabiliste. Les données peuvent être des mesures de k variables quantitatives sur i individus, et dans

ce cas l'analyste calcule la matrice des distances ou bien directement un tableau i*i des distances entre individus [3].

I-5-1-6 Analyse Factorielle Multiple :

L'analyse factorielle multiple (AFM) est dédiée aux tableaux dans lesquels un ensemble d'individus est décrit par plusieurs groupes de variables, que ces variables soient quantitatives, qualitatives ou mixtes. Cette méthode est moins connue que les précédentes mais son très grand potentiel d'application justifie une mention particulière.

L'AFM a pour intérêt de :

- pondère les variables de façon à équilibrer l'influence des différents groupes, ce qui est particulièrement précieux lorsque l'on est en présence de groupes quantitatifs et de groupes qualitatifs.
- fournit des résultats classiques des analyses factorielles : représentation des individus, des variables quantitatives et des modalités des variables qualitatives.
- fournit des résultats spécifiques de la structure en groupe : représentation des groupes eux-mêmes (un point = un groupe), des individus vus par chacun des groupes (un individu = autant de points que de groupes), des facteurs des analyses séparées des groupes (ACP ou ACM selon la nature des groupes) [3].

I-5-1-7 Autres méthodes :

- L'Analyse Factorielle Multiple Hiérarchique.
- L'Analyse Procustéenne Généralisée .
- L'Analyse Factorielle de Données Mixtes .
- L'iconographie des corrélations .
- L'Analyse en composantes indépendantes (ACI).

Ces méthodes permettent surtout de manipuler et de synthétiser l'information provenant de tableaux de données de grande taille. Pour cela, il est particulièrement important de bien estimer les corrélations entre les variables qu'on étudie. On a alors fréquemment recours à la matrice des corrélations (ou la matrice de variance-covariance) entre les variables [3].

I-5-2 Analyse par classification :

La classification des individus est le domaine de la classification automatique et de l'analyse discriminante. Classifier consiste à définir des classes, classer est l'opération permettant de mettre un objet dans une classe définie au préalable. La classification automatique est ce qu'on appelle en exploration de données (« *data mining* ») la classification non supervisée, l'analyse discriminante fait partie des techniques statistiques connues en exploration de données sous le nom de classification supervisée [3].

I-5-2-1 Classification automatique :

Le but de la classification automatique est de découper l'ensemble des données étudiées en un ou plusieurs sous-ensembles nommés classes, chaque sous-ensemble devant être le plus homogène possible. Les membres d'une classe ressemblent plus aux autres membres de la même classe qu'aux membres d'une autre classe. On a comme type de cette classification : le partitionnement hiérarchique. Et pour ce cas, classifier revient à choisir une mesure de la similarité/dis similarité, un critère d'homogénéité, un algorithme, et parfois un nombre de classes composant la partition [3].

I-5-2-1-1 Classification hiérarchique :

Les données en entrée d'une classification ascendante hiérarchique (CAH) sont présentées sous la forme d'un tableau de dis similarités ou un tableau de distances entre individus [3].

La classification ascendante se propose de classer les individus à l'aide d'un algorithme itératif. À chaque étape, l'algorithme produit une partition en agrégeant deux classes de la partition obtenue à l'étape précédente. Il se termine lorsqu'il ne reste qu'une seule classe.

I-5-2-2 Analyse factorielle discriminante :

L'analyse factorielle discriminante (AFD), qui est la partie descriptive de l'analyse discriminante, est aussi connue sous le nom d'analyse linéaire discriminante, d'analyse discriminante de Fisher et d'analyse canonique discriminante. Cette technique projette des classes prédéfinies sur des plans factoriels discriminant le plus possible. Le tableau de données décrit i individus sur lesquels k variables quantitatives et une variable qualitative avec des modalités mesurées. La variable qualitative permet de définir les classes et le regroupement des individus dans ces classes. L'AFD se propose de trouver les variables, appelées variables discriminantes, dont les axes séparent le plus les projections des classes qui découpent le nuage de points.

Comme dans toutes les analyses factorielles descriptives, aucune hypothèse statistique n'est faite au préalable ; ce n'est que dans la partie prédictive de l'analyse discriminante que des hypothèses a priori sont émises.

Une AFD est une ACP effectuée sur les barycentres des classes d'individus constituées à l'aide des modalités de la variable qualitative. C'est aussi une analyse canonique entre le groupe des variables quantitatives et celui constitué du tableau disjonctif de la variable qualitative [3].

I-6- Les logiciels utilisés :

Les méthodes d'analyse de données nées de la recherche universitaire sont depuis longtemps entrées dans le monde industriel. Il y a cependant peu de logiciels qui savent intégrer ces méthodes pour une recherche exploratoire aisée dans les données. Nous citons ici sept logiciels [5] :

I-6-1 SAS :

SAS est un logiciel de statistique très complet et très performant. Il a d'abord été développé pour l'environnement Unix, mais est maintenant accessible sous tout environnement. Il permet une puissance de calcul importante et ainsi est très bien adapté à tous traitements statistiques sur des données très volumineuses. Son manque de convivialité et surtout son prix fait qu'il est encore peu employé dans les entreprises qui ne se dédient pas complètement à la statistique [5].

I-6-2 Splus :

Splus est à la fois un langage statistique et graphique interactif interprété et orienté objet. C'est donc à la fois un logiciel statistique et un langage de programmation. La particularité de ce langage est qu'il permet de mélanger des commandes peu évoluées à des commandes très évoluées. Il a été développé par Statistical Sciences autour du langage S, conçu par les Bell Laboratories. Depuis, Splus est devenu propriété de Math soft après le rachat de Statistical Sciences [5].

I-6-3 R :

Ce logiciel est la version gratuite de Splus. Il est téléchargeable sous www.r-project.org pour tous systèmes d'exploitation. Il souffre également de peu de convivialité et semble encore très peu employé en industrie. De part sa gratuité, il est de plus en plus employé pour la réalisation de cours de statistiques [5].

I-6-4 XlStat :

Excel propose une macro payante permettant d'effectuer quelques méthodes d'analyse de données. Elle est cependant très limitée, utilisable qu'avec Excel sous Windows et de plus payante [5].

I-6-5 UniWin Plus :

Statgraphics est un logiciel de statistiques générales, qui propose un module d'analyse de données de treize méthodes. Développé uniquement pour les environnements Windows, l'accent est porté sur les interfaces graphiques. Statgraphics propose un grand nombre d'analyses statistiques et permet l'utilisation de beaucoup de format de données [5].

I-6-6 MATLAB :

C'est un langage de programmation et un environnement de développement ; il est utilisé à des fins de calcul numérique. Développé par la société The MathWorks. MATLAB permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, et de créer des interfaces utilisateurs [5].

I-6-7 SPAD :

Le logiciel SPAD est toujours maintenu à jour avec de nouvelles méthodes issues de la recherche universitaire. Sa version sous Windows est conviviale ce qui a poussé son achat par de plus en plus d'industriels.

Le souci de coller à une réalité industrielle fait qu'il est employé en enseignement [5].

I-7- Domaine d'application :

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 mis à profit en traitement du signal et des images, où elles sont souvent employées comme prétraitements (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 [5].

I-8- Les objectifs :

Les objectifs que se sont fixés les chercheurs en analyse de données sont donc de répondre aux problèmes posés par des tableaux de grandes dimensions. Les objectifs sont souvent présentés en fonction du type de méthodes, ainsi deux objectifs ressortent : la visualisation des données dans le meilleur espace réduit et le regroupement dans tout l'espace.

Les méthodes de l'analyse de données doivent donc permettre de représenter synthétiquement de vastes ensembles numériques pour faciliter l'opérateur dans ses décisions. En fait d'ensembles numériques, les méthodes d'analyse de données se proposent également de traiter des données qualitatives, ce qui en fait des méthodes capables de considérer un grand nombre de problèmes. Les représentations recherchées sont bien souvent des représentations graphiques, comme il est difficile

de visualiser des points dans des espaces de dimensions supérieures à deux, nous chercherons à représenter ces points dans des plans.

Ces méthodes ne se limitent pas à une représentation des données, ou du moins pour la rendre plus aisée, elles cherchent les ressemblances entre les individus et les liaisons entre les variables. Ces proximités entre individus et variables vont permettre à l'opérateur de déterminer une typologie des individus et des variables, et ainsi il pourra interpréter ses données et fournir une synthèse des résultats des analyses. Nous voyons donc que les deux objectifs précédemment cités sont très liés voir indissociables, ce qui entraîne souvent l'utilisation conjointe de plusieurs méthodes d'analyse de données [5].

I-9- Conclusion :

L'analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Le succès de cette discipline dans les dernières années est dû, dans une large mesure, aux représentations graphiques fournies. Ces graphiques peuvent mettre en évidence des relations difficilement saisies par l'analyse directe des données ; mais surtout, ces représentations ne sont pas liées à une opinion « a priori » sur les lois des phénomènes analysés contrairement aux méthodes de la statistique classique [5].

Parmi les méthodes d'analyse des données, on s'intéresse à l'Analyse en Composantes Principales qui sera détaillé dans le deuxième chapitre.

CHAPITRE II

Analyse En Composantes Principales

II-1- Introduction :

Lorsqu'on étudie simultanément un nombre important de variables quantitatives comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus important (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduit (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques et graphiques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

L'analyse en Composantes Principales (ACP) est un grand classique de « L'analyse des données ».

L'ACP joue dans ce chapitre un rôle central ; cette méthode sert de fondement théorique aux autres méthodes de statistique multidimensionnelle dites factorielles qui en apparaissent comme des cas particuliers [6]**.**

II-2- <u>Définition :</u>

Méthode de base de l'analyse des données, cette méthode recherche à synthétiser l'information contenue dans un tableau croissant des individus (observation) et des caractères (var), cette méthode se prête particulièrement aux données quantitatives continues.

Une analyse sert à :

- Résumer et synthétiser.
- Hiérarchiser l'information contenus dans un tableau de : n lignes (individus) et p colonnes (variables).

Les n individus sont décrits par un nuage de p variables, l'information représentée par ce nuage revient à la dispersion des n points.

- Produire un résumé de cette information c'est projeter ces points dans un espace de dimension inférieur à p le nombre de variables initiales.
- Les axes de ce sous espace sont dits « axes factoriels » ou « facteurs »
- Le résumé est possible dans la mesure ou les variables ne sont pas totalement indépendantes.
- Chaque variable p porte en elle :
  - ✓ Une part d'information originale ou part d'inertie.
  - ✓ Une part d'information redondante avec les autres venant des corrélations entre variables.

C'est cette part d'informations redondante que l'on va regrouper dans le résumé factoriel. Chaque facteur est la combinaison linéaire des « p » variables.

-Les facteurs sont hiérarchisés :

Le premier axe concentre le max d'information :

- C'est l'axe de la plus grande dimension du nuage de point.
- C'est le meilleur résumé dans un espace à une dimension.
- Mais il laisse des résidus (de l'information).

