RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI
TIZI-OUZOU

FACULTÉ DES SCIENCES

MÉMOIRE

Présenté par

M^{elle}. **TOUAT Dyhia**

Pour l'obtention du diplôme de

Master en Mathématiques

Option : Probabilités-Statistique.

Sujet

Analyse discriminante

Soutenu publiquement le 01 Octobre 2019
Devant le jury composé de

M^r. Youcef	BERKOUN	Professeur	UMMTO	Président
M^r. Mohamed	MAMOU	MAA	UMMTO	Encadreur
M^{me}. Karima	BOUALAM	MCB	UMMTO	Examinateur

Remerciements

En premier lieu, je remercie Dieu tout puissant de m'avoir accordée la puissance et la volonté pour achever ce travail.

Je tiens à exprimer toute ma reconnaissance à mon directeur de mémoire, Monsieur MAMOU Mohamed. Je le remercie de m'avoir encadrée, orientée, aidée et conseillée.

Je tiens à remercier très vivement Madame KROUCHI qui a accepté de me fournir un jeu de données très intéressant ainsi que pour son aide et sa contribution à l'aboutissement de ce projet.

J'adresse mes sincères remerciements à toute l'équipe pédagogique du département mathématiques, spécialité probabilités-statistique. Je souhaite particulièrement remercier Monsieur FELLAG Hocine pour son aide dans l'utilisation du logiciel Latex.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à mon mémoire en acceptant d'examiner mon travail.

Je désire aussi remercier mes très chers parents, Belkacem et Zahia, qui ont toujours été là pour moi. Je remercie ma sœur Fatma et mes fréres Mohand Said et Hakim pour leurs encouragements.

Enfin je remercie mes amis Sabrina, Fatima, Massiva, Lynda, Mohamed Samir, Hassane et Asma qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements m'ont été d'une grande aide.

Dédicaces

Je dédie cette thèse à :

A ma très chère mère Zahia qui m'a comblée avec sa tendresse et affection tout au long de mon parcours. Elle n'a cessé de me soutenir et de m'encourager durant toutes les années de mes études, elle a toujours été présente à mes cotés pour me consoler quand il fallait. Qu'elle trouve ici le témoignage de ma profonde reconnaissance.

A mon très cher père Belkacem : Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être. Ta patience, ta compréhension et ton encouragement sont pour moi le soutien indispensable que tu as toujours su m'apporter. Que Dieu le tout puissant te préserve, t'accorde santé, bonheur et te protège de tout mal.

A ma très chère sœur Fatma : En témoignage de l'attachement, de l'amour et de l'affection que je porte pour toi. Je te dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

Mon cher frère Mohand Said qui m'est le meilleur ami, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour toi. Je te souhaite un avenir plein de joie, de bonheur et de réussite.

A mon cher petit frère Hakim pour toute l'ambiance dont tu m'as entouré, pour toute la spontanéité et ton élan chaleureux. Je te dédie ce travail. Puisse Dieu le tout puissant exhausser tous tes vœux.

A mon frère Ahsen et mes cousines : Sabrina, Sonia, les deux Sylia et Belinda. En témoignage de l'attachement, de l'amour et de l'affection que je porte pour vous. Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A la mémoire de mes grands parents maternels qui ont été toujours dans mon esprit

et dans mon cœur, je vous dédie aujourd'hui ma réussite. Que Dieu, le miséricordieux, vous accueille dans son éternel paradis.

A mes amis : Sabrina, Fatima, Massiva, Lynda, Mohamed Samir, Hassane et Asma. En témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passé ensemble, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur.

A tous mes enseignants : J'ai eu le privilège de travailler parmi votre équipe et d'apprécier vos qualités et vos valeurs. Votre sérieux, votre compétence et votre sens du devoir m'ont énormément marqués. Veuillez trouver ici l'expression de ma respectueuse considération et ma profonde admiration pour toutes vos qualités scientifiques et humaines. Je vous dédie ce travail avec tous mes vœux de bonheur.

Table des matières

Table des matières **4**

Introduction générale **9**

1 Analyse des données **10**

1.1 Introduction 10

1.2 Un peu d'histoire 10

1.3 Objectifs 12

1.4 Domaines d'application 12

1.5 Analyse en Composantes Principales (ACP) 13

1.5.1 Tableau de données et espaces associés 14

1.5.1.1 Le tableau de données 14

1.5.2 Caractéristiques de la variable X^j 15

1.5.3 Poids et centre de gravité 16

1.5.4 La transformation des données 16

1.5.5 Matrice de variance-covariance et matrice de corrélation . . . 17

1.5.6 L'analyse des nuages 18

1.5.6.1 Analyse du nuage des points individus dans $\mathbb{R}^p$ . . . 18

1.5.6.2 Analyse du nuage des points variables dans $\mathbb{R}^n$ . . . 19

1.5.6.3 Inertie totale du nuage des individus 19

1.5.7 L'ajustement 20

1.5.7.1 Ajustement du nuage des individus 20

1.5.7.2 Ajustement du nuage des variables 27

1.5.8 Individus et variables supplémentaires 31
1.5.9 ACP et variables qualitatives 32
1.5.10 Interprétation 32
1.5.11 Limites de l'ACP 33
1.6 Analyse Factorielle des Correspondances (AFC) 34
1.6.1 Tableau de contingence, marges et profils 34
1.6.2 Liaison entre deux variables qualitatives 36
1.6.3 Métriques du Chi2 38
1.6.4 Analyse en composantes principales des deux nuages des profils 40
1.6.4.1 Interprétation des valeurs propres 41
1.6.4.2 Interprétation des plans de projection en AFC 42
1.6.4.3 Représentation simultanée (Biplot) 42
1.6.4.4 Eléments supplémentaires 44
1.7 Analyse des Correspondances Multiples (ACM) 44
1.7.1 Tableau de départ (Tableau de codage condensé) 45
1.7.2 Tableau disjonctif complet 45
1.7.3 L'AFC du tableau disjonctif complet 47
1.7.3.1 Inertie totale 49
1.7.3.2 ACP des nuages des points lignes et des points colonnes 50
1.7.3.3 Formules de transition et relations barycentriques . . 51
1.7.3.4 Biplot 51
1.7.3.5 Interprétation 53
1.7.3.6 Aides à l'interprétation 53

2 Analyse discriminante 55

2.1 Introduction 55
2.2 Données et notations 55
2.3 L'AD descriptive et l'AD prédictive 58
2.4 Principe de l'analyse discriminante 59
2.4.1 La discrimination 59

2.4.1.1 Formule de décomposition de Huygens 59
2.4.1.2 Analyse factorielle discriminante 59
2.4.1.3 Cas particulier de deux groupes - Equivalence entre régression multiple et analyse discriminante 67
2.4.2 Règle d'affectation d'un nouvel individu à l'un des K groupes . 71
2.4.2.1 Règles géométriques 71
2.4.2.2 Insuffisance des règles géométriques 73
2.4.2.3 Analyse discriminante probabiliste 73
2.4.3 Tests et sélection de variables discriminantes 79
2.4.3.1 Quelques rappels sur les tests de normalité 79
2.4.3.2 Homoscédasticité et test de Box 81
2.4.3.3 Test de Wilks 81
2.4.3.4 Evaluation individuelle des variables prédictives - Sélection de variables pas à pas 82
2.4.4 Validation d'une règle d'affectation 83
2.4.4.1 Méthode de resubstitution 83
2.4.4.2 Méthode de l'échantillon-test 84

3 Application **86**
3.1 Introduction 86
3.2 Les données 86
3.3 Nuages de points et corrélation 88
3.4 Méthode échantillon-test 89
3.4.1 Echantillon d'apprentissage 89
3.4.2 Echantillon test 90
3.5 L'analyse linéaire discriminante 90
3.6 Qualité de représentation sur les axes discriminants 91
3.7 Représentation simultanée 102
3.8 Matrice de confusion- échantillon d'apprentissage 103
3.9 Matrice de confusion- échantillon test 104

3.10 Test de Wilks 105

Conclusion **106**

Références bibliographiques **107**

Introduction générale

En statistique, les analyses multivariées ont pour caractéristique de s'intéresser à des lois de probabilité à plusieurs variables. Elles sont diverses selon l'objectif recherché, la nature des variables et la mise en œuvre formelle. On peut identifier deux grandes familles : celle des méthodes descriptives visant à structurer et résumer l'information (ACP, AFC, ACM, etc...) et celle des méthodes explicatives visant à expliquer une ou des variables dites variables à expliquer par un ensemble de variables dites variables explicatives (Analyse de régression, ANOVA, Analyse canonique des corrélations, ...etc).

Le premier chapitre est consacré aux trois méthodes les plus courantes en statistique descriptive multidimensionnelle : l'ACP, l'AFC et l'ACM.

Sous le nom d'analyse discriminante, on distingue toute une série de méthodes explicatives, descriptives et surtout prédictives destinées à étudier une population de n individus comportant K classes. Chaque individu est caractérisé par un ensemble de p variables quantitatives et une variable qualitative identifiant la classe à laquelle appartient cet individu. Cette analyse est l'objet du second chapitre.

Le troisième chapitre est réservé à l'application de l'analyse discriminante sous R sur des données réelles.

Chapitre 1

Analyse des données

1.1 Introduction

L'objet de ce chapitre est de donner quelques outils couramment employés en statistique pour traiter des données multidimensionnelles. Ces données correspondent souvent à l'observation de nombreuses variables aléatoires sur plusieurs individus. Elles sont représentées sous forme d'un tableau où chaque ligne représente les variables mesurées sur un individu. Le but est d'extraire le maximum d'informations de ce tableau de données. Les méthodes factorielles cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques et selon que l'on travaille avec un tableau de variables qualitatives ou quantitatives, on utilisera l'analyse en composante principale (variables quantitatives) ou factorielle de correspondance (voire multiples) (variables qualitatives). Par conséquent le choix d'une méthode statistique dépendra de la nature des variables.

1.2 Un peu d'histoire

Les méthodes d'analyse de données ont commencé à être développées dans les années 50 poussées par le développement de l'informatique et du stockage des données qui depuis n'a cessé de croître. L'analyse de données a surtout été développée en France par le mathématicien J.P. Benzécri ([12]).

En ce qui concerne l'analyse factorielle, il faut remonter aux travaux du psychologue anglais Ch.Spearman (1904) qui introduit pour la première fois le concept de facteur ; il cherche, derrière les notes obtenues par de nombreux sujets à de nombreux tests, une variable explicative cachée : le facteur général d'aptitude.([9])

Puis, l'analyse factorielle en composantes principales, développée par le statisticien et économiste américain H.Hotelling (1933) mais dont on peut faire remonter le principe au mathématicien britannique K.Pearson (1901) : les individus colonnes du tableau à analyser étant considérés comme des vecteurs d'un espace à p dimensions ; on proposait de réduire la dimension de l'espace en projetant le nuage des points individus sur le sous-espace de dimension k (k petit fixé) permettant d'ajuster au mieux le nuage.([4])

L'analyse factorielle des correspondances, introduite par J.P Benzécri (1962), fait l'objet d'un véritable phénomène de mode chez les statisticiens. Elle fournit, sans hypothèses a priori des représentations simplifiées dans un certain sens à l'interprétation. Laissons sur ce point la parole au professeur J.P Benzécri : l'analyse des correspondances telle qu'on la pratique en 1977 ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle aide à critiquer la validité des résultats, principalement par des calculs de contribution ; fournit des procédés efficaces de discrimination et de régression ; se conjugue harmonieusement avec la classification automatique ([6]).

Enfin, signalons l'analyse factorielle discriminante ou analyse discriminante initiée par le biologiste et statisticien britannique Fisher en 1936 qui permet de décrire la liaison entre une variable qualitative et un ensemble de variables quantitatives. Enfin l'analyse canonique introduite par Hotelling en 1936 et dont l'objectif initial était d'exprimer au mieux à l'aide d'un petit nombre de couples de variables la liaison entre deux ensembles de caractères quantitatifs, dépendant d'un même corps de résultats mathématiques.

1.3 Objectifs

L'analyse des données est un ensemble de techniques permettant d'étudier la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de la traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement.([5])

Ces techniques doivent donc permettre de représenter synthétiquement de vastes ensembles numériques pour faciliter au statisticien la prise de décisions. 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. Comme il est difficile de visualiser des points dans des espaces de dimensions supérieures à trois, nous chercherons à représenter graphiquement ces points dans des plans.

L'analyse des données ne se limite pas à une représentation des données, ou du moins à la rendre plus aisée, elle cherche 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.([12])

1.4 Domaines d'application

Les méthodes d'analyse de données sont employées dans un grand nombre de domaines.
Par exemple, l'analyse factorielle des correspondances est souvent utilisée en sciences humaines, pour cerner les résultats des enquêtes d'opinion. (On désigne par une enquête d'opinion, appelée aussi sondage d'opinion, toute application de la technique des sondages à une population humaine visant à déterminer les opinions probables des individus la composant, à partir de l'étude d'un échantillon de cette population).

Elle est aussi utilisée en sociologie pour analyser les réponses à un questionnaire.
En médecine, par exemple, pour détecter les groupes à haut risque cardiaque à partir de caractéristiques telles que l'alimentation, le fait de fumer ou pas, les antécédents familiaux etc..., on fait appel à l'analyse discriminante.
L'analyse en composantes principales est très applicable en biologie, la recherche économique et sociale et plus récemment en traitement d'images.
Ces méthodes sont aussi beaucoup utilisées en marketing, en informatique, dans le domaine bancaire etc....
En fait, tout domaine scientifique qui doit gérer de grandes quantités de données de type varié ont recours à ces approches ainsi que tout domaine industriel.

1.5 Analyse en Composantes Principales (ACP)

L'analyse en composantes principales est une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques spécifiés et que l'on utilise en général sans référence à des hypothèses de nature statistique ou à un modèle particulier.([3])

L'ACP est une des premières analyses factorielles qui constituent la plupart des analyses de données. Elle présente plusieurs variantes. On n'en décrira ici que deux :

- ***L'ACP centrée :***
  L'ACP centrée est celle où on centre les variables.

- ***L'ACP normée :***
  L'ACP normée est celle où on centre et réduit les variables : on opère une transformation linéaire sur chaque variable ramenant sa moyenne à zéro et sa variance à l'unité.

Remarque
On recommande l'ACP seulement centrée lorsque les variables sont homogènes (*i.e.* même système d'unités) et on recommande l'ACP normée lorsque les variables sont hétérogènes (*i.e.* dans le cas contraire).

1.5.1 Tableau de données et espaces associés

Les données se présentent généralement sous la forme d'un tableau rectangulaire de mesures numériques continues, à n lignes correspondant à des individus ou unités statistiques et à p colonnes représentant les variables appelées aussi caractères ou caractéristiques.

L'analyse factorielle consiste à faire la recherche des axes factoriels (ou axes principaux) des deux nuages. On cherche donc à ajuster le nuage des n points par un sous-espace vectoriel de $\mathbb{R}^p$ et le nuage des p points par un sous-espace vectoriel de $\mathbb{R}^n$.

1.5.1.1 Le tableau de données

On note X la matrice de dimension (n,p) contenant les observations :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_i^1 & x_i^j & x_i^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

où x_i^j est la valeur de l'individu i pour la variable j que l'on notera X^j et qui sera identifiée au vecteur de n composantes $(x_1^j, ..., x_n^j)'$.
De même l'individu i sera identifié au vecteur X_i à p composante avec $X_i = (x_i^1, ..., x_i^p)'$.([13],[10])

1.5.2 Caractéristiques de la variable X^j

– ***La moyenne :***

Supposons que chaque individu i est muni d'un poids p_i tel que : $p_i > 0, \forall i = \overline{1,n}$ et $\sum_{i=1}^{n} p_i = 1$.

La moyenne de la variable X^j est définie par :

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

– ***La variance :***

La variance de la variable X^j est définie par :

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

exprime la moyenne des carrés des écarts à la moyenne. Elle est aussi égale à la différence entre la moyenne des carrés et le carré de la moyenne.

La variance de X^j peut donc s'écrire comme suit :

$$\sigma_j^2 = \sum_{i=1}^{n} p_i x_i^{j2} - (\overline{x^j})^2$$

Sa racine carrée définit l'écart type σ_j.

– ***La covariance entre deux variables X^j et $X^{j'}$:***

La covariance est une extension de la notion de variance, elle est définie par :

$$cov(X^j, X^{j'}) = \sum_{i=1}^{n} p_i (x_i^j - \overline{x^j})(x_i^{j'} - \overline{x^{j'}})$$

– ***Le coefficient de corrélation entre deux variables X^j et $X^{j'}$:***

Le quotient de la covariance entre deux variables aléatoires par le produit de leurs écarts types s'appelle le coefficient de corrélation entre ces deux variables. Il s'écrit comme suit :

$$r(X^j, X^{j'}) = \frac{cov(X^j, X^{j'})}{\sigma_j \sigma_{j'}}$$

Remarque : $r(X^j, X^{j'})$ prend ses valeurs entre -1 et 1.

La variable centrée et réduite :
On définit la variable centrée réduite ainsi :

$$X_i^j = \frac{x_i^j - \overline{x^j}}{\sigma_j}$$

où σ_j est l'écart type de X^j.

1.5.3 Poids et centre de gravité

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales, les n individus ont tous la même importance dans les calculs des caractéristiques de l'échantillon. On leurs affecte donc un poids $p_i = \frac{1}{n}$.
Sinon on leurs affecte des poids $p_i / \ p_i > 0, \forall i = \overline{1,n}$ et $\sum_{i=1}^{n} p_i = 1$.
Ces poids qui sont comparables à des fréquences peuvent être regroupés dans une matrice diagonale de taille n :

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

Dans le cas le plus usuel de poids égaux, $D_p = \frac{1}{n} I_n$.
Où I_n est la matrice d'identité d'ordre n.

- **Remarque**
  Ces poids ne modifient en rien les principes de l'ACP, nous considérons donc par la suite les cas où $p_i = \frac{1}{n}$.

Le vecteur G des moyennes arithmétiques de chaque variable $G = (\overline{x^1}, ...\overline{x^p})'$ définit le point moyen ou centre de gravité du nuage.

1.5.4 La transformation des données

Les données pour l'ACP sont généralement présentées sous la forme de la matrice X vue précédemment.

Il existe plusieurs transformations utilisées. L'analyse centrée consiste à modifier les données de la matrice X en remplaçant les valeurs des x_i^j par $x_i^j - \overline{x^j}$.
Le fait de centrer les données entraîne que l'origine des axes est confondu avec le centre de gravité G.

L'analyse centrée réduite ou encore normée est liée à la transformation des données de la matrice X en remplaçant les valeurs des x_i^j par $\frac{x_i^j - \overline{x^j}}{\sigma_j}$.
Réduire les données permet d'uniformiser les unités de mesures.

Le tableau Y tel que $y_i^j = x_i^j - \overline{x^j}$ est le tableau centré associé à X.
Le tableau Z tel que $z_i^j = \frac{x_i^j - \overline{x^j}}{\sigma_j}$ est le tableau des données centrées et réduites associé à X.

1.5.5 Matrice de variance-covariance et matrice de corrélation

La matrice carrée V telle que $V = Y' D_p Y$ (avec Y'est la transposée de Y) est appelée matrice de variance-covariance. Cette matrice permet de quantifier la variation de chaque variable par rapport à chacune des autres.

$$V = \begin{pmatrix} var(X^1) & \dots & cov(X^1, X^p) \\ \vdots & \ddots & \vdots \\ cov(X^p, X^1) & \dots & var(X^p) \end{pmatrix}$$

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux, notée R, est définie comme suit : $R = Z' D_p Z$.

$$R = \begin{pmatrix} 1 & \dots & r(X^1, X^p) \\ \vdots & \ddots & \vdots \\ r(X^p, X^1) & \dots & 1 \end{pmatrix}$$

Rappelons que : cov(X,Y)=cov(Y,X) et r(X,Y)=r(Y,X).
Les matrices V et R sont donc symétriques et semi-définies positives (diagonalisables

et de valeurs propres réelles positives ou nulles).
On peut donc former une base orthonormée de $\mathbb{R}^p$ avec les valeurs propres de V ou R. Classons les valeurs propres suivant un ordre décroissant en notant λ_1 la première valeur propre et λ_p la dernière selon leur ordre de multiplicité.

Remarques

1. Les matrices V et R sont aussi appelées matrices d'inertie du nuage des points.
2. $R = D'_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}}$ où $D_{\frac{1}{\sigma}}$ désigne la matrice diagonale des inverses des écarts types. ([2])

1.5.6 L'analyse des nuages

L'analyse en composantes principales est particulièrement adaptée aux variables quantitatives, continues, a priori corrélées entre elles. Une fois les données projetées dans différents plans, les proximités entre variables s'interprètent en termes de corrélations, tandis que les proximités entre individus s'interprètent en termes de similitudes globales des valeurs observées.

1.5.6.1 Analyse du nuage des points individus dans $\mathbb{R}^p$

Pour établir un bilan des ressemblances entre individus, nous cherchons à répondre à des questions du type :

- Quels sont les individus qui se ressemblent ?
- Quels sont ceux qui sont différents ?
- Existe-t-il des groupes homogènes d'individus ?
- Est-il possible de mettre en évidence une typologie des individus ?

Définition 1.1. *Deux individus se ressemblent, ou sont proches, s'ils possèdent des valeurs proches pour l'ensemble des variables.*
Cette définition sous entend une notion de proximité qui se traduit par une distance. Ainsi, nous définissons la distance entre deux individus X_i et $X_{i'}$ par :

$$d^2(X_i, X_{i'}) = \sum_{j=1}^{p} (x_i^j - x_{i'}^j)^2$$

La métrique utilisée ici est donc euclidienne mais, de manière plus générale, nous pouvons définir cette distance par :

$$d^2(X_i, X_{i'}) = (x_i^j - x_{i'}^j)' M (x_i^j - x_{i'}^j)$$

où M est une matrice symétrique définie positive de taille p.([12])

– **Remarque** ([16])

La distance utilisée par l'ACP dans l'espace où sont représentés les individus, est la distance euclidienne classique.

