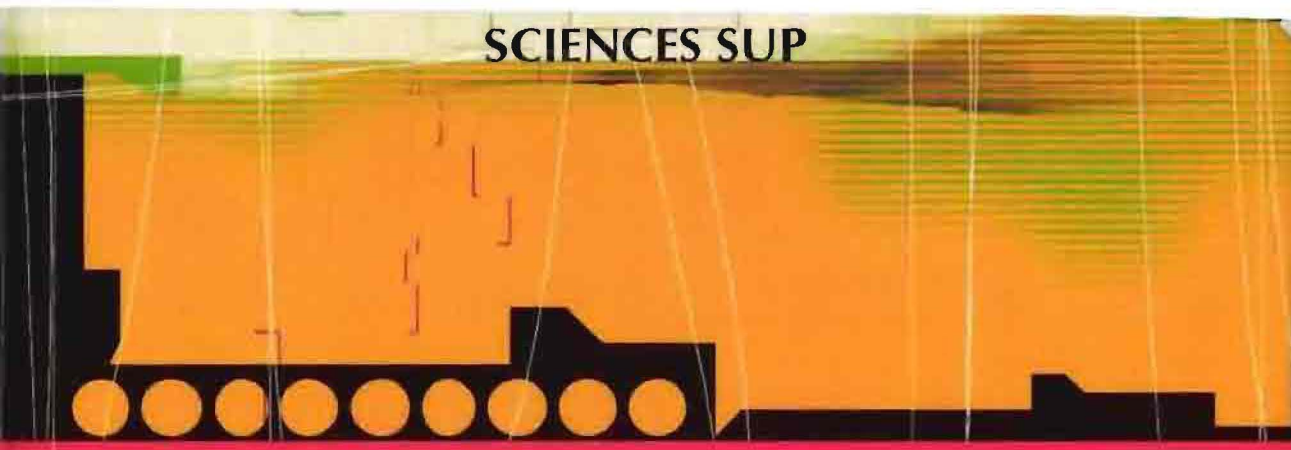
SCIENCES SUP



Cours et exercices corrigés

Masters • Écoles d'ingénieurs

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisation et inférence en fouilles de données

4e édition

Ludovic Lebart
Marie Piron
Alain Morineau

DUNOD

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisations et inférences en fouille de données

Ludovic Lebart
Directeur de recherches CNRS
à l'École nationale supérieure des télécommunications (ENST)

Marie Piron
Chargée de recherche
à l'Institut de recherche pour le développement (IRD)

Alain Morineau
Chercheur au Centre international de statistiques
et d'informatique appliquées (CISIA)

4e édition

DUNOD

Illustration de couverture : *Digitalvision*®

ISBN 978-2-10-049616-7

Cet ouvrage est dédié à la mémoire de Brigitte Escofier, Jean-Pierre Fénelon, *et* Chikio Hayashi, *pionniers de l'exploration statistique des données.*

AVANT-PROPOS

Cet ouvrage s'adresse aux praticiens, scientifiques et étudiants de toutes disciplines qui ont à analyser et traiter de grands ensembles de données multidimensionnelles, c'est-à-dire des recueils de données statistiques se présentant, totalement ou partiellement, sous forme de tableaux rectangulaires.

Le domaine d'application, limité au départ aux sciences de la vie (biométrie, agronomie, écologie) et aux sciences humaines (psychométrie, socio-économie), s'est étendu en raison des possibilités offertes par les outils de calcul qui ont suscité de nouveaux recueils de mesures et de nouvelles exigences de résultats. Les applications industrielles se développent rapidement et le contrôle de qualité, l'analyse des processus de production, la veille technologique, la recherche documentaire font de plus en plus appel à des ensembles de mesures multidimensionnelles.

Nous avons tenté de faire le point sur les développements récents de la *statistique exploratoire multidimensionnelle* en s'efforçant d'intégrer la substance de plusieurs centaines de publications. Ces analyses et ces traitements se rattachent au champ disciplinaire de l'extraction des connaissances à partir de données (ECD) désigné également par l'expression *Data Mining*, traduite souvent par *Fouille de Données*, termes retenus dans le sous-titre de cette nouvelle version de l'ouvrage.

Cette édition est entièrement refondue, révisée et complétée : le plan même de l'ouvrage a été modifié de façon à pouvoir accueillir les développements récents dans un cadre adapté.

Comme toujours pour ce type d'ouvrage qui s'adresse simultanément à des praticiens et des chercheurs de disciplines diverses, plusieurs lectures devraient être possibles selon les connaissances du lecteur notamment en mathématique et statistique : une lecture pratique, d'utilisateur, pour les personnes spécialisées dans les divers domaines d'application actuels et potentiels ; une lecture plus technique, complète, pour une personne ayant une formation en mathématique appliquée et en statistique.

La statistique exploratoire multidimensionnelle se prolonge naturellement et se diversifie en des outils et des modèles évidemment plus complexes que les méthodes de base. Mais l'essentiel des applications relèvent en fait de la partie la plus accessible. On a fait preuve d'une grande parcimonie dans l'utilisation de l'outil mathématique : le niveau d'abstraction choisi est toujours le niveau minimal compatible avec une présentation exacte, et la communication a été favorisée au détriment de la généralisation. Les lecteurs mathématiciens sauront sans difficulté ajouter les notions et notations qui permettent des formulations parfois plus élégantes.

L'ensemble doit beaucoup à des collaborations et des cadres de travail divers : à l'Ecole Nationale Supérieure des Télécommunications, au sein du département Sciences Economiques et Sociales et du Laboratoire de traitement et communication de l'information (LTCI) du Centre National de la Recherche Scientifique (UMR 5141 du CNRS), et à l'Institut de Recherche pour le Développement (UR013 de l'IRD).

Nous remercions les collègues, chercheurs ou professeurs, auprès desquels nous avons puisé collaboration et soutien, ou simplement eu d'intéressants débats ou discussions, ou encore accès à des documents. Citons, sans être exhaustif, Mireille Bardos, Abdelhalim Bouamaine, Bernard Burtschy, Pierre Cazes, Fréderic Chateau, Christian Mullon, Jérôme Pagès, André Salem, Gilbert Saporta, Michel Tenenhaus et Wenhua Zhu. Les auteurs sont redevables à Karmele Fernández Aguirre et François Micheloud pour les corrections apportées lors de la seconde, puis de la troisième édition.

Une mention particulière doit être faite de la collaboration avec Belaïd Ghermani (Université Paris XII), qui fût un lecteur pointilleux de la quatrième édition en même temps qu'un critique avisé.

Nous sommes enfin heureux de remercier Anne Bourguignon, des Editions Dunod, pour l'accueil qu'elle a réservé à cette nouvelle version de l'ouvrage.

L. L., M. P.

Paris, Mars 2006

Sommaire

Introduction

La statistique descriptive permet de représenter de façon vivante et assimilable des informations statistiques en les simplifiant et les schématisant. La *statistique descriptive multidimensionnelle* en est la généralisation naturelle lorsque ces informations concernent plusieurs variables ou dimensions.

Mais le passage au multidimensionnel induit un changement qualitatif important. On ne dit pas en effet que des microscopes ou des appareils radiographiques sont des instruments de description, mais bien des instruments d'observation ou d'exploration, et aussi des outils de recherche. La réalité multidimensionnelle n'est pas seulement simplifiée parce que complexe, mais aussi explorée parce que cachée.

Le travail de préparation et de codage des données, les règles d'interprétation et de validation des représentations fournies par les techniques utilisées dans le cas multidimensionnel n'ont pas la simplicité rencontrée avec la statistique descriptive élémentaire. Il ne s'agit pas seulement de présenter mais d'analyser, de découvrir, parfois de vérifier et prouver, éventuellement de mettre à l'épreuve certaines hypothèses. C'est pourquoi nous avons choisi de parler plutôt dans cet ouvrage de *statistique exploratoire multidimensionnelle*. Ces méthodes occupent de plus une place centrale dans la démarche « *Data mining* » traduite ici par « Fouille de Données », en permettant des visualisations fondées sur des principes géométriques et algébriques simples sous le contrôle de méthodes inférentielles souples et robustes.

La statistique du vingt et unième siècle

Née au tout début du vingtième siècle, notamment à la suite des travaux de précurseurs comme l'astronome Quételet et les démographes et biométriciens Galton, Pearson, puis Fisher, la science statistique aura manipulé des chiffres pendant un demi-siècle sans disposer de véritables outils de calcul. Les appareils que l'on trouve maintenant dans la poche des écoliers et dans tous les bureaux auraient comblé les aspirations les plus insensées des statisticiens

jusqu'en 1960. "Il est impensable d'utiliser des méthodes conçues avant l'avènement de l'ordinateur, il faut complètement réécrire la statistique", écrivait en substance Jean-Paul Benzécri dès 1965 dans son cours à la Sorbonne sur *l'Analyse des données et la reconnaissance des formes*.

Cet auteur, qui a profondément marqué le développement des recherches statistiques au cours de la seconde moitié du vingtième siècle, préconise aussi, de manière un peu provocante pour une discipline où la notion de modèle a joué un rôle central : "le modèle doit suivre les données et non l'inverse".

Aux États-Unis, John Tukey, le fondateur du courant désigné par *Exploratory Data Analysis* (*EDA*), a une attitude aussi radicale (cf. Mallows et Tukey, 1982). Il s'en faut cependant de beaucoup que ces deux pionniers aient été unanimement entendus. A défaut d'être repensée, la statistique s'est cependant considérablement enrichie. La période récente a connu des changements tout à fait notables du fait de la diffusion des moyens de calcul : les outils existants ont été améliorés, de nouveaux outils sont apparus, de nouveaux domaines d'application ont été explorés. Passons en revue quelques-unes de ces innovations avant de présenter l'ouvrage lui-même.

Meilleurs graphiques

L'informatique, surtout la micro-informatique, a popularisé tous les outils graphiques de la statistique descriptive élémentaire. Autrefois fruits d'un travail laborieux et coûteux, ces représentations sont immédiatement accessibles dans pratiquement toutes les suites logicielles (*open office, microsoft office* ®). Les techniques de statistique exploratoire multidimensionnelle mettent à profit ces interfaces graphiques pour représenter, par exemple, les espaces factoriels et les arbres de classification : c'est là l'une de leurs fonctions iconographiques qui généralise la statistique descriptive au cas de variables nombreuses.

Désuétude des tables statistiques

Classiquement, pour savoir si une quantité, dont la distribution est connue, ne dépasse pas les limites que lui assignent certaines hypothèses, on consultait la table donnant les valeurs que cette quantité ne dépassera que dans 5% ou 1% des cas. Le choix de seuils était imposé par la nécessité de limiter le volume des tables. A partir du moment où la quantité à tester est elle-même calculée sur ordinateur, il est facile d'adjoindre au programme une procédure le calcul de la probabilité de dépassement de la valeur calculée. On pourra désormais comparer et trier des statistiques grâce aux probabilités de dépassement.

Emphase sur la robustesse, le non-paramétrique

La mise en œuvre de la plupart des procédures inférentielles classiques est hypothéquée par la pertinence des hypothèses techniques et par la sensibilité éventuelle des résultats à la non-vérification de ces hypothèses. Contrairement aux hypothèses générales qui sont les hypothèses d'ordre scientifique qui régissent l'étude d'un phénomène et qui précèdent la phase d'observation ou d'expérimentation statistique, les hypothèses techniques interviennent dans la

mise en œuvre pragmatique des méthodes statistiques. Elles concernent principalement la spécification des modèles et des distributions statistiques impliquées dans ces modèles. Certaines hypothèses techniques n'ont aucun lien avec les hypothèses générales, mais sont au contraire simplement des exigences du modèle utilisé[1]. L'un des principaux obstacles à l'utilisation d'estimateurs robustes, c'est-à-dire peu sensibles à la présence de points aberrants (vis-à-vis des distributions étudiées), était la difficulté des calculs à mettre en œuvre. Les panoplies existantes se sont donc enrichies de procédures plus robustes dès l'apparition de moyens de calcul plus puissants. Pour des raisons analogues, les techniques non-paramétriques qui s'affranchissent des hypothèses les plus lourdes ont connu un regain d'intérêt, comme ce fût le cas des techniques non-paramétriques de discrimination. Les test "Fishériens", ou tests de permutation[2], connaissent également un renouveau important. Les hypothèses statistiques sont éprouvées par permutations aléatoires de l'ensemble fini des observations effectivement disponibles : il y aura donc coïncidence entre les distributions théoriques et observées. Seul l'obstacle du calcul pouvait faire écarter des techniques fondées sur des hypothèses qui épousent aussi étroitement la réalité.

Méthodes de validation

Les techniques de simulation (ou de Monte-Carlo) connaissent des applications à grande échelle dans tous les domaines où les hypothèses distributionnelles usuelles sont inadaptées. La simulation permet de construire de l'inférence "sur-mesure" en combinant des sources, des formes et des niveaux de variabilité dans des processus complexes dont la formalisation est impossible. Les techniques de rééchantillonnage telles que les techniques de "*bootstrap*" (la variabilité est étudiée en procédant à des tirages pseudo-aléatoires avec remise dans l'échantillon) ont le mérite d'avoir donné lieu à des développements théoriques. Comme on le verra, le *bootstrap*, qui présente de notables avantages (facilité de mise en œuvre, propriétés théoriques satisfaisantes) est largement utilisé. Les techniques de *validation croisée* sont surtout utilisées en analyse discriminante (famille des méthodes dites *supervisées*).

Taille et complexité des problèmes

Il n'est pas rare maintenant de traiter des tableaux correspondant à des dizaines de milliers d'observations et des centaines, voire des milliers de variables. Très vite, l'adage: "c'est l'échelle qui fait le phénomène" s'est trouvé vérifié. Le changement d'échelle des données a rapidement conduit à modifier les outils eux-mêmes et à imaginer de nouveaux outils et de nouvelles approches. Mais la statistique nous rappelle qu'il est parfois vain de vouloir traiter des millions d'observations lorsqu'il y a des possibilités d'échantillonnage.

---

[1] Exemple: dans le cas de la régression linéaire multiple, les résidus doivent (le plus souvent) être indépendants et suivre une loi normale.

[2] Cf. sur les tests dits "exacts" : Mehta *et al.* (1991), Agresti (1992), Good (1994).

Méthodes algorithmiques

La levée de l'obstacle du calcul a eu pour effet de diffuser l'emploi des techniques de type algorithmique, au premier rang desquelles se trouvent les techniques de classification automatique et les méthodes impliquant des algorithmes coûteux. D'autres techniques, comme les techniques de sélection pas-à-pas, les techniques d'estimation par la méthode du maximum de vraisemblance, de programmation dynamique, de recherches automatiques de règles dans des bases de données connaissent des utilisations à grande échelle.

Traitement des variables qualitatives

L'étude statistique des variables qualitatives est par nature plus complexe que celle des variables numériques continues, qui s'appuie généralement sur la loi normale et sur les formalismes simples qui en dérivent (maximum de vraisemblance, moindres carrés, par exemple). Il n'est donc pas étonnant que les possibilités de calcul aient permis de fortes avancées dans ce domaine : analyse des correspondances simples et multiples dans le cas descriptif, modèles log-linéaires, discrimination et modèles logistiques dans le cas inférentiel.

Réseaux neuronaux

Les techniques neuronales ou connexionnistes ont une large intersection avec les méthodes classiques d'analyse des données [1], intersection peu visible de prime abord en raison d'une terminologie et d'un cadre conceptuel tout à fait spécifiques. Inspirées à l'origine par des modèles de fonctionnement du cerveau, les méthodes connexionnistes peuvent être considérées comme des méthodes d'analyse non-linéaire des données. L'analyse en composantes principales, les méthodes de classification du type k-means ou nuées dynamiques sont des méthodes neuronales non supervisées ; la régression, l'analyse discriminante linéaire sont des cas particuliers de méthodes neuronales supervisées. Cette terminologie bipolaire supervisé / non-supervisé appartient à la théorie de l'apprentissage (cf. Vapnik, 1995, 1998 ; Hastie *et al.*, 2001) à laquelle les méthodes neuronales se rattachent. Elle est proche de la bi-polarité : descriptif / inférentiel, ou encore exploratoire / confirmatoire.

Les logiciels

Une des innovations de la statistique moderne aura été la matérialisation des techniques sous forme de "produits", les logiciels, développés avec des contraintes économiques et commerciales de conception, de production, de distribution. Comme tout produit fini, le logiciel a l'avantage de diffuser et l'inconvénient de figer. Comme tout produit coûteux, il introduit une discrimination par les moyens financiers disponibles. Comme tout produit à

[1] L'expression anglaise *data analysis* a un sens très général de statistique appliquée (avec une connotation d'approche pragmatique et informatisée). L'équivalent anglais de l'analyse des données serait à peu près *exploratory multivariate data analysis.*

l'usage de spécialistes, il induit de nouvelles divisions du travail, parfois peu souhaitables dans un processus de connaissance. Si cette division du travail se fait à l'échelle internationale, de nouvelles dépendances sont créées dans des secteurs sensibles : l'acquisition de connaissances, la recherche fondamentale. Ces avantages et inconvénients sont indissolublement liés dans les logiciels statistiques. Les logiciels accessibles et faciles à utiliser permettront une large diffusion des méthodes, mais donneront parfois lieu à des utilisations inconsidérées dans des domaines où une réflexion minutieuse et une grande prudence seraient de mise. La médiation des logiciels est un nouveau paramètre dont il faut tenir compte.

L'interactivité

Obtenir un arbre de longueur minimale dans un plan factoriel en cliquant sur un bouton, ou obtenir par un autre clic les plus proches voisins d'un individu, ou le profil de base de cet individu, voire le texte intégral de sa réponse à une question ouverte, sont des opérations devenues banales. L'exploration n'est plus la lecture d'un rapport, mais un cheminement complexe dont seules les étapes marquantes seront retenues. Les possibilités de l'animation en matière de description de données commencent à être exploitées.

Nouveaux domaines d'application, Data Mining ou « Fouille de Données »

L'informatisation et les outils qu'elle a suscité ou dont elle a stimulé le développement (gestionnaires de base de données relationnelles, systèmes d'informations géographiques par exemple) ont pour effet le plus évident de permettre le traitement statistique de recueils plus grands et plus complexes, donnant lieu à de véritables systèmes d'information. Les méthodes d'analyse des données sont des outils performants pour exploiter au mieux la structure organisée de ces systèmes. On peut citer parmi les domaines récemment abordés : les analyses d'images, de séquences d'images (données de télédétection par exemple); les analyses de signaux, de processus, de systèmes; la recherche documentaire; les analyses de données textuelles ; les analyses de grandes enquêtes, enfin les données transactionnelles des entreprises.

L'approche *Data Mining* (cf. par exemple : Berry et Linoff, 1997 ; Tuffery, 2005) correspond en bref à un traitement exploratoire de grands tableaux pas ou peu structurés à partir d'algorithmes de recherche de règles ou de méthodes d'analyses des données (méthodes factorielles de base, segmentation, classification). Les logiciels correspondants sont souvent conçus à l'usage des données transactionnelles du monde industriel. Ils ont le mérite de faire le pont entre les bases de données et la statistique, dont la panoplie de méthodes n'est pas directement interfacée avec les données des gestionnaires.

Le contexte de la statistique multidimensionnelle

On distingue deux types d'approches en statistique multidimensionnelle : les *approches descriptives et exploratoires* (qui sont souvent les approches non-supervisées de la théorie de l'apprentissage) et les *approches inférentielles et*

confirmatoires (dont font partie les approches supervisées) qui constituent le volet le plus ample et le plus classique de la science statistique.

Rappelons brièvement les caractéristiques de ces deux familles de méthodes, qui correspondent à des approches complémentaires.

- *La statistique descriptive et exploratoire* permet, par des résumés et des graphiques plus ou moins élaborés, de décrire des ensembles de données statistiques, d'établir des relations entre les variables sans faire jouer de rôle privilégié à une variable particulière. Classiquement, les conclusions ne portent dans cette phase de travail que sur les données étudiées, sans être inférées à une population plus large. L'analyse exploratoire s'appuie essentiellement sur des représentations graphiques et sur les techniques descriptives multidimensionnelles (analyse en composantes principales, analyse des correspondances, classification). Les méthodes de ré-échantillonnage actuelles permettent de valider des structures et donc d'articuler exploration et inférence.

- *La statistique inférentielle et confirmatoire* permet de valider ou d'infirmer, à partir de tests statistiques ou de modèles probabilistes, des hypothèses formulées *a priori* (ou après une phase exploratoire), et d'extrapoler, c'est-à-dire d'étendre certaines propriétés d'un échantillon à une population plus large. Les conclusions obtenues à partir des données vont au-delà de ces données. La statistique confirmatoire fait surtout appel aux méthodes dites explicatives[1] et prévisionnelles destinées, comme leurs noms l'indiquent, à expliquer puis à prévoir, suivant des règles de décision, une variable privilégiée à l'aide d'une ou de plusieurs variables explicatives (régressions multiples et logistiques, analyse de la variance, analyse discriminante, segmentation, etc.).

Les démarches sont complémentaires, l'exploration et la description devant en général précéder les phases explicatives et prédictives. En effet, une exploration préliminaire est souvent utile pour avoir une première idée de la nature des liaisons entre variables, et pour traiter avec prudence les variables corrélées et donc redondantes qui risquent de charger inutilement les modèles.

Cependant, les démarches elles-mêmes ne sont pas toujours faciles à discerner, à identifier. L'exploration pure est très rare, et correspond à une situation limite et irréaliste, un peu comme les gaz parfaits en physique... car il existe toujours des informations et des connaissances *a priori* sur le tableau de données, et donc des hypothèses générales, des attentes de la part de l'utilisateur. Les instruments d'observation correspondent d'ailleurs eux-mêmes à des modèles généraux : ainsi, les axes factoriels de l'analyse en composantes principales sont proches de ceux de l'analyse factorielle classique des psychologues (cf. section 3.2.9) qui représentent les variables latentes d'un modèle a priori. Inversement,

---

[1] La statistique n'explique rien mais fournit des éléments potentiels d'explication. Aussi le terme de variable explicative ou variable à expliquer n'est sans doute pas le plus judicieux. On dit aussi indépendante et dépendante, ou exogène et endogène. Ces deux derniers termes sont plus adéquats mais peu évocateurs. Le terme *indépendant* est, cependant, source de confusions.

la régression multiple, méthode explicative par excellence, peut aussi être utilisée pour explorer des structures de corrélation.

D'où l'intérêt d'éclaircir cette relation entre *instruments d'observation* et *modèles*, en insistant sur l'insertion, théorique et pratique, des outils exploratoires dans l'arsenal des techniques statistiques disponibles[1].

Cet ouvrage voudrait précisément montrer que l'exploration ne se réduit pas à une phase de contemplation des données, et que des inférences sont possibles pour valider les structures observées. L'étude de *la validité et la portée des résultats* a donné lieu à des recherches nombreuses au cours des dernières années. La validation et l'inférence statistique constituent toujours un domaine de recherche en pleine effervescence, mais elles ne sont plus un complément coûteux et facultatif : elles sont devenues indispensables dans les applications quotidiennes et constituent une exigence des utilisateurs éclairés. Contrairement à la disposition des versions antérieures de cet ouvrage, cette étude ne constitue plus un chapitre séparé, mais fait partie intégrante de chacun des chapitres consacrés à une famille particulière de méthodes.

Thèmes connexes

D'autres méthodes de description qui ne rentrent pas dans les deux grandes familles étudiées ici (axes principaux et classification) ne seront évoquées que brièvement, comme les méthodes purement graphiques, dévolues à la représentation de tableaux de petites dimensions, les méthodes de sériation, les méthodes de *multidimensional scaling* [2].

Parmi les méthodes purement graphiques, citons, surtout pour leur importance historique, la méthode des visages de Chernoff (1973), pour laquelle chaque visage correspond à un individu et chaque trait du visage à une variable; la méthode des courbes d'Andrews (1972), où les différents paramètres des courbes sont les variables; la méthode des constellations de Wakimoto et Taguri (1978)[3].

Les méthodes de sériations visent à faire apparaître des structures particulières de tableaux par simple réordonnancement de lignes et de colonnes. Pour des exposés de synthèse sur ce sujet, cf. par exemple Arabie (1978), Caraux (1984), Marcotorchino (1987). L'analyse des correspondances peut d'ailleurs contribuer à la résolution de ces problèmes (Hill, 1974). Ces méthodes interviennent

---

[1] On évoquera les difficultés de l'articulation exploration – inférence au chapitre 5, à propos des liens entre analyse des correspondances multiples et modèles log-linéaires.

[2] Cf. Shepard (1974), Kruskal et Wish (1978), Schiffman *et al.* (1981).

[3] Pour cette méthode, après conversion de chaque x_{ij} (valeur de la variable j pour l'individu i) en un $\cos\theta_{ij}$, chaque individu i est représenté par un point du plan complexe comme une somme de variables de modules constants et d'arguments θ_{ij} .

souvent dans des contextes particuliers d'application et sont moins adaptées aux traitements des très grands tableaux.

L'analyse des données symboliques (cf. Bock et Diday (2000)) est un effort pour formaliser la prise en compte de certains types de méta-information en analyse des données, prise en compte faisant actuellement partie de la phase obligatoire de réflexion et de décision concernant le codage adéquat des données brutes. Cette phase est cruciale si les données se présentent sous la forme d'intervalles, d'histogramme, de distribution. Les outils de base utilisables après cette phase sont principalement ceux abordés dans cet ouvrage, aussi considérons-nous, dans l'état actuel des recherches, ces méthodes comme complémentaires plutôt qu'alternatives.

Panorama du contenu de ce manuel

Les avancées et innovations qui ont été évoquées se retrouvent à des degrés divers dans le développement et la mise en œuvre de la statistique exploratoire multidimensionnelle.

La gamme des méthodes qui permettent de décrire et d'explorer des tableaux de données statistiques (tableaux mesures-observations, tableaux de contingence ou tableaux croisés, tableaux de présence-absence ou tableaux d'incidence) est assez étendue.

Les méthodes que nous retenons sont choisies en fonction de *leur aptitude à traiter de tableaux volumineux*, de la *transparence de leur fonctionnement*, de leur bonne insertion dans l'éventail des *méthodes réellement applicables et appliquées*.

Trois grandes familles de méthodes satisfont à ces exigences : les méthodes dites factorielles ou en axes principaux, les méthodes de classification, les méthodes de discrimination (appelées encore classifications supervisées). A ces trois grandes familles on peut ajouter les méthodes d'analyse de données structurées, qui selon les cas, spécifient ou généralisent les précédentes.

Les principes de base

Les principes algébriques et géométriques de base de ces familles de méthodes font l'objet des deux premiers chapitres.

Le premier chapitre traite d'un des théorème de base : la décomposition aux valeurs singulières, avec une brève introduction aux méthodes de validation par rééchantillonnage (techniques de *bootstrap*).

Une autre méthode de base du point de vue théorique est l'*analyse canonique*, inséparable de la *régression multiple* qui en est un cas particulier. Ces deux techniques sont présentées dès le deuxième chapitre parce qu'elles seront sollicitées au cours de la plupart des chapitres suivants, à propos de la procédure très utilisée de projection d'*éléments supplémentaires*, de l'*analyse des correspondances multiples*, de l'*analyse discriminante*, etc. Ce deuxième chapitre

clôture ce que l'on pourrait appeler la partie « rappels » de l'ouvrage, ou encore, plus familièrement, le « ticket d'entrée théorique » pour la suite.

Méthodes factorielles

- *les méthodes factorielles* [1], ou encore *analyses en axes principaux*, opèrent une réduction de certaines représentations "multidimensionnelles" et produisent essentiellement des *visualisations graphiques* planes ou parfois tridimensionnelles des éléments à décrire.

L'analyse en composantes principales (chapitre 3) est la technique de visualisation en axes principaux la plus ancienne et probablement la plus répandue. Sont rattachées à ce chapitre les méthodes que l'on peut considérer comme parentes, comme l'analyse des rangs, l'analyse en facteurs communs et spécifiques, la régression sur composantes principales, l'analyse en composantes indépendantes. L'utilisation de la validation des visualisations par ré-échantillonnage (bootstrap) est présentée de façon plus détaillée dans ce chapitre, dont la lecture est de ce fait indispensable pour la suite.

L'analyse des correspondances (chapitre 4) constitue l'autre technique factorielle fondamentale. La plupart des autres techniques dérivent de ces deux techniques de base pour s'adapter à des domaines d'application spécifiques. L'une des plus utilisées est l'analyse des correspondances multiples (chapitre 5) en raison notamment du caractère répandu dans les applications des grands fichiers de variables nominales.

Méthodes de classification

Les *méthodes de classification non supervisées* (appelées souvent *méthodes de classification* dans le monde francophone) (chapitre 6) produisent des groupements en classes d'objets pour les méthodes de partitionnement, ou en familles de classes hiérarchisées pour les méthodes de classification hiérarchiques. Les éléments à décrire sont groupés de la manière la moins arbitraire possible à partir de leurs vecteurs de description.

On insiste sur la complémentarité de ces méthodes avec les méthodes factorielles, qu'il s'agisse de la possibilité d'appréhender des structures très diverses, ou simplement d'aider à la lecture des résultats. Les cartes auto-organisées (*Self Organising Maps*) qui font historiquement partie des méthodes neuronales sont évoquées dans ce chapitre. Des notions sur les algorithmes de recherche de règles figurent également dans ce chapitre dévolu aux méthodes non-supervisées.

---

[1] Les techniques d'analyse factorielle comprennent dans la littérature statistique française toutes les techniques de représentation utilisant des "axes principaux": analyse en composantes principales, des correspondances simples et multiples, analyse factorielle dite classique ou des psychologues — alors que l'expression correspondante en anglais (*factor analysis*) ne désigne de façon assez stricte que cette dernière technique : analyse en facteurs communs et spécifiques de Spearman, Thurstone, utilisée principalement par les psychologues et les psychométriciens.

Méthodes de discrimination ou de classification supervisée

Les *méthodes de classification supervisées* ou encore *méthodes de discrimination* (septième chapitre) sont des méthodes de *classement* ou d'affectation d'individus dans des classes préexistantes. Un large éventail de techniques (analyse linéaire discriminante, analyses régularisées, régression logistique, segmentation) recouvre une part très importante des applications réelles et potentielles de la statistique.

Données structurées

Lorsqu'on a peu d'information *a priori* sur les données (on parlera alors de données non structurées ou amorphes) l'application des techniques exploratoires multidimensionnelles est gratifiante. Mais il est plus difficile d'utiliser ce que l'on sait pour essayer d'en savoir plus. Et si l'information *a priori* sur les données est considérable, d'autres techniques faisant appel à des modèles qui utilisent effectivement cette information sont alors compétitives. Les méthodes d'analyse de tableaux ayant une structure *a priori* constituent le complément naturel ou le prolongement des analyses exploratoires. Les méthodes les plus utiles sont présentées au huitième et dernier chapitre. Elles tentent d'intégrer en leur sein même une éventuelle information externe : les analyses partielles ou conditionnelles permettent de prendre en compte l'effet de certaines variables ; les analyses locales et les analyses de contiguïté mettent à profit des structures de graphes sur les observations (contenant comme cas particulier les partitions et les séries chronologiques); enfin les analyses de tableaux multiples étudient le cas de tableaux comportant plusieurs groupes de variables (tableaux à trois dimensions, séries de tableaux, etc.).

La plupart des jeux de données utilisés dans ce livre et le logiciel académique (DTM), permettant les calculs et les tracés des exemples correspondants aux *méthodes non supervisées*, peuvent être librement téléchargés à partir de la rubrique *logiciel* du site (http://www.lebart.org). Citons également, parmi les logiciels de traitement, les logiciels libres TANAGRA (http://eric.univ-lyon2.fr/~ricco/tanagra/), et la plateforme polyvalente « R » (http://www.r-project.org/). Mentionnons enfin, parmi les logiciels du commerce les plus utilisés, pour tout ou partie des traitements proposés dans les chapitres qui suivent, le logiciel SPAD (distribué par la société *Test and Go*), le plus proche, dans sa conception, de la méthodologie présentée dans cet ouvrage, et les logiciels SAS ® et SPSS ®.

Chapitre 1

Analyses en axes principaux : principes de base

Les méthodes d'analyses en axes principaux (appelées méthodes factorielles dans la littérature francophone) se proposent de fournir des représentations synthétiques de vastes ensembles de valeurs numériques, en général sous forme de visualisations graphiques. Pour cela, elles s'efforcent de réduire les dimensions du tableau de données en représentant les associations entre individus et entre variables dans des espaces de faibles dimensions.

Les techniques les plus utilisées dérivent des deux méthodes fondamentales que sont l'analyse en composantes principales et l'analyse des correspondances, qui, elles-mêmes, reposent sur des notions et un formalisme communs qui font l'objet de ce premier chapitre.

On introduit tout d'abord la notion même de tableau de données et les concepts et notations de base qui en découlent (Section 1.1 : *Le tableau de données*). Quelle que soit la constitution du tableau de données, toutes les techniques d'analyses en axes principaux ont un noyau commun d'algèbre linéaire que nous désignons sous le nom d'*Analyse générale* (section 1.2).

Cette même *analyse générale*, qui est une présentation géométrique d'un théorème connu sous le nom de « décomposition aux valeurs singulières » (*Singular value decomposition*) peut faire intervenir une métrique plus générale, ou accueillir des éléments supplémentaires (section 1.3 : *Diversification de l'analyse générale*).

Certaines procédures de validation de base qui seront déclinées dans les chapitres suivants sont présentées en section 1.4 (*Méthodes de validation empiriques, calculs de sensibilité et de stabilité*), enfin en section 1.5, une annexe (*Annexe technique du chapitre 1*) explicite certaines démonstrations qui auraient alourdi l'exposé.

1.1 Le tableau de données

Le tableau de données dispose la masse d'information sous forme rectangulaire. Pour fixer les idées, les lignes *(i=1,...,n)* peuvent représenter les *n individus* ou *observations*, appelés plus généralement *unités statistiques* ; les colonnes *(j=1,...,p)* sont alors les *p variables*, qui peuvent être des *mesures* (numériques) ou des *attributs* ou *caractères* observés sur les individus (cas de variables nominales)[1].

1.1.1 Représentation géométrique de base

Afin de comprendre le principe des méthodes de statistique exploratoire multidimensionnelle, il est utile de représenter géométriquement les *n* lignes et les *p* colonnes du tableau de données par des points dont les coordonnées sont précisément les éléments de ce tableau (figure 1.1-1).

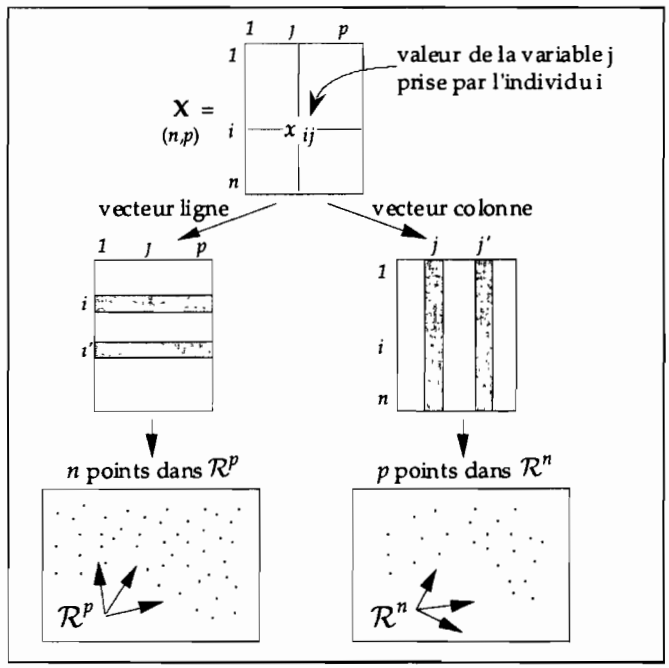


Figure 1.1-1. Principe de représentation géométrique

[1] Cette distinction entre variables et individus est commode parce qu'elle se réfère à une situation classique en statistique. Elle correspond au contexte de l'analyse en composantes principales (chapitre 2) qui précède historiquement l'analyse des correspondances et ses variantes. Cette distinction n'a pas de sens dans le cas de tables de contingence pour lesquelles lignes et colonnes jouent des rôles symétriques.

Deux nuages de points sont alors construits :

- le nuage des n individus (le nuage des points-lignes) situé dans l'espace à p dimensions $\mathcal{R}^p$ des variables (des colonnes); chacune des n lignes est représentée par un point à p coordonnées.
- le nuage des p variables (le nuage des points-colonnes) situé dans l'espace à n dimensions $\mathcal{R}^n$ des individus (des lignes); chacune des p colonnes est représentée par un point à n coordonnées.

Le tableau de données noté **X** est donc une matrice dans laquelle chaque vecteur, ligne ou colonne, représente un point soit dans $\mathcal{R}^p$ soit $\mathcal{R}^n$.

Chacune des deux dimensions du tableau de données permet de définir des distances (ou des proximités) entre les éléments définissant l'autre dimension. L'ensemble des colonnes permet de définir, à l'aide de formules appropriées, des distances entre lignes. De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes.

Les *proximités géométriques* usuelles entre points-lignes et entre points-colonnes traduisent en fait des *associations statistiques* soit entre les individus, soit entre les variables. Les tableaux de distances associés à ces représentations géométriques (simples dans leur principe, mais complexes en raison du grand nombre de dimensions des espaces concernés) pourront alors être décrits par les deux grandes familles de méthodes que sont les méthodes en axes principaux ou méthodes factorielles et la classification (figure 1.1-2).

Ces représentations géométriques du tableau de données nous conduisent naturellement à utiliser les notions d'espaces vectoriels, de nuages de points, de métriques (permettant de calculer des distances entre points-lignes ou entre points-colonnes) mais aussi de masses affectées aux points si l'on ne leur accorde pas la même importance dans le nuage[1].

Les développements théoriques des méthodes de statistique exploratoire multidimensionnelle vont reposer sur ces notions.

Ces méthodes impliquent souvent de la même manière les individus (lignes) et les variables (colonnes). Les individus ne sont plus de simples intermédiaires utilisés pour calculer des moyennes ou des corrélations sur les variables, suivant le schéma de la statistique traditionnelle où ils sont modélisés comme des réalisations d'épreuves indépendantes.

La confrontation des espaces d'individus et de variables enrichira les interprétations.

---

[1] C'est le cas courant des fichiers d'enquêtes dans lequel les individus sont affectés de *poids de redressement* ou de *poids d'extrapolation*.

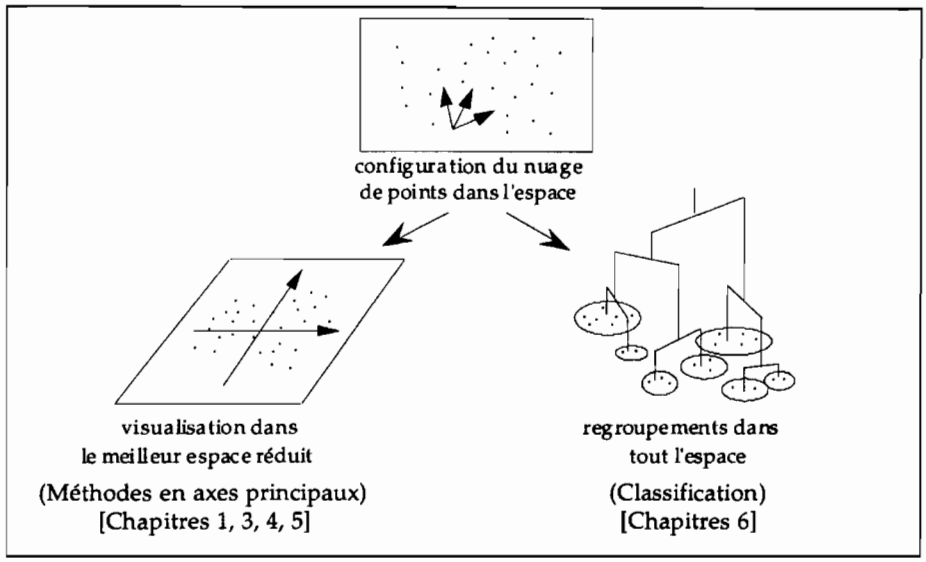


Figure 1.1-2. Les deux grandes familles de méthodes

1.1.2 Principaux types de tableaux et de méthodes

La géométrie des nuages de points et les calculs de proximités ou de distances qui en découlent diffèrent selon la nature des lignes et des colonnes du tableau analysé. Les colonnes peuvent être des variables continues ou des variables nominales ou des catégories dans le cas des tables de contingences. Les lignes peuvent être des individus ou des catégories.

La nature des informations, leur codage, les spécificités du domaine d'application vont introduire des variantes au sein des méthodes factorielles.

• *Les tableaux de type "variables-individus"*

Leurs colonnes sont des variables à valeurs numériques continues et leurs lignes sont des individus, des observations, des objets, etc. Ils constituent le domaine d'application de l'Analyse en Composantes Principales (ACP) (chapitre 3).

Les proximités entre variables s'interprètent alors en termes de corrélation ; les proximités entre individus s'interprètent en termes de similitudes globales des valeurs observées.

L'ACP peut donner lieu à de nombreuses variantes en s'appliquant par exemple à un tableau de rangs (diagonalisation de la matrice de corrélation des rangs de Spearman), ou encore en éliminant l'effet de certaines variables (analyses locales ou partielles).

• *Les tableaux de contingences, les tableaux binaires*

Ce sont les tableaux de comptages obtenus par le croisement de deux variables nominales ou les tableaux de présence-absence, domaine privilégié de l'Analyse des Correspondances (chapitre 4). Ces tableaux ont la particularité de faire jouer un rôle identique aux lignes et aux colonnes. L'analyse fournit des représentations des associations entre lignes et colonnes de ces tableaux, fondées sur une distance entre profils (qui sont des vecteurs de fréquences conditionnelles) désignée sous le nom de distance du Chi 2 .

• *Les tableaux disjonctifs complets*

Ce sont de grands tableaux de variables nominales dont les fichiers d'enquêtes socio-économiques ou médicales constituent des exemples privilégiés. Les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers); les colonnes sont des modalités de variables. L'Analyse des Correspondances Multiples (chapitre 5), adaptée à l'analyse de ces tableaux, est une extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

Notations de base

Malgré leur partielle inadaptation aux éléments mathématiques que l'on va traiter, les notations matricielles seront souvent utilisées par souci de cohérence et volonté de communication avec l'essentiel de la littérature statistique disponible.

Le tableau des données soumis à l'analyse est désigné par la lettre majuscule grasse **X**. La matrice **X** est d'ordre (n,p), autrement dit, elle a n lignes et p colonnes. Son terme générique est x_{ij} ($i^{ème}$ observation de la $j^{ème}$ variable). Une colonne de **X** sera désignée par la lettre minuscule grasse $\mathbf{x}_j$.

La transposée de **X** est notée **X'**; cette matrice a donc p lignes et n colonnes.

Pour les notations utilisant des caractères latins, les matrices sont représentées par des lettres majuscules grasses; les vecteurs par des lettres minuscules grasses, sauf dans certains contextes géométriques où ils pourront être désignés par leurs origines et leurs extrémités (exemple : AB, dont la longueur est alors notée : AB) ; les scalaires sont désignés par des lettres minuscules en italique.

La convention retenue dans cet ouvrage est de noter en italique les constantes numériques exprimées en chiffres arabes (à l'exception des dates, des numéros de section ou de chapitre, des tableaux issus des exemples).

1.2 Analyse générale, décomposition aux valeurs singulières

Considérons un tableau de valeurs numériques **X** ayant n lignes et p colonnes. Pour prendre un exemple, le tableau **X** a $n = 1000$ lignes et $p = 100$ colonnes. Il représente les p variables observées sur n individus constituant un échantillon statistique.

Le tableau **X** possède donc $np = 100\ 000$ éléments. Pour des raisons diverses, il peut exister des liaisons fonctionnelles ou stochastiques entre certaines variables. Peut-on résumer ces np données par un nombre inférieur de valeurs sans perte notable d'information compte tenu des liaisons et interrelations entre les valeurs ?

Nous recherchons en fait une technique de réduction s'appliquant de façon systématique à divers types de tableaux et conduisant à une reconstitution rapide mais approximative du tableau de départ.

1.2.1 Notions élémentaires et principe d'ajustement

On a vu précédemment comment les lignes et les colonnes d'un tableau rectangulaire permettaient de définir des nuages de points. La position des points dans le nuage est donnée par l'ensemble des distances entre tous les points et détermine la *forme du nuage*. C'est elle qui caractérise la nature et l'intensité des relations entre les individus (lignes) et entre les variables (colonnes) et révèle les structures de l'information contenues dans les données.

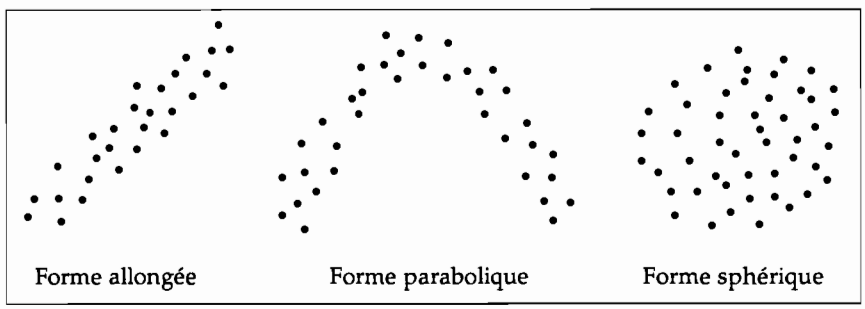


Figure 1.2 – 1. Différentes formes de nuages

Par exemple, si le nuage de points est uniformément allongé le long d'une droite, il existe un support linéaire dominant pour les points. Une forme

parabolique traduira une relation non linéaire tandis qu'un nuage de forme sphérique marquera plutôt une absence de relation (cf. figure 1.2 - 1).

Une façon simple de rendre compte visuellement de la forme d'un nuage est de le projeter sur des droites, ou mieux sur des plans, en minimisant les déformations que la projection implique. Pour cela, on peut chercher le sous-espace à une dimension H qui maximise la somme des carrés des distances entre les projections sur H de tous les couples de points *(k,k')* :

$$\underset{(H)}{Max}\left\{\sum_{k}\sum_{k'} d^2(k,k')\right\}$$

Si chaque point est muni d'une masse, c'est la somme pondérée que l'on pourra chercher à maximiser :

$$\underset{(H)}{Max}\left\{\sum_{k}\sum_{k'} p_k\, p_{k'}\, d^2(k,k')\right\}$$

On calcule ainsi le sous-espace vectoriel qui ajuste au mieux le nuage de points.

Nous verrons plus loin, à propos de l'analyse en composantes principales, que ce dernier critère est équivalent au critère ci-dessous (où G désigne le point moyen ou centre de gravité des projections) :

$$\underset{(H)}{Max}\left\{\sum_{k} p_k\, d^2(k,G)\right\}$$

Toutefois, on ne s'intéresse pas toujours à la forme d'un nuage, mais quelques fois à sa position par rapport à l'origine. Ainsi, en analyse en composantes principales, on s'intéresse bien à la forme du nuage des points-observations dans un espace, mais c'est la position par rapport à l'origine des points-variables qui aura du sens dans l'autre espace.

Le modèle d'analyse par rapport à l'origine désigné ici sous le nom d'*analyse générale* permet de rendre compte de ces diverses situations. Il n'est qu'une présentation sous forme géométrique de la *décomposition aux valeurs singulières* présentée pour la première fois par Eckart et Young (1936, 1939) pour les tableaux rectangulaires, généralisant les travaux de Sylvester (1889) relatifs aux matrices carrées[1]. Le problème que l'on se propose de résoudre est alors un problème de réduction purement numérique, autrement dit, un problème de compression de données.

Pour exposer cette technique de réduction factorielle, nous nous plaçons successivement dans les espaces vectoriels $\mathcal{R}^p$ et $\mathcal{R}^n$, avec pour notre exemple : $p = 100$, $n = 1000$.

---

[1] . Gifi (1990) mentionne également les travaux antérieurs et indépendants de Beltrami (1873) et Jordan (1874). Cf. également Gower (1966), Gabriel (1971).

1.2.2 Ajustement du nuage des individus dans l'espace des variables

On envisage ici le nuage de n points-individus définis dans l'espace des variables $\mathcal{R}^p$ et qui sont non pondérés (pour simplifier la formulation). Chacune des n lignes du tableau **X** est considérée comme un vecteur ou encore un point de $\mathcal{R}^p$. Si ce nuage est contenu dans un sous-espace vectoriel à q dimensions de $\mathcal{R}^p$ et si q est notablement inférieur à p, autrement dit, si le tableau **X** est de rang q, le problème d'approximation est pratiquement résolu[1].

a – Droites d'ajustement

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

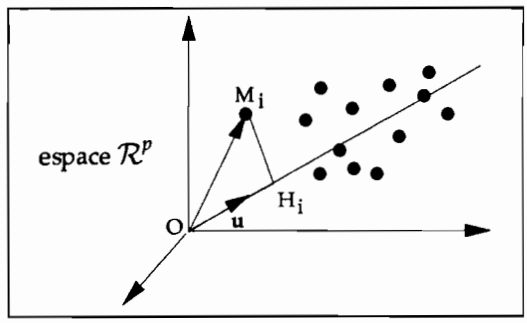


Figure 1.2 – 2. Meilleur ajustement du nuage de points

Il faut pour cela définir le vecteur directeur unitaire de cette droite. Soit **u** ce vecteur. On désignera également par **u** la matrice colonne associée, et par **u'** sa transposée. On exprime que **u** est unitaire par la relation **u'u** = *1*. La longueur de la projection OH_i d'un vecteur OM_i sur le sous-espace à une dimension porté par **u** (figure 1.2 - 2) n'est autre que le produit scalaire de OM_i par **u**, somme des produits terme à terme[2] des composantes de OM_i et de **u** :

$$OH_i = \mathbf{x}_i'\mathbf{u} = \sum_j^p x_{ij}u_j$$

---

[1] Par exemple, si les n = *1000* points-individus se trouvent dans un sous-espace à q = *10* dimensions, il suffit, pour retrouver les positions relatives de ces points dans $\mathcal{R}^p$, de connaître la nouvelle base (soit q = *10* vecteurs à p = *100* dimensions) et les nouvelles coordonnées des points dans cette base (soit n = *1000* vecteurs à q = *10* dimensions). On pourrait dans ce cas reconstituer les np = *100 000* nombres à partir de *11 000* nombres ($qp + nq$ = *11 000*).

[2] On suppose implicitement (et provisoirement) que la métrique dont est muni cet espace est la métrique euclidienne usuelle.

Chacune des n lignes du tableau **X** est un vecteur-individu $\mathbf{x_i}$ dans $\mathcal{R}^p$. Or le produit matriciel **Xu** est la matrice-colonne à n éléments, dont chaque terme est le produit scalaire d'une ligne de **X** par **u** :

$$\mathbf{Xu} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \cdots & \\ \cdots & x_{ij} & \cdots \\ & \cdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} u_1 \\ \cdots \\ \\ \cdots \\ u_p \end{bmatrix} = \begin{bmatrix} \sum x_{1j}u_j \\ \cdots \\ \sum x_{ij}u_j \\ \cdots \\ \sum x_{nj}u_j \end{bmatrix}$$

Ce sont les n composantes de la matrice colonne **Xu** qui repèrent sur **u** les n projections OH_i des points du nuage.

Parmi les critères d'ajustement d'un sous-espace à un nuage de n points, celui que l'on retient et qui conduit aux calculs analytiques sans doute les plus simples, est le critère classique *des moindres carrés*. Il consiste à rechercher la *droite d'allongement maximum* du nuage de points et donc à rendre minimale la somme des carrés des écarts :

$$\sum_{i=1}^{n} M_i H_i^2$$

Le théorème de Pythagore appliqué à chacun des n triangles rectangles du type H_iOM_i conduit à la relation :

$$\sum_i M_i H_i^2 = \sum_i OM_i^2 - \sum_i OH_i^2$$

Comme $\sum_i OM_i^2$ est une quantité fixe, indépendante du vecteur **u** cherché, il est équivalent de rendre maximale la quantité :

$$\sum_i OH_i^2$$

qui s'exprime en fonction de **X** et **u** par :

$$\sum_i OH_i^2 = \mathbf{(Xu)'Xu = u'X'Xu}$$

Pour trouver **u**, on est donc conduit à chercher le maximum de la forme quadratique **u'X'Xu** :

$$\begin{cases} Max_{(\mathbf{u})}\{\mathbf{u'X'Xu}\} \\ \text{sous la contrainte} \quad : \mathbf{u'u} = 1 \end{cases}$$

Soit $\mathbf{u_1}$ le vecteur qui réalise ce maximum. Le sous-espace à deux dimensions s'ajustant au mieux au nuage contient nécessairement le sous-espace engendré par $\mathbf{u_1}$ [1]. On cherche ensuite $\mathbf{u_2}$, le second vecteur de base de ce sous-espace,

---

[1] Le raisonnement par l'absurde prouve que s'il ne contenait pas $\mathbf{u_1}$, il en existerait un meilleur contenant $\mathbf{u_1}$.

orthogonal à $\mathbf{u}_1$ et rendant maximal $\mathbf{u}_2'\mathbf{X}'\mathbf{X}\mathbf{u}_2$. On recherche de façon analogue le meilleur sous-espace au sens des moindres carrés à q dimensions ($q \leq p$).

b – Caractéristiques du sous-espace d'ajustement

Les démonstrations qui figurent en annexe (§ 1.5.1 ci-après) conduisent à l'énoncé suivant :

> le vecteur unitaire $\mathbf{u}_1$ qui caractérise le sous-espace à une dimension ajustant au mieux le nuage des n points individus dans $\mathcal{R}^p$, est le *vecteur propre* de la matrice **X'X** correspondant à la plus grande *valeur propre* λ_1.

