Méthodes multi variable et multifactorielles

I. Introduction

Les ***méthodes factorielles***, ou ***méthodes exploratoires multivariées***, fournissent des représentations graphiques qui constituent la meilleure synthèse visuelle de l'information contenue dans les grands tableaux de données. Il s'agit de représentations approchées car il faut consentir à une perte d'information afin de gagner en lisibilité, mais la méthode factorielle assure une approximation optimale.

La nature du tableau de données à analyser guide la méthode factorielle à utiliser : **A**nalyse en **C**omposantes **P**rincipales **(ACP)**, pour les variables quantitatives, **A**nalyse **F**actorielle des **C**orrespondances **(AFC)** pour les tableaux de contingence et l'**A**nalyse des **C**orrespondances **M**ultiples **(ACM)** pour les variables qualitatives.

Il existe d'autres méthodes d'analyses factorielles, comme l'analyse factorielle d'un tableau de distances, l'analyse des tableaux ternaires, etc. Toutes les méthodes d'analyses factorielles ont un tronc commun de calculs et reposent sur les mêmes principes géométriques que l'Analyse en Composantes principales, qu'on appelle aussi ACP.

II. Les méthodes d'analyse des données

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

- Les méthodes de classification,
- Les méthodes factorielles ou méthodes exploratoires multivariées.

- Les ***méthodes de classification*** (partitionnement, Classification Ascendante Hiérarchique CAH), visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes. Du point de vue des individus - c'est-à-dire les lignes du tableau - la synthèse se fait par regroupement des individus en classes d'individus semblables : c'est la famille ***des algorithmes de classification*** :
    - Les méthodes de partitionnement ;
    - Les algorithmes d'arbre hiérarchique (CAH)

- Les ***méthodes factorielle***s [méthodes d'analyse factorielle (ACP, AFCM,…)], cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l'algèbre linéaire et donnant lieu à des représentations graphiques dans lesquelles les objets à décrire se transforment en des points sur des axes et des plans.

Les ***principales techniques (méthodes)*** sont : ***Quelles méthodes pour quelles données ?***

- ✓ ***L'analyse en composantes principales*** (***ACP***), employée pour des données quantitatives, et ses méthodes dérivées :
    - ***l'analyse factorielle des correspondances*** (***AFC***) utilisée sur des données qualitatives (tableau d'association) ;
    - ***l'analyse factorielle des correspondances multiples*** (***AFCM*** ou ***ACM***) généralisant la précédente.
- ✓ ***L'analyse canonique*** et ***l'analyse canonique généralisée***, qui sont plus des cadres théoriques que des méthodes aisément applicables, étendent plusieurs de ces méthodes et vont au-delà des techniques de description. L'Analyse Factorielle Multiple est adaptée aux tableaux dans lesquels les variables sont structurées en groupes et peuvent être quantitative et/ou qualitatives.
- ✓ ***La classification automatique***, ***l'analyse factorielle discriminante*** (***AFD***) ou ***analyse discriminante*** permettent d'identifier des groupes homogènes au sein de la population du point de vue des variables étudiées.

II.1.Analyse en composantes principales (ACP) - Avec des données quantitatives, on fait de l'ACP

L'analyse en composantes principales est notée ACP. Elle s'applique à tous les tableaux de données où les variables sont de type quantitatif.

L'analyse en composantes principales consiste en l'étude des projections des points de ce nuage sur un axe (axe factoriel ou principal).

II.1.1. Présentation générale de l'ACP

On dispose d'un tableau de données où les lignes représentent des individus et les colonnes, des variables observées sur ces individus. La propriété essentielle de l'analyse factorielle ACP est d'analyser les variables quantitatives dans leur ensemble, plus précisément en tenant compte de toutes les liaisons simultanées entre les variables. On peut aborder l'ACP sous différents angles, dont les deux principaux sont :

- Soit la recherche d'un ensemble réduit de variables non-corrélées, appelées « ***Composantes Principales***», qui sont les combinaisons linéaires des variables initiales résumant au mieux l'ensemble des variables initiales.
- Soit la recherche de représentations graphiques, appelées aussi « ***mapping*** », où les individus et les variables constituent des nuages de points dont on saura interpréter les proximités en terme de liaison statistique ou en terme de similitude.