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. On associe à cette distance un produit scalaire entre deux vecteurs :

$$< \overrightarrow{OX_i}, \overrightarrow{OX_{i'}} >= X_i' X_{i'}$$

ainsi que la norme d'un vecteur :

$$\| \overrightarrow{OX_i} \|^2 = X_i' X_i$$

1.5.6.2 Analyse du nuage des points variables dans $\mathbb{R}^n$

Il est essentiel de définir la liaison entre des variables en répondant à des questions du type :

– Quelles sont les variables qui sont liées positivement entre elles ?
– Quelles sont celles qui s'opposent (*i.e.* liées négativement) ?
– Existe-t-il des groupes de variables corrélées entre elles ?
– Est-il possible de mettre en évidence une typologie des variables ?

Définition 1.2. *Deux variables sont liées si elles ont un fort coeficient de corrélation linéaire.*

1.5.6.3 Inertie totale du nuage des individus

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

$$I_G = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_i^j - \overline{x_j})^2$$

Ce moment d'inertie total est intéressant car il mesure la dispersion du nuage des individus par rapport à son centre de gravité. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré sur son centre de gravité.([16])

– **Remarque**

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

$$I_G = \sum_{j=1}^{p}[\frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x_j})^2] = \sum_{j=1}^{p} var(X^j)$$

Sous cette forme, on constate que l'inertie totale est égale à la trace de la matrice de variance-covariance V.

$$I_G = trace(V)$$

1.5.7 L'ajustement

L'approche factorielle consiste donc à approcher ces nuages de points dans des sous-espaces vectoriels permettant de fournir des images planes de ces nuages.

1.5.7.1 Ajustement du nuage des individus

Axes principaux et composantes principales

Définition 1.3. *On appelle premier axe principal la direction de l'espace qui rend maximum l'expression suivante :*

$$I_n(a_1) = a_1' V a_1$$

avec $I_n(a_1)$ est l'inertie expliquée par la direction a_1.

Le problème à résoudre est donc la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte $|| a_1 ||= 1$ (les inconnues sont les composantes de a_1). La méthode des multiplicateurs de Lagrange peut alors être utilisée.
Dans le cas de la recherche de a_1, il suffit de calculer les dérivées partielles de :

$$L(a_1) = a_1' V a_1 - \lambda_1(a_1' a_1 - 1)$$

En utilisant la dérivée matricielle, on obtient :

$$\frac{\partial L(a_1)}{\partial(a_1)} = 2Va_1 - 2\lambda_1 a_1 = 0$$

Le système à résoudre est :

$$\begin{cases} Va_1 - \lambda_1 a_1 = 0...(1) \\ a_1' a_1 - 1 = 0...(2) \end{cases}$$

De l'équation matricielle (1) de ce système on déduit que a_1 est un vecteur propre de la matrice V associé à la valeur propre λ_1.
En multipliant à gauche par a_1' les deux membres de l'équation (1) on obtient :

$$a_1' V a_1 - \lambda_1 a_1' a_1 = 0$$

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

$$a_1' V a_1 = \lambda_1$$

Le premier membre de l'équation précédente est égal à l'inertie $I_n(a_1)$ qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice de covariance V et que cette valeur propre est égale à l'inertie portée par l'axe a_1.([16])

On vient de démontrer le résultat suivant :

Proposition 1.1. *Le premier axe principal est engendré par le vecteur propre normé a_1 correspondant à la plus grande valeur propre λ_1. L'inertie expliquée par cet axe est égale à λ_1.*

Le sous-espace à une dimension optimale au sens de l'inertie maximale est donc l'axe engendré par a_1.

Cherchons maintenant à déterminer le sous-espace à deux dimensions s'ajustant au mieux au nuage des individus.

Le sous-espace à deux dimensions qui s'ajuste au mieux au nuage des individus contient a_1. En effet, par un raisonnement par l'absurde, si ce sous-espace ne contient pas a_1, alors il est défini par deux vecteurs $a^{'}$ et $a^{''}$ différents de a_1. L'inertie le long des deux droites portées par $a^{'}$ et $a^{''}$ est donc inférieure à celle de l'inertie le long de la droite portée par a_1. Il existe donc un sous-espace de dimension deux meilleur que celui défini par les deux vecteurs $a^{'}$et $a^{''}$. Nous montrons ainsi que le sous-espace à deux dimensions qui ajuste au mieux le nuage des individus contient a_1.

Le sous-espace à deux dimensions est donc caractérisé par le vecteur a_1 orthogonal à a_2 vérifiant :

- $a_2' V a_2$ est maximal.
- $a_2' a_2 = 1$ (contrainte de normalité).
- $a_2' a_1 = 0$ (contrainte d'orthogonalité).([12])

Par récurrence, le sous-espace à p dimensions s'ajustant au mieux au nuage des individus contient les vecteurs $a_1, ..., a_p$ vérifiant :

- $a_p' V a_p$ est maximal.
- $a_p' a_p = 1$ (contrainte de normalité).
- $a_p' a_q = 0$ pour tout $q = \{1, ..., p-1\}$ (contrainte d'orthogonalité).

Proposition 1.2. *Une base orthonormée du sous-espace vectoriel de dimension q, s'ajustant au mieux au nuage des individus dans R^n est constituée par les q vecteurs propres $(a_1, ..., a_q)$ correspondant aux q plus grandes valeurs propres de la matrice V. L'inertie expliquée par ce sous-espace est égale à la somme des q premières valeurs propres de V.*

- **Remarque**

  Le plan engendré par les vecteurs propres normés a_1 et a_2 est appelé le premier plan principal, le plan engendré par les vecteurs propres normés a_1 et a_3 est

appelé le deuxième plan principal,...etc

On cherche des combinaisons linéaires des variables initiales, appelées *facteurs* ou encore *composantes principales*.
Les composantes principales sont les variables artificielles, non corrélées et définies par les axes principaux :

$$c^j = Xa_j$$

elles donnent les coordonnées des projections orthogonales des individus sur les axes définis par les a_j.

– **Remarque**

La variance d'une composante principale est égale à la valeur propre λ correspondante : $var(c^j) = \lambda_j$. En effet, par définition on a $a'_j a_j = 1$ et donc

$$var(c^j) = c^{j'} D_p c^j = a'_j X' D_p X a_j = a'_j V a_j = a'_j(\lambda_j a_j) = \lambda_j a'_j a_j = \lambda_j$$

Qualité des représentations sur les plans principaux

La contribution absolue de l'axe a_j à l'inertie totale du nuage des individus est égale à λ_j, la valeur propre qui lui est associée.
Sa contribution relative est égale à :

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

On emploie souvent l'expression *pourcentage ou taux d'inertie* expliquée par l'axe engendré par a_j.
On peut étendre ces définitions à tous les sous-espaces engendrés par les p axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes a_1 et a_2 est égal à :

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres sont très faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.([16])

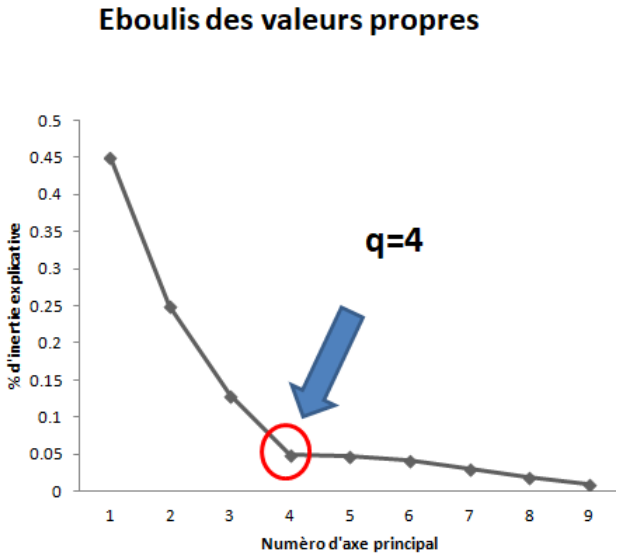
– **Remarque**

On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les q premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse à un sous-espace de dimension $q < p$.

Nombre d'axes à retenir

Le nombre q d'axes principaux à retenir peut être choisi par les méthodes suivantes :

1. *Règle de la part d'inertie :* on se fixe un seuil correspondant au pourcentage minimum d'inertie que l'on veut restituer et on retiendra le nombre q d'axes nécessaires pour atteindre ce seuil, par exemple si on fixe un seuil de 80%, on choisit q tel que :
$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_p} \geq 80\%$$
2. *Critère du coude :* sur le diagramme des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement. Pour cela il suffit de porter sur un graphique en abscisse, le numéro de l'axe principal, et en ordonnée, le pourcentage d'inertie qu'il restitue (voir l'éboulis des valeurs propres suivant) :



3. *Critère de Kaiser :* on ne retient que les axes dont l'inertie I_n est supérieure à l'inertie moyenne $\frac{I_n}{p}$.

Dans le cas d'une ACP normée : $\frac{I_n}{p} = 1$.
En effet, dans l'ACP normée : $p = \lambda_1 + \lambda_2 + ... + \lambda_p = I_n$. On ne retiendra donc que les axes associés à des valeurs propre supérieures à 1.

Qualité de la représentation des individus

Lorsque des points projections des individus sont éloignés sur un axe (ou dans un plan), on peut assurer que les points représentant ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace.
Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut donc s'assurer que ces individus sont bien représentés dans le plan. Définissons alors quelques aides à l'interprétation :

a- Contribution relative d'un individu à un axe

Il est très utile de calculer pour chaque axe la contribution apportée par les divers individus à cet axe. Considérons la jième composante c^j ; soit c_i^j la valeur de cette composante pour le ième individu.

La contribution de l'individu X_i à l'axe engendré par a_j est définie par :

$$CTR_j(X_i) = \frac{p_i (c_i^j)^2}{\lambda_j}$$

On peut aussi définir une mesure de la contribution de l'individu X_i à l'inertie du nuage des individus comme le rapport :

$$CTR(X_i) = \frac{p_i \sum_{j=1}^{p} (c_i^j)^2}{\sum_{j=1}^{p} \lambda_j}$$

Ces contributions permettent de détecter les observations les plus influentes et éventuellement aberrantes qui peuvent déterminer à eux seuls l'orientation des axes et plus globalement l'ensemble des résultats de l'ACP. En effet : il n'est pas souhaitable qu'un individu ait une contribution excessive car cela serait un facteur d'instabilité, le fait de retirer cet individu modifie profondément le résultat de

l'analyse. Si ce cas se produisait il y aurait intérêt à effectuer l'analyse en éliminant cet individu puis le réintroduire a posteriori comme élément supplémentaire (que nous allons définir dans la suite).

– **Remarque**

Par construction on a :

$$\sum_{i=1}^{n} CTR_j(X_i) = 1$$

et

$$\sum_{i=1}^{n} CTR(X_i) = 1$$

b-Qualité de projection d'un individu sur un axe

La qualité de représentation de l'individu X_i sur l'axe engendré par a_j est mesurée par le cosinus carré de l'angle formé par X_i et l'axe engendré par a_j :

$$CO2_j(X_i) = \frac{(c_i^j)^2}{\sum_{j=1}^{p} (c_i^j)^2}$$

– Si $CO2_j(X_i)$ est proche de 1, l'individu X_i est bien représenté sur l'axe engendré par a_j.
– Si $CO2_j(X_i)$ est proche de 0, l'individu X_i est mal représenté sur l'axe engendré par a_j.

Remarque

Il est facile de voir que :

$$\sum_{i=1}^{n} CO2_j(X_i) = 1$$

Par orthogonalité des axes principaux, la qualité de projection d'un individu sur un plan principal est additive :

$$CO2_{j+j'}(X_i) = CO2_j(X_i) + CO2_{j'}(X_i)$$

D'une façon plus générale : la qualité de projection d'un individu sur le sous-espace de dimension q formé par les q premiers axes principaux peut être mesurer par la quantité suivante :

$$QLT_q(X_i) = \sum_{j=1}^{q} CO2_j(X_i)$$

– Si $QLT_q(X_i) > 90\%$, par exemple, l'individu X_i est bien représenté sur le sous-espace à q dimension formé par les q premiers axes principaux.

Orientation des plans principaux

Chaque point individu X_i est un vecteur de $\mathbb{R}^p$ qu'on peut écrire sous la forme :

$$X_i = \sum_{j=1}^{p} X_i^j e_j$$

où e_j est la base canonique de $\mathbb{R}^p$.

Les axes engendrés par les $e_j, j = \overline{1,p}$ sont appelés les axes variables de départ.

Dans un plan principal, on peut faire apparaître les projections de ces p axes en cherchant les coordonnés du vecteur e_j dans le plan principal engendré par les vecteurs propres normés a_j et $a_{j'}$, elles sont données par :

$$< e_j, a_j > \qquad et \qquad < e_j, a_{j'} >$$

La projection du point individu X_i sur le plan principal éloigné dans la direction de la variable X^j indique une valeur X_i^j au dessus de la moyenne, à condition que le point X_i et sa projection soient assez voisins *i.e.* que la qualité de projection de l'individu X_i sur le plan soit supérieur à 90%.

1.5.7.2 Ajustement du nuage des variables

De même que nous avons réalisé l'ajustement des points individus, nous pouvons procéder à l'ajustement des points variables dans un espace de dimension réduit.

Mathématiquement, cela conduit à diagonaliser la matrice $\Gamma = YD_pY'$ dans le cas d'une ACP centrée ou $\Sigma = ZD_pZ'$ dans le cas d'une ACP normée.

Cet ajustement est appelé *dual* de l'ajustement du nuage des points individus.

Comme précédemment, on obtient la décomposition de l'inertie selon les directions définies par les vecteurs propres de la matrice Γ. L'inertie projetée sur chaque direction est égale à la valeur propre associée.

La droite d'inertie maximale se trouve définie par le vecteur propre b_1, associé à la plus grande valeur propre. Le plan d'inertie maximale s'obtient en ajoutant à cette droite, la direction du vecteur propre b_2 associé à la seconde valeur propre, et ainsi de suite.

La représentation des variables sur un axe s'obtient par projection des points variables sur le vecteur unitaire b_j qui définit la direction de l'axe.
Soit F^j la coordonnée de la variable X^j sur l'axe b_j. On aura :

$$F^j = X' b_j$$

- **Remarque**

  Entre les axes principaux des deux nuages de points, il existe des relations qui permettent, connaissant les directions dans un espace, d'obtenir les directions dans l'autre espace. On les appelle *relations de transition.*

Proposition 1.3. *Si a_j est un vecteur propre normé de V associé à la valeur propre non nulle λ_j, alors $b_j = \frac{1}{\sqrt{\lambda_j}} X' a_j$ est un vecteur propre unitaire de Γ associé à la valeur propre λ_j.*

Démonstration

Supposons que a_j est un vecteur propre de V associé à la valeur propre non nulle λ_j, alors par définition d'un vecteur propre on a :

$$V a_j = \lambda_j a_j$$

En remplaçant V par $\frac{1}{n} X'X$ on aura :

$$\frac{1}{n} X'X a_j = \lambda_j a_j$$

En multipliant à gauche par X les deux membres de l'équation ci-dessus on obtient :

$$X \frac{1}{n} X'X a_j = X \lambda_j a_j$$

Du fait que le produit matricielle est associatif, on aura :

$$(\frac{1}{n}XX')Xa_j = \lambda_j Xa_j$$

Remplaçons $\frac{1}{n}XX'$ par Γ :

$$\Gamma Xa_j = \lambda_j Xa_j$$

D'où : Xa_j est un vecteur propre de Γ associé à la valeur propre non nulle λ_j.
Mais Xa_j n'est pas unitaire. On doit donc trouver un vecteur du type kXa_j qui soit unitaire, *i.e.* $(kXa_j)'(kXa_j) = 1$.
$\iff k^2(a_j'X'Xa_j) = 1$
$\iff k^2(a_j'\lambda_j a_j) = 1$ (a_j vecteur propre de $X'X$)
$\iff k^2\lambda_j a_j'a_j = 1$ (le produit matricielle est associatif)
$\iff k^2\lambda_j = 1$ ($a_j'a_j = 1$)
$\iff k^2 = \frac{1}{\lambda_j} \iff k = \frac{1}{\sqrt{\lambda_j}}$
Donc $\frac{1}{\sqrt{\lambda_j}}Xa_j$ est un vecteur propre unitaire de Γ associé à la valeur propre λ_j.

- **Remarque**

  On vient de démontrer que les matrices V et Γ ont les mêmes valeurs propres non nulles.

Dans la pratique, il suffit de réaliser un seul ajustement et à partir de celui-ci, on obtient l'autre.
On réalise en général l'ajustement de plus faible coût, celui du nuage de points individus si $p < n$ (on diagonalise une matrice de dimension plus petite).
La projection des variables est calculée en fonction des axes principaux du nuage des points individus :

$$F^j = a_j\sqrt{\lambda_j}$$

La formule ci-dessus permettra ultérieurement d'interpréter la représentation simultanée des individus et des variables (cette manière de représentation s'appelle *biplot*).

Remarques

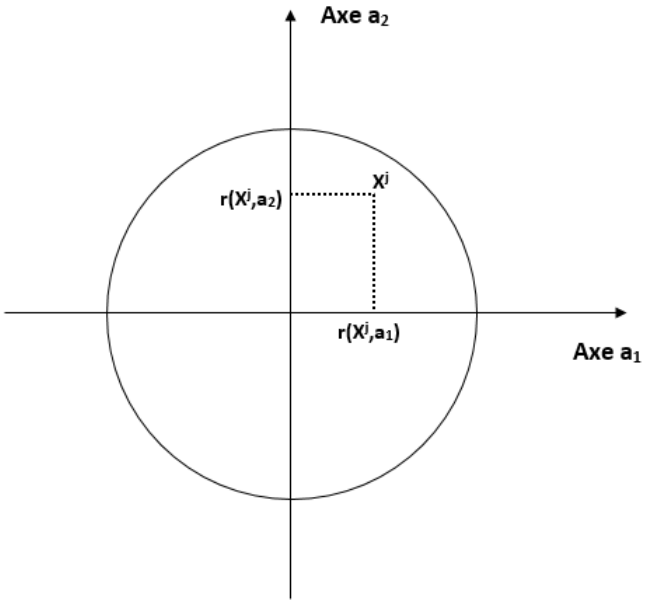
1. Dans le cas d'une analyse normée, la coordonnée d'une variable X^j sur l'axe a_j coïncide avec la corrélation de cette variable avec l'axe a_j :

$$F^j = r(X^j, a_j)$$

Cette formule joue un rôle important pour l'interprétation des résultats. Elle met en relation les représentations des points individus avec les représentations des points variables.

Une corrélation élevée indique que la disposition des individus sur l'axe principal est semblable à l'ordonnance des individus selon la variable (la corrélation unité signifierait que la composante principale est fonction linéaire de la variable). Une corrélation proche de zéro indique qu'il n'y a pas d'association linéaire entre la composante principale et la variable.

2. En travaillant sur des données centrées réduites : pour un couple d'axes principaux engendrés par a_1 et a_2, par exemple, on synthétise usuellement les corrélations sur une figure appelée *cercle des corrélations* (de rayon 1) où chaque variable X^j est repérée par un point d'abscisse $r(X^j, a_1)$ et d'ordonnée $r(X^j, a_2)$ (voir la figure suivante) :



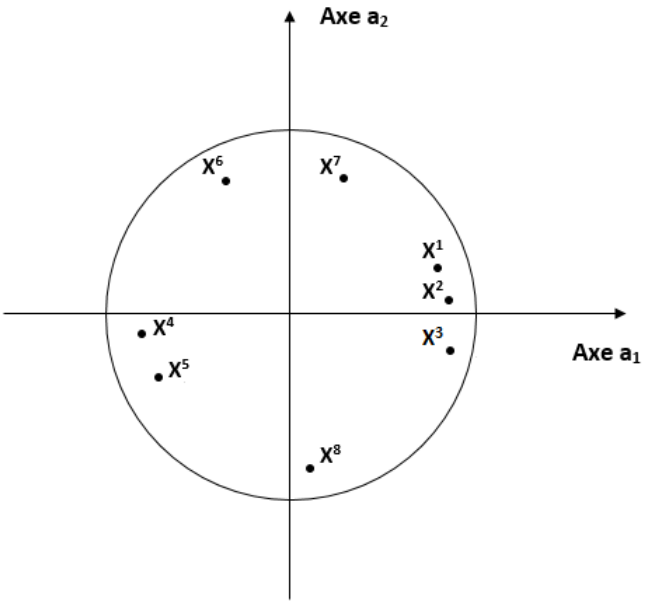
3. Si deux variables X^j et $X^{j'}$ sont bien représentées sur le cercle des corrélations (proches de l'axe et du bord du cercle) alors :

   * La proximité des projections de X^j et $X^{j'}$ indique une forte corrélation linaire entre ces deux variables.

   * Si les projections des points X^j et $X^{j'}$ sont opposées, X^j et $X^{j'}$ sont corrélées négativement.

   * Des directions de X^j et $X^{j'}$ presque orthogonales indiquent une faible corrélation entre ces variables.

4. On remarque, en visualisant la figure ci-dessous, que l'axe principal a_1 est très corrélé positivement avec les variables X^1, X^2 et X^3, anticorrélé (corrélé négativement) avec les variables X^4 et X^5 et non corrélé avec X^6, X^7 et X^8.



Par contre l'axe principal a_2 oppose la variable X^8 aux variables X^6 et X^7.