L'axe porté par le vecteur $\mathbf{u}_1$ est le *premier axe principal* ou encore le *premier axe factoriel.*

Plus généralement, le sous-espace à q dimensions qui ajuste au mieux (au sens des moindres carrés) le nuage dans $\mathcal{R}^p$ est engendré par les q premiers vecteurs propres de la matrice symétrique **X'X** correspondant aux q plus grandes valeurs propres. On diagonalisera, par conséquent, la matrice **X'X** d'ordre (p,p).

L'analyse générale effectue donc une rotation du repère autour de l'origine **O** et fournit un système de vecteurs orthonormés dont $\mathbf{u}_1$ puis $(\mathbf{u}_1, \mathbf{u}_2)$, ..., $(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_\alpha, ..., \mathbf{u}_p)$ passent "au plus près" du nuage.

On notera $\mathbf{u}_\alpha$ le vecteur propre de la matrice **X'X** correspondant à la valeur propre λ_α.

1.2.3 Ajustement du nuage des variables dans l'espace des individus

Plaçons-nous maintenant dans l'espace des individus $\mathcal{R}^n$, où le tableau **X** peut être représenté par un nuage de p points-variables dont les n coordonnées représentent les colonnes de **X**. La démarche pour ajuster le nuage des p points-variables dans cet espace est exactement la même que pour le nuage des points-individus et consiste à rechercher le vecteur unitaire **v**, puis le sous-espace à q dimensions dans $\mathcal{R}^n$ qui ajuste au mieux le nuage de points.

Cela conduit à rendre maximale la somme des carrés des p projections sur **v**, qui sont les p composantes du vecteur **X'v**. On maximise la quantité :

$$(\mathbf{X'v})'\mathbf{X'v} = \mathbf{v'XX'v} \quad \text{avec la contrainte} \quad \mathbf{v'v} = 1$$

Comme précédemment, nous sommes amenés à retenir les q vecteurs propres de **XX'** correspondant aux q plus grandes valeurs propres. La matrice à diagonaliser sera cette fois la matrice **XX'** d'ordre (n,n). On notera $\mathbf{v}_\alpha$ le vecteur propre normé de **XX'** correspondant à la valeur propre μ_α.

1.2.4 Relation entre les ajustements dans les deux espaces

Recherchons les relations dites de transition entre les deux espaces.

Dans $\mathcal{R}^p$, nous avons :

$$\mathbf{X}'\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \qquad [1.2 - 1]$$

et dans $\mathcal{R}^n$:

$$\mathbf{X}\mathbf{X}'\mathbf{v}_\alpha = \mu_\alpha \mathbf{v}_\alpha \qquad [1.2 - 2]$$

En prémultipliant les deux membres de [1.2 - 1] par $\mathbf{X}$, on obtient :

$$(\mathbf{X}\mathbf{X}')\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha(\mathbf{X}\mathbf{u}_\alpha)$$

Cette relation montre qu'à tout vecteur propre normé $\mathbf{u}_\alpha$ de $\mathbf{X}'\mathbf{X}$ relatif à une valeur propre λ_α non nulle, correspond un vecteur propre $\mathbf{X}\mathbf{u}_\alpha$ de $\mathbf{X}\mathbf{X}'$, relatif à la même valeur propre λ_α. Comme on a appelé μ_1 la plus grande valeur propre de $\mathbf{X}\mathbf{X}'$, on a nécessairement $\lambda_1 \leq \mu_1$.

En prémultipliant les deux membres de [1.2 - 2] (pour $\alpha = 1$) par $\mathbf{X}'$, on voit de même $\mathbf{X}'\mathbf{v}_1$ est vecteur propre de $\mathbf{X}'\mathbf{X}$ relativement à la valeur propre μ_1, d'où la relation $\mu_1 \leq \lambda_1$, ce qui prouve finalement que $\lambda_1 = \mu_1$.

On verrait de la même façon que toutes les valeurs propres non nulles des deux matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}\mathbf{X}'$ sont égales[1] (avec le même ordre de multiplicité éventuellement) :

$$\lambda_\alpha = \mu_\alpha$$

Remarquons que le vecteur $\mathbf{X}\mathbf{u}_\alpha$ a pour norme λ_α (on a $\mathbf{u}'_\alpha\mathbf{X}'\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha$) et donc le vecteur $\mathbf{v}_\alpha$ unitaire correspondant à la même valeur propre λ_α est facilement calculable en fonction de $\mathbf{u}_\alpha$. On obtient ainsi, pour $\lambda_\alpha > 0$, les *formules de transition* entre les deux espaces, $\mathcal{R}^p$ et $\mathcal{R}^n$:

$$\begin{cases} \mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\mathbf{u}_\alpha & [1.2-3] \\ \mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}'\mathbf{v}_\alpha & [1.2-4] \end{cases}$$

Dans $\mathcal{R}^p$, $\mathbf{u}_\alpha$ est le $\alpha^{\text{ième}}$ *axe factoriel* et l'on calcule le vecteur ψ_α des coordonnées sur cet axe par :

$$\psi_\alpha = \mathbf{X}\mathbf{u}_\alpha$$

---

[1] Il est donc inutile de refaire les calculs de diagonalisation sur $\mathbf{X}\mathbf{X}'$, puisqu'une simple transformation linéaire, associée à la matrice $\mathbf{X}$ de départ, nous permet d'obtenir les directions propres $\mathbf{X}\mathbf{u}_\alpha$ cherchées dans $\mathcal{R}^n$. Il suffit de diagonaliser la matrice $\mathbf{X}'\mathbf{X}$ (p,p) ou $\mathbf{X}\mathbf{X}'$ (n,n) ayant la plus petite dimension.

De même dans $\mathcal{R}^n$, $\mathbf{v}_\alpha$ est le $\alpha^{\text{ième}}$ axe factoriel et l'on construit φ_α le vecteur des coordonnées sur cet axe :

$$\varphi_\alpha = \mathbf{X}'\mathbf{v}_\alpha$$

Compte tenu de [1.2 - 3] et [1.2 - 4], les facteurs peuvent se calculer par :

$$\psi_\alpha = \mathbf{v}_\alpha\sqrt{\lambda_\alpha} \qquad \text{et} \qquad \varphi_\alpha = \mathbf{u}_\alpha\sqrt{\lambda_\alpha}$$

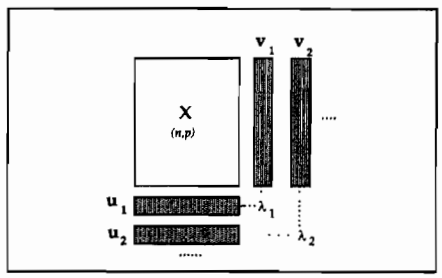


Figure 1.2 – 3. Relations de transitions

Sur le sous-espace de $\mathcal{R}^p$ engendré par $\mathbf{u}_\alpha$, les coordonnées des points du nuage des individus sont les composantes de $\mathbf{X}\mathbf{u}_\alpha$. Ce sont aussi les composantes de $\mathbf{v}_\alpha\sqrt{\lambda_\alpha}$. Les coordonnées des points sur un axe factoriel dans $\mathcal{R}^p$ sont donc proportionnelles aux composantes de l'axe factoriel dans $\mathcal{R}^n$ correspondant à la même valeur propre. De même, les coordonnées des points du nuage des variables sur $\mathbf{v}_\alpha$ sont proportionnelles aux composantes de l'axe factoriel dans $\mathcal{R}^p$.

Remarques

1) L'orientation des axes est arbitraire. En effet, les vecteurs propres sont définis au signe près. La figure 1.2 - 4, concernant trois points, montre que toutes les images, obtenues suivant des orientations différentes des facteurs, respectent la forme du nuage c'est-à-dire les distances entre les points.

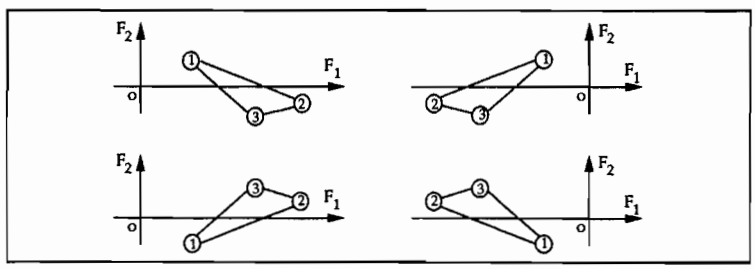


Figure 1.2 – 4. Orientation arbitraire des axes

2) Les vecteurs de coordonnées dans $\mathcal{R}^p$ et $\mathcal{R}^n$ ont pour norme :

$$\psi'_\alpha\psi_\alpha = \sum_i^n \psi_{\alpha i}^2 = \lambda_\alpha \qquad \text{et} \qquad \varphi'_\alpha\varphi_\alpha = \sum_j^p \varphi_{\alpha j}^2 = \lambda_\alpha$$

1.2.5 Reconstitution des données de départ

Nous désignons toujours par $\mathbf{u}_\alpha$ le $\alpha^{\text{ième}}$ vecteur propre de norme *1* de la matrice **X'X**, correspondant à la valeur propre λ_α; $\mathbf{v}_\alpha$ le $\alpha^{\text{ième}}$ vecteur propre de norme *1* de **XX'**. Nous avons :

$$\psi_\alpha = \mathbf{X}\mathbf{u}_\alpha = \mathbf{v}_\alpha\sqrt{\lambda_\alpha}$$

a – Reconstitution exacte

Postmultiplions les deux membres de cette relation par $\mathbf{u}'_\alpha$ et sommons sur l'ensemble des axes[1] :

$$\mathbf{X}\left\{\sum_{\alpha=1}^{p}\mathbf{u}_\alpha\mathbf{u}'_\alpha\right\} = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\mathbf{v}_\alpha\mathbf{u}'_\alpha$$

Désignons par **U** la matrice d'ordre (p,p) ayant en colonnes les vecteurs propres $\mathbf{u}_\alpha$ de **X'X**. Ces vecteurs étant orthogonaux et de norme *1*, on a :

$$\mathbf{UU'} = \mathbf{I}_p \quad \text{et donc} \quad \mathbf{U'U} = \mathbf{I}_p$$

où $\mathbf{I}_p$ est la matrice unité d'ordre p. Or : $\sum_{\alpha=1}^{p}\mathbf{u}_\alpha\mathbf{u}'_\alpha = \mathbf{UU'}$

Les valeurs propres λ_α étant toujours rangées par ordre décroissant, la formule précédente devient :

$$\mathbf{X} = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\mathbf{v}_\alpha\mathbf{u}'_\alpha \qquad [1.2 - 5]$$

et apparaît comme une formule de reconstitution du tableau **X**, à partir des λ_α et des vecteurs $\mathbf{u}_\alpha$ et $\mathbf{v}_\alpha$ associés (figure 1.2 - 5).

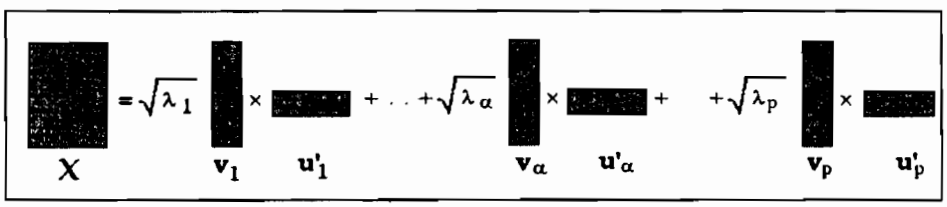


Figure 1.2 – 5. Reconstitution du tableau de données, ou décomposition aux valeurs singulières

Remarque

Les méthodes d'analyse factorielle reposent toutes sur une propriété mathématique des tableaux rectangulaires : la décomposition aux valeurs singulières. Cela signifie principalement que toute matrice peut être écrite de façon unique comme une

[1] Certains d'entre eux peuvent correspondre à une valeur propre nulle; ils sont alors choisis de façon à compléter la base orthonormée formée par les axes précédents.

"somme optimale" de matrices de rang *1* (produits d'une matrice ligne par une matrice colonne). Ce qui signifie que la première matrice de rang *1* constitue la meilleure approximation de rang *1* de la matrice initiale (au sens des moindres carrés), que la somme des deux premières constituent la meilleure approximation de rang 2, etc..

b – Reconstitution approchée

Si les *p-q* plus petites valeurs propres sont jugées très faibles ou "négligeables", on peut limiter la sommation aux q premiers termes :

$$\mathbf{X} \approx \mathbf{X}^* = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\, \mathbf{v}_\alpha \mathbf{u}'_\alpha \qquad [1.2 - 6]$$

Si q est notablement inférieur à p, on apprécie le gain réalisé en comparant les deux membres de cette relation : le vecteur $\sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha$ a n composantes et le vecteur $\mathbf{u}_\alpha$ a p composantes.

Les np termes de $\mathbf{X}$ sont donc approchés par des termes construits à partir des $q(n+p)$ valeurs contenues dans le membre de droite.

c – Qualité numérique de l'approximation

La qualité de la reconstitution peut être évaluée par la quantité :

$$\tau_q = \frac{\sum_i \sum_j x_{ij}^{*2}}{\sum_i \sum_j x_{ij}^{2}}$$

On a encore (*tr* désignant l'opérateur trace) :

$$\tau_q = \frac{tr\, \mathbf{X}^{*\prime}\mathbf{X}^*}{tr\, \mathbf{X}'\mathbf{X}}$$

Remplaçant $\mathbf{X}$ et $\mathbf{X}^*$ par leurs valeurs tirées de [1.2 - 5] et [1.2 - 6], on obtient immédiatement :

$$\tau_q = \frac{\sum_{\alpha \le q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}$$

Le coefficient τ_q, inférieur ou égal à *1*, sera appelé *taux d'inertie* ou encore *pourcentage de variance* relatif aux q premiers facteurs. Son interprétation comme mesure de la qualité numérique de la reconstitution est assez claire, mais nous verrons plus loin que le problème de sa signification statistique est délicat et peut donner lieu à des interprétations erronées.

1.3 Diversification de l'analyse générale

La métrique (c'est-à-dire la formule de distance) et le critère d'ajustement (c'est-à-dire la pondération des points) varient suivant le problème et donc suivant la nature des variables.

1.3.1 _ Analyse générale avec des métriques et des critères quelconques

Jusqu'à présent, nous avons considéré les espaces munis de la métrique euclidienne usuelle dont la forme quadratique est associée à la matrice **I** (matrice identité) et nous avons supposé que tous les points du nuage avaient la même importance.

Cependant il arrive que l'on ait à travailler avec une métrique plus générale et avec des individus dont les masses sont différentes (pondérations calculées après un redressement d'échantillon, regroupements divers d'individus, etc.). Ces masses vont intervenir dans les calculs de moyennes et lors de l'ajustement des sous-espaces.

Généralisons le principe d'analyse factorielle présenté ci-dessus à des métriques et des critères quelconques.

Plaçons-nous dans l'espace $\mathcal{R}^p$ et considérons le nuage de n points-lignes pesants.

Soit **X** la matrice d'ordre (n,p) représentant le tableau de données, **M** la matrice symétrique définie positive d'ordre (p,p) définissant la métrique dans $\mathcal{R}^p$, et **N** la matrice diagonale d'ordre (n,n) dont les éléments diagonaux sont les masses m_i des n points.

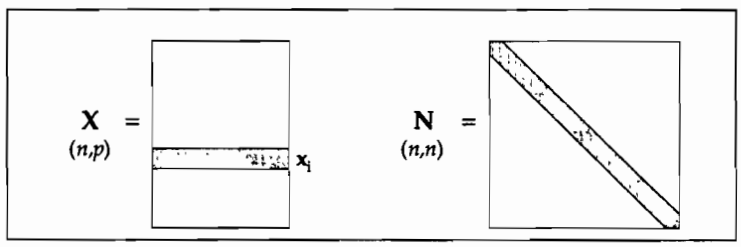


Figure 1.3 - 1 . X, tableau de coordonnées et N, matrice diagonale des masses

Un vecteur unitaire **u** de $\mathcal{R}^p$ vérifie maintenant la relation de normalisation **u'Mu** = *1*.

La coordonnée de la projection H_i du point *i* sur l'axe **u** vaut :

$$OH_i = \mathbf{x}_i'\mathbf{Mu}$$

et l'ensemble **F** des coordonnées des projections sur l'axe **u** des *n* points-lignes s'exprime par :

$$\mathbf{F = XMu}$$

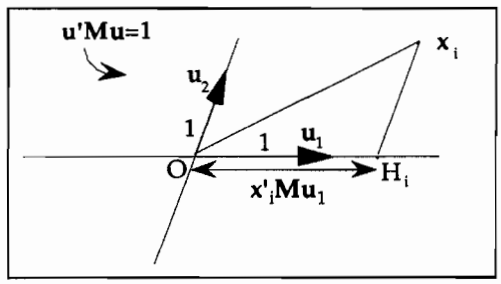


Figure 1.3 – 2. Métrique M dans $\mathcal{R}^p$

Compte tenu du critère d'ajustement, on veut trouver le vecteur **u** qui rende maximale la somme pondérée des carrés des projections :

$$Max_{(\mathbf{u})}\{\sum_i m_i OH_i^2\} = Max_{(\mathbf{u})}\{\mathbf{u'MX'NXMu}\}$$

sous la contrainte : $\mathbf{u'Mu} = 1$

Les calculs de l'annexe de cette section montrent que **u** est le vecteur propre de la matrice **A=X'NXM** correspondant à la plus grande valeur propre λ.

L'équation de l'axe factoriel **u** dans $\mathcal{R}^p$ s'écrit :

$$\mathbf{X'NXMu} = \lambda\mathbf{u}$$

et les coordonnées factorielles des *n* points sont données par la relation :

$$\psi = \mathbf{XMu}$$

- Relation entre $\mathcal{R}^p$ et $\mathcal{R}^n$

Si les masses et les métriques dans $\mathcal{R}^p$ (**N** et **M**) et dans $\mathcal{R}^n$ (**P**, matrice des masses des *p* points-colonnes et **Q**, métrique dans $\mathcal{R}^n$) n'ont pas de relations privilégiées entre elles, on perd les relations de transition et la formule de reconstitution[1].

---

[1] En analyse en composantes principales normées, on utilise la même métrique dans les deux espaces. En analyse des correspondances, on verra que la matrice des masses dans un espace est liée à la métrique de l'autre espace, ce qui permettra de conserver les relations de transition.

- Axes principaux ou axes d'inertie

La quantité :

$$\mathbf{u'MX'NXMu} = \psi'\mathbf{N}\psi = \sum_i m_i \psi_i^2$$

représente l'inertie du nuage de points pesants le long de l'axe d'allongement maximal, l'axe factoriel **u**. Elle est égale à la valeur propre λ associée au vecteur propre **u**. Les p vecteurs propres définissent donc des axes d'inertie du nuage de points et on les obtient par ordre d'inerties décroissantes. La somme de toutes les valeurs propres donne l'inertie totale du nuage. C'est la trace de la matrice diagonalisée, appelée *matrice d'inertie*, **A** = **X'NXM** :

$$Trace(\mathbf{A}) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

1.3.2 Principe des éléments supplémentaires

L'analyse factorielle permet de trouver des sous-espaces de représentation des proximités entre vecteurs de description d'observations. Elle s'appuie, pour cela, sur des éléments (variables et individus) appelés *éléments actifs*. Mais elle permet aussi de positionner, dans ce sous-espace, des éléments (points-lignes ou points-colonnes du tableau de données) n'ayant pas participé à l'analyse qui sont appelés *éléments supplémentaires* ou *illustratifs*.

Les éléments supplémentaires interviennent *a posteriori* pour caractériser les axes. Leur introduction dans l'analyse factorielle constitue un apport fondamental car elle permettra de conforter et d'enrichir l'interprétation des facteurs. En effet, il est fréquent, dans la pratique, que l'on dispose d'informations complémentaires élargissant le tableau de données.

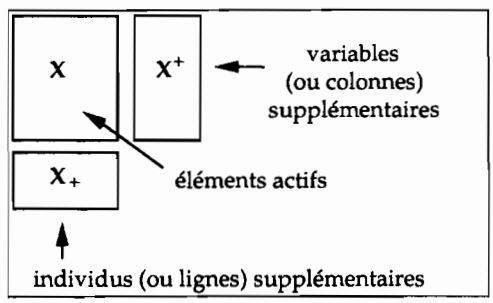


Figure 1.3 – 3. Représentation des éléments supplémentaires

Très souvent dans les applications, ce ne sont pas les individus par eux-mêmes qui sont intéressants mais certaines de leurs caractéristiques connues par ailleurs; on positionnera alors comme "individus" supplémentaires les centres de gravité des individus appartenant à une même catégorie. Ce peut être aussi de nouvelles variables (colonnes supplémentaires); on peut disposer d'un ensemble de variables nominales qu'il est intéressant de faire apparaître dans

l'analyse réalisée sur des variables continues (et réciproquement). Par ailleurs certaines variables observées sur l'échantillon initial peuvent être disponibles alors qu'on les a volontairement écartées de l'analyse pour ne conserver qu'un corpus homogène de caractéristiques.

Les éléments supplémentaires n'interviennent pas dans les calculs d'ajustement et ne participent donc pas à la formation des axes factoriels. On cherche uniquement à les positionner dans le nuage des individus ou dans celui des variables en calculant *a posteriori* leurs coordonnées sur les axes factoriels.

Les coordonnées des nouvelles variables sur l'axe α sont les composantes du vecteur :

$$(\mathbf{X}^{+})'\mathbf{v}_{\alpha}$$

et les coordonnées des nouveaux individus sur l'axe α sont :

$$(\mathbf{X}_{+})\mathbf{u}_{\alpha}$$

Les éléments actifs[1], définis dans un espace et servant à calculer les plans factoriels, doivent former un ensemble homogène en texture (c'est-à-dire doivent être de même nature, continues ou nominales) pour que les distances entre éléments aient un sens. Mais pour interpréter les similitudes entre ces éléments, ils doivent aussi être homogènes en contenu, c'est-à-dire relatifs à un même thème ; on compare les objets selon un certain point de vue et non pas en utilisant sans différenciation tous les attributs connus et souvent disparates. Les variables supplémentaires ne sont pas nécessairement homogènes.

1.3.3 Autres approches

La décomposition aux valeurs singulières est une propriété de tous les tableaux rectangulaires. Elle fait appel à des distances euclidiennes, c'est-à-dire à des formes quadratiques définies positives, et à des ajustements de sous-espaces vectoriels par minimisation d'un critère lié à ces distances. D'autres approches sont possibles, qui modifient le type de distance, ou la nature des sous-espaces, ou les deux. Il faut s'attendre à perdre certaines des propriétés mathématiques simples de l'analyse générale : unicité de la décomposition, symétrie des rôles joués par les lignes et les colonnes, simplicité de la formule de reconstitution, positionnement aisé de variables supplémentaires.

D'autres critères d'ajustements peuvent tout d'abord être utilisés. A la méthode des moindres carrés $min\{\sum e_i^2\}$ (norme dite "L_2"), on peut par exemple

---

[1] Cette dichotomie entre variables actives et variables illustratives est analogue à la distinction établie entre les variables explicatives (exogènes) et les variables à expliquer (endogènes) dans les modèles de régression multiple (cf. chapitre 2). D'un point de vue géométrique, nous verrons que les deux situations sont d'ailleurs très similaires. Notons que les points supplémentaires peuvent être considérés comme des points actifs affectés d'une masse nulle.

substituer celle des moindres valeurs absolues $min\{\sum |e_i|\}$ (norme dite "L_1"). Nous évoquerons à nouveau ces normes à propos de la régression, chapitre 2. Sur les méthodes d'analyse des données utilisant la norme L_1 (dite aussi *city-block distance*) on consultera les contributions et points de vue de Fichet (1987, 1988, ainsi que dans Van Cutsem *et al.*, 1994), Arabie (1991) et le recueil édité par Dodge (1987).

Dans un esprit un peu différent, Meyer (1994) donne un algorithme pour ajuster (au sens des moindres carrés, c'est-à-dire de L_2) une matrice de distances de type L_p à une matrice de dissimilarité donnée. Pour étudier certaines tables de contingence, notamment les tableaux d'échanges, Domenges et Volle (1979) proposent d'utiliser la distance de Hellinger :

$$d^2(x, y) = \sum_i (\sqrt{x_i} - \sqrt{y_i})^2$$ ("analyse factorielle sphérique").

Enfin, sans changer la métrique ni le critère d'ajustement, on peut songer à ajuster d'autres surfaces que des hyperplans. Ainsi, dans le cas de l'analyse en composantes principales normée qui est, dans l'espace $\mathcal{R}^n$, l'analyse générale de points situés sur une sphère (cf. chapitre 3), Falissard (1995) propose d'ajuster non pas un plan, mais une hypersphère.

1.4 Méthodes de validation empiriques : calculs de stabilité et de sensibilité

Cette section est consacrée aux questions de sensibilité et de stabilité des résultats obtenus à l'issue de méthodes dérivant de l'analyse générale. Elle constitue également une introduction aux méthodes pragmatiques de validation qui seront systématiquement utilisées dans les chapitres suivants.

1.4.1 Aspects théoriques

D'un point de vue théorique, Escofier et Le Roux (1972), Escofier (1979) ont traité de la stabilité des facteurs en analyse en axes principaux (analyse en composantes principales et en analyse des correspondances). Ces auteurs étudient les variations maximales des facteurs et des valeurs propres lorsque l'on apporte des modifications bien déterminées aux données : suppression ou ajout d'éléments au tableau de données, influence du regroupement de plusieurs éléments ou de petites modifications des valeurs du tableau, influence du choix de la métrique et de la pondération.

Les sous-espaces correspondant au haut du spectre sont les plus stables vis-à-vis des éventuelles perturbations de la matrice à diagonaliser (cf. Wilkinson,

1965 ; Kato, 1966). De plus cette matrice elle-même (par exemple la matrice des corrélations expérimentales en analyse en composantes principales normée) est moins sensible aux fluctuations d'échantillonnage que les moments d'ordre *1* (moyennes ou pourcentages)[1]. Ces perturbations ne doivent pas affecter l'orientation des axes ni les configurations si on les suppose stables, et la structure mise en évidence sera alors significative.

Des résultats assez forts existent dans le cas de perturbations symétriques de matrices symétriques, comme par exemple le théorème de Wielandt-Hoffman (cf. Wilkinson, *op.cit.*) qui énonce que si **A**, **B**, **C**, sont des matrices symétriques (p,p) ayant respectivement pour valeurs propres classées par ordre décroissant α_i, β_i, γ_i et telles que **C** = **A** + **B** (**B** représente une perturbation additive, **C** est la matrice perturbée), alors :

$$\sum_{i=1}^{p} (\gamma_i - \alpha_i)^2 \leq \sum_{i=1}^{p} \beta_i^2$$

Un autre théorème classique très utilisé dans les travaux précités énonce que, avec les mêmes notations, pour tout i tel que $1 \leq i \leq p$:

$$\alpha_i + \beta_p \leq \gamma_i \leq \alpha_i + \beta_1$$

▸ **Petites variations des valeurs et vecteur propres d'une matrice symétrique A** (voir le détail des calculs en annexe 1.5.2)

Montrons dans ce paragraphe comment des variations de la matrice symétrique à diagonaliser **A**, de terme général a_{ij}, (variations supposées ici infinitésimales) influencent les éléments propres $\mathbf{u}_r$ et λ_r.

Nous partons de la relation $\mathbf{A}\,\mathbf{u}_r = \lambda_r \mathbf{u}_r$. La matrice **A**, et par conséquent les vecteurs propres $\mathbf{u}_r$ et les valeurs propres λ_r sont supposés dépendre continûment d'un paramètre s.

A la suite d'un calcul que nous avons reporté dans l'annexe 1.5.2 de ce chapitre, nous obtenons, pour la variation des valeurs propres :

$$\frac{\partial \lambda_r}{\partial s} = \sum_{i,j} u_{ir} u_{jr} \frac{\partial a_{ij}}{\partial s}$$

Pour les vecteurs propres, le calcul, plus complexe, conduit à :

$$\frac{\partial u_{jr}}{\partial s} = \sum_{t \neq r} u_{jt} \frac{q_{rt}}{\lambda_r - \lambda_t}$$

---

[1] Les travaux de Tanaka (1984) concernent également l'analyse des correspondances (connue également au Japon sous le nom de *méthode de quantification n°3* de Hayashi). Sur l'analyse en composantes principales, on mentionnera les travaux de Krzanowski (1984), Critchley (1985), Benasseni (1986a, 1986b).

Ces formules nous montrent donc, d'une part que la partie principale de la variation des valeurs propres ne dépend pas des variations des vecteurs propres (la variance le long d'un axe dépend plus, par exemple, de l'adjonction ou du retrait d'un élément que de petites variations d'angle de l'axe), d'autre part que les variations des composantes d'un vecteur propre dépendent des écarts entre la valeur propre correspondante et les autres valeurs propres, c'est-à-dire de l'isolement de cette valeur propre, résultat également intuitif et récurrent dans tous les calculs de perturbation[1].

Ainsi, beaucoup des résultats de Escofier et Le Roux (*op. cit.*) se fondent sur un théorème que ces auteurs établissent à partir de résultats de Davis et Kahan (1970), qui s'énonce, avec les mêmes notations que pour le théorème de Wielandt-Hoffman :

Soit deux sous-espaces invariants de **A** et de **C** (**C** = **A** + **B**) correspondant à des valeurs propres de mêmes rangs $s, s+1,..., s+r$. Si θ est le plus grand angle canonique entre ces sous-espaces, on a la majoration :

$$\sin 2\theta \leq \frac{\beta_1 - \beta_n}{\varepsilon}, \quad \text{avec } \theta \leq \frac{\pi}{4} \text{ si } \beta_1 - \beta_n < \varepsilon$$

avec :
$$\varepsilon = inf\{\alpha_{s-1} - \alpha_s, \alpha_{s-r} - \alpha_{s+r+1}\} \text{ si } s \neq 1$$
$$\varepsilon = inf\{\alpha_{r+1} - \alpha_{r+2}\} \text{ si } s = 1$$

Ce sont donc les écarts entre les valeurs propres qui "bordent" le sous-espace qui définissent la stabilité de ce sous-espace. Dans le cas du sous-espace engendré par les premiers facteurs (cas $s = 1$), c'est l'écart entre la dernière valeur propre correspondant à ce sous-espace et la valeur propre immédiatement consécutive qui compte. L'angle entre le sous-espace du tableau initial **A** et le sous-espace homologue du tableau perturbé **C** sera d'autant plus petit que cet écart entre valeurs propres est grand.

1.4.2 Techniques de *bootstrap*

Le présent chapitre 1 dévolu aux principes des analyses en axes principaux met l'accent sur les aspects mathématiques et numériques des réductions et des compressions de données. Bien que les justifications des techniques de bootstrap soient plutôt d'ordre statistique, il a paru utile de présenter ici les principes de cet outil qui, comme l'analyse générale, sera commun à toutes les méthodes d'analyses en axes principaux traitées aux chapitres suivants. On pourrait en fait parler de calcul de stabilité par simulation de perturbations des données. L'idée est simple : compte tenu des erreurs attendues sur les données,

---

[1] Deux valeurs propres voisines correspondent à un nuage de points projetés approximativement circulaire, et donc à une incertitude sur la détermination des axes principaux dans ce plan.

on génère des tableaux perturbés, et l'on vérifie la stabilité des représentations obtenues. Le *bootstrap* met en oeuvre des perturbations particulières d'inspiration statistique, qui seront, pourrait-on dire, les « perturbations standards » utilisées dans cet ouvrage.

Le bootstrap (noté dorénavant en caractères romains) n'est rien d'autre qu'une technique de simulation particulière, fondée sur la distribution empirique de l'échantillon de base. Efron et Tibshirani (1993) réservent le nom de *non-parametric bootstrap* à ce type de simulation, et qualifie de *parametric bootstrap* les simulations qui mettent en jeu une distribution théorique et des paramètres calculés à partir de l'échantillon (simulations classiques)[1].

Le bootstrap est employée pour analyser la variabilité de paramètres statistiques simples en produisant des intervalles de confiance de ces paramètres. Il peut aussi être appliqué à de nombreux problèmes pour lesquels on ne peut pas estimer analytiquement la variabilité d'un paramètre. Ceci est le cas pour les caractéristiques des méthodes multidimensionnelles où les hypothèses de multinormalité sont rarement vérifiées.

Cette technique, introduite par Efron (1979), consiste, pour estimer la confiance que l'on doit accorder à l'estimation θ^* d'un paramètre inconnu θ, à simuler m (m généralement supérieur à *30*) échantillons de même taille n que l'échantillon initial. Ces échantillons, ou réplications, sont obtenus par tirage au hasard avec remise parmi les n individus observés au départ, ceux-ci ayant tous la même probabilité $1/n$ d'être choisis. Certains individus auront de ce fait un poids élevé *(2, 3,...)* alors que d'autres seront absents (poids nul). Chaque réplication k donnera lieu à une estimation $\theta^*(k)$ du paramètre θ. Quelle que soit la complexité du calcul de θ^*, et surtout du calcul de sa distribution statistique, on va pouvoir travailler sur le nouvel échantillon des m réplications de θ^*, et, par exemple calculer la variance empirique des m quantités $\theta^*(k)$. Celle-ci, sous des conditions assez générales, estimera de façon satisfaisante la variance inconnue de l'estimateur.

En anticipant sur les chapitres suivants, prenons l'exemple de l'estimation du coefficient de corrélation r entre deux variables. Le principe consiste à calculer le coefficient de corrélation pour chaque échantillon simulé (pour lequel on effectue un tirage avec remise des *couples* d'observations). On établit alors la distribution des fréquences du coefficient de corrélation (on porte en ordonnée le nombre d'échantillons ayant une même valeur de r, laquelle est représentée en abscisse). Puis on calcule la probabilité pour que le coefficient de corrélation d'un échantillon soit compris dans différentes fourchettes de valeurs définissant ainsi les intervalles de confiance. On obtient une estimation de la précision de la valeur de r obtenue sur l'échantillon de base sans faire l'hypothèse d'une

---

[1] Sur les techniques de simulation et de génération de variables pseudo-aléatoires, cf. par exemple Newman et Odell (1971), Ripley (1983).

distribution normale des données. Les bornes de l'intervalle de confiance peuvent être estimées directement par les quantiles de la distribution simulée.

Pour estimer les valeurs propres, les taux d'inertie et les coordonnées factorielles issus d'une analyse en composantes principales, par exemple, le principe est le même que pour le coefficient de corrélation ; on effectue sur chaque échantillon simulé, une analyse en composantes principales puis on établit une distribution de fréquences pour chacune des composantes.

La méthode de bootstrap donne dans la plupart des cas une bonne image de la précision statistique de l'estimation sur un échantillon. Les recherches théoriques menées par Efron en particulier montrent que, pour de nombreux paramètres statistiques, l'intervalle de confiance correspondant à la distribution simulée par bootstrap et celui correspondant à la distribution réelle sont généralement de même amplitude. Un exemple classique d'échec du bootstrap est l'estimation des bornes d'un intervalle pour une loi uniforme dans cet intervalle. Il est clair que dans ce cas, l'estimation classique (valeurs extrêmes) ne sera pas améliorée par des tirages à l'intérieur de l'échantillon de base. Pour une revue critique de l'utilisation du bootstrap (avec discussions), on consultera Young (1994).

1.5 Annexe technique du chapitre 1

1.5.1 Démonstration sur les extrema de formes quadratiques sous contraintes quadratiques

Le problème est la recherche du vecteur **u** qui rend maximale la quantité **u'Au**, avec la contrainte **u'Mu** = 1, expression où **A** et **M** sont des matrices symétriques; **M** est de plus définie non-négative et définit la métrique dans $\mathcal{R}^p$.

On donnera deux démonstrations élémentaires pour la solution de ce problème.

L'une fait appel aux Lagrangiens (calcul classique d'extremum sous contrainte), l'autre suppose connues les propriétés spectrales des matrices symétriques.

Le problème est ici plus général que celui rencontré précédemment, pour lequel **A** = **X'X** et **M** = **I** où **I** est la matrice unité.

Mais cette formulation avec une métrique et des masses affectées aux points sera utile pour l'analyse des correspondances et de l'analyse discriminante.

Elle n'introduit pas de difficulté supplémentaire dans les démonstrations.

- Démonstration directe

La forme quadratique **u'Au** s'écrit : $\mathbf{u'Au} = \sum_{ij} a_{ij} u_i u_j$

En dérivant cette quantité successivement par rapport aux p composantes du vecteur **u**, on voit que le vecteur des dérivées partielles de **u'Au** s'écrit sous forme matricielle :

$$\frac{\partial(\mathbf{u'Au})}{\partial \mathbf{u}} = 2\mathbf{Au}, \quad \text{et de même:} \quad \frac{\partial(\mathbf{u'Mu})}{\partial \mathbf{u}} = 2\mathbf{Mu},$$

La recherche d'un maximum lié implique que s'annulent les dérivées du Lagrangien :

$$\mathcal{L} = \mathbf{u'Au} - \lambda(\mathbf{u'Mu} - 1)$$

λ étant un multiplicateur de Lagrange.

Par la suite :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = 2\mathbf{Au} - 2\lambda\mathbf{Mu} = \mathbf{0}$$

exprime la condition d'extremum. On en déduit la relation :

$$\mathbf{A\,u} = \lambda\,\mathbf{M\,u} \qquad [1.5 - 1]$$

Prémultipliant les deux membres de cette relation par **u'**, et tenant compte du fait que **u'Mu** = 1, il vient : $\lambda = \mathbf{u'\,A\,u}$

La valeur du paramètre λ est donc le maximum cherché.

Lorsque la matrice **M** est définie positive, donc inversible, la relation [1.5 - 1] s'écrit alors :

$$\mathbf{M^{-1}A\,u} = \lambda\,\mathbf{u}$$

u est le *vecteur propre* de la matrice $\mathbf{M^{-1}A}$ correspondant à la plus grande *valeur propre* λ (si celle-ci est unique, ce qui sera le cas général).

Appelons désormais $\mathbf{u}_1$, le vecteur **u** correspondant à la plus grande valeur λ_1 telle que la relation [1.5 - 1] soit vérifiée. Cherchons le vecteur $\mathbf{u}_2$, M-unitaire et M-orthogonal à $\mathbf{u}_1$ (c'est-à-dire tel que $\mathbf{u}_2'\mathbf{M}\mathbf{u}_2 = 1$ et $\mathbf{u}_1'\mathbf{M}\mathbf{u}_2 = 0$), qui rend maximale la forme quadratique $\mathbf{u}_2'\mathbf{A}\mathbf{u}_2$.

On est conduit à annuler les dérivées du Lagrangien :

$$\mathcal{L} = \mathbf{u}_2'\mathbf{A}\mathbf{u}_2 - \lambda_2(\mathbf{u}_2'\mathbf{M}\mathbf{u}_2 - 1) - \mu_2\mathbf{u}_2'\mathbf{M}\mathbf{u}_1$$

où λ_2 et μ_2 sont deux multiplicateurs de Lagrange.

La condition d'extremum s'écrit pour $\mathbf{u}_2$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}_2} = 2\mathbf{A}\mathbf{u}_2 - 2\lambda_2\mathbf{M}\mathbf{u}_2 - \mu_2\mathbf{M}\mathbf{u}_1 = 0$$

En multipliant les membres de cette relation par $\mathbf{u}_1'$ on voit que $\mu_2 = 0$

Il reste donc comme précédemment :

$$\mathbf{A}\,\mathbf{u}_2 = \lambda_2 \mathbf{M}\,\mathbf{u}_2$$

Comme **M** est inversible, $\mathbf{u}_2$ est le second vecteur propre de $\mathbf{M}^{-1}\mathbf{A}$, relatif à la seconde plus grande valeur propre λ_2 si celle-ci est unique.

La démonstration s'étend aisément au cas d'un vecteur unitaire $\mathbf{u}_\alpha$ pour $\alpha \leq p$ (i.e. : $\mathbf{u}'_\alpha \mathbf{M}\,\mathbf{u}_\alpha = 1$), M-orthogonal aux vecteurs $\mathbf{u}_\beta$ trouvés précédemment ($\mathbf{u}'_\alpha \mathbf{M}\,\mathbf{u}_\beta = 0$, pour $\beta < \alpha$) et rendant maximale la forme $\mathbf{u}'_\alpha \mathbf{A}\,\mathbf{u}_\alpha$.

On a alors : $\mathbf{A}\,\mathbf{u}_\alpha = \lambda_\alpha \mathbf{M}\,\mathbf{u}_\alpha$ et si **M** est inversible :

$$\mathbf{M}^{-1}\mathbf{A}\,\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$$

- Seconde démonstration

Nous ne ferons qu'esquisser cette démonstration, dans le cas où **M** est définie positive. On peut alors décomposer cette matrice sous la forme classique **M = L'L**, où **L** est inversible puisque **M** est supposée définie positive.

Posant alors $\mathbf{u} = \mathbf{L}^{-1}\mathbf{y}$, la contrainte de normalisation $\mathbf{u'Mu} = 1$ s'écrit maintenant $\mathbf{y'y} = 1$, et la quantité à rendre maximale $\mathbf{u'Au}$ devient $\mathbf{y'Sy}$, avec

$$\mathbf{S} = \mathbf{L}'^{-1}\mathbf{A}\mathbf{L}^{-1}$$

Soit **T** la matrice orthogonale (p,p) dont les colonnes sont les vecteurs propres $\mathbf{t}_\alpha$ de **S**, normés et ordonnés suivant les valeurs propres λ_α décroissantes, et soit Λ la matrice diagonale dont le $\alpha^{\text{ième}}$ élément vaut λ_α.

Posons encore $\mathbf{z} = \mathbf{T'y}$ (ce qui implique $\mathbf{y} = \mathbf{T}\,\mathbf{z}$ car $\mathbf{T}' = \mathbf{T}^{-1}$).

On a alors :

$$\mathbf{y'Sy} = \mathbf{y'T\Lambda T'y} = \mathbf{z'\Lambda z}$$

avec la contrainte $\mathbf{z'z} = 1$.

On remarque que $\lambda_1 \geq \mathbf{z'\Lambda z}$; en effet :

$$\lambda_1 - \mathbf{z'\Lambda z} = \mathbf{z'}\,(\lambda_1 \mathbf{I} - \Lambda)\,\mathbf{z} \geq 0$$

Le maximum λ_1 est effectivement atteint pour $\mathbf{z}' = (1, 0, 0, 0, \ldots, 0)$, donc pour $\mathbf{y} = \mathbf{t}_1$ et pour $\mathbf{u}_1 = \mathbf{L}^{-1}\mathbf{t}_1$.

De la relation $\mathbf{S}\,\mathbf{t}_1 = \lambda_1 \mathbf{t}_1$, on tire :

$$\mathbf{L}'^{-1}\mathbf{A}\mathbf{L}^{-1}\mathbf{t}_1 = \lambda_1 \mathbf{t}_1$$

D'où, finalement[1] :

$$\mathbf{M}^{-1}\mathbf{A}\,\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

---

[1] On note au passage qu'il suffit ici de procéder à la diagonalisation d'une matrice symétrique **S** (après avoir décomposé **M** sous la forme : **M = L'L**), alors que la matrice précédente $\mathbf{M}^{-1}\mathbf{A}$ est en général non-symétrique. Cette propriété est utilisée dans les programmes de calcul.

1.5.2 Variations des valeurs et vecteurs propres en fonction des éléments de la matrice à diagonaliser

Montrons brièvement, en utilisant une formulation empruntée à Gifi (1990), comment des variations de la matrice symétrique à diagonaliser **A**, supposées ici infinitésimales, influencent les éléments propres $\mathbf{u}_r$ et λ_r.

La relation $\mathbf{A}\,\mathbf{u}_r = \lambda_r\,\mathbf{u}_r$ se note, pour l'ensemble du spectre :

$$\mathbf{A}\,\mathbf{U} = \mathbf{U}\,\Lambda \qquad [1.5.2 - 1]$$

avec, rappelons-le, les contraintes $\mathbf{U'U} = \mathbf{I}$. La matrice **A**, et par conséquent **U** et Λ, sont supposées dépendre continûment d'un paramètre s.

La dérivation des relations [4.2 - 1] par rapport à s donne les systèmes :

$$\frac{\partial \mathbf{A}}{\partial s}\mathbf{U} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial s} = \frac{\partial \mathbf{U}}{\partial s}\Lambda + \mathbf{U}\frac{\partial \Lambda}{\partial s}$$
$$\text{et} \quad \frac{\partial \mathbf{U}'}{\partial s}\mathbf{U} + \mathbf{U}'\frac{\partial \mathbf{U}}{\partial s} = \mathbf{0} \qquad [1.5.2 - 2]$$

Prémultipliant les deux membres de la première relation par **U'**, il vient après simplification (mettant à profit les relations précédentes) :

$$\mathbf{U}'\frac{\partial \mathbf{A}}{\partial s}\mathbf{U} = \frac{\partial \Lambda}{\partial s} + \left\{\mathbf{U}'\frac{\partial \mathbf{U}}{\partial s}\Lambda - \Lambda\mathbf{U}'\frac{\partial \mathbf{U}}{\partial s}\right\}$$

Or la matrice entre accolades a ses éléments diagonaux nuls (comme toute matrice de la forme ($\mathbf{B}\Lambda - \Lambda\mathbf{B}$), avec Λ diagonale), d'où l'expression de $\frac{\partial \Lambda}{\partial s}$:

$$\frac{\partial \Lambda}{\partial s} = diag\left(\mathbf{U}'\frac{\partial \mathbf{A}}{\partial s}\mathbf{U}\right)$$

(où le symbole *diag* signifie "diagonale de...").
Posant :

$$\mathbf{Q} = \mathbf{U}'\frac{\partial \mathbf{A}}{\partial s}\mathbf{U} \quad , \quad \text{d'où:} \qquad q_{rr'} = \sum_{i,j} u_{ir}u_{jr'}\frac{\partial a_{ij}}{\partial s}$$

on peut écrire :

$$\frac{\partial \lambda_r}{\partial s} = q_{rr} = \sum_{i,j} u_{ir}u_{jr}\frac{\partial a_{ij}}{\partial s} \qquad [1.5.2 - 3]$$

Pour les vecteurs propres, le calcul, plus complexe, conduit à :

$$\frac{\partial u_{jr}}{\partial s} = \sum_{t \neq r} u_{jt}\frac{q_{rt}}{\lambda_r - \lambda_t} \qquad [1.5.2 - 4]$$

C'est cette dernière formule qui montre que des valeurs propres voisines créent une instabilité des vecteurs propres correspondants. Les écarts entre valeurs propres figurent en effet au dénominateur dans le membre de droite.

Chapitre 2

Analyse canonique et régression linéaire

L'*analyse canonique* joue un rôle théorique important dans les méthodes multidimensionnelles et permet de jeter un pont entre les formalismes des méthodes explicatives et descriptives. C'est pourquoi nous commencerons par exposer ses principes dans ce chapitre (section 2.1).

On verra que l'analyse canonique, qui étudie les liaisons entre deux groupes de variables, contient comme cas particuliers la *régression multiple* si l'un des deux groupes est réduit à une seule variable numérique, l'*analyse discriminante* (étudiée au chapitre 7) lorsque les variables de l'un des deux groupes sont les variables indicatrices d'une partition des individus (la variable à expliquer est nominale), enfin *l'analyse des correspondances simple* (étudiée au chapitre 4) si les deux groupes sont constitués par les variables indicatrices des deux partitions.

La *régression linéaire ou régression multiple* (section 2.2) est le modèle de prédiction le plus ancien et le plus populaire. Il se situe directement dans le cadre théorique du *modèle linéaire*, lorsque la variable à expliquer y est une variable continue (ou numérique). On réserve en général le nom de régression multiple au cas où les variables explicatives sont continues. Lorsque celles-ci sont des variables nominales, on parle *d'analyse de la variance* et pour un ensemble de variables mixtes, *d'analyse de la covariance* (variables explicatives nominales et continues).

La régression linéaire mérite d'être présentée à ce stade au titre des méthodes de base car elle interviendra à différents niveaux par la suite : la technique des variables supplémentaires, fondamentale en analyse des données, est une forme de régression multiple visualisée sur les axes principaux. L'analyse linéaire discriminante (chapitre 7) se réduit à la régression multiple dans le cas, fréquent en pratique, de deux classes. Les analyses en composantes principales partielles (chapitre 8) font également intervenir des régressions, etc.

2.1 Analyse canonique

La méthode d'analyse canonique développée par Hotelling (1936) constitue un cadre théorique général important dont la régression multiple, l'analyse discriminante et l'analyse des correspondances simple sont des cas particuliers.

Sous sa forme générale, l'analyse canonique ne présente cependant qu'un intérêt assez limité pour les applications, car elle conduit à de grandes difficultés d'interprétation.

L'analyse canonique cherche à synthétiser les interrelations existant entre *deux groupes* de variables, en mettant en évidence les combinaisons linéaires des variables du premier groupe les plus *corrélées* à des combinaisons linéaires des variables du second groupe.

2.1.1 Formulation du problème et notations

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

$$\mathbf{R} = [\mathbf{X}, \mathbf{Y}]$$

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

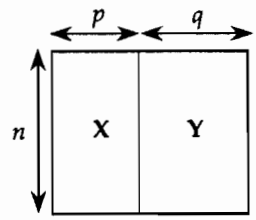


Figure 2.1 – 1. Tableau des données R

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

$$V(\mathbf{R}) = \frac{1}{n}\mathbf{R}'\mathbf{R}$$

Elle a pour terme général :

$$v_{jj'} = \frac{1}{n}\sum_{i} r_{ij} r_{ij'}$$

soit, en faisant apparaître les blocs :

$$V(\mathbf{R}) = \frac{1}{n}\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{bmatrix}$$

Considérons l'individu i, caractérisé par la $i^{ième}$ ligne de **R** :

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

Soient **a** et **b** deux vecteurs à p et q composantes, définissant deux combinaisons linéaires $a(i)$ et $b(i)$:

$$a(i) = \sum_{j=1}^{p} a_j x_{ij} \ , \qquad b(i) = \sum_{j=1}^{q} b_j y_{ij} \ .$$

Les n valeurs de $a(i)$ pour tous les individus i sont les composantes de **Xa**. De même, les n valeurs de $b(i)$ sont les composantes de **Yb**. Les vecteurs **Xa** et **Yb** représentent aussi deux points de $\mathcal{R}^n$ appartenant respectivement aux sous-espaces V_X et V_Y engendrés par les colonnes de **X** et **Y**.

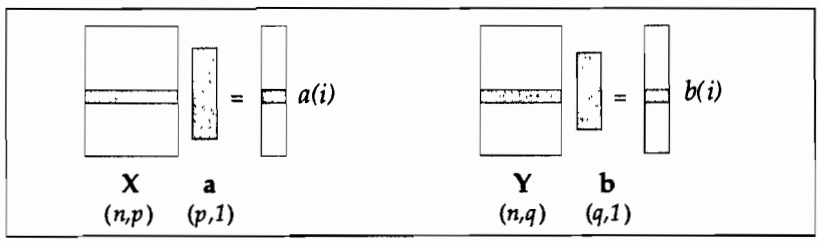


Figure 2.1 - 2. Variables canoniques $a(i)$ et $b(i)$

Nous nous proposons de chercher les deux combinaisons linéaires $a(i)$ et $b(i)$ les plus corrélées sur l'ensemble des valeurs de i. Puisque les variables initiales sont centrées, leurs combinaisons linéaires sont également centrées. Comme le coefficient de corrélation ne dépend pas de l'échelle des variables, nous imposerons aux deux combinaisons linéaires d'avoir une *variance unité*. La variance de l'ensemble des valeurs de $a(i)$ pour $i = 1, 2, \ldots, n$ sera notée $var(\mathbf{a})$; elle s'écrit :

$$var(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n} a^2(i) = \frac{1}{n}(\mathbf{Xa})'\mathbf{Xa} = \frac{1}{n}\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}$$

de la même façon :

$$var(\mathbf{b}) = \frac{1}{n}\mathbf{b}'\mathbf{Y}'\mathbf{Y}\mathbf{b}$$

Dans ces conditions, le coefficient de corrélation entre les combinaisons linéaires $a(i)$ et $b(i)$ coïncide avec la covariance :

$$cov(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} a(i)b(i)$$

soit :

$$cov(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b}$$

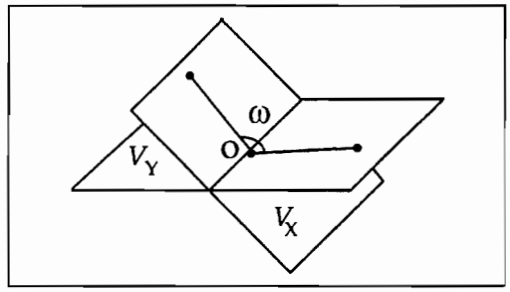


Figure 2.1 – 3. Représentation géométrique des sous-espaces V_X et V_Y

Finalement le problème de la recherche de la corrélation maximale s'écrira, après s'être affranchi des coefficients $\frac{1}{n}$ (rappelons que **X** et **Y** sont centrés) :

- trouver **a** et **b** qui rendent maximal : $\mathbf{a'X'Yb}$
- avec les contraintes : $\begin{cases} \mathbf{a'X'Xa} = 1 \\ \mathbf{b'Y'Yb} = 1 \end{cases}$

Les données étant centrées, le coefficient de corrélation n'est autre que le cosinus de l'angle entre les sous-espaces V_X et V_Y. La recherche des coefficients **a** et **b** revient donc à minimiser l'angle ω entre les sous-espaces V_X et V_Y.

On appellera *variables canoniques* le couple (**a**, **b**), **a** et **b** ayant respectivement p et q composantes.