II.1.2. Les principaux calculs de l'ACP

Les calculs à mener ont toujours une interprétation concrète, soit en terme de géométrie (une composante principale est une direction de grand allongement du nuage de points) soit en terme statistique (la coordonnée d'un point est une corrélation). Ces calculs conduisent aux représentations graphiques du nuage des individus et du nuage des variables, ainsi qu'aux divers critères permettant de juger de la qualité et de la stabilité des résultats.

II.1.3. Représentation simultanée en ACP

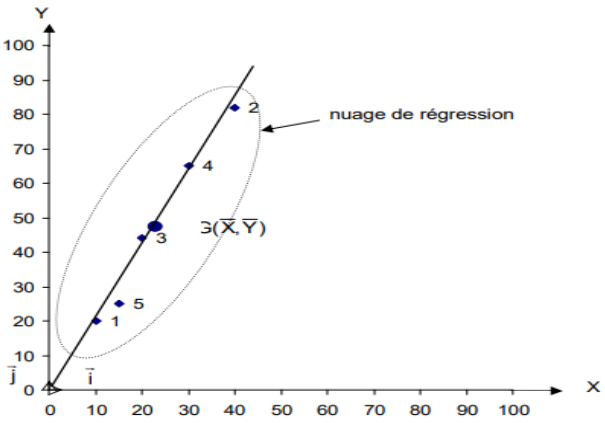
Une analyse factorielle fournit un nuage de points pour les individus (lignes du tableau) et un autre nuage pour les variables (colonnes du tableau). On connait moins bien la possibilité de faire une représentation simultanée des individus et des variables en ACP. Il est même fréquent, lorsqu'il en est question, de trouver des représentations erronées consistant à superposer sans précaution les deux nuages de points.

Exemple :

On considère le tableau suivant :

Individus\variables	Y	X
1	20	10
2	82	40
3	44	20
4	65	30
5	25	15
Somme	236	115

La représentation graphique des individus dans l'espace R^2 des deux variables, en utilisant une base orthonormée $\left((\vec{i},\vec{j}),\|\vec{i}\|=\|\vec{j}\|=1,\vec{i}*\vec{j}=0\right)$, conduit au nuage des individus (nuage de régression) suivant :



Avec la régression, il est parfois possible de visualiser l'information contenue dans le nuage de régression (les proximités relatives des 5 points).

II.2. Analyse Factorielle des Correspondances (AFC) - Avec des données qualitatives, on fait de l'AFC

L'analyse factorielle des correspondances (AFC) est une méthode statistique d'analyse des données, ou plus précisément une méthode exploratoire d'analyse des *tableaux de contingence*. On l'appelle également analyse des correspondances simples.

- ***AFC*** (analyse factorielle des correspondances) et l'***ACM*** (analyse des correspondances multiples) **=** Étude des liaisons contenues dans un tableau individus*variables, lorsque toutes les variables sont qualitatives.
- L'**AFC** est l'étude des correspondances entre les modalités de deux variables qualitatives.
- L'**ACM** est une généralisation de l'**AFC** à plus de deux variables qualitatives.

II.2.1. But de l'AFC

Résumer et représenter les principales liaisons pouvant exister entre les modalités de 2 variables qualitatives.

II.2.2. Méthode

Réduction de la dimension en effectuant la décomposition factorielle des nuages de points associés aux profils lignes et aux profils colonnes du tableau de contingence croisant les modalités des deux variables (L'AFC est une double ACP sur les deux tableaux de profils).

II.2.3. Données

Dans toute la suite, on considère un tableau de deux variables qualitatives X et Y , observées simultanément sur n individus telles que

- X a I modalités.
- Y a J modalités.
- On suppose I > J.

