REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENET SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE SAAD DAHLAB DE BLIDA

FACULTE DES SCIENCES

DEPARTEMENT D'INFORMATIQUE

MEMOIRE DE PROJET DE FIN D'ETUDE EN VU D'OBTENTION DU DIPLOME D'INGENIEUR D'ETAT EN GENIE INFORMATIQUE

Option : systèmes d'information avancés

Thème

Application d'une méthode d'analyse de données pour la prévention des accidents du travail

REALISE PAR :
-OULD MOHAMED ELHOSSEIN Cheikhna

Proposé par :

- Mme S. BENSETTITI
- Mr F. HANNANE

ANNEE UNIVERSITAIRE : 2003-2004

Remerciements

Je tiens tout particulièrement à remercier mes promoteurs : Madame **BENSETTITI** et Monsieur **HANNANE** de m'avoir guidé dans mon travail, pour les encouragements qu'ils m'ont toujours donnés, ainsi que pour leur constante disponibilité.

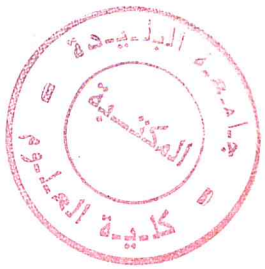
Je remercie aussi Monsieur Taleb : chef du département de statistiques et enquêtes des risques professionnels de la Caisse Nationale de Sécurité Sociale pour ses conseils et ses encouragements.

Je remercie aussi Monsieur Djamal GACEB.

Mes remerciements s'adressent également aux enseignants du département d'informatique et tous ceux qui m'ont aidé de prés ou de loin dans la réalisation de mon projet de fin d'étude.

Enfin Je remercie l'Algérie pour sa générosité et son grand cœur, de nous avoir accueilli pendant les cinq années précédentes.

OULD MOUHAMED ELHOUSSEIN Cheikhna



Tables des Matières

Introduction

Dans ce travail nous avons utilisé et programmé l'analyse canonique qui est une méthode d'analyse des données afin de traiter les problèmes liés à la prévention des accidents du travail.
L'analyse canonique permet l'étude de deux groupes de caractères mesurés sur un ensemble d'individus. Il s'agit de voir le lien qui existe entre ces deux groupes de caractères.
Le logiciel établi permet l'étude des liaisons entre les deux groupes de caractères : les matériel utilisé et les différentes lésions du corps humain. D'une façon générale le logiciel permet aussi de traiter également n'importe quel exemple d'analyse canonique.
Les trois parties I, II et III présentent La Caisse Nationale de Sécurité Sociale (C.N.S.S), les accidents du travail et le but de l'étude.
La partie IV est un rappel de la nature des données que nous utilisons dans notre travail.
La partie V et VI est une présentation de l'Analyse en Composantes Principales (A.C.P) et de l'analyse Factorielle des Correspondances (A.F.C).
L'essentiel de notre travail est constitué par les parties : VII pour l'Analyse Canonique, VIII pour l'implémentation du logiciel que nous avons établi et enfin IX pour l'utilisation et la description du logiciel.

I- PRESENTATION DE LA C.N.S.S

La Direction Générale de la Caisse National des Sécurité Social est chargée, notamment :

- D'organiser, de coordonner et de contrôler
  - ❖ Les activités des agences de wilaya et de leurs d'antennes
  - ❖ d'administration, d'entreprise et des établissements
  - ❖ La gestion des moyens humains et matériels de la caisse
- De gérer le budget de la caisse, de coordonner les opérations financières et de centraliser la comptabilité générale
- D'organiser le contrôle médical
- D'attribuer un numéro d'immatriculation national aux assurés sociaux et aux employeurs
- D'organiser l'information des assurés sociaux et des employeurs
- De suivre l'application des conventions et accords en matière de sécurité sociale
- De promouvoir la prévention des accidents du travail et de maladies professionnelles

La direction générale comprend les structures suivantes:

- La Direction des prestations
- La Direction du recouvrement et du contentieux
- La Direction de l'inspection générale
- La Direction du contrôle médical
- La Direction de la prévention des accidents du travail et des maladies professionnelles
- La Direction des études, de l'organisation et des statistiques
- La Direction de l'informatique
- La Direction des opérations financières
- La Direction des réalisations, équipements et moyens généraux
- La Direction des personnels et de la formation

- La Direction de l'action sociale et sanitaire

Il s'avère important de dresser les taches essentielles des principales directions notamment :

1) **la direction de la prévention des accidents du travail et des maladies professionnelles** est chargée :

- ❖ De contribuer à mettre en oeuvre les mesures arrêtées en matière de prévention des risques professionnels
- ❖ D'élaborer et de proposer le programme d'action de la caisse en matière de prévention
- ❖ De gérer le fonds de prévention
- ❖ De centraliser et d'exploiter les enquêtes effectuées auprès des entreprises
- ❖ D'organiser des séminaires de sensibilisation sur les questions relevant de ses attributions

2) **la direction des études, de l'organisation des statistiques** est chargée :

- ❖ D'effectuer des études et de faire des propositions en matière d'investissements, dans le cadre des procédures établies
- ❖ D'étudier, d'élaborer et de proposer des ratio-types de gestion,
- ❖ D'effectuer des études
- ❖ De collecter, de centraliser et de traiter les données et les informations statistiques
- ❖ D'élaborer et de mettre en oeuvre des programmes d'informations en direction des assurés sociaux et des employeurs
- ❖ De mettre en place des procédures d'information en direction de travailleurs de la caisse

3) **la direction de l'informatique** est chargée :

- ❖ d'élaborer le plan informatique de la caisse et de mettre en œuvre le dispositif et de l'adapter aux besoins de la caisse
- ❖ de mener les études informatiques et d'assurer la réalisation des applications informatiques
- ❖ de gérer la maintenance des équipements informatiques et l'assistance technique pour leur manipulation

II- Le phénomène accident du travail

1-Définition

- **Accident du travail**

D'après la loi 83-13 du 2 juillet 1983 est considéré comme accident du travail, tout accident ayant entraîné une lésion corporelle, imputable à une cause soudaine extérieure et survenue dans le cadre de la relation de travail.

Est assimilé à un accident du travail, l'accident survenu pendant le trajet effectué par l'assuré pour se rendre à son travail ou en revenir quel que soit le mode de transport utilisé, à condition que le parcours n'ait pas été, sauf urgence ou nécessité, cas fortuit ou force majeure, interrompu ou détourné.

Le parcours ainsi garanti est compris entre d'une part le lieu de travail et d'autre part le lieu de résidence ou un lieu assimilé tel que celui où le travailleur se rend habituellement soit pour prendre ses repas, soit pour des motifs d'ordre familial.

- **Maladie professionnelle**

Sont considérées comme maladies professionnelles : les intoxications, infections et affections présumées d'origine professionnelle particulière.

La liste des maladies présumées d'origine professionnelle probable ainsi que la liste des travaux susceptibles de les engendrer et la durée d'exposition aux risques correspondant à ces travaux a été fixée par arrêté interministériel du 5 mai 96.

2-Problematique de l'accident du travail

Les accidents du travail, qui sont des phénomènes liés à l'industrialisation d'un pays, imposent un lourd tribut de vies humaines et de souffrances et une charge financière considérable aux régimes de sécurité sociale. Le dépenses en matière d'accident de travail et supportées par la sécurité sociale sont les huit milliards de dinar octroyés chaque années. L'évaluation de ces coûts est faite sur les frais directs uniquement (indemnités journalières, frais médicaux etc...). Le coût dû aux frais indirects, supporté par les entreprises et les sociétés est trois à quatre fois plus important (temps perdu, matériel détérioré, matières premières abîmées, etc ....).

Sur le plan social, l'accident du travail se traduit au niveau du travailleur par des dommages physiques entraînant la nécessité de soins et aussi par des pertes des salaires. Dans de nombreux cas il y a réduction de la capacité du travail pouvant aller jusqu'à l'arrêt total et définitif de toute activité professionnelle.

La prévention est donc à la fois une nécessité humaine et une nécessité économique. Mais pour prévenir les accidents, encore faut-il connaître leurs causes qui sont généralement multiples et complexes. En effet, un accident ne se produit jamais par une cause unique ; il résulte de plusieurs causes qui s'interpénètrent et dont l'importance est difficile à évaluer. On s'accorde généralement à reconnaître que 21 à 25% seulement des accidents sont imputables aux causes techniques (défaillance ou mauvais état du matériel, insuffisance de protection ou mauvaise conception des machines, ...) et que 75 à 80% de ceux-ci sont dus à des causes humaines. Pour qu'une prévention soit efficace et rentable, elle doit porter aussi bien sur les facteurs techniques que sur les facteurs humains.

La prévention humaine portera essentiellement sur la sélection d'une main-d'œuvre qualifiée, sur sa formation aussi, et doit veiller surtout à ce qu'il y ait généralement une prise de conscience.

La prévention technique, dont le domaine est peut-être plus restreint doit cependant précéder la prévention sur les facteurs humains. Elle portera essentiellement sur l'établissement des consignes de sécurité et leur respect, sur la signalisation, la conception des bâtiments, la conception des machines. Depuis de nombreuses années déjà, des organismes se sont penchés sur l'étude de ce phénomène et s'attellent à prévenir ces accidents du travail.

Citons :

Pour la France I.N.R.S (Institut National des Recherches et de Sécurité).

Pour l'Italie le C.I.D.A.T (Centre d'Information et de Documentation sur les Accidents du Travail).

Pour l'Algérie : la direction de la prévention de la caisse national de Sécurité sociale.

Dans les pays occidentaux, des moyens scientifiques importants sont utilisés pour analyser ces accident du travail et essayer de relier les causes aux effets pour tenter de dégager une politique de prévention dans chaque secteur, chaque branche de l'industrie. Ces moyens scientifiques sont :

1) La mise en place d'un système de données sur les accidents du travail

2) L'utilisation des résultats développés dans la branche mathématique qui est l'analyse des données

3) L'utilisation de l'informatique

Le but de l'étude

Le décret 97 424 attribue à la C.N.S.S non seulement la tache de réparer et d' indemniser les **accidents du travail** et les **maladies professionnelles**, mais aussi et surtout de les **prévenir**. Ceci parallèlement ou en collaboration avec d'autres organismes tels que l'inspection du travail, les commissions d'hygiène,...

Pour la sécurité sociale, les missions de prévention s'articulent principalement à la promotion de la prévention en milieu de travail par l'étude de toutes mesures susceptibles de réduire les accidents et les charges financières.

Ainsi, la **direction de la prévention** est tenue à mettre en place un système de données pour la collecte des renseignements, d'assurer son pilotage et d'établir des statistiques

Les statistiques permettent une meilleure connaissance de la fréquence, de la nature, de la gravité des accidents dans chaque branche d'activités économiques. Elles permettent aussi de connaître les catégories de travailleurs qui ont été victimes s'autres facteurs influant sur le risque auquel sont exposés les travailleurs sont déterminés. La politique de prévention dépend en partie des renseignements fournis par ces statistiques qui permettent de se rendre compte des progrès ou au contraire du recul de la prévention, de mesurer les résultats obtenus par rapport aux efforts engagés et d'ajuster les programmes d'actions. Une tentative d'analyse de ces accidents du travail a été faite avec l'**analyse statistique classique** : croisement de critères, pourcentages, test du khi-2, écart-types, moyennes,...Mais cette analyse est limitée en soit du fait de sa dispersion et de la difficulté d'interpréter de trop nombreuses données.

IV- Nature des données et concepts fondamentaux

1- les tableaux de données

On distingue généralement deux ensembles : **les individus** et **les caractères** relatifs à ces individus. Les caractères observés peuvent être **quantitatifs** s'ils prennent des valeurs numériques. ils sont dits **qualitatifs** lorsqu'ils possèdent des modalités non numériques.

1.1-Tableaux individus-caractères quantitatifs

Les p caractères quantitatifs $x^1,x^2,\ldots,x^p$ sont observés sur un ensemble de n individus $e_1, e_2,\ldots,e_n$. On obtient une matrice Individus-caractères que l'on notera $x=(x_{ij})$ et dont la dimension est (np) (n lignes et p colonnes)

$$x = \begin{array}{c} \\ e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{cccccc} x^1 & \ldots & x^j & \ldots & x^p \end{array} \\ \begin{bmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & & x_{nj} & & x_{np} \end{bmatrix} \end{array}$$

1.2-Tableaux individus-caractères qualitatifs

Sur les mêmes individus, on aurait pu observer les caractères qualitatifs : sexe, niveau hiérarchique, lieu de l'accident, la nature de lésions. Ces caractères peuvent avoir plusieurs modalités. Pour le traitement numérique, ces caractères qualitatifs sont représentés sous forme d'un tableau de **variables indicatrices** prenant les valeurs 0 ou 1. On dit alors que les données sont représentées sous **forme disjonctive**. Cette représentation des

caractères qualitatifs permet de les assimiler à des caractères quantitatifs prenant les valeurs 0 et 1.En effet un caractère quantitatif peut être rendu qualitatif par découpage en classes de ses valeurs, puis représenté sous forme de variables indicatrices (classes de revenus, classes d'ages etc....).

Sexe

Individus	Mdalité1	Modalité2
	Masculin	Féminin
e1	0	1
e2	1	0
	...	...
e_i	1	0
	...	...
	1	0
e_n	0	1

1.3-Tableaux de Contingence

Un tableau de contingence ou tableau croisé contient **les effectifs ou fréquences d'association** entre les modalités de deux caractères qualitatifs observés sur un ensemble de n individus. On peut par exemple considérer le tableau croisé des catégories socioprofessionnelles (p modalités) avec les quartiers d'une ville(q modalités).Une case (i,j) de ce tableau contient le nombre n_{ij} d'individus exerçant la profession i et habitant le quartier j.ce tableau sera noté $N = (n_{ij})$ $i = 1,...,p$ et $j = 1,...,q$. Sa dimension sera pq (p lignes et q colone

	Quartier		
	... modalité j ...		modalité q
Modalité1			
Modalité i	... ...	n_{ij}	
Modalité p			

Dans un tel tableau, les individus ont été regroupés et ne peuvent plus être distingués. On peut concevoir une autre représentation en dissociant les deux caractères. A chacun d'eux on associe un tableau de variables indicatrices X_1 et X_2. En ligne on représente les individus et en colonne les modalités de chaque caractère. Une ligne ne contient alors que des 0 sauf dans la colonne correspondant au quartier considéré où l'on trouve des 1. Notons que le tableau de contingence N est le résultat du produit matriciel ${}^tX_1 X_2$ où ${}^tX_1$ désigne la matrice transposée de la matrice X_1.

$$N = {}^tX_1 X_2$$

exemple :considérons huit individus possédant trois fonctions et habitant deux quartiers.