2.1.2 Les variables canoniques

a – Calcul des variables canoniques

La démonstration est analogue à celle rencontrée lors de l'analyse générale. Deux *multiplicateurs de Lagrange* λ et μ interviennent. Il faut rendre maximal :

$$\mathcal{L} = \mathbf{a'X'Yb} - \lambda(\mathbf{a'X'Xa} - 1) - \mu(\mathbf{b'Y'Yb} - 1)$$

L'annulation des dérivées de ce lagrangien par rapport aux vecteurs **a** et **b** conduit au système :

$$\begin{cases} \mathbf{X'Y\,b} - 2\lambda\,\mathbf{X'Xa} = 0 \\ \mathbf{Y'X\,a} - 2\mu\,\mathbf{Y'Yb} = 0 \end{cases}$$

Prémultiplions les membres de ces deux relations respectivement par **a'** et **b'**. En tenant compte des contraintes :

$$\mathbf{a'X'Xa} = \mathbf{b'Y'Yb} = 1$$

Elles se simplifient en :

$$\begin{cases} \mathbf{a'X'Yb} = 2\lambda \\ \mathbf{b'Y'Xa} = 2\mu \end{cases}$$

Par conséquent $\lambda = \mu$. Nous poserons dorénavant : $\beta = 2\lambda$. On remarquera que β est la valeur du *coefficient de corrélation maximal* recherché. Le système précédent s'écrit alors :

$$\begin{cases} \mathbf{X'Yb} = \beta \mathbf{X'Xa} & [2.1-1] \\ \mathbf{Y'Xa} = \beta \mathbf{Y'Yb} & [2.1-2] \end{cases}$$

La résolution est immédiate quand les matrices **X'X** et **Y'Y** sont *inversibles*. En reportant la valeur de **a** tirée de [2.1 - 1] dans la relation [2.1 - 2], par exemple, on obtient :

$$\mathbf{Y'X(X'X)^{-1}X'Yb} = \beta^2 \mathbf{Y'Yb} \qquad [2.1 - 3]$$

Ceci montre que **b** est *vecteur propre* de la matrice :

$$\mathbf{M} = \mathbf{(Y'Y)^{-1}Y'X(X'X)^{-1}X'Y}$$

relatif à la plus grande *valeur propre* notée β^2, carré du coefficient de corrélation entre les combinaisons linéaires **a** et **b** et carré du cosinus maximum de l'angle entre les sous-espaces V_X et V_Y. Cette valeur β^2 est la première *racine canonique*, ou carré du premier *coefficient de corrélation canonique* entre les deux variables.

De façon analogue, on calcule **a** à partir de la relation [2.1 - 1] ou en considérant directement **a** comme vecteur propre de :

$$\mathbf{N} = \mathbf{(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X} \qquad [2.1 - 4]$$

Si **X** est de plein rang, alors **X'X** est inversible et la relation [2.1 - 1] permet d'écrire :

$$\mathbf{a} = \frac{1}{\beta}\mathbf{(X'X)^{-1}X'Yb}$$

Un raisonnement analogue à celui fait lors de l'analyse générale permet de généraliser ce résultat à la recherche des r variables canoniques, r étant le plus petit des deux entiers p et q : les r vecteurs propres successifs, dans l'ordre des valeurs propres décroissantes, correspondent aux couples de combinaisons linéaires les plus corrélées entre elles, les combinaisons linéaires successives relatives à un même ensemble étant assujetties à être non corrélées.

b – Interprétation géométrique

Les relations [2.1 - 1] et [2.1 - 2] peuvent s'écrire :

$$\mathbf{a} = \frac{1}{\beta}\mathbf{(X'X)^{-1}X'Yb}, \quad \text{et} \quad \mathbf{b} = \frac{1}{\beta}\mathbf{(Y'Y)^{-1}Y'Xa}$$

Prémultipliant les deux membres de chacune d'elles respectivement par **X** et **Y** on obtient :

$$\mathbf{Xa} = \frac{1}{\beta}\mathbf{X(X'X)^{-1}X'Yb} \qquad [2.1 - 5]$$

$$\mathbf{Yb} = \frac{1}{\beta}\mathbf{Y(Y'Y)^{-1}Y'Xa} \qquad [2.1 - 6]$$

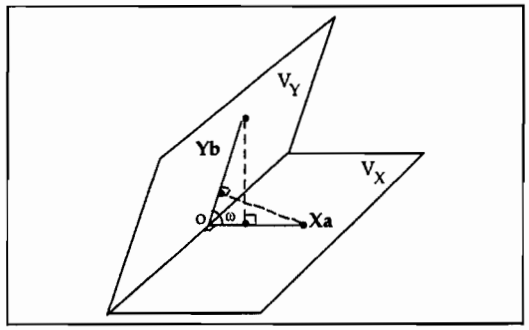


Figure 2.1 - 4 . Interprétation géométrique de l'analyse canonique

Les matrices symétriques et idempotentes :

$$\mathbf{P_X} = \mathbf{X(X'X)^{-1}X'} \qquad \text{et} \qquad \mathbf{P_Y} = \mathbf{Y(Y'Y)^{-1}Y'}$$

sont les *opérateurs de projection orthogonale* respectivement sur les sous-espaces V_X et V_Y.

Autrement dit les relations [2.1 - 5] et [2.1 - 6] expriment que chacun des vecteurs **Xa** et **Yb** est colinéaire à la projection de l'autre.

Les vecteurs **Xa** et **Yb** étant unitaires, les formules montrent en effet que :

$$\beta = cos(\omega) = cos(\mathbf{Xa}, \mathbf{Yb})$$

Il apparaît que la première racine canonique β^2 est le carré du cosinus du plus petit angle[1] entre les sous-espaces V_X et V_Y.

c – Cas de matrices non inversibles

Examinons le cas où les matrices **X'X** ou **Y'Y** sont singulières. Prenons **Y'Y** pour fixer les idées. Cela signifie que la matrice **Y** d'ordre (n, q) a un rang inférieur à q ; soit $q - s$ son rang.

Il y a deux façons de procéder pour résoudre le système des équations matricielles [2.1 - 1] et [2.1 - 2] :

- on prend dans $\mathcal{R}^n$ une *base* du sous-espace V_Y à $q - s$ dimensions engendrée par **Y**, base décrite par les $q - s$ colonnes d'une matrice $\hat{\mathbf{Y}}$; on choisira de préférence une base orthogonale, obtenue, par exemple, par le procédé d'orthogonalisation de Gram-Schmidt, ou une base issue d'une analyse

---

[1] Notons que ces considérations géométriques nous auraient permis d'écrire *directement* les formules [2.1 - 5] et [2.1 - 6], et donc de procéder au calcul des variables canoniques : on remplace, par exemple dans la relation [2.1 - 6], **Xa** par sa valeur tirée de la relation [2.1 - 5].

générale de **Y**. A **Yb**, on substitue dans les calculs $\hat{\mathbf{Y}}\hat{\mathbf{b}}$ où $\hat{\mathbf{b}}$ est un vecteur à $q - s$ composantes. La matrice $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ est maintenant inversible.

- Comme cela est fréquent dans le cas du modèle linéaire général, on construit une matrice $\mathbf{Y}_0$ de *plein rang* d'ordre (n, q), telle que $V_Y \subset V_{Y_0}$. Pour retrouver le sous-espace V_Y, il est alors nécessaire d'imposer à **b** une *contrainte*, à savoir : $\mathbf{Y}_0\mathbf{b}$ devra appartenir à V_Y. Si $\mathbf{Y}_1$ désigne une matrice d'ordre (n,s), telle que $\mathbf{Y}_1'\mathbf{Y} = \mathbf{0}$ et que $\mathbf{Y}_1\mathbf{b} \in V_{Y_0}$, la contrainte sur **b** s'écrira :

$$\mathbf{Y}_1'\mathbf{Y}_0\mathbf{b} = \mathbf{0}$$

Remarque :

Cette situation se présentera également en *analyse discriminante* dans un contexte simple : la matrice **Y** d'ordre (n, q) est singulière, alors que la matrice initiale $\mathbf{Y}_0$ (avant centrage) est de plein rang. Ceci résulte du fait que le sous-espace V_{Y_0} engendrée par $\mathbf{Y}_0$ contient le vecteur $\mathbf{e}_n$ de $\mathcal{R}^n$ dont toutes les composantes valent *1*. On travaillera alors avec la matrice $\mathbf{Y}_0$ sachant que **b** est assujetti à vérifier :

$$\mathbf{e}_n'\mathbf{Y}_0\mathbf{b} = 0$$

relation qui s'écrit :

$$\sum_{j=1}^{q} y_{.j}\, b_j = 0$$

($y_{.j}$ désignant la somme de la colonne j de la matrice $\mathbf{Y}_0$).

2.2 Régression multiple, modèle linéaire

La régression multiple vise à expliquer ou prédire une variable continue (dite variable dépendante ou à expliquer ou encore endogène) à l'aide d'un ensemble de variables dites explicatives (ou exogènes). C'est sans doute la méthode statistique la plus utilisée bien que sa portée et ses limites ne soient pas toujours bien connues. De ce fait, elle n'est pas toujours pratiquée à bon escient. La littérature sur la régression et le modèle linéaire est extrêmement abondante[1]. C'est en économétrie, champ d'application privilégié du modèle linéaire, que

[1] La littérature en anglais sur le modèle linéaire est particulièrement vaste : on trouvera une bibliographie commentée (déjà ancienne) de plusieurs centaines d'articles et ouvrages dans Harter (1974 -1975). Searle (1971) et Seber (1977) traitent de façon extensive les problèmes d'analyse de la variance et de covariance; Theil (1971) situe le modèle linéaire dans un cadre économétrique général; l'ouvrage de Rao (1973), réédition d'un manuel classique, est consacré à l'opération d'induction statistique sur le modèle linéaire. Un autre manuel classique est l'ouvrage de Draper et Smith (1981). Mosteller et Tukey (1977), Besley *et al* (1980), Atkinson (1985) présentent des points de vue un peu plus modernes, incluant diverses méthodes de sélection de variables, alors que Chatterjee et Price (1991) insistent sur la mise en oeuvre pratique.

l'on trouve les premiers manuels généraux en langue française exposant les méthodes et les principaux types de résultats (Malinvaud, 1964; Fourgeaud et *al.*, 1978). On citera également l'ouvrage de Tomassone et *al.* (1983), exposé complet, simple et opérationnel sur tous les aspects de la régression. Pour un exposé plus concis, on renverra à Saporta (1990). Mais ces quelques titres ne sauraient rendre justice de la profusion des excellents manuels sur ce sujet.

2.2.1 Formulation du problème : le modèle linéaire

On dispose d'un ensemble de n observations sur lesquelles ont été effectuées $p+1$ mesures des variables $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. On veut expliquer ou prévoir $\mathbf{y}$ à l'aide des variables explicatives ou prédicteurs, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, lesquels sont supposés connus sans erreur.

Un exemple : Supposons par exemple qu'une personne désire acquérir un magasin ayant une surface S dans une zone où la population environnante est P. Des études antérieures montrent que le chiffre d'affaires d'un magasin dépend linéairement de la surface et de la population, et les données relatives à *30* magasins du même type sont disponibles. Quel chiffre d'affaires peut espérer l'acheteur ? Le chiffre d'affaires est la variable à prévoir et les variables explicatives ou prédicteurs sont la population et la surface. Ce type de problème trouve une solution dans le cadre de la régression, technique de prévision linéaire, qui consiste tout d'abord à procéder à une estimation des paramètres d'un modèle, puis à utiliser le modèle estimé pour le calcul de la valeur attendue.

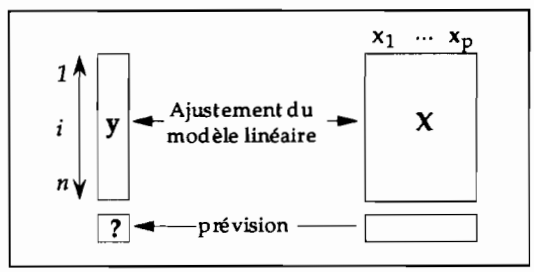


Figure 2.2 – 1. Prévision linéaire

On cherche à approcher $\mathbf{y}$ par une combinaison linéaire des variables explicatives $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

Pour cela, on pose le modèle[1] :

---

[1] La linéarité des relations par rapport aux coefficients $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_p$ peut n'apparaître qu'après transformations des données. Par exemple :

$$y = \alpha_3 x_1^{\alpha_1} x_2^{\alpha_2} (1+\varepsilon) \text{ avec } x_1 > 0, x_2 > 0, \alpha_3 > 0$$

deviendra linéaire après la transformation logarithmique :

$$log(y) = \alpha_1 log(x_1) + \alpha_2 log(x_2) + log(\alpha_3) + log(1+\varepsilon)$$

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_p x_{ip} + \varepsilon_i$$

où α_0, α_1, α_2,..., α_p sont les coefficients inconnus du modèle. Le terme constant α_0 peut être considéré comme coefficient d'une variable explicative particulière artificielle x_0 dont les valeurs x_{i0} seraient toujours égales à *1*. ε_i est le résidu représentant l'écart entre la valeur observée y_i et la partie "expliquée" de l'observation $(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_p x_{ip})$. On suppose, dans la plupart des spécifications du modèle, que tous les résidus ε_i sont des quantités aléatoires indépendantes. Ce modèle s'exprime sous forme matricielle :

$$\underset{(n,1)}{\mathbf{y}} = \underset{(n,p+1)}{\mathbf{X}} \underset{(p+1,1)}{\boldsymbol{\alpha}} + \underset{(n,1)}{\boldsymbol{\varepsilon}}$$

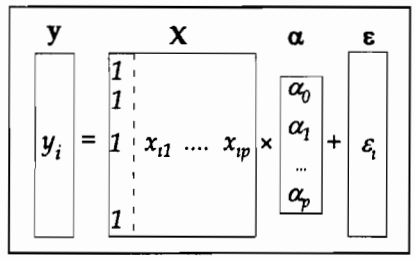


Figure 2.2 - 2 · Schématisation du modèle linéaire

On dispose, pour évaluer les coefficients inconnus du modèle, d'un système de n équations linéaires ayant $n + p + 1$ inconnues ($p+1$ coefficients α_0, α_1, α_2,..., α_p et les n résidus). Le système admet donc une infinité de solutions.

Soient $a_0, a_1, a_2, \ldots, a_p$ les coefficients correspondant à une des solutions possibles. On cherchera la solution qui minimise globalement, suivant un critère à définir, l'ensemble des écarts à la linéarité, c'est-à-dire :

$$\begin{cases} \text{choisir } (a_0, a_1, a_2, \ldots, a_p) \text{ qui minimisent l'ensemble des } e_i \\ \text{avec } e_i = y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip}) \end{cases}$$

Parmi les critères possibles de minimisation, citons la méthode des moindres carrés $min\{\sum e_i^2\}$ (norme dite "L_2") celle des moindres valeurs absolues $min\{\sum |e_i|\}$ (norme dite "L_1"), celle du minimax $min\{\max_{(i)} e_i\}$ (norme dite "L_∞")[1].

Le critère des moindres carrés conduit à des calculs algébriques simples, se prête à une interprétation géométrique claire, et donne lieu à des interprétations statistiques intéressantes.

---

[1] Plus généralement, la norme L_k correspond au critère $min\{\sum |e_i|^k\}$. La norme L_1, qui privilégie moins les écarts importants, est à la base de méthodes de régression plus *robustes* (cf. Huber, 1981; 1987). Sur le rôle de cette norme en analyse descriptive des données, cf. Fichet (1987), et Le Calvé (1987). L'utilisation de la norme L_1 dans le cas de la régression linéaire remonte à Laplace (1793). Une étude historique de l'utilisation des normes L_1 et L_∞ a été réalisée par Farebrother (1987).

2.2.2 Ajustement par la méthode des moindres-carrés

On appelle ajustement du modèle linéaire toute solution du système d'équations :

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip} + e_i \qquad (i = 1, 2, \ldots, n)$$

ce qui correspond sous forme matricielle à :

$$\underset{(n,1)}{\mathbf{y}} = \underset{(n,p+1)}{\mathbf{X}}\ \underset{(p+1,1)}{\mathbf{a}} + \underset{(n,1)}{\mathbf{e}}$$

Pour la $i^{ème}$ observation, la valeur prédite par le modèle est :

$$\tilde{y}_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip}$$

le résidu du modèle correspondant vaut donc[1] :

$$e_i = y_i - \tilde{y}_i$$

D'une manière générale, on cherche $\tilde{\mathbf{y}}$ le plus proche possible de $\mathbf{y}$:

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{a} = a_0\mathbf{x}_0 + a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_p\mathbf{x}_p$$

L'ajustement par la méthode des moindres carrés est celui qui fournit les coefficients $a_0, a_1, a_2, \ldots, a_p$ conduisant au minimum de la somme des carrés des écarts :

$$min\{\sum e_i^2\}$$

Dans la suite, nous allons supposer que les variables sont centrées, ce qui implique $a_0 = 0$. Une des propriétés de la régression multiple est que les estimations des coefficients autres que a_0 sont les mêmes, que les variables soient centrées *a priori* ou pas.

a – Calcul et propriétés de l'ajustement des moindres-carrés

Il s'agit de déterminer le vecteur $\mathbf{a}$ des coefficients qui minimise :

$$\mathbf{e}'\mathbf{e} = \sum e_i^2 = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2$$

Le vecteur de coefficients $\mathbf{a}$ doit vérifier la condition d'extremum[2] :

$$\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{X}'\mathbf{y} \qquad [2.2 - 1]$$

qui est un système de p équations à p inconnues.

---

[1] Le vocabulaire et les notations distinguent les résidus définis par le modèle théorique $\varepsilon_i = y_i - \sum_k \alpha_k x_{ik}$ et les écarts définis par un ajustement $e_i = y_i - \sum_k a_k x_{ik}$

[2] La quantité scalaire $\mathbf{e}'\mathbf{e}$ étant une fonction des inconnues $(a_1, a_2, \ldots, a_p)$, une condition nécessaire d'extremum est l'annulation des dérivées partielles premières, soit :

$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{e}'\mathbf{e}) = \underset{(p,1)}{\mathbf{0}}$$

on a : $$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{a})'(\mathbf{y} - \mathbf{X}\mathbf{a}) = \mathbf{y}'\mathbf{y} - 2\mathbf{a}'\mathbf{X}'\mathbf{y} + \mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}$$

d'où : $$\frac{\partial}{\partial \mathbf{a}}(\mathbf{e}'\mathbf{e}) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{a}$$

on en tire la condition d'extremum : $\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{X}'\mathbf{y}$

Si n est supérieur ou égal à p (plus d'équations que d'inconnues) et si $\mathbf{X}$ est de plein rang (c'est-à-dire de rang p), alors $\mathbf{X'X}$ est inversible.

On tire de la relation [2.2 - 1] la solution :

$$\mathbf{a} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad [2.2 - 2]$$

Le vecteur $\mathbf{a}$ est le vecteur des coefficients de régression multiple[1].

Il reste à vérifier que l'extremum atteint par $\mathbf{e'e}$ est bien un minimum.

Soit $\tilde{\mathbf{a}}$ une autre solution et $\tilde{\mathbf{e}}$ le vecteur correspondant des écarts :

$$\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\mathbf{a}} = (\mathbf{y} - \mathbf{X}\mathbf{a}) + (\mathbf{X}\mathbf{a} - \mathbf{X}\tilde{\mathbf{a}}) = \mathbf{e} + \mathbf{X}(\mathbf{a} - \tilde{\mathbf{a}})$$

et

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \mathbf{e'e} + 2(\mathbf{a} - \tilde{\mathbf{a}})\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{a}) + (\mathbf{a} - \tilde{\mathbf{a}})'\mathbf{X'X}(\mathbf{a} - \tilde{\mathbf{a}})$$

Dans le membre de droite, le terme central est nul d'après [2.2 - 1]; il reste donc :

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \mathbf{e'e} + (\mathbf{X}(\mathbf{a} - \tilde{\mathbf{a}}))' (\mathbf{X}(\mathbf{a} - \tilde{\mathbf{a}}))$$

Il est clair que le dernier terme est une somme de carrés et ne peut être que positif ou nul. Par conséquent $\mathbf{e'e}$ est bien la plus petite somme de carrés d'écarts.

b – Approche géométrique dans $\mathcal{R}^n$

Les propriétés algébriques de l'ajustement vont nous permettre d'interpréter géométriquement l'opération effectuée.

Plaçons-nous dans l'espace $\mathcal{R}^n$ où n est le nombre des observations effectuées sur $p+1$ variables : $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

La recherche de $\mathbf{y}$ comme combinaison linéaire des $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ revient à définir $\tilde{\mathbf{y}}$ dans le sous-espace engendré par les variables explicatives V_X. La technique d'ajustement des moindres-carrés consiste alors à approcher $\mathbf{y}$ par sa projection orthogonale $\tilde{\mathbf{y}}$ sur le sous-espace V_X.

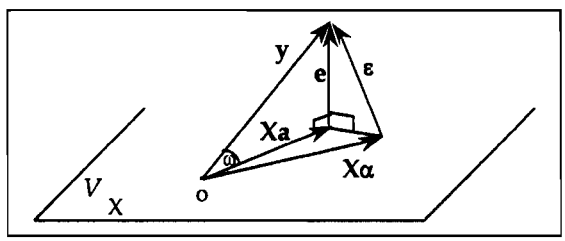


Figure 2.2 - 3 · Projection de la variable à expliquer y sur le sous-espace explicatif V_X

[1] La *régression simple* correspond au modèle $\mathbf{y} = \alpha\mathbf{x} + \varepsilon$ (une seule variable explicative, $\mathbf{y}$ et $\mathbf{x}$ centrés). La formule [2.2 - 2] devient $a = \mathbf{x'y}/\mathbf{x'x}$ ou $a = cov(x,y)/var(x)$.

En remplaçant **a** par sa valeur obtenue dans [2.2 - 2], on obtient :

$$\tilde{\mathbf{y}} = \mathbf{Xa} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_X\mathbf{y}$$

avec :

$$\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad [2.2 - 3]$$

où la matrice $\mathbf{P}_X$ désigne l'opérateur de projection orthogonale[1] sur V_X. Comme le montre la figure 2.2 - 3, le modèle théorique $\mathbf{y} = \mathbf{X}\alpha + \boldsymbol{\varepsilon}$ définit une décomposition de **y** en deux termes inconnus, l'un $\mathbf{X}\alpha$ dans V_X et l'autre $\boldsymbol{\varepsilon}$ dans $\mathcal{R}^n$. La technique des moindres-carrés propose pour solution la décomposition $\mathbf{y} = \mathbf{Xa} + \mathbf{e}$ qui minimise la "longueur" de **e** en projetant orthogonalement **y** en **Xa** sur V_X et $\boldsymbol{\varepsilon}$ en **e** sur le sous-espace orthogonal à V_X dans $\mathcal{R}^n$. Les deux vecteurs **Xa** et **e** sont orthogonaux.

c – Le coefficient de corrélation multiple

Remarquons que les variables étant centrées, les longueurs dans l'espace $\mathcal{R}^n$ s'interprètent en termes de variances. Le théorème de Pythagore appliqué au triangle rectangle de la figure 2.2 - 3 dont les côtés sont **e** et **Xa** et l'hypoténuse **y**, peut s'écrire :

$$\mathbf{y}'\mathbf{y} = \mathbf{e}'\mathbf{e} + \mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}$$

En divisant par n chacun de ces termes, on obtient la relation :

$$\frac{1}{n}\sum(y_i)^2 = \frac{1}{n}\sum(y_i - \tilde{y})^2 + \frac{1}{n}\sum(\tilde{y}_i)^2$$

variance totale — *variance résiduelle* — *variance expliquée*

Afin d'avoir une idée globale de la qualité de l'ajustement, on définit le coefficient de corrélation multiple R comme le cosinus de l'angle ω entre **y** et **Xa** qui n'est autre que le coefficient de corrélation entre les valeurs initiales et les valeurs ajustées :

$$R = cor(\mathbf{y}, \tilde{\mathbf{y}}) = cor(\mathbf{y}, \mathbf{Xa})$$

Son carré peut s'exprimer sous différentes formes :

$$R^2 = \frac{cov^2(\mathbf{y}, \tilde{\mathbf{y}})}{var(\mathbf{y})var(\tilde{\mathbf{y}})} = \frac{var(\tilde{\mathbf{y}})}{var(\mathbf{y})} = \frac{\sum(\tilde{y}_i)^2}{\sum(y_i)^2} = \frac{variance\ expliquée}{variance\ totale}.$$

De façon explicite en fonction des données initiales **X** et **y**, R^2 s'écrit :

$$R^2 = \frac{\mathbf{a}'\mathbf{X}'\mathbf{X}\,\mathbf{a}}{\mathbf{y}'\mathbf{y}} = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$$

---

[1] Cet opérateur, symétrique et idempotent, a déjà été rencontré à propos de l'analyse canonique (cf. § 2.1.2).

Ce coefficient décrit donc le partage de la variance totale en variance "expliquée" et "résiduelle" :

$$\begin{cases} \text{variance expliquée} & R^2\text{var}(\mathbf{y}) = \text{var}(\tilde{\mathbf{y}}) \\ \text{variance résiduelle} & (1 - R^2)var(\mathbf{y}) = \text{var}(\mathbf{e}) \\ \hline \text{variance totale} & \text{var}(\mathbf{y}) = \text{var}(\tilde{\mathbf{y}}) + \text{var}(\mathbf{e}) \end{cases}$$

Ainsi, en minimisant $\sum e_i^2$, on maximise R^2. En d'autres termes, l'ajustement des moindres-carrés détermine la combinaison linéaire des variables explicatives ayant une corrélation maximale[1] avec la variable à expliquer **y**.

2.2.3 Lien avec l'analyse canonique

La régression multiple est un cas particulier de l'analyse canonique quand la matrice **Y** n'a qu'une colonne **y** ($q = 1$), et donc le sous-espace V_Y est réduit à une droite. La variable canonique **b** n'a alors qu'une composante notée b. Le produit **y'y** étant maintenant un scalaire, la relation [2.1-3] (cf. § 2.1.2.a) devient :

$$\beta^2 = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$$

L'unique racine canonique β^2 est *le carré du coefficient de corrélation multiple* entre la colonne **y** et les colonnes de **X** c'est-à-dire entre la variable à expliquer et les variables explicatives.

Compte tenu de la relation [2.1-1], la variable canonique **a** s'écrit :

$$\mathbf{a} = \frac{b}{\beta}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Cette relation montre que le vecteur **a** est proportionnel (au coefficient $\frac{b}{\beta}$ près) au vecteur des coefficients de la régression multiple expliquant la variable **y** par les p variables colonnes de **X**.

Le coefficient $\frac{b}{\beta}$ est d'ailleurs facile à calculer puisque, d'après la contrainte de normalisation, $b = \frac{1}{\sqrt{\mathbf{y}'\mathbf{y}}}$.

---

[1] On remarquera par ailleurs que l'introduction dans le modèle d'une nouvelle variable explicative quelconque ne peut que diminuer la somme des carrés des écarts et par conséquent augmenter R. En ajoutant en effet une dimension à V_X, on ne peut que diminuer la distance de **y** à ce sous-espace. Dans ces conditions, la valeur prise par R ne peut être un critère absolu pour apprécier la qualité de l'ajustement.

2.2.4 Qualité de l'ajustement

Jusqu'à présent, on s'est borné à résoudre un problème purement numérique d'ajustement, avec une mesure globale de qualité fournie par le coefficient de corrélation multiple. Il s'agit maintenant de tester la signification statistique des coefficients de régression, ce qui nécessite de faire des hypothèses sur $\mathbf{y}$ et $\boldsymbol{\varepsilon}$.

a – Spécification du modèle

On suppose que le résidu ε_i est l'effet résultant d'un grand nombre de causes non identifiées, et à ce titre, on le considérera comme une perturbation *aléatoire*. Ce point de vue étendu aux n relations du modèle introduit un vecteur aléatoire de résidus $\boldsymbol{\varepsilon}$ (ayant n composantes) et, par cet intermédiaire, définit $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ comme vecteur aléatoire. Le tableau 2.2 - 1 résume les caractéristiques des différents éléments du modèle :

Tableau 2.2 – 1. Caractéristiques des éléments du modèle

$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$	Observé	Non observable
Aléatoire	$\underset{(n,1)}{\mathbf{y}}$	$\underset{(n,1)}{\boldsymbol{\varepsilon}}$
Non aléatoire	$\underset{(n,p)}{\mathbf{X}}$	$\underset{(p,1)}{\boldsymbol{\alpha}}$

On supposera que les résidus ε_i ont une espérance nulle, qu'ils ont tous la même variance σ^2 et sont deux à deux non corrélés :

$$E(\boldsymbol{\varepsilon}) = \underset{(1,n)}{\mathbf{0}} \qquad \text{et} \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \underset{(n,n)}{\sigma^2\mathbf{I}}$$

ce qui implique les relations :

$$E(\mathbf{y}) = \underset{(1,n)}{\mathbf{X}\boldsymbol{\alpha}} \quad \text{et} \quad \mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \underset{(n,n)}{\sigma^2\mathbf{I}} \qquad [2.2 - 4]$$

Sous ces hypothèses, les coefficients de régression a_k, $(k=1,\ldots,p)$, fournis pas la technique des moindres-carrés sont les "meilleurs" estimateurs[1] des coefficients inconnus α_k.

b – Moyenne et variance des coefficients

Le vecteur $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ des coefficients de régression étant une fonction de $\mathbf{y}$, est lui même un vecteur aléatoire. Les formules [2.2 - 2] et [2.2 - 4] nous montrent immédiatement que son espérance mathématique s'écrit : $E(\mathbf{a}) = \boldsymbol{\alpha}$.

---

[1] Il s'agit plus précisément d'estimateurs à *variance minimale* sur l'ensemble des estimateurs linéaires, cette propriété étant connue sous le nom de théorème de Gauss-Markov. On renvoie aux ouvrages cités au début de ce chapitre pour plus de détails sur ce théorème et ses généralisations.

Un calcul élémentaire[1] montre que la matrice des covariances des coefficients s'écrit :

$$\mathbf{V(a)} = \sigma^2(\mathbf{X'X})^{-1}$$

Notons que σ^2 est la variance théorique des résidus et n'est donc pas connue. On peut estimer σ^2 par $ns^2/(n-p)$, proche de la variance empirique s^2 des écarts calculés après l'ajustement.

Si l'on désigne par **V** la matrice des covariances empiriques des variables explicatives supposées centrées ($\mathbf{V} = \frac{1}{n}\mathbf{X'X}$), on a la relation :

$$\mathbf{V(a)} = \frac{\sigma^2}{n}\mathbf{V}^{-1}$$

On remarque la dualité qui existe entre les variables explicatives et les coefficients de ces variables dans le modèle. Des variables explicatives non corrélées (matrice **V** diagonale) conduiront à des coefficients de régression non-corrélés. Ce lien entre structure des prédicteurs et structure des coefficients sera précisé dans le paragraphe du chapitre 3 consacré à la régression sur composantes principales.

c – Tests sous l'hypothèse de normalité des résidus

Les résultats précédents (coefficient de corrélation multiple, matrices des covariances des coefficients) permettent d'imaginer des procédures de validation sous des hypothèses assez générales. Le fait de spécifier la loi des résidus autorise des épreuves de validation classiques que l'on rappelle ici, sans démonstration.

1- Test sur les coefficients de régression

Pour savoir si une variable explicative $\mathbf{x}_k$ a une influence réelle sur la variable à expliquer **y**, on procède à un test d'hypothèse sur le coefficient de régression α_k.

L'hypothèse nulle (H_0) est l'éventuelle non-influence qui se traduit par :

(H_0) $\alpha_k = 0$ (les autres coefficients sont quelconques)

On écrit alors la statistique de Student :

$$t = \frac{a_k}{s_k}$$

---

[1] La variance de **a** s'écrit $\mathbf{V(a)} = E[(\mathbf{a} - \alpha)(\mathbf{a} - \alpha)']$.

Or, $\mathbf{a} - \alpha = (\mathbf{X'X})^{-1}\mathbf{X'y} - \alpha$

d'où : $\mathbf{a} - \alpha = (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}\alpha + \varepsilon) - \alpha$

soit : $\mathbf{a} - \alpha = (\mathbf{X'X})^{-1}\mathbf{X'}\varepsilon$

On obtient donc : $E[(\mathbf{a} - \alpha)(\mathbf{a} - \alpha)'] = (\mathbf{X'X})^{-1}\mathbf{X'}E(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X'X})^{-1}$

Finalement : $\mathbf{V(a)} = \sigma^2(\mathbf{X'X})^{-1}$

où s_k est l'estimation de l'écart-type du $k^{ème}$ coefficient de régression a_k :

$$s_k = \sqrt{\frac{\|\mathbf{y} - \mathbf{Xa}\|^2}{n-p} a_{kk}}$$

où a_{kk} désigne le $k^{ème}$ élément diagonal de $(\mathbf{X'X})^{-1}$. Si (H_0) est vraie, la statistique suit une loi de Student à $(n-p)$ degrés de liberté[1]. Soit p_c la probabilité tirée de la loi de Student correspondant à la valeur t_c prise par t :

$$p_c = P(|t| \geq t_c)$$

Si cette probabilité est jugée "trop faible", on rejette[2] l'hypothèse (H_0). On peut étendre la procédure de ce test à une combinaison linéaire quelconque des coefficients.

2- Test sur un sous-ensemble de coefficients

On vient de voir comment tester l'un après l'autre la nullité de chaque coefficient. Cependant, les réponses à des questions telles "$\alpha_1 = 0$ sans rien supposer sur α_2"? puis "$\alpha_2 = 0$ sans rien supposer sur α_1 ?" ne déterminent pas la réponse à cette autre question : "$\alpha_1 = 0$ et simultanément $\alpha_2 = 0$?" D'où l'utilité de savoir tester la *nullité simultanée* de plusieurs coefficients de régression.

On se place ici, sans perte de généralité, dans le cas où les q coefficients sont les premiers des p coefficients. L'hypothèse H_0 se traduit par :

- (H_0) $\quad \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$ (les autres α_k quelconques)
- (H_1) $\quad$ un au moins des q premiers α_k n'est pas nul

Convenons de noter $\mathbf{X}_{H_0}$ les $p-q$ dernières colonnes de $\mathbf{X}$ et $\boldsymbol{\alpha}_{H_0}$ les $p-q$ dernières composantes de $\boldsymbol{\alpha}$. L'écriture matricielle des modèles sera :

$$\begin{cases} \text{modèle (complet) sous } \mathbf{H}_1 : & \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ \text{modèle (réduit) sous } \mathbf{H}_0 : & \mathbf{y}_0 = \mathbf{X}_{H_0}\boldsymbol{\alpha}_{H_0} + \boldsymbol{\varepsilon} \end{cases}$$

On considère la statistique F qui suit une loi de Fisher[3] à q et $n-p$ degrés de liberté :

---

[1] Le modèle contient $p+1$ coefficients à estimer : le terme constant et les coefficients des p variables explicatives.

[2] On effectue par exemple le test au seuil de confiance $0,05$: si $p_C < 0,05$ on rejette l'hypothèse selon laquelle la variable x_K n'a pas d'influence réelle (avec moins de 5 chances sur 100 de se tromper) ; alors que si $p_C \geq 0,05$, on ne peut pas rejeter cette hypothèse.

[3] Le principe de tous ces tests est très simple : les statistiques F sont des quotients de χ^2 indépendants. Les χ^2 sont indépendants car ils correspondent à des composantes normales orthogonales du vecteur résiduel (ici : côté de l'angle droit du triangle rectangle $(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{y}}_0)$ dont l'hypothénuse est $(\mathbf{y}, \tilde{\mathbf{y}}_0)$).

$$F = \frac{\left(\|\mathbf{y} - \tilde{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \tilde{\mathbf{y}}\|^2\right)/q}{\|\mathbf{y} - \tilde{\mathbf{y}}\|^2/(n-p)} \qquad [2.2 - 5]$$

On note les sommes des carrés des écarts :

$$S_0 = \|\mathbf{y} - \tilde{\mathbf{y}}_0\|^2 \quad \text{et} \quad S_1 = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2$$

Si la différence entre les deux quantités S_0 et S_1 est grande (F grand) alors l'effet des q premières variables est important et on devra rejeter l'hypothèses nulle; les q variables $\mathbf{x}_1,\ldots, \mathbf{x}_q$ ont simultanément une influence sur $\mathbf{y}$. On effectue donc deux ajustements successifs pour calculer d'une part S_1 sur le modèle complet et d'autre part S_0 sur le modèle pour lequel sont exclues les q variables explicatives en cause.

2.2.5 Régression sur variables nominales : l'analyse de la variance

Lorsque les variables explicatives sont nominales, la régression multiple n'est autre que *l'analyse de la variance*, technique liée aux plans d'expériences et aux traitements statistiques des données expérimentales[1].

Il est courant d'opposer données d'observation et données expérimentales, en réservant les méthodes exploratoires pour les premières, et les méthodes inférentielles ou confirmatoires pour les secondes. La distinction n'est pas si nette en pratique : d'une part, nous l'avons vu, beaucoup de concepts et d'outils sont communs ; d'autre part, les champs d'application peuvent fréquemment se recouvrir, et une attitude méthodologique trop rigide pourrait être néfaste. D'où l'intérêt de connaître les principes et les possibilités des outils de l'analyse des données expérimentales.

a – Codage des variables nominales

Supposons que l'on dispose sur une variable $\mathbf{y}$ de n observations classées selon p variables nominales $\mathbf{x}_1,\ldots, \mathbf{x}_l, \ldots, \mathbf{x}_p$ à respectivement $m_1,\ldots, m_l, \ldots, m_p$ modalités. Le tableau des variables explicatives $\mathbf{X}$ se présente maintenant sous la forme d'un tableau disjonctif complet[2] noté $\left[\mathbf{x}_1,\ldots,\mathbf{x}_l,\ldots,\mathbf{x}_p\right]$.

---

[1] C'est R.A. Fisher qui est à l'origine de l'analyse de la variance et des plans d'expérience, dans une série d'articles datant des années vingt, repris dans l'ouvrage historique "The Design of Experiments" (Fisher, 1935). Citons également sur ce sujet les traités de Cochran et Cox (1957), de Cox (1958). Bailey (1981) présente un exposé synthétique plus récent. En langue française, on pourra consulter les chapitres consacrés à ce thème dans les ouvrages de Dagnélie (1981 / 1998) et Tomassone *et al.* (1993).

[2] Cf. la section 5.1.1 du chapitre 5 pour la définition du codage disjonctif complet.

Cependant, pour chaque sous-tableau $\mathbf{X}_l$, la somme des colonnes vaut *1*. Il existe donc p relations linéaires entre les colonnes de $\mathbf{X}$. Le tableau $\mathbf{X}$ n'est pas de plein rang et la matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible.

Le problème peut être résolu par une régularisation de la régression (cf. § 3.3.6 du chapitre 3). Mais le fait que la nature des relations linéaires entre variables explicatives soit connue *a priori* (structure disjonctive complète du tableau) suggère d'autres possibilités de solutions.

Pour éliminer la multicolinéarité, on peut ne retenir que $m_l - 1$ modalités pour chaque variable $\mathbf{x}_l$ à m_l modalités. Une autre possibilité est également de supprimer une colonne de chaque sous-tableau mais après l'avoir retranchée aux colonnes restantes. Nous retiendrons ce deuxième codage mieux adapté au modèle linéaire avec interaction entre les variables explicatives.

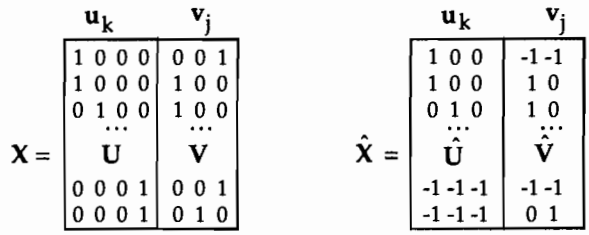


Figure 2.2- 4. Tableaux des variables explicatives initial et recodé

Le tableau des variables explicatives ainsi recodé $\hat{\mathbf{X}}$ est de plein rang :

$$rang\,(\hat{\mathbf{X}}) = \sum_{l=1}^{p} (m_l - 1)$$

Pour simplifier l'exposé, on se placera par la suite dans le cas où l'on dispose de deux variables nominales **u** et **v** ayant respectivement q et r modalités.

Notons $\mathbf{u}_k$ et $\mathbf{v}_j$, les indicatrices des variables **u** et **v** avec $1 < k < q$ et $1 < j < r$, $[\mathbf{U}, \mathbf{V}]$ le tableau disjonctif complet correspondant de dimension $(n\,,\, q + r)$ et $[\hat{\mathbf{U}}, \hat{\mathbf{V}}]$ le tableau disjonctif complet de plein rang et de dimension $(n\,,\, q + r - 2)$ obtenu après recodage. La généralisation se fera sans difficulté.

b – Modèle linéaire sans interaction

On cherche à déterminer s'il existe un effet dû à la variable **u** et un effet dû à la variable **v**, autrement dit, si **u** et **v** ont une influence sur **y**.

Les variables sont ici considérées sans interaction et l'on dispose d'un modèle linéaire où les effets sont par conséquent additifs :

$$y_{ikj} = \mu + \alpha_k + \beta_j + \varepsilon_{ikj}$$

avec $i = 1,\ldots, n$; $k = 1,\ldots, q - 1$ et $j = 1,\ldots, r - 1$. Ce modèle s'exprime sous forme matricielle par :

$$\mathbf{y} = \mu\mathbf{1}_n + (\alpha_1\mathbf{u}_1 .. + \alpha_k\mathbf{u}_k ... + \alpha_{q-1}\mathbf{u}_{q-1}) + (\beta_1\mathbf{v}_1 .. + \beta_j\mathbf{v}_j ... + \beta_{r-1}\mathbf{v}_{r-1}) + \boldsymbol{\varepsilon}$$

soit encore :

$$\mathbf{y} = \mu\mathbf{1}_n + \hat{\mathbf{U}}\boldsymbol{\alpha} + \hat{\mathbf{V}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où $\mathbf{1}_n$ est un vecteur de n composantes égales à 1 et μ un coefficient scalaire.

Rassemblons dans un tableau $\mathbf{L}$ de dimension $(n , q + r - 1)$ l'ensemble des variables explicatives artificielles et dans le vecteur $\boldsymbol{\delta}$ à $(q + r - 1)$ composantes les coefficients α_k, β_j et μ du modèle. Il prend la forme matricielle :

$$\mathbf{y} = \mathbf{L}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Le problème est de tester si les α_k (puis les β_j) sont égaux entre eux, l'hypothèse alternative étant que l'un au moins des coefficients dans chaque groupe diffère des autres[1].

On teste en d'autres termes les effets des variables $\mathbf{u}$ et $\mathbf{v}$.

On réalise alors le test de nullité simultanée des coefficients α_k, $(k = 1,\ldots, q - 1)$.

Pour cela, on effectue successivement deux ajustements pour calculer d'une part $S(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$ sur le modèle complet $\mathbf{y} = \mathbf{L}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$ et d'autre part $S(\mu, \boldsymbol{\beta})$ sur le modèle réduit obtenu en supprimant dans $\mathbf{L}$ les $q-1$ colonnes correspondant aux α_k. La statistique du test sera d'après [2.2 - 5] :

$$F = \frac{(S(\mu,\boldsymbol{\beta}) - S(\mu,\boldsymbol{\alpha},\boldsymbol{\beta}))/(q-1)}{S(\mu,\boldsymbol{\alpha},\boldsymbol{\beta})/(n-q-r+1)}$$

On rejettera l'hypothèse nulle d'absence d'effet de la variable $\mathbf{u}$ si la probabilité de dépasser la valeur F, pour une variable de Fisher à $(q - 1)$ et $(n - q - r + 1)$ degrés de liberté, est jugée trop petite. Pour tester l'existence d'un effet dû à la variable $\mathbf{v}$, on procédera de façon analogue.

c – Modèle linéaire avec interaction

Si l'on pense maintenant que l'effet de la modalité k de $\mathbf{u}$ peut être différent selon la modalité j de $\mathbf{v}$, il faut ajouter au modèle l'effet d'interaction entre les deux variables $\mathbf{u}$ et $\mathbf{v}$.

Cela peut se faire en juxtaposant au tableau disjonctif complet $[\hat{\mathbf{U}}, \hat{\mathbf{V}}]$ le sous-tableau $\hat{\mathbf{U}} \times \hat{\mathbf{V}}$ des interactions. On obtient $\hat{\mathbf{U}} \times \hat{\mathbf{V}}$ en faisant le produit terme à terme des colonnes $\mathbf{u}_k$ par les colonnes $\mathbf{v}_j$.

---

[1] La spécification du modèle est la même que lors de la régression multiple (résidus indépendants entre eux, de même variance). Pour procéder aux tests statistiques, il est nécessaire de supposer la normalité de la distribution des résidus.

Puisque $1 < k < q - 1$ et $1 < j < r - 1$, on engendre ainsi $(q-1)\times(r-1)$ colonnes contenant les produits de deux indicatrices correspond à la conjonction des présences d'effet. On vérifie que le nouveau tableau ainsi construit $[\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{U}} \times \hat{\mathbf{V}}]$ est bien de plein rang $q \times r$. Le modèle s'exprime alors par :

$$\mathbf{y} = \mu \mathbf{1}_n + \hat{\mathbf{U}}\alpha + \hat{\mathbf{V}}\beta + (\hat{\mathbf{U}} \times \hat{\mathbf{V}})\gamma + \varepsilon$$

où γ est un vecteur à $(q - 1) \times (r - 1)$ composantes.

Remarque :

La procédure développée dans le cas d'une interaction entre deux variables nominales peut être généralisée à des modèles comprenant plus de deux critères (**u**, **v**, **w**, ...), des interactions d'ordre *1* (**uv**, **uw**, **vw**,...), des interaction d'ordre *2* (**uvw**,...), etc. Cependant une certaine prudence s'impose pour plusieurs raisons. Tout d'abord, il est de plus en plus difficile d'apprécier et d'énoncer clairement la nature des hypothèses testées. D'autre part les interactions d'ordre élevé peuvent conduire à des tests "en chaîne" d'interprétation délicate (**uv** significatif, **vw** non significatif, **uvw** significatif, etc.). Enfin la procédure n'est pas robuste.

2.2.6 Régression sur variables mixtes : analyse de la covariance

Dans un modèle d'analyse de la variance, la valeur de la variable à expliquer est déterminée, à l'aléa ε près, par les classes dans lesquelles sont faites les mesures ou observations. On peut cependant imaginer un modèle où cette valeur est, à l'intérieur de chaque classe k, fonction également d'une ou plusieurs variables explicatives continues. On dira par exemple que la dépense individuelle en habillement est fonction du sexe **u** et pour chaque sexe fonction du revenu **x** de l'individu i.

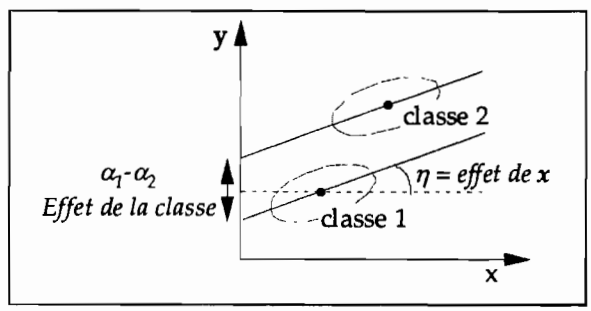


Figure 2.2 – 5. Un modèle d'analyse de la covariance : variable nominale sans effet sur la pente de la régression

La figure 2.2 - 5 illustre un modèle où l'observation i dans la classe k serait déterminée par :

$$y_{ik} = \mu + \alpha_k + \eta\, x_{ik} + \varepsilon_{ik}$$

En donnant la même pente η aux deux droites passant par les centres de classe, on suppose ici que le revenu a le même effet quel que soit le sexe; la distance $(\alpha_1 - \alpha_2)$ entre les deux droites mesure "l'effet du sexe". On aurait pu supposer un effet du revenu différencié suivant le sexe en traçant des droites non parallèles. De tels modèles, où interviennent des variables nominales et continues, sont appelés modèles d'*analyse de la covariance*. Ils sont de la forme :

$$\mathbf{y} = \mathbf{L}^*\delta + \varepsilon$$

où $\mathbf{L}^*$ est le tableau de plein rang des variables explicatives.

a – Modèles d'analyse de la covariance

Plaçons-nous, pour simplifier l'exposé, dans le cas où le modèle contient une variable nominale **u** à q modalités et une variable continue **x**.

Le modèle complet suppose à la fois un effet dû à la variable nominale **u** et un effet **x** différencié pour chaque catégorie k, $1 < k < q - 1$, ce qui s'exprime par :

$$y_{ik} = (\mu + \alpha_k) + (\eta + \beta_k)\, x_{ik} + \varepsilon_{ik} \qquad [2.2 - 6]$$

Le tableau **L** est construit en deux parties : les q premières colonnes correspondent à l'analyse de la variance à un critère; les $q - 1$ colonnes suivantes expriment de façon analogue l'effet différencié de **x** suivant la catégorie k de la variable **u**, mesuré autour de l'effet général représenté par la dernière colonne.

On remarquera que l'on obtient les q dernières colonnes comme une interaction entre la variable nominale **u** et la variable continue **x**, c'est-à-dire par multiplication terme à terme des q premières colonnes par **x**.

On notera $S(\mu, \alpha, \eta, \beta)$ la somme de carrés d'écarts des ajustements sur le modèle complet [2.2 - 6].

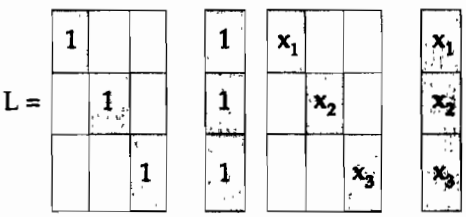


Figure 2.2 – 6· Tableau des variables explicatives : cas d'une variable nominale u à 3 modalités et d'une variable continue x

b – Test d'un effet différencié de x dans chaque classe k

Pour tester l'existence d'un effet différencié de **x** dans chaque classe k, on effectuera un deuxième ajustement sur le modèle :

$$y_{ik} = (\mu + \alpha_k) + \eta\, x_{ik} + \varepsilon_{ik}$$

Ce modèle est la réduction du modèle complet [2.2 - 6], obtenu par introduction de l'hypothèse nulle :

$$(H_0) \qquad \begin{cases} \beta_k = 0 & (k = 1,\ldots,q-1) \\ \mu, \eta, \alpha_k & \text{quelconques} \end{cases}$$

La statistique du test s'obtient par application de la formule [2.2 - 5] :

$$F = \frac{(S(\mu,\alpha,\eta) - S(\mu,\alpha,\eta,\beta))/(q-1)}{S(\mu,\alpha,\eta,\beta)/(n-2q)}$$

On rejettera l'hypothèse nulle si la probabilité de dépasser la valeur de F calculée, lue dans la table de Fisher-Snedecor à $(q-1)$ et $(n-2q)$ degrés de liberté, est jugée trop petite.

c – Test de l'effet de la variable u

Pour tester l'existence de l'effet de la variable nominale **u** (en supposant un effet différencié de **x** dans les classes), on calculera $S(\mu, \eta, \beta)$ sur le modèle :

$$y_{ik} = \mu + (\eta + \beta_k)\, x_{ik} + \varepsilon_{ik}$$

pour le comparer à $S(\mu, \alpha, \eta, \beta)$. Ce modèle est la réduction du modèle complet [2.2 - 6] obtenu par introduction de l'hypothèse nulle :

$$(H_0) \qquad \begin{cases} \alpha_k = 0 & (k = 1,\ldots,q-1) \\ \mu, \eta, \beta_k & \text{quelconques} \end{cases}$$

La statistique du test fait référence à la formule [2.2 - 5] pour laquelle les degrés de liberté sont $(q-1)$ et $(n-2q)$.

2.2.7 Choix des variables, généralisations du modèle

L'exposé qui précède ne fait que situer les principes de base du modèle linéaire par rapport aux méthodes descriptives de la première partie. Les méthodes présentées correspondent à une part notable des applications les plus courantes, mais à une part infime de la littérature théorique et technique sur le sujet, pour laquelle nous renvoyons le lecteur à la bibliographie citée au début du chapitre. On évoquera brièvement deux points dans ce paragraphe de conclusion : le problème de la sélection des variables dans les modèles et celui de la généralisation du modèle.

a – Sélection et choix des variables explicatives

La qualité de l'ajustement dépend également du choix des prédicteurs et il est souhaitable de retenir un nombre limité de variables, non redondantes et ayant un pouvoir prédictif.

Une technique souvent utilisée pour sélectionner les variables explicatives est la méthode pas-à-pas ou *stepwise*[1]. Elle consiste à effectuer une première régression simple sur une variable puis à ajouter successivement celles qui font augmenter le plus le coefficient de corrélation multiple R^2, avec éventuellement remise en question des choix antérieurs. A chaque étape sont réalisés des tests sur les coefficients de régression ou sur des sous-ensembles afin de rejeter la variable ou d'éliminer éventuellement certaines variables introduites dans les étapes précédentes. Les critères d'Akaike (1973), de Mallows (1973), sont fréquemment utilisés pour sélectionner les modèles lors de ces procédures. Une revue des critères usuels se trouve dans Atkinson (1981). L'exploration des résidus est très utilisée pour choisir ou compléter les variables du modèle, en général par des procédés graphiques (cf. Cook et Weisberg, 1982, 1994).

Les modèles graphiques (cf. par exemple : Whittaker, 1990 ; Wermuth et Cox, 1992 ; Fine, 1992) permettent, lorsque le nombre de variables explicatives n'est pas trop élevé, d'étudier les liaisons conditionnelles entre variables. Variables et liaisons sont représentées respectivement par les sommets et les arêtes de graphes de liaisons conditionnelles qui ont le mérite de conduire l'utilisateur à réfléchir sur la pertinence et les implications des modèles possibles.