Le deuxième axe concentre le max de l'information restante :

- Il est orthogonal au premier (par construction).
- C'est l'axe de la plus grande dimension résiduelle du nuage de point.
- Associé au 1^{ier} axe ,c'est le meilleur résumé dans un espace à 2 dimension.
- Mais il laisse aussi des résidus.

Le troisième axe prend encore une part d'information moindre :

- Il est orthogonal au deux premiers (toujours par construction)

Et ainsi de suite pour les axes suivants tout que l'on pense qu'iles apportent encore de l'information [7].

II-3- Approche ACP :

II-3-1- Analyse d'un nuage de points :

II-3-1-1- Visualisation d'un nuage de points (inertie) :

La distribution d'individus ou observation à travers deux caractère X_1 et X_2 permet dans un plan (graphe) permet de juger instinctivement la liaison entre les deux variables.

Ainsi dans la représentation de la figure 1.a le plan nous laisse l'idée d'absence de liaison ou plus ou moins de l'absence du constat de liaison dans une première approche rapide, à l'inverse de celui de la figure 1.b indique une présomption de forte liaison [3].

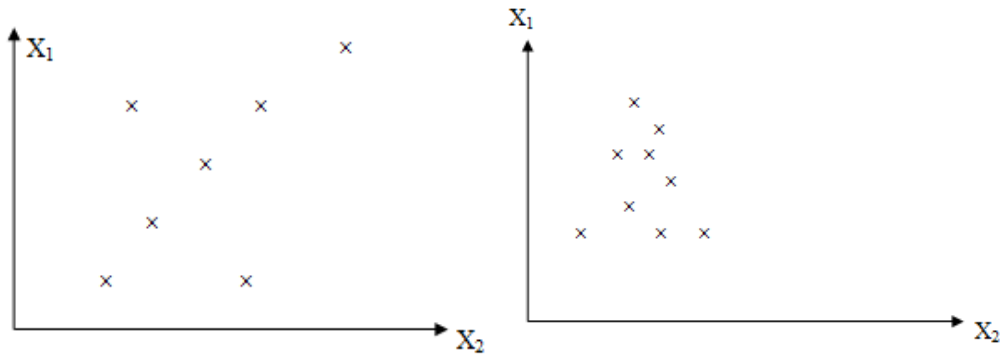


Figure 1.a Figure 2.b

Figure II .1. Nuage de point

II-3-1-2- Précision sur la méthode :

L'ACP sert a traiter les données multidimensionnelles, ceci impose de disposer pour toutes les variables observées qui sont de type numérique et que l'on veut voir s'il y'a des liaisons entre ces variables. Les variables sont supposées définies sur R^n.

Dans le cas le plus général, le tableau X de données initiale est supposé comprendre n individus soit des lignes indicées i de i = 1..n. Les caractères variables seront disposés en colonnes indicées j avec j = 1…p. De manière générale, p sera de dimension largement inférieure à n .

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

Le but d'ACP revient de trouver une bonne représentation des individus dans un espace plus restreint que celui des p variables d'origine (k < p) .Souvent même, on se limitera a une représentation dans le plan principal (k =2) .Ceci revient a rechercher le sous espace affine ou espace E_K de dimension k (k < p souvent k = 2) .

Un espace affine de K dimension est obtenu par une transformation linéaire à partir d'un espace plus riche de p dimension à savoir un centre de gravité [3].

II-3-1-3- Choix de l'origine :

Le point O correspond aux vecteurs de coordonnées toutes nulles, n'est pas forcément une origine satisfaisante car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cet origine .Il apparait plus judicieux de choisir une origine liée au nuage lui-même [3].

II-3-1-4- Le centre de gravité du nuage :

Pour définir, il faut choisir un système de pondération des unités :

$\forall$ i $\epsilon\{1..n\}$, P_i est le poids de l'unité u_i tel que $\sum^{n}_{i=1} P_i = 1$

Pour l'ACP, on choisit de donner le même poids 1/n pour tous les individus. Le centre de gravité G du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables [3] :

$$G_j = 1/n \sum^{n}_{i=1} x_{ij}$$

$$Gj = x_{.j}$$

Prendre G comme origine revient a travailler sur les tableau de données centrées.

$X_c = \{ c_{ij} \in R$ sachant que $c_{ij} = x_{ij} - x_{.j} \}$

On note I_G le moment d'inertie du nuage de points des individus par rapport au centre de gravité $G.I_G = 1/n \sum^{n}_{i=1} d^2(G,u_i) = 1/n \sum^{n}_{i=1} \sum^{p}_{j=1} (x_{ij} - x_{.j})^2$

L'inertie est une mesure de dispersion de nuage des individus par rapport au centre de gravité. Si ce moment d'inertie est petit , alors le nuage est bien centré sur l'origine. Tandis que s'il est grand le nuage est néanmoins écarté ou éloigné du centre de gravité [3].

On peut écrire I_G comme suit :

$$I_G = \sum^{p}_{j=1} \left(1/n \sum^{n}_{i=1} (x_{ij} - x_{.j})^2\right) = \sum^{p}_{j=1} var(v_j)$$

Ou $var(v_j)$:est la variance empirique de la variable v_j .

Sous cette forme on constate que l'inertie totale est égale à la trace de la matrice de covariance V des p variables $I_G = Trace(V) = \sum \lambda i$

$$\begin{pmatrix} Var(v_1) & cov(v_1,v_2) & \dots & cov(v_1,v_j) & \dots & cov(v_1,v_P) \\ cov(v_2,v_1) & Var(v_2) & \dots & cov(v_2,v_j) & \dots & cov(v_2,v_P) \\ \vdots & \vdots & & \vdots & & \\ cov(v_j,v_1) & cov(v_j,v_2) & \dots & var(v_j) & \dots & cov(v_j,v_p) \\ \vdots & \vdots & & \vdots & & \\ cov(v_p,v_1) & cov(v_p,v_2) & & cov(v_p,v_j) & \dots & var(v_p) \end{pmatrix}$$

$$V = \frac{1}{n} X_c^t X_c$$

$$Var(v^j) = 1/n \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2$$

$$Cov(v_j, v_{j'}) = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})(x_{ij'} - \bar{x}_{.j'})$$

$$\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{.j}$$

- Si on veut travailler avec des variables centré et réduit on passe de tableau des valeurs centrés au tableau des valeurs centrés et réduit comme suit :

$$X_{cr} = X_c D$$

$$D = \begin{pmatrix} \frac{1}{\sigma 1} \dots & 0 \dots \dots & 0 \\ & \ddots & \\ 0 \dots & \frac{1}{\sigma j} \dots \dots & 0 \\ & \ddots & \\ 0 \dots & 0 \dots \dots & \frac{1}{\sigma p} \end{pmatrix}$$

Avec l'écart type $\sigma = \sqrt{var\ v_j}$

« D » est la matrice diagonal qui sur sa diagonal les inverse des écarts types.

Si on calcule la matrice de covariance à partir d'un tableau de donnée centrées et réduites X_{cr} on obtient la matrice de corrélation [3].

$$R = \frac{1}{n} X_{cr}^t X_{cr}$$

II-3-1-5- Chois des axes factoriels :

Pour déterminer l'espace de projection à inertie expliqué il faut déterminer ses « K » axes.

- Le premier c'est l'axe à inertie expliqué maximum et pour le déterminer il suffit de calculer l'axe associé au premier vecteur propre de la matrice « V »ou « R », on le désigne par U_1 associe a la plus grande valeur propre λ_1 [3].

L'inertie expliqué par cet axe est égale au % d'information restituée par

$$U_1 = \lambda_1 / (\Sigma^n_{i=1} \lambda i)$$

- Le deuxième c'est l'axe a inertie expliqué maximum et pour le déterminer il suffit de calculer l'axe associe au deuxième vecteur propre de la matrice « V » ou « R », on le désigne par U_2 associe a la plus grande valeur propre λ_2 [3].

L'inertie expliqué par cet axe est égale au % d'information restituée par

$$U_2 = \lambda_2 / \Sigma^n_{i=1} \lambda i$$

Remarque : En pratique on utilise la matrice de corrélation lorsque les variables n'ont pas les mêmes ordres de grandeurs (unités différente par exemple).

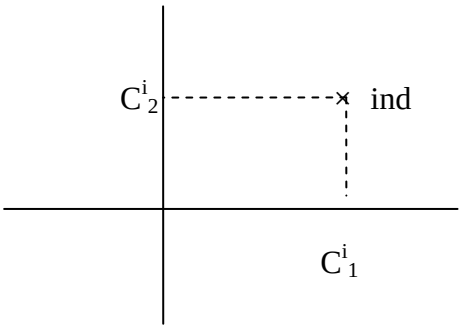
II-3-1-6- Les composantes principales :

A chaque axe est associée une composante principale :

- ✓ La composante C_1 est le vecteur renferment les coordonnées de la projection des individus sur l'axe 1.
- ✓ La composante C_2 est le vecteur renferment les coordonnées de la projection des individus sur l'axe 2 [3].

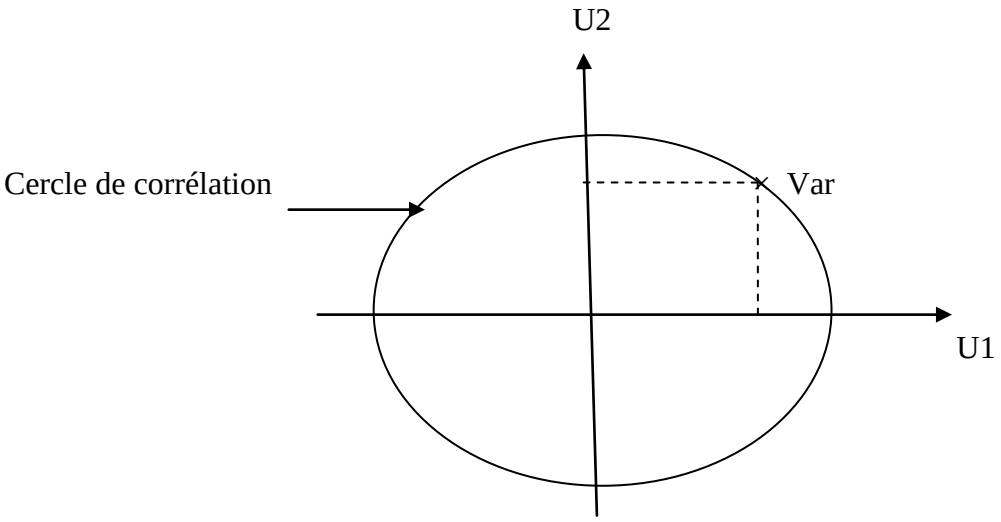
II-3-1-7- Représentation des individus sur le plan principale :

C'est une représentation ou pour deux composantes principales C_1,C_2 on représente chaque individus « i » par un point d'abscisse C_1^i et d'ordonné C_2^i [3].