X_1 =

	Professions		
e1	1	0	0
e2	0	1	0
e3	0	1	0
e4	1	0	0
e5	0	0	1
e6	0	0	1
e7	0	1	0
e8	0	0	1

X_2=

	quartiers	
e1	1	0
e2	1	0
e3	0	1
e4	1	0
e5	0	1
e6	0	1
e7	1	0
e8	1	0

Quartier

$$N = {}^tX_1X_2 = \text{profession} \begin{bmatrix} 2 & 0 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}$$

1.4 tableaux de proximité

Etant donné un ensemble d'individus. On dispose d'une mesure de **ressemblance** entre ces individus pris deux à deux. On peut alors construire un tableau de proximité dans lequel les lignes ainsi que les colonnes représentent ces individus. un tel tableau est carré, symétrique et généralement ne contient que des nombres positifs. On peut citer par exemple, le tableau où figurent les distances entre les principales villes d'un pays.

2- Réduction des données

- L'ensemble des mesures d'un caractère quantitatif x^j peut être résumé par trois nombres : la moyenne, la variance et l'écart type.

-La moyenne arithmétique $\overline{x^j}$ du caractère x^j : est un paramètre de position car elle permet de localiser le nuage des points des individus sur un axe représentatif. Elle donne l'abscisse du centre de gravit é du nuage de points.

$$\overline{x}^j = \frac{1}{n}\sum_{i=1}^{i=n} x_{ij}$$

-La variance s^2_j : permet quant-à-elle de mesure la dispersion de ce nuage autour du centre de gravité. Plus ce nombre est grand et plus le nuage est dispersé par rapport à son centre de gravité. Une valeur de variance voisine de zéro, signifie que tous les points sont concentrés autour du centre de gravité.

$$s_j^{\,2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}^j\right)^2$$

-L'écart -type s_j :est défini comme étant la racine carrée de la variance. On le mesure avec la même unité que le caractère correspondant.

Dans les formules précédentes chaque individu était considéré comme ayant le même poids $p_i = 1/n$. On peut privilégier certains individus par rapport à d'autres en leur affectant des poids différents. Les formules précédentes deviennent alors :

$$\bar{x}^j = \frac{1}{n}\sum_{i=1}^{i=n} p_i x_{ij} \qquad s_j^2 = \frac{1}{n}\sum_{i=1}^{n} p_i \left(x_{ij} - \bar{x}^j\right)^2 \text{ avec } \sum_{i=1}^{n} p_i = 1$$

- **La covariance** entre deux caractères quantitatif est définie par la relation :

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}^j\right)\left(x_{ik} - \bar{x}^k\right)$$

- **Histogramme.** Les observations recueillies sur un caractère quantitatif, peuvent être également synthétisées à l'aide d'un histogramme

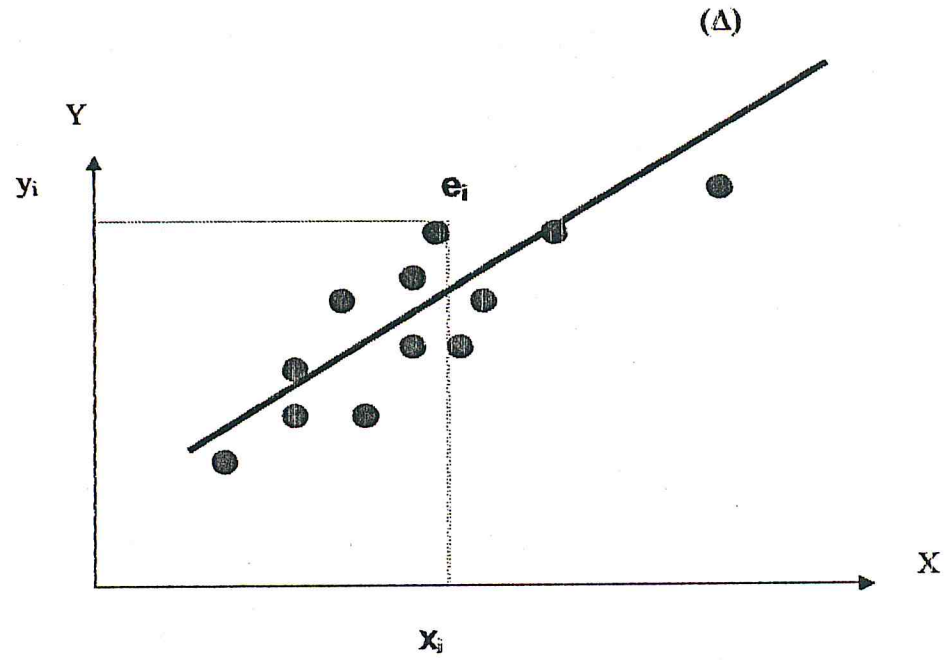
Lorsqu'on observe un caractère qualitatif sur un ensemble d'individus, la première idée qui vient à l'esprit est de compter le nombre d'individus dans chaque modalité. On peut ensuite calculer la fréquence ou la pourcentage de chaque modalité.

3-Liaison entre deux caractères

- **Liaison entre deux caractères quantitatifs**

Considérons en ensemble de n individus { $e_1,\ldots,e_n$} sur lequel nous avons observé deux caractères quantitatifs x et y.

	X	y
e_1	x_1	y_1
e_2	x_2	y_2
.	.	.
.	.	.
.	.	.
e_n	x_n	y_n



Chaque individu, ayant deux coordonnées est un point de R^2. L'ensemble des individus constitue **un nuage de points** dans le plan xoy. Si ces points sont répartis de façon linéaire on essayera de tracer **la droite de régression** qui passe le plus prés possibles de tous les points et ensuite de mesurer la dépendance entre les deux caractères en calculant **le coefficient de corrélation r.**

L'équation de la droite s'écrit : y = ax+b. Les coefficients a et b sont obtenus par **la méthode des moindres carrés,** c'est à dire choisis de façon à rendre minimale la somme quadratique $\sum_{i=1}^{n} (u_i)^2$ où $u_i = y_i - (ax+b)$ est l'écart entre le point e_i et la droite cherchée. On montre que cette droite passe par le centre de gravité g $(\overline{x}, \overline{y})$ où $\overline{x}$ et $\overline{y}$ désignent les moyenne des caractères x et y. On montre également que le rapport $\frac{\sum_{i=1}^{n} (u_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})}$

Est toujours inférieur à $1-r^2$

$$1-r^2 = \frac{\sum_{i=1}^{n} (u_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})}$$

r est un nombre appartenant à l'intervalle [-1,1]. On l'appelle coefficient de corrélation des deux caractères x et y. C'est un nombre qui mesure en quelque sorte **la force de la liaison entre deux caractères** : plus il est grand en valeur absolue et plus les points du nuage sont plus proches de la droite de régression(la connaissance de la valeur d'un caractère implique celle de l'autre). Si r = 0, la droite est parallèle à l'axe de x (la valeur de x ne joue aucune rôle pour prévoir y). Si $r=\mp 1$ la précision est parfaite, les écarts sont tous nuls et les points se trouvent tous sur la droite de régression.

- **Liaison entre deux caractères qualitatifs**

Pour mesurer la dépendance entre deux caractères qualitatifs d'un tableau croisé, la statistique classique propose de calculer le **X^2 de contingence** largement utilisé en analyse de donnés,

$n = \sum_{i,j} n_{ij}$ effectif total dans le tableau

$n_i = \sum_{j=1}^{q} n_{ij}$ effectif marginal ligne

$n_j = \sum_{i=1}^{p} n_{ij}$ effectif marginal colonne

On calcule la quantité :

$$D^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i . n_j}{n}\right)^2}{\frac{n_i . n_j}{n}} = n\left[\sum_{i,j} \frac{n_{ij}^2}{n_i . n_j} - 1\right]$$

Si on suppose que les deux caractères observés soient indépendants. Dans ce cas la connaissance de l'un d'entre eux n'apporte rien à la connaissance de l'autre. La probabilité $p_{ij} = n_{ij}/n$ D'avoir simultanément les modalités i et j ne dépend que des probabilités $p_{ij} = p_i p_j$ où $p_i = n_i/n$ et $p_j = n_j/n$. Si les deux caractère sont indépendants, le numérateur du D^2 sera voisine de zéro :

$$\frac{n_{ij}}{n} = \frac{n_i . n_j}{n\ n} \Rightarrow \frac{1}{n}\left(n_{ij} - \frac{n_i . n_j}{n}\right) = 0$$

V- Analyse en Composantes Principales

Cette méthode a pour objet la description des données contenues dans un tableau individus-caractères numériques: p caractères sont mesurés sur n individus.

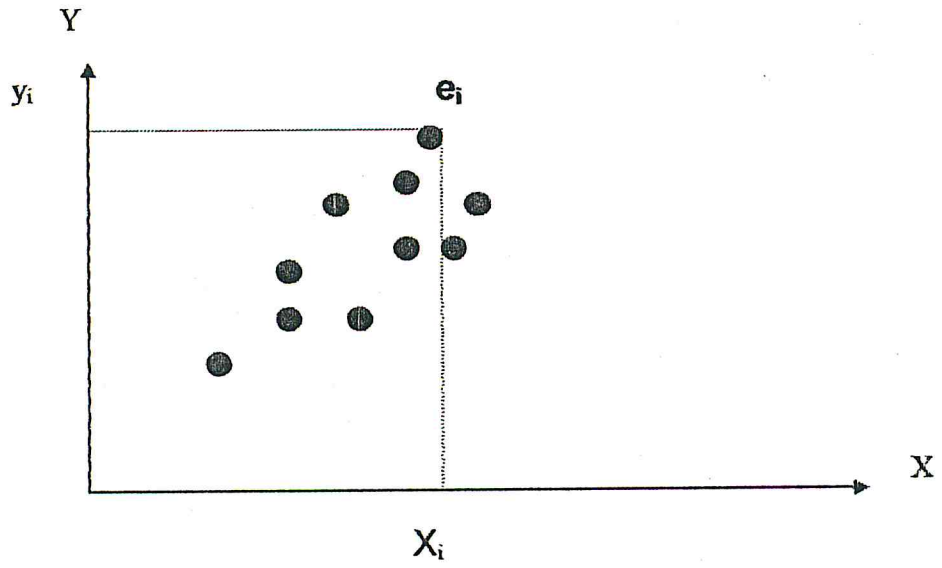
1- Présentation de la méthode :

Supposons que l'on ait observé deux caractères quantitatifs x^1 et x^2 sur un ensemble de n individus $\{e_1, e_2,\ldots,e_n\}$.Nous obtenons un tableau individus-caractères quantitatifs x à n lignes et 2 colonnes :

$$X = \begin{array}{c} \\ e_1 \\ e_2 \\ \\ \\ \\ e_n \end{array} \begin{array}{c} \begin{array}{cc} x^1 & x^2 \end{array} \\ \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ . & . \\ . & . \\ . & . \\ x_{n1} & x_{n2} \end{bmatrix} \end{array}$$

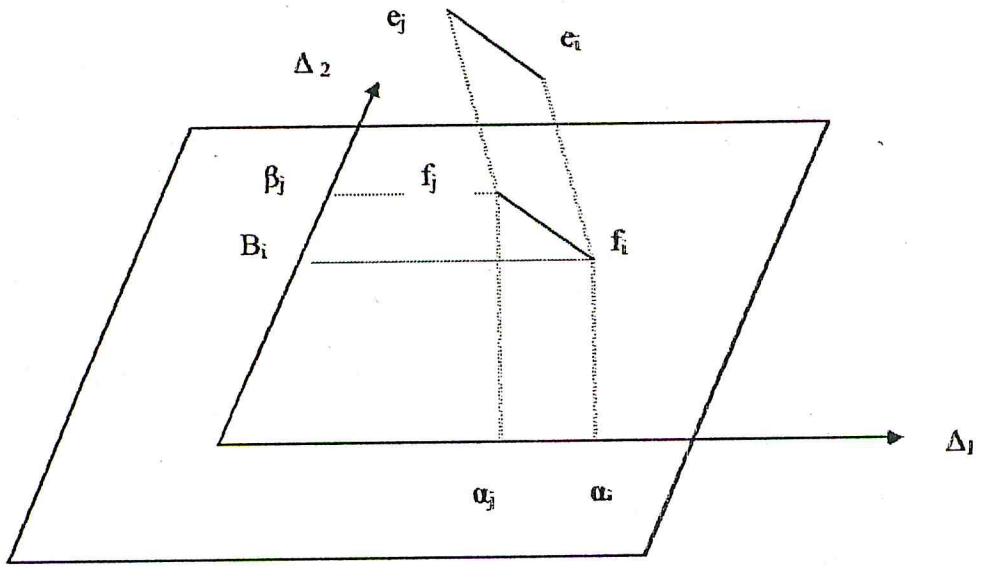
Il est facile de représenter, sur un graphique plan l'ensemble de ces données afin de repérer les individus ou groupes d'individus présentant des caractéristiques voisines, et d'étudier la liaison entre les deux caractères x^1 et x^2.On peut également essayer d'expliciter les variations d'un des deux caractères en fonction de celle de l'autre.

Chaque individu e_i est un point de R^2,de coordonnées x_{i1}et x_{i2}.



Supposons que l'on veuille quand même avoir une représentation de ces n individus en les projetant sur un graphique plan. Les distances entre les n individus sur le plan ne peuvent être toutes égales aux distances entre les n points sur l'espace complet à p dimensions. Il y aura forcement des distorsions que l'on cherchera à rendre minimales.

Projetons les points individus $e_1, e_2, \ldots, e_n$ sur un plan comme le montre le figure suivante :



Il faudra évidemment choisir le plan de projection sur lequel les distances serons en moyenne le mieux conservées. Comme l'opération de projection raccourcit toujours les distances.

$d(f_i,f_j) \leq d(e_i,e_j)$

On se fixera pour critère derendre maximale la moyenne des carrés des distances entre les projections $f_1,f_2,\ldots,f_n$.

Pour déterminer ce plan que l'on appelle le **plan principal**, il suffit de trouver deux droites $\Delta_1$et Δ_2 sont perpendiculaire on a alors :

$d^2(f_i,f_j)=d^2(\alpha_i, \alpha_j)+d^2(\beta_i, \beta_j)$

Où α_i et α_j sont les projections de e_i et e_j sur l'axe Δ_1 et β_i et β_j les projections sur Δ_2.