1.5.8 Individus et variables supplémentaires

Il arrive que l'on veuille faire apparaître dans les représentations graphiques certains individus sans qu'ils interviennent dans la détermination des axes. Cela peut être le cas de nouveaux individus que l'on veut simplement positionner par rapport aux autres sans que les positions de ceux-ci soient influencées par les nouveaux. On

dit d'eux qu'ils sont *des individus supplémentaires.*

Il en est de même pour les variables. On peut, par exemple, vouloir représenter une variable qui dépend de façon synthétique des p variables choisies pour faire l'ACP, afin de mieux comprendre comment cette variable est liée aux anciennes, mais on ne souhaite pas qu'elle intervienne dans la construction des axes car ses liaisons avec les p variables de départ fausseraient la représentation si elle faisait partie intégrante de l'ACP. Elles sont appelées *variables supplémentaires.*

Pour représenter un individu supplémentaire, il suffit d'exprimer les coordonnées de cet individu dans la nouvelle base des axes principaux. Pour une variable supplémentaire, il suffit de calculer ses coefficients de corrélation empiriques avec les composantes principales.([16])

1.5.9 ACP et variables qualitatives

En ACP, il est fréquent que l'on veuille introduire des variables qualitatives supplémentaires. Pour cela il suffit de ramener la variable qualitative ayant m modalités à m groupes d'individus définis par la modalité de la variable. On traite ensuite ces m groupes d'individus comme des individus supplémentaires. Ce sont les centres de gravité de ces groupes d'individus qui vont être positionnés dans l'espace $\mathbb{R}^p$.

Remarques

- On peut représenter avec des symboles différents ces groupes d'individus définis par chaque modalité.
- L'analyse d'une variable nominale supplémentaire ne se fait pas dans $\mathbb{R}^n$ mais dans $\mathbb{R}^p$.

1.5.10 Interprétation

Définissons quelques règles pour l'interprétation :

- Un individu sera du côté des variables pour lesquelles il a de fortes valeurs, inversement il sera du côté opposé des variables pour lesquelles il a de faibles valeurs.
- Plus les valeurs d'un individu sont fortes pour une variable plus il sera éloigné de l'origine suivant l'axe principal décrivant le mieux cette variable.
- Deux individus à une même extrémité d'un axe (*i.e.* éloignés de l'origine) sont proches (*i.e.* se ressemblent).
- Deux variables très corrélées positivement sont du même côté sur un axe.
- Il n'est pas possible d'interpréter la position d'un individu par rapport à une seule variable et réciproquement, il n'est pas possible d'interpréter la position d'une variable par rapport à un seul individu. Les interprétations doivent se faire de manière globale.([12])

1.5.11 Limites de l'ACP

La principale faiblesse de l'ACP est la sensibilité aux points extrêmes (manque de robustesse) qui est notamment lié au rôle central qu'y joue le coefficient de corrélation : les points extrêmes, en perturbant les moyennes et corrélations, faussent fortement l'analyse ; on peut cependant envisager de les déplacer en point supplémentaire.
L'ACP est aussi inadaptée aux phénomènes non linéaires. Pour ce genre de problème, d'autres méthodes ont été développées, comme l'ACPN (Analyse en Composantes Principales par Noyau).

En résumé, on peut dire que l'ACP consiste à transformer les variables initiales X^j corrélées en de nouvelles variables, les composantes principales c^j, combinaisons linéaires des X^j non corrélées entre elles, de variance maximale et les plus liées aux X^j : l'ACP est une méthode factorielle linéaire.

1.6 Analyse Factorielle des Correspondances (AFC)

L'analyse factorielle des correspondances (ou analyse des correspondances binaires) a été conçue pour l'étude des tableaux de contingence obtenus par croisement de variables qualitatives. Cette analyse permet donc de traiter des variables qualitatives et est surtout adaptée à ce type de variables. Comme l'ACP, l'AFC a pour but de décrire sous forme graphique, le maximum d'information contenu dans un tableau de données. Dans cette approche, les lignes et les colonnes ont un rôle symétrique et s'interprètent de la même façon.

- **Remarque**
  L'AFC peut également être étendue aux variables quantitatives homogènes (*i.e.* même système d'unités), en définissant simplement quelques modalités pour ces variables. Par extension, elle s'applique aussi aux tableaux individus-variables pour des variables quantitatives homogènes, dans ce cas les individus sont considérés comme des variables.

1.6.1 Tableau de contingence, marges et profils

On considère deux variables qualitatives observées simultanément sur k individus. On suppose que la première variable, notée X, possède n modalités notées $x_1, ..., x_i, ..., x_n$ et que la seconde, notée Y, possède p modalités notées $y_1, ..., y_j, ..., y_p$. On présente usuellement les données sous la forme d'un tableau croisé appelé *tableau de contingence* (ou *tableau de dépendance*) à n lignes et p colonnes renfermant les effectifs k_{ij} d'individus (voir le tableau suivant) :

$X\|Y$	y_1	y_2	.....	y_j	.....	y_p	
x_1	k_{11}	k_{12}				k_{1p}	$k_{1.}$
x_2	k_{21}	k_{22}				k_{2p}	$k_{2.}$
...							
x_i				k_{ij}			$k_{i.}$
...							
x_n	k_{n1}	k_{n2}				k_{np}	$k_{n.}$
	$k_{.1}$	$k_{.2}$		$k_{.j}$		$k_{.n}$	k

Avec des notations standards on a : $k_{i.} = \sum_{j=1}^{p} k_{ij}$ et $k_{.j} = \sum_{i=1}^{n} k_{ij}$, avec k_{ij} est le nombre d'individus possédant à la fois la modalité i de la première variable et la modalité j de la seconde variable. Nous avons donc :

$$\sum_{i=1}^{n}\sum_{j=1}^{p} k_{ij} = k$$

Les $k_{i.}$ et les $k_{.j}$ s'appellent respectivement *marges en lignes* et *marges en colonnes*.

Davantage que le tableau précédent, c'est le tableau des fréquences (ou des probabilités) relatives suivant qui est considéré :

$X\|Y$	y_1	y_2	.....	y_j	.....	y_p	
x_1	f_{11}	f_{12}				f_{1p}	$f_{1.}$
x_2	f_{21}	f_{22}				f_{2p}	$f_{2.}$
...							
x_i				f_{ij}			$f_{i.}$
...							
x_n	f_{n1}	f_{n2}				f_{np}	$f_{n.}$
	$f_{.1}$	$f_{.2}$		$f_{.j}$		$f_{.n}$	1

Les fréquences f_{ij} sont données par :

$$f_{ij} = \frac{k_{ij}}{k}$$

et les marges par :

$$f_{i.} = \sum_{j=1}^{p} f_{ij} \quad et \quad f_{.j} = \sum_{i=1}^{n} f_{ij}$$

Nous avons ainsi :

$$\sum_{i=1}^{n} f_{i.} = \sum_{j=1}^{p} f_{.j} = \sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = 1$$

Définition 1.4. *On appelle* tableau des profils-lignes *le tableau des frequences conditionnelles* $\frac{k_{ij}}{k_{i.}} = \frac{f_{ij}}{f_{i.}}$ *(la somme de chaque ligne est ramenée à* 100%*) et* tableau des profils-colonnes *le tableau des fréquences conditionnelles* $\frac{k_{ij}}{k_{.j}} = \frac{f_{ij}}{f_{.j}}$ *(le total de chaque colonne est alors ramené à* 100%*).([13])*

- **Remarque**

  Pour analyser un tableau de contingence ce n'est pas le tableau brut qui est intéressant mais les tableaux des profils lignes et des profils colonnes (*i.e.* les répartitions en pourcentage à l'intérieur d'une ligne ou d'une colonne).

1.6.2 Liaison entre deux variables qualitatives

L'étude traditionnelle d'un tableau de contingence se concentre le plus souvent sur la dépendance ou l'indépendance entre les deux caractères : X et Y.

Définition 1.5. *Il y a indépendance entre les deux variables considérées si :*

$$f_{ij} = f_{i.}f_{.j}, \quad \forall i = \overline{1,n} \quad et \quad \forall j = \overline{1,p}$$

Nous dirons qu'il y a liaison entre ces deux variables ou que ces deux variables sont liées si elles ne sont pas indépendantes.([12])

Ainsi nous pouvons dire que :

- Si f_{ij} est supérieur au produit des marges, les modalités i et j s'associent plus que sous l'hypothèse d'indépendance. Nous dirons que les deux modalités i et j *s'attirent.*

– Si f_{ij} est inférieur au produit des marges, les modalités i et j s'associent moins que sous l'hypothèse d'indépendance. Nous dirons qu'il y a *répulsion* entre les deux modalités i et j.

Remarque

Sous l'hypothèse d'indépendance nous avons :

– En considérant le tableau comme un ensemble de lignes :

$$\frac{f_{ij}}{f_{i.}} = f_{.j}, \quad \forall i = \overline{1,n} \;\; et \;\; \forall j = \overline{1,p}$$

– En considérant le tableau comme un ensemble de colonnes :

$$\frac{f_{ij}}{f_{.j}} = f_{i.}, \quad \forall i = \overline{1,n} \;\; et \;\; \forall j = \overline{1,p}$$

le terme $f.j$ s'interprète comme le pourcentage de la population totale possédant la modalité j et le terme $\frac{f_{ij}}{f_{i.}}$ représente ce même pourcentage dans la sous-population possédant la modalité i. Ainsi il y a indépendance lorsque les lignes du tableau de fréquences sont proportionnelles.
Par symétrie il en est de même pour les colonnes.

On peut aussi parler d'indépendance entre X et Y si tous les profils lignes sont identiques puisque la connaissance de X ne change pas les distributions conditionnelles de Y. Il s'ensuit d'ailleurs que tous les profils colonnes sont également identiques.
On doit donc avoir

$$\frac{k_{1j}}{k_{1.}} = \frac{k_{2j}}{k_{2.}} = ... = \frac{k_{nj}}{k_{n.}}, \quad \forall j = \overline{1,p}$$

ce qui entraîne, par sommation des numérateurs et dénominateurs,

$$\frac{k_{ij}}{k_{i.}} = \frac{k_{.j}}{k}$$

L'indépendance empirique se traduit donc par :

$$k_{ij} = \frac{k_{i.}k_{.j}}{k}$$

Test de Chi2

Comme en ACP, on s'intéresse alors aux directions de plus grande dispersion de chacun des deux nuages de points, mais on utilise la distance du χ^2 entre ces deux variables (à la place de la distance euclidienne). Cette distance permet de comparer l'effectif de chacune des cellules du tableau de contingence à la valeur qu'elle aurait si les deux variables étaient indépendantes.

Soient les hypothèses suivantes :

- L'hypothèse nulle H_0 : X et Y sont indépendantes en probabilités.
- L'hypothése altérnative H_1 : les variables X et Y ne sont pas indépendantes.

La statistique de test est alors

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(k_{ij} - \frac{k_{i.}k_{.j}}{k})^2}{\frac{k_{i.}k_{.j}}{k}} = k\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

elle suit asymptotiquement et si l'hypothèse H_0 est vraie, une loi de χ^2 à (n-1)(p-1) degrés de liberté. On rejette donc H_0 si χ^2 dépasse une valeur particulière (on prend généralement la valeur 5%).

- **Remarque**

  Les $\frac{k_{i.}k_{.j}}{k}$ sont les effectifs théoriques sous H_0.

1.6.3 Métriques du Chi2

Sur le plan mathématique, on peut considérer l'analyse des correspondances comme une analyse en composantes principales avec une métrique spéciale, la métrique du χ^2.

Les profils lignes forment un nuage de n points dans $\mathbb{R}^p$, chacun de ces points est muni d'un poids $f_{i.}$.
Le centre de gravité (le barycentre) de ce nuage de points est :

$$G_n = (f_{.1}, ..., f_{.j}, ..., f_{.p})'$$

Réciproquement, les profils colonnes forment un nuage de p points dans $\mathbb{R}^n$, chacun de ces points est muni d'un poids $f_{.j}$.
Le centre de gravité du nuage des points colonnes est donné par :

$$G_p = (f_{1.}, ..., f_{i.}, ..., f_{n.})'$$

Les espaces $\mathbb{R}^p$ et $\mathbb{R}^n$ sont respectivement munis des métriques, dites du χ^2, de matrices diagonales D_p^{-1} et D_n^{-1} données comme suit :

$$D_p^{-1} = \begin{pmatrix} \frac{1}{f_{.1}} & & \\ & \ddots & \\ & & \frac{1}{f_{.p}} \end{pmatrix}$$

et

$$D_n^{-1} = \begin{pmatrix} \frac{1}{f_{1.}} & & \\ & \ddots & \\ & & \frac{1}{f_{n.}} \end{pmatrix}$$

Remarques

1. Le tableau des profils lignes d'éléments $\frac{f_{ij}}{f_{i.}}$ est alors $D_n^{-1}F$, où F est le tableau des fréquences.
2. Le tableau des profils colonnes d'éléments $\frac{f_{ij}}{f_{.j}}$ est alors FD_p^{-1}.

On utilise la métrique de χ^2 plutôt que la métrique euclidienne pour deux raisons fortes :

a- Avec la métrique du χ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes.

b- La métrique du χ^2 possède la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils colonnes, ou entre les autres profils lignes restent inchangées.([15])

1.6.4 Analyse en composantes principales des deux nuages des profils

L'AFC étant une ACP particulière, nous ne donnons pas trop de détails techniques sur cette méthode. On en donne juste les grandes lignes dans ce paragraphe.
En pratique, nous devons centrer le nuage des points lignes, ainsi le centre de gravité G_n devient l'origine des axes.
Deux ACP sont donc possibles :

(1)- ACP du nuage des points lignes dans $\mathbb{R}^p$ avec :

- tableau de données $X = D_n^{-1}F$.
- métrique $M = D_p^{-1}$.
- poids $D = D_n$.

(2)- ACP du nuage des points colonnes dans $\mathbb{R}^n$ avec :

- tableau de données $X = D_p^{-1}F'$.
- métrique $M = D_n^{-1}$.
- poids $D = D_p$.

Les axes principaux sont les vecteurs propres de $MX'DX$ et les composantes principales sont les vecteurs propres de $XMX'D$.
Dans le cas de l'analyse des profils lignes, on a :

$$MX'DX = D_p^{-1}F'D_n^{-1}F$$

et

$$XMX'D = D_n^{-1}FD_p^{-1}F'$$

Remarques

1. Pour les profils colonnes il suffit de transposer F et d'inverser les indices n et p,

comme le montre le tableau suivant :

	ACP des profils lignes	ACP des profils colonnes
Axes principaux	Vecteurs propres de $D_p^{-1}F'D_n^{-1}F$	Vecteurs propres de $D_n^{-1}FD_p^{-1}F'$
Composantes principales	Vecteurs propres de $D_n^{-1}FD_p^{-1}F'$ normalisés par $a'D_n^{-1}a$	Vecteurs propres de $D_p^{-1}F'D_n^{-1}F$ normalisés par $b'D_p^{-1}b$

où a et b sont respectivement des vecteurs propres de $D_n^{-1}FD_p^{-1}F'$ et $D_p^{-1}F'D_n^{-1}F$.

On constate que les deux analyses conduisent aux mêmes valeurs propres et que les facteurs principaux de l'une sont les composantes principales de l'autre (à un coefficient multiplicateur près).([13])

2. Les coordonnées des points lignes et les coordonnées des points colonnes sont reliées par des formules de transition (appelées également barycentriques ou encore quasi-barycentriques) dont le premier intérêt est d'éviter de réaliser deux diagonalisations.

   On diagonalisera la matrice de dimension plus petite en formant la matrice de dimensions (p,p) plutôt que la matrice de dimensions (n,n) en supposant $(n \geq p)$.

   Ces formules sont données par :

$$a = \frac{1}{\sqrt{\lambda}} D_n^{-1} F b \quad soit \quad a_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{f_{ij}}{fi.} b_j$$

   et

$$b = \frac{1}{\sqrt{\lambda}} D_p^{-1} F' a \quad soit \quad b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \frac{f_{ij}}{f.j} a_i$$

   où λ est une valeur propore de $D_p^{-1}F'D_nF$ et de $D_n^{-1}FD_pF'$.

1.6.4.1 Interprétation des valeurs propres

– Notons tout d'abord que la première valeur propre est une valeur propre triviale égale à 1. En effet, le premier vecteur propre associé à cette première valeur

propre définit un axe principal pour lequel les projections des points lignes et des points colonnes possèdent une variance (dispersion) nulle. Ce qui signifie que toutes les projections possèdent les mêmes coordonnées. L'axe principal correspondant à cette valeur sera donc exclu de l'analyse.

Les autres valeurs propres sont toutes inférieures à 1.

- Les valeurs propres proches de 1 traduisent une forte liaison entre les lignes et les colonnes.
- Si une valeur propre est proche de 1, cela indique l'existence de deux sous groupes de modalités dans les données. Il est alors intéressant de réaliser des AFC indépendamment sur les deux sous groupes.
- De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne associée.

Remarque

Dans la pratique, on ne considère que $d = inf(n-1, p-1)$ plus grandes valeurs propres différentes de 1, ainsi que les vecteurs propres associés.

1.6.4.2 Interprétation des plans de projection en AFC

Pour une bonne interprétation des plans de projection en AFC, nous proposons de suivre la même démarche qu'en ACP *i.e.* choisir le nombre d'axes de projection à étudier, calculer les contributions relatives des lignes et des colonnes à un axe ou à un plan, étudier la qualité de projection des lignes et des colonnes sur un axe ou sur un plan,...etc.

1.6.4.3 Représentation simultanée (Biplot)

La parfaite symétrie entre ACP des profils lignes et ACP des profils colonnes conduit alors à superposer les plans principaux des deux ACP afin d'obtenir une représentation simultanée des catégories des deux variables croisées dans le tableau de contingence (voir la figure suivante) :

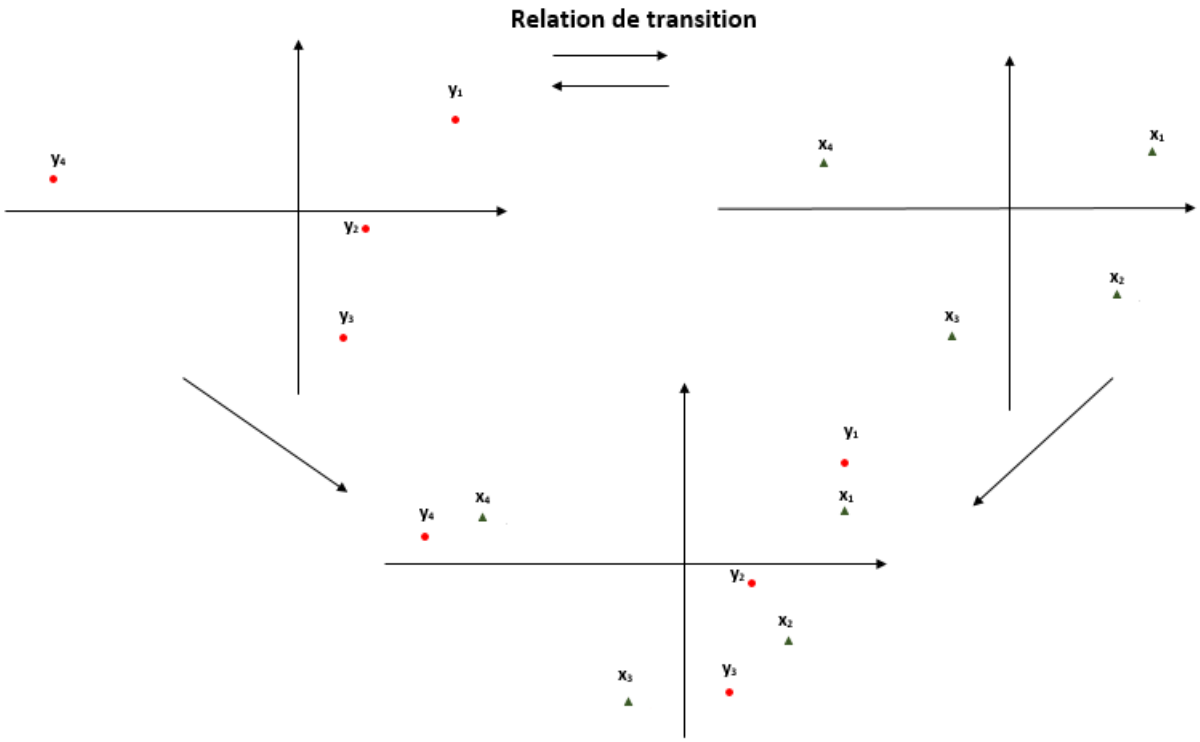


FIG. 1.1 – Représentation simultanée sur le premier plan principal.

– **Remarque**

En AFC, la représentation simultanée des deux nuages repose sur une dualité plus riche qu'en ACP car les lignes et les colonnes représentent des éléments de même nature.

Interprétation

La position relative de deux points d'un même ensemble (ligne ou colonne), s'interprète en tant que distance. La position d'un point d'un ensemble et tous les points d'un autre ensemble s'interprète en tant que barycentre.

La représentation simultanée dans le premier plan principal (FIG.1.1) montre, par exemple, que les modalités x_4 et x_1 sont éloignées, nous remarquons que x_4 dépend de y_4, alors que x_1 dépend de y_1. La modalité y_3 dépend des modalités x_3 ou x_2.

La modalité y_2 est proche de l'origine, elle représente donc un profil moyen et n'est rattachée à aucune variable y_j, $\forall j = \{1, 3, 4\}$.

1.6.4.4 Eléments supplémentaires

Il est aussi possible, comme pour l'ACP, d'ajouter des éléments supplémentaires, illustratifs qui sont projetés sur les plans étudiés. Leur utilisation pour l'AFC est plus fréquente que pour l'ACP car il peut y avoir beaucoup de variables pour une étude donnée qui ne sont pas considérées dans cette analyse. Les projections sur les axes principaux des profils lignes ou des profils colonnes de ces éléments n'interviennent pas dans les calculs de ces axes.
Soit x_i une ligne supplémentaire, pour visualiser x_i sur un axe principal : on projette son profil sur cet axe (même chose pour une colonne supplémentaire).

En résumé, l'analyse factorielle des correspondances est la méthode privilégiée d'étude des relations entre deux variables qualitatives et l'une de ses principales propriétés est la faculté de représenter simultanément lignes et colonnes d'un tableau de contingence.