II-3-1-8- Représentation des variables :

Les proximités entre les composantes principales et les variables initiaux sont mesurés par les covariances et surtout les corrélations [3].



Remarque :

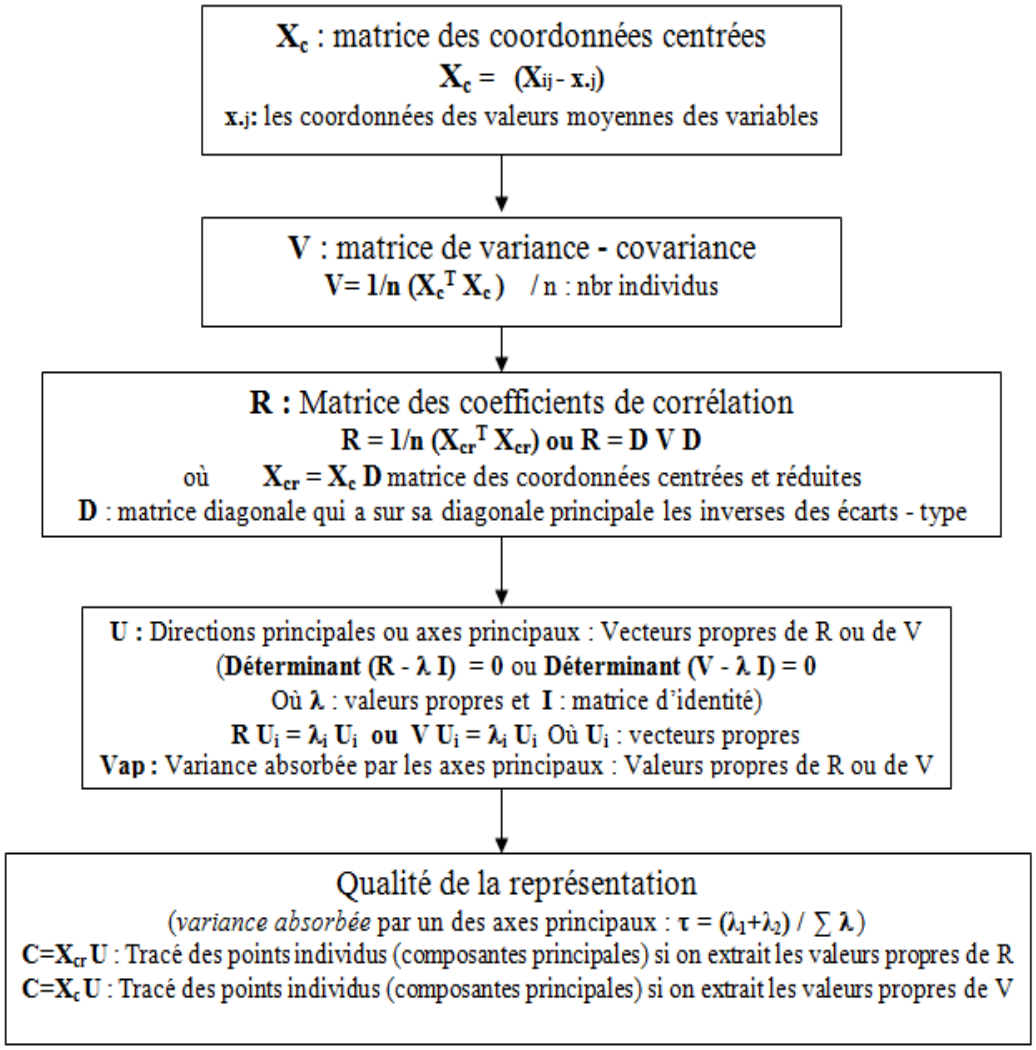
Pour que deux variables soient effectivement corrélées il faut d'une part qu'ils soient proches du cercle de corrélation et d'autre part proche entre eux.

Si elles sont proches du cercle, elles sont bien représentées. Sinon, elles sont mal représentées.

Et aussi, le cosinus de l'ongle compris entre les deux variables représente la corrélation entre eux. Donc si les variables sont non corrélées, elles sont perpendiculaires. Sinon, sur un axe (ou presque). Pour la corrélation négative, chacune d'un coté, et pour la corrélation positive, elles forment un seul point (ou proche) [3].

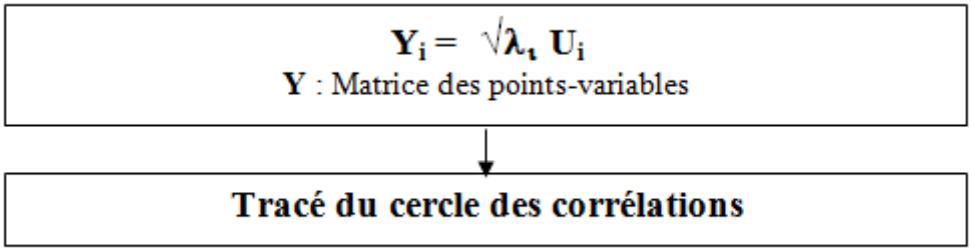
II-4- Organigramme de l'ACP :

Analyse dans R^P



Org I.1. Organigramme de l'ACP dans R^P [3]

Analyse dans R^N



Org I.2. Organigramme de l'ACP dans R^N [3]

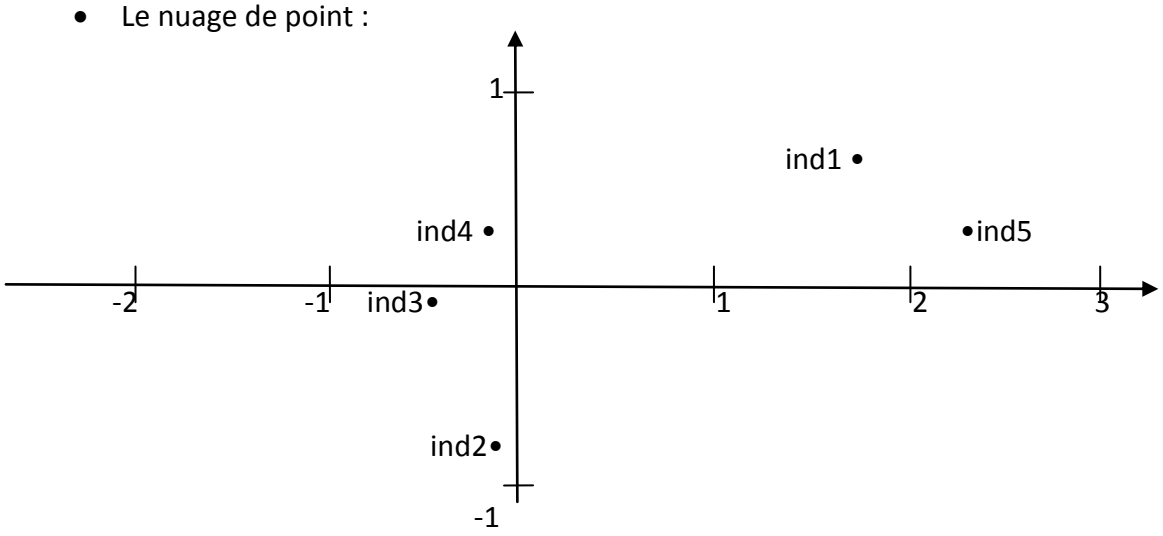
Exemple numérique :

Considérons deux variables x_1 et x_2 mesurés par cinq individus :

Var / Ind	X_1	X_2
1	1.00	5.00
2	2.00	10.00
3	3.00	8.00
4	4.00	8.00
5	9.00	12.00

- Les détails des calculs sont présentés en annexe.

- Le nuage de point :



- Contribution à la variance

	Axe 1	***Axe 2***
Ind 1	0.363	0.136
Ind 2	0.0001	0.732
Ind 3	0.016	0.0004
Ind 4	0.002	0.051
Ind 5	0.619	0.079

- Qualité de la représentation :

Axe 1	*Axe 2*
0.96	0.04
0.01	0.99
0.99	0.003
0.24	0.76
0.98	0.02

- Interprétation:

Les individus 1 et 5 sont ceux qui contribuent le plus fortement à la variance sur le premier axe.

L'interprétation est simple: le premier axe est lié fortement aux variables x_1 et x_2(voir le cercle de corrélation) et représente « la taille » des individus ; les individus 1 et 5 sont ceux qui connaissent des valeurs extrêmes à la fois.

Pour x_1 et x_2 petites pour l'individu 1, grandes pour l'individu 5.

Notons que l'individu 3 est presque parfaitement représenté sur l'axe 1, sa position correspond aux valeurs qu'il prend pour x_1 et x_2 ; c à d légèrement au dessous de la moyenne pour chacune des 2 variables.

C'est surtout l'individu 2 qui contribue à la variance de l'axe2, en fait cet axe et lié positivement à x_1 et négativement à x_2 (voir cercle des corrélations) et la position de l'individu 2 est due à la faible valeur qu'il prend pour x_1 par rapport à la forte valeur prise par x_2 .

A l'opposé, l'individu 4 bien représenté sur le second axe, doit sa position à une valeur de x_2 relativement forte par rapport à la valeur prise par x_1 .

II-5 – <u>Objectifs de l'ACP :</u>

Puisqu'il s'agit d'une méthode d'analyse de données multifactorielle, son but est de résumer cet ensemble de données. Ceci se fait selon les modalités suivantes :

- fournir des outils simples et lisibles de représentation des informations traitées, permettant de faire ressortir des données brutes les éventuels liens existant entre les variables (en terme de corrélation),

- donner des indications sur la nature, la force et la pertinence de ces liens, afin de faciliter leur interprétation et découvrir quelles sont les tendances dominantes de l'ensemble de données,
- réduire efficacement le nombre de dimensions étudiées (et ainsi simplifier l'analyse), en cherchant à exprimer le plus fidèlement possible l'ensemble original de données grâce aux relations détectées entre les variables.
- Traiter et manipuler des images en tirant les informations utiles de celle-ci sous forme des matrices [8]**.**

II-6-Avantages et inconvénients de l'ACP :

II-6-1- Avantages :

- **Simplicité mathématique**: L'ACP est une méthode factorielle car la réduction du nombre des caractères ne se fait pas par une simple sélection de certains d'entre eux, mais par la construction de nouveaux caractères synthétiques obtenus en combinant les caractères initiaux au moyen des "facteurs". Cependant, il s'agit seulement de combinaisons linéaires. Les seuls véritables outils mathématiques utilisés dans l'ACP sont le calcul des valeurs/vecteurs propres d'une matrice, et les changements de base.Sur le plan mathématique, l'ACP est donc une méthode simple à mettre en œuvre [8].
- **Simplicité des résultats** : Grâce aux graphiques qu'elle fournit, l'Analyse en Composantes Principales permet d'appréhender une grande partie de ses résultats d'un simple coup d'œil [8].
- **Puissance** : L'ACP a beau être simple, elle n'en est pas moins puissante. Elle offre, en quelques opérations seulement, un résumé et une vue complète des relations existant entre les variables quantitatives d'une population d'étude, résultats qui n'auraient pas pu être obtenus autrement, ou bien uniquement au prix de manipulations fastidieuses [8].
- **Flexibilité** : L'ACP est une méthode très souple, puisqu'elle s'applique sur un ensemble de données de contenu et de taille quelconques, pour peu qu'il s'agisse de données quantitatives organisées sous forme individus/variables. Cette souplesse d'utilisation se traduit surtout par la diversité des applications

de l'ACP, qui touche tous les domaines, comme exposé dans la partie précédente [8].