La méthode consiste donc à chercher d'abord Δ_1 rendant maximale la moyenne des $d^2(\alpha_i, \alpha_j)$puis un axe Δ_2 perpendiculaire à Δ_1 et rendant maximale la moyenne des distances $d^2(\beta_i, \beta_j)$.

On peut continuer en dehors du plan et chercher alors Δ_3 et $\Delta_4,\ldots,\Delta_p$, perpendiculaires entre eux et qui sont appeler **axe principaux** du nuages.

La projection du point individu e_i , de coordonnées initiales $x_i^1,x_i^2,\ldots,x_i^p$, sur les axes principaux conduit a l'obtention de nouvelle coordonnées $c_i^1,c_i^2,\ldots,c_i^p$. on construit ainsi des nouveaux caractères $c^1,c^2,\ldots,c^p$ qui sont appelées **composantes principales**.

Chaque composante s'écrite comme une combinaison linéaire des caractères initiaux :

$c^k=u_1^k x^1+u_2^k x^2+\cdots+u_p^k x^p$

Où $(u_1^k,u_2^k \cdots u_p^k)$ représente le k-ième facteur principal u^k.

Tel est le schéma de l'analyse en composantes principales qui est une méthode d'obtenir un résumé descriptif (sous forme graphique le plus souvent) de l'ensemble de n observations effectuées sur p

caractères numériques, et on considère que les individus et les caractères sont des éléments de deux espaces vectorielles euclidiens à p et n dimensions respectivement.

Les outils mathématiques utilisées sont ceux de l'algèbre linéaire et du calcule matriciel.

Le problème sera :

- comment calculer la distance entre deux individus, entre deux caractères ?
- Comment résumer les caractéristiques du tableaux des données ?

C'est les problèmes qui nous allons traiter dans ce qui suit.

2- Résumé du tableau de données

Les caractères $x^1,x^2,\ldots,x^p$ peuvent être résumés en calculant respectivement leur moyenne, leur variance et leur écart type.

Nous pouvons calculer également les covariances s_{jk} entre les différents caractères pris deux à deux :

$$S_{jk} = \frac{1}{N}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}^{j}\right)\left(x_{ik} - \overline{x}^{k}\right)$$

Les variances et les covariances sont rassemblées dans un même tableau V appelé **matrice de variances-covariances** :

$$V = \begin{bmatrix} s^2_1 & s_{12} & \cdots & s_{1p} \\ & s^2_2 & & \vdots \\ & & \ddots & \vdots \\ & & & s^2_p \end{bmatrix}$$

V est une matrice carrée d'ordre p .Elle est symétrique car $s_{jk} = s_{kj}$, définie positive. Considérons les vecteurs lignes $e_i = (x_{i1},\ldots,x_{ip})$ et $g = (\overline{x^1},\ldots,\overline{x^p})$.

Nous pouvons représenter également la matrice V par la relation matricielle suivante

$$V = \frac{1}{n}\sum_{i=1}^{n} {}^{t}(e_i - g)(e_i - g)$$

Si X est le tableau à n lignes et p colonnes des données centrées, on a la relation matricielle :

$$V = {}^{t}XDX$$

Où ${}^{t}X$ est la matrice transposée de X et D la matrice (d'ordre n) diagonale de poids :

$$D = \begin{bmatrix} \frac{1}{n} & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \frac{1}{n} \end{bmatrix}$$

Nous supposons pour la suite que les caractères sont centrés. De même l'ensemble des coefficients de corrélation est regroupé dans une matrice de corrélation R dont les termes diagonaux valant 1 puisque $r(x^j, x^j) = 1$.

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \vdots \\ \vdots & & \ddots & \\ r_{p1} & \cdots & & 1 \end{bmatrix}$$

perpendiculaires et les unités de mêmes natures, on utilise simplement la formule de Pythagore pour calculer la distance entre les deux individus e_1 et e_2 :

$$d^2(e_1,e_2)=(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+....+(x_{1p}-x_{2p})^2$$

En Analyse des données on ne peut pas utiliser une telle formule puisque les unités avec lesquelles s'expriment les caractéristiques sont différentes. Si on utilise une formule de forme :

$$d^2(e_1,e_2)=a_1(x_{11}-x_{21})^2+a_2(x_{12}-x_{22})^2+....+a_p(x_{1p}-x_{2p})^2$$

Cela reviendrait à multiplier par $\sqrt{a_j}$ le caractère x^j. D'une façon générale nous pouvons prendre une formule de la forme :

$$d^2(e_1-e_2)=\sum_{k=1}^{p}\sum_{j=1}^{p}m_{kj}(x_{1k}-x_{2k})(x_{1j}-x_{2j}).$$

Qui s'écrit sous la forme matricielle :

$$d^2(e_1-e_2)={}^t(e_1-e_2)M(e_1-e_2)$$

la distance entre ces deux individus s'écrira alors :

$$d^2(e_1-e_2)=\|e1-e2\|_M^2={}^t(e_1-e_2)M(e_1-e_2)$$

Où M=(m_{kj}) désigne une matrice symétrique définie positive d'ordre p

$$M=\begin{bmatrix} m_{11} & \cdots & m_{1p} \\ \vdots & \cdots & \vdots \\ m_{p1} & \cdots & m_{pp} \end{bmatrix}$$

Ceci revient à définir un **produit scalaire** sur l'espace des individus R^p :

$$\langle e_1,e_2\rangle_M={}^t e_1Me_2$$

On dit alors que l'on a muni l'espace des individus R^p d'une structure euclidienne. La matrice M s'appelle alors la **matrice de l'espace.** La longueur du vecteur e_1 se note $\|e_1\|_M$ et s'appelle la

M-norme de e_1 . Nous avons $\|e_1\|_M^2=\langle e_1,e_2\rangle_M$. Les métriques les plus utilisées en ACP sont les **métriques diagonales :**

$$D_a=\begin{bmatrix} a_1 & & 0 \\ & a_j & \\ & & \\ 0 & & a_p \end{bmatrix}$$

Utiliser une métrique diagonale revient à multiplier le caractère x^j par $\sqrt{a_j}$ et utiliser ensuite la métrique usuelle M=I. en particulier on utilise fréquemment la métrique $M=D_{\frac{1}{s2}}$, appelée **métrique de l'analyse en composantes Principales:**

$$D_{\frac{1}{s^2}}=\begin{bmatrix} \frac{1}{s_1^2} & & & & 0 \\ & \frac{1}{s_2^2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ 0 & & & & \frac{1}{s_p^2} \end{bmatrix}$$

Chaque caractère est divisé par son écart type.

4- Ajustement dans R^p

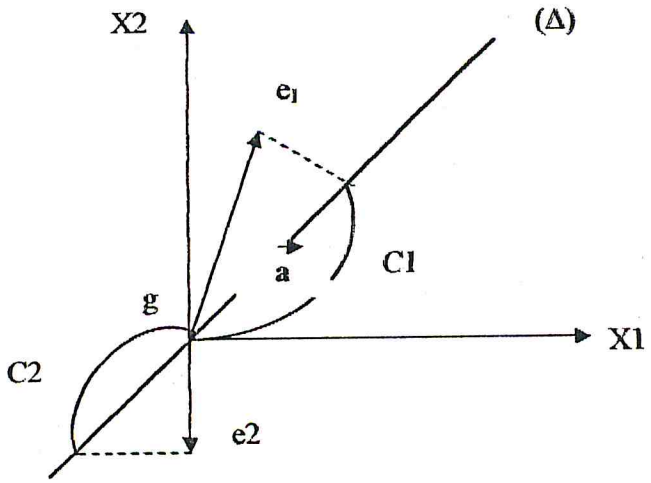
Commençons par chercher un sous-espace vectoriel à une dimension. C'est à dire une droite passant par l'origine, qui réalise le meilleur ajustement possible du nage de points.

. Soit $\vec{u}$ un vecteur unitaire on désignera également par u la matrice colonne associée et par u sa transposée. puisque $\vec{u}$ est unitaire on a :

$$u^t u=\sum_{j=1}^{p} u_j^2=1$$

4.1 Comment calculer les coordonnées des individus sur un nouvel axe

Considérons le système d'axe orthonormé représentant les caractères initiaux $x^1,x^2,....,x^p$ .En projetant les individus sur une droite quelconque Δ on crée un nouveau caractère c dans le valeur $(c^1,c^2,....,c_n)$ sont les mesures algébriques des projection des point e_i sur cette droite c est appelée composante principale.



Soit le vecteur unitaire $\bar{a}$ de (Δ) de M- norme 1.la mesure algébrique c_1 de la projection de l'individus e_1 est alors égale au produit scolaire e_1 par a c'est à dire :

$c_1={}^t e_1Ma={}^taMe_1={}^t(Ma)e_1$ car M est symétrique

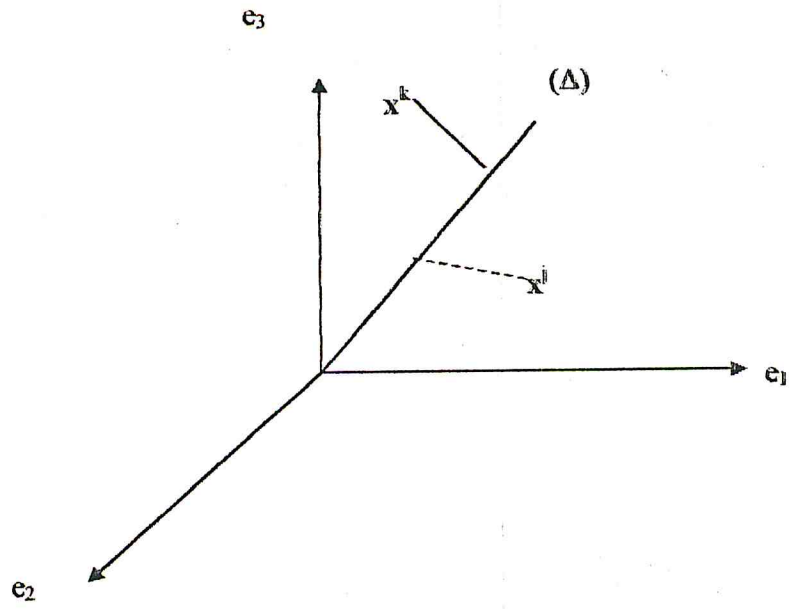
On posant $u=Ma$ alors on peut écrire que la composante c_1 de e_1 sur Δ vaut

$$c={}^tue_i=\sum_{j=1}^{p}u_jx_{ij}$$

le caractère c dont les valeurs sont les n coordonnées $c_1,c_2,...,c_n$ s'obtient alors directement par la formule $c=Xu$,c'est donc une combinaison linéaire des p caractère initiaux au moyen du facteur u ,si $M=I$ il y aura égalité entre les facteur u et le vecteur a.

5. L'espace des caractères R^n

Chaque caractère x^j possède n coordonnées $(x_{1j},x_{2j},...,x_{nj})$ .C'est donc un point de l'espace vectoriel R^n que l'on appelle espace des caractère. les p caractères forment un nuage de p point de R^n.

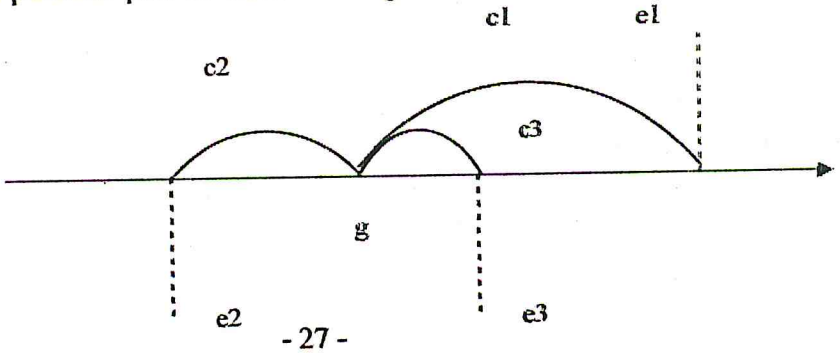


Soit $\bar{v}$ le vecteur unitaire d'une droite Δ_1 ,pour que ce vecteur ajuste au mieux au sens de moindre carrées le nuage des p points de R^n ,il faut comme précédemment que la somme des carrées des projections sur Δ_1 soit maximale, ce qui équivaut à chercher le maximum de la forme quadratique $^t(xv)(xv)$ avec contrainte $v^tv=1$

Nous savons maintenant que v est le valeur propre v_1 de la matrice X^tX relatif à la plus grande valeur propre λ_1 ,on calculerait de même les vecteurs propres $v_2,..v_p$ engendrant le meilleur sous espace d'ajustement à q dimension $(q \leq n)$ Au sens moindre carrée.

6. Composantes, axes et facteurs principaux :

Nous avons que la 1er axe principale Δ_1 avait pour propriété de rendre maximale la moyenne des carrées des distances des projections des points du nuage au centre de gravité. Ceci équivaut à rendre maximale l'inertie des projections qui vaut $p_i c_i^2$ ou les c_i sont les mesures algébriques des projections des e_i sur Δ_1 car on choisit de faire passer par le centre de gravité g du nuage.



Δ_1 est l'axe d'allongement principal du nuage en ce sens que sur cette axe les c_i sont les plus dispersées possible c'est à dire : c est combinais linéaire des x^j de variance maximale.
Pour trouver explicitement facteurs et composantes principales et pour alléger la démonstration on peut toujours se ramener au cas $M = I$ et en raisonnant sur la tableau des données transformées $Y = X\,{}^tT$ avec $M = {}^tTT$

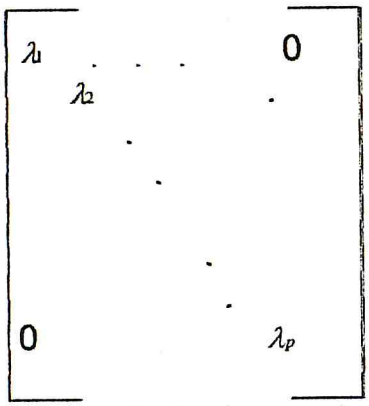
La matrice de variance covariance de Y sera $V_y={}^tYDY=R$ matrice des corrélations. La composante principale c à pour variance $s_c^2={}^tvV_y v$ ou v est égale au vecteur unitaire de l'axe principale. Le problème est de trouver le vecteur v, de norme 1 tel que ${}^tvV_y v$ soit maximale ,ceci est équivalent à rendre maximale le quotient : $\frac{v^tV_y v}{v^tv}$ car ${}^tvv=1$.

Le maximum est atteint si $d\left[\frac{v^tV_y v}{{}^tvv}\right]/dv=0$ c'est à dire $2({}^tvv)V_y v-2(vV_y)v=0$.