1.7 Analyse des Correspondances Multiples (ACM)

L'analyse des correspondances multiples (ACM) ou l'analyse factorielle des correspondances multiples (AFCM) est une extension de l'AFC dont le mot multiple signifie que l'on dispose de plusieurs caractéristiques (variables qualitatives) sur la population au lieu de 2 pour l'AFC.
On considère ici n individus décrits par p variables qualitatives à M catégories (modalités). Cette méthode est particulièrement bien adaptée à l'exploration d'enquêtes où les questions sont à réponses multiples.

Nous aurons donc ici trois familles d'éléments à étudier, les individus, les va-

riables et les modalités des variables. Afin d'établir un bilan des ressemblances entre individus, comme en ACP nous cherchons à répondre à des questions du type :

- Quels sont les individus qui se ressemblent ?
- Quelles sont ceux qui sont différents ?
- Existe-t-il des groupes homogènes d'individus ?
- Est-il possible de mettre en évidence une typologie des individus ?

Les mêmes types de questions se posent pour les variables et les modalités.

1.7.1 Tableau de départ (Tableau de codage condensé)

Le tableau de départ est souvent le tableau d'une enquête ou d'un sondage. Il se présente avec en lignes n individus enquêtés et en colonnes p questions posées à ces individus (variables). Chacune de ces questions possède plusieurs modalités de réponses. Le nombre total de modalités est M.
Les variables qualitatives peuvent être codées par un codage condensé qui attribue une valeur à chaque modalité. Les données peuvent donc être représentées sous la forme de la matrice X suivante :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_i^1 & x_i^j & x_i^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

où x_i^j est le codage condensé de l'individu i pour la variable j.

1.7.2 Tableau disjonctif complet

Les données précédentes ne peuvent pas être traitées par l'ACP ou l'AFC précédemment étudiées. On effectuera donc une transformation pour modifier la codification en nombres binaires. L'analyse utilise ensuite le même principe que l'AFC, en transformant le tableau disjonctif complet (qui est une autre représentation de nos données) en profils lignes et en profils colonnes.

Le tableau disjonctif complet représente les n individus en lignes, alors que les colonnes représentent les M modalités des variables (et non plus les p variables) (voir la matrice suivant) :

$$Y = \begin{pmatrix} x_1^1 & \dots & x_1^M \\ \vdots & \vdots & \vdots \\ x_i^1 & x_i^m & x_i^M \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^M \end{pmatrix}$$

A l'intersection de la ligne i avec la colonne m, la valeur x_i^m vaut 1 si l'individu i possède la modalité m et 0 sinon. Ce tableau porte le nom de disjonctif complet, car l'ensemble des valeurs x_i^m d'un même individu pour les modalités d'une même variable, comporte la valeur 1 une fois (complet) et une fois seulement (disjonctif). Chaque modalité m est relié à une variable j.
Ainsi le tableau brut suivant :

$$X = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \end{pmatrix}$$

correspondant à 5 observations de trois variables à 3, 2, 3 modalités respectivement engendre le tableau disjonctif Y à 5 lignes et 8 colonnes :

$$Y = \left(\begin{array}{ccc|cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

Notons m_j le nombre de modalités de la variable j. Ainsi $M = \sum_{j=1}^{p} m_j$. Nous avons donc les égalités suivantes :

$$\sum_{m=1}^{m_j} x_i^m = 1, \forall (i,j)$$

$$\sum_{m=1}^{M} x_i^m = p, \forall i$$

$$\sum_{i=1}^{n} x_i^m = n_m, \forall m$$

où n_m est l'effectif marginal de la modalité m.

$$\sum_{i=1}^{n}\sum_{m=1}^{M} x_i^m = np$$

– **Remarque**

En pratique, l'utilisateur n'a jamais besoin de calculer explicitement le tableau disjonctif complet (les programmes d'ACM acceptent en entrée le tableau de codage condensé et calculent eux-mêmes le tableau disjonctif complet).

L'ACM peut être vue comme une AFC du tableau disjonctif complet.

1.7.3 L'AFC du tableau disjonctif complet

En ACM, on traite le tableau disjonctif complet Y comme un tableau de contingence, nous allons donc considérer le tableau disjonctif complet en profils lignes et en profils colonnes. Pour se faire nous modifions ce tableau pour considérer les fréquences f_{im} données par :

$$f_{im} = \frac{x_i^m}{np}$$

On a donc :

$f_{im} = \frac{1}{np}$ si l'individu i possède la modalité m et $f_{im} = 0$ sinon.

De plus les marges sont données par :

$$f_{i.} = \sum_{m=1}^{M} \frac{x_i^m}{np} = \frac{1}{n} \quad et \quad f_{.m} = \sum_{i=1}^{n} \frac{x_i^m}{np} = \frac{n_m}{np}$$

Considérons maintenant le tableau des profils lignes L et le tableau des profils colonnes C dont leurs cases sont composées respectivement de :

$$\frac{f_{im}}{f_{i.}} = \frac{x_i^m}{p} \quad et \quad \frac{f_{im}}{f_{.m}} = \frac{x_i^m}{n_m}$$

- Chaque individu du nuage des points lignes est représenté par les modalités qu'il possède. C'est un point de $\mathbb{R}^m$ qui a pour coordonnée sur l'axe m la valeur $\frac{x_i^m}{p}$ muni d'un poids constant $f_{i.} = \frac{1}{n}$.
  Le barycentre G_n du nuage des points lignes a pour coordonnée $f_{.m} = \frac{n_m}{np}$ sur l'axe m.
- Chaque modalité peut être représentée par le profil colonne, *i.e.* par les valeurs prises par tous les individus pour la modalité considérée. Ainsi une modalité m est un point de l'espace $\mathbb{R}^n$ et a pour coordonnée $\frac{x_i^m}{n_m}$ sur l'axe i avec un poids de $f_{.m} = \frac{n_m}{np}$.
  Le barycentre G_m du nuage des points colonnes a pour coordonnée $f_{i.} = \frac{1}{n}$ sur l'axe i.

Définition 1.6. *La ressemblance entre deux individus est définie par les modalités de chacun des individus. Si deux individus présentent globalement les mêmes modalités, alors ils se ressemblent.*
La distance qui caractérise la ressemblance entre deux individus i et i′ est la distance de χ^2, elle est définie par :

$$d^2(i,i') = \sum_{m=1}^{M} \frac{1}{f_{.m}} (\frac{x_i^m}{p} - \frac{x_{i'}^m}{p})^2 = \frac{n}{p} \sum_{m=1}^{M} \frac{1}{n_m} (x_i^m - x_{i'}^m)^2$$

Cette expression est remarquable car $(x_i^m - x_{i'}^m)^2 = 1$ si un seul individu possède la modalité m et 0 sinon.
Le poids de la modalité m dans la distance est l'inverse de sa fréquence : $\frac{n}{n_m}$. Ainsi si un individu possède une modalité rare (n_m petit), il sera éloigné de tous les autres individus et du centre de gravité.([12])

- **Remarque**
  En pratique, on évite de conserver dans l'analyse des modalités rares et des variables ayant des nombres de modalités trés différents.

Définition 1.7. *La ressemblance entre deux modalités m et m′ est donnée par la distance :*

$$d^2(m,m') = \sum_{i=1}^{n} \frac{1}{f_{i.}} (\frac{x_i^m}{n_m} - \frac{x_i^{m'}}{n_{m'}})^2 = n \sum_{i=1}^{n} (\frac{x_i^m}{n_m} - \frac{x_i^{m'}}{n_{m'}})^2$$

Deux modalités sont proches si elles sont possédées par les mêmes individus, tandis que les modalités rares sont éloignées de toutes les autres et du centre de gravité G_m.

Remarque

En notant F le tableau des fréquences, on a :

$$L = D_n^{-1}F \quad et \quad C = D_m^{-1}F'$$

avec D_n^{-1} et D_m^{-1} sont les métriques de χ^2 relatives respectivement aux espaces $\mathbb{R}^n$ et $\mathbb{R}^m$, elles sont données par :

$$D_n^{-1} = \begin{pmatrix} \frac{1}{f_{1.}} & & \\ & \ddots & \\ & & \frac{1}{f_{n.}} \end{pmatrix} = \begin{pmatrix} n & & \\ & \ddots & \\ & & n \end{pmatrix} = nI_n$$

où I_n est la matrice d'identité d'ordre n.

et

$$D_m^{-1} = \begin{pmatrix} \frac{1}{f_{.1}} & & \\ & \ddots & \\ & & \frac{1}{f_{.M}} \end{pmatrix} = \begin{pmatrix} \frac{np}{n_1} & & \\ & \ddots & \\ & & \frac{np}{n_M} \end{pmatrix} = p \begin{pmatrix} \frac{n}{n_1} & & \\ & \ddots & \\ & & \frac{n}{n_M} \end{pmatrix} = p\Delta$$

avec

$$\Delta = \begin{pmatrix} \frac{n}{n_1} & & \\ & \ddots & \\ & & \frac{n}{n_M} \end{pmatrix}$$

1.7.3.1 Inertie totale

L'inertie totale du nuage des modalités et, par dualité, celle du nuage des individus en ACM dépend du nombre moyen de modalités par variables ($\frac{M}{p}$), elle vaut $\frac{M}{p} - 1$. De plus, elle est égale à $\lambda_1 + \lambda_2 + ... + \lambda_r$ où $r = min(n-1, M-p)$ est le nombre de valeurs propres non nulles. Le pourcentage d'inertie expliquée par un axe α est donc :

$$\frac{\lambda_\alpha}{\lambda_1 + \lambda_2 + ... + \lambda_r}$$

Remarque

En ACM, les pourcentages d'inertie expliqués par les axes sont par construction petits et ne peuvent donc pas être interprétés comme en AFC ou en ACP. Le nombre d'axes retenus pour l'interprétation ne peut pas être choisi à partir de ces pourcentages.

1.7.3.2 ACP des nuages des points lignes et des points colonnes

Nous allons effectuer une ACP des nuages des points individus et des points modalités (centrés). Reprenons donc les résultats vus en AFC :
On réalise l'ACP des profils lignes dans $\mathbb{R}^m$ avec :

- tableau de données $X = L = D_n^{-1}F = nF$.
- métrique $M = D_m^{-1} = p\Delta$.
- poids $D = D_n = \frac{1}{n}I_n$.

Les axes principaux sont les vecteurs propres de $MX'DX = np\Delta F'F$ et les composantes principales sont les vecteurs propres de $XMX'D = npF\Delta F'$ normalisés par $a'D_n^{-1}a$, où a est un vecteur propre de $npF\Delta F'$.

Nous allons suivre la même démarche pour les profils colonnes *i.e.* effectuer une ACP du nuage des points colonnes dans $\mathbb{R}^n$ avec :

- tableau de données $X = C = D_m^{-1}F' = p\Delta F'$.
- métrique $M = D_n^{-1} = nI_n$.
- poids $D = D_m = \frac{1}{p}\Delta^{-1}$.

Les axes principaux sont donc les vecteurs propres de $npF\Delta F'$ et les composantes principales sont les vecteurs propres de $np\Delta F'F$ normalisés par $b'D_m^{-1}b$, où b est un vecteur propre de $np\Delta F'F$.

- **Remarque**

  On remarque que :

$$MX'DX = CL$$

et

$$XMX'D = LC$$

1.7.3.3 Formules de transition et relations barycentriques

En appliquant les relations de transition de l'AFC, on obtient les deux relations fondamentales de l'ACM :

$$a_i = \frac{1}{\sqrt{\lambda}} \sum_{m=1}^{M} \frac{x_i^m}{p} b_m = \frac{1}{\sqrt{\lambda}} \frac{1}{p} \sum_{m=1}^{M} x_i^m b_m$$

et

$$b_m = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \frac{x_i^m}{n_m} a_i = \frac{1}{\sqrt{\lambda}} \frac{1}{n_m} \sum_{i=1}^{n} x_i^m a_i$$

La première formule s'interprète comme suit :

- A $\frac{1}{\sqrt{\lambda}}$ près la coordonnée d'un individu est égale à la moyenne arithmétique simple des coordonnées des catégories auxquelles il appartient.

La deuxième formule montre que :

- A $\frac{1}{\sqrt{\lambda}}$ près la coordonnée d'une catégorie m est égale à la moyenne arithmétique des coordonnées des n_m individus de cette catégorie.([13])

A un coefficient près ($\frac{1}{\sqrt{\lambda}}$), sur chaque axe, un individu est au barycentre des modalités qu'il possède et une modalité est au barycentre des individus qui la possèdent. Cette double propriété barycentrique est particulièrement simple du fait que x_i^m vaut 0 ou 1. Elle suffit presque pour interpréter les graphiques, ce qui fait d'ailleurs de l'ACM la méthode factorielle dont les graphiques sont les plus faciles à interpréter.([10])

1.7.3.4 Biplot

Les relations barycentriques données ci-dessus donnent trois modes de représentation simultanée des individus et des modalités :

1. Les individus au barycentre des modalités.

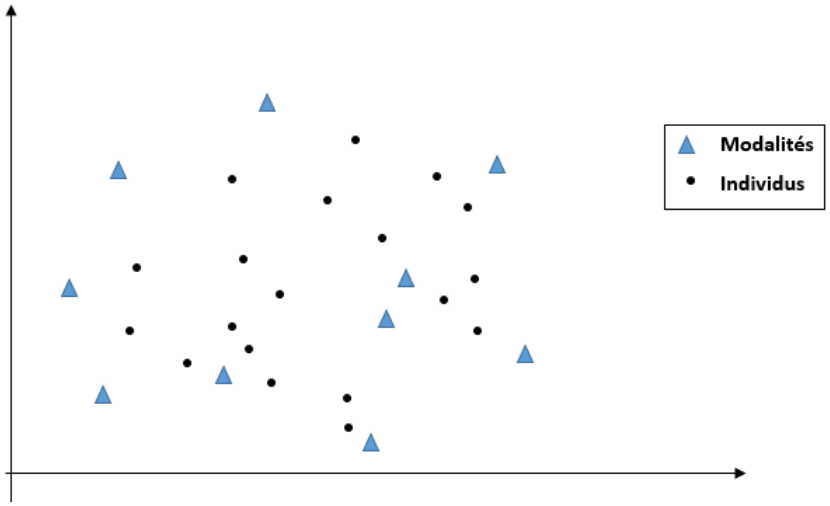


FIG. 1.2 – Représentation des individus au milieu des modalités (Représentation barycentrique)

2. Les modalités au barycentre des individus.

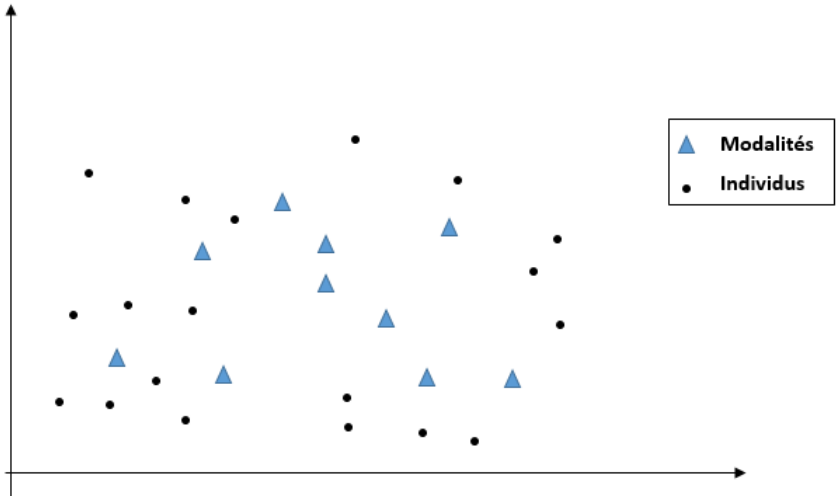


FIG. 1.3 – Représentation des modalités au milieu des individus (Représentation barycentrique)

3. ▶ Les individus sont au centre des modalités qu'ils ont choisis.

▶ Les modalités sont au centre des individus qui les ont choisis.

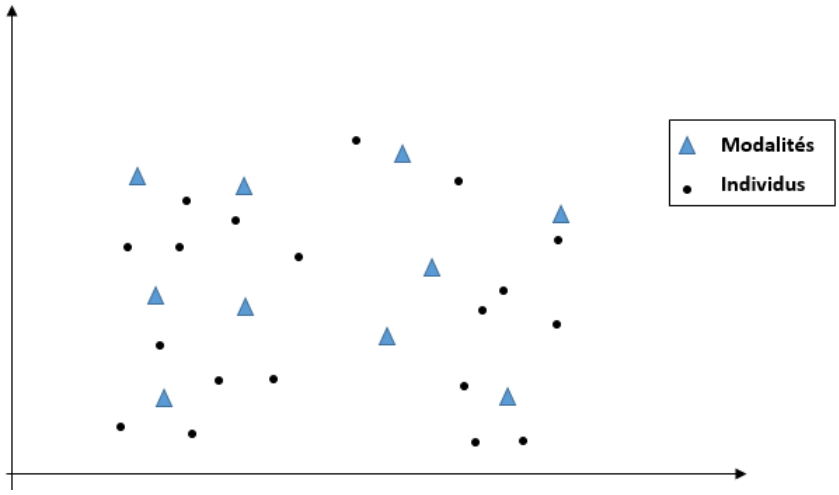


FIG. 1.4 – Représentation simultanée des individus et des modalités (Représentation pseudo-barycentrique)

1.7.3.5 Interprétation

Nous avons vu que deux individus se ressemblent s'ils présentent globalement les mêmes modalités. Cette ressemblance se traduit par une proximité des individus dans l'espace de projection choisi pour la représentation simultanée. De même si deux modalités d'une même variable sont proches dans l'espace de projection, ceci se traduit par une ressemblance entre les groupes d'individus qui les ont choisies. La proximité de deux modalités de variables différentes s'interprète en terme d'association. Ainsi une modalité apparaît du côté des modalités avec lesquelles elle s'associe le plus ; deux modalités apparaissent proches si elles s'associent de la même manière aux autres modalités.
En ce qui concerne la proximité entre modalités et individus, l'interprétation peut se faire en considérant les modalités comme barycentre de classe d'individus.

1.7.3.6 Aides à l'interprétation

Rappelons que sur la représentation simultanée, les nuages des points lignes et des points colonnes ne sont pas dans les mêmes espaces. Il est donc important d'avoir

recours à des indicateurs sur la qualité de représentation pour l'interprétation de l'ACM. Ces indicateurs sont les mêmes que ceux de l'ACP déjà donnés :

- Choisir le nombre d'axes de projection : on retient les axes associés à des valeurs propres supérieures à $\frac{1}{p}$.
- Etudier les valeurs propres qui représentent l'inertie de chaque axe.
- Etudier la contribution des lignes et des modalités.
- Etudier la contribution des variables en sommant les contributions des modalités d'une variable pour un facteur donné.
- Etudier les coordonnées des modalités et des individus actifs : celles et ceux qui déterminent les axes.
- Etudier les coordonnées des variables, des modalités et des individus supplémentaires s'il y en a.

Chapitre 2

Analyse discriminante

2.1 Introduction

L'analyse discriminante est une technique d'analyse des données connue sous l'abréviation AD. Elle est utilisée dans le cadre de la modélisation d'une variable qualitative Y à K catégories (modalités) dite variable à expliquer (ou variable endogène ou encore variable à prédire), à partir de p variables quantitatives appelées variables explicatives (ou variables exogènes ou encore prédicteurs).
On peut considérer l'analyse discriminante comme une extension du problème de la régression au cas où la variable à expliquer est qualitative ; on verra d'ailleurs que dans le cas de deux catégories, on peut se ramener exactement à une régression linéaire multiple. Elle peut aussi être vue comme un cas particulier de l'analyse en composantes principales. En effet, son aspect descriptif décrit dans le paragraphe 2.3 fait appel à des calculs d'axes principaux appelés axes factoriels discriminants.

2.2 Données et notations

Nous disposons de n individus (ou observations) décrits par p variables et répartis en K classes (groupes) données par la variable qualitative Y. Les K classes sont a priori connues. La variable qualitative Y possède donc K modalités (ce sont les modalités de Y qui définissent les classes). En notant X la matrice des p variables explicatives et A le tableau disjonctif associé à la variable qualitative Y, on obtient ainsi la matrice

de données suivante :

$$\begin{array}{c} \\ 1 \\ 2 \\ \\ \\ \\ \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccc|cccc} 1 & 2 & \dots & K & 1 & 2 & \dots & p \end{array} \\ \left(\begin{array}{cccc|cccc} 1 & 0 & \dots & 0 & & & & \\ & & & & & & & \\ & & & & & & & \\ & & A & & & X & & \\ & & & & & & & \\ & & & & & & & \\ 0 & 0 & \dots & 1 & & & & \end{array}\right) \end{array}$$

Notons :

. x_i^j la valeur de la jème variable explicative mesurée sur le ième individu.

. $X_i = (x_i^1, ..., x_i^p)' \in \mathbb{R}^p$ une ligne de X décrivant le ième individu.

. $X^j = (x_1^j, ..., x_n^j)' \in \mathbb{R}^n$ une colonne de X décrivant la jème variable.

. G_k le groupe des individus de l'échantillon qui possèdent la modalitè k.

. $n_k = card(G_k)$ le nombre d'individus qui possèdent la modalité k.

Les ensembles $\{X_i \setminus i = \overline{1,n}\} \subset \mathbb{R}^p$ et $\{X^j \setminus j = \overline{1,p}\} \subset \mathbb{R}^n$ désignent respectivement les nuages des individus et des variables.