II-6-2-Inconvénients :

En tant que méthode d'analyse de données, l'ACP n'a pas réellement d'inconvénients en soi. Elle s'applique simplement sur des cas précis et pour générer un type de résultat particulier. Ca n'aurait donc aucun sens de dire que c'est un inconvénient de l'ACP qu'elle ne s'applique pas en dehors de ce contexte. De même, étant donné qu'il s'agit avant tout d'une technique de résumé de données, la perte d'information forcément engendrée n'est pas un inconvénient, mais plutôt une condition d'obtention du résultat, même si elle occulte parfois des caractéristiques pourtant représentatives dans certains cas particuliers [8].

II-7 – Conclusion :

L'ACP est une méthode puissante pour synthétiser et résumer de vastes populations décrites par plusieurs variables quantitatives. Elle permet entre autre de dégager de grandes catégories d'individus et de réaliser un bilan des liaisons entre les variables. Par cette analyse nous pouvons mettre en évidence de grandes tendances dans les données telles que des regroupements d'individus ou des oppositions entre individus (ce qui traduit un comportement radicalement différent de ces individus) ou entre variables (ce qui traduit le fait que les variables sont inversement corrélées). Les représentations graphiques fournies par l'ACP sont simples et riches d'informations. L'ACP peut être une première analyse pour l'étude d'une population dont les résultats seront enrichis par une autre analyse factorielle ou encore une classification automatique des données [5].

CHAPITRE III

Application

III-1- Introduction :

Dans ce chapitre nous présentons notre application et les résultats obtenus.

III-2- Langage de programmation :

Notre application est implémentée sous l'environnement de programmation MATLAB version 7.6.0 (R2008a).

Ce dernier est l'abréviation de Matrix LABoratory. Écrit à l'origine, en Fortran, par Cleve. Moler.

MATLAB est un environnement complet, ouvert et extensible pour le calcul et la visualisation. Il dispose de plusieurs centaines de fonctions mathématiques, scientifiques et techniques. L'approche matricielle de MATLAB permet de traiter les données sans aucune limitation de taille et de réaliser des calculs numériques et symboliques de façon fiable et rapide. Grâce aux fonctions graphiques de MATLAB, il devient très facile de modifier interactivement les différents paramètres des graphiques pour les adapter selon nos souhaits [9].

III-3- *Les fenêtres usuelles de MATLAB :*

Après le démarrage du logiciel, quelques fenêtres apparaissent, ou peuvent être ouvertes à partir du menu déroulant (view). Nous introduisons les fenêtres
"Répertoire courant" (current directory), "Commandes" (command) et la fenêtre "Historique" (command history) [10].

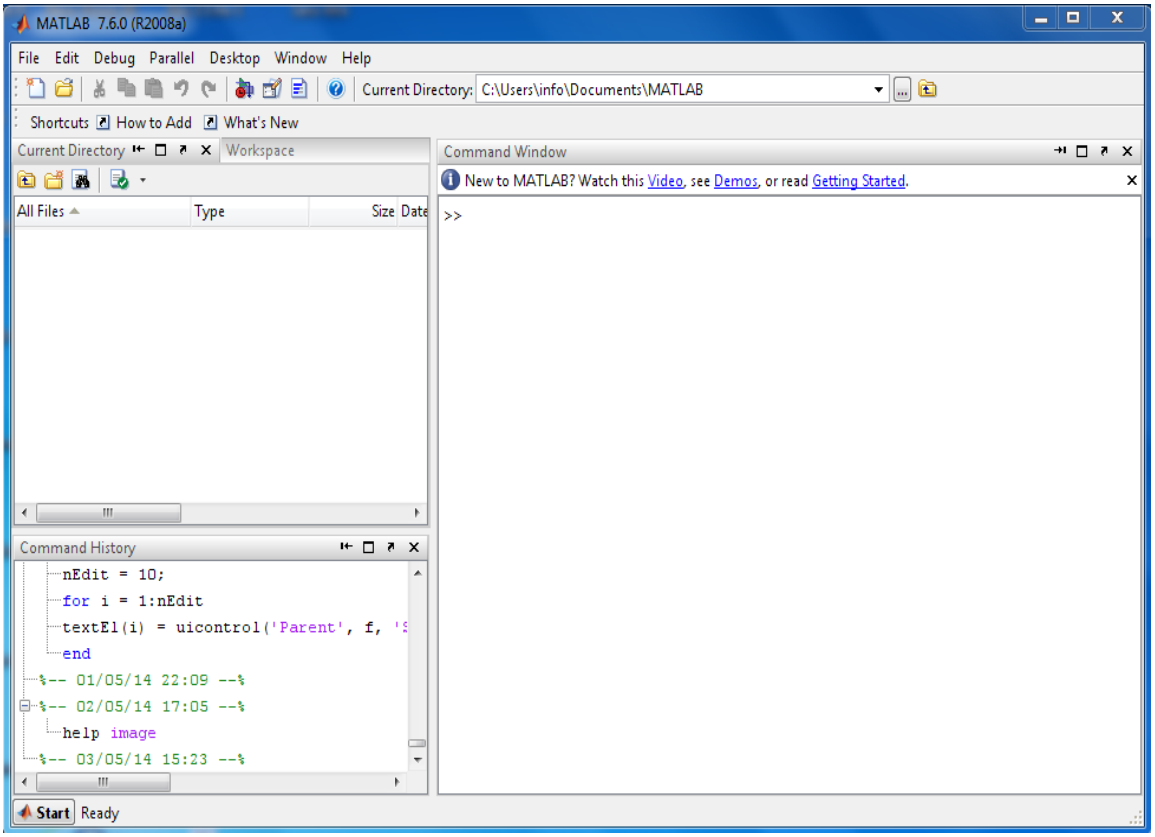


Figure III.1. Fenêtre de MATLAB

III-3-1 Fenêtre "Répertoire courant" :

Elle permet de sélectionner le répertoire de travail et de visualiser ce qu'il contient. Avant toute utilisation, le répertoire dans lequel on souhaite travailler doit être sélectionné. Seuls les programmes listés dans ce répertoire seront reconnus [10].

III-3-2 "Fenêtre de commandes" :

Elle permet de saisir des valeurs, de taper des instructions ligne par ligne et d'exécuter des programmes. L'exécution d'un programme se fait directement dans cette fenêtre en tapant le nom du programme puis ENTREE [10].

III-3-3 Fenêtre "Historique" :

Toutes les instructions saisies dans la fenêtre de commandes sont stockées dans une mémoire constituant ce qu'on appelle l'historique. C'est une liste des commandes déjà tapées, y compris lors des utilisations antérieures de Matlab.

Il est possible de voir cette liste en ouvrant la fenêtre command history.

Il est également possible, à partir de la fenêtre de commande, d'accéder à une commande déjà tapées et donc stockées dans cette liste d'historiques, il est possible

d'effacer le contenu de cette fenêtre en sélectionnant l'option "clear command history" dans le menu déroulant "Edit" [10].

III-4- Description de l'application :

L'exécution de notre application se présente comme suit :

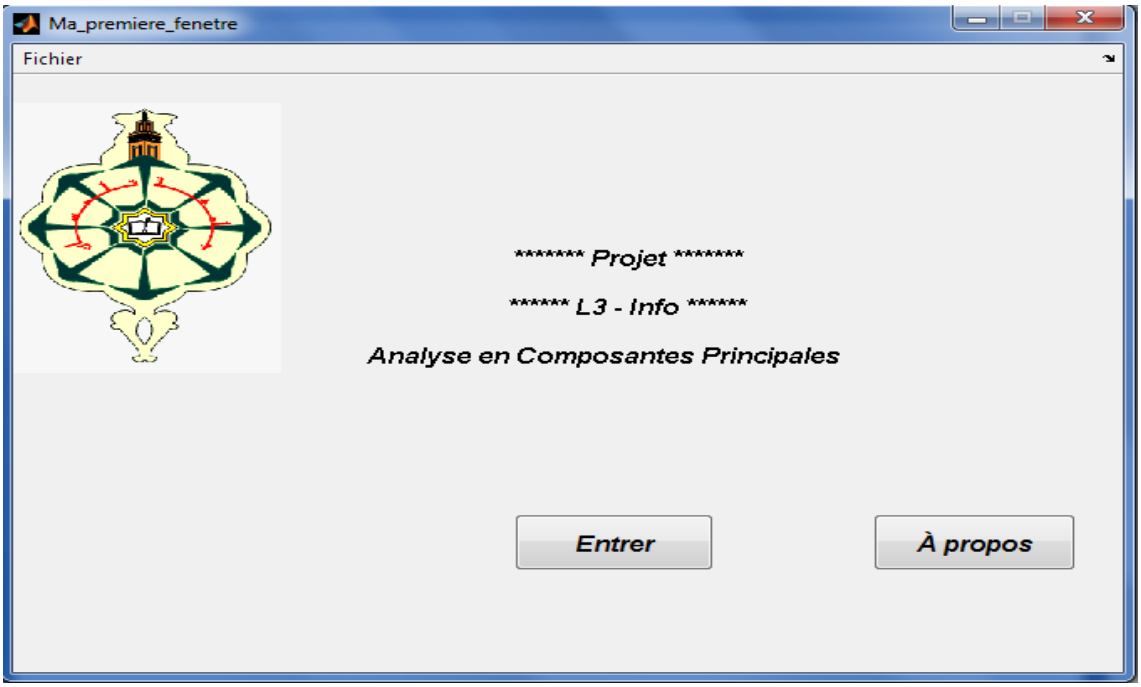


Figure III.2. La première fenêtre.

- « À propos » : bouton qui affiche les informations concernant l'application.