Soit : $V_y v=({}^tvV_y v)v=\lambda v$

v doit être donc vecteur propre de V_y et sa valeur propre λ doit être la plus grande puisqu'elle représente la quantité à maximiser, la variance de c vaut alors λ car v est de norme 1 ($s_c^2=(c,c)M={}^tcMc={}^t(Xv)TT(Xv)=v^tV_y v=\lambda$). Comme la matrice de variance est symétrique et définie positif, alors elle possède p vecteur propres orthogonaux deux à deux et ses valeurs propres sont toutes positives ou nulles.

Prendre comme nouveaux axes de l'espace des individus des vecteurs de la matrice de variance revient à diagonaliser l'opérateur linéaire associé à V_y ,la matrice de variance des composantes principales V_c est égale à :

$V_c =$ 

Les composantes principales sont donc corrélés 2 à 2 $cov(c_i,c_i)=0$. L'ACP remplace alors les p caractères initiaux par des caractères non corrélé de variance maximale pour trouver directement axes, facteurs et composantes en fonction de X il suffit écrire :

$$V_{y\mu} = \lambda v = TV\,{}^tTv$$

Et en multiplier à gauche par ${}^tT$ on obtient :

$${}^tTTVT{}^tv = \lambda\,{}^tTv \text{ on pose } ({}^tTv = u)$$

$$\Rightarrow Mvu = \lambda u$$

L'axe a est tel que =Ma donc $MVMa = \lambda Ma$

Si M est régulière (il existe M^{-1}) alors on peut simplifier par M^{-1} on obtient les axes principaux sont donc les vecteurs propres de VM et les facteurs principaux u ceux de MV, et les composants principaux s'obtient par c = Xu

On remarque que $MV = M\,{}^tXDX \Rightarrow M\,{}^tXDX = \lambda u$

Multiplions à gauche par X on obtient :$(XM\,{}^tXD)c = \lambda c$, donc c est un valeur propre de la matrice $XM\,{}^tXD$ associée à la valeur propre λ.

$(\lambda_1+\lambda_2+\ldots+\lambda_p) = \text{trace}V_y = \text{trace } MV = \text{Inertie.}$

Inertie

L'équation $\frac{\lambda_k}{\sum\lambda}$ est appelé la part d'inertie ou de variance expliquée par l'axe numéro k et $\frac{\lambda_1+\lambda_2}{\sum\lambda}$ est appelé l'inertie cumulée des deux premiers axes principaux elle mesure l'aplatissement du nuage sur le plan principal. Plus cette part est grande, la représentation du nuage sur ce plan est meilleure.

VI- L'Analyse factorielle des correspondances

Les principales sous-jacents au aspects technique de cette méthode, ont des antécédents assez anciens : on peut les faire remontrer à certains travaux de **HIRSHFLED** (1935) et **FICHER** (1990). On consultera sur ce point les notes historique de **HILL** (1974) et de **BENZERAI** (1976), ce sont les travaux de dernier auteur qui, avec l'avènement des ordinateurs, sont responsables du développement de cette méthode. Du point de vue des analyste de données, il s'agit d'une méthode statistique permettons d'analyser et de décrire graphiquement de manière synthétique de grand tables de contingences.

Par ces propriétés mathématiques et la richesse de ses interprétation, l'analyse de correspondances est devenue la méthode privilégiée des description des données qualitatives.

1-Présantation de la méthode :

L'analyse factorielle des correspondances (en abrégé AFC) s'applique à des tableaux des contingences ou tableaux croisés. Comme nous l'avons vu, ce tableau est un tableau N d'effectif n_{ij} correspondant à la ventilation des individus selon deux caractères qualitatifs dont le nombre de modalités sont p pour le premier et q pour le seconde. Le nombre n_{ij} du tableau représente le nombre d'individus ayant à la fois la modalité i et la modalité j.

Prenons l'exemple d'un disquaire qui répartit la vente de 1000 disques, suivant la catégorie de la musique (trois modalités: chansons(C), Jazz(J) et musique classique(Mc)) et la population des utilisateurs (quatre modalités : jeunes sans distinction du sexe(JSDS), adultes féminins(AF), adultes masculins(Ms) et personnes âgées sans distinction de sexe(PASDS)).il a obtenu le tableau suivant :

	JSDS	AM	AF	PASDS	Total
C	69	172	133	27	401
Ja	41	84	118	11	254
Mc	18	127	157	42	345
Total	128	383	408	81	1000

Il est claire qu'il s'agit bien d'un tableau de contingence. Appliquée à un tableau, l'objectif de l'AFC revient à analyser la structure de la dépendance entre les deux caractères qualitatifs et à faire ressortir les traits principaux de cette dépendance.

Remarquons tout d'abord qu'un tableau de contingence peut se lire de deux manières différentes : selon ses lignes ou selon ses colonnes. Cela réponde à deux préoccupations différentes.

a) si on désire s'avoir pour chaque catégorie de musique comment se répartit la population des utilisateurs, on calculera le pourcentages en lignes en le divisant les effectifs n_{ij} de la ligne n°i par le total n_i de la ligne. On obtient se qu'on appelle tableau de profile ligne :

	JSDS	AM	AF	PASDS	$\frac{n_k}{n}$
C	0.17207	0.42893	0.33167	0.06733	0.401
Ja	0.16142	0.33071	0.46457	0.04331	0.254
Mc	0.05217	0.36812	0.45507	0.12464	0.345
$\frac{n_j}{n}$	0.128	0.383	0.408	0.081	1

Le profil marginal n_j est aussi le profil moyen car il est la moyenne des profils des lignes pondérés par le poids $p_i = \frac{n_k}{n}$ de chaque ligne

$$\bar{x}=\sum_{i=1}^{p} p_i \frac{n_{ij}}{n_i} = \sum_{i=1}^{p} {}_i\frac{n_{ij}}{n_i} = \frac{1}{n}\sum_{i=1}^{p} n_{ij} = \frac{n_{ij}}{n}$$

b) si réciproquement ont veut savoir pour une catégorie de population données, comment se répartissent les différentes catégories de musique, on calculera les profile des colonnes en divisant les effectifs n_{ij} de la colonne j par n_{ij} total de la colonne. Le tableau des profils colonnes est donné dans le tableau suivant :

	JSDS	AM	AF	PASDS	$\frac{n_{ih}}{n}$
C	0.069	0.172	0.133	0.027	0.401
Ja	0.041	0.084	0.118	0.011	0.254
Mc	0.018	0.127	0.157	0.042	0.345
$\frac{n_j}{n}$	0.128	0.383	0.408	0.081	1

Si on appelle D_1 et D_2 les matrices diagonales des effectifs marginaux :

$$D_1 = \begin{bmatrix} n_1 & & 0 \\ & n_2 & \\ & & \ddots \\ 0 & & n_p \end{bmatrix} \qquad D_2 = \begin{bmatrix} n_1 & & 0 \\ & n_2 & \\ & & \ddots \\ 0 & & n_q \end{bmatrix}$$

Le tableau renfermant les p profils des lignes est le produit matriciel :

$$D_1^{-1}N = \left(\frac{n_{ij}}{n_i}\right) \qquad \text{i=1...p et j=1...q.}$$

Le tableau des profils des colonnes est le produit matriciel :

$ND_2^{-1} = \left(\frac{n_{ij}}{n_j}\right)$ i=1...p et j=1...q.

1.1 Analyse dans R^q

Si on s'intéresse aux lignes de N, on peut considérer le tableau $D_1^{-1}N$ des profils des lignes comme un tableau individus-caractères particulier, et effectuer une analyse en composantes principales. Les individus de cette analyse sont les profils des lignes , chaque individu i ayant pour coordonnées les quantités $\frac{n_{ij}}{n_i}$ (pour j = 1...q), affecté de la masse $f_i = \frac{n_i}{n}$.

L'ACP revient alors à étudier la dispersion du nuage des p profils dans R^q autour de leur centre de gravité qui n'autre que le profil marginal colonne ($\frac{n_{.1}}{n}$, $\frac{n_{.2}}{n}$,..., $\frac{n_{.q}}{n}$).

En d'autres termes on cherche a rendre compte de l'écartement entre les $\frac{n_{ij}}{n_i}$ et les $\frac{n_{.j}}{n}$, ce qui est une façon d'analyser la dépendance entre les deux caractères qualitatifs.

1.2 Analyse dans R^p

Inversement si on s'intéresse aux colonnes de N, c'est à dire le tableau ND_2^{-1} ou plutôt son transposé $D_2^{-1}N^t$ qui jouera de rôle de tableau individus-caractères, on étudiera alors la configuration des profils des colonnes dans R^p . Chaque individu, j ayant pour coordonnées les quantités $\frac{n_{ij}}{n_{.j}}$ (pour i= 1..p), affecté de la masse $f_j = \frac{n_{.j}}{n}$.

2. Analyse en composantes principales des tableaux deux profils

Pour effectuer une ACP sur ces tableaux, il faut définir une formule de distance entre objets, en d'autre terme une métrique.

2.1 La métrique χ^2 (khi-deux)

On appelle métrique de χ^2 pour les lignes, la matrice diagonale :

$$M_1 = nD_1^{-1} = \begin{bmatrix} \frac{n}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{n}{n_q} \end{bmatrix}_{qxq}$$

Elle est définie par l'inverse du profil marginal des colonnes de N. De la même façon, la métrique de χ^2 pour les colonnes, est définie par l'inverse du profil marginal des lignes de N.

$$nD_1^{-1} = \begin{bmatrix} \frac{n}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{n}{n_p} \end{bmatrix}_{pxp}$$

2.2 ChoiX des distances

Le fait de travailler sur des profils dans les deux espaces R^p et R^q, nous incite à munir ces espaces d'une distance différente de la distance euclidienne usuelle. La distance entre deux catégories i et j sera donné par la formule suivant appelée distance du χ^2.

$$D^2(i,k) = \sum_{j=1}^{q} \frac{n}{n_j}\left(\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}\right)^2$$

De façon symétrique, la distance entre deux points_colonnes j et j' s'écrire :

$$D^2(j,k) = \sum_{i=1}^{p} \frac{n}{n_i}\left(\frac{n_{ij}}{n_j} - \frac{n_{ik}}{n_k}\right)^2$$

2.3. ACP du nuage des profils

Afin de se ramener à la métrique usuelle, on peut écrire la formule exprimant la distance entre profils lignes de manière suivante :

La formule exprimant la distance entre deux profils lignes i et k s'écrit :

$$d^2 = \sum_{j=1}^{q} \frac{n}{n_j}\left(\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}\right)^2$$

$$= \sum_{j=1}^{q} \frac{1}{\frac{n_j}{n}}\left(\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}\right)^2$$

$$= \sum_{j=1}^{q} \left[\frac{1}{\sqrt{\frac{n_{.j}}{n}}}\left(\frac{n_{ij}}{n_i} - \frac{n_{kj}}{n_k}\right)\right]^2$$

avec $f_{ij} = \frac{n_{ij}}{n}$, $f_i = \frac{n_i}{n}$ et $f_j = \frac{n_j}{n}$ alors on obtient

$$d^2(i,k) = \sum_{j=1}^{q} \left[\left(\frac{f_{ij}}{\sqrt{f_j}\, f_i} - \frac{f_{kj}}{\sqrt{f_j}\, f_k}\right)\right]^2$$

Cette dernière formule permet de se ramener à la métrique usuelle où le profil ligne devient alors : $\frac{f_{ij}}{f_i\sqrt{f_j}}$. Calculons les coordonnées du centre de gravité du nuage :

$$g_j = \sum_{i=1}^{q} p_i\left(\frac{f_{ij}}{f_i}\sqrt{f_j}\right)$$

$$= \sum_{i=1}^{q} f_i\left(\frac{f_{ij}}{f_i}\sqrt{f_j}\right)$$

$$= \sqrt{f_i}$$

Les coordonnées centrées sont :Appliquons

$(\frac{f_{ij}}{f_i}\sqrt{f_j} - \sqrt{f_i})$.

Appliquons au tableau des profils lignes X de terme général $x_{ij}=(\frac{f_{ij}}{f_i}\sqrt{f_j} - \sqrt{f_i})$ le résultat du chapitre sur l'ACP . Les facteurs principaux sont les vecteurs propres de MV la métrique est ici ND_2^{-1}, et la matrice V est égale au produit matriciel ${}^tXDX$ où (X est tableau du profils ainsi obtenus et D la matrice diagonale des poids $d_{ii}=\frac{n_i}{n}=f_i$).

Donc on a : T = MV = ${}^tXDX$.

$$t_{ik}=\sum_{i=1}^{q} f_i\left(\frac{f_{ij}}{f_i}\sqrt{f_j}-\sqrt{f_j}\right)\left(\frac{f_{kj}}{f_i}\sqrt{f_k}-\sqrt{f_k}\right)$$

il est possible de donner à cette matrice T une forme simple posons en effet :

$$x_{ij}^{*}=\left(\frac{f_{ij}-f_i.f_j}{\sqrt{f_i.f_j}}\right)$$

Alors la matrice T à diagonaliser s'exprime en fonction du tableau noté X^*. $T= {}^tX^*X^*$. On projetons le point i sur l'axe factorielle U_α en obtient les coordonnée de profils lignes .

$$c_i=\sum_{j=1}^{p}\left(\frac{f_{ij}-f_i.f_j}{\sqrt{f_i.f_j}}\right)U_{aj}$$

2.4 ACP de nuage de colonnes R^p :

Les ensembles mis en correspondance dans le tableau des fréquences jouent des rôles analogues. L'analyse dans R^q peut donc se déduire de celle menée dans R^p par permutation des rôle des indices i et j.

Ainsi les coordonnées du point j seront maintenant les quantités

$\frac{f_{ij}}{f_j}\sqrt{f_j}$ ce point j sera muni de la masse f_j . La i éme coordonnée du centre gravité H du nuage des p points s'écrit :$h_i = \sqrt{f_i}$

La matrice MV à diagonaliser aura de terme générale :

$$w_{ik}=\sum f_j\left(\frac{f_{ij}}{f_j\sqrt{f_i}}\sqrt{f_i}\right)\left(\frac{f_{ij}}{f_j\sqrt{f_k}}\sqrt{f_k}\right)$$

Donc la matrice MV est le produit matriciel de $W = {}^tX^*X^*$.En fin les coordonnées des profils lignes s'obtiennent par la formule suivante :

$$c_{ic}=\sum\left(\frac{f_{ij}}{f_i}\sqrt{f_i}-\sqrt{f_i}\right)V_{ik}^1$$

3-Relation entre les deux espace R et R^p

On montre qu'il y a une dualité entre les deux analyse et que les matrices T et W ont les mêmes valeurs propres. On s'aperçoit que l'ACP du nuage de profils des lignes est équivalente à l'ACP du nuage des profils des colonnes : les facteurs principales d'une analyse sont à $\sqrt{\lambda}$ prés les composantes principales de l'autre :

$$U_\alpha=\frac{1}{\sqrt{\lambda_\alpha}}X^*V_\alpha$$

$$V_\alpha=\frac{1}{\sqrt{\lambda_\alpha}}X^*U_\alpha$$

VII- ANALYSE CANONIQUE

1-Introduction

L'analyse canonique, proposée en 1936 par H. Hotelling, est d'un intérêt théorique essentiel. Elle englobe en effet la plupart des méthodes d'analyse des données comme cas particulier : qu'il s'agisse de la régression multiple, de l'analyse de la variance, de l'analyse des correspondances ou de l'analyse discriminante, ces méthodes peuvent être considérées comme des applications spécifiques de l'analyse canonique.