Enfin on verra qu'une analyse en composantes principales de tout ou partie des variables explicatives $\mathbf{x}_k$, avec positionnement de la variable à expliquer **y** en élément supplémentaire, permet de positionner la ou les estimations **y** parmi les $\mathbf{x}_k$. Il est également possible de positionner différents changements de variables, voire de nouvelles variables fonctions de plusieurs prédicteurs, et donc de porter une appréciation critique sur les redondances et complémentarités au sein du modèle et de ses extensions.

b– Modèles linéaires généralisés

Ces modèles, présentés pour la première fois sous ce nom par Nelder et Wedderburn (1972), exposés de façon complète par McCullagh et Nelder (1989), généralisent le modèle linéaire de base sur deux points :

1- La combinaison linéaire notée $\omega_i = a_0 x_{i0} + a_1 x_{i1} + \ldots + a_p x_{ip}$ des variables explicatives n'est pas nécessairement l'espérance mathématique $E(y_i)$ de la variable y_i mais peut être plus généralement une fonction $g(.)$ de $E(y_i)$ (appelée *fonction lien*) et notée :

$$\omega_i = g[E(y_i)]$$

---

[1] La méthode de Furnival et Wilson (Furnival, 1971 ; Furnival and Wilson, 1974) permet de calculer les meilleures régressions pour *1, 2, ..., p* variables explicatives, par une exploration optimisée de toutes les possibilités. En pratique, *p* ne doit pas dépasser *40* pour que le volume de calcul reste raisonnable. Une telle procédure est recommandable car elle ne fait pas intervenir de critères externes (peu ou mal justifiés) pour inclure ou exclure des variables dans le modèle.

Pour le modèle linéaire classique :

$$\omega_i = E(y_i)$$

2- La loi des composantes de **y** appartient à la famille des lois exponentielles[1] (dont la loi normale est un cas particulier). Elle fait intervenir deux paramètres θ et φ, et trois fonctions $a(.)$, $b(.)$, et $c(.)$.

$$f_Y(y,\theta,\varphi) = e^{\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + C(y,\varphi)\right\}}$$

On voit que l'on obtient la fonction de densité de la loi normale :

$$f_Y(y;\theta,\varphi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{\frac{(y-\mu)^2}{2\sigma^2}\right\}}$$

pour les spécifications suivantes des paramètres et des fonctions :

$$\theta=\mu;\ \varphi=\sigma^2;\ a(\varphi)=\varphi;\ b(\theta)=\theta^2/2;\ c(y,\varphi)=-1/2\left\{(y^2/\sigma^2)+log(2\pi\sigma^2)\right\}$$

D'autres valeurs des paramètres et des fonctions conduisent aux lois binomiales, de Poisson, gamma. L'ajustement du modèle se fait par la méthode du maximum de vraisemblance[2], qui coïncide avec les moindres carrés dans le cas de la loi normale. En faisant varier la loi de **y** et la *fonction lien*, le modèle linéaire généralisé inclut comme cas particulier une famille de modèles mettant en jeu des variables nominales, parmi lesquels les modèles log-linéaires (cf. section 5.6 du chapitre 5).

---

[1] Cf. un exposé général dans : Dempster (1971) ; Berk (1972).

[2] La méthode numérique de résolution est une méthode des moindres carrés pondérés itératifs très voisine de la méthode de Newton-Raphson.

Chapitre 3

Analyse en Composantes Principales

L'analyse en composantes principales, qui est probablement la méthode d'analyse en axes principaux la plus utilisée actuellement, va être présentée tout au long des six sections de ce chapitre. Après un bref survol historique (section 3.1 : Histoire, domaine, principes), on présente les modalités de calcul dans les deux espaces des individus et des variables (section 3.2 : Individus et variables). Puis, dans la section 3.3 (Compléments et diversifications), on évoque les problèmes posés par les éléments supplémentaires, la possibilité de représentation simultanée des individus et des variables, les méthodes connexes, incluant l'analyse factorielle en facteurs communs et spécifiques. La section suivante est dévolue à l'interprétation des résultats et aux problèmes de validation, pour lesquels le bootstrap joue un rôle essentiel (section 3.4 : Interprétation et validation). La section 3.5 développe de façon assez détaillée deux exemples d'applications. Enfin, l'annexe technique (section 3.6) contient certains développements théoriques ou techniques du chapitre.

3.1 Histoire, domaine, principes

Conçue pour la première fois par Karl Pearson en 1901, intégrée à la statistique mathématique par Harold Hotelling en 1933, l'analyse en composantes principales n'est vraiment utilisée que depuis l'avènement et la diffusion des moyens de calculs actuels. L'analyse en composantes principales peut être présentée de divers points de vue. Pour le statisticien classique, il s'agit de la recherche des axes principaux de l'ellipsoïde indicateur d'une distribution normale multidimensionnelle, ces axes étant estimés à partir d'un échantillon. C'est la présentation initiale de Hotelling (1933), puis celle des manuels

classiques d'analyse multivariée, comme l'ouvrage fondamental d'Anderson (1958). Pour les factorialistes classiques, il s'agit d'un cas particulier de la méthode d'analyse factorielle des psychométriciens (cas de variances spécifiques nulles ou égales ; cf. Horst, 1965; Harman, 1967 ; et la présentation de cette méthode au § 3.3.5 de la section 3 de ce chapitre).

Enfin, du point de vue plus récent des analystes de données, il s'agit d'une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques et que l'on utilise en général sans référence à des hypothèses de nature statistique ni à un modèle particulier. Ce point de vue, fort répandu actuellement est peut-être le plus ancien. C'est celui qui avait été adopté par Pearson (1901). Bien entendu, il ne s'agissait pas de l'analyse en composantes principales telle que nous la présentons, mais les idées essentielles de la méthode étaient déjà entrevues par cet auteur. On trouvera une présentation plus proche de nos préoccupations dans l'article de synthèse de Rao (1964).

L'analyse en composantes principales présente de nombreuses variantes selon les transformations apportées au tableau de données : le nuage des points-individus peut être centré ou non, réduit ou non. Parmi ces variantes, l'analyse en composantes principales normée (nuage centré-réduit) est certainement la plus utilisée et c'est celle-ci que nous choisirons pour commencer.

3.1.1 Domaine d'application

L'utilisateur éventuel de l'analyse en composantes principales se trouve dans la situation suivante : il possède un tableau rectangulaire de mesures, dont les colonnes figurent des variables à valeurs numériques continues (des mensurations, des taux, etc.) et dont les lignes représentent les individus sur lesquels ces variables sont mesurées. En biométrie, il est fréquent de procéder à de nombreuses mensurations sur certains organes ou certains animaux. En micro-économie, on aura par exemple à relever les dépenses des ménages en divers postes. D'une manière générale, la condition que doivent remplir ces tableaux numériques pour être l'objet d'une description par l'analyse en composantes principales est la suivante : l'une au moins des dimensions du tableau (les lignes en général) est formée d'unités ayant un caractère répétitif, l'autre pouvant être éventuellement plus hétérogène.

Dans les exemples cités, les lignes ont ce caractère répétitif : on les désignera en général sous le nom d'individus ou d'observations, les colonnes étant désignées sous le nom de variables. Quelquefois, ces lignes pourront être considérées comme des réalisations indépendantes de vecteurs aléatoires.

Un exemple « fil d'Ariane » pour ce chapitre

Pour fixer les idées, nous considérons le tableau **R** des mesures prises sur quelques milliers d'hommes actifs concernant leurs temps d'activités

quotidiennes. On dispose de *16* variables décrivant des temps d'activités, en minutes par jour (sommeil, repos, repas chez soi, etc.). Les personnes enquêtées sont regroupées en *27* groupes selon l'âge, le niveau d'éducation et le type d'agglomération. Ce sont ces groupes qui sont observés et sont ici considérés comme des "individus" (cf. en section 3.5 le tableau 3.5 - 1).

Il s'agit de disposer d'un tableau de dimensions raisonnables dans le cadre d'un exposé pédagogique, et non pas d'un exemple ayant une portée méthodologique générale, une des attitudes de base en analyse descriptive des données étant au contraire "*de ne pas réduire a priori le champ de l'observable*".

Le tableau **R** aura en colonnes les *16* mesures caractérisant les *27* observations. Le terme général r_{ij} de ce tableau décrit la durée moyenne de l'activité j de l'observation i (constituant un groupe d'individus).

Nous voulons avoir une idée de la structure de l'ensemble des *16* activités, ainsi que des similitudes éventuelles de comportement entre les groupes d'individus.

Un autre exemple portant sur un tableau individuel de *notes sémiométriques* sera traité en section 5. Cet exemple sera repris au chapitre 6 (Classification).

3.1.2 Interprétations géométriques

Les représentations géométriques entre les lignes et entre les colonnes du tableau de données permettent de restituer visuellement les proximités entre les individus et entre les variables.

a – Pour les *n* individus

Dans $\mathcal{R}^p$, les *n(n-1)* distances attachées aux couples de points qui représentent des individus ont une interprétation directe pour l'utilisateur :

$$d^2(i,i') = \sum_{j=1}^{p} (r_{ij} - r_{i'j})^2 \qquad [3.1-1]$$

Il s'agit ici de la distance euclidienne classique. Deux points sont très voisins si, dans l'ensemble, leurs p coordonnées sont très proches. Les deux individus concernés sont alors caractérisés par des valeurs presque égales pour chaque variable.
Dans l'exemple évoqué ci-dessus, deux individus représentés par des points proches consacrent les mêmes temps aux mêmes activités.

b – Pour les *p* variables

Si les valeurs prises par deux variables particulières sont très voisines pour tous les individus, ces variables seront représentées par deux points très proches dans $\mathcal{R}^n$. Cela peut vouloir dire que ces variables mesurent une même chose ou encore qu'elles sont liées par une relation particulière. Toutefois la définition de ces proximités dans les deux espaces est assez fruste. Des

problèmes d'échelle de mesure se posent d'emblée : le temps consacré au sommeil est toujours beaucoup plus important que le temps passé à la lecture.

Par ailleurs, dans un cadre plus général, comment calculer la distance entre deux variables si l'une est exprimée en euros et l'autre en litres ? Comment interpréter un éloignement moyen dans $\mathcal{R}^p$? Est-ce que deux individus assez proches dans $\mathcal{R}^p$ ont des valeurs assez voisines pour chacune des variables, ou au contraire très proches pour certaines et éloignées pour d'autres ? L'analyse en composantes principales normée permet de donner des éléments de réponses à ces questions.

3.2 Individus et variables

3.2.1 Analyse du nuage des individus

Nous considérons tout d'abord ici le nuage des n individus non pondérés. Nous voulons, dans l'espace des variables, ajuster le nuage de n points par un sous-espace à une, puis deux dimensions, de façon à obtenir sur un graphique une représentation visuelle la plus fidèle possible des proximités existant entre les n individus vis-à-vis des p variables.

a – Principe d'ajustement

Ce n'est donc plus la somme des carrés des distances à l'origine en projection qu'il faut rendre maximum (cf. section [1.2] du chapitre 1), mais la somme des carrés des distances entre *tous les couples d'individus* :

$$\underset{(H)}{Max}\left\{\sum_{i}^{n}\sum_{i'}^{n} d_H^2(i,i')\right\}$$

Autrement dit, la droite d'ajustement H_1 ne doit pas être astreinte à passer par l'origine, comme H_0 dans l'analyse générale (figure 3.2 - 1).

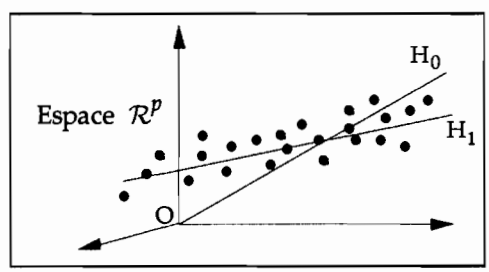


Figure 3.2 – 1. Droite d'ajustement du nuage de n points

Si h_i et $h_{i'}$ désignent les valeurs des projections de deux points-individus i et i' sur H_1, on a la relation classique :

$$\sum_{i,i'}^{n} d_H^2(i,i') = \sum_{i,i'}^{n} (h_i - h_{i'})^2 = n\sum_{i,i'}^{n} h_i^2 + n\sum_{i,i'}^{n} h_{i'}^2 - 2\sum_{i}^{n} h_i \sum_{i'}^{n} h_{i'}$$

$$= 2n^2\left(\frac{1}{n}\sum_{i}^{n} h_i^2 - \bar{h}^2\right) = 2n\sum_{i}^{n} (h_i - \bar{h})^2$$

où $\bar{h}$ désigne la moyenne des projections des n individus :

$$\bar{h} = \frac{1}{n}\sum_{i}^{n} h_i$$

et correspond à la projection sur H_1 du centre de gravité G du nuage dont la $j^{ème}$ coordonnée vaut :

$$\bar{r}_j = \frac{1}{n}\sum_{i}^{n} r_{ij}$$

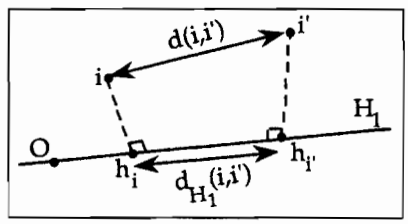


Figure 3.2 – 2. Projections sur H_1

Par conséquent, on a :

$$\sum_{i,i'}^{n} d^2(i,i') = 2n\sum_{i}^{n} d^2(i,G)$$

Rendre maximum la somme des carrés des distances entre tous les couples d'individus revient à maximiser la somme des carrés des distances entre les points et le centre de gravité du nuage G :

$$Max_{(H)}\left\{\sum_{i,i'}^{n} d_H^2(i,i')\right\}$$

est équivalent à :

$$Max_{(H)}\left\{\sum_{i}^{n} d_H^2(i,G)\right\}$$

Si l'origine est prise en G, la quantité à maximiser sera à nouveau la somme des carrés des distances à l'origine, ce qui correspond au problème de l'analyse générale dans $\mathcal{R}^p$.

Le sous-espace cherché résulte de l'*analyse générale* du tableau transformé **X**, de terme général :

$$x_{ij} = r_{ij} - \bar{r}_j$$

b – Distance entre individus

La distance entre deux individus i et i' est la distance euclidienne usuelle donnée par la formule [3.1 - 1].

Il peut exister des valeurs de j pour lesquelles les variables correspondantes sont d'échelles très diverses, (exemple : temps passé au sommeil, temps passé à la lecture) ; on veut que la distance entre deux points soit indépendante des unités sur les variables. On peut parfois désirer, surtout lorsque les unités de mesures ne sont pas les mêmes, faire jouer à chaque variable un rôle identique dans la définition des proximités entre individus : on parle alors d'*analyse en composantes principales normée.* Pour cela on corrige les échelles en adoptant la distance :

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{r_{ij} - r_{i'j}}{s_j \sqrt{n}} \right)^2$$

s_j désignant l'écart-type de la variable j dont le carré (variance) vaut :

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)^2$$

Finalement, nous retiendrons que l'analyse normée dans $\mathcal{R}^p$ du tableau brut **R** est l'analyse générale de **X**, de terme général :

$$x_{ij} = \frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}} \qquad [3.2 - 1]$$

Toutes les variables ainsi transformées sont "comparables" et ont même dispersion :

$$s^2(x_j) = 1$$

Les variables sont *centrées réduites.* On mesure l'écart à la moyenne en nombre d'écarts-types de la variable j.

c – Matrice à diagonaliser

En résumé, l'analyse du nuage des points-individus dans $\mathcal{R}^p$ nous a amené à effectuer une translation de l'origine au centre de gravité de ce nuage et à changer, dans le cas de l'analyse normée, les échelles sur les différents axes.

L'analyse du tableau transformé **X** nous conduit à diagonaliser la matrice **C** = **X'X**. Le terme général $c_{jj'}$ de cette matrice s'écrit :

$$c_{jj'} = \sum_{i}^{n} x_{ij} x_{ij'}$$

soit :

$$c_{jj'} = \frac{1}{n} \sum_{i}^{n} \frac{(r_{ij} - \overline{r}_j)(r_{ij'} - \overline{r}_{j'})}{s_j s_{j'}}$$

c'est-à-dire :

$$c_{jj'} = cor(j, j')$$

$c_{jj'}$ n'est autre que le coefficient de corrélation empirique entre les variables j et j' (d'où l'utilité du coefficient $\sqrt{n}$ au dénominateur de la relation [3.2 - 1]).

La matrice à diagonaliser est donc la *matrice de corrélations* **C**.

d – Axes factoriels

Les coordonnées des n points-individus sur l'axe factoriel $\mathbf{u}_\alpha$ normé ($\alpha^{\text{ième}}$ vecteur propre de la matrice **C** associé à la valeur propre λ_α) sont les n composantes du vecteur :

$$\psi_\alpha = \mathbf{X}\mathbf{u}_\alpha$$

Le facteur ψ_α est une combinaison linéaire des variables initiales.

Puisque le nuage des individus est centré sur le centre de gravité (les masses affectées aux individus étant égales à $1/n$), la moyenne du facteur est nulle :

$$\sum_i^n \psi_{\alpha i} = 0$$

et sa variance vaut :

$$var(\psi_\alpha) = \lambda_\alpha$$

La coordonnée du point-individu i sur cet axe s'écrit explicitement :

$$\psi_{\alpha i} = \sum_{j=1}^{p} u_{\alpha j} x_{ij} = \sum_{j=1}^{p} u_{\alpha j} \frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}}$$

3.2.2 Analyse du nuage des points-variables

L'analyse générale développée dans la section précédente nous a montré qu'en effectuant un ajustement dans un espace, on effectuait implicitement un ajustement dans l'autre espace. Nous avons volontairement choisi de commencer en travaillant dans $\mathcal{R}^p$. Dans cet espace, la transformation du tableau **R** initial selon la relation [3.2 - 1] avait deux objectifs :

- d'une part obtenir un ajustement qui respecte dans la mesure du possible les distances entre points-individus ;
- d'autre part, faire jouer des rôles similaires à toutes les variables dans la définition des distances entre individus.

Notons que la formule [3.2 - 1] ne fait pas intervenir de façon symétrique les lignes et les colonnes du tableau initial **R**. Que signifie, dans $\mathcal{R}^n$, la proximité entre deux points-variables j et j' si l'on prend comme coordonnées de ces variables les colonnes du tableau transformé **X** ?

a – distances entre points-variables

La distance entre variables découle de l'analyse dans $\mathcal{R}^p$.

Calculons la distance euclidienne usuelle entre deux variables j et j':

$$d^2(j, j') = \sum_{i=1}^{n} (x_{ij} - x_{ij'})^2$$

soit :

$$d^2(j, j') = \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ij'}^2 - 2\sum_{i=1}^{n} x_{ij} x_{ij'}$$

Remplaçant x_{ij} par sa valeur tirée de [3.2 - 1] et tenant compte du fait que :

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n} (r_{ij} - \bar{r}_j)^2$$

on obtient : $\sum_{i=1}^{n} x_{ij}^2 = \sum_{i=1}^{n} x_{ij'}^2 = 1$ et également : $\sum_{i=1}^{n} x_{ij} x_{ij'} = c_{jj'}$

D'où la relation liant la distance dans $\mathcal{R}^n$ entre deux points-variables j et j' et le coefficient de corrélation $c_{jj'}$ entre ces variables :

$$d^2(j, j') = 2\,(1 - c_{jj'}) \qquad [3.2 - 2]$$

ce qui implique :

$$0 \leq d^2(j, j') \leq 4$$

Dans l'espace $\mathcal{R}^n$, le cosinus de l'angle de deux vecteurs-variables est le coefficient de corrélation entre ces deux variables ($c_{jj'} = cos\ (j, j')$). Si ces deux variables sont à la distance 1 de l'origine (i.e. si elles sont de variance unité), le cosinus n'est autre que leur produit scalaire.

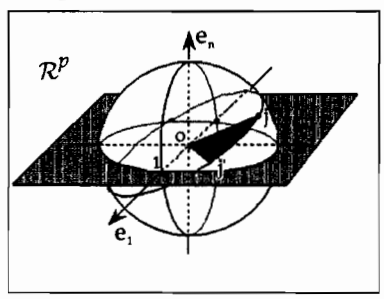


Figure 3.2 – 3. Système de proximités entre deux points-variables j et j'

Le système de proximités entre points-variables induit par la relation [3.2 - 2] est familier au statisticien :

- Deux variables centrées réduites fortement corrélées sont très proches l'une de l'autre ($c_{jj'} = 1$) ou au contraire les plus éloignées possible ($c_{jj'} = -1$) selon que la relation linéaire qui les lie est directe ou inverse :

- Deux variables orthogonales ($c_{jj'} = 0$) sont à distance moyenne et signifie qu'elles sont *linéairement* indépendantes.

Les proximités entre points-variables s'interprètent donc en termes de corrélations.

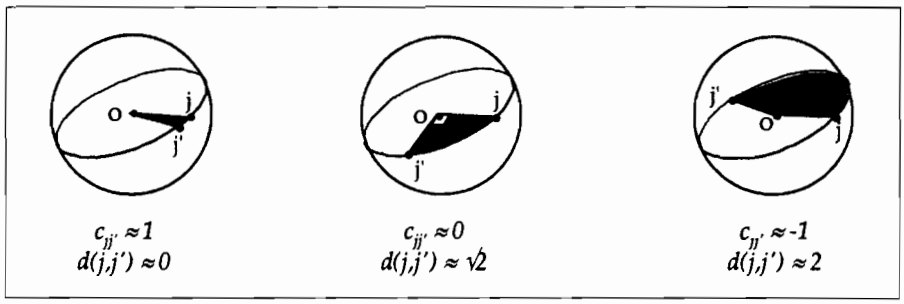


Figure 3.2 – 4. Corrélations et distances entre points-variables

b – Distance à l'origine

L'analyse dans $\mathcal{R}^n$ ne se fait pas par rapport au centre de gravité du nuage de points-variables, contrairement au cas du nuage des points-individus, mais par rapport à l'origine. La distance d'une variable j à l'origine O s'exprime par :

$$d^2(O,j) = \sum_{i=1}^{n} x_{ij}^2 = 1$$

Tous les points-variables sont sur une sphère de rayon *1* centrée à l'origine des axes, la *sphère des corrélations*. Les plans d'ajustement couperont la sphère suivant de grands cercles (de rayon *1*), les *cercles des corrélations*, à l'intérieur desquels se trouveront les points-variables.

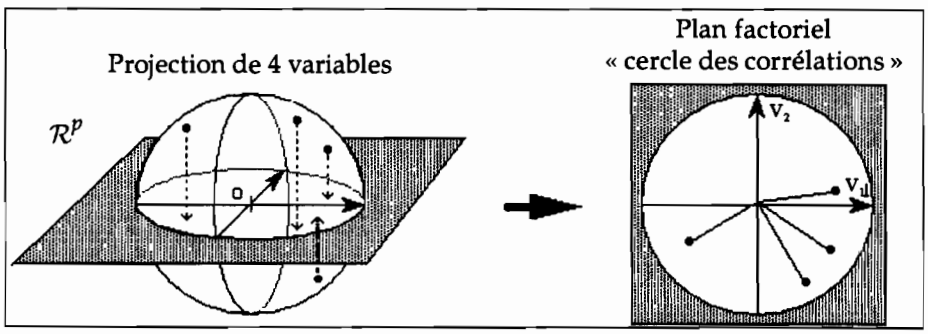


Figure 3.2 – 5. Représentation de la sphère et du cercle des corrélations

Remarque

La transformation analytique simple [3.2 - 1] a dans les espaces $\mathcal{R}^p$ et $\mathcal{R}^n$ des interprétations géométriques totalement différentes. Considérons par exemple l'opération de centrage des variables $r_{ij} \rightarrow (r_{ij} - \bar{r}_j)$:

- Dans $\mathcal{R}^p$, cette transformation équivaut à une translation de l'origine des axes au centre de gravité (ou point moyen) du nuage (cf. figure 3.2 - 6).
- Dans $\mathcal{R}^n$, cette transformation est une projection parallèlement à la première bissectrice des axes sur l'hyperplan qui lui est orthogonal[1] (cf. figure 3.2 - 7).

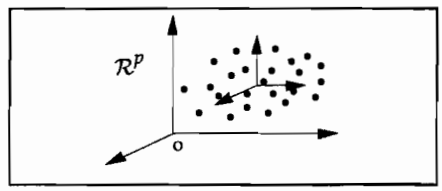


Figure 3.2 – 6. Transformation dans $\mathcal{R}^p$: simple translation

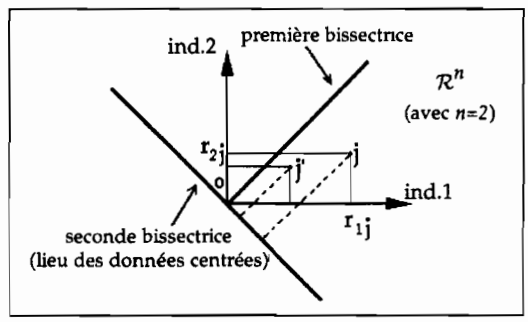


Figure 3.2 – 7. Transformation dans $\mathcal{R}^n$: projection parallèlement à la première bissectrice

c – Axes factoriels ou composantes principales

Nous avons vu dans l'analyse générale (chapitre 1) qu'il est inutile de procéder à la diagonalisation de la matrice **XX'** d'ordre (n,n) une fois connus les vecteurs propres $\mathbf{u}_\alpha$ et les valeurs propres λ_α de la matrice $\mathbf{C} = \mathbf{X'X}$ d'ordre (p,p).

Le vecteur $\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\mathbf{u}_\alpha$ est en effet un vecteur propre unitaire de **XX'**, relativement à la même valeur propre λ_α. Le $\alpha^{\text{ième}}$ facteur dans $\mathcal{R}^n$ s'écrit :

$$\varphi_\alpha = \mathbf{X'}\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X'X}\mathbf{u}_\alpha = \mathbf{u}_\alpha\sqrt{\lambda_\alpha}$$

comme $\psi_\alpha = \mathbf{X}\mathbf{u}_\alpha$, on a :

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}'\psi_\alpha$$

---

[1] La matrice **P** d'ordre (n,n) associée à cette transformation a pour terme général $p_{ii'} = \delta_{ii'} - \frac{1}{n}$ où $\delta_{ii'} = 1$ si i = i', et 0 sinon. **P** est idempotente : $\mathbf{P}^2 = \mathbf{P}$.

alors les coordonnées factorielles $\varphi_{\alpha j}$ des points-variables sur l'axe α sont les composantes de $\mathbf{X'v}_\alpha$ soit encore de $\mathbf{u}_\alpha\sqrt{\lambda_\alpha}$:

$$\varphi_{\alpha j} = \sum_{i=1}^{n} \left(\frac{r_{ij} - \bar{r}_j}{s_j\sqrt{n}}\right)\left(\frac{\psi_{\alpha i}}{\sqrt{\lambda_\alpha}}\right)$$

et l'on a :

$$\varphi_{\alpha j} = cor(j, \psi_\alpha) \qquad [3.2 - 3]$$

La coordonnée d'un point-variable sur un axe n'est autre que le *coefficient de corrélation* de cette variable avec le facteur ψ_α (combinaison linéaire des variables initiales) considéré lui-même comme variable artificielle dont les coordonnées sont constituées par les n projections des individus sur cet axe.

Les axes factoriels étant orthogonaux deux à deux, on obtient ainsi une série de variables artificielles non corrélées entre elles, appelées *composantes principales*, qui synthétisent les corrélations de l'ensemble des variables initiales.

Remarques

1) L'analyse en composantes principales ne traduit que des liaisons linéaires entre les variables. Un coefficient de corrélation faible entre deux variables signifie donc que celles-ci sont indépendantes linéairement alors qu'il peut exister une relation de degré supérieur à *1* (liaison non linéaire).

2) La coordonnée d'un point-variable sur l'axe α est nécessairement inférieure à *1* en valeur absolue :

$$\left|\varphi_{\alpha j}\right| \leq 1$$

et :

$$\sum_{\alpha=1}^{p} cor^2(j, \psi_\alpha) = 1$$

3) Le nuage de points-variables dans $\mathcal{R}^n$ n'est pas centré sur l'origine.

3.3 Compléments et variantes

L'analyse en composantes principale étant au carrefour de plusieurs méthodes et de plusieurs pratiques, on regroupera dans cette section quelques thèmes qui se rattachent aussi bien à son environnement théorique qu'à son utilisation concrète. Le paragraphe 3.3.1 sera ainsi consacré à la méthodologie des éléments supplémentaires, déjà évoquée à propos de l'analyse générale. On décrira ensuite un mode de représentation simultanée des variables et des individus (cf. § 3.3.2) et l'on exposera les précautions à prendre lorsque l'on procède à une analyse en composantes principales non normée (cf. § 3.3.3). On traitera ensuite des variantes non-paramétriques de l'analyse en composantes

principales (cf. § 3.3.4). Au paragraphe 3.3.5, on posera les bases de l'analyse factorielle en facteurs communs et spécifique, utilisée depuis le début du 20ème siècle par les psychométriciens, qui est étroitement apparentée à l'analyse en composantes principales. Enfin le dernier paragraphe évoquera certaines méthodes dérivées.

3.3.1 Individus et variables supplémentaires

On dispose d'informations complémentaires que l'on veut rapporter à l'analyse des temps d'activités des hommes actifs regroupés en catégories. Par exemple, on veut enrichir cette analyse par une série d'indicateurs d'habitudes de fréquentation-média, constituant des variables continues et par le niveau d'éducation et l'âge qui sont des variables nominales. On désire également positionner des groupes d'individus (lignes supplémentaires).

Le tableau de données **R** peut être ainsi complété en colonnes par un tableau à n lignes et p_s colonnes $\mathbf{R}^+$ et en lignes par un tableau $\mathbf{R}_+$ à n_s lignes et p colonnes. Il n'est pas nécessaire de connaître le tableau $\mathbf{R}_+^+$ à n_s lignes et p_s colonnes croisant individus et variables supplémentaires (cf. figure 3.3 - 1).

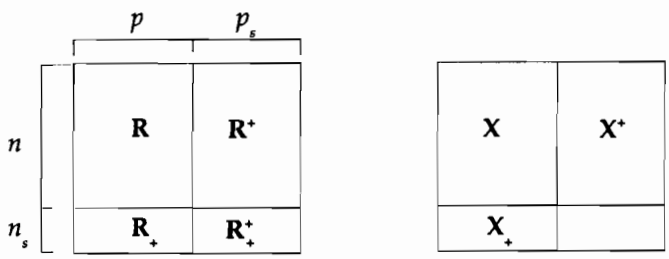


Figure 3.3 – 1. Lignes et colonnes supplémentaires

Les tableaux $\mathbf{R}^+$ et $\mathbf{R}_+$ vont être respectivement transformés en tableaux $\mathbf{X}^+$ et $\mathbf{X}_+$ de façon à rendre ces nouvelles lignes et colonnes comparables à celles de **X**. Dans l'espace $\mathcal{R}^n$ les p_s variables supplémentaires peuvent être continues ou nominales.

a – Individus supplémentaires

Pour situer les individus supplémentaires par rapport aux autres dans l'espace $\mathcal{R}^p$ il est nécessaire de les positionner par rapport au centre de gravité du nuage (déjà calculé sur les n individus), de diviser les coordonnées par les écarts-types des variables (déjà calculés sur les n individus), et de faire intervenir le coefficient $\sqrt{n}$. D'où la transformation :

$$x_{+ij} = \frac{r_{+ij} - \overline{r}_j}{s_j \sqrt{n}}$$

Les coordonnées des nouveaux points-individus sont donc les n_s lignes du vecteur $\mathbf{X}_+\mathbf{u}_\alpha$.

En appelant $\mathbf{X}_s$ le tableau $\begin{bmatrix}\mathbf{X}\\ \mathbf{X}_+\end{bmatrix}$ on obtient simultanément les $n+n_s$ coordonnées des individus analysés et supplémentaires en effectuant le produit $\mathbf{X}_s\mathbf{u}_\alpha$.

b – Variables continues supplémentaires

Dans $\mathcal{R}^n$, pour que les distances entre variables s'interprètent encore en termes de corrélations, ces variables doivent être à valeurs numériques continues et il est indispensable d'effectuer la transformation :

$$x_{ij}^{+} = \frac{r_{ij}^{+} - \overline{r}_j^{+}}{s_j^{+}\sqrt{n}}$$

On calcule donc les nouvelles moyennes et les nouveaux écarts-types correspondant aux variables supplémentaires, pour positionner celles-ci sur la sphère de rayon unité. Les coordonnées des p_s variables supplémentaires sur cet axe sont donc les p_s lignes du vecteur $\mathbf{X}^{+\prime}\mathbf{v}_\alpha$ et correspondent chacune au coefficient de corrélation entre la variable et le facteur (le facteur est la variable artificielle « coordonnée sur l'axe factoriel »).

c – Variables nominales supplémentaires

Si la variable à mettre en supplémentaire est nominale, on ne peut plus effectuer la même transformation.

Dans ce cas, on ramène la variable nominale ayant m modalités, à m *groupes d'individus* définis par les modalités 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 $\mathcal{R}^p$.

Supposons, par exemple, que l'on mesure la taille et le poids de *10* individus et que l'on désire mettre en supplémentaire la variable sexe. Nous disposons du tableau de mesures représenté par la figure 3.3 - 2.

On calcule alors la taille et le poids moyens des hommes (*177; 75*) et celui des femmes (*167; 59*). Ce sont ces points moyens qui vont être positionnés parmi les points-individus.

L'analyse d'une variable nominale supplémentaire ne se fait donc pas dans $\mathcal{R}^n$ mais dans $\mathcal{R}^p$. La figure 3.3 - 3 schématise le positionnement des variables supplémentaires.

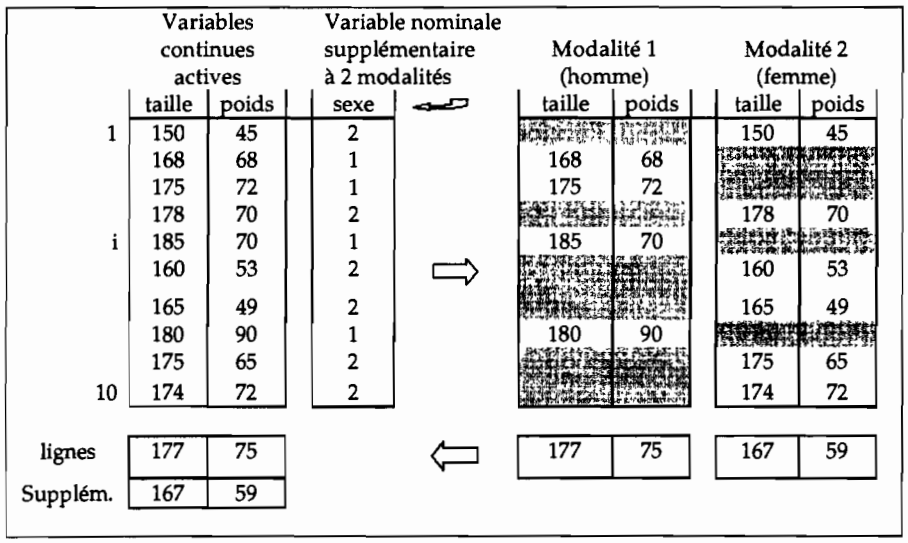


Figure 3.3 – 2. Les modalités de la variable nominale supplémentaire sont des individus supplémentaires

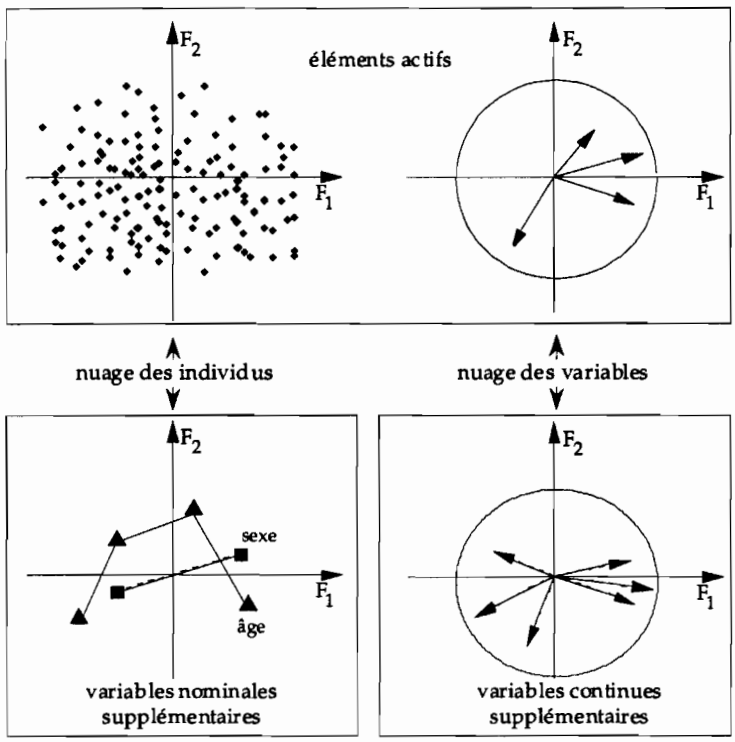


Figure 3.3 – 3. Représentation des variables supplémentaires

3.3.2 Représentation simultanée

L'analyse du nuage des variables est déduite de celle du nuage des individus : la représentation des variables sur les axes factoriels dans $\mathcal{R}^n$ aide l'interprétation des axes factoriels dans $\mathcal{R}^p$ et réciproquement.

a – Représentation séparée des deux nuages

Mais les deux nuages ne sont pas dans le même repère, ce qui rend impossible la représentation simultanée des individus et des variables. Les proximités entre individus s'interprètent en termes de similitudes de comportement vis-à-vis des variables et les proximités entre variables en termes de corrélations. Il faut bien se garder d'interpréter la distance séparant un point-variable d'un point-individu car ces deux points ne font pas partie d'un même nuage dans un même espace : la superposition de ces deux plans factoriels est dénuée de sens.

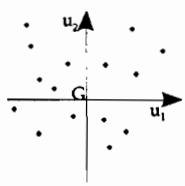
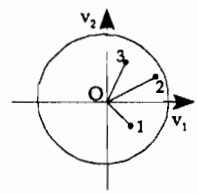
Dans l'espace $\mathcal{R}^p$	Dans l'espace $\mathcal{R}^n$
L'analyse du nuage des n points-individus se fait dans le repère : $\{G, u_1, \ldots, u_\alpha, \ldots, u_p\}$  La représentation des individus sur les axes factoriels fournit la meilleure visualisation approchée des distances entre les individus.	L'analyse du nuage des p points-variables se fait dans le repère : $\{O, v_1, \ldots, v_\alpha, \ldots, v_n\}$  La représentation des variables sur les axes factoriels fournit une synthèse graphique de la matrice de corrélations.

Figure 3.3 – 4.
Nuage des individus dans $\mathcal{R}^p$

Figure 3.3 – 5.
Nuage des variables dans $\mathcal{R}^n$

b – Justification d'une autre représentation simultanée

Cependant si l'on considère non plus des points-variables mais des directions de variables dans $\mathcal{R}^p$, on peut alors envisager de représenter simultanément, dans cet espace, à la fois les points-individus et des vecteurs représentant les variables. Dans l'espace $\mathcal{R}^p$ des n points-individus, après transformation du tableau de données, on dispose de deux systèmes d'axes :

- les anciens axes unitaires $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_j, \ldots \mathbf{e}_p)$ correspondant aux p variables avant l'analyse où :

$$\mathbf{e}_j = (0, 0, \ldots, 1, 0, \ldots, 0)$$

$\{\mathbf{e}_j, (j = 1, ..., p)\}$ est le système d'axes de référence pour les coordonnées initiales des individus.

- les nouveaux axes unitaires $\{\mathbf{u}_\alpha, (\alpha = 1, ..., p)\}$ constitués des axes factoriels. La possibilité d'une représentation simultanée réside alors dans la projection (en ligne supplémentaire) de l'ancien axe $\mathbf{e}_j$ sur le nouvel axe $\mathbf{u}_\alpha$.

La coordonnée de la projection de $\mathbf{e}_j$ sur $\mathbf{u}_\alpha$ vaut :

$$\mathbf{e}_j' \mathbf{u}_\alpha = u_{\alpha j}$$

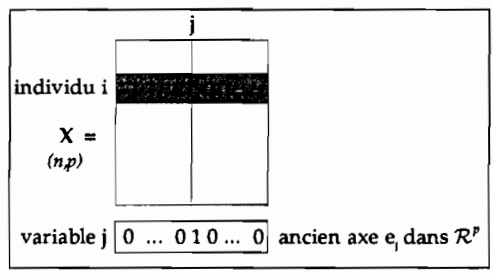


Figure 3.3 – 6. Ancien axe dans $\mathcal{R}^p$ en supplémentaire
La variable j est un individu particulier

Il est ainsi possible de représenter dans $\mathcal{R}^p$ les directions données par les variables d'origine sur le plan factoriel du nuage des individus ; ces directions peuvent être matérialisées par des vecteurs unitaires. Ces vecteurs constituent le repère d'origine dans lequel on a construit le nuage des individus. Ils sont donc orthogonaux deux à deux [1]. Ce qui s'appellera *représentation simultanée* est donc la projection du repère orthonormé des axes d'origine sur le plan factoriel du nuage des individus.

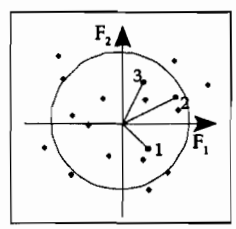


Figure 3.3 – 7. Projection de 3 anciens axes (3 variables initiales) sur le plan factoriel (F_1, F_2) du nuage des individus

Rappelons que, dans $\mathcal{R}^n$, la coordonnée de la variable j sur l'axe α est égale au coefficient de corrélation (cf. formule [3.2 - 4]) entre la variable et le facteur et vaut :

$$\varphi_{\alpha j} = \sqrt{\lambda_\alpha} u_{\alpha j}$$

---

[1] Il apparaît donc clairement que cette représentation des variables est distincte du nuage de variables décrit précédemment.

Les deux nuages des variables ne coïncident donc pas. Ils diffèrent l'un de l'autre par une dilatation définie sur chaque axe par le coefficient $\sqrt{\lambda_\alpha}$. Dans le cas de la représentation simultanée, qui est en fait une représentation dans $\mathcal{R}^p$, on n'interprète pas la distance entre deux variables en terme de corrélation, puiqu'il s'agit en réalité des extrémités de deux vecteurs unitaires orthogonaux[1]. L'interprétation de la distance entre deux variables (en terme de corrélation) ne peut se faire[2] que dans $\mathcal{R}^n$. En tenant compte de ces considérations, il est licite de comparer, sur la représentation simultanée, les positions respectives de deux individus vis-à-vis de l'ensemble des variables, ou de deux variables vis-à-vis de l'ensemble des individus. On dispose ainsi d'une perspective déformée du système d'axes originel tenant compte des liaisons existant entre les variables initiales. La direction d'une variable définit des zones pour les individus : d'un côté, ceux qui prennent des fortes valeurs pour cette variable et, à l'opposé, ceux qui prennent des valeurs faibles.

Remarques:

1) Si l'échelle des coordonnées des points-variables a une interprétation en termes de corrélations, il n'en est pas de même pour les points-individus. On appliquera à leurs coordonnées un coefficient de dilatation convenable. La valeur $\sqrt{n/p}$ assure souvent un positionnement dans le plan compatible avec la répartition des points-variables et permet ainsi une représentation équilibrée des deux nuages.

2) Dans la représentation simultanée, il ne peut y avoir de variables continues supplémentaires (elles ne constituent pas des axes d'origine pour le positionnement des individus). Il peut y avoir des variables nominales supplémentaires car ce sont des individus supplémentaires.

3.3.3 Analyse en composantes principales non normée

L'analyse en composantes principales non normée revient à considérer le nuage de points centré et non réduit. On généralisera cependant l'analyse en faisant jouer maintenant à chaque point-individu un rôle proportionnel à sa masse (ce que l'on aurait évidemment pu faire à propos de l'analyse normée).

a – Principe de l'analyse et nuage des individus

Plaçons-nous dans l'espace $\mathcal{R}^p$ et considérons le nuage des points-individus pesants, centré sur le centre de gravité G.

---

[1] Toutes ces distances sont égales à $\sqrt{2}$ dans l'espace complet.

[2] On note toutefois que le nuage projeté des extrémités des vecteurs unitaires dans $\mathcal{R}^p$ et le nuage des extrémités des vecteurs variables dans $\mathcal{R}^n$ ont généralement des allures voisines, surtout si les valeurs propres sont presque égales, car alors la dilatation est peu déformante.

L'analyse en composantes principales revient à effectuer une analyse générale de points pondérés avec comme origine le centre de gravité du nuage.

Le tableau de données initiales **R** subit plusieurs transformations : on construit le tableau **X** de données centrées et chaque individu i est affecté d'une masse (ou d'un poids[1]) p_i. Ces masses constituent les éléments diagonaux de la matrice diagonale **N**. Le tableau **Z** soumis à l'analyse en composantes principales non normée est par conséquent de la forme :

$$\mathbf{Z} = \mathbf{N}^{1/2}\mathbf{X}$$

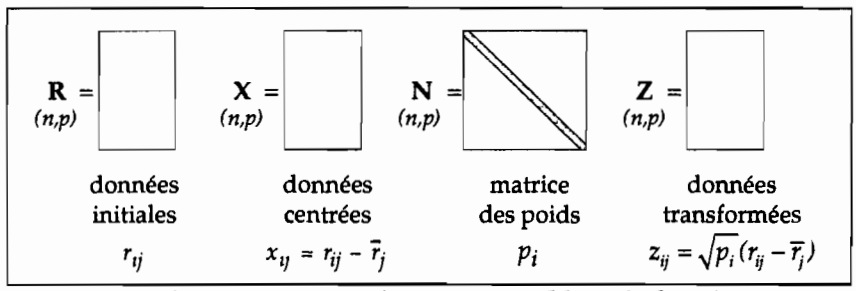


Figure 3.3 – 8. Transformation du tableau de données en analyse en composantes principales non normée

La matrice à diagonaliser est la matrice d'inertie autour du centre de gravité du nuage G :

$$\mathbf{A} = \mathbf{Z}'\mathbf{Z} = \mathbf{X}'\mathbf{N}\mathbf{X}$$

de terme général :

$$a_{jj'} = \sum_{i=1}^{n} p_i (r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})$$

Si les masses représentent des fréquences, alors la matrice à diagonaliser est la *matrice des covariances*A partir de là, on détermine les axes factoriels $\mathbf{u}_\alpha$ tels que $\mathbf{u}'_\alpha\mathbf{u}_\alpha = 1$. Les coordonnées factorielles sur ces axes sont données par :

$$\psi_\alpha = \mathbf{X}\mathbf{u}_\alpha$$

dont les composantes s'écrivent :

$$\psi_{\alpha i} = \sum_{j=1}^{p} (r_{ij} - \bar{r}_j) u_{\alpha j}$$

avec :

$$\sum_{i=1}^{n} p_i \psi_{i\alpha}^2 = \lambda_\alpha$$

b – Nuage des variables

L'analyse du nuage des p variables dans $\mathcal{R}^n$ revient à faire l'analyse générale du tableau **Z** :

---

[1] Les termes de masse et de poids sont utilisés indifféremment en statistique. Ils désignent souvent des fréquences relatives ou des probabilités *a priori*.

$$z_{ij} = \sqrt{p_i}(r_{ij} - \bar{r}_j)$$

avec :

$$\sum_{i=1}^{n} p_i = 1 \quad \text{et} \quad \bar{r}_j = \sum_{i=1}^{n} p_i r_{ij}$$

La distance induite entre deux variables s'exprime par :

$$d^2(j,j') = \sum_{i=1}^{n} (z_{ij} - z_{ij'})^2$$

soit :

$$d^2(j,j') = \sum_{i=1}^{n} z_{ij}^2 + \sum_{i=1}^{n} z_{ij'}^2 - 2\sum_{i=1}^{n} z_{ij} z_{ij'}$$

Par conséquent[1] :

$$d^2(j,j') = var(j) + var(j') - 2cov(j,j') \qquad [3.3 - 1]$$

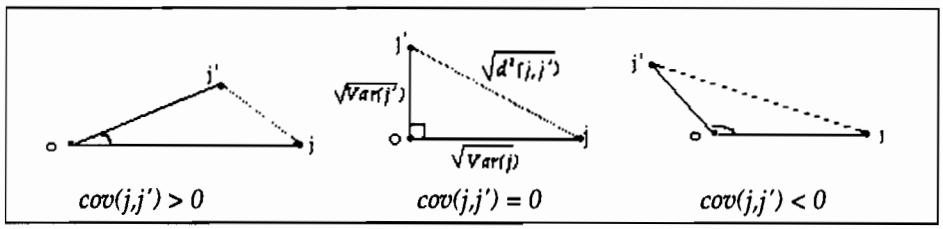


Figure 3.3 – 9. Distance entre deux variables

La distance entre deux variables s'exprime en terme de covariance et augmente avec les variances. Elle diminue si la liaison est positive et augmente si la liaison est négative. La distance d'une variable à l'origine des axes est sa variance :

$$d^2(O,j) = var(j) = \sum_{i=1}^{n} z_{ij}^2 = \sum_{i=1}^{n} p_i (r_{ij} - \bar{r}_j)^2$$

Par conséquent, pour l'analyse en composantes principales non normée, la sphère de corrélations n'est plus l'espace de départ[2].

3.3.4 Analyses non-paramétriques

Ces méthodes ne diffèrent de la précédente que par une transformation préliminaire des données. Elles sont recommandées lorsque les données de base sont hétérogènes. Elles donnent des résultats d'une grande robustesse, se prêtant par ailleurs à des interprétations simples en termes statistiques.

---

[1] La formule [3.2 - 2] est un cas particulier lorsque *var(j)=var(j')=1*, c'est-à-dire lorsqu'il s'agit d'une analyse en composantes principales normée.

[2] Dans une représentation simultanée, les anciens axes (distance *1* de l'origine) seront toujours dans un cercle de corrélations (cf. § 1.2.2).

a – Analyse des rangs

Le tableau initial des données est transformé en tableau de rangs. L'observation i de la variable j consiste alors en un classement q_{ij} : c'est le rang de l'observation i lorsque les n observations sont classées par ordre de grandeur (avec une convention *ad hoc* pour le classement des ex-aequos). Dans ces conditions, la distance entre deux variables j et j' est définie par la formule[1] :

$$d^2(j,j') = \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (q_{ij} - q_{ij'})^2$$

L'utilisation des rangs sera justifiée dans les contextes suivants :

- Les données de base peuvent être elles-mêmes des classements, auquel cas ce type d'analyse s'impose.
- Les échelles de mesure des variables peuvent être si différentes que l'opération de réduction pratiquée par l'analyse en composantes principales normée reste insuffisante. De plus cette opération ne remédie pas par exemple à la dissymétrie des distributions. Il paraît enfin plus justifié de synthétiser une famille de classements qu'un ensemble très hétérogène de mesures.
- Les hypothèses *a priori* faites implicitement sur les mesures sont plus faibles et par conséquent moins arbitraires : la loi des distances est maintenant non-paramétrique; nous disposerons donc de seuils de confiance qui ne dépendront que de l'hypothèse de continuité des lois des observations, plus plausible que celle de normalité.
- Enfin, les représentations fournies sont robustes, peu sensibles à l'existence de valeurs aberrantes, ce qui sera souvent une qualité appréciable.

Les règles d'interprétation se déduisent de celles de l'analyse en composantes principales puisque c'est cette analyse que l'on effectue après l'opération de transformation en rangs. La proximité entre deux variables s'interprète en terme de corrélation de rangs : deux variables seront très proches pour des classements voisins des observations ; au contraire, deux variables éloignées correspondront à des classements pratiquement inverses. Deux observations seront proches si elles ont des rangs similaires pour chacune des variables. Enfin, dans la représentation simultanée, on a une idée de l'ensemble du classement des observations pour une variable en examinant les positions respectives de cette variable et de l'ensemble des observations [2].

---

[1] On reconnaît dans cette formule le complément à *1* du coefficient de corrélation des rangs de Spearman (cf. Kendall, 1962).

[2] Ajoutons enfin que le caractère non-paramétrique de la représentation obtenue permet de procéder à des tests de validité sur les valeurs propres. La loi des valeurs propres issues de l'analyse d'un tableau de rangs ne dépend en effet que des paramètres n et p, nombres de lignes et de colonnes du tableau. Il est donc possible de procéder à une tabulation permettant de connaître les seuils de signification des valeurs propres .

b – Analyse en composantes robustes

Le critère d'ajustement des moindres-carrés est particulièrement bien adapté à la distribution normale. Dans le cas d'une distribution uniforme (cas de l'analyse des rangs), il tend à donner une importance excessive aux observations extrêmes. On rendra donc plus robuste l'analyse par une transformation qui "normalise" la distribution uniforme des rangs.

Considérons la $k^{ième}$ observation de n observations rangées et soit F la fonction de répartition de la loi Normale. On remplacera l'observation de rang k par la valeur y_k tirée de la *fonction de répartition inverse* de la loi Normale[1] :

$$y_k = F^{-1}(\frac{k}{n+1})$$

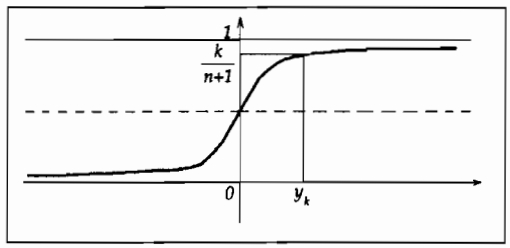


Figure 3.3 – 10. Transformation selon la fonction de répartition inverse de la loi Normale

Pour n grand, la transformation est équivalente au remplacement de la $k^{ième}$ observation par l'espérance de la $k^{ième}$ observation dans un échantillon rangé de n valeurs normales.

3.3.5 L'analyse factorielle en facteurs communs et spécifiques

L'analyse factorielle en facteurs communs et spécifiques (*factor analysis*) est un modèle très ancien [2]. Bien qu'il s'agisse d'un modèle statistique particulier, et non d'une méthode exploratoire, ses liens profonds avec l'analyse en composantes principales nous incitent à le présenter dans ce chapitre. Ce modèle est utilisé principalement par les psychologues et psychométriciens. Les développements auxquels il donne lieu sont complexes et diversifiés. On pourra consulter sur ce point les ouvrages de Harman (1967), Mulaik (1972).

---

[1] On trouve déjà ce type de transformation dans Fisher et Yates (1949).