II.2.4. Définitions et notations

- On dispose d'un nombre important d'observations. Elles sont caractérisées par deux variables, éventuellement trois si les croisements de deux d'entre elles peuvent être considérés comme nouvelles variables. Il est possible de croiser plus de variables mais il faut alors utiliser alors la procédure de l'ACM (analyse des correspondances multiples) et non l'AFC.

Table de contingence (permet d'analyser la correspondance entre X et Y)

X / Y	y_1		y_j		y_J	Total
x_1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
x_i	n_{i1}		n_{ij}		n_{iJ}	$n_{i.}$
x_I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

n_{ij} = Nombre d'observations ayant la modalité xi de x et yj de y.
$n_{i.}$ = effectif marginal : Nombre d'observations ayant la modalité xi de x
$n_{.j}$ = effectif marginal :Nombre d'observations ayant la modalité yj de y

- Ces observations sont synthétisées dans un tableau de contingence. Dans chaque case figure le nombre d'unités statistiques qui présentent telle modalité d'une première variable placée en ligne et telle modalité d'une autre variable présentée en colonne. Si les variables sont numériques, elles doivent être découpées en classes.

- Quand il y a peu de modalités, un simple coup d'œil suffit pour estimer s'il y a liaison et une brève présentation de pourcentages fait l'affaire. Afin d'évaluer de façon plus sérieuse si globalement il existe des liens, il faut procéder à un test d'indépendance du khi².
- En revanche, lorsque les modalités sont très nombreuses, une montée en puissance s'avère nécessaire. C'est ainsi que l'on passe des statistiques descriptives à l'analyse des données, comme on passe de la vrille à la perceuse…
- Que faire de toutes ces données ? Les passez aux rayons X pour découvrir l'invisible, c'est-à-dire les liens qui existent dans une pléthore de modalités.
- Evidemment, et contrairement à l'ACP, on ne peut pas visualiser le nuage représentatif des individus à l'aide d'un système de coordonnées cartésiennes puisque la population est définie par des critères nominaux. Mais comme toute analyse factorielle qui se respecte, l'analyse des correspondances permet de visualiser des liaisons sur un ou deux plans factoriels car on va définir une métrique…

II.2.5. Comment ça se passe ?

Il existe différentes façons d'expliquer le mécanisme de l'AFC et plusieurs algorithmes sont possibles (à partir des données brutes, des centrées réduites, des fréquences…). Rassurez-vous, c'est votre logiciel qui s'occupe de tout.

D'abord, le tableau des effectifs observés est comparé à un tableau des effectifs théoriques dans lequel il n'y aurait aucun lien entre telle modalité de la variable **x** et telle modalité de la variable **y**, comme pour le test du χ^2. Ce tableau d'effectifs théoriques est construit à partir des fréquences marginales.

Ces deux tableaux, réel et théorique, vont bénéficier des joies du calcul matriciel, ainsi que le tableau qui en montre les différences (soustraction réelle – théorique). C'est bien sûr cette troisième matrice que l'on va explorer en la décomposant en une somme de nn matrices (**n** étant le nombre de colonnes moins 1. S'il y a plus de colonnes que de lignes, **n** est alors le nombre de lignes moins 1). Ces nouvelles décompositions sont réalisées de la même façon que la première. Les tableaux obtenus les uns à partir des autres expliquent ainsi de moins en moins d'inertie, comme une orange qui donne moins de jus à chaque tour de presse-agrume.

A chacune de ces matrices est associé un vecteur propre.

Aux vecteurs propres qui déterminent les directions des axes factoriels sont associées des valeurs propres qui indiquent dans quelle mesure chaque axe explique une part de l'inertie totale (égale à la somme de ces valeurs propres). La métrique utilisée pour mesurer les éloignements ou proximités entre modalités est la distance du χ2 (pour être plus précis du phi², c'est-à-dire$\frac{\chi 2}{n}$.

A chaque tableau résiduel est associé un test d'indépendance du χ^2 et le ϕ^2 obtenu est égal à la valeur propre.