L'analyse canonique est davantage une curiosité mathématique qu'une méthode des données : les résultats qu'elle procure sont, en effet, difficilement interprétable.

L'analyse canonique joue donc dans l'analyse des données un rôle d'une sorte de plaque tournante théorique, à partir de laquelle on peut donner une présentation unifiée de la plupart des autre méthodes.

2-Le problème de l'analyse canonique

L'analyse canonique à pour but d'étudier les relations linéaires existant entre deux groupes de caractères observés sur le même ensemble d'individus. De façon plus précisé on cherche une combinaison linéaire des caractères de 1^{er} groupe et une combinaison linéaire du 2^{eme} groupe qui soient les plus corrélés possibles.

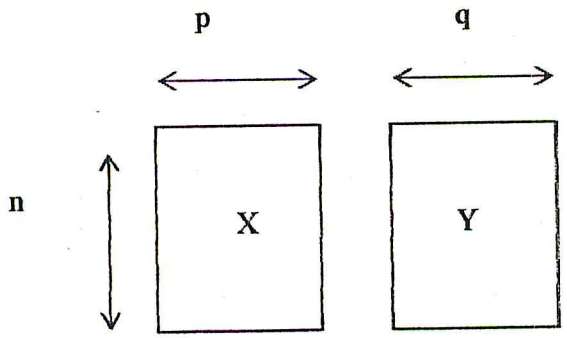
Considérons une population de n individus, et supposons que l'on ait observé sur chaque individu les valeurs de p variables $x_1, \dots, x_p$ et de q variables $y_1, \dots, y_q$.

Nous utiliserons pour repérer individus et variables les trois indices i, j et k ($i \in [1,n]$; $j \in [1,p]$; $k \in [1,q]$).

Nous notons x_{ij} (resp. y_{ik}) la valeur de x_j(resp. y_k) pour l'individu i.

- Le regroupement des variables en deux catégories distinctes se justifie dans chaque cas particulier. Par exemple, si l'on considère une population d'êtres humains, les x_j pourront désigner des variables de situation (revenu , age , etc .), les y_k des variables de comportement (emploi du temps, répartition des consommations, etc.) .
- Les variables x_i et y_j peuvent être quantitatives, ou bien repérer des modalités d'une variable qualitative(dans ce cas, $x_{ij} = 1$ si l'individu i possède la modalité repéré par la colonne j, $x_{ij} = 0$ sinon).
- On centre parfois les variables x_i et y_j avant de procéder à l'analyse (c'est à dire que l'on remplace, dans le tableau des x_{ij}, le nombre x_{ij} par $x_{ij} - \overline{x}_j$, avec $\overline{x}_j = \frac{1}{n}\sum_i x_{ij}$). Cette opération n'est nullement indispensable, et nous supposerons que les variables ne sont pas centrées, sauf mention du contraire.

  On peut représenter les données par le tableau suivant :



3- Formulation mathématique

Le tableau de données R à n ligne et p+q colonnes, est partitionné en deux sous-tableaux X et Y, ayant respectivement p et q colonnes.

$$R = [X \mid Y]$$

Les lignes représentent les individus ou observations ; les p premières colonnes sont les variables du premier groupe, et les q suivantes sont celles du second groupe.

Nous supposerons, sans perte de généralité, que les variables sont centrées, ce qui signifie que chaque colonne de R est telle que la somme de ses éléments vaut 0.alors la matrice des covariances expérimente les des p+q variables s'écrit :

$$V(R)=\frac{1}{n}R'R \qquad (V_{jj'}=\frac{1}{n}\sum n_j n_{j'})$$

soit : $$V(R)=\frac{1}{n}\begin{vmatrix} x'x & x'y \\ y'x & y'y \end{vmatrix}$$

considérons l'individu i, caractérisé par la i-éme ligne de R :

$(x_{i1}, x_{i2},\ldots..,x_{ip},y_{i1},y_{i2},\ldots..y_{iq})$

pour rapprocher ce eux ensembles de caractères x et y, on calcule une combinaison linéaire des caractère du premier groupe

$\xi = a_1x^1 + \ldots + a_jx^j + \ldots + a_px^p$

et une combinaison linéaire des caractère du deuxième groupe

$\eta = b_1 y^1 + \ldots + b_k y^k + \ldots + b_qy^q$

on cherchera les coefficients

${}^ta = (a_1, \ldots a_j , \ldots ,a_p)$

${}^tb = (b_1, \ldots ,b_k , \ldots,b_q)$

Qui maximisent le carré de corrélation entre ξ et η on appelle

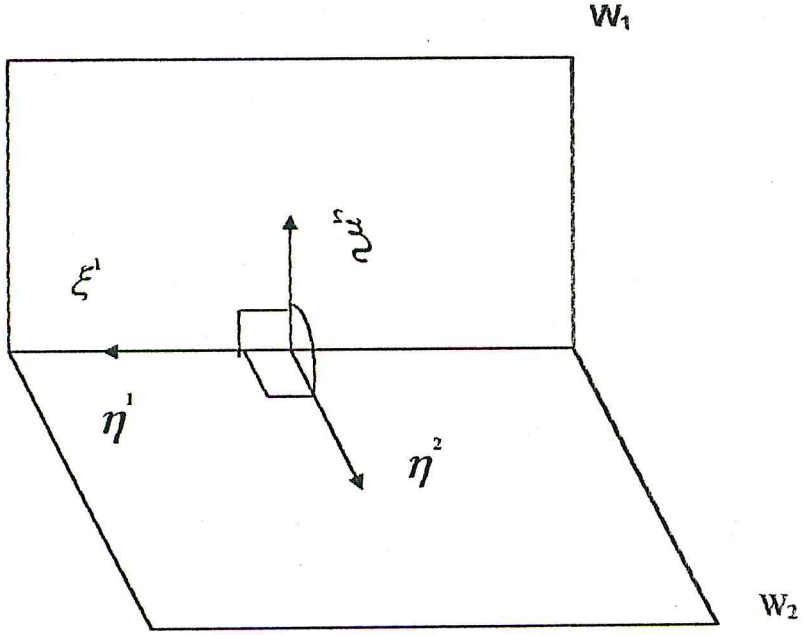
caractères canoniques les vecteurs $(\xi$ et $\eta) \in \mathbb{R}^n$

facteurs canoniques les vecteurs de coefficients a $\in R^p$ et b $\in R^q$ et **corrélation canonique** le coefficient de corrélation entre ξ et η .

L'ensemble des caractères ξ combinaisons linéaires des $x^1, \ldots, x^j, \ldots,x^p$, forme un sous espace vectoriel W_1 de R^n que l'on appelle « potentiel de prévision » du premier groupe.

De même au seconde groupe, on associe W_2, sous espace vectoriel de R^n.

Il s'agit donc de trouver deux vecteurs $\xi \in W_1$ et $\eta \in W_2$ faisant un angle minimum



Dans le schéma précédent, il existe une solution très simple ξ^1 et η^1 tels que $\cos^2(\eta^1, \xi^1) = 1$

En effet, dans R^3, l'intersection de deux plan est de dimension inférieure ou égale à 2.

Lorsqu'un premier couple de variable s canoniques a été obtenu, on recherche, dans un deuxième temps, un autre couple de caractères ξ^2 et η^2 tels que $r(\xi^2, \eta^2)$ soit maximum et tels que ξ^1 et η^1 (respectivement η^1 et η^2) aient une corrélation nulle et ainsi de suite, ξ^3 et η^3, etc...

4- Interprétation géométrique :

4.1-Projection orthogonale sur un sous-espace vectoriel

A) la régression multiple

Avant de résoudre le problème de l'analyse canonique il est nécessaire d'effectuer quelques rappels sur la régression multiple, et particulier sur la projection orthogonale d'un vecteur sur un sous espace vectoriel.

Considérons le cas d'un caractère « à expliquer » y et de p caractères « explicatifs » $x^1,\dots,x^j,\dots,x^p$.

Nous supposons que ces p+1 caractères sont observés sur le même ensemble de n individus, chaque individu étant muni du poids $p_i > 0$ avec $\sum p_i = 1$

Il s'agit de trouver une combinaisons linéaire des p caractères explicatifs

$$\xi = a_1x^1 + \dots + a_jx^j + \dots + a_px^p$$

Telle que ξ soit le plus proche possible de y au sens de la distance dans l'espace des caractères (critère des moindres carrés).

Nous allons maintenant présenter géométriquement problème de la régression multiple.

Chaque des p+1 caractères peut être représenté par un vecteur de R^n :

$$y=\begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_i \\ . \\ . \\ . \\ y_n \end{bmatrix} \in R^n \text{ et } \qquad X^j=\begin{bmatrix} x^j_1 \\ . \\ . \\ . \\ x^j_i \\ . \\ . \\ . \\ x^j_n \end{bmatrix} \in R^n \; j=1,\ldots,p$$

On suppose que ces p+1 caractères sont centrés :

$$\sum_{i=1}^{n} p_i y_i = 0 \qquad \sum_{i=1}^{n} p_i x_i{}^j = 0 \quad j=1,\ldots,p$$

Nous considérons le sous espace vectoriel W de R^n engendré par les combinaisons linéaires des caractères x^j :

$$\xi \in W \Leftrightarrow \quad \xi = a_1x^1 + \ldots + a_jx^j + \ldots + a_px^p$$

Nous supposons par la suite que la dimension de W est égale à p, ce qui revient à dire que les p caractères x^j forment une base de W, ou encore que le rang de la matrice :

G={$x^1,\ldots,x^p$} engendre W et G possède p élément } => G est une base de W.

$$X_{n,p} = \begin{bmatrix} x_1{}^1,\ldots, & x_1{}^j,\ldots, & x_1{}^p \\ . & & \\ . & & \\ . & & \\ x_i{}^1,\ldots, & x_i{}^j,\ldots, & x_i{}^p \\ . & & \\ . & & \\ . & & \\ x_n{}^1,\ldots, & x_n{}^j,\ldots, & x_n{}^p \end{bmatrix}$$

En notation abrégée, on pose :

$W = \{ \xi \in R^n / \xi = Xa, a \in R^p \}$

Comme dans l'analyse en composantes principales, nous supposons que l'espace des caractères est muni du produit scalaire associé à la matrice diagonale des poids :

$$D = \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ & & p_i \\ & \ddots & \\ 0 & & p_n \end{bmatrix}$$

Sur l'espace caractères centrés, on a vu que le produit scalaire et la covariance sont identiques :

${}^tX^j D X^k = s_{jk}$

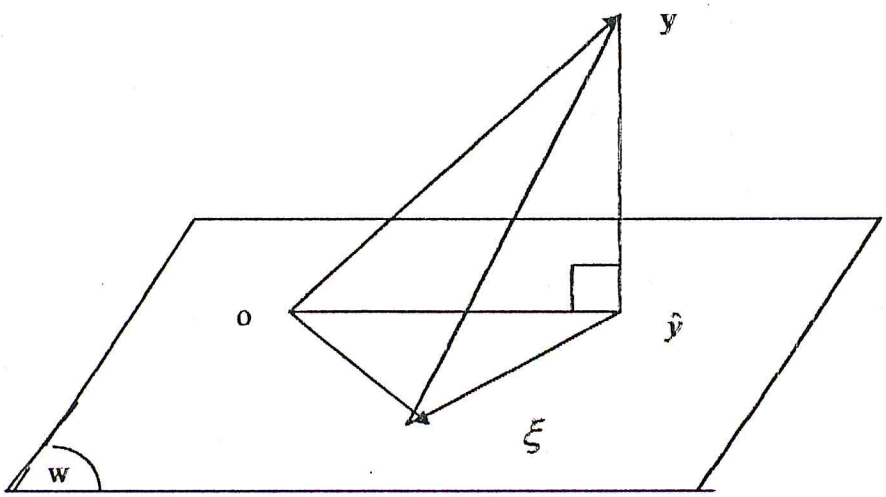
de même la norme et la variance :

$$\left\| x^j \right\|^2 = Sj^2$$

La distance entre deux caractères est donnée par

$$D^2(x^j,x^k) = \left\| x^j - x^k \right\|^2 = {}^t(x^j-x^k).D.(x^j-x^k)$$

Dans l'espace des caractères on peut schématiquement représenter $W \subset R^n$ et $y \in R^n$ par la figure suivante :



$y \in R^n$ est donné, on cherche $\xi \in W$ tel que la distance entre y et ξ soit minimum, le critère des moindres carrés peut s'écrite :

$$\min \left\| y - \xi \right\|^2 , \xi \in W ; \cos(\theta_{\xi,y}) = r(\xi, y)$$

Dans la suite nous noterons $\hat{y}$ le point de W le plus proche de y :

$\hat{y}$ est la projection orthogonale de y sur W

B) Recherche du projection orthogonal sur un sous espace vectoriel

Nous appelons projecteur orthogonal sur W l'application linéaire de R^n dans R^n faisant correspondre à tout vecteur de R^n sa projection orthogonale sur W.

Notons A la matrice de cette application :

$$R^{n} \rightarrow R^{n}$$

$$y \rightarrow Ay = \hat{y}$$

avec ${}^t(y - \hat{y})D\hat{y} = 0$ (orthogonalité)

Nous allons maintenant voir comment A peut construit à partir des vecteurs $x^1, \ldots, x^j, \ldots, x^p$ base de W.

Tout vecteur $\xi \in W$ peut s'écrire sous la forme :

ξ = Xa, en particulier $\hat{y} \in W$, pour lequel nous posons :

$\hat{y}$ =X$\hat{a}$.

y -$\hat{y}$ doit être orthogonal à tout vecteur de W, donc en particulier, aux vecteurs de base. On a par conséquence p équation :

${}^tX^j D (y - \hat{y}) = 0$; j=1,...,p

ou encore, puisque $\hat{y}$=X$\hat{a}$, j=1,...,p

Ces p équations s'écrivent sous la forme d'une seule équation matricielle :

$${}^tXD\,X\hat{a} = {}^tXDy$$

Le vecteur $\hat{a}$ contient donc les p coefficients de la combinaisons linéaire $\hat{y} = \hat{a}_1x^1 + ... + \hat{a}_jx^j + ... + \hat{a}_px^p \in W$

La plus proche de y.