[2] A l'origine des principes de la méthode se trouvent Spearman (1904) (analyse monofactorielle), puis Garnett (1919) et Thurstone (1947) (analyse multifactorielle).

a – Le modèle

Cette méthode se propose de reconstituer, à partir d'un petit nombre q de *facteurs*, les corrélations existant entre p variables observées. On suppose l'existence d'un modèle *a priori* :

$$\underset{(p,1)}{\mathbf{x}_i} = \underset{(p,q)}{\Gamma} \underset{(q,1)}{\mathbf{f}_i} + \underset{(p,1)}{\mathbf{e}_i} \qquad [3.3 - 2]$$

Dans cette écriture $\mathbf{x}_i$ représente le $i^{ième}$ vecteur observé des p variables; Γ est un tableau (p, q) de coefficients inconnus (avec $q < p$); $\mathbf{f}_i$ est la $i^{ième}$ valeur du vecteur aléatoire et non observable de q *facteurs communs;* et $\mathbf{e}_i$ la $i^{ième}$ valeur du vecteur non observable de résidus, lesquels représentent l'effet combiné de *facteurs spécifiques* et d'une perturbation aléatoire.

Ainsi par exemple, dans le cas des facteurs communs "f_1 = intelligence" et "f_2 = mémoire" que cherchaient les psychologues, le système [3.3 - 2] s'écrit pour le $i^{ième}$ individu :

$$\begin{cases} x_{i1} = \gamma_{11} f_{i1} + \gamma_{12} f_{i2} + e_{i1} \\ x_{i2} = \gamma_{21} f_{i1} + \gamma_{22} f_{i2} + e_{i2} \\ \qquad\qquad \cdots \\ x_{ip} = \gamma_{p1} f_{i1} + \gamma_{p2} f_{i2} + e_{ip} \end{cases}$$

Chaque observation de chaque variable est considérée comme une réalisation d'une variable aléatoire déterminée, par addition au résidu aléatoire spécifique, des deux variables aléatoires que sont les facteurs communs (avec des pondérations qui dépendent de chaque variable) [1].

Désignons par $\mathbf{X}$ le tableau (n,p) dont la $i^{ième}$ ligne est le vecteur transposé $\mathbf{x}_i'$ qui représente l'observation i. De même $\mathbf{F}$ désigne le tableau (n,q) non observable dont la $i^{ième}$ ligne est $\mathbf{f}_i'$; et $\mathbf{E}$ le tableau (n,p) non observable dont la $i^{ième}$ ligne est $\mathbf{e}_i'$. Le modèle liant l'ensemble des observations aux facteurs hypothétiques s'écrit :

$$\underset{(n,p)}{\mathbf{X}} = \underset{(n,q)}{\mathbf{F}} \underset{(q,p)}{\Gamma'} + \underset{(n,p)}{\mathbf{E}} \qquad [3.3 - 3]$$

Dans cette écriture, seul $\mathbf{X}$ est observable, et le modèle est par conséquent indéterminé. Son identification et l'estimation des paramètres posent des problèmes complexes, sources d'une abondante littérature [2]. Une cascade d'hypothèses *a priori* supplémentaires va permettre d'écrire le problème sous une forme simplifiée, la seule que nous aborderons ici.

---

[1] Ainsi, on reconstitue approximativement les p notes d'un individu i dans p matières à partir de ses deux notes factorielles et de coefficients qui ne dépendent que des matières.

[2] Voir par exemple la synthèse et les références très complètes de Fine (1993). Il existe de nombreuses variantes de la méthode : axes obliques, rotations selon différents critères (varimax, quartimax, oblimax), recherches de structures simples, pour lesquelles on peut citer globalement l'ensemble des parutions de la revue *Psychometrika*. Cf. aussi la section 3.3.6.

Sans perte de généralité, nous supposerons *centrées* les variables dont les observations sont les colonnes de **X**, ainsi que les variables aléatoires que constituent les facteurs communs et les facteurs spécifiques. Nous utiliserons les notations suivantes:

- **W** matrice (p,p) des covariances théoriques entre variables;
- Φ matrice (q,q) des covariances théoriques entre facteurs communs;
- Δ matrice (p,p) des covariances théoriques entre facteurs spécifiques.

Appelons **S** la matrice des covariances empiriques des observations **X**, que nous supposerons également centrées. Par définition et en vertu de [3.3 - 3], on a :

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}(\mathbf{F}\Gamma' + \mathbf{E})'(\mathbf{F}\Gamma' + \mathbf{E})$$

c'est-à-dire :

$$\mathbf{S} = \frac{1}{n}\Gamma\mathbf{F}'\mathbf{F}\Gamma' + \frac{1}{n}\Gamma\mathbf{F}'\mathbf{E} + \frac{1}{n}\mathbf{E}'\mathbf{F}\Gamma' + \frac{1}{n}\mathbf{E}'\mathbf{E} \qquad [3.3 - 4]$$

Aux hypothèses du modèle, nous ajouterons l'hypothèse *a priori* que les facteurs résiduels sont non corrélés aux facteurs communs; la matrice des covariances théoriques correspondantes étant nulle, nous considérerons comme négligeables dans [3.3 - 4] les matrices $\frac{1}{n}\Gamma\mathbf{F}'\mathbf{E}$ et $\frac{1}{n}\mathbf{E}'\mathbf{F}\Gamma'$ dont les espérances doivent être nulles. La relation [3.3 - 4] prend la forme simplifiée :

$$\mathbf{S} = \frac{1}{n}\Gamma\mathbf{F}'\mathbf{F}\Gamma' + \frac{1}{n}\mathbf{E}'\mathbf{E}$$

correspondant à la relation théorique suivante pour le modèle :

$$\mathbf{W} = \Gamma\Phi\Gamma' + \Delta \qquad [3.3 - 5]$$

Le problème d'estimation consiste à ajuster dans [3.3 - 5] une matrice $\tilde{\mathbf{W}}$ qui, au regard d'un critère choisi par ailleurs, soit proche de la matrice des covariances empiriques **S**. Mais afin d'obtenir une solution unique pour les paramètres de Γ, Φ et Δ, il est nécessaire d'introduire des contraintes supplémentaires dans le modèle.

On suppose en général que les facteurs spécifiques sont non corrélés, c'est-à-dire que la matrice Δ est diagonale. On impose de plus généralement que les facteurs communs soient orthogonaux et de variance unité, autrement dit la matrice Φ est la matrice identité **I** d'ordre q. La relation [3.3 - 5] du modèle s'écrit alors :

$$\mathbf{W} = \Gamma\Gamma' + \Delta$$

Sur cette relation le lien avec l'analyse en composantes principales apparaît clairement. Il s'agit dans ce cas de décomposer la matrice des covariances empiriques **S** sous la forme:

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}'$$

où Λ est la matrice diagonale des valeurs propres (rangées) et $\mathbf{U}$ le tableau des vecteurs propres unitaires correspondant. Cette relation s'écrit encore :

$$\mathbf{S} = (\mathbf{U}\Lambda^{1/2})(\mathbf{U}\Lambda^{1/2})' = \hat{\mathbf{U}}\hat{\mathbf{U}}'$$

où $\hat{\mathbf{U}}$ est le tableau des vecteurs propres multipliés par les racines carrées des valeurs propres correspondantes.

Avec ce point de vue, l'analyse en facteurs communs et spécifiques suppose qu'en retranchant une matrice diagonale à éléments positifs ($\tilde{\Delta}$ estimant Δ), on obtient une décomposition de la matrice des covariances empiriques sous la forme :

$$\mathbf{S} - \tilde{\Delta} = \Gamma\Gamma'$$

où Γ ne contient que q colonnes alors que dans $\mathbf{S} = \hat{\mathbf{U}}\hat{\mathbf{U}}'$ le tableau $\hat{\mathbf{U}}$ contenait p colonnes. On voit au passage qu'une analyse en composantes principales où les $p - q$ dernières valeurs propres sont proches et voisines de 0, donnera des résultats très voisins de ceux d'une analyse à q facteurs communs orthogonaux.

b– Estimation des paramètres inconnus

On n'insistera pas ici sur les problèmes posés par un tel modèle, qui font l'objet d'une abondante littérature. On donnera seulement quelques moyens pratiques de calcul.

Le problème essentiel est d'estimer Δ, matrice diagonale des variances des résidus spécifiques. Une fois Δ estimée par $\tilde{\Delta}$, il suffit de chercher les composantes principales (vecteurs propres) de $(\mathbf{S} - \tilde{\Delta})$; on ne doit normalement trouver qu'un petit nombre de composantes différentes (statistiquement) de 0.

Nous allons examiner ici une spécification particulière du modèle, puis donner un algorithme de calcul dans le cas général.

- Cas de variances spécifiques égales

On suppose *a priori* que les facteurs spécifiques ont tous même variance théorique σ^2 ; autrement dit par hypothèse $\Delta = \sigma^2\mathbf{I}$:

$$\mathbf{W} = \Gamma\Gamma' + \sigma^2\mathbf{I}$$

et, si on note s^2 une estimation de σ^2, la relation [3.3.2] devient :

$$\mathbf{x}_i = \Gamma\mathbf{f}_i + s\mathbf{e}_i$$

On obtiendrait une estimation de Γ en cherchant les composantes principales de la matrice $(\mathbf{S} - s^2\mathbf{I})$. En effet, effectuant l'analyse de $\mathbf{S}$, on écrit :

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}'$$

et par conséquent :

$$\mathbf{S} - s^2\mathbf{I} = \mathbf{U}\Lambda\mathbf{U}' - s^2\mathbf{U}\mathbf{U}' = \mathbf{U}\left(\Lambda - s^2\mathbf{I}\right)\mathbf{U}'$$

Les valeurs propres de $(\mathbf{S} - s^2\,\mathbf{I})$ sont celles de $\mathbf{S}$ diminuées de s^2 (les vecteurs propres étant identiques). Puisque $(\mathbf{S} - s^2\,\mathbf{I})$ doit être de rang q, il est nécessaire que s^2 soit valeur propre multiple d'ordre $p - q$ pour $\mathbf{S}$.

En particulier si, dans une analyse en composantes principales, les petites valeurs propres sont sensiblement égales, on peut considérer que les données sont engendrées par un modèle factoriel à variances spécifiques égales [1].

- une méthode de calcul dans le cas général

La méthode que nous donnons ici est simple [2]. Elle procède de façon itérative, en posant au départ $\tilde{\Delta} = \mathbf{0}$. On calcule les vecteurs propres unitaires de $\mathbf{S}$ rangés dans le tableau $\mathbf{U}$:

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}' = \hat{\mathbf{U}}\hat{\mathbf{U}}'$$

Si l'on veut retenir q facteurs communs, on ne garde que les q premières colonnes de $\hat{\mathbf{U}}$, tableau que l'on notera $\hat{\mathbf{U}}_1$. On devrait pouvoir écrire :

$$\mathbf{S} = \hat{\mathbf{U}}_1\hat{\mathbf{U}}_1' + \tilde{\Delta}$$

On estimera donc provisoirement $\tilde{\Delta}$ par les éléments diagonaux $\tilde{\Delta}_1$ de $(\mathbf{S} - \hat{\mathbf{U}}_1\hat{\mathbf{U}}_1')$, et on calculera les q premiers vecteurs propres $\hat{\mathbf{U}}_2$ de $(\mathbf{S} - \tilde{\Delta}_1)$.

A l'itération suivante on estime $\tilde{\Delta}$ par les éléments diagonaux $\tilde{\Delta}_2$ de $(\mathbf{S} - \hat{\mathbf{U}}_2\hat{\mathbf{U}}_2')$ et l'on poursuit les opérations jusqu'à observer une convergence raisonnable du processus. On aura alors obtenu la décomposition cherchée :

$$\mathbf{S} = \Gamma\Gamma' + \tilde{\Delta}\,.$$

Mentionnons pour conclure ce bref aperçu les travaux historiques d'Anderson et Rubin (1956) et de Lawley et Maxwell (1963) qui ont placé l'analyse factorielle en facteurs communs et spécifiques dans un cadre inférentiel classique.

---

[1] Ce modèle à variances spécifiques égales peut être justifié lorsque les p variables sont mesurées avec le même instrument (exemples : mensurations anthropométriques), et donc avec la même erreur.

[2] Cette procédure est parfois appelée analyse en facteurs principaux. Pour une première estimation de Δ, on peut également prendre (Joreskög, 1963), lorsque $\mathbf{S}$ est une matrice des corrélations, $\delta_{jj} = 1 - R_j^2$, où la quantité R_j^2 est le coefficient de corrélation multiple de la variable j avec toutes les autres. Ainsi, une variable très peu corrélée avec les autres aura une variance spécifique forte. Une variable qui peut s'exprimer comme combinaison linéaire des autres aura une variance spécifique nulle. Notons que $1 - R_j^2$ est l'inverse du $j^{ème}$ élément diagonal de $\mathbf{S}^{-1}$.

3.3.6 L'analyse en composantes indépendantes

L'analyse en composantes indépendantes (ACI) repose sur un modèle qui est une généralisation de l'analyse en facteurs communs et spécifiques et de l'analyse en composantes principales. Ce modèle a été suscité par les exigences des applications dans le domaine du traitement du signal. La formulation est très simple : au lieu de chercher des composantes non-corrélées, on s'attachera à chercher des composantes indépendantes. La non-corrélation est une condition suffisante d'indépendance dans le cas Gaussien (normal), mais pas dans le cas général. Il s'agit bien, comme dans la section précédente, d'estimer des facteurs latents qui sont censés exister, donc d'un modèle, mais l'analyse en composantes indépendante peut être aussi un outil d'exploration.

La non-corrélation des facteurs cachés est insuffisante dans la « séparation aveugle de sources », dont le problème de la *cocktail party* est l'illustration emblématique. On analyse un signal complexe qui provient de la superposition des sons émis par plusieurs locuteurs, et l'on veut retrouver les signaux émis par chacun des locuteurs. La supériorité de l'ACI est indiscutable pour ce type de traitement. Le modèle peut s'écrire comme celui de la section 3.3.5 précédente (avec ou sans terme résiduel $\mathbf{e}_i$) :

$$\underset{(p,1)}{\mathbf{x}_i} = \underset{(p,q)}{\Gamma}\ \underset{(q,1)}{\mathbf{f}_i} + \underset{(p,1)}{\mathbf{e}_i}$$

Mais ici, le vecteur $\mathbf{f}_i$ a ses composantes indépendantes, hypothèse plus forte que la non-corrélation, conduisant à des calculs plus complexes. Le modèle peut être rejeté s'il n'y a pas de terme résiduel car il se peut que l'on ne puisse pas trouver de combinaisons linéaires de facteurs indépendants qui satisfassent l'équation précédente.

a_ Différents points de vue

Les travaux de pionniers dans ce domaine sont ceux de Jutten (1987) et Jutten et Hérault (1991) dans le domaine de l'application des réseaux de neurones au traitement du signal. Citons également Comon (1994), Cardoso (1989, 1996), Hyvärinen (1996). Outre le tutoriel de Cardoso (1998), on trouvera dans Hyvärinen (1999) un article de synthèse clair sur le sujet. Comme les méthodes de *projection-pursuit* (Friedman et Tukey, 1974 ; Friedman, 1987) ont une certaine parenté avec les méthodes d'ACI, on peut également considérer ces travaux comme des travaux de pionniers. La *projection-pursuit*[1] cherche des « directions intéressantes » et non pas de façon explicite des composantes indépendantes. Mais pour trouver ces directions, elle utilise, comme le fera l'ACI, des critères de non-normalité. Il est intéressant de voir que les « factorialistes classiques », c'est-à-dire les psychologues utilisant le modèle de

[1] Evoquée incidemment au chapitre 8, section 8.2, à propos des analyses locales. Cf. aussi « Projections révélatrices et exploratoires », par Caussinus et Ruiz-Gaussens, in : Govaert(2003)

la section 3.3.5, ont cherché, en procédant à des rotations, des directions interprétables dans l'espace des premiers facteurs : les critères qu'ils utilisent pour ce faire ont une certaine parenté avec ceux de l'ACI.

b_ Comment s'éloigner de la normalité ?

Nous ne pouvons qu'esquisser quelques unes des approches, renvoyant au *survey* précité de Hyvärinen pour une formulation plus complète.

L'approche la plus répandue fait intervenir l'information mutuelle, ou divergence de Kullback-Leibler $I(\mathbf{y})$ mesurant la distance entre la distribution $f(y)$ du vecteur aléatoire $\mathbf{y}$ et le produit des distributions marginales de ses composantes $\prod_{j=1}^{p} f_j(y_j)$ (différences entre les entropies différentielles H).

$$I(\mathbf{y}) = \sum_{j=1}^{p} H(y_j) - H(\mathbf{y}), \qquad \text{avec : } H(t) = -\int f(t)\log f(t)dt$$

On montre alors que chercher une matrice orthogonale $\mathbf{U}$ telle que $\mathbf{y} = \mathbf{Ux}$ qui minimise $I(\mathbf{y})$ (la variance de $\mathbf{y}$ étant fixée) conduit à une indépendance maximale des composantes, en même temps qu'à un éloignement maximal de la normalité (ou : *gaussiannité*). Parmi les nombreuses techniques de résolution proposées et les approximations auxquelles elles donnent lieu, citons la maximisation de la valeur absolue du coefficient d'aplatissement (*kurtosis*)[1] des valeurs d'une composante (E désigne l'espérance mathématique).

$$kurt(z) = E(z^4) - 3\left[E(z^2)\right]^2$$

Ce critère semble manquer de robustesse, par excès de sensibilité aux *outliers*.

Le critère de rotation des factorialistes classiques le plus utilisé (critère dit *Varimax*, de Kayser, 1958, proportionnel à la variance empirique des carrés des coordonnées) porte de façon globale sur les k composantes cherchées (z_{ij} étant la nouvelle coordonnée de la variable i sur la composante j)

$$\max_{\mathbf{u}} \left\{ \sum_{i=1}^{p}\sum_{j=1}^{k} \left(z_{ij}^2 - \frac{1}{p}\sum_{r=1}^{p} z_{rj}^2 \right)^2 \right\}, \quad \text{avec } \mathbf{U'U}=\mathbf{I_p}$$

Ainsi, on rejoint le critère varimax si l'on remplace le coefficient 3 de $kurt(z)$ par 1, et si l'on fait une moyenne sur les composantes. Hommage soit rendu à l'intuition des factorialistes qui, guidés par leur profonde connaissance des phénomènes qu'ils étudiaient, ont frôlé un résultat théorique important !

---

[1] Le coefficient $kurt(z)$ est le cumulant d'ordre 4 de la loi de z ((il intervient dans le 4ème terme du développement en série du logarithme de la fonction caractéristique de cette loi). Pour la loi normale : $kurt(z) = 0$; il est positif pour des distributions « plus pointues » que la loi normale, et négatif pour des distributions plus aplaties. Sa valeur absolue est un bon indice de « non-normalité ».

3.3.7 Régression sur composantes principales et régression régularisée

On a vu lors de la régression (chapitre 2) que la structure du tableau à n lignes et p colonnes $\mathbf{X}$ des variables explicatives (structure décrite par la matrice des covariances) avait des répercussions sur la qualité des coefficients de régression (§ 2.2.4.b). Le calcul des coefficients de régression requiert une matrice $\mathbf{X'X}$ inversible et donc des vecteurs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ linéairement indépendants.

Si les variables explicatives sont fortement corrélées (autrement dit si certains des vecteurs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ ont des directions voisines) alors l'inversion de la matrice $\mathbf{X'X}$ est difficile. Le vecteur $\mathbf{a}$ dont les composantes sont les coordonnées de la projection de $\mathbf{y}$ dans la base de V_X formée par $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ est mal spécifié. Les résultats de la régression seront instables[1].

On a également évoqué le fait que la méthode des moindres carrés pouvait donner un poids excessif à des points éloignés (pouvant parfois être erronés ou aberrants).

On vient de voir d'autre part que l'analyse en composantes principales décrit la structure d'un tableau $\mathbf{X}$ en mettant en évidence les interrelations entre variables (colonnes de $\mathbf{X}$) ; elle permet également de visualiser les points-observations (points-lignes de $\mathbf{X}$) et donc d'aider à repérer d'éventuelles anomalies dans leur distribution. Enfin, on a vu que l'analyse fournit une base orthogonale hiérarchisée du sous-espace de $\mathcal{R}^n$ appelé V_X.

Il est clair dans ces conditions qu'une analyse en composantes principales préalable permettra d'apprécier l'existence de colinéarités entre les variables explicatives, de détecter les redondances et compétitions entre prédicteurs; de repérer les individus occupant des positions aberrantes ou simplement suspectes. Il s'agit là d'une phase descriptive qui doit précéder la régression.

L'analyse peut également fournir des variables artificielles orthogonales (les coordonnées des points-observations sur les nouveaux axes) comme nouveaux prédicteurs : c'est la régression sur composantes principales, recommandée

---

[1] La décomposition en éléments propres de $\mathbf{X'X}$ s'écrit : $\mathbf{X'X} = \mathbf{U}\Lambda\mathbf{U}' = \sum_{\alpha=1}^{p} \lambda_\alpha \mathbf{u}_\alpha \mathbf{u}'_\alpha$, où Λ est la matrice diagonale dont le $\alpha^{ième}$ élément est la valeur propre λ_α et $\mathbf{U}$ le tableau des vecteurs propres unitaires (en colonnes) correspondants. On a donc également :

$$(\mathbf{X'X})^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}' = \sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha} \mathbf{u}_\alpha \mathbf{u}'_\alpha$$

L'estimation de la matrice de covariances du vecteur $\mathbf{a}$ des coefficients vaut :

$$Var(\mathbf{a}) = s^2(\mathbf{X'X})^{-1} = s^2 \sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha} \mathbf{u}_\alpha \mathbf{u}'_\alpha$$

Sous cette forme on voit comment une ou plusieurs valeurs propres presque nulles rendent imprécis l'ajustement.

lorsque les variables explicatives sont nombreuses ou fortement corrélées entre elles. L'analyse factorielle joue donc un double rôle : un rôle d'exploration préalable et un rôle de régularisation[1].

a – Principe de la régression régularisée

Le principe revient à remplacer les p variables explicatives $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, définies dans $\mathcal{R}^n$, par leurs p composantes principales qui engendrent le même sous-espace V_X à p dimensions. S'il existe r relations linéaires entre les variables explicatives, alors la transformation des p variables fournira $q = p - r$ composantes principales. Il est possible ensuite d'exprimer les résultats de la régression en fonction des variables initiales. Nous nous plaçons dans $\mathcal{R}^n$ où un point $\mathbf{y}$ est projeté sur le sous-espace V_X engendré par les vecteurs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

Les p vecteurs propres $\mathbf{v}_k$ auxquels correspondent p composantes principales constituent une base orthonormée du sous-espace V_X sur lequel on veut projeter $\mathbf{y}$.

On élimine le problème posé par la quasi-colinéarité en supprimant de cette base les $p - r$ vecteurs $\mathbf{v}_k$ correspondant à des valeurs propres λ_k nulles ou très faibles. Autrement dit on ne retient que les q premières composantes principales de variances non négligeables.

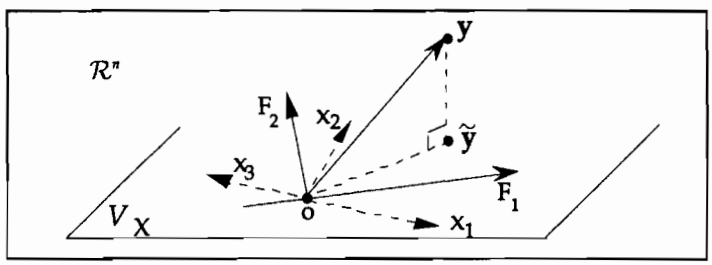


Figure 3.3 – 11. Régression sur composantes principales

Les variables étant centrées, nous sommes dans le cas de l'analyse générale du chapitre 1. Le tableau $\mathbf{X}$ est reconstitué sur les q premiers axes factoriels par la formule ($\mathbf{v}_\alpha$ et $\mathbf{u}_\alpha$ sont unitaire) :

$$\mathbf{X}^* = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\, \mathbf{v}_\alpha \mathbf{u}'_\alpha = \mathbf{V}_q \Lambda_q^{1/2} \mathbf{U}'_q \qquad \text{(avec } q < p\text{)}$$

---

[1] Les techniques de régularisation, largement utilisées en analyse discriminante (chapitre 7), participent à la résolution de problèmes *mal posés* (ici : cas de colinéarité entraînant une singularité de la matrice $\mathbf{X'X}$, et donc une impossibilité de calcul de $\mathbf{a}$) ou de problèmes *pauvrement posés* (ici : cas de quasi-colinéarité, entraînant une instabilité numérique de $(\mathbf{X'X})^{-1}$ et du vecteur $\mathbf{a}$ des coefficients de régression). Pour une revue des traitements de la colinéarité dans le cas de la régression, cf. Palm et Iemma (1995).

où $\mathbf{V}_q$ et $\mathbf{U}_q$ sont les matrice respectivement d'ordre *(n,q)* et *(p,q)* contenant en colonne les vecteur propres $\mathbf{v}_\alpha$ et $\mathbf{u}_\alpha$ et Λ_q la matrice diagonale *(q,q)* des q premières valeurs propres.

On calcule[1] à partir de ce nouveau tableau le vecteur de coefficient $\mathbf{a}^*$ tel que :

$$\mathbf{a}^* = \sum_{\alpha=1}^{q} \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{u}_\alpha \mathbf{v}'_\alpha \mathbf{y} \qquad [3.3 - 6]$$

Remarquons que $\mathbf{a}^*$ n'est plus unique, puisque tout vecteur de la forma $\mathbf{a}^* + \mathbf{c}$ (avec $\mathbf{c}$ tel que $\mathbf{U}'\mathbf{c} = 0$) satisfait aux équations [2.2 - 1].

Pour que la relation $E(\mathbf{a}^*) = \alpha$ soit vérifiée, il faut, dans le cas de l'estimation précédente, que le modèle théorique spécifie que α soit de la forme $\mathbf{U}\beta$, β étant un vecteur quelconque à q composantes. Dans ces conditions, l'estimation de la matrice des covariances de $\mathbf{a}^*$ (de rang q) sera :

$$Var(\mathbf{a}^*) = s^2 \sum_{\alpha=1}^{q} \frac{1}{\lambda_\alpha} \mathbf{u}_\alpha \mathbf{u}'_\alpha$$

Notons que $\mathbf{X} = \mathbf{X}^*$ s'il y a exactement q valeurs propres différentes de 0.

b – Variables supplémentaires et régression

La procédure de mise en éléments supplémentaires dans une analyse en composantes principales constitue une variante descriptive de la régression multiple. D'un point de vue géométrique, les deux situations sont similaires :

- les p variables explicatives engendrent un sous-espace V_X ayant au plus p dimensions sur lequel est projetée la variable à expliquer;
- les p variables actives de l'analyse engendrent aussi un sous-espace à au plus p dimensions que l'on réduit à q facteurs pour le visualiser et c'est sur ce sous-espace réduit à q dimensions que l'on projette les variables supplémentaires pour les situer par rapport aux variables actives.

La formule [3.3 - 6] permet d'expliciter ce lien. Calculons à partir d'elle la nouvelle estimation $\tilde{\mathbf{y}}^*$ de $\mathbf{y}$ en utilisant la formule [1.2 - 3] du chapitre 1 :

$$\tilde{\mathbf{y}}^* = \mathbf{X}^*\mathbf{a}^* = \sum_{\alpha=1}^{q} \mathbf{v}_\alpha \mathbf{v}'_\alpha \mathbf{y}$$

---

[1] Les équations [2.2 - 1] du chapitre 2 s'écrivent $\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{X}'\mathbf{y}$, c'est-à-dire, en abandonnant provisoirement les indices q :

$$\mathbf{U}\Lambda\mathbf{U}'\mathbf{a} = \mathbf{U}\Lambda^{1/2}\mathbf{V}'\mathbf{y}$$

Le vecteur $\mathbf{a}$ n'ayant que q composantes indépendantes peut s'écrire sous la forme : $\mathbf{a} = \mathbf{U}\mathbf{b}$, d'où puisque $\mathbf{U}'\mathbf{U} = \mathbf{I}$ (matrice unité (q,q)) :

$$\mathbf{U}\Lambda\mathbf{b} = \mathbf{U}\Lambda^{1/2}\mathbf{V}'\mathbf{y}$$

Prémultipliant les deux membres par $\mathbf{U}'$, on obtient $\mathbf{b}$:

$$\mathbf{b} = \Lambda^{-1/2}\mathbf{V}'\mathbf{y}, \quad \text{donc} \quad \mathbf{a} = \mathbf{U}\Lambda^{-1/2}\mathbf{V}'\mathbf{y}$$

On a ainsi obtenu une expression de l'opérateur-projection $\mathbf{P}_{X^*}$ sur l'espace des q premiers axes factoriels. Le dernier membre rappelle clairement que la coordonnée ($\mathbf{v}'_\alpha \mathbf{y}$) de $\tilde{\mathbf{y}}^*$ sur l'axe unitaire $\mathbf{v}_\alpha$ correspond au positionnement classique de $\mathbf{y}$ en variable supplémentaire dans l'analyse dont les variables actives sont les colonnes de $\mathbf{X}$.

c – Expression des coefficients dans la nouvelle base

Désignons par $\mathbf{z}_\alpha$ le vecteur, dans $\mathcal{R}^n$, des nouvelles coordonnées des points sur l'axe $\mathbf{u}_\alpha$. Rappelons que l'on a les relations :

$$\mathbf{z}_\alpha = \mathbf{X}^*\mathbf{u}_\alpha = \mathbf{X}\mathbf{u}_\alpha = \sqrt{\lambda_\alpha}\mathbf{v}_\alpha \qquad (\alpha = 1,2,\ldots,q)$$

L'ajustement, dans $\mathcal{R}^n$, avec les nouvelles variables explicatives ($\mathbf{v}_1$, $\mathbf{v}_2$,...,$\mathbf{v}_q$) s'écrira :

$$\mathbf{y} = \mathbf{Z}\mathbf{c} + \mathbf{e}$$

où $\mathbf{Z}$ est le tableau (n, q) des q vecteurs orthogonaux $\mathbf{z}_\alpha$ et $\mathbf{c}$ le vecteur des q nouveaux coefficients de régression cherchés.

Puisque $\mathbf{Z}'\mathbf{Z} = \Lambda$, matrice diagonale dont les éléments diagonaux sont les valeurs propres, on a :

$$\mathbf{c} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \Lambda^{-1}\mathbf{Z}'\mathbf{y}$$

Cette situation idéale pour laquelle les variables explicatives sont orthogonales revient d'ailleurs à faire q régressions simples, car chacun des p coefficients peut être estimé séparément. On a en effet :

$$c_\alpha = \frac{\mathbf{z}'_\alpha \mathbf{y}}{\lambda_\alpha} = \frac{cov(\mathbf{z}_\alpha, \mathbf{y})}{var(\mathbf{z}_\alpha)}$$

La matrice des covariances des coefficients $\mathbf{c}$ sera estimée par :

$$Var(\mathbf{c}) = s^2(\mathbf{Z}'\mathbf{Z})^{-1} = s^2\Lambda^{-1}$$

autrement dit ces coefficients de régression sont non corrélés et ont pour variances les quantités :

$$var(c_\alpha) = \frac{s^2}{\lambda_\alpha}$$

3.3.8 Aperçu sur les autres méthodes dérivées

De nombreuses techniques sont directement dérivées de l'analyse en composantes principales. Les variantes non-paramétriques du paragraphe précédent en sont des exemples. Certaines présentations de l'analyse des correspondances (cf. chapitre 4) considèrent cette méthode comme une analyse en composantes principale particulière. Cela est possible si l'on traite les deux

espaces (lignes et colonnes) séparément, ce qui n'est pas l'optique choisie ici. En effet, ce traitement séparé masque un des apports méthodologiques fondamentaux des analyses factorielles descriptives. L'analyse en composantes principales, qu'il s'agisse d'analyse normée ou non-normée, analyse les individus par rapport à leur *centre de gravité* et les variables par rapport à *l'origine des axes*. Cette dissymétrie de traitement des lignes et des colonnes correspond à des domaines d'applications spécifiques et induit des règles d'interprétation particulières. La décomposition aux valeurs singulières (ou encore analyse générale, ou théorème d'Eckart et Young) est bien le noyau théorique commun aux deux méthodes.

Citons parmi les méthodes dérivées l'*analyse des corrélations partielles* ou *analyse avec variables instrumentales* (Rao, 1964), qui sera abordée au chapitre 8. Dans ce cas, on ne se contente plus d'éliminer les effets de l'hétérogénéité des variables (opérations de centrage et de réduction) mais on se propose d'éliminer également l'effet d'autres variables, en procédant à une régression multiple préalable. L'analyse logarithmique (Kazmierczak, 1985) est une analyse en composantes principales non-normée du tableau (doublement centré en lignes et en colonnes) des logarithmes des variables initiales. Cette variante possède d'intéressantes propriétés de stabilité et de robustesse.

3.4 Interprétation, validation

Devant les résultats d'une analyse en composantes principales, on est naturellement conduit à poser un certain nombre de questions sur la qualité des représentations. Ces questions concernent d'ailleurs toutes les analyses en axes principaux :

- Observe-t-on vraiment quelque chose? Les données ont-elle une structure? Ou, au contraire, de simples fluctuations d'échantillonnage suffiraient-elles à expliquer les valeurs obtenues pour les valeurs propres et les taux d'inertie?
- Les premières valeurs propres sont-elles significatives? Que représente le taux d'inertie en terme d'information? Comment apprécier la position d'un point dans l'espace factoriel?

Cette section va tenter de donner des réponses à ces questions. Après le paragraphe 3.4.1 consacré à la lecture élémentaire des graphiques, on aborde le problème du choix du nombre d'axes à retenir (§ 3.4.2) en se référant surtout aux pratiques empiriques existantes. Le paragraphe 3.4.3 présente les seuls résultats théoriques vraiment utilisables en pratique : les intervalles de confiance asymptotiques d'Anderson pour les valeurs propres. Enfin le paragraphe 3.4.4 présente la mise en oeuvre de la méthodologie *bootstrap* dans le cadre de l'analyse en composantes principales.

3.4.1 Eléments pour l'interprétation

Les axes principaux (ou axes factoriels) permettent d'obtenir la meilleure visualisation approchée, au sens des moindres carrés, des distances entre les individus d'une part et entre les variables d'autre part. Pour interpréter (éventuellement) ces directions principales et ces distances, il faut apprécier correctement cette approximation. On considérera surtout le cas, le plus fréquent, de l'analyse en composantes principales normée.

a – les variables

Nous nous plaçons ici dans le nuage des points-variables (p points de $\mathcal{R}^n$).

Les variables fortement corrélées avec un axe vont contribuer à la définition de cet axe. Cette corrélation se lit directement sur le graphique puisqu'il s'agit de la coordonnée du point-variable sur l'axe (formule [3.2 - 4]).

On s'intéresse par conséquent aux variables présentant les plus fortes coordonnées (ce qui les situent proches du cercle de corrélations) et l'on interprétera les composantes principales en fonction des regroupements de certaines de ces variables et de l'opposition avec les autres.

Rappelons que le cosinus de l'angle sous lequel on voit deux points-variables actives dans $\mathcal{R}^n$ n'est autre que le coefficient de corrélation de ces deux variables. Selon la qualité de l'ajustement, cette propriété sera plus ou moins bien conservée en projection. On se gardera d'interpréter la distance entre deux variables actives qui ne seraient pas proches du cercle de corrélation. Les figures du paragraphe 3.5.1 de ce chapitre donnent un exemple de cercle des corrélations dans le plan des deux premiers facteurs.

Dans le cas des variables continues supplémentaires, les corrélations n'étant pas transitives, il est prudent de ne pas interpréter abusivement les proximités entre variables en terme de corrélation, bien que celles-ci en soient souvent de bonnes images.

b – Les individus

Si les points-individus ne sont pas anonymes pour l'étude, on s'intéresse à ceux qui participent à la formation des axes. On calcule la *contribution* de chaque point i (de masse m_i) à l'inertie de l'axe α. Celle-ci s'exprime par la formule :

$$Cr_\alpha(i) = \frac{m_i \psi_{\alpha i}^2}{\lambda_\alpha}$$

où λ_α est l'inertie de l'axe α et $m_i \psi_{\alpha i}^2$ est la contribution de l'individu i à l'inertie de cet axe. On a:

$$\sum_{i=1}^{n} Cr_\alpha(i) = 1$$

On s'intéressera surtout aux individus qui ont les plus fortes contributions relatives aux axes.

Lorsque les n individus sont affectés d'une même masse égale à $1/n$, l'inertie d'un point varie comme sa distance au centre de gravité. Les individus qui contribuent le plus à la détermination de l'axe sont les plus excentrés et l'examen des coordonnées factorielles ou la lecture du graphique suffisent à interpréter les facteurs dans ce cas.

c – Possibilité d'apparition d'un facteur "taille"

L'analyse du nuage des variables se faisant à partir de l'origine, les variables peuvent être toutes situées du même côté d'un axe factoriel. Une telle disposition apparaît lorsque toutes les variables sont corrélées positivement entre elles.

Si pour un individu, une variable prend une valeur forte, toutes les autres variables prennent également des valeurs fortes. Cette caractéristique apparaît le plus souvent sur le premier axe, que l'on appelle alors "*facteur taille*".

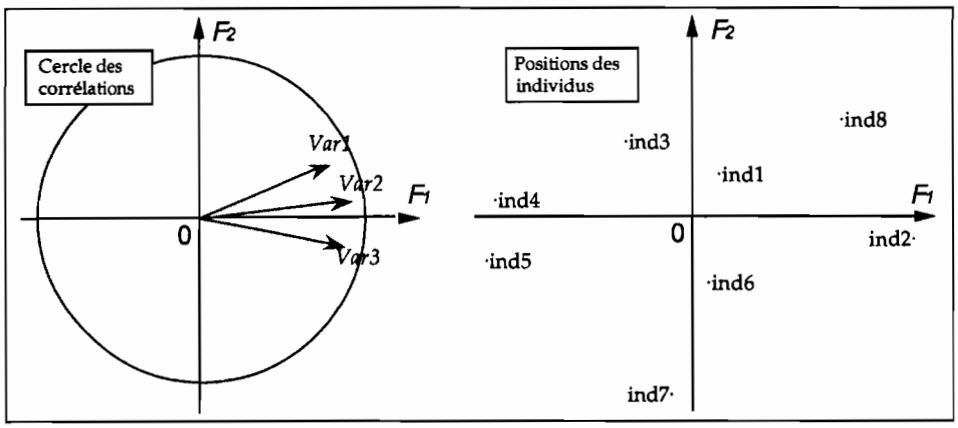


Figure 3.4-1. Exemple de *Facteur taille*

On peut lire, par exemple sur la figure 3.4-1, que les individus 4 et 5 ont des comportements semblables caractérisés par des valeurs faibles pour les trois variables, alors que les individus 2 et 8 ont au contraire simultanément des bons "scores" pour ces mêmes variables.

d – Méthodes empiriques de validation

En pratique, les méthodes empiriques de validation qui permettent un premier contrôle de la qualité des résultats et de leur stabilité font partie intégrante du processus d'analyse des données.

Quels sont les différents éléments qui peuvent conditionner la qualité et la stabilité des résultats d'une analyse en composantes principales ?

Nous en distinguons quatre :

1 - le choix et le poids des variables,
2 - le codage des variables,
3 - les erreurs de mesure,
4 - les poids des individus, les fluctuations d'échantillonnage [1].

Les quatre sources de perturbation peuvent donner lieu à des modifications du tableau de données et permettent de vérifier la permanence de la configuration initiale. L'usage exploratoire des méthodes factorielles nécessite, non pas une analyse, mais une série d'analyses : à chaque étape, le tableau de données sera modifié par le choix des variables ou d'individus (avec ajout ou retrait de certains éléments), les corrections d'éventuelles erreurs, le recodage des données, etc.. Cette démarche proche de la "structuration en escalade" décrite par Mallows et Tukey (1982) permet une connaissance progressive du phénomène et constitue en soi une procédure de validation des résultats.

- Le choix et le poids des variables

Le problème se pose lorsque le statisticien a la possibilité d'échantillonner dans l'espace des variables, ce qui n'est pas toujours le cas. On pourra alors effectuer des "ponctions aléatoires" dans l'ensemble des variables, afin d'éprouver la sensibilité des résultats vis-à-vis de la composition de cet ensemble. Si les variables sont nombreuses, un *bootstrap* des variables est possible (section 3.4.4).

- Le codage des variables

Le codage apparaîtra comme source de perturbation éventuelle des résultats dans le cas des notes, des échelles ou des classements. Il est alors important de vérifier que les configurations obtenues résistent à des changements de variables monotones très déformants (logarithme, exponentiel, etc.), afin de s'assurer que *l'ordre* des notes est plus important que les *propriétés métriques* particulières à l'échelle utilisée. Une exemple de telles déformations est donné au paragraphe 3.5.2, à propos du second exemple.

- Les erreurs de mesure

L'ordre de grandeur de ces erreurs, ainsi que leur distribution approximative dans la population, doivent être spécifiés par l'utilisateur en fonction de sa propre connaissance du domaine étudié. Par exemple dans le cas classique des réponses ordonnées du type :"pas du tout d'accord; pas d'accord; assez d'accord; tout à fait d'accord", on peut supposer que l'individu enquêté a une chance sur deux d'avoir exprimé exactement ce qu'il ressentait, une chance sur quatre (sauf aux extrémités) de répondre à une modalité immédiatement

---

[1] Les trois premiers points correspondent à ce que Greenacre (1984) désigne par *stabilité interne* (l'univers est constitué par le tableau analysé, sans référence à une population plus large). Le quatrième point répond plutôt aux demandes de *stabilité externe* (visant à étendre les faits structuraux observés à partir du tableau analysé à une population plus générale).

contigüe. Les programmes de calcul permettront en général de simuler une grande variété de situations dont la traduction analytique serait inextricable. De ce fait les hypothèses que l'on soumet à l'épreuve d'un test peuvent être mieux adaptées aux situations réelles et aux préoccupations des utilisateurs que les hypothèses classiques donnant lieu à une formulation analytique. En revanche la mise en œuvre de ces validations sur-mesure est coûteuse.

- Les poids des individus, les fluctuations d'échantillonnage

Les typologies obtenues par analyse factorielle n'exigent pas une représentativité de l'échantillon aussi stricte[1] que les estimations de pourcentages ou de moments d'ordre *1* (moyennes, fréquences). Cette relative stabilité vis-à-vis de la représentativité de l'échantillon est un fait d'expérience, étayé par les considérations sur la stabilité (section 1.4 du chapitre 1).

Dans les enquêtes par sondage, lorsque l'échantillon n'est pas représentatif et privilégie par exemple une sous-population de la population mère, chaque individu de l'échantillon est alors affecté d'un "coefficient de redressement" qui permet d'ajuster les moyennes ou les marges sur des valeurs connues dans la population parente[2]. Il n'est pas rare que les typologies obtenues fassent preuve d'une bonne stabilité et qu'elles soient les mêmes que l'échantillon soit "redressé", ou que l'analyse soit faite sur les données brutes (alors que les *variables de niveau*, moyennes ou pourcentages, peuvent enregistrer ds variations importantes). Mais les méthodes privilégiées pour étudier la stabilité des résultats vis-à-vis de fluctuations d'échantillonnage sont les techniques de rééchantillonnage examinées en section 3.4.4.

3.4.2 Choix du nombre d'axes : règles empiriques, validation externe

Qu'il s'agisse d'une simple visualisation des données ou de l'utilisation des axes factoriels en vue d'une analyse ultérieure (classification sur facteurs, régression ou analyse discriminante sur facteurs), il reste important de savoir combien d'axes retenir, autrement dit de connaître la dimension de l'espace de représentation.

Il existe quatre types de procédures pour guider le choix de ce nombre d'axes :

---

[1] Bien entendu, un échantillon où certains aspects de la population parente sont absents, ne pourra pas fournir de résultats "extrapolables", même si les configurations obtenues sont stables.

[2] Ces redressements de tableaux à partir de leur marge se font en général à partir d'algorithmes itératifs (iterative proportional fitting) proposés à l'origine par Deming et Stephan (1940). Pour une vision historique et générale, cf. Thionet (1976) et d'autres publications sur ce thème, notamment celles de Deville et Särndal (1992), et les travaux publiés dans les ouvrages édités par Brossier et Dussaix (1999), Lejeune (2001), Droesbeke et Lebart (2001), Ardilly (2004).

a - des règles empiriques.
b - des procédures de validation externe.
c - des critères fondés sur certaines propriétés statistiques des valeurs propres.
d - des méthodes de calcul de stabilité, de rééchantillonnage ou de simulation.

Dans ce paragraphe, nous évoquerons ici brièvement les points (a) et (b). Le point (c) fera l'objet du paragraphe 3.4.3 suivant, consacré aux « critères statistiques pour les valeurs propres ». Le point (d) sera traité dans le paragraphe 3.4.4 dévolu aux méthodes de bootstrap utilisables en analyse en composantes principales.

a – Règles empiriques

Les règles empiriques sont fondées sur l'allure de la séquence des valeurs propres [1]. Deux règles, attribuées respectivement à Cattell et Kaiser, seront citées, surtout à titre historique.

- Critère de Cattell

Lorsqu'un tableau est généré suivant un modèle stipulant l'indépendance de ses lignes et de ses colonnes, on observe une décroissance régulière des valeurs propres. Cette remarque est à l'origine de procédures empiriques pour juger du nombre d'axes à retenir. On étudie l'histogramme de décroissance des valeurs propres pour y déceler un changement de pente.

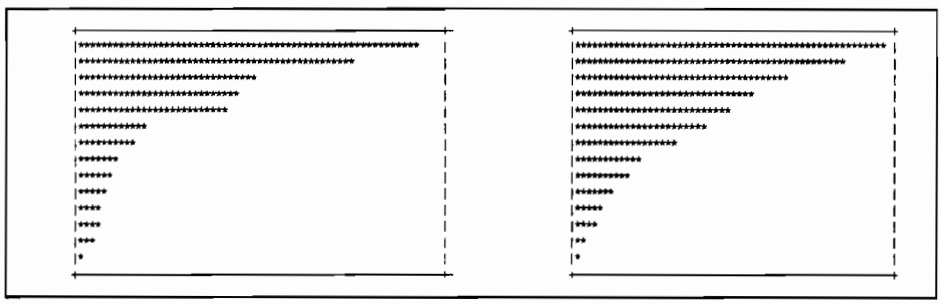


Figure 3.4-2.
Paliers dans la décroissance des valeurs propres

Figure 3.4-3.
Décroissance régulière des valeurs propres

Chaque fois que l'histogramme des valeurs propres présente un "décrochage" ou une discontinuité, on peut supposer que quelque chose de non aléatoire intervient. Ce repérage des "coudes" était préconisé par Cattell (1966).

---

[1] Ces procédures ont surtout concerné en pratique l'analyse factorielle en facteurs communs et spécifiques ou « analyse des psychologues », dont l'analyse en composantes principales est un cas particulier - cas des variances spécifiques égales ou nulles (cf. § 3.3.5).

Dans une analyse normée, la somme des inerties est égale au nombre de variables (trace de la matrice des corrélations) et donc l'inertie moyenne vaut *1*. Chaque axe étant une combinaison particulière des variables d'origine, on s'intéresse en général aux axes ayant une inertie "notablement" supérieure à la moyenne[1]. On observe souvent une décroissance assez irrégulière des premières valeurs propres (Figure 3.4-2).

Si les données sont peu structurées (les variables ne sont pas fortement corrélées entre elles), le nuage a une forme "régulière". Dans ce cas, les valeurs propres sont régulièrement décroissantes (Figure 3.4-3) et, en général, l'analyse factorielle ne fournira pas des résultats intéressants.

Les pourcentages d'inertie des axes (chapitre 1, paragraphe 1.2-5) définissent les "pouvoirs explicatifs" des facteurs : ils représentent la part de la variance (ou inertie) totale prise en compte par chaque facteur. Son appréciation doit cependant tenir compte du nombre de variables et du nombre d'individus. Un taux d'inertie (relatif à un axe) égal à 10% peut être une valeur importante si le tableau possède 100 variables et très faible s'il n'en a que 10. Comme nous le signalerons à propos de l'analyse des correspondances (chapitre 4), l'inertie est une mesure pessimiste du pouvoir explicatif des facteurs, liée parfois de façon assez arbitraire au codage des données.

***Remarque* : Influence du choix des variables sur les taux d'inertie**

Si l'on complète un tableau à n lignes et p colonnes, par q nouvelles colonnes formées de nombres pseudo-aléatoires, l'analyse en composantes principales normée du nouveau tableau à $p+q$ colonnes donnera les mêmes premiers axes (s'ils prédominent) que l'analyse du tableau initial. Les pourcentages de variance expliquée seront cependant plus faibles (car la trace qui valait p, vaut maintenant $p+q$). Pourtant la part d'information dont les axes rendent compte reste naturellement la même, puisque l'on a complété le tableau par du « bruit ».En pratique, on est dans une situation analogue lorsque le nombre potentiel des variables est très grand (cas par exemple de la présence d'espèces animales ou végétales dans les relevés écologiques). Une certaine discipline dans le choix du recueil des données, dictée par les critères d'homogénéité, devrait en principe permettre d'éviter ces inconvénients.

- Critère de Kayser

Le second critère empirique est le critère de Kaiser (1961), qui stipule de ne retenir que les valeurs propres supérieures à la moyenne des valeurs propres (c'est-à-dire à *1* dans le cas d'une analyse en composantes principales sur matrices de corrélation), en s'appuyant notamment sur des travaux de Guttman (1954)[2]. D'un emploi très répandu à cause de son extrême simplicité,

---

[1] Cette règle empirique est encore adoptée par certains utilisateurs.

[2] Cf. également, parmi de très nombreuses publications sur ce thème, les articles de synthèse de Anastassakos et d'Aubigny (1984), Francisco et Finch (1980) et la revue faite par Jolliffe (1986).

il peut être facilement mis en défaut. Ainsi, une analyse en composantes principales sur matrice des corrélations en biométrie peut produire un facteur de taille très dominant. Comme la trace est constante, les autres valeurs propres sont condamnées à être très petites, ce qui peut conduire à sous-estimer l'importance d'autres dimensions.

b – Procédures externes

Les procédures externes sont fondées sur des connaissances extérieures au tableau de données (interprétabilité de certains résultats, informations apportées par le positionnement de certaines variables supplémentaires [1]).

De telles procédures externes peuvent être couplées avec les procédures de rééchantillonnage dont on parlera au paragraphe 3.4.4. Ainsi, prenons le cas d'une variable supplémentaire nominale, n'ayant donc pas participé à la construction des axes. Si sa projection s'avère avoir une position significative sur un axe, cela suffit à valider l'axe, même si cet axe ne correspond pas à une valeur propre particulièrement élevée. Il faut cependant tenir compte des effets de « comparaisons multiples » lorsque les variables supplémentaires sont nombreuses. C'est donc ici le pouvoir de prédiction sur la variable externe qui permet de choisir la dimension de l'espace des prédicteurs[2]. Les procédures externes jouent un rôle important dans la méthodologie de l'analyse des données.

3.4.3 Critères statistiques pour les valeurs propres

Dans sa publication, citée dans l'annexe technique de ce chapitre, donnant l'expression de la densité des valeurs propres d'une matrice de Wishart, Girshick (1939) calcule les variances et covariances asymptotiques (quand le nombre d'observations n tend vers l'infini) des valeurs propres et vecteurs propres de la matrice des covariances expérimentales $\mathbf{S}$, ceci dans le cas où la matrice des covariances théoriques $\boldsymbol{\Sigma}$ a toutes ses valeurs propres distinctes. Il donne également les variances et covariances théoriques des valeurs propres de la matrice des corrélations expérimentales, lorsque la matrice de corrélation théorique $\mathbf{R}$ a également toutes ses valeurs propres distinctes.

Bartlett (1950, 1951) propose une méthode pour tester l'égalité de p-q valeurs propres des matrices $\boldsymbol{\Sigma}$ ou $\mathbf{R}$. Lawley (1956) approfondit le cas des p-q plus

---

[1] C'est le cas notamment en analyse des correspondances multiples quand les modalités d'une variable nominale supplémentaire possède des valeurs-test très significatives sur un ou plusieurs axes.

[2] Plus généralement, ce type de procédure permet de sélectionner un sous-espace qui n'est pas nécessairement engendré par des axes consécutifs.

petites valeurs propres de $\mathbf{\Sigma}$. Anderson (1963) a généralisé ces résultats, en déterminant les lois limites des valeurs propres sans nécessairement supposer que les valeurs théoriques correspondantes sont distinctes.

Intervalles de confiance d'Anderson

Anderson démontre en particulier, pour tester l'égalité des r plus petites valeurs propres $\hat{\lambda}_\alpha$ de la matrice des covariances expérimentales **S**, que la statistique :

$$X^2 = nr \log \frac{\left(\frac{1}{r}\right) \sum_{\alpha=p-r+1}^{\alpha=p} \hat{\lambda}_\alpha}{\left(\prod_{\alpha=p-r+1}^{\alpha=p} \hat{\lambda}_\alpha \right)^{1/r}}$$

(nr fois le logarithme du rapport de la moyenne arithmétique des r plus petites valeurs propres à leur moyenne géométrique) est asymptotiquement distribué comme un χ^2 à $[\frac{r(r+1)}{2} - 1]$ degrés de liberté.

Les intervalles de confiance asymptotiques d'Anderson utilisés en pratique pour les valeurs propres remontent en fait aux travaux précités de Girshick. Le résultat utile est le suivant : si les valeurs propres théoriques λ_α de $\mathbf{\Sigma}$ sont distinctes, les valeurs propres $\hat{\lambda}_\alpha$ de la matrice des covariances empiriques **S** suivent asymptotiquement des lois normales d'espérance λ_α et de variance $2\lambda_\alpha^2/(n-1)$ où n est la taille de l'échantillon.