II.2.6. L'interprétation

Chaque axe factoriel supporte une part de l'inertie totale. Cette part est mesurée par les valeurs propres, inférieures ou égales à 1. Des valeurs proches de 1 indiquent d'intéressants liens entre modalités de variables différentes. Donc, la seule valeur propre du premier axe donne une bonne indication sur l'intérêt de l'étude…

Après la qualité globale de l'AFC, l'explication fournie par chaque axe. En d'autres termes, après la taille du nuage, voyons sa forme. Votre logiciel vous donne le pourcentage d'inertie expliqué par les axes factoriels. Vous en retenez un certain nombre en fonction des pourcentages cumulés et de leur courbe représentative qui peut montrer une inflexion.

On observe ensuite les contributions.

Puis on visualise les proximités et éloignements de modalités sur les plans factoriels. L'origine est le centre de gravité, aussi bien des lignes que des colonnes et, même si les échelles des deux variables sont différentes, on peut rapprocher facilement les modalités qui peuvent l'être (inter ou intra variables).

Cela étant, s'il n'existe pas de structure dans vos données, l'AFC ne l'inventera pas ! (voir Résultats d'une AFC peu informative).

Enfin, vous rédigez un beau rapport.

II.3. Analyse discriminante

L'**analyse** factorielle **discriminante** (**AFD**) ou simplement **analyse discriminante** (**AD**) est ***une technique statistique*** qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire…) d'un ensemble d'observations (individus, exemples…) à partir d'une série de variables prédictives (descripteurs, variables exogènes…).

L'analyse discriminante est à la fois une ***méthode*** prédictive (**analyse discriminante linéaire** – **ADL**) et descriptive (**analyse factorielle discriminante** – **AFD**). Elle vise à expliquer et prédire l'appartenance des individus à des groupes (des classes), représentés par une variable cible catégorielle, à partir d'une collection de variables explicatives/descriptives, principalement quantitatives, mais qui peuvent être qualitatives moyennant un aménagement. Cette méthode, à la fois explicative et prédictive, peut être utilisée pour :

- Vérifier sur un graphique à deux ou trois dimensions si les groupes auxquels appartiennent les observations sont bien distincts,
- Identifier quelles sont les caractéristiques des groupes sur la base de variables explicatives,
- Prédire le groupe d'appartenance pour une nouvelle observation.

II.3.1.Approche Statistique

Du point de vue des calculs, l'analyse discriminante est très proche de l'analyse de variance (***ANOVA***). Considérons un exemple simple. Supposons que nous mesurons la taille en centimètres dans un échantillon aléatoire de 50 hommes et de 50 femmes. Les femmes sont, en moyenne, plus petites que les hommes, et cette différence va se refléter dans les moyennes (de la variable *Taille*). Par conséquent, la variable "taille" nous permet de mieux discriminer les hommes des femmes que le simple hasard : si une personne est grande, il s'agit plus probablement d'un homme ; en revanche, si elle est petite, il s'agit plus vraisemblablement d'une femme.

Il est possible de généraliser ce raisonnement à des groupes et des variables moins "triviales". Par exemple, supposons que nous ayons deux groupes de bacheliers jeunes diplômés : ceux qui choisissent de poursuivre vers l'université après le baccalauréat et ceux qui envisagent une autre orientation. Nous pourrions avoir demandé aux étudiants leur intention de poursuivre ou non à l'université un an avant l'obtention de leur diplôme. Si les moyennes des deux groupes (ceux qui sont effectivement allés à l'université et ceux qui n'y sont pas allés) sont différentes, nous pouvons alors dire que l'intention de rejoindre l'université exprimée un an avant l'obtention du diplôme nous permet de bien discriminer entre ceux qui sont et ceux qui ne sont pas aux portes de l'université (cette information peut être utilisée par des conseillers d'orientation pour offrir des pistes appropriées aux étudiants respectifs).

Pour résumer la présentation faite jusqu'à présent, l'idée de base d'une analyse discriminante est de déterminer si des groupes sont différents par rapport à la moyenne qu'ils prennent sur une variable particulière, et d'utiliser cette variable pour prédire l'appartenance à un groupe (par exemple, pour de nouvelles observations).