De l'expression de $\hat{a}$, on déduit l'expression de $\hat{y}$=X$\hat{a}$:

$$\hat{y} = X({}^tXDX)^{-1}\,{}^tXDy$$

La matrice $X({}^tXDX)^{-1}\,{}^tXD$ fait donc correspondre à y sa projection orthogonale sur W.On déduit l'expression de A :

$$\boxed{A = X({}^tXDX)^{-1}\,{}^tXD}$$

4.2-Les caractères canoniques

A) présentation géométrique

Revenons maintenant au problème de l'analyse canonique. Nous disposons maintenant de deux ensembles de caractères x^1,... ,x^j,... ,x^p et y^1,..., y^k,..., y^q.

De même qu'en régression multiple, nous supposons que ces p + q caractères sont observés sur le même ensemble de n individus munis de poids $p_i > 0$, $i = 1,\ldots,n$

Avec $$\sum_{i=1}^{n} p_i = 1$$

Nous supposons également que les p + q caractères sont centré. Chacun des p + q caractères peut être représenté par un vecteur de R^n :

$$y^k = \begin{bmatrix} y^k_1 \\ \cdot \\ \cdot \\ \cdot \\ y^k_i \\ \cdot \\ \cdot \\ \cdot \\ y^k_n \end{bmatrix} \in R^n \quad k=1,\ldots,q \quad \text{et} \quad X^j = \begin{bmatrix} x^j_1 \\ \cdot \\ \cdot \\ \cdot \\ x^j_i \\ \cdot \\ \cdot \\ \cdot \\ x^j_n \end{bmatrix} \in R^n \quad j=1,\ldots,p$$

Aux vecteurs x^j et y^k nous associons respectivement les sous espaces vectoriels de R^n W_1 et W_2

$$W_1 = \{ \xi \in R^n / \ \xi = Xa, a \in R^p \}$$

$$W_2 = \{ \eta \in R^n /, b \in R^q \}$$

Les vecteurs X^j (et Y^k) étant centrés, les sous espaces vectoriels W_1 (et W_2) contiennent de vecteurs centrés.

L' encore, nous supposons que les X^j (les Y^k) forment une base de W_1 (de W_2) et donc que :

Dim (W_1) = p et dim (W_2) = q

Rang(X) = p et rang(Y) = q

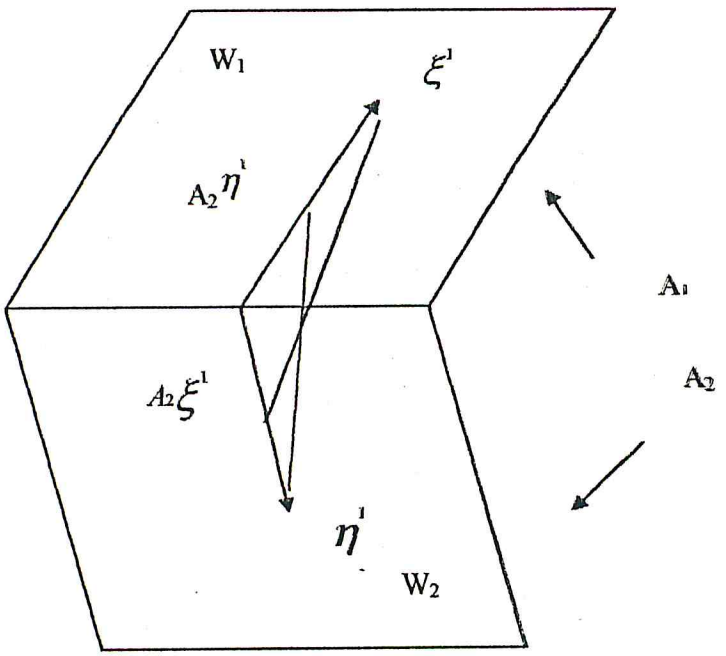
Géométriquement, le problème de l'analyse canonique peut être formulé de la façon suivante :

Il s'agit de trouver $\xi \in W_1$ et $\eta \in W_2$ tel que :

$Cos^2(\xi, \eta) = r^2(\xi, \eta)$ soit maximum.

B) Recherche des caractères canoniques

Supposons que les caractères ξ^1 et η^1 soient solution du problème.



Puisque l'angle entre ξ et η ne dépend pas se leur norme, on suppose que $\|\xi\| = \|\eta\| = 1$

η^1 doit être colinéaire avec la projection orthogonale de ξ^1 sur W_2 faisant un angle minimum avec ξ^1.

Cette condition s'écrit :

$$A_2\xi^1 = r_1\eta^1$$

où $r_1 = \cos(\xi^1, \eta^1)$ et où A_2 est l'opérateur de projection orthogonale sur W_2 on a de même :

$$A_1\eta^1 = r_1\xi^1$$

On en déduit de ces deux équations les systèmes :

$$\boxed{\begin{aligned} A_1A_2\xi^1 &= \lambda_1\xi \\ A_2A_1\eta^1 &= \lambda_1\eta^1 \end{aligned}}$$

Où $\lambda_1 = \cos^2(\xi^1, \eta^1)$ on en déduit que ξ^1 et η^1 sont respectivement les vecteurs propres de A_1A_2 et A_2A_1 associes a la même plus grande valeur propre λ_1, égale à leur cosinus carré (à leur corrélation carrée).

Les caractères ξ^1 et η^1 se déduisent l'un de l'autre par une simple application linéaire:

$$\boxed{\begin{aligned} \eta^1 &= \frac{1}{\sqrt{\lambda_1}}A_2\xi^1 \\ \xi^1 &= \frac{1}{\sqrt{\lambda_1}}A_1\eta^1 \end{aligned}}$$

Les caractères canoniques suivants sont les vecteurs propres de A_1A_2 (resp A_2A_1) associes aux valeurs propres rangées en ordre décroissant. On peut en effet montrer que les vecteurs propres de A_1A_2 sont orthogonaux pour D et que, par conséquent, $\cos^2(\xi^i, \xi^j) = \cos^2(\eta^i, \eta^j) = 0$ lorsque $i \neq j$. A chaque étape, on choisit le couple des caractères canonique ξ^i, η^i associé à la plus grande valeur propre λ_i non encore sélectionnée.

On remarque que le nombre maximum de caractères canoniques est égal à min (p,q). En effet, en supposant que p<q les ξ^i i=1,...,p forment une base de W_1 et il n'est pas possible d'obtenir d'autres vecteurs appartenant à W_1 et orthogonaux aux ξ^i.

c) Recherche des facteurs canoniques

Nous avons vu que, puisque $\xi \in W_1$, ξ peut s'écrire comme une combinaisons linéaire des caractères $x^1,...,x^p$:

$$\xi = a_1x^1 + ... + a_jx^j + ... + a_px^p$$

Ou encore, en posant

$${}^ta = (a_1,...,a_p)$$

De même

$$\boxed{\begin{aligned}\xi &= Xa\\ \eta &= Yb\end{aligned}}$$

Les facteurs canoniques a et b peuvent être calculés directement.

En posant :

$$A_1 = X({}^txDX)^{-1}\,{}^tXD$$
$$A_2 = Y({}^tyDY)^{1-}\,{}^tYD$$

Et en remplaçons dans les équations donnant ξ et η il vient :

$$X({}^txDX)^{-1}\,{}^tXD\;Y({}^tyDY)^{1-}\,{}^tYDXa = \lambda Xa$$
$$Y({}^tyDY)^{1-}\,{}^tYD\;X({}^txDX)^{-1}\,{}^tXD\;Yb = \lambda Yb$$

Posons :

$$V_{11} = {}^tXDX$$
$$V_{22} = {}^tYDY$$
$$V_{12} = {}^tXDY = {}^tV_{21}$$

Nous avons déjà vu que V_{11} été identique à la matrice de variance –covariance des caractères ^jX, de même V_{22} est la matrice de variance –covariance de y^k. En fin V_{12} contient les covariance entre les X^j et les y^k.

Les équations précédentes se simplifient :

$$X V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}\, a = \lambda Xa$$

$$Y\, V_{22}^{-1}\, V_{21}\, V_{11}^{-1}\, V_{12} b = \lambda Y b$$

Puisque les matrices X et Y sont respectivement de rang p et q, on peut simplifier les équations précédentes qui deviennent :

$$V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}\, a = \lambda a$$

$$V_{22}^{-1}\, V_{21}\, V_{11}^{-1}\, V_{12} b = \lambda b$$

Nous avons ainsi une manière de calculer les facteurs canoniques comme vecteurs propres de produits de matrices de covariance.

Les conditions de normalisation $\|\eta^2\| = \|\xi^2\| = 1$ deviennent :

$\xi' D \xi = {}^t a\, {}^t X D X a = {}^t a\, V_{11} a = 1$

$\eta' D \eta = {}^t b\, {}^t Y D Y b = {}^t b V_{22} b = 1$

Enfin a et b se déduisent l'un de l'autre par transformation linéaire :

$$\eta = \frac{1}{\sqrt{\lambda}} v_{22}^{-1} V_{21} a$$

devient :

$$Yb = \frac{1}{\sqrt{\lambda}} Y ({}^t Y D Y)^{-1}\, {}^t Y D X a$$

Et en simplifiant :

$$b = \frac{1}{\sqrt{\lambda}} v_{22}^{-1} V_{21} a \quad \frac{1}{\sqrt{\lambda}} v_{22}^{-1} V_{21} a$$

$$a = \frac{1}{\sqrt{\lambda}} v_{11}^{-1} V_{12} b \quad \frac{1}{\sqrt{\lambda}} v_{11}^{-1} V_{12} b$$

On cherchera d'abord a si p<q pour travailler sur la matrice de plus faible taille, et on en déduira ensuite b.

VIII- Implémentation

Ce présent chapitre est consacré à l'implémentation informatique de la méthode de l'AC et à pour but l'élaboration d'un logiciel en langage c++ Builder. Nous avons testé les programmes sur prés de vingt tableaux des données prélevés dans la littérature.

Notre problème se ramène à la recherche des valeurs et vecteurs propres d'une matrice quelconque. Il existe plusieurs méthodes de calcul de ces valeurs et vecteurs propres. Nous allons exposer ici un algorithme assez général emprunté à Rutishauser qui permet de trouver les valeurs propres d'une matrices quelconque.

1. Méthode de RUTISHAUSER

Une matrice carrée A quelconque d'ordre n peut, en général, être décomposé en un produit de deux matrices triangulaires :

A=M*S

Si l'une des matrices M ou S possède des 1 sur sa diagonale principale, la décomposition est unique.

Etant donné une matrice A d'ordre 4 et a_{ij} posons :

$$M = \begin{bmatrix} m_{11} & 0 & 0 & 0 \\ m_{21} & m_{22} & 0 & 0 \\ m_{31} & m_{31} & m_{33} & 0 \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \quad S = \begin{bmatrix} 1 & s_{12} & s_{13} & s_{14} \\ 0 & 1 & s_{23} & s_{24} \\ 0 & 0 & 1 & s_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

En effectuant le produit et en identifiant, on obtient les relations :

$s_{11}=1$
$m_{11}=a_{11}$
$s_{12}=a_{12}/m_{11}$
$m_{32}*s_{23}$
$m_{21}=a_{21}$
$s_{13}=a_{13}/m_{11}$
$m_{31}=a_{31}$
$m_{13}=0$
$m_{23}=0$
$m_{33}=a_{33}-m_{31}*s_{13}-$
$m_{43}=a_{43}-m_{41}*s_{13}-m_{42}*s_{23}$
$s_{31}=0$
$s_{32}=0$

$s_{14}=a_{14}/m_{11}$ $s_{33}=(a_{33}-m_{31}*s_{13}-m_{32}*s_{23})/m_{33}$

$m_{41}=a_{41}$ $s_{34}=(a_{34}-m_{31}*s_{14}-m_{32}*s_{24})/m_{33}$

$m_{12}=0$ $m_{14}=0$

$m_{22}=a_{22}-m_{21}*s_{12}$ $m_{24}=0$

$m_{32} = a_{32}-m_{31}*s_{12}$ $m_{34}=0$

$m_{42} = a_{42}-m_{41}*s_{12}$ $m_{44}= a_{44}-m_{41}*s_{14}-m_{42}*s_{24}-m_{43}*s_{34}$

$s_{21}=0$ $s_{41}=0$

$s_{22}=(a_{22}-m_{21}*s_{12})/m_{22}$ $s_{42}=0$

$s_{23}=(a_{23}-m_{21}*s_{13})/m_{22}$ $s_{43}=0$

$s_{24}=(a_{24}-m_{21}*s_{14})/m_{22}$ $s_{44}=(a_{44}-m_{41}*s_{14}-m_{42}*s_{24}-m_{43}*s_{34})/m_{44}$

On constate qu'il faut calculer pas à pas, car on utilise toujours les valeurs précédentes.

L'expression générale de la matrice M s'écrit :

$m_{ji} = a_{ji}$-

Tandis que celui de la matrice S s'écrit :

S_{ij} = 'reste)

On décompose la matrice A en un produit de deux matrices triangulaires

$A = M_1*S_1$

M_1 est une matrice triangulaire inférieure et S_1 une matrice triangulaire supérieure.

On construit une matrice B_1 à partir de M_1 et S_1 de la manière suivante :

$B_1 = S_1* M_1$

On opère sur B_1 comme sur A, c'est-à-dire qu'on la décompose également en un produit de matrices triangulaire, soit :

$B_1 = M_2*S_2$

Puis, on permute à nouveau les deux matrices obtenues formant ainsi une nouvelle matrice B_2 :

$B_2 = S_2 * M_2$

Et on continue ainsi le processus. A la $i^{éme}$ application de l'algorithme, on a donc une matrice :

$B_i = S_i * M_i$

Qui peut s'exprimer en fonction des matrices B précédentes, on peut en effet écrire :

$B_i * S_i = S_i * M_i * S_i = S_i * B_{i-1}$

Ou :

$$B_i * S_i * S_i^{-1} = B_i = S_i * B_{i-1} * S_i^{-1}$$

Les valeurs propres de B_i sont telles que le déterminant : $|B_i - \lambda * I|$ soit nul, où I est la matrice unité de même ordre que la matrice A.