On en déduit les intervalles de confiance approchés au seuil 95% :

$$\lambda_\alpha \in \left[\hat{\lambda}_\alpha \left(1 - 1.96\sqrt{2/(n-1)}\right) ; \hat{\lambda}_\alpha \left(1 + 1.96\sqrt{2/(n-1)}\right) \right]$$

L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage supposé laplacien. L'empiétement des intervalles de deux valeurs propres consécutives suggérera donc l'égalité de ces valeurs propres. Les axes correspondants sont alors définis à une rotation près. Ainsi l'utilisateur pourra éviter d'interpréter un axe instable selon ce critère. Il existe des généralisations de ces résultats asymptotiques au cas non normal (Waternaux, 1976; Davis, 1977), mais leur utilisation n'est guère pratique[1].

Les intervalles de confiance d'Anderson concernent en pratique aussi bien les valeurs propres des matrices des covariances que des matrices de corrélations. Les simulations entreprises montrent que les intervalles de confiance obtenus

---

[1] On trouvera une revue de résultats asymptotiques relatifs à l'analyse en composantes principales dans Muirhead (1982), Anderson (1958 - seconde édition : 1984), Flury (1988), Pousse (1992).

sont en général "prudent" : le pourcentage de couverture de la vraie valeur est le plus souvent supérieur au seuil de confiance annoncé (Morineau, 1983).

Dans tous les cas, la nature asymptotique des résultats et l'hypothèse sous-jacente de normalité[1] font considérer les résultats comme indicatifs. On trouvera un exemple d'intervalles de confiance d'Anderson en section 3.5.2.

3.4.4. Bootstrap pour l'analyse en composantes principales

Les résultats fournis par les méthodes factorielles ne sont pas des assertions, mais des représentations, c'est-à-dire des objets complexes, auxquels s'appliquent mal les différentes techniques de mesure d'information usuelles en statistique.

Comment valider une forme observée dans un plan factoriel ?

- Par des procédures externes, analogues à celles mentionnées pour le choix du nombre d'axes : connaissances a priori, positionnement de variables supplémentaires.
- Par des calculs de stabilité adaptés (exploration d'un voisinage des données construit à partir des erreurs de mesure ou de réponses).
- Par des calculs de zones de confiance pour les positions des points-lignes et des points-colonnes. Ces calculs peuvent être analytiques, fondés sur des hypothèses probabilistes, ou au contraire, fondés sur les techniques de rééchantillonnage dont le principe a été évoqué au chapitre 1.

On commencera par présenter le cadre de l'utilisation des simulations pour le calcul de ces zones de confiance. Dans ce cadre, le bootstrap, qui constitue une méthode de simulation non paramétrique d'une grande souplesse, jouera un rôle de premier plan.

a – Premier travaux

L'analyse en composantes principales est le domaine d'application qui a donné lieu au plus grand nombre de travaux, utilisant notamment des méthodes de rééchantillonnage antérieures au bootstrap, comme la validation croisée et ses variantes. C'est ainsi que Wold (1978), puis Eastment et Krzanowski (1982), Krzanowski (1987) proposent des méthodes de validation croisée (cf. § 7.3.4) pour déterminer la dimension de l'espace de représentation. Ces auteurs utilisent la théorie de la perturbation pour alléger les calculs (l'omission d'un ou plusieurs éléments, qui est à la base de la validation croisée, est considérée comme une perturbation du tableau de données (cf. § 1.4), et une formule approchée permet d'éviter de refaire une analyse complète). Besse et Ferré

---

[1] Muirhead (1982) a montré que l'hypothèse d'existence des quatre premiers moments pour la loi théorique de l'échantillon suffisait pour valider ces intervalles.

(1993) ont montré que la réitération de ces approximations revenait en fait à utiliser le critère classique de la part de variance expliquée par les axes.

Si l'on excepte les travaux de Gifi (1981) qui concernent plus spécifiquement l'analyse des correspondances (le principe du bootstrap est en fait sensiblement différent selon les méthodes factorielles) les premiers travaux d'application du bootstrap à la validité des résultats en analyse en composantes principales sont ceux de Diaconis et Efron (1983), Stauffer *et al.* (1985), Holmes (1985), Beran et Srivastava (1985), Daudin *et al.* (1988), Holmes (1989), qui construisent des intervalles de confiance pour les valeurs propres et les composantes, ou étudient les propriétés asymptotiques des intervalles ou des estimations obtenus.

b – Diverses possibilités de bootstrap

Il existe plusieurs variantes possibles quant au choix de l'espace factoriel commun. Holmes (1985) applique une méthode d'analyse conjointe de tableaux (méthode STATIS, cf. L'Hermier des Plantes, 1976; Lavit, 1988) au tableau initial et à l'ensemble de ses réplications.

– ***Le bootstrap partiel***

Selon nous, les axes principaux de la matrice de corrélation initiale, calculés sur les données originales non perturbées, doivent jouer un rôle privilégié (la matrice des corrélations initiale **C** est en effet l'espérance mathématique des matrices $\mathbf{C}_k$ « perturbées » par la réplication k). Pourquoi calculer des sous-espaces de représentation prenant en compte des perturbations, et donc moins exacts que le sous-espace calculé sur les données initiales? La variabilité bootstrap s'observe mieux sur le repère fixe initial, non perturbé. Cette technique de bootstrap que l'on appellera *bootstrap partiel* (sans recalcul des valeurs propres) proposée notamment par Greenacre (1984) dans le cadre de l'analyse des correspondances, répond à la plupart des préoccupations des utilisateurs dans le cas de l'analyse en composantes principales[1].

L'algorithme qui nous paraît le mieux adapté pour les intervalles de confiance est analogue à celui évoqué en section 1.4 du chapitre 1 : Une réplication consiste en un tirage avec remise des n individus (vecteurs-observations), suivi du positionnement des p nouvelles variables ainsi obtenues en "variables supplémentaires" sur les q premiers axes de l'analyse de base.

Les procédures décrites ci-dessus peuvent être mises en oeuvre avec un programme classique de projection d'éléments supplémentaires.

En analyse en composantes principales normée, par exemple, on peut, de façon équivalente, utiliser le fait que la coordonnée d'une variable sur un axe factoriel

---

[1] Pour une discussion du bootstrap partiel en analyse en composantes principales, cf. Chateau et Lebart (1996).

n'est autre que son coefficient de corrélation avec la variable "coordonnées des individus sur l'axe" (cf. § 3.2.2). On calcule donc les réplications de ce coefficient, ce qui revient à repondérer les individus avec les "poids Bootstrap" *0, 1, 2, ...* qui caractérisent un tirage avec remise. On obtient, comme sous-produit, des réplications de la variance sur l'axe, qui sont évidemment distinctes de ce que seraient des réplications des valeurs propres.

Le bootstrap partiel se fonde donc sur la projection en tant qu'*éléments supplémentaires* des points répliqués sur les sous-espaces de référence fournis par les axes principaux de la matrice de corrélation **C=X'X**, provenant de l'échantillon initial. Rappelons l'équation [3.2 - 3] de la section 3.2 :

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}' \psi_\alpha$$

formule dans laquelle la coordonnée factorielle $\varphi_{\alpha j}$ d'un point-variable j sur l'axe α n'est autre que le coefficient de corrélation de la variable j avec les coordonnées des individus sur l'axe α :

$$\varphi_{\alpha j} = \sum_{i=1}^{n} \left(\frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}}\right)\left(\frac{\psi_{\alpha i}}{\sqrt{\lambda_\alpha}}\right)$$

La projection de la $k^{ème}$ réplication des p variables de $\mathcal{R}^n$ est maintenant donnée par le vecteur $\varphi_\alpha(k)$ dans $\mathcal{R}^p$ tel que :

$$\varphi_\alpha(k) = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}' \mathbf{D}_k \psi_\alpha$$

et $\mathbf{D}_k$ désigne la matrice diagonale (n, n) des n *poids bootstrap* associée à la $k^{ème}$ réplication.

Les s réplications étant projetées sur un repère commun (celui de l'analyse initiale), on caractérisera graphiquement la dispersion des réplications d'une variable donnée soit par l'enveloppe convexe de l'ensemble de ses réplications, soit par un ellipsoïde d'ajustement du nuage des réplications, qui résultera en fait d'une petite analyse en composantes principales des ce dernier nuage. L'enveloppe convexe a l'avantage de l'exhaustivité (toutes les réplications sans exception sont enveloppées), l'ellipsoïde a l'avantage de prendre en compte la densité du nuage des réplications, et d'être moins sensible à d'éventuelles rares réplications aberrantes. L'exemple d'application 3.5.2 comporte des exemples de tels tracés.

– ***Le bootstrap total***

Le *bootstrap total* consiste à réaliser autant d'analyses en composantes principales qu'il y a de réplications. Mais le système d'axes n'est plus le même d'une analyse à une autre. Il peut y avoir des changements de signes (les axes factoriels sont définis aux signes près), des interversions d'axes, des rotations

d'axes[1]. Il faut donc procéder à une série de transformations afin de retrouver des axes homologues au cours des diagonalisations successives des s matrices de corrélation répliquées $\mathbf{C}_k$ ($\mathbf{C}_k$ correspond à la $k^{ème}$ réplication).

Les trois types de transformations possibles, conduisant à trois types de tests de stabilité, sont :

1. Bootstrap total de type 1 (épreuve trop sévère, très pessimiste) : simple changement (éventuel) de signe des axes homologues pour les réplications. Il s'agit de remédier au fait que les axes sont définis au signe près. Un simple produit scalaire entre axes principaux initiaux et axes principaux répliqués de mêmes rangs permet de rectifier le signe de ces derniers.
2. Bootstrap total de type 2 (épreuve sévère, plutôt pessimiste) : changement de signe et correction des interversions d'axes. Les axes répliqués sont affectés (séquentiellement, sans remise en cause d'affectations antérieures) du rang des axes originaux avec lesquels ils sont les plus corrélés en valeur absolue. Puis on procède à un éventuel changement de signe des axes, comme en bootstrap de type 1.
3. Bootstrap total de type 3 (épreuve plutôt laxiste si on s'intéresse à la stabilité des axes, mais apte à décrire la stabilité des sous-espaces de dimension supérieure à 1) : une rotation dite procrustéenne (voir chapitre 8, § 8.3.2) permet de rapprocher de façon optimale les systèmes d'axes répliqués et les systèmes d'axes initiaux.

Le bootstrap total de type 1 ignore les possibles interversions d'axes et rotations d'axes. Il permet de valider des structures stables et robustes. Chaque réplication doit produire les axes initiaux avec les mêmes rangs (ordre des valeurs propres).

Le bootstrap total de type 2 est idéal si on veut valider des axes, c'est-à-dire des dimensions cachées, sans attacher une importance particulière aux rangs de celles-ci.

Enfin le bootstrap de type 3 permet de valider globalement un sous-espace engendré par les axes principaux correspondant aux premières valeurs propres. Si par exemple le sous-espace des quatre premiers axes répliqués coïncide avec celui des quatre premiers axes initiaux, on pourra trouver une rotation dans $\mathcal{R}^4$ qui fera coïncider les axes (ce qui nous ramène au cas du bootstrap partiel).

Comme le bootstrap partiel, le bootstrap total de type 3 peut être qualifié de laxiste par les utilisateurs qui s'intéressent à l'individualité des axes, et pas seulement aux sous-espaces engendrés par plusieurs axes consécutifs. On peut effectivement discuter le principe de compensation des rotations accidentelles, qui sont la cause de l'instabilité des axes. L'exemple du paragraphe 3.5.2 va illustrer ces différentes situations.

---

[1] Cf. Milan et Whittaker (1995).

– ***Le bootstrap sur variables***

Une telle procédure n'a de sens que si les variables sont assez nombreuses et si il existe un « univers des variables » pour lequel la notion de « tirage de variables » a un sens. Ceci peut arriver dans des circonstances où les variables sont par exemple des événements nombreux, des instants, des zones échantillonnées, où, comme dans le cas de l'exemple du paragraphe 3.5.2, des mots.

Classiquement, les réplications sont obtenues par des tirages avec remises dans l'ensemble des n individus. Pour tester la stabilité des structures vis-à-vis de l'ensemble des variables, nous proposons de répliquer l'ensemble des variables lui-même par la méthode du *bootstrap total*.

Nous supposons ainsi implicitement que l'ensemble des variables actives constitue un échantillon de m variables extrait aléatoirement d'un ensemble de variables potentielles.

Nous cherchons à perturber cet échantillon de variables selon les mêmes principes que le *bootstrap* opéré sur les individus.

Pour cela, on appelle $\mathbf{B_k}$ la matrice diagonale (p, p) dont les éléments diagonaux sont les p poids des variables de la *k-ème* réplication bootstrap (*1,0,2,0,...*). La matrice $\mathbf{X}$ d'ordre (n, p) initiale étant supposée centrée, la matrice à diagonaliser est, dans cet espace, la matrice $\mathbf{T_k}$ qui vaut :

$$\mathbf{T_k} = \mathbf{X}\mathbf{B_k}\mathbf{X}' = \mathbf{X}\mathbf{B_k}^{1/2}\mathbf{B_k}^{1/2}\mathbf{X}'$$

On obtient donc :

$$\mathbf{X}\mathbf{B_k}\mathbf{X}'\mathbf{v_q}(k) = \lambda_q\mathbf{v_q}(k)$$

en multipliant chaque terme par $\mathbf{B_k}^{1/2}\mathbf{X}'$ on a :

$$\mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{X}\mathbf{B_k}^{1/2}\mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{v_q}(k) = \lambda_q\mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{v_q}(k)$$

et en posant $\mathbf{u_q}(k) = \mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{v_q}(k)$ alors :

$$\mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{X}\mathbf{B_k}^{1/2}\mathbf{u}(k) = \lambda_q\mathbf{u}(k)$$

La matrice $\mathbf{T_k} = \mathbf{X}\mathbf{B_k}^{1/2}\mathbf{B_k}^{1/2}\mathbf{X}'$ a les mêmes valeurs propres non nulles que la matrice $\mathbf{B_k}^{1/2}\mathbf{X}'\mathbf{X}\mathbf{B_k}^{1/2}$.

On diagonalisera cette dernière matrice de dimension (p, p)

En pratique, on remplace les *poids bootstrap* nuls (éléments diagonaux de $\mathbf{B_k}$) par des poids infinitésimaux (numériquement parlant), de façon à ce que les variables absentes d'une réplication apparaissent quand même avec le statut de variable supplémentaire. Cette épreuve de validation est évidemment très sévère. Le tirage avec remise suscite approximativement, en moyenne, l'abandon de plus d'un tiers des éléments à chaque réplication. La probabilité qu'un point (parmi n) soit absent des n réplications vaut en effet : $p = [(n-1)/n]^n$, quantité dont la limite, lorsque n tend vers l'infini, est : $1/e$ $(= 0.368)$. On trouvera des exemples de *bootstrap sur variables* dans : Lebart *et al.* (2003).

3.5 Deux exemples d'application

Le premier exemple d'application (§ 3.5.1) est celui dont les données nous ont servi de « fil d'Ariane » depuis le début de ce chapitre. Les dimensions modestes (n = 27, p = 16) nous ont permis de publier le tableau de données complet, et d'examiner les résultats numériques relatifs aux deux dimensions.

Le second exemple (§ 3.5.2) est un tableau individuel de notes (n = 300, p = 70) propre à illustrer les différentes options de validation par bootstrap.

3.5.1 Exemple d'application 1

Nous présentons ici l'exemple (cf. tableau 3.5 - 1) relatif aux temps d'activités quotidiennes évoqué en section 3.1.

Le CESP (Centre d'Étude des Supports de Publicité) a relevé, dans son *Enquête Budget-temps Multimédia* de 1991/1992 auprès de *17 665* personnes, des descripteurs de fréquentation de divers médias (radio, télévision, presse) et des temps d'activités quotidiennes (cf. Boeswillwald, 1992).

Ont été également relevées de nombreuses caractéristiques socio-économiques, parmi lesquelles l'âge, le sexe, l'activité, le niveau d'éducation, et le lieu de résidence de ces personnes, ce qui a conduit à créer *96* catégories en croisant ces divers critères.

Nous nous intéressons seulement ici à la sous-population des hommes actifs, soit *27* groupes qui seront, pour cet exemple, les "individus".

On cherche à connaître les associations entre les temps consacrés à différentes activités par les "individus" observés et à étudier les liens entre ces familles d'activités et les caractéristiques de base des individus.

Enfin, on se propose d'étudier le lien entre les activités quotidiennes et la fréquentation de divers médias (presse, radio, télévision, cinéma). Pour ce faire, on fera intervenir les caractéristiques socio-économiques (variables nominales) et les habitudes de fréquentation des médias (variables numériques continues) en tant que variables supplémentaires.

L'analyse du tableau de données (tableau 3.5 - 1) nous conduit tout d'abord à calculer les paramètres descriptifs élémentaires regroupés dans le tableau 3.5 - 3.

Les moyennes et écarts-types vont servir à transformer les variables de base et n'interviendront plus directement dans la suite. Il importe donc de prendre connaissance de ces mesures de niveau et de dispersion. Les valeurs extrêmes sont également utiles pour apprécier la qualité de l'information recueillie.

Le tableau 3.5-2 donne les moyennes, écarts-types et valeurs extrêmes relatifs aux variables continues supplémentaires.

Lecture du tableau 3.5 - 1

(*16* variables continues actives)

Les *27* "individus" (qui sont en réalité dans le cadre de cet exemple des groupes d'individus) sont repérés par un identificateur en 4 caractères :

- le 1er caractère est l'âge du groupe (1=jeune, 2=moyen, 3=âgé)
- le 2ème caractère est ici toujours égal à 1 (car il s'agit ici d'une sélection d' hommes actifs)
- le 3ème est le niveau d'éducation (1=primaire, 2=secondaire, 3=supérieur)
- le 4ème est le type d'agglomération (1=communes rurales; 2=villes moyennes; 3=villes importantes; 4=agglomération parisienne; 5,6,7 = groupes mixtes).

(On trouvera des libellés plus détaillés des variables dans le tableau 3.5 - 2 ci-après.)

On lit par exemple sur la première ligne du tableau 3.5 - 1 que le groupe '1111' (jeunes, actifs, peu instruits, ruraux) consacre en moyenne par jour 463.8 minutes au "sommeil", 23.8 minutes à des activités regroupées sous la rubrique "repos", 107.3 minutes pour les "repas chez soi", etc.

Pour le thème "budget-temps", trois variables seront projetées a posteriori : autres activités, total des activités à domicile, total des activités déclarées en déplacement, ces deux dernières étant des regroupements de variables actives.

Pour le thème "fréquentation média" (qui donne lieu à une mesure de durée globale au niveau des variables actives) six variables décrivent les intensités de contacts avec le cinéma, la radio, la télévision, les presses quotidiennes et magazines, en isolant dans celle-ci les hebdomadaires dits "News".

La matrice des corrélations (tableau 3.5 - 3) nous fournit des éléments de description des associations entre variables actives. Sa lecture nous donne une première idée du réseau d'interrelations existant entre les variables, mais l'analyse en composantes principales va permettre d'obtenir une synthèse de ces liaisons.

Le premier résultat est constitué par la liste des valeurs propres et des pourcentages de variance (cf. tableau 3.5 - 3). La somme des valeurs propres est égale au nombre de variables soit *16*.

Les deux premiers axes fournissent presque la moitié de l'inertie (47%) mais l'on sait que ces quantités sont d'interprétation délicate. On note cependant, à la vue de l'histogramme, qu'il existe une concentration nette du nuage dans un sous-espace à deux dimensions, le plan factoriel principal.

Tableau 3.5 - 1 : Budget-temps agrégé quotidien de $p = 16$ activités (colonnes) pour $n = 27$ groupes d'hommes actifs (lignes)

IDENT	Somm	Repo	Reps	Repr	Trar	Ména	Visi	Jard	Lois	Disq	Lect	Cour	Prom	A pi	Voit	Fréq
1111	463.8	23.8	107.3	4.8	300.0	21.3	51.0	82.3	10.0	1.2	.0	41.3	6.9	7.1	52.1	135.8
1115	515.6	58.5	102.7	10.4	208.8	41.9	30.0	32.9	2.1	4.6	.6	33.7	8.3	24.6	29.4	225.8
1121	463.3	34.2	84.8	17.1	298.3	18.1	37.8	55.8	18.4	5.9	2.6	30.7	5.9	8.8	56.7	135.8
1122	456.4	43.1	74.2	21.9	239.0	26.0	51.2	59.7	18.4	3.6	4.6	52.2	9.5	10.8	72.7	142.3
1123	478.0	44.2	76.7	15.2	212.3	22.3	42.0	43.7	18.4	2.3	6.4	48.3	14.7	15.5	72.8	167.7
1124	465.1	41.6	85.2	23.7	226.0	37.0	42.5	16.3	10.7	8.7	9.4	44.3	13.7	19.8	59.0	145.1
1136	458.4	47.4	94.7	15.1	314.3	25.3	39.1	42.4	16.9	.9	16.7	34.5	4.6	6.4	61.5	103.4
1133	457.2	30.7	82.0	26.2	269.8	52.1	37.6	35.6	25.6	6.0	8.0	42.8	10.4	12.0	81.4	107.6
1134	465.2	40.2	78.6	31.1	268.6	36.3	21.6	4.0	19.4	6.0	14.8	46.9	10.7	21.9	48.3	82.4
2111	449.0	42.1	86.2	7.9	312.5	15.1	16.1	112.9	15.4	.0	2.2	32.1	7.6	8.1	60.1	153.9
2112	450.2	63.1	86.7	9.8	249.6	40.4	55.6	83.3	3.0	2.2	.0	45.0	9.4	10.4	61.9	145.4
2117	455.2	47.4	95.6	9.0	250.8	30.4	13.5	57.3	7.9	2.9	7.0	52.2	15.1	15.7	49.1	194.8
2121	461.9	39.3	90.3	8.5	323.5	14.9	21.7	81.8	15.4	1.2	5.3	26.0	3.8	7.4	59.6	130.8
2122	453.7	44.7	97.5	18.7	269.0	23.1	39.6	93.5	3.1	3.4	12.1	42.0	12.1	10.6	62.4	129.1
2123	433.1	49.8	91.7	12.6	283.7	22.4	21.0	62.9	13.1	6.2	7.3	38.1	11.6	11.7	47.6	168.6
2124	438.3	32.8	102.3	11.1	338.3	28.0	6.5	64.8	13.8	1.4	19.8	34.9	7.4	14.1	53.2	130.5
2131	457.7	44.0	87.9	6.9	313.0	24.4	23.2	63.8	9.2	.6	11.8	30.0	7.3	7.5	69.7	108.3
2132	455.0	47.0	78.9	31.6	380.6	23.9	7.5	40.0	13.0	.0	10.3	23.3	1.4	9.4	59.4	100.0
2133	467.3	37.5	86.9	21.9	264.0	40.8	27.6	33.4	11.9	1.6	10.8	45.3	6.7	10.7	72.8	135.2
2134	433.5	35.6	76.1	17.1	355.0	34.1	13.4	31.7	12.6	3.2	13.2	37.5	8.5	22.3	57.5	96.5
3116	473.0	51.5	99.3	6.3	356.3	21.2	27.6	82.1	8.6	.0	1.5	35.7	13.4	7.1	40.6	107.7
3117	461.9	60.0	103.7	9.1	240.5	35.3	14.5	83.4	1.4	2.0	7.4	46.1	5.7	16.6	53.3	183.7
3121	453.4	45.6	86.2	7.8	358.7	12.9	18.5	54.4	4.2	.0	4.9	34.3	3.3	10.3	48.7	143.1
3122	485.1	53.5	86.0	.3	222.4	24.7	23.2	91.9	8.5	.0	3.7	52.9	7.1	9.9	75.3	166.3
3123	456.7	43.2	94.6	12.1	265.3	30.5	23.7	61.1	9.1	2.3	11.6	50.1	17.6	13.2	46.3	185.3
3136	444.2	53.6	90.7	7.2	302.4	31.7	16.4	97.6	4.7	2.4	4.3	38.8	13.6	11.4	61.8	127.2
3137	438.4	50.7	81.0	11.2	306.6	19.3	23.8	10.5	13.6	.0	18.4	67.6	8.3	18.6	63.1	143.3

Tableau 3.5 - 2
Statistiques sommaires des variables continues
effectif total : 27

IDEN – LIBELLE	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
Variables actives				
Somm – Sommeil	458.91	16.47	433.10	515.60
Repo – Repos	44.63	8.90	23.80	63.10
Reps - Repas chez soi	89.18	8.90	74.20	107.30
Repr - Repas restaurant	13.87	7.82	.30	31.60
Trar - Travail rémunéré	286.27	46.75	208.80	380.60
Ména - Ménage	27.90	9.29	12.90	52.10
Visi - Visite à amis	27.64	13.26	6.50	55.60
Jard - Jardinage, Bricolage	58.49	27.39	4.00	112.90
Lois - Loisirs extérieur	11.42	5.95	1.40	25.60
Disq - Disque cassette	2.54	2.32	.00	8.70
Lect - Lecture livre	7.95	5.47	.00	19.80
Cour - Courses démarches	40.99	9.47	23.30	67.60
Prom - Promenade	9.06	3.88	1.40	17.60
A pi - Déplacement à pied	12.66	5.01	6.40	24.60
Voit - Déplacement en Voiture	58.38	11.29	29.40	81.40
Fréq - Fréquentation Média	140.58	32.56	82.40	225.80
Variables continues supplémentaires				
Autr – Autres activités	12.71	5.70	2.10	25.90
Domi – Total Domicile	928.73	49.92	826.00	1034.00
Tdep – Total Déplacement	88.45	14.65	67.50	122.10
Habitudes Cinema	.14	.14	.00	.60
Habitudes Radio	1.92	.23	1.49	2.64
Habitudes Télévision	3.20	.37	2.13	3.90
Habitudes Presse Quotidienne	.18	.14	.03	.53
Habitudes Presse magazine	3.56	.74	2.00	5.31
Habitudes Hebdomadaires News	.31	.18	.00	.67

On lira sur le tableau 3.5 - 4 les coordonnées des points variables sur les trois premiers axes ainsi que les coordonnées des extrémités des axes unitaires (cf. § 3.3.2) destinés à une éventuelle représentation simultanée des individus et des variables. Les deux premières valeurs propres étant voisines (*3.871* et *3.660*), leurs racines carrées le sont également (*1.97* et *1.91*) et donc les nuages bidimensionnels des points variables et des anciens axes unitaires auront des allures très voisines.

Tableau 3.5-3 : Matrice des corrélations, et valeurs propres correspondantes

Matrice des corrélations

Sommeil	1.00														
Repos	.21	1.00													
Repas c.	.21	.10	1.00												
Repas r.	−.08	−.30	−.53	1.00											
Travail	−.52	−.28	−.02	−.01	1.00										
Ménage	.20	.08	−.01	.39	−.46	1.00									
Visites	.27	−.08	−.07	.10	−.47	.15	1.00								
Jardin.	−.09	.19	.43	−.64	.08	−.37	−.02	1.00							
Loisirs	−.17	−.61	−.55	.52	.10	−.01	.12	−.39	1.00						
Disques	.07	−.17	−.15	.52	−.46	.50	.30	−.42	.25	1.00					
Lecture	−.44	−.21	−.15	.38	.24	.08	−.36	−.51	.27	−.01	1.00				
Courses	−.04	.18	−.17	−.03	−.56	.23	.24	−.24	−.01	.08	.18	1.00			
Promen.	.00	.09	.04	−.02	−.45	.27	.18	−.01	−.05	.40	−.03	.48	1.00		
A pied	.17	.15	−.14	.28	−.38	.49	−.18	−.62	−.09	.48	.27	.37	.30	1.00	
Voiture	−.19	−.22	−.55	.21	−.15	.10	.27	.03	.44	−.09	.15	.23	−.11	−.33	1.00
Fréq.med	.40	.42	.37	−.44	−.62	.05	.01	.18	−.45	.07	−.38	.30	.28	.28	−.33	1.00
	Somm	Repo	Reps	Repr	Trar	Ména	Visi	Jard	Lois	Disq	Lect	Cour	Prom	A pi	Voit	Fréq

Histogramme des 16 valeurs propres

NUMER.	VALEUR PROPRE	POURCENTAGES	POURCENTAGES CUMULES	
1	3.871	24.20	24.20	**********
2	3.660	22.88	47.07	**********
3	2.006	12.54	59.61	**********
4	1.514	9.47	69.08	**********
5	1.126	7.04	76.12	**********
6	.837	5.23	81.35	**********
7	.766	4.79	86.15	**********
8	.596	3.73	89.87	**********
9	.444	2.78	92.65	**********
10	.374	2.34	94.99	**********
11	.246	1.54	96.53	**********
12	.222	1.39	97.92	**********
13	.161	1.01	98.93	**********
14	.114	.72	99.64	**********
15	.037	.23	99.88	**********
16	.019	.12	100.00	**********

Tableau 3.5 – 4. Coordonnées des variables actives sur les axes 1 à 3

Variables	Coordonnées			Anciens axes unit.		
	1	2	3	1	2	3
Sommeil	.22	-.52	.18	.11	-.27	.13
Repos	.46	-.40	-.17	.23	-.21	-.12
Repas chez soi	.67	-.15	-.23	.34	-.08	-.17
Repas restaurant	-.84	.00	-.07	-.43	.00	-.05
Travail rémunéré	.05	.88	-.34	.03	.46	-.24
Ménage	-.40	-.57	-.08	-.20	-.30	-.06
Visite à amis	-.13	-.33	.73	-.07	-.17	.52
Jardinage, Bricolage	.76	.22	.35	.39	.11	.25
Loisirs extérieur	-.72	.30	.30	-.37	.16	.21
Disque cassette	-.53	-.53	.01	-.27	-.27	.01
Lecture livre	-.54	.24	-.50	-.27	.12	-.36
Courses démarches	-.21	-.54	.11	-.11	-.28	.08
Promenade	-.10	-.58	.04	-.05	-.30	.03
A pied	-.37	-.62	-.57	-.19	-.33	-.40
En Voiture	-.41	.22	.65	-.21	.11	.46
Fréquentation Média	.49	-.68	-.05	.25	-.36	-.03

La figure 3.5 - 1 donne une représentation des variables sur les deux premiers axes factoriels. Les données étant ici centrées réduites, les coordonnées des variables sur les axes sont les coefficients de corrélations entre ces variables et les facteurs.

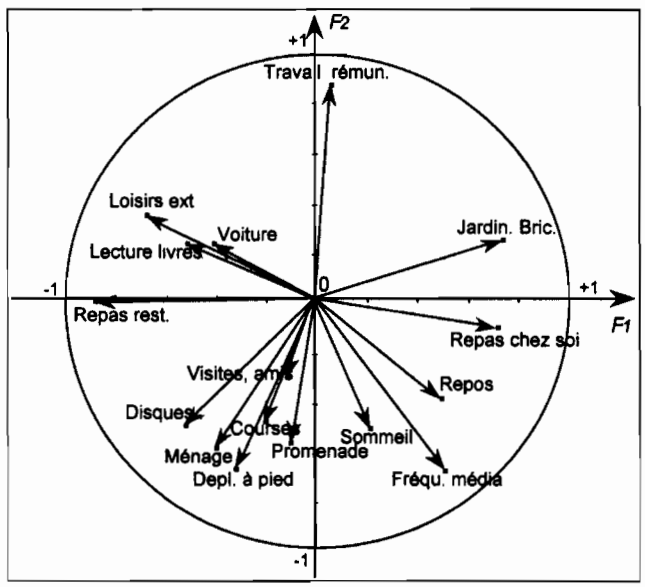


Figure 3.5-1. Représentation des *16* variables actives dans le plan des facteurs 1 et 2

Le premier axe oppose les activités extérieures ou d'ouverture (lecture, loisir extérieur, repas restaurant, déplacement en voiture) à des activités plus intérieures (jardinage, jeux, bricolage, repas chez soi). Le deuxième axe oppose

essentiellement l'activité professionnelle (travail rémunéré) aux activités de temps disponible ou libre (promenade, disque cassette, fréquentation média) mais aussi le temps passé au ménage et au sommeil.

Les variables supplémentaires (tableau 3.5 - 5 et figure 3.5 - 2) relatives aux déplacements et aux médias illustrent ces propos. Les activités "total déplacement" et "total domicile" caractérisent bien le premier axe.

Tableau 3.5 – 5. Coordonnées des variables supplémentaires (ou illustratives) sur les axes 1 à 3

VARIABLES	COORDONNEES		
	1	2	3
Autres activités	.08	.16	.04
Total Domicile	.67	-.50	-.21
Total Déplacement	-.72	.05	.14
Habitudes Cinema	-.87	-.11	-.14
Habitudes Radio	-.27	-.57	.07
Habitudes Télévision	.04	-.55	.34
Habitudes Presse Quot	-.39	.01	-.70
Habitudes Presse mag	-.24	-.38	-.26
Habitudes Hebdo-News	-.46	.20	-.48

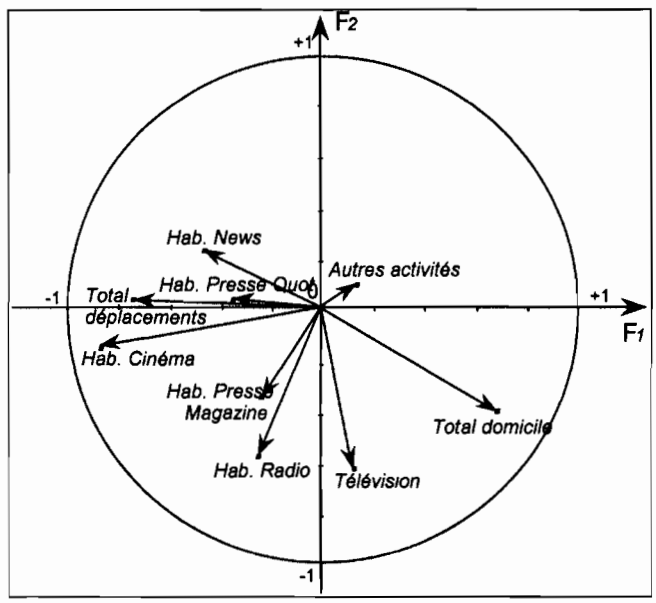


Figure 3.5-2. Positionnement des variables supplémentaires (plan de la figure 3.5-1)

La presse quotidienne et surtout le cinéma sont corrélés aux activités dites d'ouvertures, pour lesquelles le temps passé en déplacement est important. Le temps passé au domicile est pratiquement point moyen des points : activités Repos, Jardinage-bricolage, Repas chez soi, Télévision, qui est le media

dominant en durée. On pressent le rôle de certaines caractéristiques socio-économiques, qui seront positionnées plus loin dans l'espace des individus. Les positions des individus dans le plan factoriel (tableau 3.5 - 6 et figure 3.5 - 3) vont permettre d'expliquer certaines des corrélations observées.

Ainsi, deux groupes (1133 et 1134) se distinguent à l'extrême gauche du premier axe : il s'agit de jeunes actifs instruits des grandes métropoles régionales ou de Paris, qui ont un profil d'activité typé (lecture, repas au restaurant, ...) expliquant à eux deux *35%* de la variance le long de cet axe.

Le groupe 1115 (jeunes peu instruits habitant dans des communes de profils variés) se distingue sur le deuxième axe (contribution de *26%*). Remarquons aussi que ce même groupe a une distance à l'origine des axes (colonne DISTO = carré de la distance à l'origine, point moyen des individus) anormalement élevée (*47.51*) qui confirme son atypicité.

Tableau 3.5 - 6
Coordonnées, contributions et cosinus carrés des individus sur les axes 1 et 2

INDIVIDUS		COORDONNEES		CONTRIBUT.		COS. CARRE	
IDENTIF.	DISTO	1	2	1	2	1	2
1111	19.89	2.01	.85	3.8	.7	.20	.04
1115	47.51	2.26	-5.11	4.9	26.4	.11	.55
1121	10.55	-.71	1.01	.5	1.0	.05	.10
1122	13.29	-1.86	-.64	3.3	.4	.26	.03
1123	14.49	-1.28	-1.81	1.6	3.3	.11	.23
1124	19.06	-2.72	-2.93	7.1	8.7	.39	.45
1136	10.68	-.56	1.97	.3	3.9	.03	.36
1133	27.04	-4.21	-.30	17.0	.1	.66	.00
1134	25.35	-4.29	-.91	17.6	.8	.73	.03
2111	12.86	1.91	2.12	3.5	4.5	.28	.35
2112	17.27	1.43	-1.68	2.0	2.8	.12	.16
2117	10.89	1.03	-2.16	1.0	4.7	.10	.43
2121	10.96	1.27	2.55	1.5	6.6	.15	.59
2122	7.92	.62	-.21	.4	.0	.05	.01
2123	8.33	.30	-.33	.1	.1	.01	.01
2124	15.54	-.12	2.06	.0	4.3	.00	.27
2131	7.39	.55	2.03	.3	4.2	.04	.56
2132	24.45	-1.17	3.53	1.3	12.6	.06	.51
2133	7.85	-1.63	-.11	2.5	.0	.34	.00
2134	17.19	-2.54	1.36	6.2	1.9	.37	.11
3116	16.19	2.68	.96	6.9	.9	.45	.06
3117	15.96	2.43	-1.84	5.7	3.4	.37	.21
3121	13.00	1.90	2.11	3.4	4.5	.28	.34
3122	17.31	2.12	-.95	4.3	.9	.26	.05
3123	10.26	.56	-1.74	.3	3.1	.03	.30
3136	9.09	1.56	.09	2.3	.0	.27	.00
3137	21.68	-1.55	.08	2.3	.0	.11	.00

On vérifie sur le tableau de données 3.5 - 1 que ce groupe a un temps de travail moyen exceptionnellement faible (*208.8*, valeur qui est d'ailleurs le minimum de cette variable donné par le tableau 3.5 - 2) et des temps maxima pour

"déplacement à pied" et "fréquentation média" (il s'agit essentiellement d'écoute télévision).

Souvent, dans les applications en vraie grandeur, les individus sont beaucoup plus nombreux et les identificateurs renvoient en général à un numéro de questionnaire ou d'observation.

Les variables nominales sont alors projetées selon la procédure indiquée au paragraphe 3.3.1.

Le tableau 3.5 - 7 fournit les coordonnées des modalités (ou catégories) de ces variables qui sont, rappelons-le, les centres de gravité des individus concernés.

Ces centres de gravité ont été portés sur la figure 3.5-3 et les modalités contiguës d'une même variable nominale (il s'agit en fait de variables ordinales) ont été jointes par des lignes polygonales.

Dans l'hypothèse où les groupes correspondant à une modalité particulière pourraient être considérés comme tirés au hasard parmi les 27 groupes, ces centres de gravité ne devraient pas s'éloigner beaucoup du centre de gravité du nuage (origine des axes factoriels).

On peut convertir cette distance au centre de gravité en "valeur-test"[1], qui sera alors la réalisation d'une variable normale centrée réduite (deuxième bloc du tableau 3.5 - 7).

Comme on le lit sur le tableau 3.5 - 7, la valeur-test du point "Paris" sur l'axe horizontal est de -2.6. C'est une modalité dont la position est significativement différente de l'origine.

La figure 3.5 - 3, tout comme le tableau 3.5 - 7, montrent que les trois variables nominales permettent surtout d'identifier le premier axe, opposant les jeunes instruits urbains aux personnes plus âgées et moins instruites. Seules les communes rurales (Agglo1) semblent liées au second axe.

Le lecteur de ces graphiques doit garder à l'esprit le fait qu'il s'agit ici d'identification passive par des variables nominales d'une analyse réalisée uniquement à partir des temps d'activité (variables actives).

Il ne s'agit pas d'une étude des liaisons existant entre ces variables nominales, même si certaines proximités peuvent paraître familières.

---

[1] Ces aides à l'interprétation sont abordées dans un cadre plus général à l'occasion de l'analyse des correspondances multiples (chapitre 4). Brièvement, la valeur-test d'une modalité supplémentaire sur un axe est une conversion de la coordonnée de cette modalité sur l'axe en une variable normale, centrée, réduite, dans l'hypothèse d'indépendance. Cette hypothèse stipule ici que les individus concernés par cette modalité sont répartis aléatoirement autour de l'origine de l'axe (il n'est pas nécessaire de préciser plus la distribution à l'origine de cette répartition car la normalité de la valeur-test sera une conséquence du théorème de la limite centrale). Cette hypothèse d'indépendance implique que, pour un test unique, une valeur test a 95 chances sur 100 d'appartenir à l'intervalle [-1.96, +1.96].

Tableau 3.5 – 7. Valeurs-test et coordonnées des modalités supplémentaires sur les axes 1 et 2

Modalités		Valeurs-test		Coordonnées	
Libellé	Effectif	Axe 1	Axe 2	Axe 1	Axe 2
. Age en trois classes					
A-35 - Jeunes	9	-2.3	-1.6	-1.26	-.87
A+35 - Age-Moy	11	.3	1.8	.15	.83
A+50 - Ages	7	2.1	-.3	1.39	-.18
. Niveau d'éducation					
prim - primaire	7	3.0	-1.5	1.96	-.98
seco - secondaire	11	.0	-.2	.01	-.08
supe - superieur	9	-2.8	1.6	-1.54	.86
. Agglomération (*extraits*)					
AGG1 - - de 20 000	6	1.6	2.5	1.15	1.78
AGG2 - de 20 a 100 000	5	.3	.0	.23	.01
AGG3 - Plus de 100 000	5	-1.5	-1.1	-1.25	-.86
AGG4 - Paris	4	-2.6	-.1	-2.42	-.11

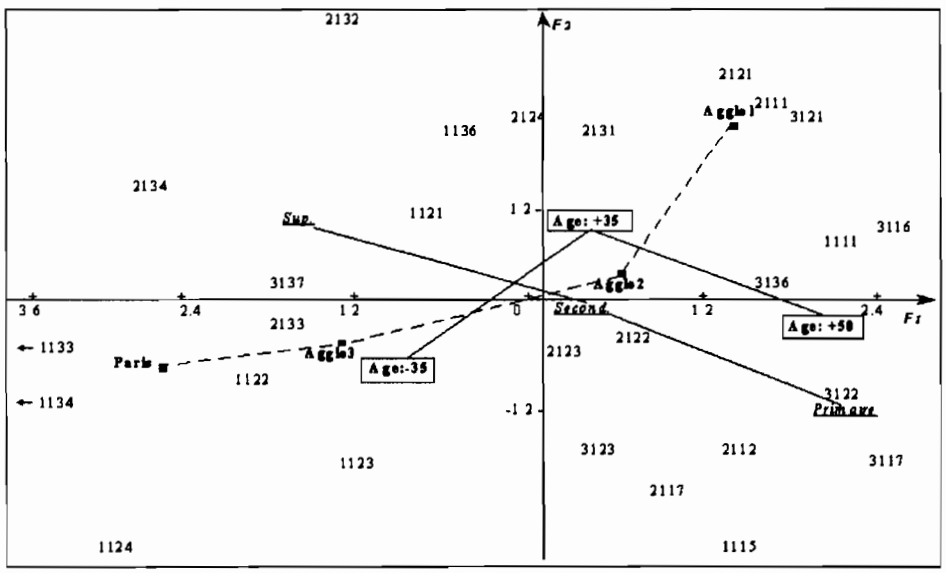


Figure 3.5 – 3. Positionnement des individus (symboles à 4 chiffres) et des variables nominales dans le premier plan de l'ACP

3.5.2 Exemple d'application 2

Le premier exemple portait sur des données agrégées ; le second porte sur des réponses individuelles, à propos d'enquêtes *sémiométriques*.

Le questionnaire de ces enquêtes peut être considéré comme une forme simplifiée de questionnaire « style de vie » car il ne s'agit pas de donner son opinion sur des assertions, mais d'attribuer des notes à des mots en fonction des sensations agréables (note *7*) ou désagréables (note *1*) que provoquent leurs évocations. Le questionnaire comporte ainsi une liste de *210* mots (*e.g.* : *aventure, morale, travail...*) censés être, directement ou indirectement, porteurs de valeur. On va travailler ici sur un ensemble plus restreint de *70* mots, et sur un sous-échantillon de *300* individus[1].

Tableau 3.5-8. Liste de la sélection de *70* mots (notes moyennes entre parenthèses)

Mot	Note moyenne	Mot	Note moyenne	Mot	Note moyenne
absolu	(4.53)	évasion	(5.31)	or	(5.73)
admirer	(5.87)	féconder	(5.78)	paix	(6.80)
âme	(5.19)	feu	(3.98)	partir	(5.68)
animal	(5.84)	fleuve	(5.40)	perfection	(5.67)
armure	(2.97)	fusil	(1.98)	précieux	(5.64)
attachement	(6.03)	gratuit	(5.79)	produire	(5.13)
aventurier	(4.68)	hériter	(5.31)	prudence	(4.82)
bleu	(5.73)	honneur	(5.69)	pureté	(6.02)
campagne	(6.31)	île	(5.60)	raison	(5.50)
certitude	(5.27)	inconnu	(3.96)	réfléchir	(5.55)
charnel	(5.19)	interdire	(2.72)	rêver	(6.26)
commander	(4.34)	inventeur	(5.75)	rigide	(3.26)
confiance	(6.56)	justice	(5.41)	rompre	(2.07)
consoler	(5.75)	livre	(6.05)	sacré	(5.04)
créateur	(5.62)	lune	(5.15)	science	(5.88)
danger	(2.19)	maîtriser	(5.57)	sensuel	(5.83)
désir	(6.10)	matériel	(4.43)	sommet	(5.31)
Dieu	(4.98)	métallique	(3.44)	sublime	(5.82)
douceur	(6.57)	modération	(5.08)	tradition	(5.11)
eau	(6.14)	montagne	(5.84)	utilitaire	(5.07)
écrire	(5.36)	muraille	(2.67)	vide	(2.19)
élégance	(5.92)	nager	(5.69)	vitesse	(4.39)
enfance	(6.09)	noble	(4.73)		
escalader	(4.73)	nudité	(4.73)		

Les résultats de ces notations pour des échantillons représentatifs de la population, ont été soumis à des analyses en composantes principales et ont produit six axes stables (dans le temps : périodes différentes; dans l'espace : pays différents). Alors que le premier axe est surtout un facteur de taille, le plan des axes 2 et 3, stable pour des échantillons importants (de l'ordre de *600*), est appelé « plan sémiométrique » et peut recevoir comme variables

[1] Le tableau de données et le logiciel (DTM) permettant de réaliser tous les traitements évoqués peuvent être librement téléchargés sur le site www.lebart.org.

supplémentaires des consommations de produits, des usages de services, des opinions sur des produits, des services ou des marques[1].

Cet exemple sera pour nous l'occasion de montrer et de discuter les procédures de validations exposées en section 3.4. Il n'est pas question compte tenu du volume des données de publier le tableau de notes, ni même la matrice des corrélation (*70, 70*).

a – Les valeurs propres

Nous publions les *12* premières valeurs propres (sur *70*), accompagnées des pourcentages de variance (ou d'inertie) simples et cumulés (tableau 3.5-9) et de leur histogramme horizontal (tableau 3.5-10) destiné à permettre une appréciation visuelle de leur décroissance.

Tableau 3.5-9. Liste des *12* premières valeurs propres et des pourcentages bruts et cumulés

numéro	Valeur propre	pourcentage	pourcentage cumulé
1	7.63	10.9	10.9
2	4.77	6.8	17.7
3	3.18	4.5	22.2
4	3.14	4.5	26.7
5	2.28	3.2	30.0
6	1.97	2.8	32.8
7	1.82	2.6	35.4
8	1.73	2.4	37.9
9	1.62	2.3	40.2
10	1.51	2.1	42.4
11	1.42	2.0	44.4
12	1.39	1.9	46.4

Si l'on se réfère au critère empirique de Cattell (§ 3.4.2), appelé encore *critère du coude*, la figure 3.5-4 nous suggère de garder quatre, ou peut-être cinq axes.

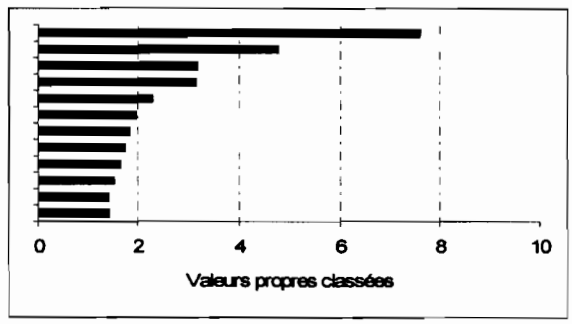


Figure 3.5-4. Esquisse de l'histogramme horizontal des *12* premières valeurs propres

[1] Pour des résultats en vraie grandeur (*210* mots, des milliers d'individus dans plusieurs pays), cf. : *La sémiométrie* (Lebart, Piron *et* Steiner, Dunod, Paris, 2003).

Le critère de Kayser (valeurs propres supérieures à leur moyenne, c'est-à-dire ici à 1) demanderait de garder 22 axes (la $22^{ème}$ valeur propre vaut *1.02*, et la $23^{ème}$ *0.97*).

Tableau 3.5-10. Intervalles de confiance d'Anderson pour les *12* premières valeurs propres

(exemple : La valeur propre 1 a 95 chances sur *100* d'appartenir à l'intervalle [*6.50, 8.95*])

nombre	minimum	valeur propre observée	maximum
1	6.50	7.63	8.95
2	4.06	4.77	5.60
3	2.71	3.18	3.73
4	2.68	3.14	3.69
5	1.94	2.28	2.68
6	1.68	1.97	2.31
7	1.55	1.82	2.13
8	1.47	1.73	2.03
9	1.38	1.62	1.90
10	1.29	1.51	1.77
11	1.21	1.42	1.67
12	1.18	1.39	1.63

Les axes 3 et 4 ont des importances comparables, mais se détachent assez sensiblement de l'axe 5, lui-même peu différent de l'axe 6. A partir du 5, les intervalles de confiance empiètent largement.

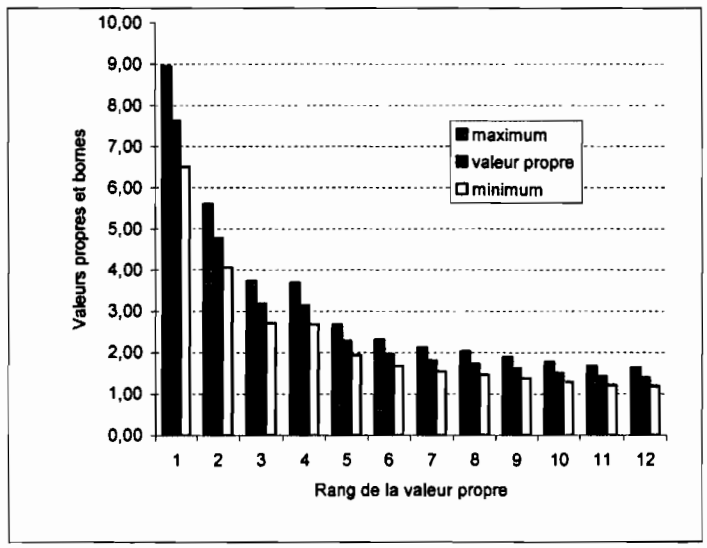


Figure 3.5-5. Intervalles de confiance d'Anderson pour les *12* premières valeurs propres

Les intervalles de confiance d'Anderson du tableau 3.5-10 et de la figure 3.5-5 semblent confirmer le critère de Cattell : les deux premiers axes sont dominants et bien caractérisés (le minimum de l'intervalle de confiance d'une valeur

propre est supérieur au maximum de l'intervalle de la valeur propre consécutive).

b– Les plans factoriels

La figure 3.5-6 qui représente le premier plan factoriel (axe 1 en abscisse) est un exemple de *facteur taille* (cf. § 3.4.1) qui s'avère être ici un effet-notation : certains individus donnent des notes élevées à la majorité des mots, d'autres utilisent systématiquement des notes plus faibles. Les points-variables sont en majorité d'un même côté de l'origine des axes, alors que les répondants sont répartis de façon beaucoup plus équilibrée. L'origine des axes est en effet le point moyen (centre de gravité) des individus, que l'on a pas représentés ici puisqu'ils sont anonymes dans l'enquête. Seuls quelques mots majoritairement mal notés (*rompre, vide, muraille, danger, fusil*) s'opposent à la grande masse des mots.

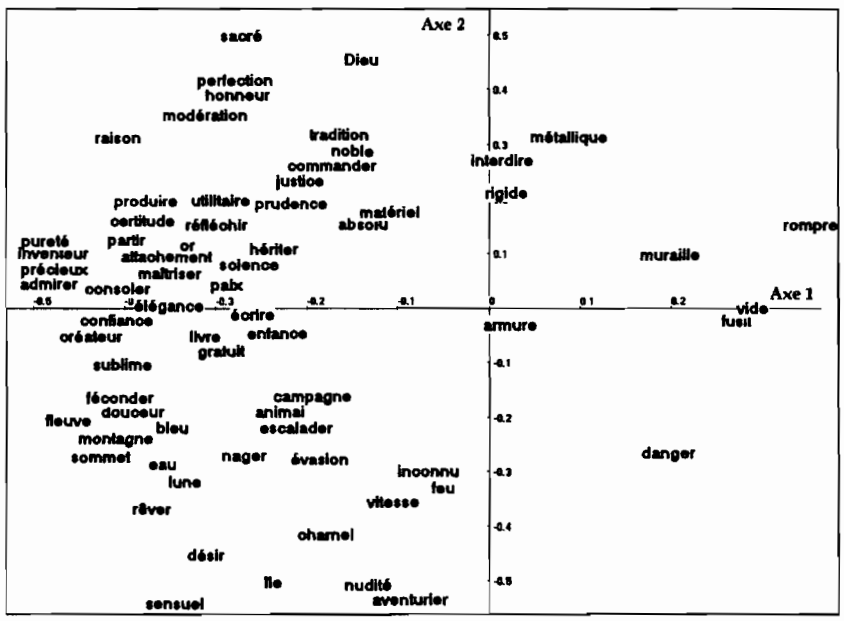


Figure 3.5-6. Plan factoriel (1, 2) de l'analyse de la table (réponses x mots)

Les aspects méthodologiques de l'effet-notation ont fait l'objet d'une étude approfondie dans l'ouvrage précité. Le second axe (vertical) est beaucoup plus riche de sens. Cet axe stable qui oppose les notes de « *sacré, Dieu, perfection, honneur, tradition* » à celles de « *sensuel, aventurier, nudité, île, désir, charnel*» est souvent désigné comme une dimension d'opposition *devoir – plaisir*.