II.3.2.Analyse Discriminante Pas-à-Pas

L'application la plus courante de l'analyse discriminante consiste à inclure de nombreuses mesures dans une étude, afin de déterminer celles qui discriminent les groupes. Par exemple, un enseignant, intéressé par la prévision des choix des nouveaux bacheliers en matière

d'éducation va probablement inclure autant de mesures de personnalité, de motivation, de performances scolaires, etc... que possible afin de savoir quelles sont celles qui donnent la meilleure prévision.

Modèle : En d'autres termes, nous souhaitons construire un "modèle" qui nous donnera la meilleure prévision de l'appartenance d'une observation à un groupe. Par la suite, nous utiliserons les termes "dans le modèle" (ou inclues/présentes dans le modèle) pour désigner les variables qui sont prises en compte dans la prévision de l'appartenance à un groupe, et "pas dans le modèle" (ou exclues/absentes du modèle) pour celles qui ne le sont pas.

Analyse pas-à-pas ascendante : Dans l'analyse discriminante pas-à-pas, *STATISTICA* "construit" un modèle de discrimination pas-à-pas. Plus précisément, à chaque étape, *STATISTICA* examine toutes les variables et évalue celle qui contribue le plus à discriminer les groupes. Cette variable est alors introduite dans le modèle, puis *STATISTICA* passe à l'étape suivante.

Analyse pas-à-pas descendante : Il est également possible de procéder en sens inverse ; dans ce cas, *STATISTICA* va tout d'abord inclure toutes les variables dans le modèle, puis, à chaque étape, éliminer la variable qui contribue le moins à la prévision d'appartenance à un groupe. Par conséquent, une analyse discriminante réussie doit conserver, en dernier ressort, les variables "importantes" dans le modèle, c'est-à-dire celles qui contribuent le plus à discriminer les groupes.

F* d'inclusion, *F* d'exclusion :** La procédure pas à pas est "guidée" par les valeurs respectives du ***F d'inclusion et du ***F*** d'exclusion. La valeur du *F* d'une variable indique sa significativité statistique dans la discrimination des groupes, c'est-à-dire qu'il s'agit d'une mesure du degré auquel une variable contribue, de façon unique (à elle seule) à la prévision de l'appartenance à un groupe. Si vous êtes familiarisé(e) avec les procédures de régression multiple pas-à-pas, vous pouvez interpréter les valeurs du ***F*** d'inclusion/d'exclusion de la même manière que dans une régression pas à pas.

D'une manière générale, *STATISTICA* continue à intégrer des variables dans le modèle, tant que les valeurs respectives du *F* pour ces variables demeurent supérieures à la valeur du *F* d'inclusion que vous aurez spécifiée ; et *STATISTICA* exclut (supprime) les variables du modèle si leur significativité est inférieure au *F* d'exclusion spécifié.

II.4. Classification

Les **méthodes de classification**, aussi appelées de partition des données, permettent de grouper des objets (observations ou individus) dans des classes (clusters) de manière à ce que les objets appartenant à la même classe sont plus similaires entre eux qu'aux objets appartenant aux autres classes.

II.4.1. Définition

- ✓ **Données** = tableau n*p individus*variables.
- ✓ **Objectif** = recherche d'une typologie ou segmentation, c'est-à-dire d'une partition ou répartition des n individus dans des classes, sur la base de l'observation de p descripteurs.
- ✓ **Moyen** = chaque classe doit être la plus homogène possible et, entre elles, les plus distinctes possibles, au sens d'un critère à définir.
- ✓ **Algorithmes** qui permettent de classer des objets observes dans des classes (appelées **clusters**).
- ✓ Les **objets** d'une même classe doivent être "similaires" et les objets de deux classes différentes doivent être "**distincts**".
- ✓ Différents algorithmes de classification :
  - Méthodes Hiérarchiques (classification hiérarchiques...).
  - Partitionnement en k-moyennes (ou **k-means** en anglais).
  - méthodes Probabilistes (EM-algorithme...) et encore d'autres.

II.4.2. Qu'est-ce que la classification ?