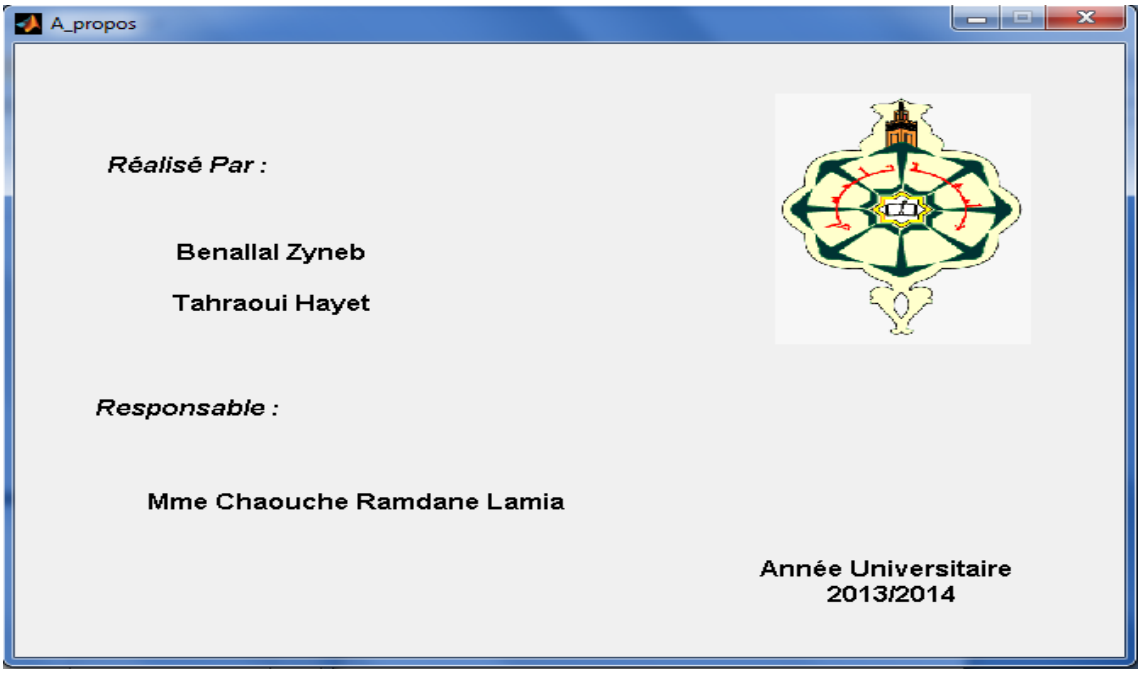


Figure III.3. À propos.

- Pour que l'utilisateur soit plus alaise à continuer l'exécution il a deux façons :
  - Accès direct par le bouton « Entrer ». ou par le menu « Fichier » :

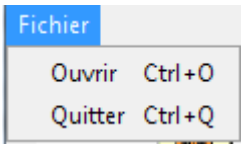


Figure III.4. Menu fichier.

« Entrer » et « Ouvrir » affichent :

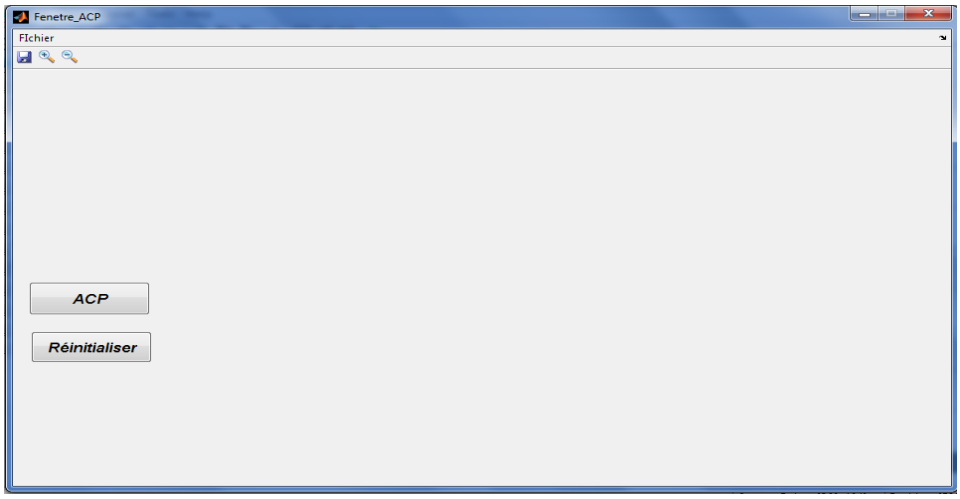


Figure III.5.Fenêtre_ACP.

En appuyant sur le bouton « ACP », l'utilisateur sélectionne un fichier sous l'extension « .txt ».

Exemple 1 : c'est l'exemple cité en annexe.

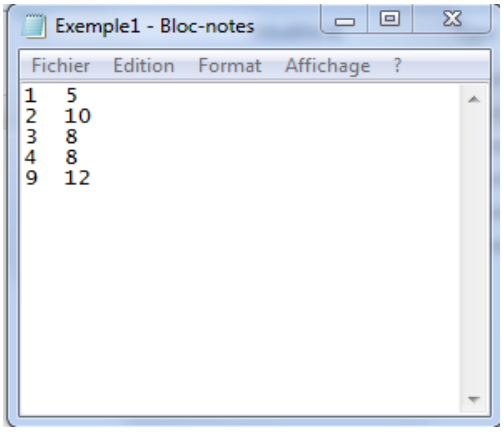


Figure III.6. Exemple 1(5 ind , 2 var).

Son ACP:

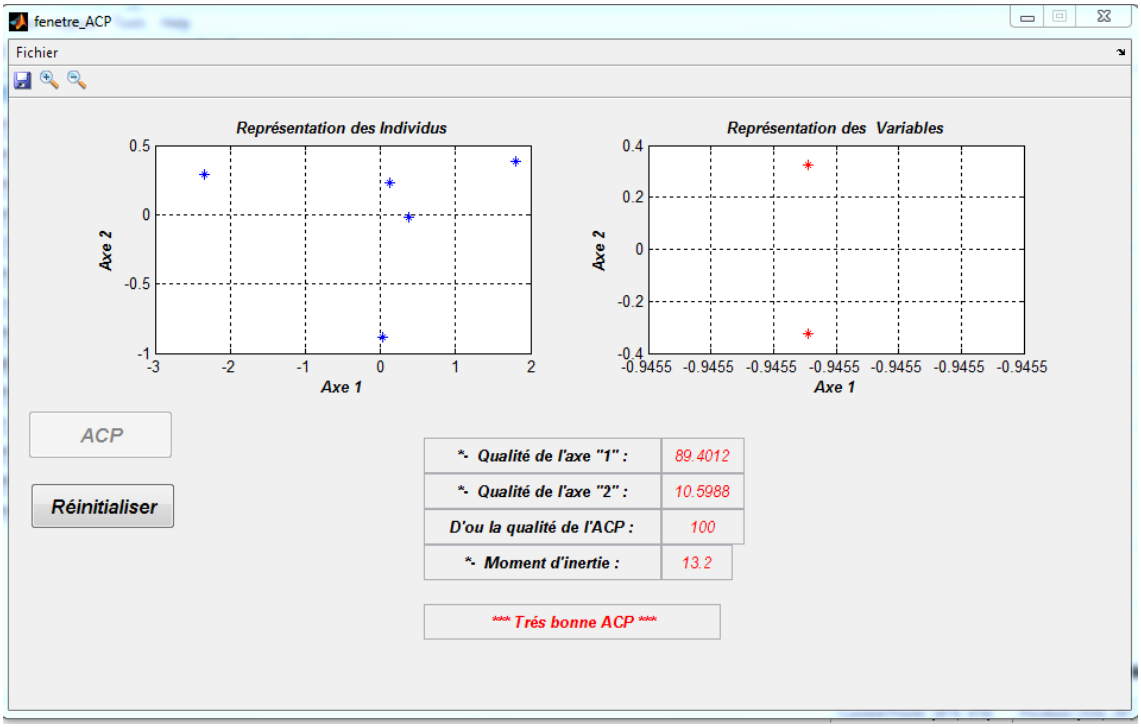


Figure III.7.Nuage de point exemple1.

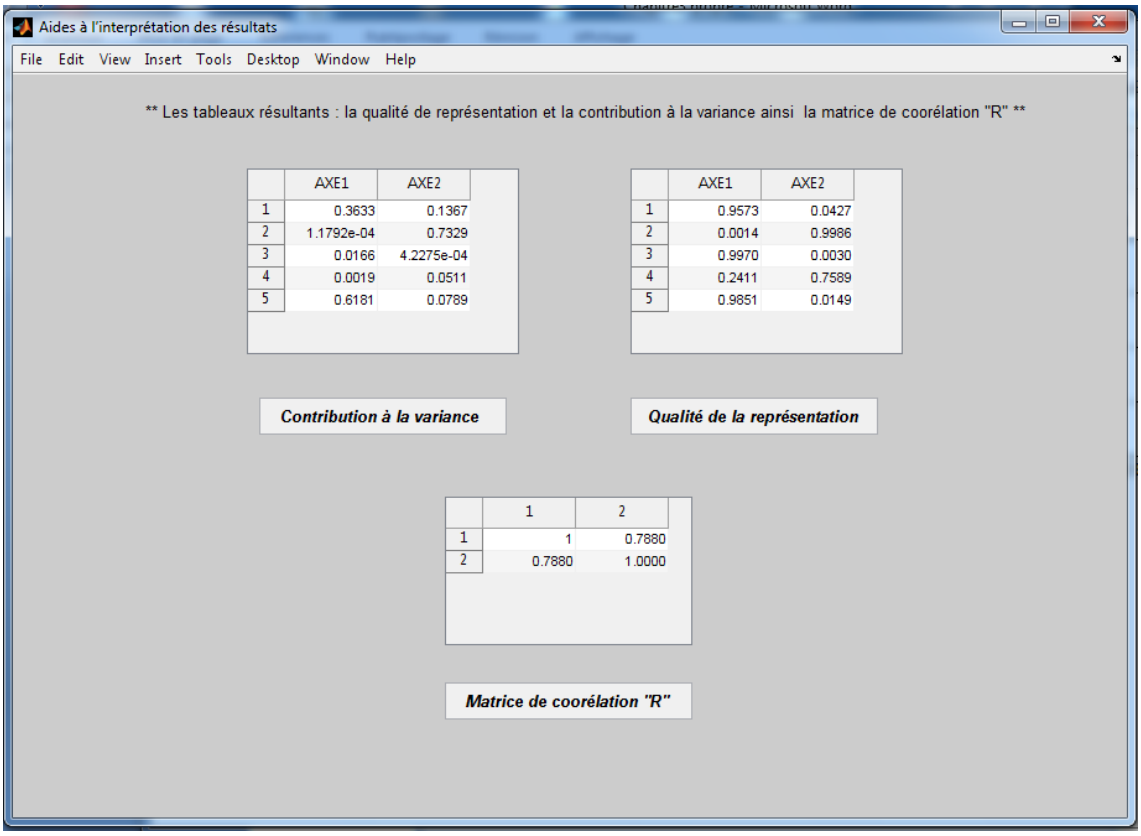


Figure III.8.Interprétation exemple1.