La variable Y permet de définir, compte-tenu des K modalités, une partition de l'ensemble des individus en K sous-ensembles $G_1, G_2, \dots, G_K$, l'individu i appartenant à G_k si c'est la kième modalité de la variable qualitative qui est réalisée. On peut ainsi considérer K sous-nuages $N_1, N_2 \dots, N_K$ tels que :

$$N_k = \{X_i \backslash i \in G_k\} \subset \mathbb{R}^p$$

Si les n individus sont affectés des poids $p_1, ..., p_n$, tels que $\forall i = \overline{1,n}$, $p_i > 0$ et $\sum_{i=1}^{n} p_i = 1$ alors le poids de chaque groupe G_k est :

$$P_k = \sum_{i \in G_k} p_i$$

En général, on prend $p_i = \frac{1}{n}$ et donc $P_k = \frac{n_k}{n}$. On a alors les définitions suivantes :

– Le centre de gravité global est le vecteur de $\mathbb{R}^p$ défini par :

$$g = \sum_{i=1}^{n} p_i x_i = \frac{1}{n} \sum_{i=1}^{n} x_i$$

– Le centre de gravité du groupe G_k est le vecteur de $\mathbb{R}^p$ défini par :

$$g_k = \frac{1}{P_k} \sum_{i \in G_k} p_i x_i = \frac{1}{n_k} \sum_{i \in G_k} x_i$$

– La matrice $p \times p$ de variance-covariance globale est définie par :

$$V = \sum_{i=1}^{n} p_i (x_i - g)(x_i - g)' = \frac{1}{n} \sum_{i=1}^{n} (x_i - g)(x_i - g)'$$

– La matrice $p \times p$ de variance-covariance du groupe G_k est définie par :

$$V_k = \frac{1}{P_k} \sum_{i \in G_k} p_i (x_i - g_k)(x_i - g_k)' = \frac{1}{n_k} \sum_{i \in G_k} (x_i - g_k)(x_i - g_k)'$$

– La matrice $p \times p$ de variance-covariance intra-groupe est définie par :

$$W = \sum_{k=1}^{K} P_k V_k = \sum_{k=1}^{K} \frac{n_k}{n} V_k$$

– La matrice $p \times p$ de variance-covariance inter-groupe est définie par :

$$B = \sum_{k=1}^{K} P_k (g_k - g)(g_k - g)' = \sum_{k=1}^{K} \frac{n_k}{n} (g_k - g)(g_k - g)'$$

Remarques

1. La matrice B est la matrice de variance-covariance des K centres de gravités g_k pondérés par P_k et la matrice W est la somme pondérée des covariances inter-groupe.
2. En règle générale, W est inversible tandis que B ne l'est pas, car les K centres de gravité sont dans un sous-espace de dimension K-1 de $\mathbb{R}^p$ (si $p > K - 1$ ce qui est généralement le cas), alors que la matrice B est de taille p.

On a la relation suivante :

$$g = \sum_{k=1}^{K} P_k g_k = \sum_{k=1}^{K} \frac{n_k}{n} g_k$$

2.3 L'AD descriptive et l'AD prédictive

On distingue deux aspects en analyse discriminante :

1. L'analyse discriminante descriptive :

L'analyse discriminante descriptive est une technique de statistique exploratoire. Elle consiste à chercher les combinaisons linéaires de variables qui permettent de séparer le mieux possible les K classes et donner une représentation graphique (ainsi qu'en ACP), qui rende compte au mieux de cette séparation. Ces combinaisons linéaires sont appelées *fonctions linéaires discriminantes*. Il s'agit donc d'une étape de discrimination des classes.

L'analyse discriminante descriptive est une technique descriptive car elle propose une représentation graphique qui permet de visualiser les proximités entre les observations, appartenant a un même groupe ou non.

C'est aussi une technique explicative car nous avons la possibilité d'interpréter les axes principaux, combinaisons linéaires des variables initiales et ainsi comprendre les caractéristiques qui distinguent les différents groupes.

Remarque

Contrairement à l'analyse discriminante prédictive, l'analyse discriminante descriptive ne repose sur aucune hypothèse probabiliste. Il s'agit essentiellement d'une méthode géométrique.

2. L'analyse discriminante prédictive :

Un nouvel individu se présente pour lequel on connait les valeurs des prédicteurs. Il s'agit alors de décider dans quelle classe il faut l'affecter. C'est un problème de classement par opposition au problème de classification qui est la construction de classes les plus homogènes possibles dans un échantillon.

Ces deux aspects correspondent donc à la distinction entre les méthodes géométriques qui sont essentiellement descriptives et qui ne reposent que sur des notions de distance et les méthodes probabilistes.

2.4 Principe de l'analyse discriminante

2.4.1 La discrimination

L'idée du principe de la discrimination repose sur le fait que la discrimination visuelle est plus aisée si :

- Les centres de gravité de chaque sous-nuage appartenant à une seule classe sont éloignés.
- Les sous-nuages appartenant à une seule classe sont les plus homogènes possibles autour de ces centres de gravité.

Pour se faire il faut maximiser les variances inter-goupe (entre les groupes) et minimiser les variances intra-groupe (à l'intérieur des groupes). Nous parlons également de variances externes et internes.

2.4.1.1 Formule de décomposition de Huygens

Proposition 2.1. *L'inertie totale du nuage des points individus est égale à la somme de l'inertie inter-groupe et de l'inertie intra-groupe.*
Cette proposition s'énonce également par le fait que la covariance totale du nuage est la somme de la covariance inter-groupe et de la covariance intra-groupe :

$$V = B + W$$

([16])

Cette proposition se démontre aisément et constitue une généralisation de la relation classique : variance totale = moyenne des variances + variance des moyennes.

2.4.1.2 Analyse factorielle discriminante

Le problème est de déterminer un vecteur u normé de $\mathbb{R}^p$ (appelé *facteur discriminant*), engendrant un axe Δ_u (appelé *axe discriminant*) passant par le centre de gravité g et tel que l'inertie des sous-nuages des individus n_k projetés sur Δ_u soit maximale (inertie inter-groupe $u'Bu$) et chaque sous-nuage soit groupé donc l'inertie intra-groupe $u'Wu$ soit minimale. En effet, B mesure la dispersion des centres de

gravité le long de l'axe Δ_u et W mesure la dispersion moyenne à l'intérieur des k groupes le long de l'axe Δ_u.

Le problème revient donc à trouver u tel que :

$$\frac{u'Bu}{u'Wu}$$

soit maximum.

La propriété V=B+W entraîne l'équivalence suivante :

$$max\frac{u'Bu}{u'Wu} \Longleftrightarrow max\frac{u'Bu}{u'Vu}$$

La fonction $\frac{u'Bu}{u'Vu}$ à rendre maximum est inchangée si u est remplacé par αu, α étant un scalaire quelconque. Par conséquent :

$$max\frac{u'Bu}{u'Vu} \Longleftrightarrow max\ u'Bu$$

sous la contrainte $u'Vu = 1$ (u normé).
La solution est obtenue en utilisant le Lagrangien $L(u)$ et en annulant sa dérivée :

$$L(u) = u'Bu - \lambda(u'Vu - 1)$$

$$\frac{\partial L(u)}{\partial u} = 0 \Longrightarrow Bu = \lambda Vu$$

Si V est une matrice inversible (ce qui est le cas en général), alors :

$$V^{-1}Bu = \lambda u$$

Ainsi u est le vecteur propre de $V^{-1}B$ associé à la plus grande valeur propre λ.

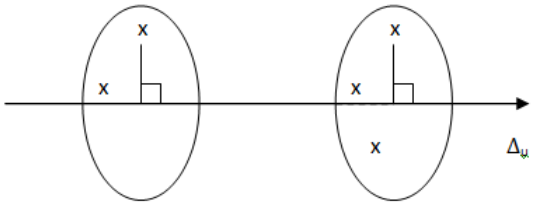
Proposition 2.2. *Le premier axe factoriel discriminant Δ_{u_1} est engendré par le vecteur propre normé u_1 de $V^{-1}B$ correspondant à la plus grande valeur propre λ_1. De même, le vecteur propre de $V^{-1}B$ relatif à la deuxième valeur propre λ_2 ($\lambda_2 \leq \lambda_1$) constitue le deuxième axe factoriel discriminant Δ_{u_2} et ainsi de suite pour chacun des vecteurs propres successifs.*

Définition 2.1. *On appelle pouvoir discriminant de l'axe Δ_{u_h} la valeur propre λ_h associée.*
La valeur λ_h est d'autant plus grande (proche de 1) que la discrimination est satisfaisante.

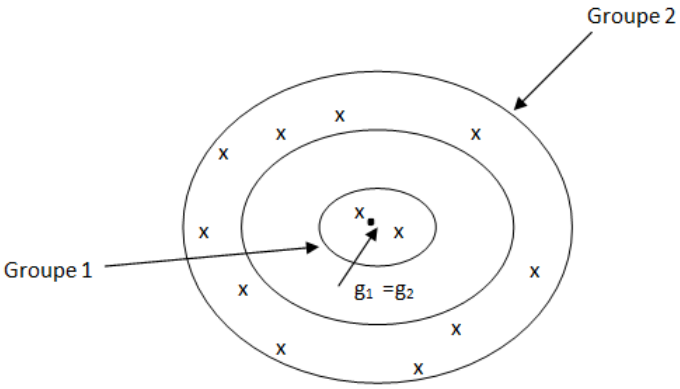
Interprétation des valeurs propres

Notons d'abord qu'on a toujours $0 \leq \lambda \leq 1$ et que le nombre des valeurs propres non nulles, donc d'axes discriminants, est égal à $K-1$ dans le cas habituel où $n > p > K$ et où les variables ne sont pas liées par des relations linéaires.

- $\lambda = 1$ correspond au cas suivant :
  En projection sur Δ_u les dispersions intra-groupe sont nulles. Les K sous-nuages sont donc chacun dans un hyperplan orthogonal à Δ_u.
  Il y a évidemment une discrimination parfaite si les centres de gravité se projettent en des points différents (voir la figure suivante) :



- $\lambda = 0$ correspond au cas où le meilleur axe ne permet pas de séparer les centres de gravité g_k. C'est le cas où ils sont confondus, les nuages sont donc concentriques et aucune séparation linéaire n'est possible (voir la figure suivante) :



Remarque

L'interprétation de ces $K-1$ valeurs propres diffère de celle des valeurs propres obtenues lors d'une analyse factorielle faite sur un ensemble non partitionné a priori (analyse en composantes principales ou analyse des correspondances). En effet, dans ce dernier cas, la somme des q premières valeurs propres indique l'inertie expliqué par les q premiers axes principaux, tandis que la somme de plusieurs valeurs propres dans une analyse factorielle discriminante n'aurait pas de sens.

Une ACP particulière

L'analyse factorielle discriminante est une analyse en composantes principales particulière d'un ensemble de n points individus X_i (de poids respectifs p_i tels que $\sum_{i=1}^{n} p_i = 1$) repérés dans l'espace $\mathbb{R}^p$ muni de la métrique M, dont le principe général est de déterminer un axe Δ_u engendré par un vecteur u normé passant par le centre de gravité du nuage des points individus, tel que la variance de l'ensemble des projections $p_u(X_i)$ des points X_i sur Δ_u soit maximum. Cela conduit à la résolution de l'équation $VMu = \lambda u$ avec $u'Mu = 1$, où V est la matrice des variances-covariances des points X_i.

En effet, $Bu = \lambda Vu$ avec $u'Vu = 1$. En posant $v = Vu$, on a $BV^{-1}v = \lambda v$ avec $v'V^{-1}v = 1$. Par conséquent, effectuer une analyse factorielle discriminante revient à faire une analyse en composantes principales à partir de l'ensemble G des centres de gravité $\{g_1, ..., g_k, ..., g_K\}$, munis des poids respectifs $\{P_1, ..., P_k, ..., P_K\} = \{\frac{n_1}{n}, ..., \frac{n_k}{n}, ..., \frac{n_K}{n}\}$ repérés dans l'espace $\mathbb{R}^p$ muni de la métrique

V^{-1}.

Comme en ACP, on pourra interpréter les p variables (dites variables discriminantes) au moyen d'un cercle des corrélations. A la différence de l'ACP, trois types de corrélation sont possible : la corrélation globale entre les axes et les variables initiales, la corrélation intra-groupe calculée à l'intérieur des groupes et la corrélation inter-groupe calculée à partir des centres de gravité des groupes pondérés par leurs poids.

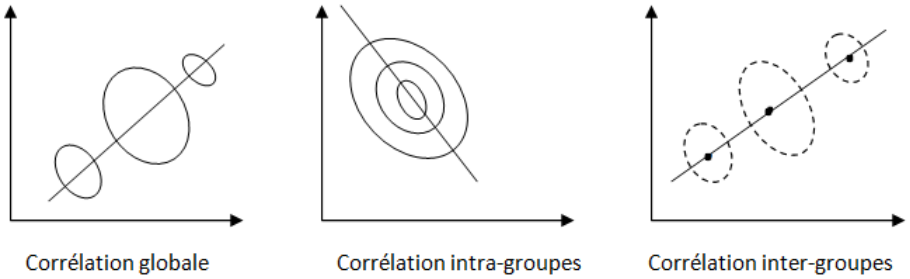


FIG. 2.1 – Les trois types de corrélations.

Equivalance des métriques V^{-1} et W^{-1}

Les vecteurs propres de $W^{-1}B$ sont les mêmes que ceux de $V^{-1}B$. En effet

$$V^{-1}Bu = \lambda u \Longleftrightarrow Bu = \lambda Vu$$

En remplaçant V par $B + W$ on a :

$$Bu = \lambda Bu + \lambda Wu \Longleftrightarrow Bu = \frac{\lambda}{\lambda - 1} Wu \Longleftrightarrow W^{-1}Bu = \frac{\lambda}{\lambda - 1} u$$

On vient de montrer que si λ et u sont valeur et vecteur propres de $V^{-1}B$ alors $\mu = \frac{\lambda}{\lambda - 1}$ et u sont respectivement valeur et vecteur propres de $W^{-1}B$. L'utilisation de V^{-1} ou de W^{-1} comme métrique est donc indifférent.

Remarques

1. La métrique W^{-1} est appelée métrique de *Mahalanobis*.
2. Dans les procédures de quelques logiciels (SAS par exemple) concernant l'analyse discriminante, les calculs sont faits en utilisant la métrique W^{-1} (ces

logiciels fournissent la valeur propre μ).

Détermination des vecteurs propres de $W^{-1}B$

Il faut diagonaliser la matrice $W^{-1}B$ qui n'est pas a priori symétrique.

La matrice B a pour expression :

$$\sum_{k=1}^{K} \frac{n_k}{n}(g_k - g)(g_k - g)'$$

de terme général :

$$b_{jj'} = \sum_{k=1}^{K} \frac{n_k}{n}(g_{kj} - g_j)(g_{kj'} - g_{j'})$$

où g_{kj} est la jième composante de g_k.

B s'écrit en fonction de la matrice C de terme général :

$$c_{jk} = \sqrt{\frac{n_k}{n}(g_{kj} - g_j)}$$

sous la forme :

$$B = CC'$$

Les matrices $W^{-1}B$ (d'ordre p) et $C'W^{-1}C$ (d'ordre k) ont les mêmes valeurs propres et leurs vecteurs propres respectifs sont liés par la relation :

$$u = W^{-1}Cw$$

En effet, soit λ une valeur propre de la matrice $C'W^{-1}C$ et w le vecteur propre associé :

$$C'W^{-1}Cw = \lambda w \Longrightarrow CC'W^{-1}Cw = \lambda Cw \Longrightarrow BW^{-1}Cw = \lambda Cw$$

et donc :

$$W^{-1}BW^{-1}Cw = \lambda W^{-1}Cw \Longleftrightarrow W^{-1}Bu = \lambda u$$

On se ramène ainsi à la diagonalisation d'une matrice symétrique d'ordre k. ([11])

Représentations graphiques ([2])

Supposons qu'on se contente des deux premiers axes discriminants pour représenter les individus et les variables sur le plan discriminant $(\Delta_{u_1}, \Delta_{u_2})$.

1. Description dans $\mathbb{R}^p$

Dans le plan discriminant $P = (\Delta_{u_1}, \Delta_{u_2})$, on représente :

– Les points individus X_i à l'aide de leurs coordonnées :

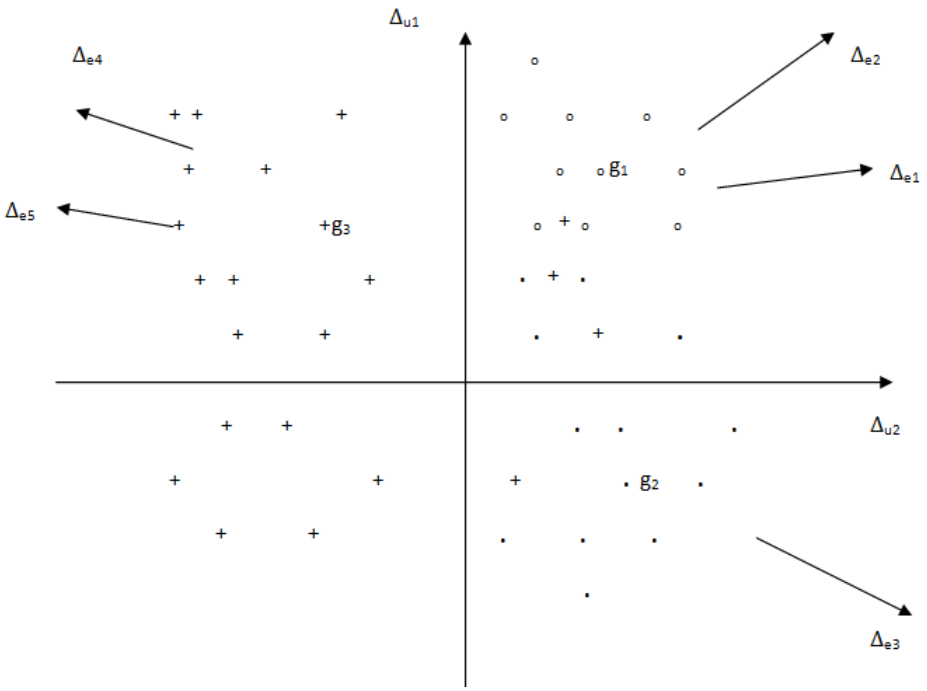
$$c^j = u_j X_i, \quad j = 1, 2$$

– Les centres de gravités g_k à l'aide de leurs coordonnées :

$$u_j g_k, \quad j = 1, 2$$

On oriente le plan discriminant P en faisant figurer la projection des axes Δ_{e_i} associés aux variables initiales, le vecteur de base e_i a pour coordonnées, dans le système des deux premiers axes discriminants :

$$u_j e_i, \quad j = 1, 2$$



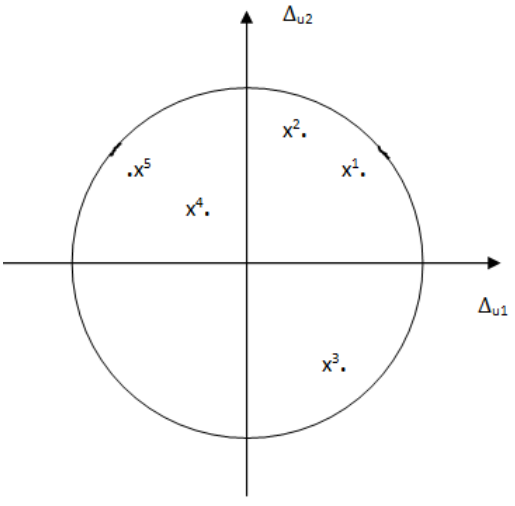
Le plan discriminant ci-dessus décrit la dépendance entre une variable qualitative Y à trois modalités 1, 2, 3 et un ensemble de cinq variables quantitatives.
La partition induite par la variable Y est mise en évidence en associant un même symbole aux individus ($_+$, $_o$, $\cdot$) appartenant à une même classe : on a ainsi une idée visuelle, compte-tenu des dispersions autour de chacun des centres de gravité, de la séparation, dans le plan discriminant, entre les groupes d'individus associés aux différentes modalités de la variable Y.
Ici les trois groupes sont assez bien séparés, il est possible de retrouver à l'aide des deux premiers facteurs discriminants la modalité prise par la variable qualitative Y.

2. Description dans $\mathbb{R}^n$

La projection des variables est calculée en fonction des axes discriminants du nuage des points individus. La coordonnée d'une variable X^j sur l'axe u_j est donnée par :

$$F^j = r(X^j, u_j)$$

Les variables X^1, X^2, X^3 et X^5 (voir la figure ci-dessous) interviennent fortement dans la description des individus contrairement à X^4.
Si on revient au graphique précédant, on constate d'une part que les variables X^3 et X^5 permettent de bien séparer les individus prenant les modalités 2 et 3 de la variable Y et que d'autre part, X^1 et X^2 permettent de repérer les individus prenant la modalité 1.



2.4.1.3 Cas particulier de deux groupes - Equivalence entre régression multiple et analyse discriminante

Lorsqu'il n'y a que deux groupes ($k = 2$) notés G_1 et G_2, le rang de la matrice $W^{-1}B$ est égal à 1, cette matrice n'a qu'une seule valeur propre non nulle et un seul vecteur propre associé, donc un seul axe discriminant, qui correspond à la *fonction linéaire discriminante de Fisher*, déterminé par la droite passant par les centres de gravité des deux groupes g_1 et g_2. La matrice de variance-covariance inter-groupe B, dans ce cas, s'écrit :

$$B = \frac{n_1}{n}(g_1 - g)(g_1 - g)' + \frac{n_2}{n}(g_2 - g)(g_2 - g)'$$

$$\Longleftrightarrow B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)' \quad (car \;\; g = \frac{n_1 g_1 + n_2 g_2}{n})$$

de terme général :

$$b_{jj'} = \frac{n_1 n_2}{n^2}(g_{1j} - g_{2j})(g_{1j'} - g_{2j'})$$

Si λ et u désignent l'unique valeur propre et l'unique vecteur propre de $W^{-1}B$, alors

$$u = W^{-1}(g_1 - g_2)$$

et

$$\lambda = \frac{n_1 n_2}{n^2}(g_1 - g_2)'W^{-1}(g_1 - g_2)$$

En effet, on a :

$$W^{-1}Bu = \lambda u \Longleftrightarrow W^{-1}\frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)'u = \lambda u$$

En remplaçant u par sa valeur on a :

$$\frac{n_1 n_2}{n^2}W^{-1}(g_1 - g_2)(g_1 - g_2)'W^{-1}(g_1 - g_2) = \lambda W^{-1}(g_1 - g_2)$$

La quantité $\frac{n_1 n_2}{n^2}(g_1 - g_2)'W^{-1}(g_1 - g_2)$ est un scalaire et c'est la valeur propre λ, qui n'est autre que *le D^2 de Mahalanobis* au coefficient $\frac{n_1 n_2}{n^2}$ près.