Regrouper des objets en groupes, ou classes, ou familles, ou segments, ou clusters, de sorte que :
2 objets d'un même groupe se ressemblent le + possible
2 objets de groupes distincts diffèrent le + possible le nombre des groupes est parfois fixé

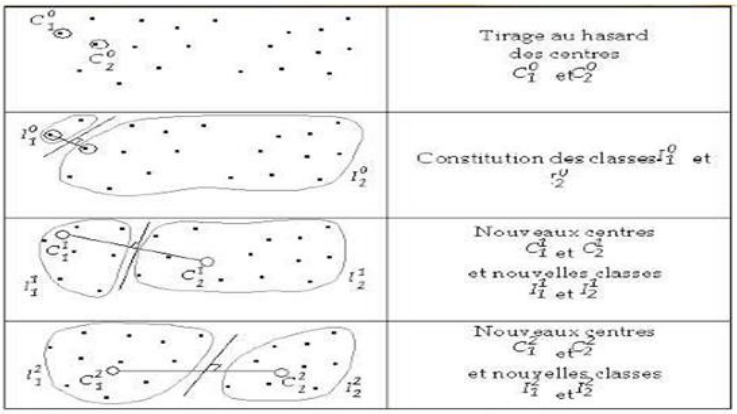
II.4.3. Les méthodes

- Méthodes hiérarchiques
  - ascendantes (agglomératives)
    - basées sur une notion de distance ou de densité
  - descendantes (divisives)
- Méthodes de partitionnement
  - centres mobiles, k-means et nuées dynamiques
  - k-modes, k-prototypes, k-représentants (k-medoids)
  - réseaux de Kohonen
  - méthodes basées sur une notion de densité
  - méthode « de Condorcet » (analyse relationnelle)
- Méthodes mixtes
- Analyse floue (fuzzy clustering)

II.4.4. Différents algorithmes de classification :

A. *Algorithmes de partitionnement :* *Centres mobiles (Forgy):*

- ✓ Initialisation : Choix aléatoire de k points de l'espace (centres des classes)
- ✓ Itérer les deux étapes suivantes jusqu'à ce que le critère à minimiser (inertie intraclasses) ne décroisse plus de manière significative (minimum local), ou bien jusqu'à atteindre un nombre d'itérations fixées:
  - Tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie. On construit ainsi k classes d'individus.
  - On calcule les barycentres des classes créées qui deviennent les k nouveaux centres

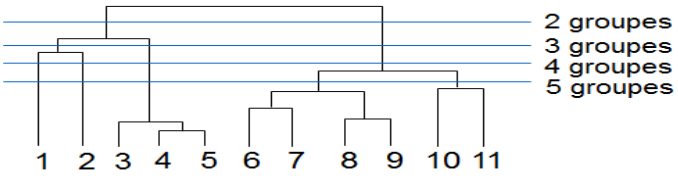


Source : Ludovic Lebart – *Analyse des données appliquée* - 2002

B. *Algorithmes hiérarchiques:*

La **C**lassification **A**scendante **Hiérarchique** (CAH) est une **méthode** de classification qui présente les avantages suivants : On travaille à partir des dissimilarités entre les objets que l'on veut regrouper. On peut donc choisir un type de dissimilaire adapté au sujet étudié et à la nature des données.

Méthodes cherchent plutôt à obtenir une *hiérarchie de regroupements*, qui fournit une information plus riche concernant la structure de similarité des données. Noter qu'à partir d'une telle hiérarchie il est facile d'extraire plusieurs partitionnements, à des niveaux de « granularité » différents, comme dans l'exemple suivant :



Exemple de regroupement hiérarchique qui peut servir à obtenir plusieurs partitionnements

Bibliographie:

- El Marhoum, A.(2005): «Analyse des données ». Toubkal.
- Hotelling, (1933) . Analysis of a complex of statistical variables into principal components.
- Labrousse, C. (1976): «Introduction à l'éconmétrie ». Dunod.
- Saporta, G. (1990): «Probabilités, Analyse des données et Statistique ». TECHNIP.