Si on remplace B_i par $S_i * B_{i-1} * S_i^{-1}$ dans l'expression matricielle $B_i - \lambda * I$, il vient :

$B_i - \lambda * I = S_i * B_{i-1} * S_i^{-1} - S_i * \lambda * I * S_i^{-1} = S_i * (B_{i-1} - \lambda * I) * S_i^{-1}$

Et en revenant aux déterminants :

$$|B_i - \lambda * I| = |S_i| * |B_{i-1} - \lambda * I| * \frac{1}{|s_i|} = |B_{i-1} - \lambda * I|$$

B_i et B_{i-1} ont donc les mêmes valeurs propres et en raisonnant par récurrence, il en est de même de B_i et A.

On peut montrer que la matrice B tend vers une matrice triangulaire inférieure quand i tend vers l'infini.

Le déterminant d'une matrice triangulaire étant égal au produit des éléments de sa diagonale principale, les valeurs propres de B,qui aussi celles de a, se trouvent sur la diagonale principale de B.

On calcule donc les b par récurrence jusqu'à ce que les termes diagonaux ne soient pas modifiés par l'itération, c' est –à- dire que l'on se donne une précision ε et que l'on s'arrête lorsque tous les termes diagonaux successifs ne diffèrent pas de plus de ε.

Détermination des vecteurs propres :

Connaissant la valeur propre λ_k da la matrice A, le calcul du vecteur propre X^k qui lui correspond est représenté par la solution du système d'équation suivant :

$$\begin{bmatrix} a_{11}-\lambda_k & a_{12} & . & a_{1,n} \\ a_{2,1} & a_{2,2}-\lambda_k & . & a_{2,n} \\ . & . & . & . \\ a_{n,1} & a_{n,2} & . & a_{n,n}-\lambda_k \end{bmatrix} \begin{bmatrix} x_1^k \\ x_2^k \\ . \\ x_n^k \end{bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ . \\ 0 \end{Bmatrix}$$

Ce système ne possède pas de solution exacte en dehors du vecteur nul (lequel ne pas considéré comme vecteur propre).Pour obtient une solution approchée, fixons arbitrairement la valeur d'une des composantes. Cela n'est possible que si cette composante réellement non nulle x_i (dans notre cas la dernière composante x_n est donnée égale à 1) et la dernière colonne de

$$\begin{bmatrix} a_{11}-\lambda_k & a_{12} & . & a_{1,n} \\ a_{2,1} & a_{2,2}-\lambda_k & . & a_{2,n} \\ . & . & . & . \\ a_{n,1} & a_{n,2} & . & a_{n,n}-\lambda_k \end{bmatrix} \begin{bmatrix} x_1^k \\ x_2^k \\ . \\ 1 \end{bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ . \\ 0 \end{Bmatrix}$$

La matrice (A-λ_k*I) devient seconde membre d'un système B*X=Y

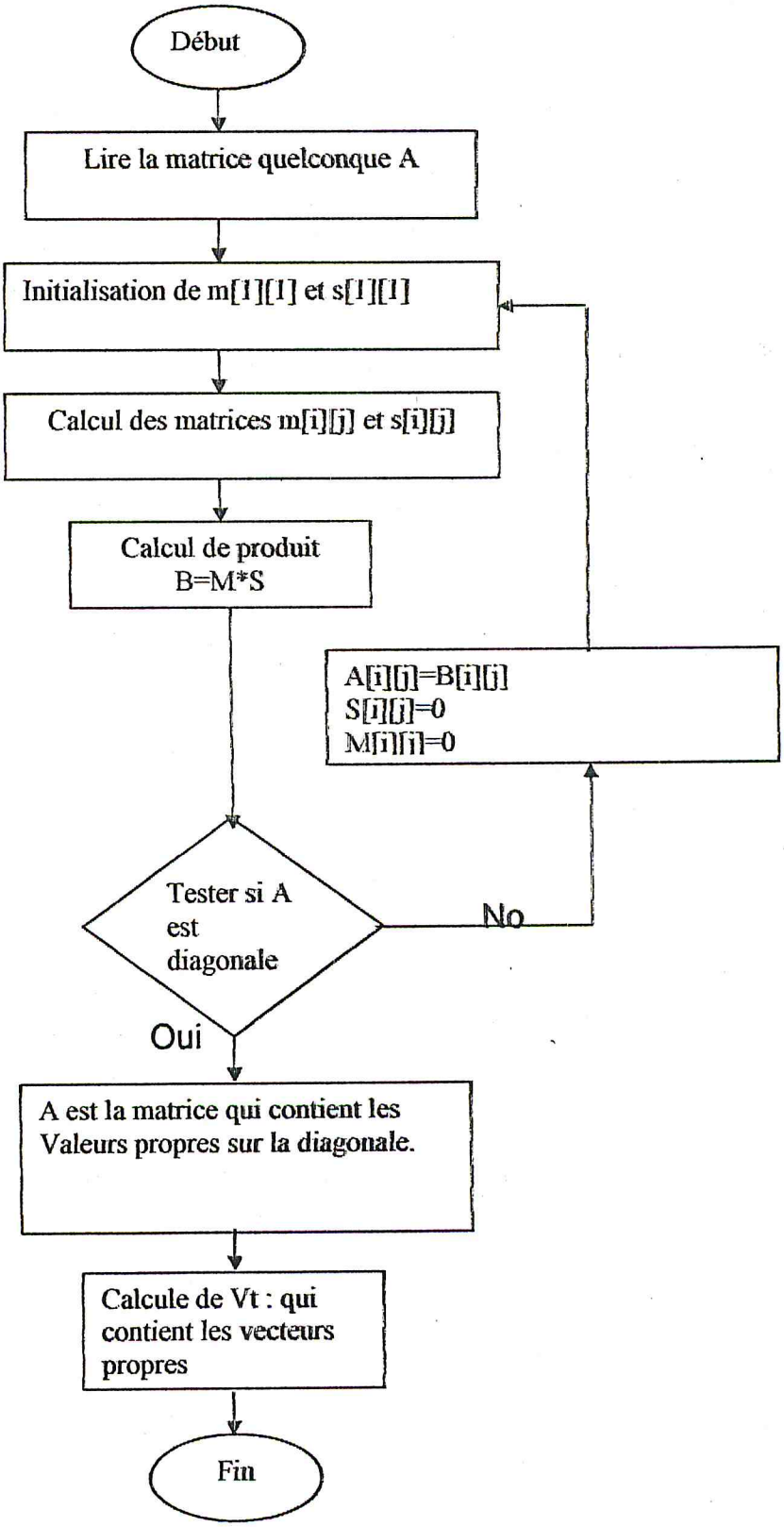
Surdéterminé de n équations à n-1 inconnues.

$$\begin{bmatrix} a_{11}-\lambda_k & a_{12} & . & a_{1,n-1} \\ a_{2,1} & a_{2,2}-\lambda_k & . & a_{2,n-1} \\ . & . & . & . \\ a_{n,1} & a_{n,2} & . & a_{n,n-1} \end{bmatrix} \begin{bmatrix} x_1^k \\ x_2^k \\ . \\ x_{n-1}^k \end{bmatrix} = \begin{Bmatrix} -a_{1,n} \\ -a_{2,n} \\ . \\ -a_{n,n}-\lambda_k \end{Bmatrix}$$

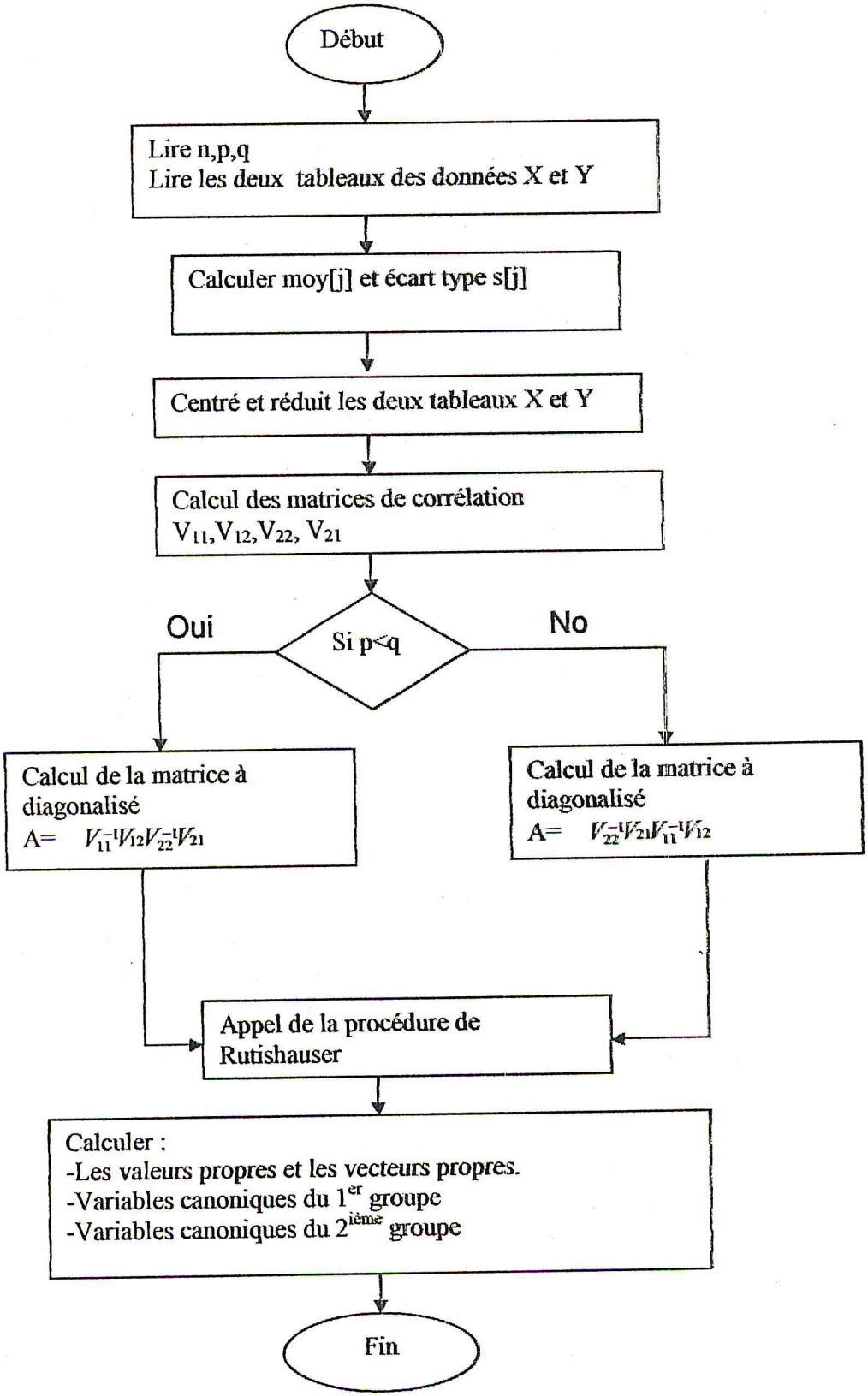
B X Y

Ce nouveau système est résolu avec l'une des méthodes de résolution des systèmes d'équations linaires (JORDAN).

Organigramme de la méthode de RUTISHAUSER



Organigramme de l'Analyse Canonique (AC)



Utilisation et Description du logiciel

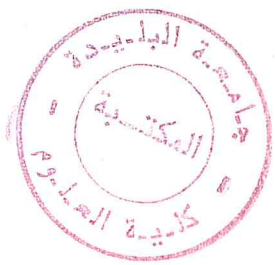
Dans cette partie nous allons décrire la conception puis l'utilisation du logiciel que nous avons établi.

1. Choix du langage de programmation

Nous avons utilisé le C++Builder version 5, comme langage de programmation pour la mise en oeuvre de la méthode **d'Analyse canonique.**

C++Builder est un environnement de programmation visuel orienté objet basé sur le langage C++. Nous pouvons l'utiliser pour développer toute sorte d'application. Il est possible de créer des applications Microsoft Windows 98 et Windows NT très efficace minimum de codage manuel. Il fournit aussi une bibiliothèque complète de composants réutilisables (Boutons, Edits, Boites de dialogue,Menu,) et un ensemble d'outils de conceptions. Ces outils simplifient le développement d'application et réduisent la durée du développement.

Lorsque on démarre C++Builder, on est immédiatement positionné dans l'environnement de programmation visuelle. C'est dans cet environnement que C++Builder fournit tous les outils dont on a besoin pour concevoir, développer,et tester les applications.



2- Description du logiciel

La fenêtre principale :

On a :

- **Le bouton Quitter :**

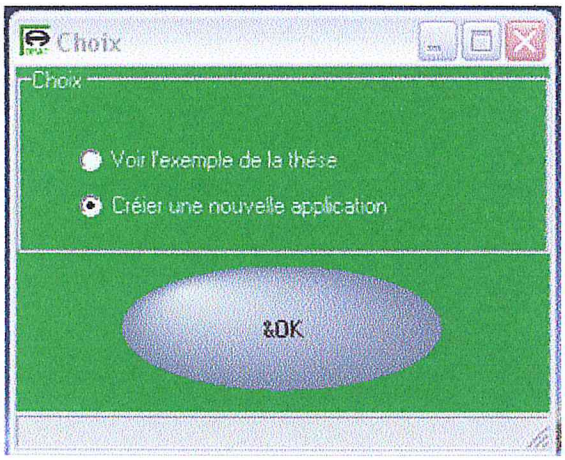
Il permet de quitter l'application

- **Le bouton d'Aide :**

Il comporte un seul champ qui permet de donner la définition et l'explication sur la méthode étudiée dans notre mémoire de fin d'étude.