– **Remarque**

Le D^2 de Mahalanobis mesure la distance entre deux groupes, plus précisément il mesure la distance entre les centres des groupes pour la métrique W^{-1}, il vaut :

$$D^2 = (g_1 - g_2)'W^{-1}(g_1 - g_2)$$

B peut se mettre sous la forme $B = CC'$ où C est une matrice colonne (p,1) de terme général :

$$c_j = \frac{\sqrt{n_1 n_2}}{n}(g_{1j} - g_{2j})$$

Nous avons donc :

$$W^{-1}Bu = \lambda u \Longleftrightarrow W^{-1}CC'u = \lambda u \Longleftrightarrow C'W^{-1}CC'u = \lambda C'u$$

Ainsi l'unique valeur propre de $W^{-1}B$ est $\lambda = C'W^{-1}C$, dont le vecteur propre associé est $u = W^{-1}C$.

λ est appelée distance généralisée entre les deux groupes ou encore distance de Mahalanobis.

Dans ce cas de deux groupes, l'AD est équivalente à la régression multiple.

Rappelons que le modèle de régression multiple s'écrit sous la forme :

$$y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + \varepsilon_i, \quad i = \overline{1, n} \tag{2.1}$$

où

- Les x_i^j sont des nombres connus, non aléatoires.
- Les paramètres β_j sont inconnus, mais non aléatoires, ce sont les paramétres à estimer (coefficients de la régression linéaire).
- Les ε_i sont des variables aléatoires inconnues, elles correspondent aux erreurs de l'estimation.
- Les y_i sont donc aléatoires.

En utilisant l'écriture matricielle de (2.1) nous obtenons la définition suivante :

Définition 2.2. *Un modèle de régression linéaire multiple est défini par une équation de la forme :*

$$Y = X\beta + \varepsilon$$

$$\Longleftrightarrow \begin{pmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_1^j & x_1^p \\ \\ 1 & x_i^1 & x_i^j & x_i^p \\ \\ 1 & x_n^1 & x_n^j & x_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \\ \beta_j \\ \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$$

où :

- *Y est un vecteur aléatoire de dimension n, c'est le vecteur des variables à expliquer.*
- *X est une matrice de taille $n \times (p+1)$ connue, appelée matrice du plan d'expérience.*
- *β est le vecteur de dimension $p+1$ des paramètres inconnus du modèle.*
- *ε est le vecteur de dimension n des erreurs.*

Les hypothèses concernant le modèle sont :

$$(H) \begin{cases} (H_1): & \text{rang(X)=p+1} \\ (H_2): & \text{E}(\varepsilon)=0_{\mathbb{R}^n}, \ \text{var}(\varepsilon)=\sigma^2 I_n \end{cases}$$

L'hypothèse (H_2) signifie que les erreurs sont centrées, de même variance σ^2 (homoscédasticité) et non corrélées entre elles.([14])

L'objectif : estimer les p+1 paramètres $\beta_0, ..., \beta_p$.

Estimateurs des Moindres Carrés Ordinaires

On cherche l'équation de la droite de régression en dimension p+1 pour laquelle les erreurs quadratiques (ε_n) sont les plus faibles.

Définition 2.3. *L'estimateur des moindres carrés $\widehat{\beta}$ est défini comme suit :*

$$\arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\beta\in\mathbb{R}^p} (y_i - \beta_0 - \beta_1 x_i^1 - ... - \beta_p x_i^p)^2 = \arg\min_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 = \arg\min_{\beta\in\mathbb{R}^p} \|\varepsilon\|^2$$

Proposition 2.3. *L'estimateur $\widehat{\beta}$ des Moindres Carrés Ordinaires a pour expression :*

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

– **Remarque**

L'hypothèse (H_1) assure que la matrice $X'X$ est bien inversible.

Preuve

On cherche $\beta \in \mathbb{R}^p$ qui minimise la fonction :

$$S(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta) = \beta'(X'X)\beta - 2Y'X\beta + \|Y\|^2$$

Or S est de type quadratique en β, avec $X'X$ symétrique définie positive, donc le problème admet une unique solution $\widehat{\beta}$: c'est le point où le gradient de S est nulle. Par dérivation (vectorielle...) par rapport à β on a :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2XX'\beta$$

Chercher à annuler cette dérivée revient à choisir l'estimateur $\widehat{\beta}$ tel que $X'X\widehat{\beta} = X'Y$ et donc :

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

L'estimateur $\widehat{\beta}$ peut s'écrire sous la forme :

$$\widehat{\beta} = \frac{1}{n}V^{-1}X'Y \quad (car \frac{1}{n}X'X = V)$$

Si Y est défini par : $\begin{cases} y_i = +\sqrt{\frac{n_2}{n_1}} & \text{si } i \in G_1 \\ y_i = -\sqrt{\frac{n_1}{n_2}} & \text{si } i \in G_2 \end{cases}$

alors $\frac{1}{n}X'Y = C$, d'où $\widehat{\beta} = V^{-1}C = W^{-1}C$.

On conclut que le vecteur $\widehat{\beta}$ des coefficients de la régression linéaire coïncide à un facteur α prés avec la forme linéaire discriminante $u = W^{-1}C$. Ce facteur α a pour expression :

$$\alpha = \frac{\frac{\sqrt{n_1 n_2}}{n}}{(1 + C'W^{-1}C)}$$

– **Remarque**

Il faut prendre garde au fait que les hypothèses habituelles de la régression ne sont pas vérifiées, bien au contraire : ici Y est non aléatoire et X l'est. Il ne faudra donc pas utiliser les statistiques usuelles fournies par un programme de

régression, en particulier les erreurs standards des coefficients et les niveaux de signification.

2.4.2 Règle d'affectation d'un nouvel individu à l'un des K groupes

Tout ce qui précède concernait la discrimination à but descriptif, nous allons maintenant nous intéresser au problème décisionnel, *i.e.* au problème de classement proprement dit.

2.4.2.1 Règles géométriques

La méthode classique consiste à comparer les distances d'un nouvel individu aux centres des groupes, distances mesurées avec une certaine métrique (la métrique W^{-1}). Cette métrique s'introduit naturellement dans l'analyse discriminante dont le but est de mettre en évidence des facteurs tels que les valeurs de ceux-ci soient aussi différentes que possible pour les individus appartenant à des groupes différents.

Règle de Mahalanobis-Fisher

Etant donné un nouvel individu a, nous voulons savoir à quel groupe il appartient. Nous supposons ici, que l'individu a appartient surement à l'un des K groupes.
La règle de Mahalanobis-Fisher consiste à utiliser la métrique W^{-1} (ou V^{-1} ce qui est équivalent), nous allons donc calculer la distance, pour la métrique W^{-1}, de a au centre g_k de la classe G_k :

$$d^2(a, G_k) = (a - g_k)' W^{-1} (a - g_k) \tag{2.2}$$

puis nous décidons d'affecter a à la classe G telle que :

$$d^2(a, G) = \min_{k=\overline{1,K}} d^2(a, G_k) \tag{2.3}$$

En développant les quantités définies ci-dessus en (2.2) on trouve :

$$d^2(a, G_k) = a' W^{-1} a + g_k' W^{-1} g_k - 2 g_k' W^{-1} a$$

Elles sont des fonctions quadratiques de a, mais elles ont toutes en commun le terme carré $a'W^{-1}a$ qui ne dépend pas de k, on pourra donc comparer les fonctions linéaires discriminantes de a, définies ci-dessous, relatives à chaque groupe G_k :

$$f_{G_k}(a) = d^2(a, G_k) - a'W^{-1}a = g_k'W^{-1}g_k - 2g_k'W^{-1}a \tag{2.4}$$

Dans ces conditions la règle de décision définie en (2.3) devient :
On décide d'affecter a à la classe G telle que :

$$f_G(a) = \min_{k=\overline{1,K}} f_{G_k}(a)$$

Cas de deux groupes

Dans ce cas il n'y a que deux fonctions $f_{G_1}(a)$ et $f_{G_2}(a)$ à comparer. La règle de décision est alors la suivante :
On affecte a au groupe G_1 si

$$f_{G_1}(a) > f_{G_2}(a)$$

i.e.

$$f_{G_1}(a) - f_{G_2}(a) > 0$$

En remplaçant $f_{G_k}(a)$ par sa valeur indiquée en (2.4), on a :

$$g_2'W^{-1}g_2 - g_1'W^{-1}g_1 - 2g_2'W^{-1}a + 2g_1'W^{-1}a > 0 \tag{2.5}$$

Les deux derniers termes s'écrivent :

$$2(g_1 - g_2)'W^{-1}a$$

et on remarque que :

$$g_1'W^{-1}g_1 - g_2'W^{-1}g_2 = (g_1 - g_2)'W^{-1}(g_1 + g_2)$$

D'où (2.5) devient, après division par 2 :

$$(g_1 - g_2)'W^{-1}a - \frac{1}{2}(g_1 - g_2)'W^{-1}(g_1 + g_2) > 0$$

La règle de décision devient donc :
On affecte a au groupe G_1 si

$$(g_1 - g_2)'W^{-1}a > \frac{1}{2}(g_1 - g_2)'W^{-1}(g_1 + g_2) \tag{2.6}$$

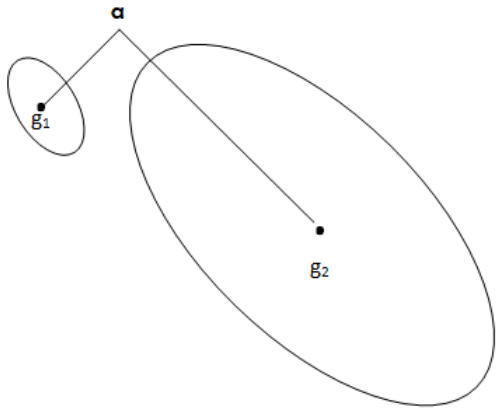
On affecte a au groupe G_2 si

$$(g_1 - g_2)'W^{-1}a < \frac{1}{2}(g_1 - g_2)'W^{-1}(g_1 + g_2) \tag{2.7}$$

On notera que le terme de gauche de (2.6) ou (2.7) n'est autre que la fonction linéaire discriminante de Fisher.([1])

2.4.2.2 Insuffisance des règles géométriques

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles : rien ne justifie alors l'usage de la même métrique pour les différents groupes.
En effet, si l'on considère la figure ci-dessous, bien que a soit plus proche de g_1 que de g_2 au sens habituel, il est plus naturel d'affecter a à la deuxième classe qu'à la première dont le pouvoir d'attraction est moindre.([13])



2.4.2.3 Analyse discriminante probabiliste

La règle bayésienne et le modèle gaussien

L'objectif est de trouver une règle d'affectation qui permet de prédire, pour un individu a donné, sa valeur associée de Y à partir des valeurs prises par les p variables quantitatives $X = (X^1, ..., X^p)$.
La règle bayésienne consiste à trouver une estimation de la probabilité a posteriori

d'affectation :

$$P(Y = G_k/X) = \frac{P(Y = G_k)P(X/Y = G_k)}{\sum_{k=1}^{K} P(Y = G_k)P(X/Y = G_k)} = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

où $P(Y = G_k) = \pi_k$ est la probabilité a priori d'appartenance à un groupe et $P(X/Y = G_k) = f_k(x)$ représente la fonction de densité des X conditionnellement au groupe G_k ($f_k : \mathbb{R}^p \longrightarrow [0.1]$).

- **Remarque**
  Les probabilités a posteriori $P(Y = G_k/X)$ sont parfois qualifiées de scores discriminants.

La règle bayésienne consiste à affecter l'individu a au groupe qui a la probabilité a posteriori maximale (pour lequel le score est le plus grand) :
$Y(a) = G_{k^*}$ si et seulement si

$$G_{k^*} = \arg \max_{k=\overline{1,K}} P(Y = G_k/X)$$

Toute la problématique de l'analyse discriminante revient donc à proposer une estimation de la quantité $P(X/Y = G_k) = f_k(x)$.
Plusieurs approches sont possibles, en supposant que l'on dispose d'un échantillon i.i.d. de même loi que (X,Y) :

a. On peut supposer que f_k a une forme paramétrique et estimer les paramètres sur l'échantillon d'apprentissage.

b. Les approches non paramétriques : on cherche à estimer directement à partir des données les densités f_k avec des méthodes d'estimation de densité à noyau.

Ici on se place dans le cadre paramétrique gaussien en supposant que $X \sim \mathcal{N}(\mu_k, \Sigma_k)$ dans chaque groupe G_k :

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1} (x - \mu_k)]$$

où

- $\mu_k \in \mathbb{R}^p$ est le vecteur des moyennes théoriques.

- Σ_k est la matrice $p \times p$ des variances-covariances théoriques.
- $|\Sigma_k|$ représente le déterminant de la matrice Σ_k.

L'objet étant de déterminer le maximum de la probabilité a posteriori d'affectation. Nous pouvons négliger tout ce qui ne dépend pas de k.
En passant au logarithme, nous obtenons le score discriminant qui est proportionnel à $P(Y = G_k/X)$:

$$D(Y = G_k, X) = 2\ln(\pi_k) - \ln|\Sigma_k| - (x - \mu_k)'\Sigma_k^{-1}(x - \mu_k) \tag{2.8}$$

En effet, maximiser $P(Y = G_k/X)$ revient à maximiser $\pi_k f_k(x)$ (car $\sum_{k=1}^{K} \pi_k f_k(x)$ ne dépend pas de k) ce qui est équivalent à maximiser $\ln(\pi_k f_k(x))$ avec :

$$\ln(\pi_k f_k(x)) = \ln(\pi_k) + \ln(f_k(x))$$

$$\iff \ln(\pi_k f_k(x)) = \ln(\pi_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k)$$

En multipliant par 2 et en remarquant que $\frac{p}{2}\ln(2\pi)$ est indépendant de k, on obtient le résultat indiqué en (2.8).

La règle d'affectation devient donc :

$$G_{k^*} = \arg\max_{k=\overline{1,K}} D(Y = G_k, X)$$

- **Remarque**

$$G_{k^*} = \arg\max_{k=\overline{1,K}} D(Y = G_k, X) = \arg\min_{k=\overline{1,K}} D_k^2(x)$$

où $D_k^2(x) = (x - \mu_k)'\Sigma_k^{-1}(x - \mu_k) - 2\ln(\pi_k) + \ln|\Sigma_k|$ est appelé le carré de la distance de Mahalanobis théorique généralisée.

Estimation des paramètres

A partir de l'échantillon d'apprentissage, on veut estimer le paramètre :

$$\theta = (\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k)$$

La méthode du maximum de vraisemblance peut être utilisée. La vraisemblance s'écrit :

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i) = \prod_{k=1}^{K} \prod_{x_i \in G_k} \pi_k f_k(x_i)$$

et on en déduit que la log-vraissemblance s'écrit :

$$\ln(L(\theta)) = \sum_{i=1}^{n} \sum_{x_i \in G_k} (\ln(\pi_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln \mid \Sigma_k \mid -\frac{1}{2}(x_i - \mu_k)'\Sigma_k^{-1}(x_i - \mu_k))$$

On obtient alors les estimateurs du maximum de vraisemblance suivants :

$$\widehat{\pi_k} = \frac{n_k}{n}$$

$$\widehat{\mu_k} = \frac{1}{n_k} \sum_{i \in G_k} x_i$$

$$\widehat{\Sigma_k} = \begin{cases} \widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in G_k} (x_i - \mu_k)(x_i - \mu_k)' & \text{dans le cas homoscédastique} \\ \widehat{\Sigma_k} = \frac{1}{n_k} \sum_{i \in G_k} (x_i - \mu_k)(x_i - \mu_k)' & \text{dans le cas hétéroscédastique} \end{cases}$$

Ces estimateurs de Σ_k sont biaisés et on a les estimateurs sans biais suivants :

$$\widehat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in G_k} (x_i - \mu_k)(x_i - \mu_k)'$$

$$\widehat{\Sigma_k} = \frac{1}{n_k - 1} \sum_{i \in G_k} (x_i - \mu_k)(x_i - \mu_k)'$$

Analyse discriminante quadratique

En se plaçant dans le cas hétéroscédastique *i.e.* le cas où $\exists G_k \neq G_{k^*}$ tel que $\Sigma_k \neq \Sigma_{k^*}$. On estime alors les paramètres sur l'échantillon d'apprentissage et en reprenant les notations de la section 2.2.

- μ_k est estimée par $g_k = \frac{1}{n_k} \sum_{i \in G_k} x_i$.
- Σ_k est estimée par $V_k = \frac{1}{n_k} \sum_{i \in G_k} (x_i - g_k)(x_i - g_k)'$ ou encore par sa version sans biais : $V_k = \frac{1}{n_k - 1} \sum_{i \in G_k} (x_i - g_k)(x_i - g_k)'$.

On obtient ainsi la règle de classement d'analyse discriminante quadratique :

$$G_{k^*} = \arg \min_{k=\overline{1,K}} Q_k(x)$$

où $Q_k = (x - g_k)' V_k^{-1} (x - g_k) - 2\ln(\widehat{\pi}_k) + \ln|V_k|$ est la fonction quadratique discriminante du groupe G_k (encore appelée fonction quadratique de classement). Chaque fonction quadratique discriminante définit une fonction score et un nouvel individu sera affecté au groupe pour lequel le score sera le plus petit.

Analyse discriminante linéaire

Dans le cas où les matrices de variance-covariance sont identiques *i.e.* $\Sigma_1 = ... = \Sigma_k = \Sigma$ (hypothèse d'homoscédasticité ou encore hypothèse d'équicovariance), les calculs seront simplifiés. On pourra interpréter géométriquement cette hypothèse en terme de forme et volume des nuages de points dans l'espace de représentation : ces nuages auront la même forme (et volume).

Dans ce cas, la règle bayésienne d'affectation s'écrit :

$$G_{k^*} = \arg \max_{k=\overline{1,K}} (x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \ln(\pi_k))$$

En effet, en développant la quantité

$$D_k^2(x) = (x - \mu_k)'\Sigma^{-1}(x - \mu_k) - 2\ln(\pi_k) + \ln|\Sigma|$$

on trouve

$$D_k^2(x) = x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu_k + \mu_k'\Sigma^{-1}\mu_k - 2\ln(\pi_k) + \ln|\Sigma|$$

Donc minimiser $D_k^2(x)$ est équivalent à maximiser $-\frac{1}{2}(-2x'\Sigma^{-1}\mu_k + \mu_k'\Sigma^{-1}\mu_k - 2\ln(\pi_k))$ (car $x'\Sigma^{-1}x$ et $\ln|\Sigma|$ ne dépendent pas de k).

Les estimateurs du maximum de la vraisemblance, en reprenant les notations de la section 2.2, sont donc :

- $\widehat{\mu_k} = g_k = \frac{1}{n_k} \sum_{i \in G_k} x_i$.
- $\widehat{\Sigma} = W = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in G_k} (x_i - \mu_k)(x_i - \mu_k)' = \frac{1}{n} \sum_{k=1}^{K} n_k V_k$ ou encore par la version sans biais $W = \frac{1}{n-K} \sum_{k=1}^{K} n_k V_k$.

On obtient ainsi la règle de classification d'analyse discriminante linéaire :

$$G_{k^*} = \arg \max_{k=\overline{1,K}} L_k(x)$$

où $L_k(x) = x'W^{-1}g_k - \frac{1}{2}g_k'W^{-1}g_k + \ln(\widehat{\pi_k})$ est la fonction linéaire discriminante du groupe G_k (encore appelée fonction linéaire de classement). Chaque fonction linéaire discriminante définit une fonction score et une nouvelle observation sera affectée au groupe pour lequel le score sera le plus grand.

- **Remarque**

  Lorsqu'on suppose que $\pi_1 = ... = \pi_k$ (égalité des probabilités à priori), la règle de l'analyse discriminante linéaire est équivalente à la règle de classement géométrique qui consiste à affecter un nouvel individu a au groupe G_k dont le centre de gravité g_k est le plus proche en terme de distance. Rappelons que la distance utilisée est celle de Mahalanobis correspondante à la métrique W^{-1}.

Deux groupes avec égalité des matrices de variance-covariance

On affectera l'individu a au groupe G_1 si :

$$p(Y = G_1/X) > p(Y = G_2/X)$$
$$\iff \ln(\pi_1) - \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - \ln(\pi_2) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) > 0$$
$$\iff x'\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) - \ln(\frac{\pi_2}{\pi_1}) > 0$$

Si $\pi_1 = \pi_2 = \frac{1}{2}$, on trouve la règle de Mahalanobis-Fisher en replaçant x par a et en estimant Σ par $\frac{n}{n-2}$W, μ_1 par g_1 et μ_2 par g_2.

Soit maintenant

$$S(x) = x'\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) - \ln(\frac{\pi_2}{\pi_1})$$

On affectera a au groupe G_1 si $S(x) > 0$ et au groupe G_2 si $S(x) < 0$.

La fonction S appelée score ou *statistique d'Anderson* est liée simplement à la probabilité a posteriori d'appartenance au groupe G_1.

En effet, on a :

$$P(Y = G_1/X) = P = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$

d'où :

$$\frac{1}{P} = 1 + \frac{\pi_2 f_2(x)}{\pi_1 f_1(x)} = 1 + \frac{\pi_2}{\pi_1} \exp(-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1))$$

d'où : $ln(\frac{1}{P} - 1) = -S(x)$

Soit

$$P = \frac{1}{\exp(-S(x)) + 1} = \frac{\exp(S(x))}{1 + \exp(S(x))}$$

P est donc fonction logistique du score.