Le cercle des corrélations de rayon *1* n'est pas représenté pour donner au graphique une échelle convenable, car les plus grandes corrélations observées (coordonnées des mots sur les axes) sont de l'ordre de 0.5.

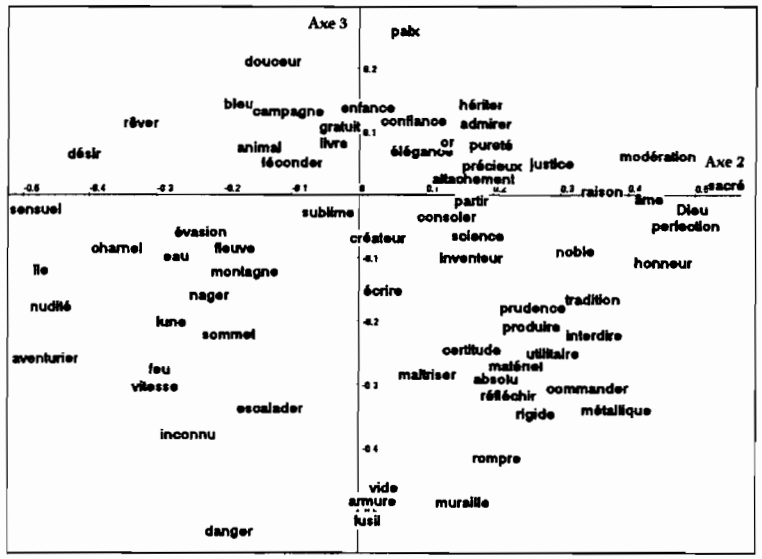


Figure 3.5-7. Plan factoriel (2, 3) de l'analyse de la table (réponses x mots)

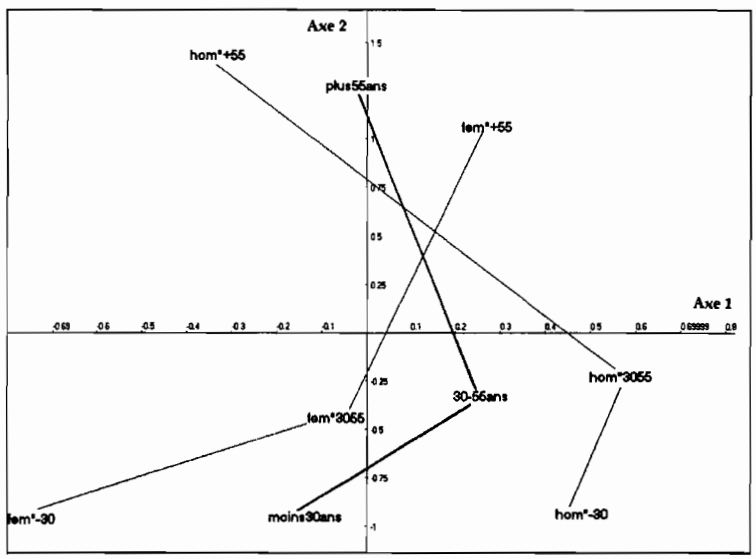


Figure 3.5-8. Positionnement de nominales supplémentaires dans le plan (1, 2)
(Age en trois catégories, sexe-âge en 6 catégories)

La figure 3.5-7 reprend l'axe 2 qui est maintenant en abscisse, en le confrontant à l'axe 3. Ce dernier axe qui oppose ici « *danger, vide, muraille, armure, fusil* » à

« *paix, douceur, campagne...* » est qualifié, lors des études en vraie grandeur portant sur *210* mots et des milliers d'individus, d'axe « *détachement - attachement* ».
La figure 3.5-8 illustre le positionnement de deux variables nominales supplémentaires dans le premier plan factoriel (même plan factoriel que celui de la figure 3.5-6). Il s'agit de l'âge en trois classes (moins de 30 ans, de 30 à 55 ans, plus de 55 ans), et du croisement sexe-âge en six catégories. Ces variables se déploient surtout le long du deuxième axe, les catégories jeunes étant dans la zone du « *plaisir* » alors que les catégories plus âgées du côté du « *devoir* ».

c – Bootstrap partiel pour la position des points- variables

Les considérations empiriques et les intervalles d'Anderson nous ont appris qu'il y avait au moins quatre dimensions principales significatives. Mais à l'intérieur de ces dimensions et des sous-espaces qu'elles engendrent, quelle confiance peut-on accorder aux positions des points ?

La figure 3.5-9 représente sept ellipses de confiance issues d'un bootstrap partiel dans le plan factoriel (1, 2), alors que la figure 3.5-10, sur la même trame et les mêmes points, représente les enveloppes convexes des mêmes réplications bootstrap.

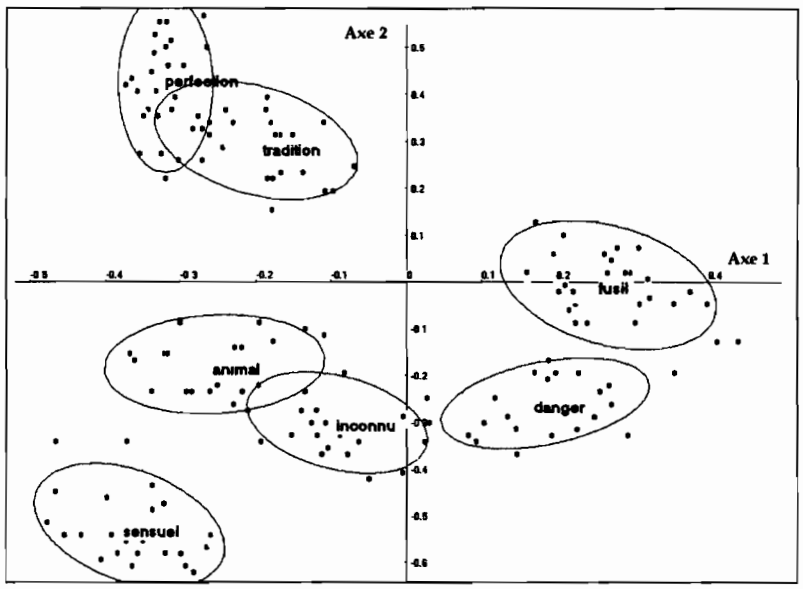


Figure 3.5-9. Bootstrap partiel : Sept zones de confiance elliptiques dans le plan (1, 2) de la figure 3.5-6

Les réplications sont représentées par des points. Conformément à la préconisation du paragraphe 3.4.4.b, les tableaux répliqués (*20*) ont été projetés en tant qu'éléments supplémentaires sur le plan initial de la figure 3.5-6.

Les sept ellipses d'ajustement de la figure 3.5-9 résultent des sept analyses en composantes principales des sept petits tableaux (*20, 2*) [*20* réplications, deux coordonnées sur les deux premiers axes factoriels] [1].

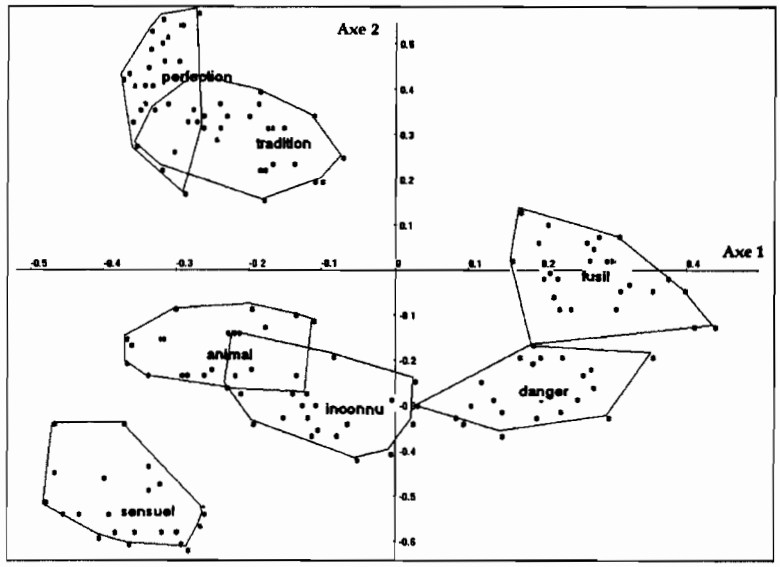


Figure 3.5-10. Bootstrap partiel : Sept zones de confiance (enveloppes convexes) dans le plan (1, 2) de la figure 3.5-6

Les enveloppes convexes de la figure 3.5-10 contiennent, elles, l'intégralité des réplications de chaque point-variable, mais ne prennent pas en compte la densité du sous-nuage des réplications[2]. Qu'elles soient elliptiques ou polygonales, ces zones nous donnent une idée de la précision qu'il faut attacher à la position de chacun des points. Sur cette projection, toutes les positions apparaissent comme significativement distinctes, à l'exception des mots « tradition » et « «perfection » dont les zones empiètent assez largement.

d – Bootstrap total de type 1 pour la position des points-variables

Les résultats du bootstrap total présenté figure 3.5-11 doivent donner, on l'a vu, une mesure très pessimiste de la qualité de la visualisation. En fait, les résultats

---

[1] Pour une variable *j* donnée, l'analyse fournit les axes principaux de l'ellipse d'ajustement des *20* réplications, avec deux valeurs propres $\lambda_1(j)$ et $\lambda_2(j)$. La longueur des axes principaux est quatre fois la racine carrée de ces valeurs propres, ce qui correspond à des intervalles de ± 2 écarts-type sur chaque axe. L'ellipse contient ainsi approximativement *86*% des réplications de la variable *j*.

[2] En fait, nous utilisons ici l'expression zone de confiance en parlant simplement des enveloppes convexes des projections des valeurs répliquées. Les enveloppes convexes, étudiées par Efron (1965) peuvent être "pelées" progressivement de façon à obtenir des estimations non-paramétriques de zones de confiances (cf. Barnett, 1976; Green, 1981; et Holmes, 1985, qui publie également des exemples et les programmes de calcul correspondant).

sont étonnamment proches de ceux du bootstrap partiel, avec des ellipses un peu plus dilatées et un peu plus empiétantes, mais sans apporter d'éléments d'interprétation nouveaux ni de remise en cause profonde du bootstrap partiel.

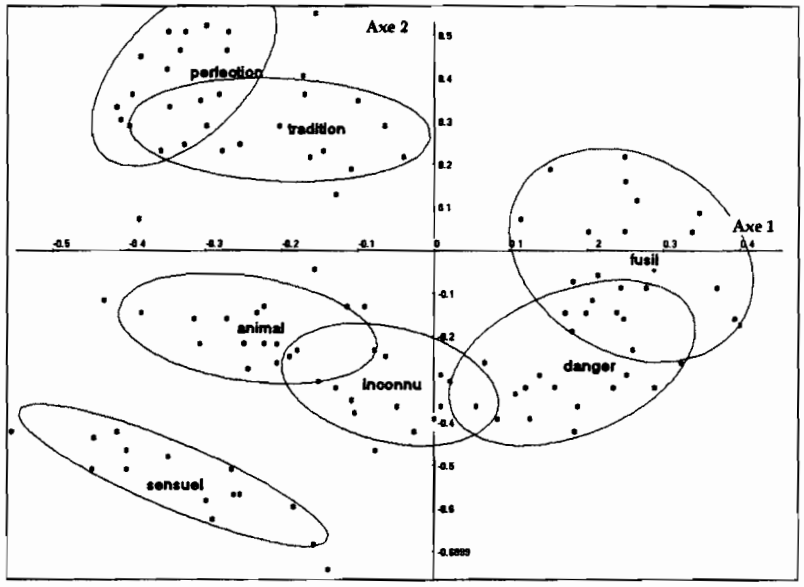


Figure 3.5-11. Bootstrap total : Sept zones de confiance elliptiques des mêmes variables dans le plan (1, 2) de la figure 3.5-6

Si les 20 diagonalisation séparées (avec simples corrections des changements éventuels de signes des axes) des tableaux répliqués du bootstrap total donnent des résultats assez voisins des *20* projections sur le même sous-espace du bootstrap partiel, c'est parce que les deux premières valeurs propres sont significativement distinctes, et très détachées des valeurs propres suivantes (cf. les intervalles de confiance de la figure 3.5-5). Les deux dimensions ont été retrouvées sans interversion pour chacune des *20* réplications.

e – Bootstrap partiel d'un plan instable : le plan (2. 3)

L'histogramme des valeurs propres et les intervalles d'Anderson nous ont montré que les axes 3 et 4, assez bien détachés des axes antérieurs et postérieurs, correspondaient à des valeurs propres voisines numériquement, et indiscernables statistiquement.

Il faut donc s'attendre, lors d'une procédure de bootstrap total, à des confusions entre ces deux axes : une réplication peut voir le $3^{ème}$ axe passer au quatrième rang, ou même remplacé par n'importe quel axe du plan (3, 4). Dans ce plan (3, 4) instable, mais considéré comme fixe par le bootstrap partiel de la figure 3.5-12, les réplications restent proches des points originaux.

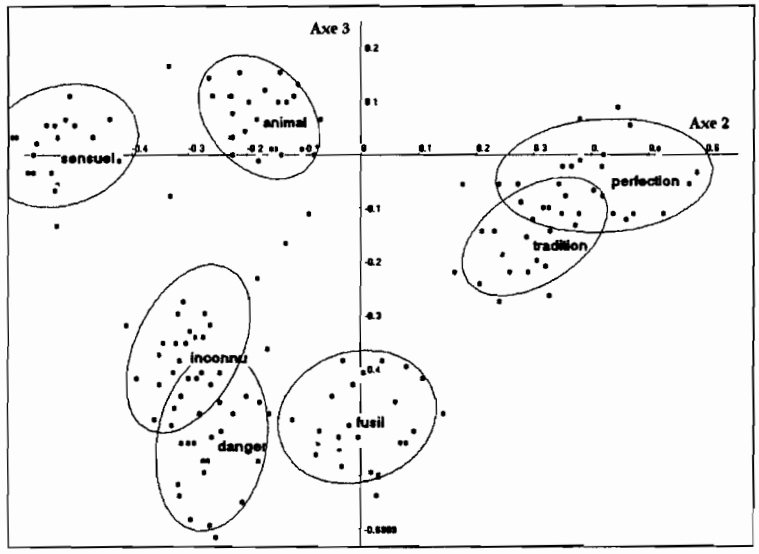


Figure 3.5-12. Bootstrap partiel : Les sept zones de confiance elliptiques des mêmes variables dans le plan (2, 3) de la figure 3.5-7

f – Bootstrap total de type 1 du même plan instable : le plan (2. 3)

En revanche, le bootstrap total de la figure 3.5-13 met en évidence l'instabilité de l'axe 3 le long duquel s'allongent pratiquement toutes les ellipses.

Cette épreuve de validité est évidemment sévère pour l'axe 3 pris isolément. En effet, l'axe 3 et l'axe 4 forment un sous espace plus stable, à une rotation près dans le plan (3, 4). Donc certaines réplications font apparaître l'axe 4 initial avec le rang 3, et vice-versa.

Ce bootstrap de type 1 superpose des axes qui ont même rang mais ne sont pas homologues.

g – Bootstrap total de type 2 du même plan instable : le plan (2. 3)

Cette fois-ci, les interversions d'axes sont prises en compte, ce qui a pour effet de diminuer considérablement la taille des ellipses tracées autour des mots « *fusil* », «*danger*», « *inconnu* », qui occupent un pôle du troisième axe.

Il existe bien une dimension stable opposant [entre autres] ces trois mots au mot « *animal* ». Cette dimension apparaît tantôt sur le troisième axe, tantôt sur le quatrième.

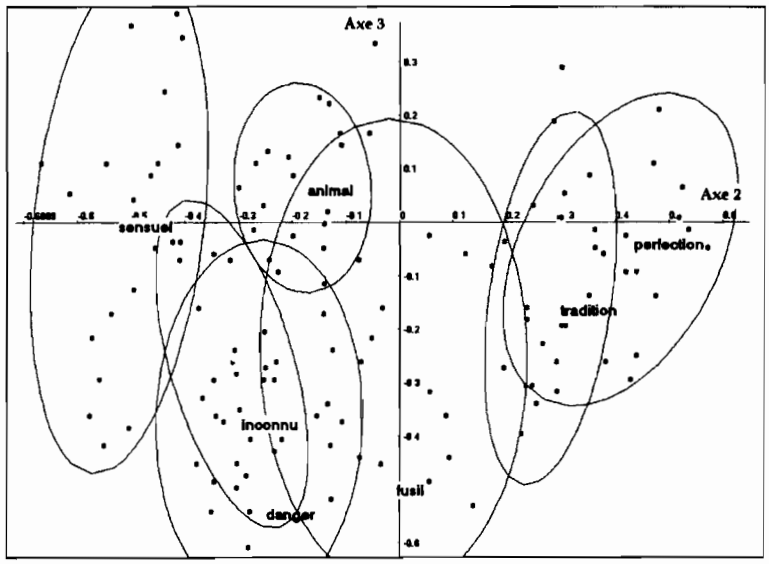


Figure 3.5-13 Bootstrap total de type 1 : Les sept zones de confiance elliptiques des mêmes variables dans le plan (2, 3) de la figure 3.5-7

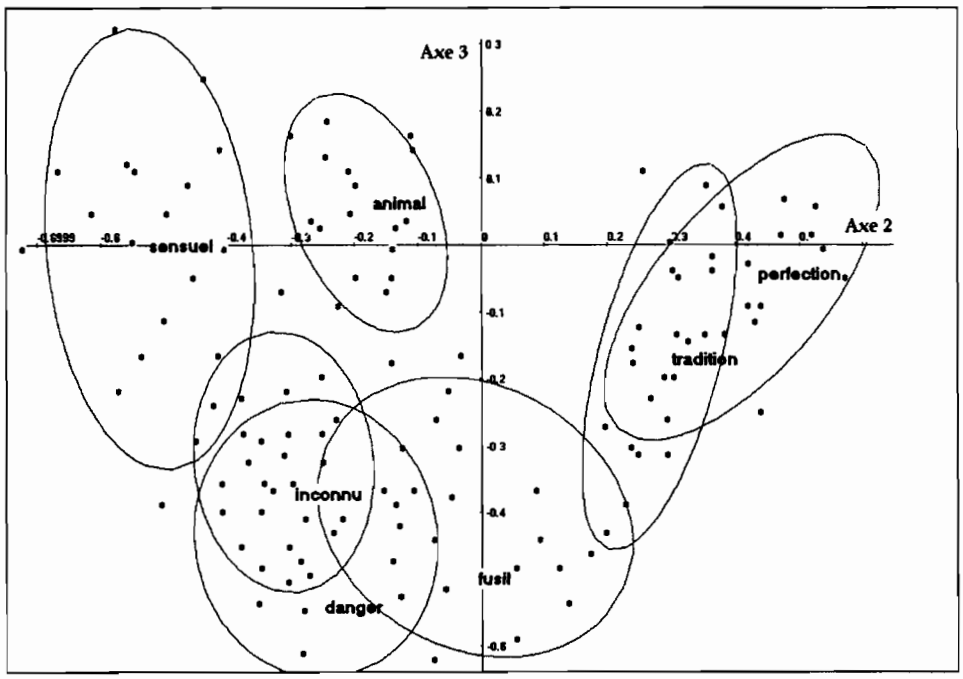


Figure 3.5-14 Bootstrap total de type 2 : avec corrections des seules interversions d'axes: Les sept zones de confiance elliptiques des mêmes variables dans le plan (2, 3) de la figure 3.5-7

h – Bootstrap total de type 3 du même plan instable : le plan (2. 3)

Pour ce type de bootstrap, une rotation est calculée pour chaque réplication de façon à la rapprocher le plus possible de l'échantillon initial. Cette opération prend aussi en compte les changements de signes et les interversions d'axes. La figure 3.5-15 ressemble beaucoup à la figure 3.5-12 qui correspondait au bootstrap partiel, avec toutefois des formes d'ellipses qui peuvent varier. On peut en conclure que l'espace des premiers axes représentés est stable pour l'ensemble des réplications. Les proximités entre les points (qui s'interprètent, rappelons-le, en termes de corrélations) sont respectées d'une réplication à une autre. Il est donc licite d'interpréter ces proximités.

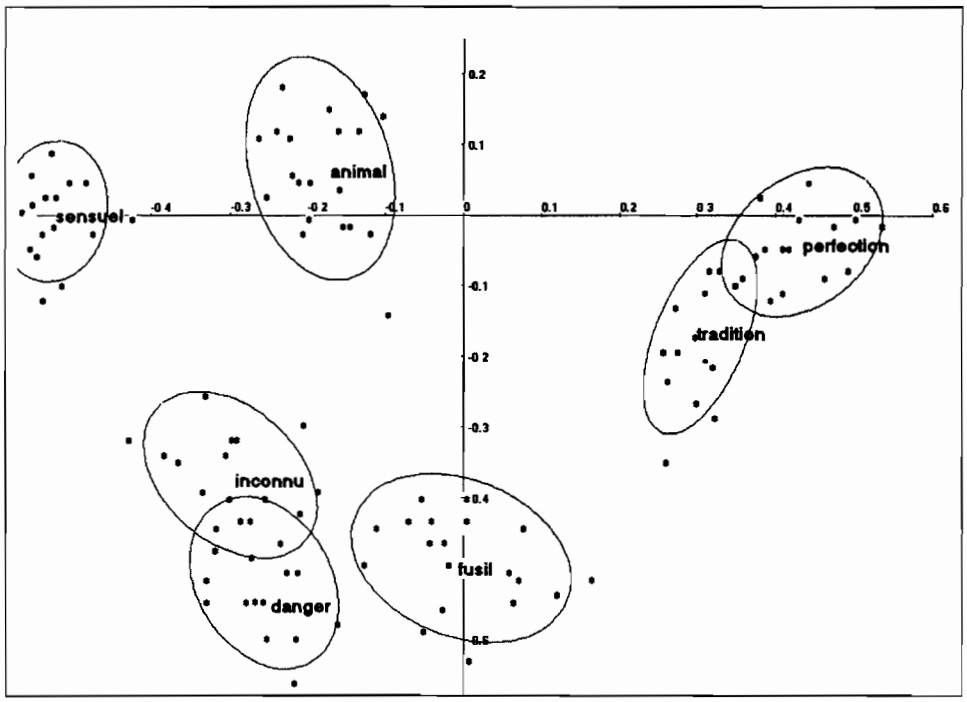


Figure 3.5-15. Bootstrap total de type 3 avec rotations procrustéennes. Les sept zones de confiance elliptiques des mêmes variables dans le plan (2, 3) de la figure 3.5-7

On voit à partir de cet exemple que la notion de validation dans le cas de l'analyse en composantes principales est complexe : les intervalles d'Anderson ont montré qu'un espace à quatre dimensions (peut-être cinq) était *saillant*, c'est-à-dire correspondait à une variance anormalement élevée statistiquement. Il peut cependant exister des rotations internes à ce sous-espace, bien que les deux premiers axes soient isolément saillants au sens précédent. Le bootstrap partiel et les trois types de bootstrap total ont confirmé ces résultats en montrant plus précisément la variabilité autour des points. On s'est limité ici à

sept points-variable et à quelques axes pour des questions de place disponible, mais la validation se fait en pratique dans un contexte plus interactif. La validation externe, lorsque des informations supplémentaires sont disponibles, est un autre outil de validation utile. C'est l'objet du sous-paragraphe suivant.

i – Bootstrap pour la position des nominales supplémentaires

Il ne peut s'agir que de bootstrap partiel, puisque les variables nominales supplémentaires ne participent pas à la construction des axes. Il s'agit, de plus, d'une validation externe (cf. § 3.4.2).

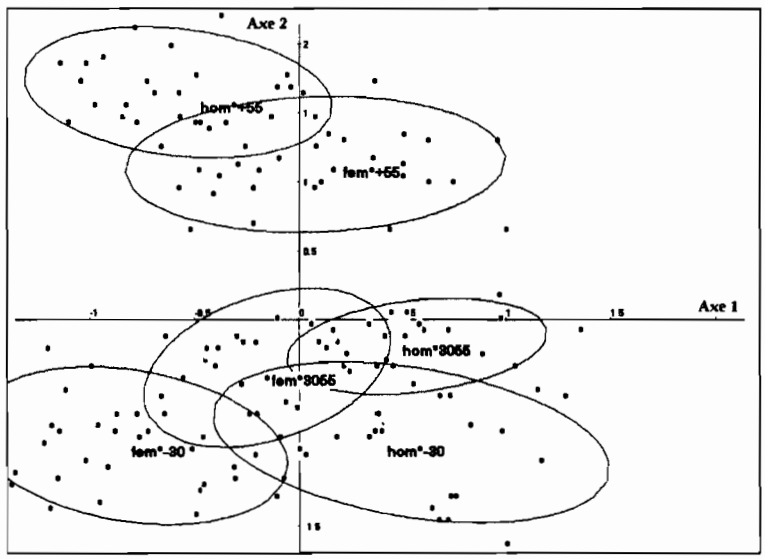


Figure 3.5-16. Bootstrap pour les variables nominales supplémentaires
Six zones de confiance elliptiques pour les modalités de la variable sexe-âgedans le plan (1, 2)

Même si un plan factoriel est instable, il peut en effet refléter une partie d'une structure stable. Si une variable externe se projette de façon significative dans ce plan, elle valide son pouvoir prédictif ou explicatif. La figure 3.5-16 nous montre clairement que l'axe 2 oppose de façon significative les hommes et les femmes de plus de 45 ans aux plus jeunes. Elle laisse entendre, de façon moins nette, que les femmes de moins de 30 ans s'opposent sur le premier axe aux hommes d'âge comparable. A partir des remarques précédentes, on peut en conclure que les hommes jeunes utilisent de façon systématique des notes plus basses que leurs homologues du sexe opposé, phénomène indécelable chez les personnes plus âgées.

j – Remarque sur le codage : Peut-on analyser des notes en ACP ?

Le codage de base utilisé dans cet exemple (notes de 1 à 7) relève d'une convention assez arbitraire : pour le répondant, c'est surtout l'ordre des notes

qui importe, et non la matérialisation de celles-ci sous forme d'échelle de *1* à *7*, avec des notes en progression arithmétique.

Est-ce que les structures observées dans les plans factoriels seraient les mêmes si le codage proposé était différent, en respectant toutefois l'ordre des notes ? Notons déjà que toute transformation y du codage initial x de la forme $y = ax + b$ laisse invariante la matrice des corrélations, et donc laisse identique toute la structure. Ainsi, on ne change pas les résultats en remplaçant (*1, 2, 3, 4, 5, 6, 7*) par (*2, 4, 6, 8, 10, 12, 14*) ou encore par (*11, 12, 13, 14, 15, 16, 17*).

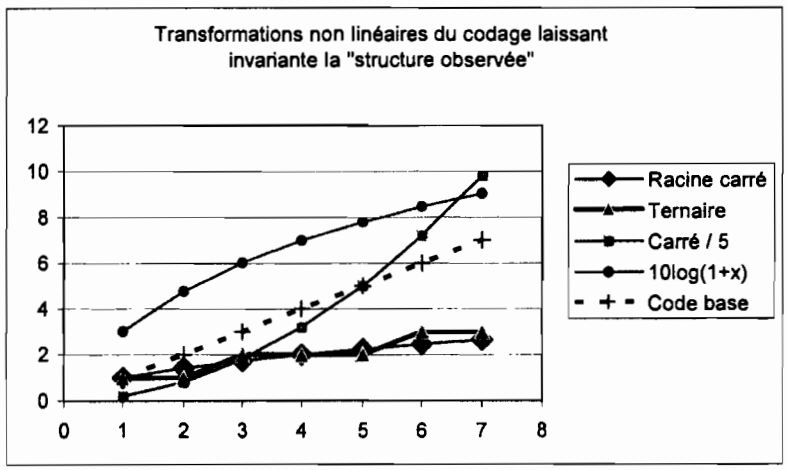


Figure 3.5-17. Diverses transformations de l'échelle des notes
Sur cette figure, le carré a été divisé par 5 et le logarithme multiplié par 10 pour ramener les différents codages à une échelle comparable à celle du codage initial (Code base, en pointillé).

Les transformations non-linéaires suivantes du codage arithmétique de départ ont été réalisées (voir figure 3.5-17) :

a) $y = \sqrt{x}$;

b) $y = x^2$;

c) $y = \log_{10}(1 + x)$;

d) codage « ternaire » selon lequel $y = 1$ pour $x = 1$ ou 2 ; $y = 2$ pour $x = 3$ ou 4 ou 5 ; $y = 3$ pour $x = 6$ ou 7.

Aucune de ces transformations n'altère profondément la structure observée sur les échantillons nationaux de tailles réelles (de l'ordre de *1000* individus). On observe cependant des interversions d'axes (attendues) dans le cas du sous-échantillon de *300* individus.

3.6 Annexe technique du chapitre 3

3.6.1 Travaux sur la loi des valeurs propres en analyse en composantes principales

La loi du χ^2, dans un cadre paramétrique classique, définit la distribution d'une *variance* empirique sous l'hypothèse d'observations indépendantes identiquement distribuées suivant une loi normale de moyenne nulle et d'écart-type σ connu. La loi de Wishart, établie par Fisher (1915) dans le cas $p = 2$, puis par Wishart (1928), plus générale, concerne la distribution d'une *matrice des covariances* empiriques.

Si les n vecteurs-lignes d'une matrice $\mathbf{X}$ d'ordre (n,p) sont des réalisations indépendantes d'un vecteur multinormal de moyenne théorique nulle, et de matrice des covariances théoriques $\mathbf{\Sigma}$ (non singulière) alors la matrice $\mathbf{S} = \mathbf{X'X}$ (qui contient $p(p+1)/2$ éléments distincts) suit une loi de Wishart[1], notée $W(p,n,\mathbf{\Sigma})$ dont la densité $f(\mathbf{S})$ est donnée par la formule :

$$f(\mathbf{S}) = C(n,p,\mathbf{\Sigma})\ |\mathbf{S}|^{\frac{n-p-1}{2}}\ exp\left\{-\frac{1}{2}trace(\mathbf{\Sigma}^{-1}\mathbf{S})\right\},$$

la constante $C(n,p,\mathbf{\Sigma})$ ayant pour valeur :

$$C(n,p,\mathbf{\Sigma}) = 2^{-\frac{np}{2}}\ |\mathbf{\Sigma}|^{-\frac{n}{2}}\ \pi^{-\frac{p(p-1)}{4}} \prod_{k=1}^{p} \Gamma\left(\frac{1}{2}(n+1-k)\right)$$

On vérifie que pour $\mathbf{\Sigma} = \mathbf{I}$ (matrice unité) et $p = 1$, notant $s = \mathbf{x'x}$, on retrouve la densité de probabilité du χ^2. En effet :

$$f(s) = C(n,1,\mathbf{I})\ s^{\frac{n}{2}-1} exp\left\{-\frac{s^2}{2}\right\} \qquad \text{avec :} \qquad C(n,1,\mathbf{I}) = 2^{-\frac{n}{2}}\ \Gamma\left(\frac{n}{2}\right)$$

La loi de la matrice $\mathbf{S}$ (loi de Wishart) intervient dans l'établissement des tests intervenant en analyse de la variance multidimensionnelle et en analyse discriminante. C'est le cas pour le test d'égalité de plusieurs matrices de covariances [*test de Box*], test d'égalité de vecteurs moyens [*test du Λ de Wilks*],

---

[1] Pour l'établissement de la densité de la loi de Wishart et de certaines lois dérivées, cf. Dugué (1958), Anderson (1958), Muirhead (1982). On note que $\mathbf{S}$ n'est pas la matrice des covariances empiriques puisque les variables ne sont pas centrées sur la moyenne empirique de l'échantillon. On montre (cf. références ci-dessus) que la loi de $\mathbf{S}$ après centrage empirique est une loi $W(p,n-1,\mathbf{\Sigma})$.

etc. (cf. Saporta, 1990). La densité de probabilité des *valeurs propres* issues d'une matrice de Wishart a été explicitée simultanément par Fisher (1939), Girshick (1939), Hsu (1939) et Roy (1939), puis par Mood (1951). On en trouve une démonstration dans Anderson (1958), Muirhead (1982).

Dans le cas où $\mathbf{\Sigma} = \mathbf{I}$, la densité de la loi de Wishart s'écrit facilement en fonction de la trace et du déterminant de **S**, c'est-à-dire de la somme et du produit des valeurs propres λ_k :

$$f(\mathbf{S}) = C(n,p,\mathbf{I}) \left(\prod_{k=1}^{p} \lambda_k \right)^{-\frac{n-p-1}{2}} exp\left\{ -\frac{1}{2} \sum_{k=1}^{p} \lambda_k \right\}$$

On retrouvera donc ces éléments (multipliés par le jacobien de la transformation qui est ici le produit de toutes les différences possibles entre valeurs propres) dans l'expression de la densité $g(\Lambda)$ des valeurs propres :

$$g(\Lambda) = D(n,p) \left(\prod_{k=1}^{p} \lambda_k \right)^{-\frac{n-p-1}{2}} exp\left\{ -\frac{1}{2} \sum_{k=1}^{p} \lambda_k \right\} \prod_{k<j}^{p} (\lambda_k - \lambda_j)$$

La constante $D(n,p)$ ayant pour valeur :

$$D(n,p) = 2^{-\frac{np}{2}} \pi^{\frac{p}{2}} \prod_{k=1}^{p} \left\{ \Gamma\left(\frac{1}{2}(n+1-k) \right) \Gamma\left(\frac{1}{2}(p+1-k) \right) \right\}$$

L'intégration de cette densité assez complexe a donné lieu à plusieurs publications ; parmi les principales, celles de Pillai (1965), Krishnaiah et Chang (1971), qui s'inspirent des travaux du physicien Mehta (1960, 1967) [1].

Les distributions ci-dessus s'appliquent à des variables indépendantes de variance théorique égale à *1* (l'hypothèse de moyenne nulle n'est pas nécessaire puisqu'il suffit de travailler avec la matrice des covariances centrées, et de changer n en $n-1$ dans la loi de **S**).

Il n'est donc pas facile d'utiliser ces résultats dans les applications usuelles de l'analyse en composantes principales. Le fait de réduire les variables (**X'X** est alors n fois la matrice des corrélations) ne résout pas le problème car **X'X**, dont les termes diagonaux sont constants et égaux à n, ne suit évidemment pas une loi de Wishart. Les éléments diagonaux d'une matrice de Wishart $W(\pi, \nu-1, I)$ sont en effet des χ^2 à $n-1$ degrés de liberté.

---

[1] Une table des seuils correspondant aux deux valeurs propres extrêmes a été publiée par Choudary, Hanumara et Thompson (1968) pour des matrices ayant leur plus petit côté p inférieur à *10* ; par Pillai (1967), Pillai et Chang (1970) et par Clemm, Krishnaiah et Waikar (1973) pour $p \leq 20$.

Chapitre 4

Analyse des correspondances

L'analyse des correspondances, présentée sous ce nom et développée par Benzécri (1969), a un certain nombre de précurseurs, parmi lesquels il faut citer Guttman (1941), Hayashi (1956). Comme l'analyse en composantes principales, elle peut être présentée selon divers points de vue. Il est pour cela difficile de faire l'historique précis de cette méthode. Les principes théoriques remontent probablement aux travaux de Fisher (1940) sur les tables de contingences, dans un cadre de statistique inférentielle classique. Depuis les travaux de Benzécri (1973) et de Escofier-Cordier (1965), on utilise surtout les propriétés algébriques et géométriques de l'outil descriptif que constitue l'analyse[1].

L'analyse des correspondances est, avec l'analyse en composantes principales, l'autre technique fondamentale des méthodes en axes principaux (ou méthodes factorielles). Elle concerne un domaine d'application différent. Alors que l'on réserve l'analyse en composantes principales aux tableaux de mesures éventuellement hétérogènes et au traitement de variables numériques continues, l'analyse des correspondances est une méthode adaptée aux *tableaux de contingence* et permet d'étudier les éventuelles relations existant entre deux variables nominales. Cette méthode n'est pas un cas particulier de l'analyse en composantes principales bien que l'on puisse se ramener à cette technique en faisant des changements de variables appropriés (à condition de traiter chaque espace séparément). On montre qu'il s'agit de la recherche de la meilleure

[1] Les ancêtres les plus lointains de l'analyse des correspondances seraient, de façon tout à fait indépendante, Richardson et Kuder (1933) et Hirschfeld (1935). Les premiers auteurs visaient une meilleure sélection des vendeurs de la société *Procter and Gamble*, alors que le dernier étudiait une propriété de statistique mathématique. Cette variété de contextes est caractéristique de l'analyse des correspondances, méthode aussi utile en pratique que stimulante du point de vue théorique. Cf. Escofier (2003) et les références historiques de Hill (1974), Benzécri (1982 a).

représentation simultanée de deux ensembles constituant les lignes et les colonnes d'un tableau de données. Nous verrons au chapitre suivant qu'elle fournit, par extension, des descriptions satisfaisantes de certains tableaux de codages discontinus.

4.1 Démarche et principe : introduction élémentaire

Nous allons utiliser, pour un premier survol élémentaire de la méthode, une table de contingence de faible dimension qui va permettre de présenter de façon simple les principes de cette méthode et les propriétés qui en découlent[1]. Bien que les lignes et les colonnes jouent un rôle similaire, nous conservons les mêmes notations que pour l'analyse générale.

4.1.1 Tableau de contingence : hypothèse d'indépendance

Le tableau de contingence (ou tableau croisé) est obtenu en ventilant une population selon deux variables nominales. L'ensemble des colonnes du tableau désigne les modalités d'une variable et l'ensemble des lignes correspond à celles de l'autre variable. De ce fait, les lignes et les colonnes, qui désignent deux partitions d'une même population, jouent des rôles symétriques et sont traitées de façon analogue.

a – Notations

Considérons le tableau de contingence **K** à n lignes et p colonnes obtenu en ventilant une population de *592* femmes suivant leurs couleurs des yeux et des cheveux.

Tableau 4.1 – 1. Tableau de contingence
répartition de 592 femmes suivant les couleurs des yeux et des cheveux.

		couleur des cheveux				
		brun	châtain	roux	blond	Total
couleur des yeux	marron	*68*	*119*	*26*	*7*	*220*
	noisette	*15*	*54*	*14*	*10*	*93*
	vert	*5*	*29*	*14*	*16*	*64*
	bleu	*20*	*84*	*17*	*94*	*215*
	Total	*108*	*286*	*71*	*127*	*592*

Source : Snee (1974), Cohen(1980)

[1] Une présentation technique plus détaillée sera l'objet des paragraphes suivants de la même section.

En lignes est présentée la variable "couleur des yeux" à $n = 4$ modalités (ou catégories) et en colonnes est donnée la variable "couleur des cheveux" à $p = 4$ modalités.

A l'intersection d'une ligne et d'une colonne, nous avons le nombre k_{ij} de femmes ayant simultanément la couleur i des yeux et la couleur j de cheveux. Le total marginal $k_{i.}$ est le nombre de femmes ayant les yeux de couleur i, alors que le total marginal $k_{.j}$ est le nombre de femmes ayant les cheveux de couleur j.

On a les relations suivantes :

$$k_{i.} = \sum_{j}^{p} k_{ij} \qquad k_{.j} = \sum_{i}^{n} k_{ij} \qquad k = \sum_{i,j} k_{ij}$$

qui, en termes de fréquences relatives, donnent lieu aux relations :

$$f_{ij} = \frac{k_{ij}}{k} \qquad f_{i.} = \sum_{j}^{p} f_{ij} \qquad f_{.j} = \sum_{i}^{n} f_{ij} \qquad \sum_{i,j} f_{ij} = 1$$

Y a-t-il indépendance entre la couleur des yeux et celle des cheveux ? Sinon quels types d'associations existent entre ces couleurs ?

b – Transformations du tableau de contingence

Pour analyser un tableau de contingence, ce n'est pas le tableau d'effectifs bruts qui nous intéresse mais les tableaux des profils-lignes et celui des profils-colonnes c'est-à-dire les répartitions en pourcentage à l'intérieur d'une ligne ou d'une colonne.

On note les profils-lignes : $\frac{f_{ij}}{f_{i.}} = \frac{k_{ij}}{k_{i.}}$

Tableau 4.1 – 2 Profils-lignes (pourcentages-lignes arrondis)

		couleur des cheveux				
		brun	châtain	roux	blond	total
couleur des yeux	marron	*31*	*54*	*12*	*3*	*100*
	noisette	*16*	*58*	*15*	*11*	*100*
	vert	*8*	*45*	*22*	*25*	*100*
	bleu	*9*	*39*	*8*	*44*	*100*
profil moyen		*18*	*48*	*12*	*22*	*100*

et les profils-colonnes : $\frac{f_{ij}}{f_{.j}} = \frac{k_{ij}}{k_{.j}}$

Le tableau 4.1 - 2 des profils-lignes (multipliés par *100*) indique la répartition de la couleur des cheveux pour chaque modalité de couleur des yeux. Ce sont en somme les probabilités conditionnelles d'avoir les cheveux de la couleur j

sachant que les yeux ont la couleur *i*. Cette répartition sur l'ensemble de la population étudiée donne le profil moyen :

$$f_{.j} = \frac{k_j}{k}$$

Tableau 4.1 – 3 Profils-colonnes (pourcentages-colonnes arrondis)

		couleur des cheveux				profil
		brun	châtain	roux	blond	moyen
couleur des yeux	marron	*63*	*42*	*37*	*6*	*37*
	noisette	*14*	*19*	*20*	*8*	*16*
	vert	*5*	*10*	*20*	*13*	*11*
	bleu	*19*	*29*	*24*	*74*	*36*
total		*100*	*100*	*100*	*100*	*100*

Le tableau 4.1 - 3 des profils-colonnes (multipliés par *100*) fournit la répartition de la couleur des yeux suivant chaque modalité de couleur des cheveux et le profil moyen de la couleur des yeux :

$$f_{i.} = \frac{k_i}{k}$$

c – Hypothèse d'indépendance

On s'intéresse aux liens éventuels entre couleurs des yeux et des cheveux.

On sait qu'il y a indépendance entre deux variables aléatoires *i* et *j* prenant leurs valeurs sur deux ensembles de tailles *n* et *p*, dont la loi jointe est p_{ij} et les lois marginales $p_{i.}$ et $p_{.j}$, si pour tout *i* et pour tout *j* on a (avec les notations usuelles) :

$$p_{ij} = p_{i.} p_{.j}$$

La traduction de cette relation en termes d'estimations empiriques est la suivante :

$$f_{ij} = f_{i.} f_{.j}$$

Naturellement, même sous l'hypothèse d'indépendance, une telle relation n'est qu'approximativement vraie. Le classique test du χ^2 de Karl Pearson pour les tables de contingence permet précisément d'apprécier l'écart entre les lois empiriques f_{ij} et $f_{i.} f_{.j}$.

Consultons le tableau 4.1 - 4 des fréquences observées f_{ij} qui n'est autre que la tableau 4.1 - 1 divisé par sa somme (*592*) et multiplié par *100* pour plus de lisibilité.

Parmi les 37% de femmes aux yeux marrons par exemple, on devrait observer, sous l'hypothèse d'indépendance, *18*% de femmes brunes (ce qui ferait alors 7%

de l'ensemble des femmes, au lieu des *11%* réellement observés), *48%* aux cheveux châtains (ce qui ferait *18%* au lieu de *20%*), etc.

Tableau 4.1 - 4 . Tableau de fréquences observées

		couleur des cheveux				profil
		brun	châtain	roux	blond	moyen
couleur des yeux	marron	*11*	*20*	*4*	*1*	*37*
	noisette	*3*	*9*	*2*	*2*	*16*
	vert	*1*	*5*	*2*	*3*	*11*
	bleu	*3*	*14*	*3*	*16*	*36*
profil moyen		*18*	*48*	*12*	*21*	*100*

Construisons le tableau de "fréquences théoriques" $f_{i.}f_{.j}$ sous l'hypothèse d'indépendance (cf. tableau [4.1 - 5]) :

Tableau 4.1 - 5 . Tableau de fréquences théoriques

		couleur des cheveux				profil
		brun	châtain	roux	blond	moyen
couleur des yeux	marron	*7*	*18*	*4*	*8*	*37*
	noisette	*3*	*8*	*2*	*3*	*16*
	vert	*2*	*5*	*1*	*2*	*11*
	bleu	*7*	*18*	*12*	*8*	*36*
profil moyen		*18*	*48*	*12*	*21*	*100*

Cette hypothèse s'exprime aussi sur les profils-lignes. En effet, il en découle que, quelque soit j :

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

Si tous les profils "couleurs des yeux" sont identiques entre eux, et par conséquent identiques au profil moyen correspondant, il y a indépendance entre les couleurs des yeux et celles de cheveux puisque la connaissance d'une couleur des yeux ne change pas la répartition de la couleur des cheveux. Il en est de même pour les profils-colonnes où, quelque soit i :

$$\frac{f_{ij}}{f_{.j}} = f_{i.}$$

Ainsi, examiner les proximités entre les profils revient à examiner la proximité entre chaque profil et son profil moyen, ce qui permet d'étudier la liaison entre deux variables nominales, c'est-à-dire l'écart à l'indépendance. Sur un tableau de dimension importante, la lecture directe des profils-lignes et des profils-colonnes est difficile, ainsi que la comparaison de ces profils avec leur profil moyen.

4.1.2 Représentation géométrique

Nous allons voir comment la construction du nuage, le choix du critère d'ajustement et celui de la distance, s'imposent de par la nature même des données analysées.

a – Construction des nuages

Pour l'analyse d'un tableau de contingence, nous raisonnerons en termes de profils, ce qui permet de rendre comparables les modalités d'une même variable. Les proximités entre les points s'interpréteront en termes de similitude.

- Nuage des n lignes

L'ensemble des profils-lignes forme un nuage de n points dans l'espace des p colonnes et représente ici le nuage des 4 modalités de couleurs des yeux. Chaque point i a pour coordonnées dans $\mathcal{R}^p$:

$$\left\{\frac{f_{ij}}{f_{i.}}; j = 1,2,\ldots,p\right\}$$

Il est affecté d'une masse $f_{i.}$ qui est sa fréquence relative.

Puisque $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} = 1$, les n points du nuage sont situés dans un sous-espace à $p-1$ dimensions.

Le centre de gravité de ce nuage est la moyenne des profils-lignes affectés de leurs masses et correspond au profil moyen, c'est-à-dire au profil de la couleur des cheveux sur l'ensemble de la population. Sa $j^{ème}$ composante vaut :

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

C'est la fréquence marginale des colonnes.

- Nuage des p colonnes

De la même façon, l'ensemble des p profils-colonnes constitue un nuage de p points dans l'espace des n lignes et représente ici le nuage des 4 modalités de couleur des cheveux.

Les coordonnées dans $\mathcal{R}^n$ du point j sont données par :

$$\left\{\frac{f_{ij}}{f_{j}}; i = 1,2,\ldots,n\right\}$$

Chaque point est affecté d'une masse $f_{.j}$. Les p points du nuage sont situés dans un sous-espace à $n-1$ dimensions puisque $\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} = 1$.

Le centre de gravité du nuage des profils-colonnes est le profil moyen de la couleur des yeux. Sa $i^{ème}$ composante vaut :

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

C'est la fréquence marginale des lignes.

b – Critère d'ajustement

On cherche à représenter géométriquement les similitudes entre les différentes modalités d'une même variable, ce qui nous conduit à représenter les proximités entre les profils et le profil moyen défini sur l'ensemble de la population[1]. Ceci nous amène, comme en analyse en composantes principales dans le cas des points-individus, à considérer le nuage de points centré sur son centre de gravité.

Dans la construction des nuages de $\mathcal{R}^p$ et de $\mathcal{R}^n$ (cf. tableaux 4.1 - 2 et 4.1 - 3), le choix des profils comme coordonnées donne à toutes les modalités de couleur des yeux et celles de cheveux la même importance. L'importance est cependant restituée au travers de la masse affectée à chaque point (proportionnelle à sa fréquence), afin de ne pas privilégier les classes d'effectifs faibles et de respecter la répartition réelle de la population. Cette masse interviendra d'une part lors du calcul des coordonnées du centre de gravité du nuage et d'autre part dans le critère d'ajustement.

Pour le calcul de l'ajustement, la quantité à rendre maximale sera donc la somme pondérée des carrés des distances entre les points et le centre de gravité du nuage (c'est-à-dire l'inertie de la droite d'allongement maximum du nuage) en utilisant une distance entre profils qu'il reste à définir.

c – Choix des distances

La distance euclidienne usuelle entre deux points-lignes exprimée sur le tableau d'effectifs bruts ne ferait que traduire les différences d'effectifs entre deux modalités de couleurs des yeux. En revanche, la distance euclidienne usuelle entre deux profils-lignes traduit bien la ressemblance ou la différence entre les deux couleurs des yeux sans tenir compte des effectifs totaux de ces modalités :

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Cependant, cette distance favorise les colonnes qui ont une masse $f_{.j}$ importante (ici : les couleurs de cheveux qui sont fréquentes dans la population étudiée).

---

[1] Un nuage de points concentré autour de son centre de gravité est un nuage dont les points-profils sont proches du profil moyen, et donc traduira une certaine indépendance entre les deux variables nominales.

Pour remédier à cela, et aussi pour d'autres propriétés qui seront développées plus bas, on pondère chaque écart par l'inverse de la masse de la colonne et l'on calcule une nouvelle distance appelée[1] *la distance du* χ^2 :

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \qquad [4.1 - 1]$$

On définit de la même manière la distance entre les profils-colonnes par :

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \qquad [4.1 - 2]$$

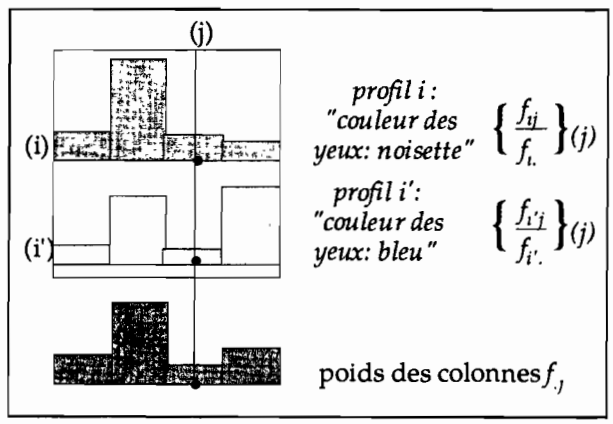


Figure 4.1 – 1. Distance du χ^2

4.1.3 Propriétés

La distance pondérée du χ^2, ainsi que le rôle symétrique joué par les lignes et les colonnes du tableau de contingence, particularisent l'analyse des correspondances et lui assurent des propriétés remarquables que ne possède pas l'analyse en composantes principales : l'équivalence distributionnelle et les relations de transition.

a – Equivalence distributionnelle

La propriété d'équivalence distributionnelle permet d'agréger deux modalités d'une même variable ayant des profils identiques en une nouvelle modalité affectée de la somme de leurs masses, sans rien changer, ni aux distances entre les modalités de cette variable, ni aux distances entre les modalités de l'autre

[1] L'inertie totale des nuages de points lignes (ou de points colonnes) calculée avec cette distance est proportionnelle au classique χ^2 de Karl Pearson utilisé pour éprouver l'indépendance des lignes et des colonnes d'une table de contingence. D'où le nom de distance du χ^2.

variable. Si par exemple les deux profils-lignes i' et i'' sont identiques dans $\mathcal{R}^p$, on les agrège en un profil-ligne i dont la masse sera la somme des fréquences des deux profils i' et i''. Les deux points i' et i'' étant confondus cela ne modifie pas la configuration du nuage de points dans $\mathcal{R}^p$.

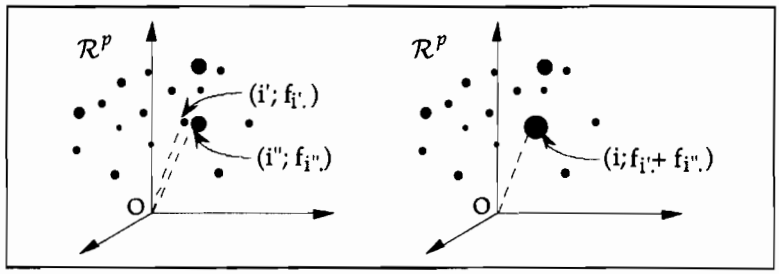


Figure 4.1 - 2 . Equivalence distributionnelle : points-lignes confondus

Mais surtout, les distances entre colonnes restent inchangées. Il en est de même pour des profils-colonnes dans $\mathcal{R}^n$ ayant les mêmes propriétés. Cette propriété est fondamentale puisqu'elle garantit une certaine invariance des résultats vis-à-vis de la nomenclature choisie pour la construction des modalités d'une variable, sous condition de regrouper des modalités aux profils semblables. On ne perd pas d'information en agrégeant certaines classes et l'on n'en gagne pas en subdivisant des classes homogènes.

Prenons le cas de deux tables de contingences issues du recensement de la population, l'une croisant professions et départements, l'autre professions et régions. Sous l'hypothèse d'homogénéité des départements d'une même région par rapport aux professions, il sera équivalent de réaliser l'analyse des correspondances sur les départements et sur les régions. Les configurations du nuage des professions, pour les deux analyses, seront semblables (voir la démonstration au § 4.2.1.a).