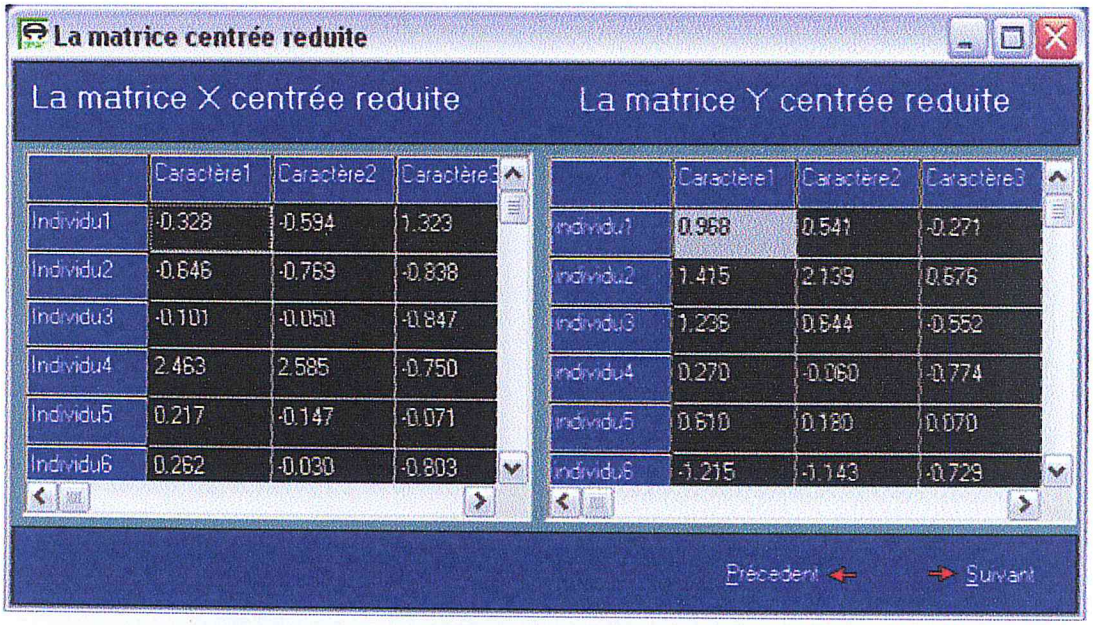
- **Le bouton Analyse canonique :**

Il permet de démarrer l'application



Dans laquelle on fait le choix d'une nouvelle application ou de voir l'exemple de la thèse, après la validité du choix une fenêtre apparaît.

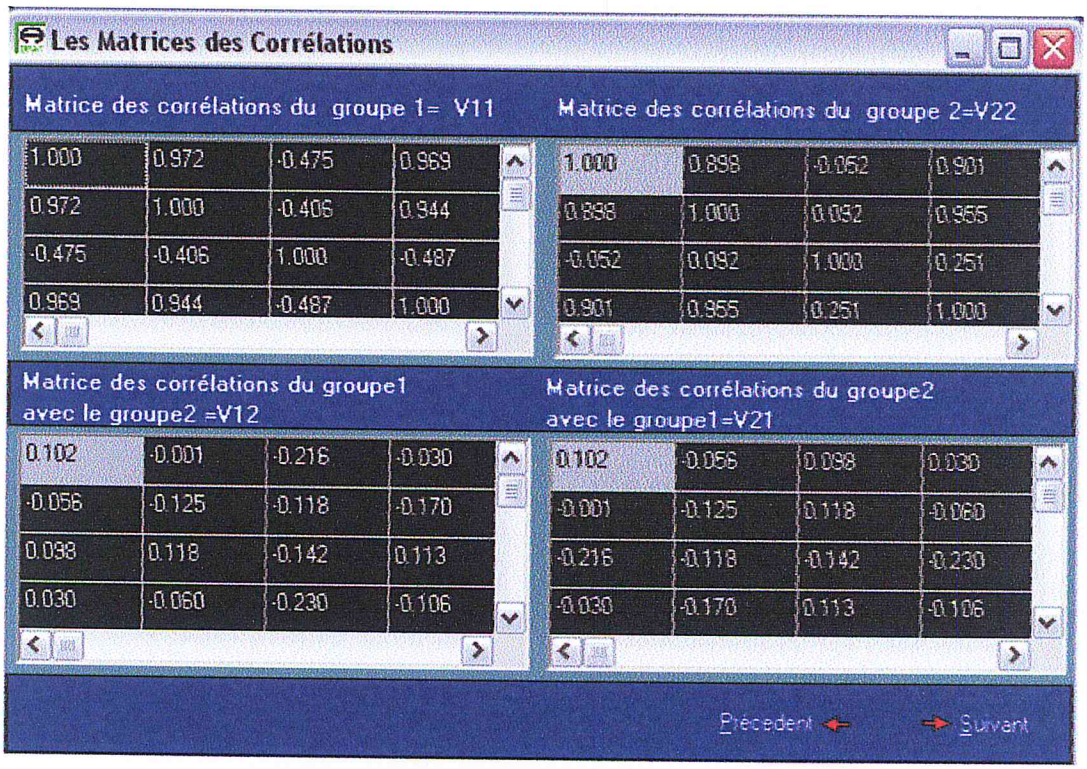
En cliquant sur le bouton **suivant** deux tableaux apparaissant qui contiennent les moyens et les écart-types pour chaque matrice, on clique sur le bouton **suivant** une fenêtre apparaît
Qui contient les matrices centrées réduites.



	Caractère1	Caractère2	Caractère3
Individu1	-0.328	-0.594	1.323
Individu2	-0.646	-0.769	-0.838
Individu3	-0.101	-0.050	-0.847
Individu4	2.463	2.585	-0.750
Individu5	0.217	-0.147	-0.071
Individu6	0.262	-0.030	-0.803

	Caractère1	Caractère2	Caractère3
Individu1	0.968	0.541	-0.271
Individu2	1.415	2.139	0.676
Individu3	1.236	0.644	-0.552
Individu4	0.270	-0.060	-0.774
Individu5	0.610	0.180	0.070
Individu6	-1.215	-1.143	-0.729

En cliquant sur le bouton **suivant** une fenêtre apparaît Qui contient les matrices corrélations.



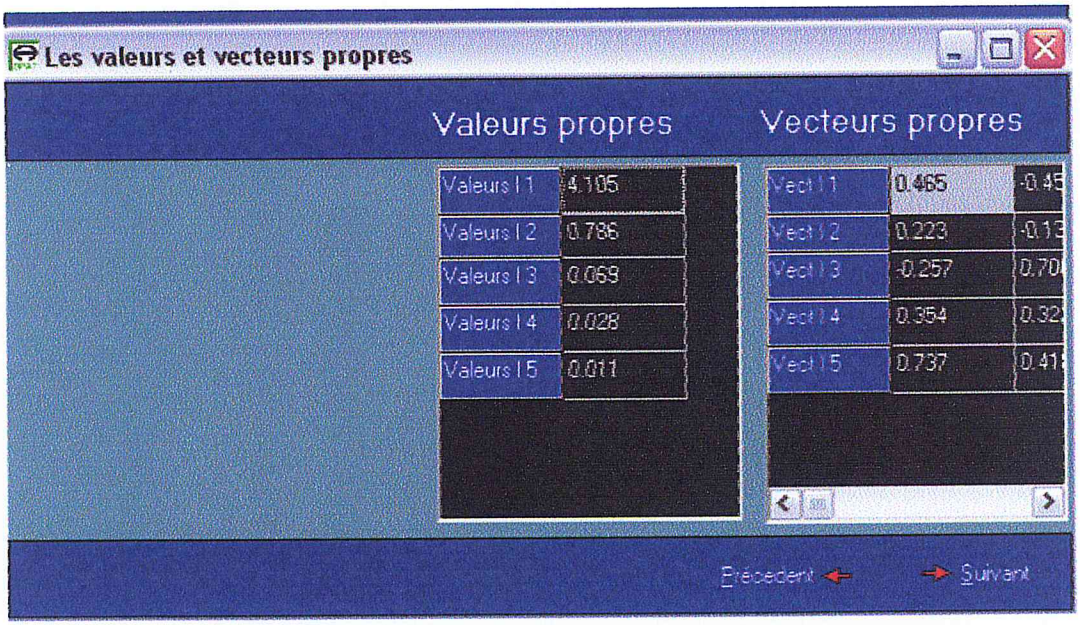
1.000	0.972	-0.475	0.969
0.972	1.000	-0.406	0.944
-0.475	-0.406	1.000	-0.487
0.969	0.944	-0.487	1.000

1.000	0.898	-0.052	0.901
0.898	1.000	0.092	0.955
-0.052	0.092	1.000	0.251
0.901	0.955	0.251	1.000

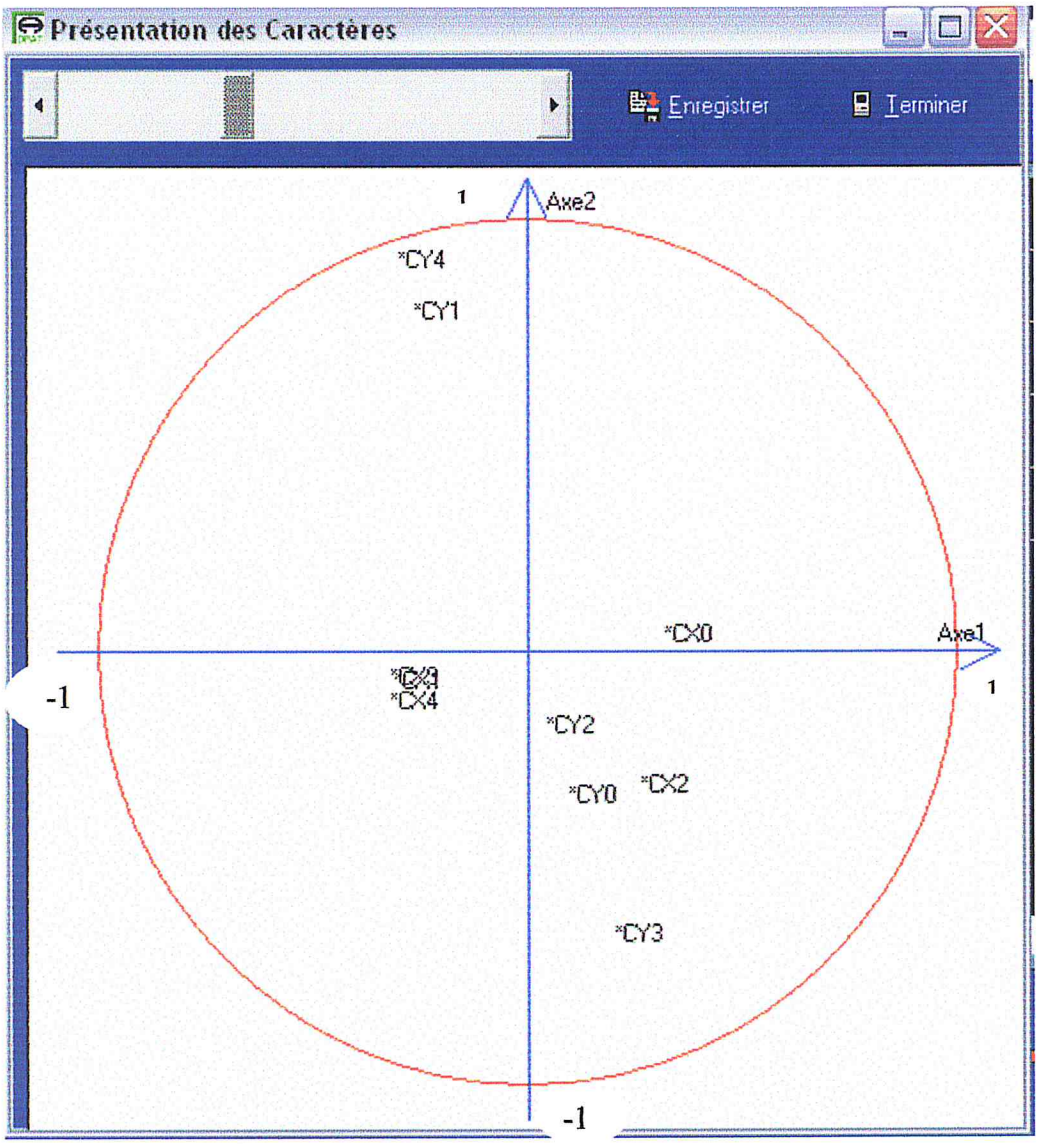
0.102	-0.001	-0.216	-0.030
-0.056	-0.125	-0.118	-0.170
0.098	0.118	-0.142	0.113
0.030	-0.060	-0.230	-0.106

0.102	-0.056	0.098	0.030
-0.001	-0.125	0.118	-0.060
-0.216	-0.118	-0.142	-0.230
-0.030	-0.170	0.113	-0.106

En cliquant sur le bouton **suivant** un tableau apparaît qui contient la matrice à diagonaliser, on clique sur le bouton **suivant** une fenêtre apparaît qui contient les valeurs et les vecteurs propres.



En cliquant sur le bouton **suivant** une fenêtre apparaît
Qui contient les caractères canoniques. Et en cliquant sur le bouton **suivant** une fenêtre apparaît
Qui contient le cercle de corrélation dans laquelle on la présentation des caractères qui permet d'avoir la relation existent entre ces caractères.



Avec deux boutons " **Enregistrer** " qui permet d'enregistrer le graphe, et une bouton " **Terminer** " qui permet de quitter l'application .

Conclusion

Cette étude nous a permis de découvrir l'analyse de données et de programmer l'une de ces méthodes, en l'occurrence **l'analyse canonique**. Que nous avons utilisé dans le cadre de la lutte contre les accidents du travail.

Les résultats que nous avons obtenus peuvent être retenus et utilisés pour la dite prévention.

Le programme d'analyse canonique écrit en langage C++Builder constitue également des outils efficaces pour le traitement d'exemples concrets.

Nous souhaitons enfin que notre travail a contribué à sensibiliser les spécialistes de la prévention sur l'utilisation des moyens scientifiques tels que l'analyse de données pour la prévention des accidents du travail.

Bibliographie

[1] P.Pelletier ; techniques numériques appliquées au calcul scientifique, Masson et Cie.

[2] Ronald CEHESSAT; Exercices commentés de statistique et informatique Appliquées, Dunod.

[3] BRIJITTE ESCOFIER (1997) ; Analyse Factorielle simples et multiples. Objectifs Méthodes et Interprétation.

[4] JEAN DE LAGARDE (1983) ; Initiation à l'analyse des données, Dunod

[5] P.Pascaux. R.Theodere (1998) ; Analyse Numérique. Matricielle Appliquée à l'Art de l'Ingénieur. (T2).

[6] BDEMIDVITCH, I.MARON ; Elément de Calcul Numérique, Dunod.

[7] L.LEBART, A. MORINEAU, J.P Fenelon ; Traitement des données Statistiques.

[8]. M.SYBONG .J.C MARDON (1988) ; Analyse Numérique (T1)

[9] L.LEBART, A.MORINEAU, NTABARD ; Technique et la description Statistique.

[10] VOLLE Michel. Analyse des données. Economica

[11] CH. BASTIN,J-P,BENZECRI , BOURGARIT, P.CAZES ; pratique de l'analyse des données abrégé théorique études de cas modèle Dunod.

[12] GILLE Jaen –Charles, Clique Marc ; Calcul matriciel et introduction à l'analyse fonctionnelle pour ingénieurs, Eyrolles

[13] G.MARCHOUK ; méthode de calcul numérique, Edition de moscou