2.4.3 Tests et sélection de variables discriminantes

Les valeurs sont-elles distribuées normalement au sein de chaque groupe ? Des droites de Henry ou des tests de normalité de type Kolmogorov-Smirnov permettent de répondre à cette question. Même si l'analyse discriminante est assez robuste à la non-normalité par rapport aux autres techniques linéaires telles que la régression logistique. La normalité autorise des tests importants, notamment la vérification de l'homoscédasticité.

Lorsque le nombre de variables, c'est-à-dire la dimension de X, est grand, il peut être judicieux de ne travailler qu'avec un sous-ensemble de variables : celles dont le pouvoir discriminant sera le plus élevé.

2.4.3.1 Quelques rappels sur les tests de normalité

a. Le diagramme quantile-quantile et la droite de Henry

Le diagramme quantile-quantile est un des tests de normalité les plus utilisés. Il permet de comparer deux distributions que l'on estime semblables. Sur l'axe des abscisses nous avons les quantiles de la distribution théorique (distribution normale dans notre cas) et sur l'axe des ordonnées nous avons les quantiles de l'échantillon que nous souhaitons comparer avec la distribution théorique.

Vu que nous avons un échantillon d'une loi normale, on devrait s'attendre à ce qu'il y ait une tendance linéaire dans le diagramme quantile-quantile et afin d'identifier cette tendance linéaire nous allons ajouter une droite, dite droite de Henry (voir la figure suivante) :

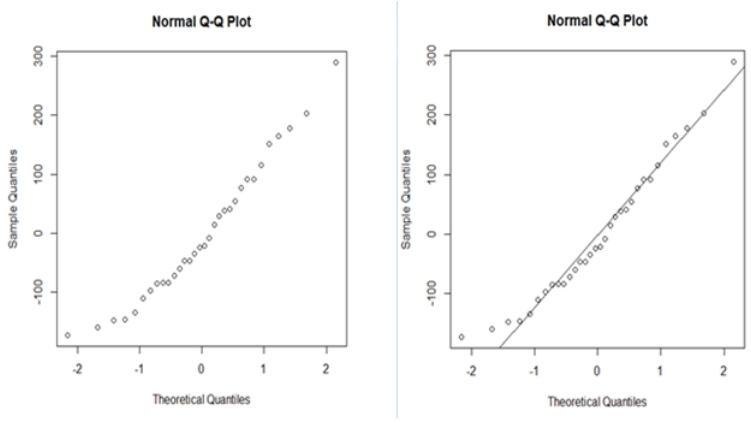


FIG. 2.2 – Diagramme quantile-quantile et droite de Henry.

b. Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov (K-S) est l'un des tests d'adéquation non paramétrique les plus courants. Il permet d'estimer si le caractère d'un échantillon est distribué de façon comparable à celui d'un autre ou s'il peut être approximé par une loi de probabilité connue. Il donne notamment une bonne indication d'ajustement à une loi normale.

Le principe est simple : on mesure l'écart maximum qui existe entre une fonction de répartition empirique F_n (donc des fréquences cumulées) et une fonction de répartition d'une loi de probabilité théorique F.

$$d = \max |F_n(x) - F(x)|$$

Précisons que le test de Kolmogorov-Smirnov est indépendant de cette loi théorique : on peut comparer la répartition empirique aussi bien à une loi normale qu'à une loi de Poisson ou autre.

Soit D la variable aléatoire qui prend la valeur d.

Sous l'hypothèse H_0, d tend vers 0. La distribution de D fait l'objet des tables de Kolmogorov, qui prennent en compte l'effectif de l'échantillon et le seuil de risque accepté : il suffit alors de comparer d à la valeur idoine de D dans la table.

2.4.3.2 Homoscédasticité et test de Box

L'hypothèse d'égalité des matrices Σ_k peut être testée au moyen du test de Box. Si l'hypothèse $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$ est vraie, la quantité :

$$(1-\frac{2p^2+3p-1}{6(p+1)(k-1)})[(\sum_k \frac{1}{n_k-1}-\frac{1}{n-K})(n-K)\ln|\frac{n}{n-K}W|-\sum_k (n_k-1)\ln|\frac{n_k}{n_k-1}V_k|]$$

suit approximativement une loi de χ^2 à $\frac{p(p+1)(K-1)}{2}$ degrés de liberté.

- **Remarque**

  Si l'on rejette l'hypothèse d'égalité, doit-on utiliser les règles quadratiques ?
  Cela n'est pas sûr dans tous les cas. Tout d'abord le test de Box n'est pas parfaitement fiable, ensuite l'usage de règles quadratiques implique l'estimation de bien plus de paramètres que la règle linéaire, puisqu'il faut estimer chaque Σ_k. Lorsque les échantillons sont de petite taille, les fonctions obtenues sont très peu robustes et il vaut mieux utiliser une règle linéaire malgré tout.
  Si, par exemple, $p = 10$ variables et $K = 4$ groupes : l'analyse discriminante linéaire demande l'estimation de 95 paramètres et l'analyse discriminante quadratique l'estimation de 260 paramètres.

2.4.3.3 Test de Wilks

Soient les hypothèses suivantes :
Hypothèse nulle : $H_0 = \{$les centres de gravité conditionnels sont confondus : indépendance entre X et Y $(\mu_1 = \mu_2 = ... = \mu_K)\}$.
Hypothèse alternative : $H_1 = \{$il existe au moins un des centres de gravité qui s'écarte significativement des autres$\}$.
La statistique du test est le lambda de Wilks, son expression est la suivante :

$$\Lambda = \frac{|W|}{|V|} = \frac{|W|}{|W+B|} = \frac{1}{|W^{-1}B+I|}$$

Elle suit la loi de Wilks de paramètres $(p, n-k, k-1)$ sous H_0
avec $|W|$ représente le déterminant de la matrice de variance-covariance intra-groupes et $|V|$ le déterminant de la matrice de variance-covariance globale.

L'hypothèse H_0 est rejetée si Λ calculé est inférieur à Λ tabulé.

Remarques

- Ce test peut s'exprimer comme une généralisation multidimensionnelle de l'analyse de variance à un facteur (ANOVA), on parle dans ce cas de MANOVA (Multidimensional Analysis of Variance).
- Il est rare de trouver la table de la loi de Wilks implémentée sous les différents logiciels de statistique existant. Par conséquent, si n est suffisamment grand, nous nous orienterons vers l'utilisation de l'approximation de Bartlett suivante :

  $$\chi^2 = -[n - \frac{1}{2}(p - K - 1)] \ln(\Lambda)$$

  qui suit une loi du χ^2 à P degrés de liberté.
- Dans le cas où $K = 2$, on peut utiliser la transformation de Rao qui suit une loi de Fisher de paramètres (p, n-p-1).

  La formule de la statistique de test devient alors :

  $$F = \frac{1 - \Lambda}{\Lambda}$$

2.4.3.4 Evaluation individuelle des variables prédictives - Sélection de variables pas à pas

Comme dans toutes les méthodes linéaires, il est possible d'évaluer individuellement chaque variable prédictive et éventuellement d'éliminer celles qui ne sont pas significatives dans la discrimination.
La statistique du test s'appuie sur la variation du Lambda de Wilks mesurée par :

$$F = \frac{n - K - p}{K - 1}(\frac{\Lambda_p}{\Lambda_{p+1}} - 1)$$

Elle suit une loi de Fisher à $(K - 1, n - K - p)$ degrés de liberté.

- **Remarque**

  La plupart des logiciels présentent des techniques de sélection ascendante, descendante ou mixte des variables.

Sélection ascendante (option Forward)
- A l'étape initiale aucune variable n'est présente.
- A chaque étape on fait entrer la variable qui contribue le plus au pouvoir discriminant du modèle, mesuré par le lambda de Wilks.
- La sélection s'arrête quand aucune des variables non sélectionnées ne convient au sens du seuil de probabilité choisi pour le F de Fisher.

Sélection descendante (option Backward)
- On démarre avec le modèle complet (construit avec toutes les variables)
- A chaque étape, la variable contribuant le moins au pouvoir discriminant du modèle est éliminée.
- La sélection s'arrête quand on ne peut plus éliminer de variables étant donné le seuil de probabilité choisi pour le F de Fisher.

Sélection mixte (option Stepwise)
- On démarre comme dans la procédure ascendante.
- Dès qu'une variable entre dans le modèle, on vérifie compte tenu de cette entrée si l'une des variables déjà présentes est susceptible d'être éliminée.
- La sélection s'arrête quand on ne plus ajouter ou éliminer de variables.

2.4.4 Validation d'une règle d'affectation

Elle consiste à estimer le taux théorique d'erreur de classement (appelé aussi risque ou coût) à partir d'un ensemble d'individus répartis en K groupes soumis à une analyse discriminante. Cette estimation s'obtient de différentes manières, que l'on peut mettre en œuvre dans les analyses.

2.4.4.1 Méthode de resubstitution

De manière classique en analyse discriminante, pour évaluer les performances d'une fonction de classement, nous confrontons ses prédictions avec les vraies valeurs de la variable à prédire sur un fichier de données. Le tableau croisé qui en résulte s'appelle une matrice de confusion de terme général n_{ks} avec : en lignes les

vrais groupes d'appartenance, en colonnes les groupes d'appartenance prédits (voir le tableau ci-dessous). Le taux d'erreur ou taux de mauvais classement est tout simplement, lorsque la prédiction ne coïncide pas avec la vraie valeur, rapporté à l'effectif du fichier de données.

		Affectation			
		G_1	G_2	G_s	G_k
Origine (échantillon total)	G_1	n_{11}	n_{12}	n_{1s}	n_{1k}
	G_2	n_{21}	n_{22}	n_{2s}	n_{2k}
	G_k	n_{k1}	n_{k2}	n_{ks}	n_{kk}

L'estimation du taux théorique d'erreur de classement est égale au pourcentage d'individus de l'échantillon mal classés par la règle d'affectation, on parle de taux d'erreur en resubstitution :

$$\widehat{C}^{res} = 1 - \frac{1}{n}\sum_{k=1}^{K} n_{kk}$$

Cette estimation, obtenue en appliquant la règle d'affectation à tous les individus ayant contribué à la construire, est une estimation dite *échantillon dépendante*. C'est une estimation biaisée.

- **Remarque**

  Cette approche n'est pas recommandée en pratique puisqu'on a tendance à sous-estimer le taux d'erreur.

Pour obtenir une estimation plus précise du taux d'erreur théorique de classement, on a recours à la méthode suivante.

2.4.4.2 Méthode de l'échantillon-test

Cette méthode est valable dans le cas des grands échantillons. La procédure est de construire la fonction de classement sur une fraction de données, dites d'apprentissage ; puis de l'évaluer sur une autre fraction de données, dites de test.
On effectue au hasard un tirage de 20 à 30% des individus dans chaque groupe, ce

qui constitue l'échantillon test, répartis en K sous-ensembles $T_1, ..., T_K$, le reste (de 70 à 80%) est l'échantillon d'apprentissage (échantillon de base) constitué de K sous-ensembles $B_1, ..., B_K$.

L'analyse discriminante est alors effectuée sur les individus de l'échantillon de base. Il en résulte une règle d'affectation qui, appliquée à l'échantillon-test, fournit une estimation du taux théorique d'erreur de classement qui est le pourcentage d'individus de l'échantillon-test mal classés par la règle d'affectation issu du tableau de classement :

		Affectation			
		G_1	G_2	G_s	G_k
Origine (échantillon-test)	G_1	n_{11}	n_{12}	n_{1s}	n_{1k}
	G_2	n_{21}	n_{22}	n_{2s}	n_{2k}
	G_k	n_{k1}	n_{k2}	n_{ks}	n_{kk}

$$\widehat{C}^{et} = 1 - \frac{1}{n}\sum_{k=1}^{K} n_{kk}$$

$\widehat{C}^{et}$ est une meilleure estimation (estimation sans biais) que celle obtenue par resubstitution, puisqu'elle est obtenue à partir d'un échantillon indépendant de l'échantillon de base.

Remarque

On peut sélectionner une règle de classement permettant de minimiser l'erreur de classement afin de sélectionner les variables discriminantes en utilisant la méthode dite *séquentielle* :

1. On sélectionne la variable j_1 qui permet d'obtenir le plus faible taux d'erreur dans un modèle à 1 variable.

2. On sélectionne la variable j_2 qui permet d'obtenir le plus faible taux d'erreur, lorsqu'elle est utilisée conjointement à j_1 dans un modèle à 2 variables.

3. On sélectionne la variable j_3 qui permet d'obtenir le plus faible taux d'erreur, lorsqu'elle est utilisée conjointement à j_1 et j_2 dans un modèle à 3 variables. etc

Chapitre 3

Application

3.1 Introduction

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 : SAS, Splus, R, XlStat, UniWin Plus, Stalab et SPAD. Dans ce chapitre nous avons fait le choix d'utiliser le logiciel R qui est la version gratuite de Splus. Il est téléchargeable sous www.r-project.org pour tous systèmes d'exploitation. Il semble très peu employé en industrie. De part sa gratuité, il est de plus en plus employé pour la réalisation de cours de statistiques.
Notre travail consiste à discriminer les arbres aux nombre 140 sur la base des caractères quantitatifs des gousses et des graines utilisées en industrie. Ces gousses ont été récoltées dans 7 localités de l'Algérie du nord. Sur chaque gousse et sur chaque graine ont été évalués les paramètres suivants : la longueur, la largeur, l'épaisseur et le poids. On s'intéresse au rendement de chaque arbre qui a été réparti en 3 classes suivantes : A, B et C.

3.2 Les données

Les données suivantes concernent 140 observations et 9 variables dont 8 sont numériques et la neuvième est qualitative à 3 modalités. Le nombre d'individus qui

constituent la première et la deuxième classe (A et B) est 44 tandis que la troisième classe (C) contient 52 individus.

```
> #exploring data
> arbre<-read.table("arbre.txt",header=TRUE)
> str(arbre)
'data.frame': 140 obs. of 9 variables:
 $ sl    : num 9.38 9.8 10.06 9.82 10.03 ...
 $ swd   : num 7.46 6.28 6.44 6.31 6.38 6.52 7.8 7.58 7.01 7.33 ...
 $ sth   : num 4.64 3.68 3.67 3.75 3.75 3.75 4.52 3.53 3.89 4.5 ...
 $ swg   : num 1.65 1.6 1.83 1.76 1.78 1.67 1.87 1.22 2.82 2.74 ...
 $ pl    : num 18.3 17.5 17.7 17.9 19 ...
 $ pwd   : num 2.34 1.91 1.99 1.97 1.88 1.93 2.46 1.94 2.26 2.11 ...
 $ pth   : num 1.16 0.76 0.82 0.78 0.75 0.78 1.06 0.91 0.99 0.96 ...
 $ pwg   : num 14.4 12 14.2 13.6 13.6 ...
 $ classes: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
> table (arbre$classes)

 A  B  C
44 44 52
```

Description des variables

sl : longueur de la graine (seed lenght).

swd : largeur de la graine (seed width).

sth : épaisseur de la graine (seed thikness).

swg : poids total des graines par gousse (seed weight).

pl : longueur de la gousse (pod lenght).

pwd : largeur de la gousse (pod width).

pth : épaisseur de la gousse (pod thikness).

pwg : poids de la gousse (pod weight).

classes : classe de rendement de chaque arbre, il en existe trois (A, B et C)

Objectif

Réduire le nombre de dimensions des données, en projetant les individus sur des axes déformant le moins possible la réalité, en prenant en compte la structuration en classes de la population totale des individus.

Principe

Recherche de nouvelles variables, dites *composantes discriminantes*, obtenues à partir des combinaisons linéaires des variables originales.

3.3 Nuages de points et corrélation

Pour chaque variable quantitative nous avons les histogrammes, les nuages des points ainsi que les coefficients de corrélation suivants (voir la figure 3.1) :

```
#scatter plots and correlations
> library(psych)
> pairs.panels(arbre[1:8], gap = 0, bg = c("red", "green", "yellow")
[arbre$classes], pch=21)
```

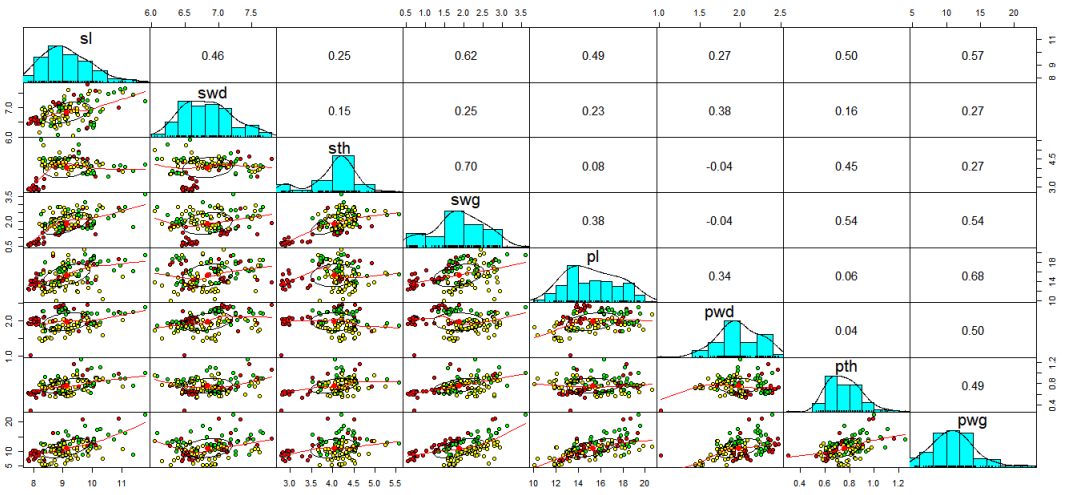


FIG. 3.1 – Nuages de points, histogrammes et corrélation.

On observe, en visualisant la figure ci-dessous, une allure approximativement normale, cela suggère que chaque variable suit une loi normale. On peut aussi remarquer que les variables qui aident le plus à séparer les classes sont "pwg" et "swg". Cependant les autres variables ne permettent pas une bonne séparation des groupes.
Signalons aussi la faible corrélation entre ces variables deux à deux.

3.4 Méthode échantillon-test

Séparons l'échantillon en deux blocs : échantillon d'apprentissage pour créer le modèle et élaborer des règles de décision ou d'affectation et échantillon test pour estimer les performances du modèle.
On prend 70% de l'échantillon global comme échantillon d'apprentissage et les 30% restante comme échantillon test.

```
> #data partition
> set.seed(555)
> ind<-sample(2, nrow(arbre), replace=TRUE, prob=c(0.7, 0.3))
```

3.4.1 Echantillon d'apprentissage

```
> #échantillon d'apprentissage
> training<-arbre[ind==1, ]
> dim(training)
[1] 96 9
```

Le nombre d'individus constituant l'échantillon d'apprentissage est 96 arbres.

3.4.2 Echantillon test

```
> #échantillon test
> testing<-arbre[ind==2, ]
> dim(testing)
[1] 44 9
```

Le nombre d'individus constituant l'échantillon test est 44 arbres.

3.5 L'analyse linéaire discriminante

Les coordonnées de chaque variable sur les deux axes discriminants ainsi que les probabilités a priori des classes et les moyennes des variables dans chacune de ces classes sont donnés comme suit :

```
> #linear discriminant analysis
> library("MASS")
> linear<-lda(classes~.,training)
> linear
Call:
lda(classes ~ ., data = training)

Prior probabilities of groups:
       A        B        C
0.3541667 0.3229167 0.3229167

Group means:
       sl      swd      sth      swg      pl      pwd      pth      pwg
A 8.795882 6.795000 3.635588 1.234706 15.28912 2.092647 0.6938235 11.226471
B 9.502581 6.856452 4.245161 2.090323 16.15226 2.078065 0.8258065 12.953548
C 9.133226 6.878387 4.272581 2.201935 14.30452 1.788387 0.7706452 9.383226
```

```
Coefficients of linear discriminants:
          LD1        LD2
sl  -0.3774525 -0.51045533
swd  0.4150403 1.21870600
sth -0.4045695 -0.88159280
swg  4.3386545 0.07298041
pl   0.1805364 -0.04576348
pwd -0.1804172 -1.42677842
pth  1.3020256 -2.05657750
pwg -0.7660769 -0.08885871

Proportion of trace:
   LD1    LD2
0.9395 0.0605
```

Nous remarquons que le premier axe explique 93.95% de l'information tandis que le deuxième explique seulement 6.05%.