Pour faire une autre ACP sur un autre exemple, il suffit de cliquer sur le bouton « Réinitialiser »

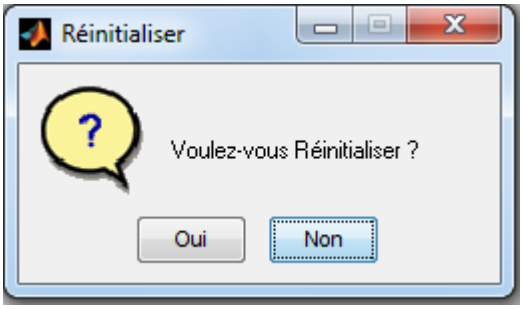


Figure III.9.Réinitialiser.

L'utilisateur clique sur 'Oui', il obtient une autre fois la 'Fenêtre_ACP' d'exécution.

Exemple 2 : Considérons les notes d'épreuves finales (de 0 à 20) obtenues par 102 étudiants dans 7 modules différents.

Nous souhaitons présenter les données dans un espace de dimension réduite (par exemple 3) avec une bonne qualité en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales.

Num Etudiant	Système d'exploitation	Programmation Logique	Théorie des Graphe	Réseaux	Compilation	Infographie	Probabilité
1	6	6	5	5.5	12	7	8
2	8	8	7	9	10	12	9.5
3	6	7	11	9.5	8	12	7.5
4	14.5	14	12	12.5	11	10.5	15
5	14	14	12	12.5	13	11	10
6	5.5	7	14	11.5	8	13	10
7	13	12.5	8.5	9.5	11	9	14
8	9	9.5	12.5	12	10	8	7
9	10	13	9	8	12	14	4.5
10	12	5.5	8	9	13	10	13
11	8	7	6	15	9.5	8	11
12	14.5	14	12	12.5	11	10 .5	15
13	7	5.5	16	11	10	8	9
14	8	8	7	9	10	8	9
15	10	13	14	9	12	9	11
16	8	9	13	10	12	6.5	11
17	10	10	11	8	14	13	9.5
18	4.5	6	10	9	8	13	7
19	13	12.5	8.5	9.5	11	9	14
20	12	5.5	8	9	13	10	13

Figure III.10. Exemple 2(extrait de 102 ind , 7 var).

Après le choix de dimension 3 :

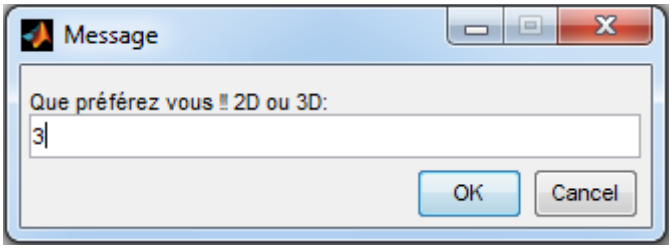


Figure III.11. Choix de dimension.

Son ACP :

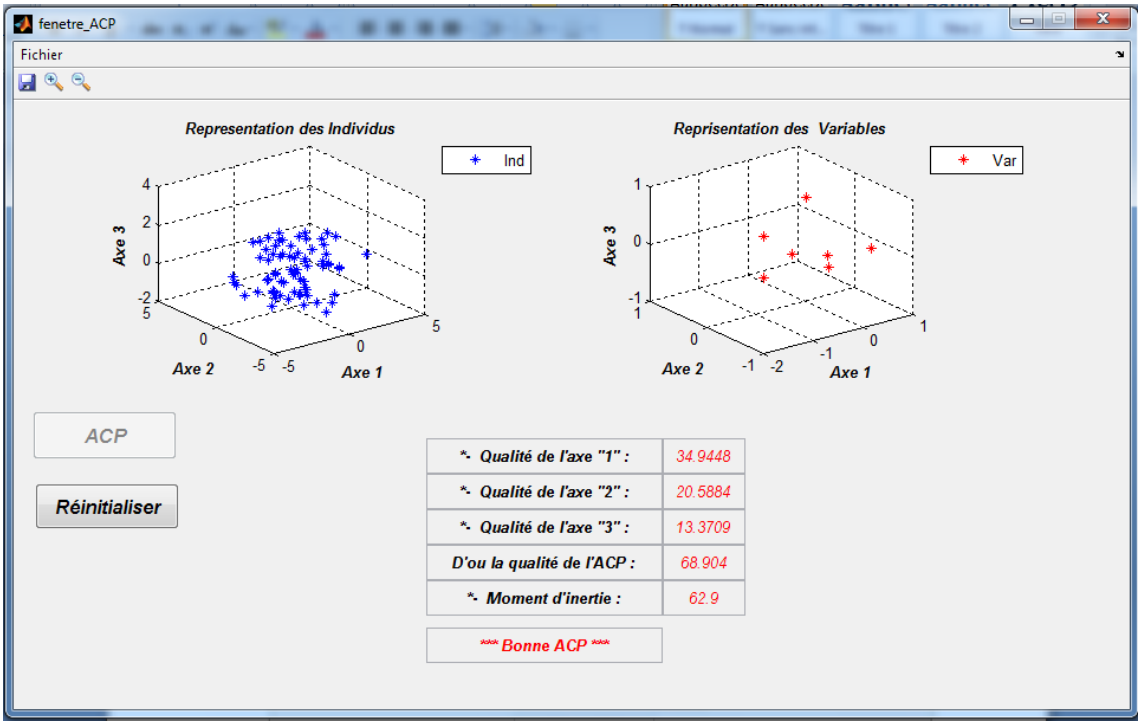


Figure III.12.Nuage de point exemple2.

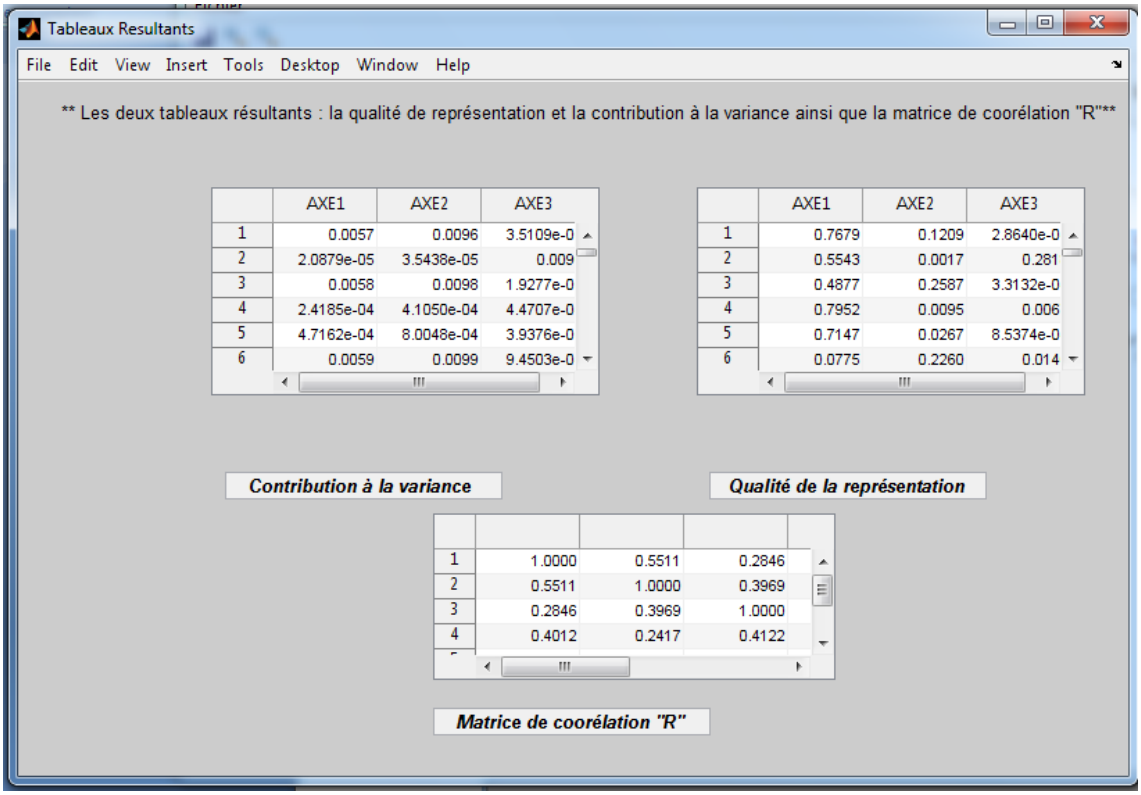


Figure III.13.Interprétation exemple2.

III-5- Notre algorithme d'ACP :

```
% A est une matrice a saisir avec plus de 2 variable
function [f,h,cp]= acp(A)
  disp('    ********calculons l''acp avec la matrice R******** ');
  B=A';              % transposé de la matrice A
  [n,m]= size(B);    % dimension de la matrice
  Xc=zeros(n,m);     %matrice vide
  for   i=1:n;
        l= B(i,:);
        a=(1/m)*sum(l);
        z=l-a;
        Xc(i,:)=z';
  end
  disp('*** la matrice centrée Xc *** ');
  Xc=Xc'              % matrice centré
  disp('*** la matrice de variance-convariance V *** ');
  V=(1/m)*(Xc'*Xc)     %matrice de variance-convariance
  K=diag(V);           % diagonal de V
  D=1 ./ sqrt(K);
  z=eye(m);            % matrice identique
  z=diag(D);
  disp('*** la matrice D *** ');
  D=z                  % matrice D
  Xcr=Xc*D;            % matrice Xcr
  disp('*** la matrice R *** ');
  R=(1/m)*(Xcr'*Xcr)      %matrice de corrélation
  disp('*** les vecteurs propre ''f'' et valeurs propres ''h'' *** ');
  [f,h]=eig(R)         %f:vecteurs propre-h:valeur propres
  a= diag(h);          % pour extraire juste les valeurs propres
  [h1,Indices]=sort(a) %ordonne les valeurs propres
  [s,Indices]=min(h1)
  s=0 ;
  qualite=[0 0];
  for i=1:n
      s=s+h1(i);
  end
  for i=1:n
      qualite(i)=(h1(i)/s)*100;
  end
  disp('*** la qualité  *** ');
  qualite              % qualité de chaque valeurs propres
  disp('*** les composantes principales *** ');
  cp=Xcr*f             %composantes principales
 choix = input(' Entrer votre choix : ');
 if (choix ==2)       %représentation sur 2D
 switch(Indices)
       case Indices==1
            k=cp(:,n-1:n)
            j=h(:,n-1:n)
       case Indices==2
            k=cp(:,[n-2,n]);
            j=h(:,[n-2,n]);
       case Indices==3;
            k=cp(:,n-2,n-1);
            j=h(:,n-2:n-1);
 end
 Xcr2=Xcr.^2; % la matrice Xcr au carré
```

```
cp2=cp.^2;   % la matrice cp au carré
a=zeros(0);
e=zeros(0);
for   i=1:m;
      q=Xcr2(i,:); % selectionner les ligne de la matrice Xcr2
      e(i)=sum(q); % faire la somme des ligne
      a(i)=e(i);  % mettre les valeurs de e(i) dans un vecteur colonn
end
a1=a';
col1=cp2(:,n);    % éxtraire la n éme colonne de la matrice cp
Q1=col1./a1 ;
col2=cp2(:,n-1);  % éxtraire la (n-1) éme colonne de la matrice cp
Q2=col2./a1;
disp('*** qualité de représentation *** ');
Q=[Q1 Q2]         % la matrice de la qualité de représentation
h2=m*h1;
co1= Xcr2(:,n);   % éxtraire la n éme colonne de la matrice Xcr2
C1=co1./h2(1);
co2= Xcr2(:,n-1); % éxtraire la (n-1) éme colonne de la matrice Xcr2
C2=co2./h2(2);
disp('*** Contribution de la variance *** ');
C=[C1 C2]         % la matrice de la contribution de la variance
       t=zeros(n,m);
      [r,indices]=max(j);
      [l,indices]=max(r);
      t(:,1)=sqrt(l)*f(:,3);
      t(:,1);
      [e,indices]=min(r);
      t(:,2)=sqrt(e)*f(:,2);
      t
elseif(choix==3)
      switch(Indices)
              case Indices==1
                   k=cp(:,n-2:n)
                   j=h(:,n-2:n)
             case Indices==2
                   k1=cp(:,n-3);
                   k2=cp(:,[n-1,n]);
                   k=[k1 k2];
                   j1=h(:,n-3);
                   j2=h(:,[n-1,n]);
                   j=[j1 j2];
             case Indices==3;
                   k1=cp(:,n-3:n-2);
                   k2=cp(:,n);
                   k=[k1 k2];
                   j1=h(:,n-3:n-2);
                   j2=h(:,n);
                   j=[j1 j2];
             case Indices==4;
                   k=cp(:,n-3,n-1);
                   j=h(:,n-3:n-1);
          end
 Xcr2=Xcr.^2;         % la matrice Xcr au carré
 cp2=cp.^2;           % la matrice cp au carré
 a=zeros(0);
 e=zeros(0);
 for   i=1:m;
       q=Xcr2(i,:);   % selectionner les ligne de la matrice Xcr2
```

```
      e(i)=sum(q);   % faire la somme des ligne
      a(i)=e(i);% mettre les valeurs de e(i) dans un vecteur colonne
end
a1=a';
col1=cp2(:,n);     % éxtraire la n éme colonne de la matrice cp
Q1=col1./a1 ;
col2=cp2(:,n-1);   % éxtraire la (n-1) éme colonne de la matrice cp
Q2=col2./a1;
col3=cp2(:,n-2);   % éxtraire la (n-2) éme colonne de la matrice cp
Q3=col3./a1;
disp('*** qualité de représentation *** ');
Q=[Q1 Q2 Q3]       % la matrice de la qualité de représentation
h2=m*h1;
co1= Xcr2(:,n);    % éxtraire la n éme colonne de la matrice Xcr2
C1=co1./h2(1);
co2= Xcr2(:,n-1); % éxtraire la (n-1) éme colonne de la matrice Xcr2
C2=co2./h2(2);
co3= Xcr2(:,n-2); % éxtraire la (n-2) éme colonne de la matrice Xcr2
C3=co3./h2(3);
disp('*** Contribution de la variance *** ');
C=[C1 C2 C3]       % la matrice de la contribution de la variance
            t=zeros(n,m);
            [r,indices]=max(j);
            [l,indices]=max(r);
            t(:,1)=sqrt(l)*f(:,3);
            t(:,1);
            [e,indices]=min(r);
            t(:,2)=sqrt(e)*f(:,2);
            [b,indices]=min(r);
            t(:,3)=sqrt(e)*f(:,1);
            t

      end
end
```

*VI- **Conclusion:***

Ce chapitre a été consacré à la modélisation et réalisation de l'ACP, les différentes étapes de réalisation ont été détaillées en appliquant quelques exemples.

De ce fait l'ACP a été programmé de façon a rendre visible les différentes composantes principales sous forme de nuage de point avec des aides à l'interprétation comprenant des tableaux correspondant à la qualité de représentation pour chaque individus pour les axes et un autre tableau pour la contribution à la variance.

Conclusion générale

Notre mémoire avait pour but d'appliquer l'une des méthodes multidimensionnelles de l'analyse des données à savoir l'Analyse en Composantes Principales.

L'objectif de cette analyse était de traiter un nombre très important de données afin de visualiser ces derniers dans le meilleur espace réduit. Cette analyse est réalisable lorsqu'il est possible de réduire l'espace multidimensionnel (où l'information n'est pas lisible) en un espace à deux ou trois dimensions (où l'information est lisible), de telle sorte que cet espace réduit conserve une part importante de l'information qui était contenue dans l'espace multidimensionnel d'origine.

Nous avons présenté dans ce mémoire le principe de l'ACP qui est une technique d'analyse statistique, principalement descriptive, qui consiste à représenter sous forme graphique le plus d'informations possibles contenues dans un tableau. Elle permet ainsi de visualiser un espace à p dimensions à l'aide d'espaces de dimensions plus petites. Cette approche facilite l'analyse en regroupant les données en des ensembles plus petits et en permettant d'éliminer les problèmes de multi colinéarité entre les variables.

Et pour finir terminé par présenté notre application sur différents exemples avec leur résultats et aide à leur interprétation.

Comme perspectives on peut s'intéressé à d'autres méthodes multidimensionnelles tel que l'analyse factorielle en correspondance AFC, analyse des correspondance multiples et d'autres.

Annexe : Détail de calcul

Considérons deux variables x_1 et x_2 mesurés par cinq individus [12]:

Var / Ind	X_1	X_2
1	1.00	5.00
2	2.00	10.00
3	3.00	8.00
4	4.00	8.00
5	9.00	12.00

Les variables x_1 et x_2 formant un plan dans l'espace des individus. L'ACP comporte deux étapes seulement :

On calculer les moyennes et les écarts types de chacune des deux variables :

Moyenne de x_1 et x_2 :

$\overline{X_1} = (1+2+3+4+9)/5$

$\overline{X_1}=3.8$

$\overline{X_2} = (5+10+8+8+12)/5$

$\overline{X_2}=8.6$

L'écart type de x_1 et x_2 :

$\sigma_{x1} = \sqrt{var(x_1)}$

$$Var(v^j) = 1/n \sum^{n}_{i=1} (x_{ij} - x_{.j})^2$$

$\rightarrow \quad \sigma_{x1} = 2.79$

$\sigma_{x2} = \sqrt{Var(x_2)}$

$\rightarrow \quad \sigma_{x2} = 2.33$

En suite :

Matrice centrée X_c :

Les coordonnées de cette matrice sont calculer de la manière suivant :

La valeur de la variable – la moyenne de cette variable on obtient la matrice :

$$X_c = \begin{pmatrix} -2.8 & -3.6 \\ -1.8 & 1.4 \\ -0.8 & -0.6 \\ 0.2 & -0.6 \\ 5.2 & 3.4 \end{pmatrix}$$

Matrice de variance / covariance « V » :

$$V = 1/n\ X_c^t\ X_c$$

$$\rightarrow \quad V = \begin{pmatrix} 7.76 & 5.12 \\ 5.12 & 5.44 \end{pmatrix}$$

Matrice centrée réduit X_{cr} :

$$X_{cr} = X_c\ D$$

Tel que $D = \begin{pmatrix} 1/\sigma_{x1} & 0 \\ 0 & 1/\sigma_{x2} \end{pmatrix}$

$$\rightarrow \quad D = \begin{pmatrix} 0.359 & 0 \\ 0 & 0.428 \end{pmatrix}$$

$$\rightarrow \quad X_{cr} = \begin{pmatrix} -1.005 & -1.543 \\ -0.646 & +0.600 \\ -0.287 & -0.257 \\ 0.072 & -0.257 \\ 1.867 & 1.458 \end{pmatrix}$$

Matrice de corrélation R :

$R = 1/n\ X_{cr}^{t}\ X_{cr}$

$$\rightarrow R = \begin{bmatrix} 1 & 0.788 \\ 0.788 & 1 \end{bmatrix}$$

Les facteurs sont les vecteurs propres normés de cette matrice de corrélation

Valeur propre λ :

$Det\ (R - \lambda I) = 0$

Nombre de λ = Nombre de variable

$$Det \begin{vmatrix} 1-\lambda & 0.788 \\ 0.788 & 1-\lambda \end{vmatrix} = 0$$

$$\rightarrow \quad (1-\lambda)^2 - (0.788)^2 = 0$$

$$\rightarrow \quad \lambda^2 - 2\lambda + 0.38 = 0$$

$\Delta = 2.48$

$\lambda_1 = 0.212 \qquad \lambda_2 = 1.788$

On les ordonne par ordre décroissant, on obtient :

$\lambda_1 = 1.788$ et $\lambda_2 = 0.212$

Vecteur propre U :

$R*U = \lambda*U$

$\lambda_1 = 1.788 \rightarrow R*U_1 = \lambda_1 * U_1$

$$\begin{bmatrix} 1 & 0.788 \\ 0.788 & 1 \end{bmatrix} * \begin{bmatrix} a \\ b \end{bmatrix} = \lambda_1 * \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\rightarrow \begin{cases} a + 0.788*b = \lambda_1 * a \\ 0.788*a + b = \lambda_1 * b \end{cases} \rightarrow \begin{cases} (1 - \lambda_1)*a + 0.788*b = 0 \\ (1 - \lambda_1)*b + 0.788*a = 0 \end{cases}$$

$$U_1 = \begin{bmatrix} a \\ a \end{bmatrix}$$

$\lambda_2 = 0.212 \rightarrow R*U_2 = \lambda_2 * U_2$

$$\begin{bmatrix} 1 & 0.788 \\ 0.788 & 1 \end{bmatrix} * \begin{bmatrix} a \\ b \end{bmatrix} = \lambda_2 * \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\rightarrow \begin{cases} a + 0.788*b = 0.212*a \\ 0.788*a + b = 0.212*b \end{cases} \rightarrow \begin{cases} 0.788*a + 0.788*b = 0 \\ 0.788*b + 0.788*a = 0 \end{cases}$$

$U_2 = \begin{pmatrix} a \\ -a \end{pmatrix}$

$U_1 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$ $U_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$

Par conséquent les composantes principales s'écrivent :

$C^1 = 0.707\ X_1 + 0.707 X_2$

$C^2 = 0.707 X_1 + 0.707 X_2$

- Et les pourcentages de variance expliqué sont 1.788 divisé par 2, soit 89.4% pour le premier axe. Et 0.212 divisé par 2 , soit 10.6% pour le second axe.