3.6 Qualité de représentation sur les axes discriminants

Les probabilités a postériori d'appartenance de chaque individu de l'échantillon d'apprentissage à chacune des trois classes et les coordonnées des projections des individus sur les deux axes discriminants sont données comme suit :

```
> p<-predict(linear,training)
> p
$class
 [1] A A A A B A A A A A A A A A A A A A A A A A A A A A B A A A A A A A A B B B B
[39] B B B B B B B B B B B B B B B B B B B A B B B C B B B C C C C C C C C C C C C
[77] C C C B B C C C C C C C C C C C C C C C C
Levels: A B C

$posterior
              A           B           C
1   6.141179e-01 3.858804e-01 1.658474e-06
3   7.093896e-01 2.906076e-01 2.891670e-06
4   6.432680e-01 3.567260e-01 6.011507e-06
5   5.310552e-01 4.689237e-01 2.102694e-05
6   3.800239e-01 6.198607e-01 1.154611e-04
8   9.942521e-01 5.747863e-03 1.871589e-08
9   7.847680e-01 2.152320e-01 4.004726e-08
12  9.980305e-01 1.969442e-03 6.052346e-08
13  9.976758e-01 2.324165e-03 2.434154e-08
14  9.997540e-01 2.459095e-04 1.094580e-07
17  9.878901e-01 1.210223e-02 7.688972e-06
18  9.846646e-01 1.533457e-02 8.062936e-07
21  9.861681e-01 1.382956e-02 2.371158e-06
```

```
22  9.992161e-01 7.839259e-04 1.761672e-09
23  9.878171e-01 1.218041e-02 2.483014e-06
24  9.995687e-01 4.313186e-04 1.279543e-09
25  9.993126e-01 6.874024e-04 2.025117e-09
26  9.960448e-01 3.955026e-03 1.277670e-07
27  9.801378e-01 1.985972e-02 2.440675e-06
29  9.592469e-01 4.072287e-02 3.026301e-05
30  9.979894e-01 2.010521e-03 3.948284e-08
31  9.717614e-01 2.823852e-02 8.203255e-08
32  9.103479e-01 8.964945e-02 2.672914e-06
33  9.878907e-01 1.210927e-02 2.248174e-08
34  6.282377e-01 3.716994e-01 6.285293e-05
35  4.851181e-01 5.147586e-01 1.232637e-04
36  8.215370e-01 1.784497e-01 1.328276e-05
37  9.531694e-01 4.683043e-02 1.270238e-07
38  7.104397e-01 2.894074e-01 1.529741e-04
39  8.751693e-01 1.248096e-01 2.113080e-05
40  6.591873e-01 3.407795e-01 3.324357e-05
42  9.793482e-01 2.065182e-02 1.132208e-08
43  7.711749e-01 2.288031e-01 2.200989e-05
44  8.564935e-01 1.435024e-01 4.104011e-06
45  2.301903e-01 7.697184e-01 9.126172e-05
46  1.788445e-01 8.206763e-01 4.792286e-04
```

```
48  1.156524e-01 8.836382e-01 7.093469e-04
49  1.722763e-01 8.275649e-01 1.587827e-04
50  7.559007e-02 9.233605e-01 1.049397e-03
51  1.908115e-01 8.011726e-01 8.015910e-03
52  1.002310e-02 9.057894e-01 8.418755e-02
54  4.496687e-03 9.052628e-01 9.024054e-02
55  3.740692e-01 6.259225e-01 8.299473e-06
56  3.540980e-02 9.624766e-01 2.113619e-03
57  1.400968e-02 9.859542e-01 3.612909e-05
58  4.088274e-03 9.911894e-01 4.722327e-03
61  5.573841e-03 9.868848e-01 7.541324e-03
62  2.771903e-03 6.263277e-01 3.709004e-01
63  4.806284e-03 6.218225e-01 3.733713e-01
64  5.678359e-03 8.161343e-01 1.781873e-01
65  2.939525e-03 7.147008e-01 2.823597e-01
67  2.380257e-03 6.340572e-01 3.635625e-01
68  4.629506e-01 5.364979e-01 5.514093e-04
69  2.178655e-01 7.808772e-01 1.257387e-03
70  4.429006e-01 5.568532e-01 2.462312e-04
71  3.321729e-01 6.660463e-01 1.780842e-03
72  3.344344e-01 6.650148e-01 5.508252e-04
73  7.260475e-01 2.738780e-01 7.452574e-05
79  9.083049e-04 4.055384e-01 5.935533e-01
```

```
77  2.253182e-02 9.154258e-01 6.204236e-02
78  2.418331e-03 9.932211e-01 4.360568e-03
79  9.083049e-04 4.055384e-01 5.935533e-01
80  1.479773e-03 8.836023e-01 1.149179e-01
87  7.522495e-02 9.240319e-01 7.431609e-04
88  2.259984e-02 9.703698e-01 7.030315e-03
89  2.294670e-04 3.112567e-01 6.885138e-01
91  2.480402e-13 9.520301e-07 9.999990e-01
92  2.162907e-09 1.558048e-04 9.998442e-01
93  1.816067e-04 4.289924e-01 5.708260e-01
95  3.632928e-09 1.742161e-04 9.998258e-01
98  4.845659e-10 1.461484e-04 9.998539e-01
99  1.840521e-08 1.367766e-03 9.986322e-01
103 8.604316e-04 2.720338e-01 7.271058e-01
105 8.393550e-04 3.785076e-01 6.206530e-01
106 1.028516e-06 2.255738e-03 9.977432e-01
107 7.529876e-07 2.313592e-03 9.976857e-01
109 1.881853e-03 3.549291e-01 6.431890e-01
110 1.588809e-04 8.875226e-02 9.110889e-01
111 7.599355e-06 9.178581e-03 9.908138e-01
114 2.292223e-03 5.458062e-01 4.519016e-01
115 3.504472e-03 6.113031e-01 3.851924e-01
117 2.119810e-07 2.018028e-03 9.979818e-01
```

```
119 2.544300e-08 8.206250e-04 9.991793e-01
120 5.838927e-06 1.281884e-02 9.871753e-01
121 2.313615e-07 1.501398e-03 9.984984e-01
124 2.747938e-04 1.158317e-01 8.838935e-01
125 6.513005e-06 6.164708e-02 9.383464e-01
126 6.662819e-07 1.113444e-02 9.888649e-01
128 8.592472e-06 3.566977e-02 9.643216e-01
130 1.408096e-07 2.116416e-03 9.978834e-01
131 1.356133e-07 3.637278e-04 9.996361e-01
132 9.598555e-16 2.226570e-08 1.000000e+00
133 3.348837e-06 1.891130e-03 9.981055e-01
134 1.017436e-07 2.925967e-04 9.997073e-01
138 5.490518e-03 3.408051e-01 6.537044e-01
139 2.410806e-07 4.372673e-04 9.995625e-01
$x
           LD1        LD2
1   -2.25479326 -1.65712748
3   -2.16861863 -1.14227913
4   -2.00950439 -1.09890317
5   -1.73184046 -0.97064965
6   -1.33942088 -0.75308161
8   -3.17126940 0.53758622
9   -3.02365124 -2.39340910
```

```
12  -2.92816641 1.88048793
13  -3.10899841 1.40762601
14  -2.78617962 3.87246874
17  -1.99781562 2.08832674
18  -2.44292051 1.06433522
21  -2.23012904 1.54565091
22  -3.61127049 1.38061968
23  -2.21980108 1.67160498
24  -3.66665616 1.77436327
25  -3.58228607 1.54350095
26  -2.78980878 1.55628310
27  -2.22784605 1.24431728
29  -1.73855041 1.53680611
30  -3.01228340 1.70745404
31  -2.89680554 -0.29457305
32  -2.21523674 -0.04508152
33  -3.14339799 -0.03478906
34  -1.54485205 -0.29172545
35  -1.36915096 -0.45117878
36  -1.89017772 -0.09958844
37  -2.81372071 -0.57666706
38  -1.38970181 0.30562925
39  -1.80619022 0.40538378
```

```
40  -1.67766088 -0.42576596
42  -3.28314062 -0.74416543
43  -1.78249675 -0.15896756
44  -2.12571184 -0.32044115
45  -1.29613360 -1.26890868
46  -0.92492350 -0.84388159
48  -0.76870591 -0.97775087
49  -1.13502699 -1.27118673
50  -0.61418090 -1.08103831
51  -0.38345809 0.23314026
52   0.61840366 -0.45880163
54   0.77940735 -0.82552134
55  -1.85353456 -1.72686885
56  -0.33780928 -1.23304479
57  -0.96653088 -3.18788659
58   0.21656208 -2.02275980
61   0.25153323 -1.69698918
62   1.15046220 -0.23386097
63   1.05066710 0.04421012
64   0.87138218 -0.37531997
65   1.08446951 -0.41697869
67   1.17439081 -0.32618303
68  -1.06693094 0.03562119
```

```
69  -0.77120148 -0.35390978
70  -1.21754136 -0.31112334
71  -0.77842529 0.11497087
72  -1.01003599 -0.30725690
73  -1.53419684 0.10173288
74  -0.59678357 0.10091267
77   0.40942018 -0.18224209
78   0.29741798 -2.31058552
79   1.45332566 -0.23816252
80   1.03151083 -1.26112672
87  -0.68104637 -1.20955611
88  -0.01939496 -1.02268114
89   1.73871015 -0.63202059
91   5.76615355 0.23127161
92   4.03434815 0.32336165
93   1.74090724 -1.08854486
95   3.93760851 0.48195566
98   4.31017892 -0.35451485
99   3.61333542 -0.48258975
103  1.50811015 0.14991750
105  1.47746640 -0.20170803
106  2.86724533 1.06006899
107  2.92424637 0.88575347
```

```
109  1.33683597 0.26148490
110  1.87695303 0.36073589
111  2.48066791 0.83909795
114  1.22589567 -0.13760077
115  1.11507483 -0.08456046
117  3.15904949 0.38185993
119  3.56028911 0.11214748
120  2.52422255 0.42364921
121  3.14675903 0.67726319
124  1.76695413 0.39071229
125  2.47456973 -0.88162497
126  2.92537817 -0.51831826
128  2.43581931 -0.26908965
130  3.23364628 0.14089129
131  3.26288975 1.62606573
132  6.83424159 0.71660524
133  2.65248436 1.78868411
134  3.31844980 1.67110681
138  1.14366741 0.82633252
139  3.15480356 1.75059327
```

```
> ldahist(data=p$x[,1],g=training$classes)
```

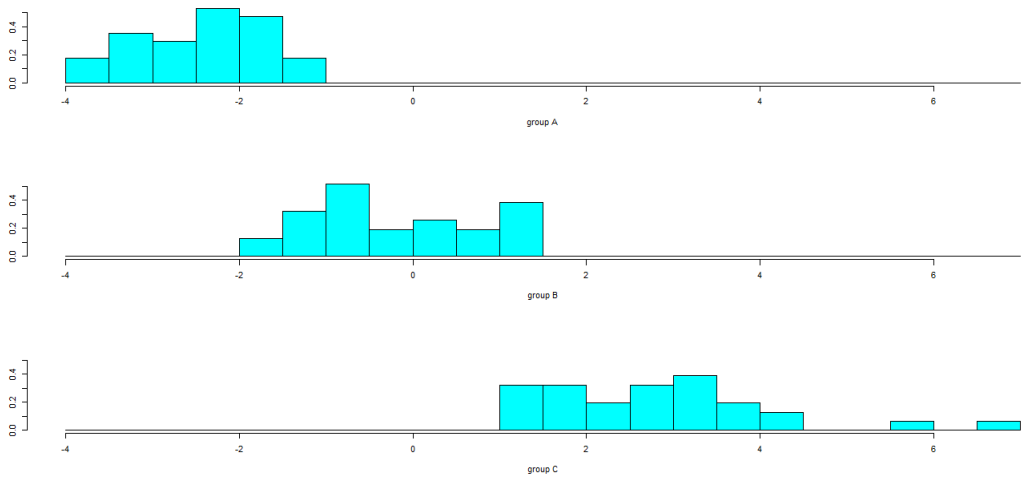


FIG. 3.2 – Qualité de représentation sur l'axe LD1.

La figure 3.2 nous montre que, sur le premier axe discriminant, la classe A est très bien séparée de la classe C mais on n'a pas une nette séparation entre les classes A et B ainsi qu'entre les classes B et C.

Tandis que la figure suivante nous montre qu'il n'y a aucune séparation entre les trois classes A, B et C sur l'axe LD2.

```
> ldahist(data=p$x[,2],g=training$classes)
```

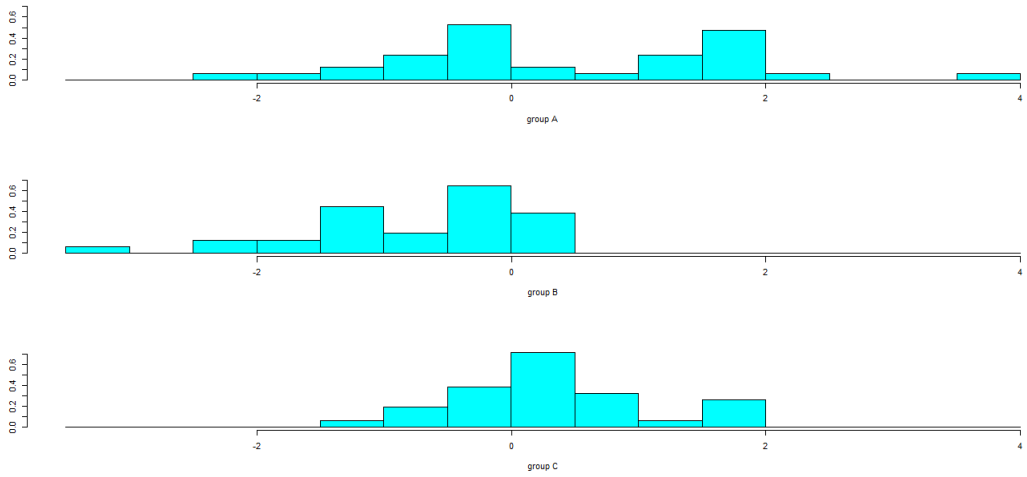


FIG. 3.3 – Qualité de représentation sur l'axe LD2.

3.7 Représentation simultanée

```
> #Biplot
> library(devtools)
> library(ggord)
> ggord(linear, training$classes)
```

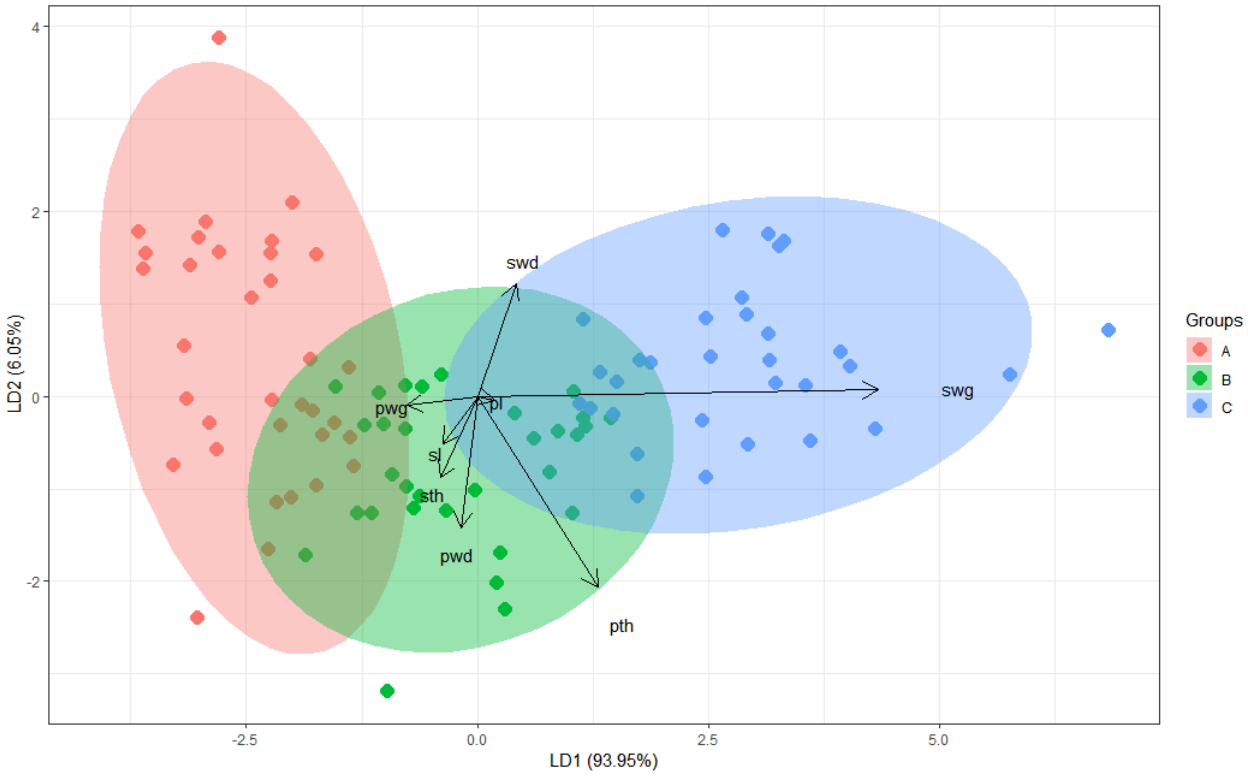


FIG. 3.4 – Biplot.

La figure ci-dessus montre que les trois classes sont assez bien séparées. L'axe "LD2" n'est guère discriminant mais il est bien utile pour faire un graphique.
Nous remarquons aussi que les variables swg, swd, pth et pwd interviennent dans la description des individus contrairement à pl, sl, sth et pwg et que la variable swg permet de bien séparer la classe C des autres classes.

3.8 Matrice de confusion- échantillon d'apprentissage

La matrice de confusion suivante donne les résultats du classement :

```
> #Cofusion matrix and accurary - Training data
> p1<-predict(linear, training)$class
> tab<-table(Predicted=p1, Actual=training$classes)
> tab
         Actual
Predicted A  B  C
        A 32  1  0
        B  2 29  2
        C  0  1 29
> sum(diag(tab))/sum(tab)
[1] 0.9375
```

On trouve qu'un seul individu de la classe A est classé dans B et que deux individus de la classe B, sont attribués à la classe A et deux autres à la classe C, alors qu'un seul individu de la classe C est mal classé.
Le pourcentage des individus bien classés est 93.75%. Ces résultats semblent excellents.

3.9 Matrice de confusion- échantillon test

On trouve, en visualisant la matrice ci-dessous, que les 9 individus de la classe A sont parfaitement classés et que seuls deux individus de la classe C et de la classe B sont mal classés.

```
> #Cofusion matrix and accurary - Testing data
> p2<-predict(linear, testing)$class
> tab1<-table(Predicted=p2, Actual=testing$classes)
> tab1
         Actual
Predicted A  B  C
        A  9  0  0
        B  1 11  1
        C  0  2 20
> sum(diag(tab1))/sum(tab1)
[1] 0.9090909
```

L'estimation du taux théorique d'erreur est : $\widehat{C^{et}} = 9.09\%$

3.10 Test de Wilks

Les centres de gravité des trois nuages de points s'écartent significativement. C'est ce que nous indique le test de Wilks suivant :

```
> #Test de Wilks
> fit <- manova(as.matrix(arbre[,1:8]) ~ classes , arbre)
> summary(fit, test = "Wilks")
          Df   Wilks approx F num Df den Df Pr(>F)
classes    2 0.12958  28.893    16    260 < 2.2e-16 ***
Residuals 137
---
Signif. codes: ‘0 ***’‘ 0.001 **’‘ 0.01 *’‘’‘ 0.05 . 0.1 ’1
```

Conclusion

Nous avons présenté dans le premier chapitre le principe général des analyses factorielles. Cette approche permet de représenter géométriquement de grands tableaux de données dans des sous-espaces sans perte d'information importante. La dimension de ces sous-espaces se fait en cherchant à minimiser la perte d'information. Une fois la dimension du sous-espace choisie, les données sont représentées graphiquement par des projections sur les différents plans qui constituent le sous-espace. Bien sûr les premiers plans principaux sont ceux contenant le plus d'information. Avant d'appliquer cette approche générale à un tableau quelconque, il est important de tenir compte des données de départ. Pour se faire, il faut appliquer des transformations en fonction de leur type. En effet, nous avons vu comment transformer des données quantitatives dans le cadre de l'analyse en composantes principales et des données qualitatives dans les cas de l'analyse factorielle de correspondances et de celle des correspondances multiples.

Dans le second chapitre nous avons présenté l'une des méthodes les plus utilisées de nos jours : l'analyse discriminante. La simplicité de sa mise en œuvre fait que nous la retrouvons dans de nombreux logiciels. Elle est adéquate pour la représentation des données dans des espaces qui discriminent au mieux les individus selon des classes connues. Cette représentation permet de dégager des informations à partir d'un grand nombre de données souvent difficile à interpréter. Elle permet également l'affectation de nouveaux individus dans les classes existantes. Il est alors possible de rendre la méthode adaptative pour tenir compte de ces nouvelles observations.

Le troisième chapitre est consacré à l'application de l'analyse linéaire discriminante sur des données d'agronomie. Les résultats obtenus semblent adaptés à la réalité.

Références bibliographiques

[1].J-M.ROMEDER., Méthode et programmes d'analyse discriminante., DUNOD Paris-Bruxelles-Montréal., 1973.

[2].F.CAILLIEZ et J-P.PAGES., Introduction à l'Analyse des Données., SMASH., 1976.

[3].L.Lebart, A.Morineau, N.Tabard., Techniques de la description statistique. Méthodes et logiciels pour l'analyse de grands tableaux., Dunod., 1977.

[4].J-P.Pages., F.Cailliez, Y.Escoufier., Analyse factorielle : un peu d'histoire et de géométrie., Revue de Statistique Appliquée, Vol XXVII, n°1 pp. 5-28., 1979.

[5].J-P.Fenelon., Qu'est-ce que l'analyse des données ?, Lefonen., 1981.

[6].J-P.Benzécri., Histoire et préhistoire de l'analyse des données., Dunod., (1982).

[7].E.Diday, J.Lemaine, J.Pouget, F.Testu., Eléments d'analyse de données., DUNOD., 1982.

[8].B.Escoffier et J. Pagès : Analyses factorielles simples et multiples objectifs, méthodes et interprétations., Dunod., 1990.

[9].S.AMBAPOUR., Introduction à l'analyse des données., BAMSI B.P. 13734 Brazzaville., 04/2003.

[10].G.Govaert., Analyse des données., LAVOISIER., 2003.

[11].J-P.Nakache, Josiane Confais., Statistique explicative appliquée., TECHNIP, 27 rue Ginoux, 75737 PARIS Cedex 15, France., 2003.

[12].A.MARTIN.,L'analyse de données Polycopié de cours ENSIETA - Réf. : 1463., Septembre 2004.

[13].G.SAPORTA., Probabilités, analyse de données et statistique, TECHNIP, 27 rue Ginoux, 75737 PARIS Cedex 15, France, 2006.

[14].A.GUYAFER. , Régression linéaire., 2013.

[15].V.Monbet., Analyse des données Master Statistique et économétrie., 2013-2014.
[16].C.Duby, S. Robin., Analyse en Composantes Principales., AgroParisTech.