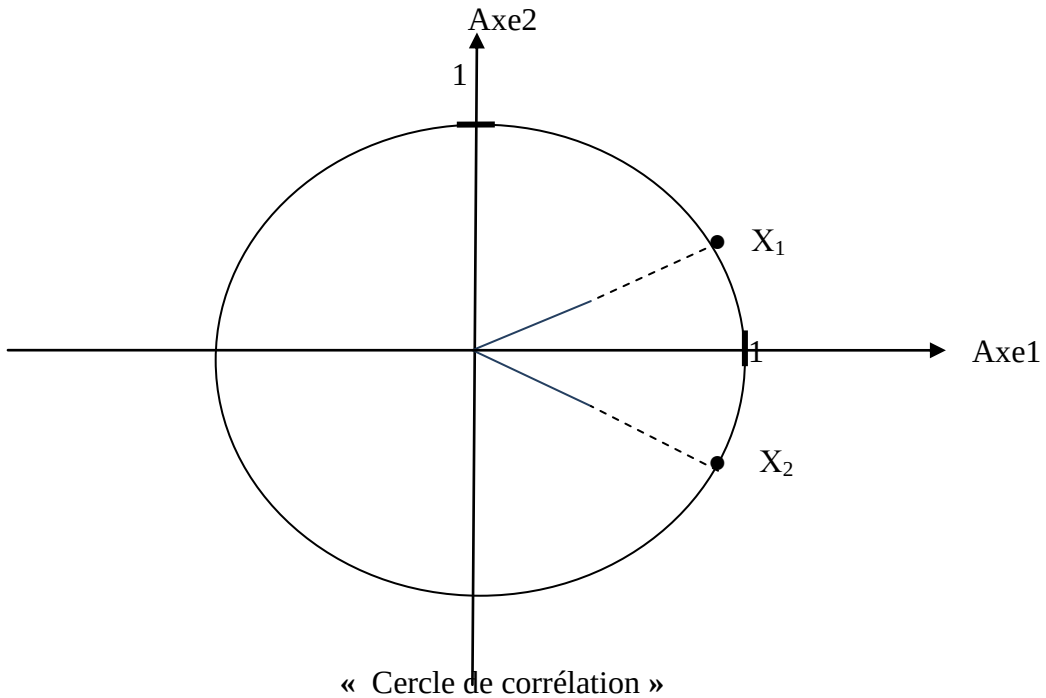
Cercle de corrélation :

$$\begin{pmatrix} R(C^1,X_1) \\ R(C^1,X_2) \end{pmatrix} = \sqrt{\lambda_1} * U_1 = \begin{pmatrix} 0.906 \\ 0.946 \end{pmatrix} \begin{matrix} X_1 \\ X_2 \end{matrix}$$

Et

$$\begin{pmatrix} R(C^2,X_1) \\ R(C^2,X_2) \end{pmatrix} = \sqrt{\lambda_2} * U_2 = \begin{pmatrix} 0.324 \\ -0.324 \end{pmatrix} \begin{matrix} X_1 \\ X_2 \end{matrix}$$

Les coordonnées de X_1 et X2 sont donc : X_1 (0.946 , 0.324) et X_2 (0.946 , -0.324)



« Cercle de corrélation »

Dans cet exemple x_1 et x_2 sont parfaitement représentées sur le cercle des corrélations, car ce cercle est situé dans le plan des deux variables x_1 et x_2 ce qui est algébriquement se traduit par les fait que $(0.964)^2 + (0.324)^2 = 1$.

L'angle au centre entre x_1 et x_2 est égal à 38 degré,et on retrouve la valeur du coefficient de corrélation entre x_1 et x_2 à partir du graphique

$R(x_1 , x_2) = \cos(38^o) = 0,79$

x_1 et x_2 sont positivement et fortement corrélés avec la première composante principale ; au contraire ces deux variables sont assez faiblement corrélés avec la seconde composante principale et s'apposent sur ce seconde axe.

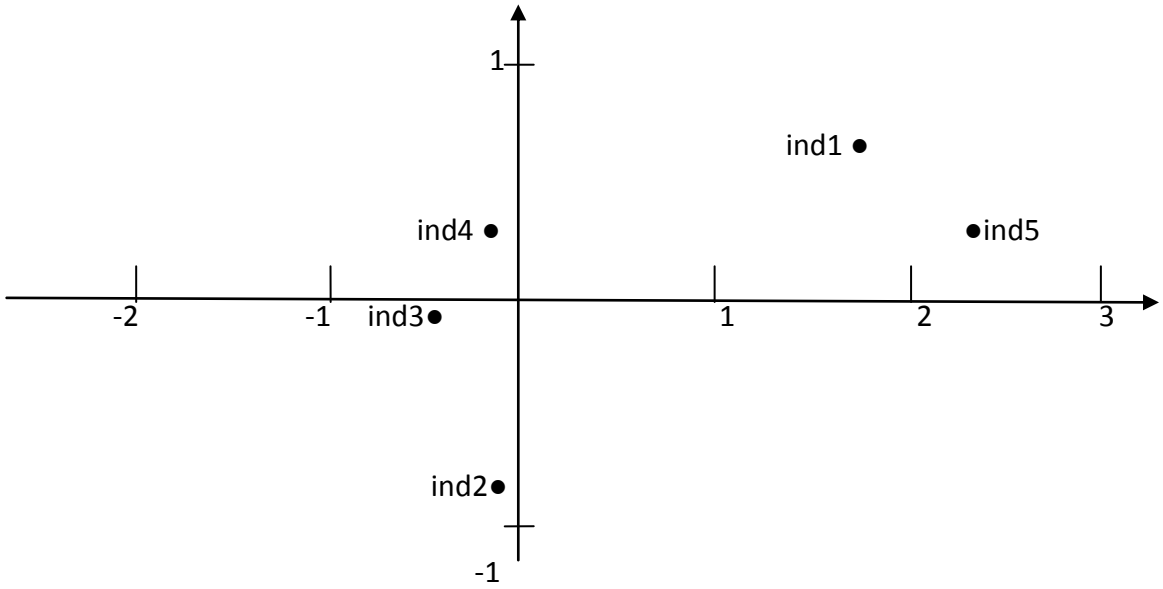
On obtient les coordonnés des individus :

$$X_{cR} \begin{pmatrix} -1{,}005 & -1{,}543 \\ -0{,}646 & 0{,}6 \\ -0{,}287 & -0{,}257 \\ 0{,}072 & -0{,}257 \\ 1{,}867 & 1{,}458 \end{pmatrix} * \begin{pmatrix} 0{,}707 & 0{,}707 \\ 0{,}707 & -0{,}707 \end{pmatrix} = \begin{matrix} U_1 & U_2 \end{matrix}$$

$$\rightarrow C= \begin{pmatrix} 1{,}802 & 0{,}381 \\ -0{,}032 & -0{,}881 \\ -0{,}385 & -0{,}021 \\ -0{,}131 & 0{,}233 \\ 2{,}351 & 0{,}289 \end{pmatrix} \begin{matrix} \text{ind 1} \\ \text{ind 2} \\ \text{ind 3} \\ \text{ind 4} \\ \text{ind 5} \end{matrix}$$

Coordonnée d'individus

Ce qui permet de tracer la représentation des individus :



Pour faciliter l'interprétation des résultats, calculons maintenant la qualité de représentation de chaque individus pour chacun des deux axes a partir des coordonnés de l'individu sur ces 2axes ; pour l'individu 4 par exemple, la qualité de représentation est :

Axe 1 : $\dfrac{(C_i)^2}{(X_{cR}(i))^2+(X_{cR}(j))^2} = \dfrac{(-0.131)^2}{(0.072)^2+(-0.257)^2} = 0.24$

Axe 2 : $\dfrac{(C_j)^2}{(X_{cR}(i))^2+(X_{cR}(j))^2} = \dfrac{(0,233)^2}{(0,072)^2+(-0,257)^2} = 0,76$

Pour ce même individu 4, la contribution à la variance est :

Axe 1 : $\dfrac{(C_i)^2}{\text{Nbr d'individu}*\lambda_1} = \dfrac{(-0.131)^2}{5*(1.788)} = 0.002$

λ_1

Axe 2 : $\dfrac{(C_j)^2}{\text{Nbr d'individu}*\lambda_2} = \dfrac{(0.233)^2}{5*(0.212)} = 0.051$

λ_2

Liste des figures :

Organigrammes :

Bibliographie :

[1] Samuel Ambapour « Introduction à l'analyse des données » document de travail pour le bureau d'application des méthodes statistiques et informatiques, 2003

[2] Fenelon, J.-P., Qu'est-ce que l'analyse des données ? : exposé accessible aux non-mathématiciens, Paris, Lefonen, 1981, 311 p.

[3] Jean-Paul Benzécri. « Histoire et Préhistoire de l'Analyse des données » Livre Partie 5 » 1977,p.9-40.

[5] Arnaud Martin. « Analyse des données » polycopié de cours ENSIETA – Réf : 1463 Septembre 2004.

[6] I. Jolliffe, Principal Component Analysis, 2nd edition éd., Springer-Verlag,

2002.

[7] F.-G. Carpentier. « Analyses multidimensionnelles et applications informatiques » Cours 2010-2011.

Web graphie :

[4] http://ebrunelle.ep.profweb.qc.ca/MQ/Chapitre2

[8] http://www.chambreuil.com/public/education/3.2/stat/projet/ACP

[9] http://www.iro.umontreal.ca/~mignotte/IFT2425/Matlab

[10]http://www.com.univ-mrs.fr/~poggiale/ens/M1/UE209/Intro_Matlab.pdf

« INITIATION A L'UTILISATION DU LOGICIEL MATLAB »

Résumé

Ce mémoire s'inscrit dans le domaine de l'Analyse des Données, il s'intéresse a l'application de l'une de ses méthodes qui est « l'Analyse en Composantes Principales ».

Grâce à cette étude nous avons atteint le but d'obtention d'une plus grande lisibilité des résultats de nos interprétations qui sont donnés pour mieux comprendre notre approche.

L'élaboration de ce travail a été implémentée sous l'environnement de programmation MATLAB version 7.6.0 (R2008a).

Mots clés : Analyse des Données, ACP, Variables quantitatives.

Abstract

This final year project is in the field of Data Analysis, he is interested in the application of one of its methods "Principal Component Analysis".

Through this study we reached the goal of obtaining greater clarity the results of our interpretations are given to better understand our approach.

The development of this work has been implemented in the programming environment MATLAB Version 7.6.0 (R2008a).

Keywords: Data Analysis, PCA, quantitative variables.

ملخص

هذه المذكرة تندرج في مجال تحليل البيانات، و تهتم بتطبيق واحدة من أساليبها ه ي " تحليل المكونات الرئيسية".

من خلال هذه الدراسة توصلنا إلى هدف الحصول على قدر أكبر من الوضوح للنتائج المعالجة مع تفسير أفضل لمجريات ه ذا الأسلوب.

وقد تم تطوير هذا العمل تحت برمجة "مطلب" إصدار 7.6.0 (R2008a)

الكلمات المفتاحية: تحليل البيانات, تحليل المكونات الرئيسية, المتغيرات الكمية