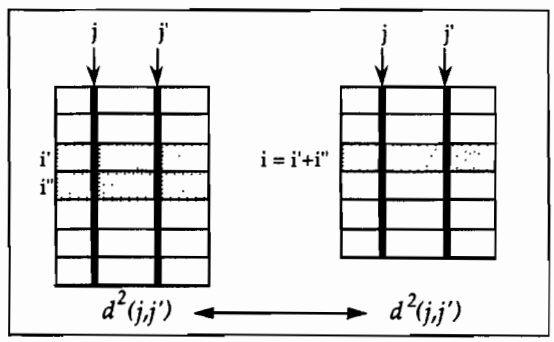


Figure 4.1 – 3. Equivalence distributionnelle : invariance des distances entre colonnes vis-à-vis de l'agrégation des lignes

b – Relations de transition ou quasi-barycentriques

Une des caractéristiques de l'analyse des correspondances est l'existence de relations de type barycentrique qui lient graphiquement les deux variables représentées en ligne et en colonne. L'idée est simple et revient à représenter les histogrammes des profils-colonnes dans le nuage des profils-lignes et réciproquement.

Supposons fixé le nuage des couleurs des yeux (nuage des profils-lignes) dans un espace à 2 dimensions comme représenté sur la figure 4.1 - 4. Le centre du graphique représente le profil moyen (la distribution marginale) des couleurs des yeux.

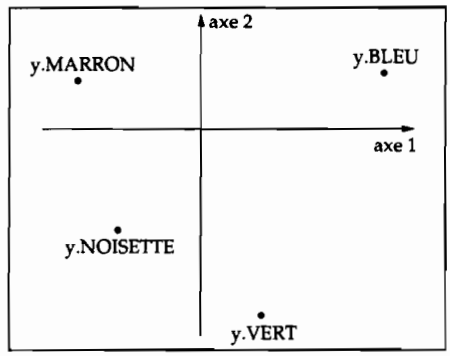


Figure 4.1 – 4. Nuage des couleurs des yeux

Considérons maintenant l'histogramme décrivant le profil des cheveux bruns suivant la couleur de yeux (cf. tableau 4.1 - 3 des profils-colonnes) représenté figure 4.1 - 5.

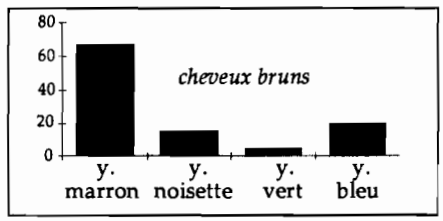


Figure 4.1 - 5. Histogramme des cheveux bruns

Cet histogramme va permettre de positionner le point-colonne "cheveux bruns" dans le nuage des points-lignes (le nuage des couleurs des yeux) : chaque point i représentant une couleur des yeux est pondéré par sa fréquence relative telle qu'elle est décrite par l'histogramme.

On construit ainsi le barycentre de ces points qui correspond au point "cheveux bruns". Il est contenu dans une enveloppe convexe constituée par l'ensemble des points pondérés (cf. figure 4.1 - 6). Cette modalité sera attirée par les yeux marrons, compte tenu de sa masse plus élevée. Elle sera par contre éloignée des yeux verts.

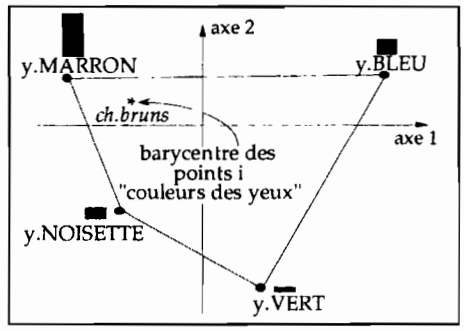


Figure 4.1 – 6. Position du point "cheveux bruns" comme barycentre des points "couleurs des yeux"

Chaque point *j* "couleur des cheveux" est ainsi un barycentre particulier des points *i* "couleur des yeux", le point *i* étant affecté de la masse "part de la couleur *i* des yeux sachant que la couleur des cheveux est *j*", (c'est-à-dire le profil-colonne $f_{ij}/f_{.j}$) (cf. figure 4.1 - 7).

Si l'on considère maintenant le nuage des profils-colonnes, c'est-à-dire le nuage des couleurs des cheveux, il est naturel de procéder de la même façon et de représenter l'histogramme de chaque couleur des yeux dans ce nuage.

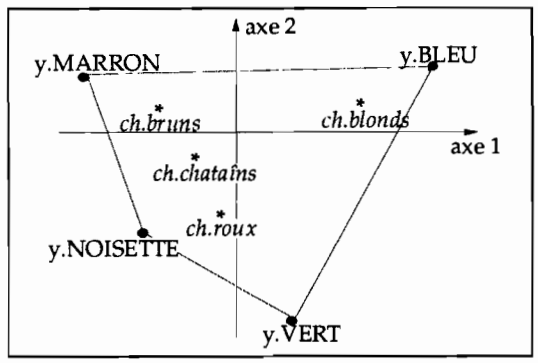


Figure 4.1 – 7. Représentation des points "couleurs des yeux" et positionnement des points "couleurs des cheveux" en barycentres

On positionne donc chaque point-ligne *i* "couleur des yeux" comme barycentre des points *j* "couleurs des cheveux" pondérés par la part de la couleur *j* des cheveux dans la couleur *i* des yeux, donnée par les profils-lignes $\{f_{ij} / f_{i.}\}$ (cf. figure 4.1 - 8).

Les relations barycentriques vont justifier et donner un sens à la représentation simultanée des deux nuages définis dans les deux espaces.

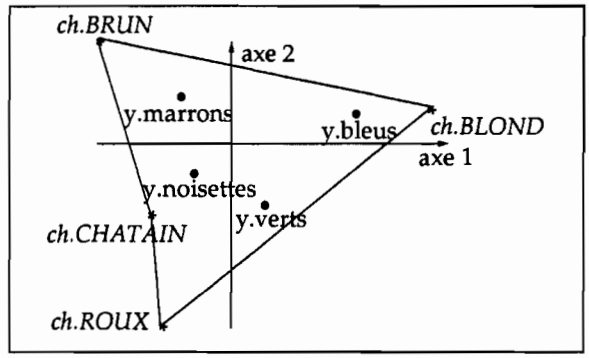


Figure 4.1 – 8. Représentation des points "couleurs des cheveux" et positionnement des points "couleurs des yeux" en barycentres

c – Justification de la représentation simultanée

D'après le schéma de l'analyse générale, on pourrait envisager l'analyse des deux nuages de points de manière indépendante et l'interpréter comme une analyse en composantes principales. Compte tenu des relations barycentriques, il en est autrement en analyse des correspondances.

Ces relations montrent qu'il existe une possibilité de représentation particulière [1] : il est possible de positionner chaque point d'un nuage parmi l'ensemble des points de l'autre nuage. Ainsi, dans le nuage des profils-lignes, chaque profil-colonne est au barycentre des points du nuage. Projeté sur un plan, nous disposons d'une première représentation simultanée (cf. figure 4.1 - 7). De même, chaque profil-ligne est barycentre de l'ensemble des profils-colonnes et constitue, avec les axes de mêmes rangs, une deuxième représentation simultanée (cf. figure 4.1 - 8).

Mais nous voulons une seule représentation simultanée des deux nuages de points. Ceci est *a priori* impossible par définition même du barycentre puisque chaque ensemble devrait alors être contenu dans l'autre. Il est cependant possible de forcer cette représentation en dilatant (sur chaque axe) les centres de gravité (figure 4.1 - 9). On pourra alors représenter sur de mêmes axes (et donc sur un même plan) l'ensemble des lignes et des colonnes. Les relations seront *quasi-barycentriques* (cf. § 4.2.2.b). Les yeux bleus s'associent aux cheveux blonds, les yeux marrons aux cheveux bruns. Les cheveux roux sont proches des yeux noisettes et verts. Les cheveux châtains sont proches de l'origine (profil moyen) et ne sont pas associés à une couleur des yeux. La représentation du nuage des points non dilaté et des barycentres correspondants (proches de l'origine) aurait fourni un graphique confus. La représentation *quasi-barycentrique*, du fait de la dilatation des nuages de points, offre une lecture plus facile.

---

[1] Cette possibilité est due au fait que les coordonnées d'origine (les profils) sont des nombres positifs dont la somme vaut *1*.

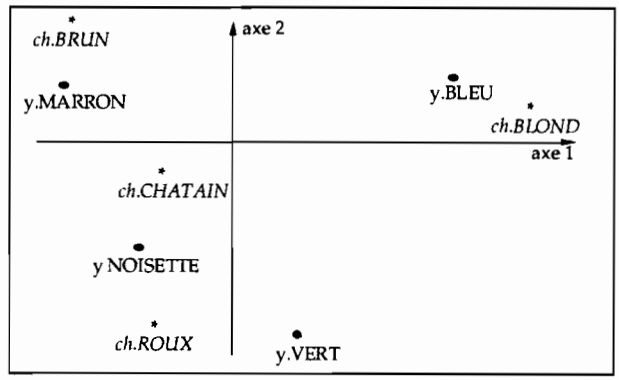


Figure 4.1 – 9. Représentation simultanée; Relations quasi-barycentriques

Nous verrons que le déroulement de l'analyse des correspondances, compte tenu des rôles symétriques des lignes et des colonnes du tableau de contingence et des propriétés de la distance du χ^2, aboutit naturellement aux relations barycentriques (à un coefficient près qui est le coefficient de dilatation permettant la représentation simultanée unique).

4.2 Schéma général de l'analyse des correspondances

L'analyse des correspondances va effectuer l'analyse générale d'un nuage de points pondérés dans un espace muni de la métrique du χ^2. On se référera donc à l'analyse générale avec des métriques et des critères quelconques (cf. § 1.3.1).

4.2.1 Eléments de base de l'analyse

a – Tableau de données, distance, géométrie des nuages

Contrairement à l'analyse en composantes principales, le tableau de données subit deux transformations, l'une en profils-lignes, l'autre en profils-colonnes, à partir desquelles vont être construits les nuages de points dans $\mathcal{R}^p$ et dans $\mathcal{R}^n$ (figure 4.2 - 1). Pour faire le lien avec l'analyse générale (cf. section 1.2), nous adopterons des notations matricielles (figure 4.2 - 2). Les transformations opérées sur le tableau des données peuvent s'écrire à partir des trois matrices **F**, $\mathbf{D}_n$ et $\mathbf{D}_p$ qui définissent les éléments de base de l'analyse. **F** d'ordre (n,p) désigne le tableau des fréquences relatives ; $\mathbf{D}_n$ d'ordre (n,n) est la matrice diagonale dont les éléments diagonaux sont les marges en lignes $f_{i.}$; $\mathbf{D}_p$ est la matrice diagonale d'ordre (p,p) des marges en colonnes $f_{.j}$.

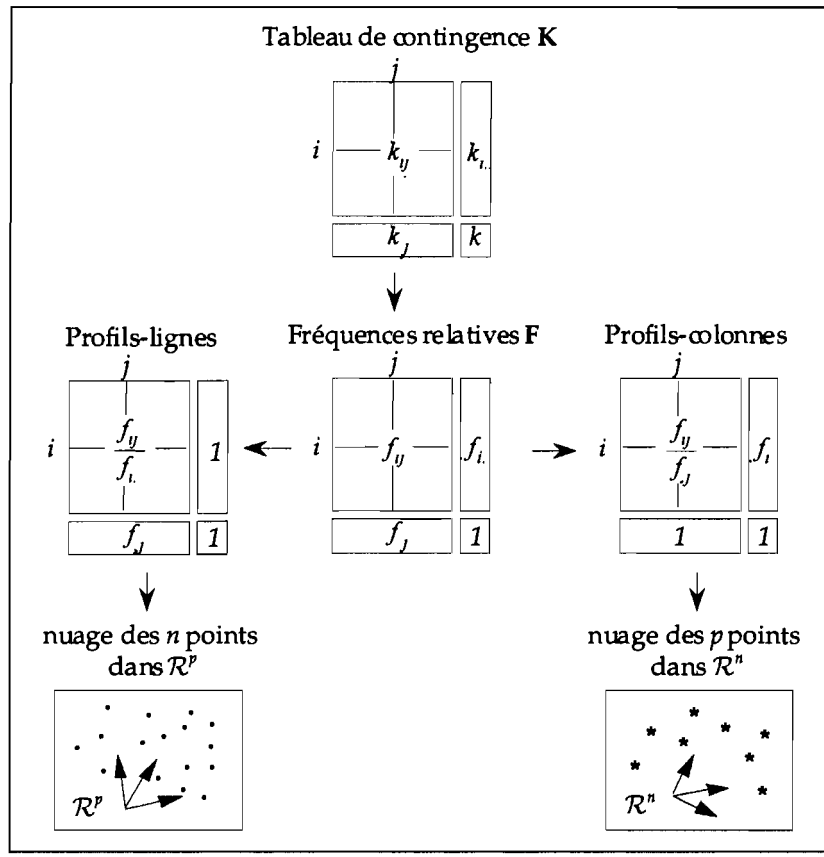


Figure 4.2 – 1. Transformations du tableau de contingence

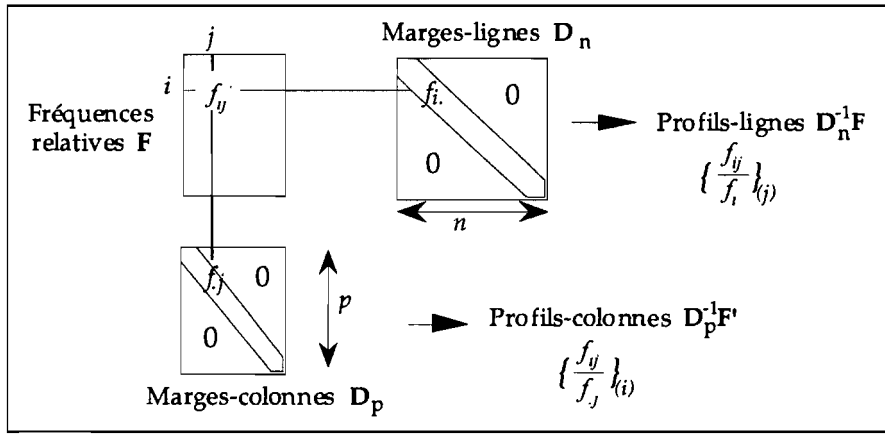


Figure 4.2 – 2. Fréquences, marges, profils

Les deux nuages de points (dans l'espace des colonnes et dans l'espace des lignes) sont construits de manière analogue. Nous récapitulons ici les éléments de base de l'analyse qui vont permettre la construction des facteurs.

Tableau 4.2 – 1. Les éléments de base de l'analyse : récapitulation

Nuage de n points-lignes dans l'espace $\mathcal{R}^p$	← Eléments → de base	Nuage de p points-colonnes dans l'espace $\mathcal{R}^n$
$\mathbf{X} = \mathbf{D}_n^{-1}\mathbf{F}$ p coordonnées (point-ligne i) $\frac{f_{ij}}{f_{i.}}$, pour j=1, 2, ...,p.	Analyse du tableau **X**	$\mathbf{X} = \mathbf{D}_p^{-1}\mathbf{F}'$ n coordonnées (point-colonne j) $\frac{f_{ij}}{f_{.j}}$, pour i=1, 2, ...,n.
$\mathbf{M} = \mathbf{D}_p^{-1}$ $d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$	avec la métrique **M**	$\mathbf{M} = \mathbf{D}_n^{-1}$ $d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$
$\mathbf{N} = \mathbf{D}_n$ masse du point i : $f_{i.}$	et le critère **N**	$\mathbf{N} = \mathbf{D}_p$ masse du point j : $f_{.j}$

Remarques

1) La matrice **N** des masses dans un espace est liée à la métrique **M** utilisée dans l'autre espace.

2) Il existe une différence fondamentale avec l'analyse en composantes principales : les transformations faites sur les données brutes dans les deux espaces sont identiques (car les ensembles mis en correspondance jouent des rôles analogues). Elles correspondent à des transformations analytiques différentes : le tableau des nouvelles coordonnées dans l'espace des colonnes n'est pas le simple transposé de celui des nouvelles coordonnées dans l'espace des lignes. En composantes principales, des transformations très différentes conduisaient à une même formule analytique.

b – Démonstration de l'équivalence distributionnelle

La distance du χ^2 a pour effet d'accorder une même importance, d'une part aux colonnes quelles que soient leurs fréquences relatives dans le calcul de la distance entre deux profils-lignes, et d'autre part aux lignes s'il s'agit du calcul de la distance entre profils-colonnes. Elle offre l'avantage de vérifier le principe d'équivalence distributionnelle[1] (cf. figure 4.1 - 2). Ce principe assure la robustesse des résultats de l'analyse des correspondances vis à vis de l'arbitraire

[1] La distance euclidienne usuelle entre profils ne possède pas la propriété d'équivalence distributionnelle, mais d'autres distances possèdent cette propriété (cf. Escofier, 1978).

du découpage en modalités des variables nominales. Il s'exprime de la façon suivante dans $\mathcal{R}^p$:

si deux points-lignes i_1 et i_2 sont confondus dans $\mathcal{R}^p$, on a pour tout j :

$$\frac{f_{i_1 j}}{f_{i_1 .}} = \frac{f_{i_2 j}}{f_{i_2 .}} = \frac{f_{i_0 j}}{f_{i_0}} \qquad [4.2 - 1]$$

On a en particulier :

$$\frac{f_{i_1 j} + f_{i_2 j}}{f_{i_1 .} + f_{i_2 .}} = \frac{f_{i_0 j}}{f_{i_0 .}}$$

D'où, puisque les dénominateurs sont égaux, on a pour tout j :

$$f_{i_1 j} + f_{i_2 j} = f_{i_0 j}$$

Les calculs des quantités $f_{.j} = \sum_i f_{ij}$ ne sont donc pas affectés et les distances $d^2(i,i')$ données par la formule [4.2 - 1] sont invariantes.

Montrons maintenant que les distances entre colonnes ne changent pas. La distance $d^2(j,j')$ donnée par la formule [4.1 - 2] contient entre autres les deux termes $A(i_1)$ et $A(i_2)$ correspondant aux indices i_1 et i_2 :

$$A(i_1) + A(i_2) = \frac{1}{f_{i_1 .}} \left\{ \frac{f_{i_1 j}}{f_j} - \frac{f_{i_1 j'}}{f_{j'}} \right\}^2 + \frac{1}{f_{i_2}} \left\{ \frac{f_{i_2 j}}{f_j} - \frac{f_{i_2 j'}}{f_{j'}} \right\}^2$$

Ces deux termes sont remplacés par un seul terme $A(i_0)$ tel que :

$$A(i_0) = \frac{1}{f_{i_0 .}} \left\{ \frac{f_{i_0 j}}{f_j} - \frac{f_{i_0 j'}}{f_{j'}} \right\}^2$$

Remarquons par exemple que :

$$A(i_1) = f_{i_1 .} \left\{ \frac{f_{i_1 j}}{f_{i_1 .} f_{.j}} - \frac{f_{i_1 j'}}{f_{i_1 .} f_{.j'}} \right\}^2$$

$A(i_1)$ et $A(i_2)$ s'écrivent de la même façon et les quantités entre accolades sont égales, d'après la relation [4.2 - 1], à un même nombre que l'on notera B. On a donc :

$$A(i_1) + A(i_2) = f_{i_1 .} B + f_{i_2} B = f_{i_0 .} B = A(i_0)$$

D'où l'invariance de la distance $d^2(j,j')$.

c – Critère à maximiser et matrice à diagonaliser

Nous voulons représenter graphiquement les proximités entre profils. Nous nous plaçons donc, dans les deux espaces, aux centres de gravité des nuages. Cependant, et c'est là une des particularités de l'analyse des correspondances, il

est équivalent de procéder à l'analyse par rapport à l'origine ou par rapport aux centres de gravité, à condition de négliger dans le premier cas l'axe factoriel qui joint l'origine au centre de gravité[1].

Nous commencerons par effectuer l'analyse générale par rapport à l'origine, l'expression des formules étant plus simple, puis nous montrerons, au paragraphe 4.6.2, l'équivalence avec l'analyse effectuée par rapport aux centres de gravité, autre propriété de l'analyse des correspondances.

Plaçons-nous dans l'espace des colonnes[2] $\mathcal{R}^p$ et cherchons l'axe d'inertie maximum du nuage des points-lignes passant par l'origine O et engendré par un vecteur-unitaire $\mathbf{u}$ pour la métrique $\mathbf{D}_p^{-1}$. Ceci nous amène à maximiser la somme pondérée des carrés des projections sur l'axe (cf. § 1.2.1) c'est-à-dire :

$$\underset{\mathbf{u}}{Max}\left\{\sum_i f_{i.}\, d^2(i,O)\right\}$$

et à rendre maximale la quantité :

$$\mathbf{u'D_p^{-1}F'D_n^{-1}FD_p^{-1}u}$$

avec la contrainte :

$$\mathbf{u'D_p^{-1}u} = 1$$

$\mathbf{u}$ est alors vecteur propre de la matrice :

$$\mathbf{S = F'D_n^{-1}FD_p^{-1}}$$

associé à la plus grande valeur propre λ différente de *1*.

La matrice à diagonaliser est donc la matrice $\mathbf{S}$ de terme général :

$$s_{ij'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}$$

De la même façon, on doit rendre maximum dans $\mathcal{R}^n$, la quantité :

$$\mathbf{v'D_n^{-1}FD_p^{-1}F'D_n^{-1}v}$$

avec la contrainte :

$$\mathbf{v'D_n^{-1}v} = 1$$

$\mathbf{v}$ est vecteur propre de la matrice :

$$\mathbf{T = FD_p^{-1}F'D_n^{-1}}$$

---

[1] Cet axe est associé à la valeur propre égale à *1*, appelée valeur propre triviale.

[2] Compte tenu du rôle symétrique des lignes et des colonnes du tableau de contingence, les démonstrations dans l'autre espace se déduisent par permutation des indices *i* et *j* (c'est-à-dire transposition de F et permutation des matrices $\mathbf{D_p}$ et $\mathbf{D_n}$).

d – Axes factoriels et coordonnées factorielles

Nous supposons ici que p correspond à la plus petite dimension du tableau de données. Après avoir écarté la valeur propre triviale égale à *1* et le vecteur propre associé, nous retenons, de la diagonalisation de la matrice, les *p-1* valeurs propres non nulles et les vecteurs propres associés. Nous obtenons ainsi au plus *p-1* axes factoriels.

Tableau 4.2 - 2
Eléments de construction de l'analyse

Dans $\mathcal{R}^p$	← Eléments de construction →	Dans $\mathcal{R}^n$
$\mathbf{S} = \mathbf{F'D}_n^{-1}\mathbf{FD}_p^{-1}$	Matrice à diagonaliser	$\mathbf{T} = \mathbf{FD}_p^{-1}\mathbf{F'D}_n^{-1}$
$\mathbf{Su}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$	Axe factoriel	$\mathbf{Tv}_\alpha = \lambda_\alpha \mathbf{v}_\alpha$
$\psi_\alpha = \mathbf{D}_n^{-1}\mathbf{FD}_p^{-1}\mathbf{u}_\alpha$ $\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_i f_{.j}} u_{\alpha j}$	Coordonnées factorielles	$\varphi_\alpha = \mathbf{D}_p^{-1}\mathbf{F'D}_n^{-1}\mathbf{v}_\alpha$ $\varphi_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.} f_j} v_{\alpha i}$

Les coordonnées factorielles sont centrées :

$$\sum_{i=1}^{n} f_i \psi_{\alpha i} = \sum_{j=1}^{p} f_{.j} \varphi_{\alpha j} = 0 \qquad [4.2 - 2]$$

et de variance égale à λ_α :

$$\sum_{i=1}^{n} f_i \psi_{\alpha i}^2 = \sum_{j=1}^{p} f_j \varphi_{\alpha j}^2 = \lambda_\alpha \qquad [4.2 - 3]$$

4.2.2. Représentation simultanée

a – Relation entre les deux espaces

L'analyse générale a montré que les matrices **S** et **T** ont les mêmes valeurs propres non nulles λ_α et qu'entre le vecteur propre unitaire $\mathbf{u}_\alpha$ de **S** associé à λ_α et le vecteur propre unitaire $\mathbf{v}_\alpha$ de **T** relatif à la même valeur propre, il existe les relations dites de transition :

$$\begin{cases} \mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{FD}_p^{-1}\mathbf{u}_\alpha & [4.2-4] \\ \mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{F'D}_n^{-1}\mathbf{v}_\alpha & [4.2-5] \end{cases}$$

La comparaison de ces relations avec les expressions des coordonnées factorielles :

$$\psi_\alpha = \mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha \qquad [4.2 - 6]$$

et

$$\varphi_\alpha = \mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha \qquad [4.2 - 7]$$

montre que celles-ci sont liées aux composantes des axes de l'autre espace par les formules :

$$\begin{cases} \psi_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_n^{-1}\mathbf{v}_\alpha & [4.2-8] \\ \varphi_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_p^{-1}\mathbf{u}_\alpha & [4.2-9] \end{cases}$$

C'est-à-dire, explicitement :

$$\begin{cases} \psi_{\alpha i} = \dfrac{\sqrt{\lambda_\alpha}}{f_{i.}} v_{\alpha i} \\ \varphi_{\alpha j} = \dfrac{\sqrt{\lambda_\alpha}}{f_{.j}} u_{\alpha j} \end{cases}$$

b – Relations de transition (ou quasi-barycentriques)

Les substitutions dans la relation [4.2 - 7] de $\mathbf{v}_\alpha$ par sa valeur tirée de [4.2 - 8] et dans la relation [4.2 - 6] de $\mathbf{u}_\alpha$ par sa valeur tirée de [4.2 - 9] conduisent aux relations fondamentales existant entre les coordonnées des points-lignes et des points-colonnes sur l'axe α, les relations quasi-barycentriques :

$$\begin{cases} \psi_{\alpha i} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} \varphi_{\alpha j} & [4.2-10] \\ \varphi_{\alpha j} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{n} \frac{f_{ij}}{f_{j}} \psi_{\alpha i} & [4.2-11] \end{cases}$$

Ainsi, au coefficient de dilatation $\dfrac{1}{\sqrt{\lambda_\alpha}}$ près, les projections des points représentatifs d'un nuage sont, sur un axe, les *barycentres* des projections des points représentatifs de l'autre nuage.

La matrice de terme général $\left(\dfrac{f_{ij}}{f_{i.}}\right)$ permettant de calculer les coordonnées d'un point i à partir de tous les points j (relation [4.2 - 10]) n'est autre que le tableau des profils-lignes.

La coordonnée de la modalité i d'une des variables est la moyenne des modalités j de l'autre variable pondérées par les fréquences conditionnelles du profil de i. De même, la relation [4.2 - 11] montre que la coordonnée de la modalité j est la moyenne de l'ensemble des modalités i pondérées par les fréquences conditionnelles du profil de j.

Remarques

1) Toutes les valeurs propres sont nécessairement inférieures ou égales à *1*.

En effet puisque :

$$\sqrt{\lambda_\alpha}\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} \varphi_{\alpha j}$$

on a :

$$\min_{(j)}\{\varphi_{\alpha j}\} \le \sqrt{\lambda_\alpha}\psi_{\alpha i} \le \max_{(j)}\{\varphi_{\alpha j}\}$$

d'où :

$$\max_{(i)}\{\sqrt{\lambda_\alpha}\psi_{\alpha i}\} \le \max_{(j)}\{\varphi_{\alpha j}\}$$

De la même manière, on a :

$$\max_{(j)}\{\sqrt{\lambda_\alpha}\varphi_{\alpha j}\} \le \max_{(i)}\{\psi_{\alpha i}\}$$

prémultipliant par $\sqrt{\lambda_\alpha}$:

$$\max_{(j)}\{\lambda_\alpha\varphi_{\alpha j}\} \le \max_{(j)}\{\varphi_{\alpha j}\}$$

et finalement:: $$\lambda_\alpha \le 1$$

2) Les relations quasi-barycentriques ne sont pas des cas particuliers des relations de transitions établies lors de l'analyse générale car les matrices "de passage" ne sont pas transposées l'une de l'autre.

c – Représentation simultanée des lignes et colonnes

Les relations quasi-barycentriques justifient la représentation simultanée des lignes et des colonnes. La figure 4.2 - 3 illustre schématiquement le processus de l'analyse des correspondances.

Si les méthodes factorielles sont fondées sur le calcul des distances entre points-lignes et entre points-colonnes, la distance entre un point-ligne et un point-colonne n'a pas de sens puisque ces points sont dans des espaces différents.

L'analyse des correspondances offre cependant la possibilité de positionner et d'interpréter un point d'un ensemble relatif à un espace par rapport à l'ensemble des autres points définis dans l'autre espace.

d – Formule de reconstitution des données

Les calculs du paragraphe 1.1.5 s'appliquent également au cas de l'analyse des correspondances, en notant toutefois que les vecteurs $\mathbf{u}_\alpha$ et $\mathbf{v}_\alpha$ sont maintenant orthonormés pour les métriques $\mathbf{D}_p^{-1}$ et $\mathbf{D}_n^{-1}$. En partant des relations [4.2 - 4] et [4.2 - 5] et en suivant un raisonnement analogue à celui du paragraphe 1.2.5, on obtient la formule, pour des vecteurs φ_α et ψ_α, normés à 1 et notés $\hat{\varphi}_\alpha$ $\hat{\psi}_\alpha$:

$$f_{ij} = f_{i.}f_{.j}\sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\,\hat{\varphi}_{\alpha j}\hat{\psi}_{\alpha i} \qquad [4.2 - 12]$$

qui s'écrit aussi, en faisant intervenir la première valeur propre qui vaut *1*, et les facteurs correspondants (voir plus bas, paragraphe 4.6.2 - a) :

$$f_{ij} = f_i f_j (1 + \sum_{\alpha=2}^{p} \sqrt{\lambda_\alpha}\, \hat{\varphi}_{\alpha j} \hat{\psi}_{\alpha i}) \qquad [4.2 - 13]$$

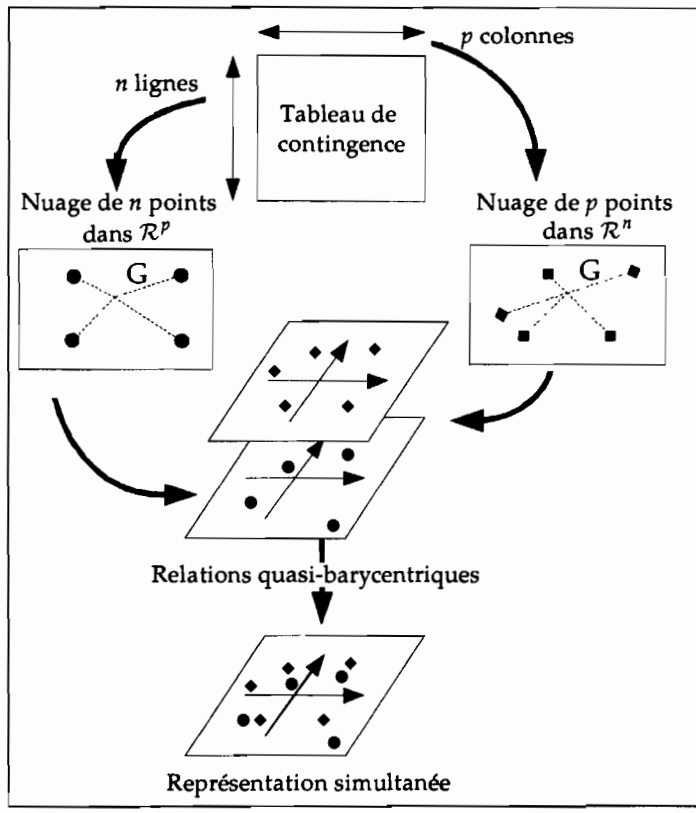


Figure 4.2 – 3. Schéma de la représentation simultanée

4.2.3 Autre présentation de l'analyse des correspondances

La *recherche de la meilleure représentation simultanée des lignes et colonnes* est une autre façon de présenter l'analyse des correspondances qui conduit directement aux formules de calculs analytiques des facteurs.

Cherchons donc à représenter sur un même axe l'ensemble des lignes et l'ensemble des colonnes, pour approcher la situation idéale suivante :

1. Chaque point-colonne j est barycentre des points-lignes i, ceux-ci étant affectés d'une masse p_i proportionnelle à leur importance dans la modalité j c'est-à-dire de la masse : $p_i = \dfrac{f_{ij}}{f_j}$

Ces masses constituent, pour chaque colonne j, les profils-colonnes du tableau de données avec $\sum_{i=1}^{n} p_i = 1$.

2. Chaque point-ligne i est barycentre des points-colonnes j, chaque point-colonne étant affecté de la masse q_j représentant la part de la modalité j dans la modalité i c'est-à-dire : $q_j = \frac{f_{ij}}{f_i}$

   Ces masses constituent, pour chaque ligne i, les profils-lignes du tableau de données avec $\sum_{j=1}^{p} q_j = 1$.

Nous définissons ainsi des relations strictement barycentriques entre les deux ensembles. Si φ_j désigne la coordonnée du point-colonne j sur un axe, et si ψ_i désigne la coordonnée du point-ligne i sur ce même axe, les conditions [i] et [ii] s'écrivent respectivement :

$$\begin{cases} \varphi = \mathbf{D}_p^{-1}\mathbf{F}'\psi \\ \psi = \mathbf{D}_n^{-1}\mathbf{F}\varphi \end{cases} \qquad \text{soit} \qquad \begin{cases} \varphi_j = \sum_{i=1}^{n} \frac{f_{ij}}{f_j}\psi_i \\ \psi_i = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}}\varphi_j \end{cases}$$

Ces relations sont en général impossibles à réaliser simultanément, car elles impliquent que chaque ensemble soit contenu dans l'autre. (Il existe une solution triviale, pour laquelle tous les points des deux ensembles sont confondus avec le point d'abscisse *1*).

Pour approcher cette situation idéale, nous cherchons un coefficient β positif et le plus proche possible de 1, tel que l'on ait les relations :

$$\begin{cases} \varphi = \beta\ \mathbf{D}_p^{-1}\mathbf{F}'\psi & [4.2-13] \\ \psi = \beta\ \mathbf{D}_n^{-1}\mathbf{F}\,\varphi & [4.2-14] \end{cases}$$

Remarquons que β est nécessairement supérieur (ou égal) à *1* sinon les relations [4.2 – 13] et [4.2 – 14] impliqueraient encore que chacun des deux ensembles recouvre un intervalle de l'axe strictement contenu dans l'intervalle recouvert par l'autre. On est donc conduit à chercher le plus petit β positif tel que ces relations soient vérifiées.

Dans [4.2 – 13], par exemple, remplaçons ψ par sa valeur tirée [4.2 – 14] :

$$\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\varphi = \frac{1}{\beta^2}\varphi$$

Prémultipliant l'équation de l'axe factoriel $\mathbf{u}$ dans $\mathcal{R}^p$ par $\mathbf{D}_p^{-1}$:

$$\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u} = \lambda\mathbf{D}_p^{-1}\mathbf{u}$$

On rappelle que les coordonnées factorielles dans $\mathcal{R}^n$ valent (formule [4.2 – 9]):

$$\varphi = \sqrt{\lambda}\mathbf{D}_p^{-1}\mathbf{u}$$

On a donc :

$$\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\varphi = \lambda\varphi$$

Et par identification, on obtient :

$$\lambda = \frac{1}{\beta^2}, \quad \text{d'où :} \quad \beta = \frac{1}{\sqrt{\lambda}}$$

Les relations [4.2 – 13] et [4.2 – 14] ne sont autres que les relations quasi-barycentriques [4.2 – 10] et [4.2 – 11] définies précédemment.

Puisque le coefficient β doit être supérieur ou égal à *1*, on démontre également de cette façon le résultat déjà établi selon lequel, en analyse des correspondances, toutes les valeurs propres sont inférieures ou égales à *1*.

On peut étendre la recherche de la meilleure représentation β-barycentrique sur un axe, à celle de la meilleure représentation (β_1,β_2)-barycentrique dans un plan repéré par deux axes orthogonaux, puis généraliser à un sous-espace de dimension quelconque.

On trouve alors la représentation simultanée fournie par l'analyse des correspondances.

Nous verrons également au chapitre 5 d'autres présentations de l'analyse des correspondances (cas particulier des analyses canoniques, discriminantes). D'autres présentations furent également données par Escoufier (1985, 1988).

4.3. Eléments pour l'interprétation des résultats

4.3.1 Inertie et formes de nuages

Les nuages de points-lignes et de points-colonnes vont être représentés dans les plans de projection formés par les premiers axes factoriels pris deux à deux. La lecture des graphiques nécessite cependant des règles d'interprétation, en particulier pour apprécier les proximités, identifier les éléments responsables de la formation des facteurs et ceux qui en sont des caractéristiques.

Ces règles s'appuient sur le bilan de l'opération de réduction que constitue la séquence des valeurs propres et des pourcentages d'inertie, ainsi que sur un ensemble de coefficients classiques : les contributions absolues et les cosinus carrés, qui seront étudiés au paragraphe 4.3.2.

La valeur de l'inertie globale n'a pas toujours une interprétation intéressante. En analyse en composantes principales normée (§ 2.4.1 – a) et, nous verrons, en analyse des correspondances multiples (chapitre 5), l'inertie totale dépend uniquement du nombre de variables. En analyse des correspondances

appliquée aux tables de contingences, l'inertie globale a une intéressante interprétation statistique.

a – Inertie et test d'indépendance

En analyse des correspondances, nous l'avons vu (cf. note du § 4.1.2 – c), la valeur de l'inertie globale est liée au test classique du χ^2.

L'inertie totale I du nuage de points par rapport au centre de gravité s'écrit par définition :

$$I = \sum_{i=1}^{n} f_i d^2(i,G) = \sum_{j=1}^{p} f_j d^2(j,G) = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}}\right)^2$$

L'effectif total étant k, on reconnaît en kI la statistique qui est asymptotiquement distribuée suivant la loi du χ^2 à $(n-1)(p-1)$ degrés de liberté (sous l'hypothèse d'indépendance) :

$$\chi^2 = kI$$

L'inertie s'exprime également par :

$$I = \sum_{\alpha=1}^{p-1} \lambda_\alpha$$

La somme des valeurs propres non triviales d'une analyse des correspondances a donc une interprétation statistique simple.

On pourra rejeter l'hypothèse nulle d'indépendance des variables en lignes et en colonnes si la valeur observée χ^2 dépasse la valeur χ_0^2 qui a une probabilité d'être dépassée inférieure à un seuil fixé au préalable[1].

La valeur de l'inertie est un indicateur de la dispersion du nuage et mesure la liaison entre les deux variables.

Cependant, on ne s'intéresse pas seulement à la dispersion globale du nuage mais surtout à l'existence de directions privilégiées dans ce nuage.

On consulte alors les inerties de chaque axe (valeurs propres) ainsi que les taux d'inertie correspondants. Cet examen nous renseigne sur la forme du nuage : forme "sphérique" (pas de direction privilégiée) ou forme non sphérique (directions privilégiées).

Le tableau 4.3 - 1 donne les valeurs des trois valeurs propres non nulles de l'analyse de la table 4.1 - 1.

---

[1] Cette façon d'opérer un test d'hypothèse correspond à l'usage des tables statistiques donnant les valeurs χ_0^2 pour chaque degré de liberté et pour certains seuils conventionnels (*0.05* ou *0.01* en général). Les logiciels donnent directement la probabilité que le χ^2 calculé soit dépassé. Il suffit alors, sans recours à une table, de comparer cette probabilité aux seuils précédents.

L'inertie totale *(0.2336)*, somme des trois valeurs propres, multipliée par l'effectif total de la table *(592)* donne la valeur *138.29* qui doit être une réalisation d'un χ^2 à *9* degrés de liberté dans l'hypothèse d'indépendance des lignes et des colonnes de la table.

Un tel χ^2 ne dépasse *21.7* que dans *1%* des cas (seuil *0.01*).

Tableau 4.3 – 1. Valeurs propres, pourcentages d'inertie pour la table 4.1 - 1

NO	VALEUR PROPRE	POUR-CENTAGE	POURCENT. CUMULE	
1	.2088	89.37	89.37	**********
2	.0222	9.51	98.89	***
3	.0026	1.11	100.00	*
Trace .2336	(= INERTIE TOTALE)			

L'hypothèse d'indépendance des couleurs des yeux et des cheveux est donc rejetée. C'est dans une telle circonstance (de *violent rejet* de l'hypothèse d'indépendance) que le recours à l'analyse des correspondances est indispensable.

D'une façon générale, deux variables sont indépendantes si les profils de leurs modalités sont identiques (aux fluctuations d'échantillonnage près) aux profils moyens : l'inertie totale est faible et il n'y a pas de direction privilégiée. Géométriquement, cela signifie que les points sont concentrés autour du centre de gravité suivant une forme sphérique.

On schématise les principaux cas sur la figure 4.3 - 1. On remarque que, dans les situations 2 et 4, les nuages ont des taux d'inertie identiques mais une inertie totale différente. Par ailleurs, les situations 3 et 4 révèlent deux nuages de même inertie totale et des taux d'inertie différents. Le test du χ^2 permet de détecter ces deux dernières situations, mais ne permet pas de mettre en évidence la situation 2 (cf. § 4.4.2).

Enfin, l'inertie d'un facteur mesure la liaison qu'il met en évidence. Elle ne peut être supérieure à *1* (cf. remarque du paragraphe 4.2.2 – b). Une valeur propre qui tend vers *1* indique une dichotomie au niveau des données ; on obtient pour chaque variable deux groupes de modalités séparant le nuage de points en deux sous-nuages. Cela peut signifier également l'existence d'un groupe de points isolés des autres points (constituant alors l'autre groupe).

Si toutes les valeurs propres sont proches de *1*, chaque modalité d'une variable est en correspondance presque exclusive avec une seule modalité de l'autre variable.

Cependant des valeurs propres faibles (signifiant que les profils sont proches du profil moyen) ne doivent pas empêcher une interprétation des axes d'inertie associés. Ceux-ci peuvent révéler une structure intéressante et plus difficilement perceptible. Ce point sera repris au paragraphe 4.4.2.

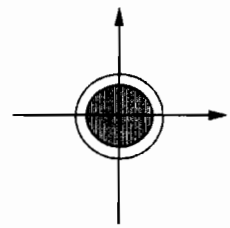
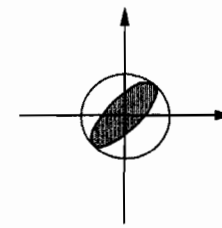
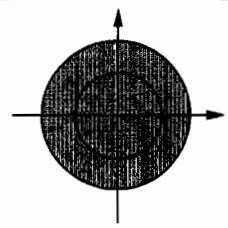
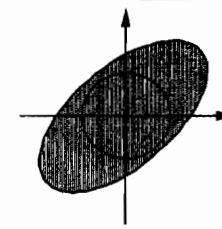
Nuage		Directions Taux d'inerties des axes *Forme "sphérique"*	*Forme "non-sphérique"*
Inertie	*Faible inertie*	 1- INDEPENDANCE • faible inertie totale • pas de direction privilégiée	 2- DEPENDANCE • faible inertie totale • direction privilégiée
	Forte inertie	 3- DEPENDANCE • forte inertie totale • pas de direction privilégiée	 4- DEPENDANCE • forte inertie totale • direction privilégiée

Figure 4.3 – 1. Indépendance et dépendances

b – Quelques formes caractéristiques de nuages de points

Envisageons quelques formes classiques de nuages afin de montrer comment la configuration du nuage de points projeté permet de réorganiser le tableau de données, par permutation des lignes et des colonnes (cf. Bastin *et al.* , 1980 ; cf. aussi : § 6.4.3, chapitre 6).

- Le nuage de points est scindé en deux sous-nuages

Le tableau de données peut être réorganisé en ordonnant les coordonnées des lignes et des colonnes sur le premier facteur (cf. figure 4.3-2). Il y a dans ce cas une valeur propre qui vaut 1 en sus de la valeur propre triviale.

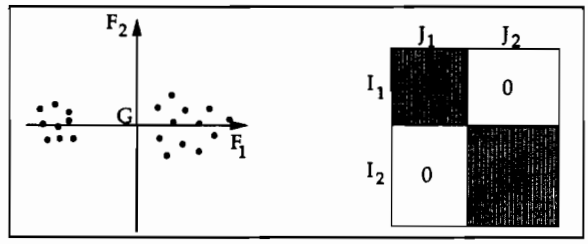


Figure 4.3 – 2. Nuage de points scindé en deux sous-nuages

- Le nuage se décompose en k sous-nuages de points (k>2)

On réorganise de la même manière le tableau de données par permutation des lignes et des colonnes. Dans un tel cas, k valeurs propres valent 1, en dehors de la valeur propre triviale.

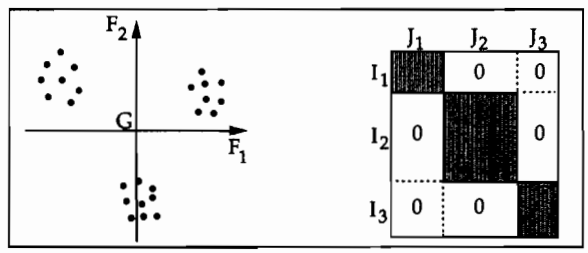


Figure 4.3 – 3. Nuage de points scindé en trois sous-nuages

- "L'effet Guttman"

On peut rencontrer aussi la situation où le nuage de points a une forme parabolique. Le tableau correspondant peut souvent être réordonné suivant une diagonale relativement chargée. Cette situation met en évidence "l'effet Guttman" qui traduit souvent une « sériation » et/ou un facteur dominant[1]. Une partie importante de l'information est alors donnée par le premier facteur.

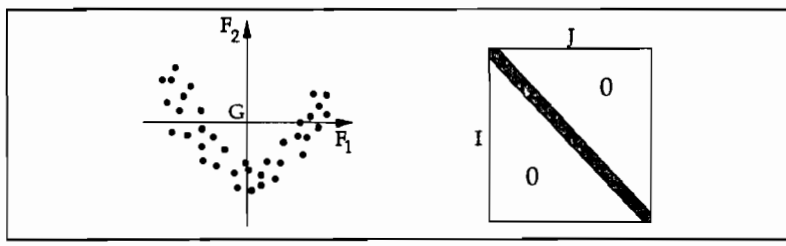


Figure 4.3 – 4. Effet Guttman et structure possible du tableau

[1] Sur l'effet Guttman en analyse des correspondances, cf. Benzécri (1973, chapitre II.B-7 et II.B-10), Heiser (1986), Van Rijckevorsel (1987) ; Tenenhaus (1994, chapitre 7, §9). Lors de l'analyse des réponses à un questionnaire, l'effet Guttman apparaît parfois lorsque les variables sont ordonnées (variables continues ou ordinales transformées en variables nominales). Un axe (souvent le premier) oppose les modalités extrêmes et un autre axe oppose les valeurs intermédiaires aux valeurs extrêmes. L'effet Guttman met parfois en évidence une structure triviale, ou une structure qui pourra être intéressante si la forme parabolique n'est pas parfaite.

Pourtant le tableau n'est pas de rang *1* et l'on disposera de *p-1* facteurs. Mais le deuxième facteur est une fonction du second degré du premier facteur, le troisième est une fonction du troisième degré, etc. L'information donnée par les axes de rang ultérieurs traduit le même phénomène. Cependant l'examen du deuxième facteur affine l'interprétation du premier axe.

4.3.2 Contributions absolues et relatives

Deux séries de coefficients apportent une information supplémentaire par rapport aux coordonnées factorielles :

- les *contributions*, appelées aussi *contributions absolues*, qui expriment la part prise par une modalité de la variable dans l'inertie (ou variance) "expliquée" par un facteur;
- les *cosinus carrés*, appelés aussi *contributions relatives* ou qualité de représentation, qui expriment la part prise par un facteur dans la dispersion d'une modalité de la variable.

C'est après l'examen de ces coefficients que l'on pourra interpréter les graphiques factoriels en tenant compte des relations de transition.

a – Contributions (ou : contributions absolues)

On cherche à connaître les éléments responsables de la construction de l'axe α. Calculons la variance des coordonnées des n points-lignes i sur l'axe α, chacun d'eux étant muni de la masse $f_{i.}$.

L'origine étant prise au centre de gravité, les coordonnées factorielles sont centrées (cf. formule [4.2 - 2]) et la variance vaut λ_α (cf. formule [4.2 - 3]).
Ainsi le quotient :

$$Cr_\alpha(i) = \frac{f_{i.}\psi_{\alpha i}^2}{\lambda_\alpha}$$

mesure la part de l'élément i dans la variance prise en compte sur l'axe α. Ce quotient est appelé *contribution* de l'élément i à l'axe α et permet de savoir dans quelle proportion un point i contribue à l'inertie λ_α du nuage projeté sur l'axe α.
On notera que pour tout axe α :

$$\sum_{i=1}^{n} Cr_\alpha(i) = 1$$

De la même façon on définit la contribution de l'élément j à l'axe α par :

$$Cr_\alpha(j) = \frac{f_{.j}\varphi_{\alpha j}^2}{\lambda_\alpha}$$

avec la relation :

$$\sum_{j=1}^{p} Cr_\alpha(j) = 1$$

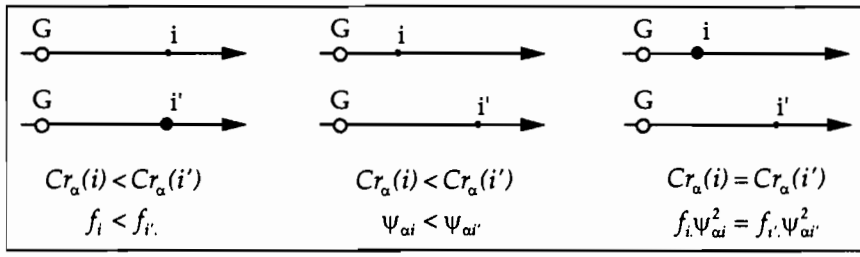


Figure 4.3 – 5. Contributions à l'axe α : trois cas de figure.

Pour trouver une éventuelle signification à un axe, on s'intéresse d'abord aux points ayant une forte contribution. Ce sont eux qui déterminent la position de l'axe (dans $\mathcal{R}^p$ pour les points *i*, et dans $\mathcal{R}^n$ pour les points *j*).

b – Cosinus carrés (ou : contributions relatives)

On cherche à apprécier si un point est bien représenté sur un sous-espace factoriel. Les axes factoriels de chaque espace constituent des bases orthonormées. Le carré de la distance d'un point au centre de gravité se décompose en somme de carrés des coordonnées sur ces axes.

Pour un point *i* de $\mathcal{R}^p$, on a :

$$d^2(i,G) = \sum_{j=1}^{p} \frac{1}{f_j} \left(\frac{f_{ij}}{f_{i.}} - f_j \right)^2$$

On remarque que la distance s'annule lorsque le profil du point est égal au profil moyen. Le carré de la projection de la variable *i* sur l'axe α vaut :

$$d^2_\alpha(i,G) = \psi^2_{\alpha i}$$

Notons que :

$$\sum_{\alpha=1}^{p-1} d^2_\alpha(i,G) = d^2(i,G)$$

Un point *i* dans $\mathcal{R}^p$ est plus ou moins proche de l'axe α. La proximité entre deux points projetés sur l'axe α correspond d'autant mieux à leur distance réelle que les points sont plus proches de l'axe.

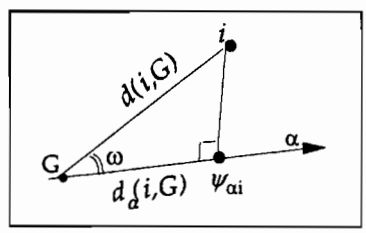


Figure 4.3 – 6. Projection du point *i* sur l'axe α

La "qualité" de la représentation du point i sur l'axe α peut être évaluée par le cosinus de l'angle entre l'axe et le vecteur joignant le centre de gravité du nuage au point i :

$$Cos^2_\alpha(i) = \frac{d^2_\alpha(i,G)}{d^2(i,G)} = \frac{\psi^2_{\alpha i}}{d^2(i,G)}$$

Cette quantité, appelée *cosinus carré*, représente la part de la distance au centre prise en compte dans la direction α. On l'appelle aussi la *contribution relative* du facteur à la position du point i. Plus le cosinus carré est proche de *1*, plus la position du point observé en projection est proche de la position réelle du point dans l'espace (figure 4.3 - 7). Notons que pour tout i :

$$\sum_{\alpha=1}^{p-1} Cos^2_\alpha(i) = 1$$

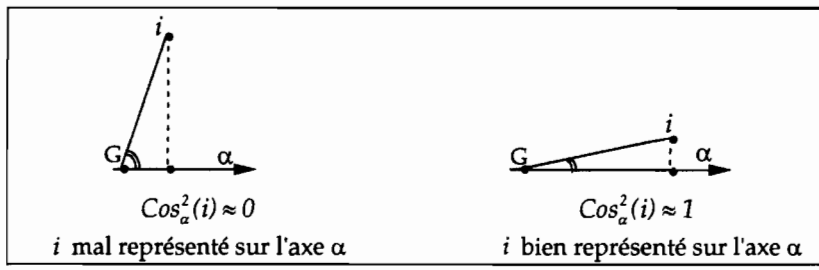


Figure 4.3 – 7. Qualité de représentation d'un point i sur l'axe α

Ce qui vient d'être dit des n points-lignes peut être transposé aux p éléments de l'autre ensemble. On mesure la contribution relative du facteur α à la position du point j par le cosinus carré de j :

$$Cos^2_\alpha(j) = \frac{\varphi^2_{\alpha j}}{d^2(j,G)}$$

et l'on a également pour tout j :

$$\sum_{\alpha=1}^{p-1} Cos^2_\alpha(j) = 1$$

Pour analyser les proximités entre points, on s'intéresse surtout aux points ayant un cosinus carré élevé. Les proximités entre ces points, observés dans le sous-espace factoriel, donnent une bonne image de leurs proximités réelles.

Remarque

Pour les contributions ainsi que pour les cosinus carrés, il n'y a pas de valeurs "seuils" à partir desquelles on peut dire que telle ou telle valeur est "forte" ou "faible". Les appréciations se font empiriquement, en fonction de l'ensemble des valeurs calculées et varient d'un jeu de données à un autre[1].

[1] Notons qu'il est usuel de multiplier par *100* les contributions de façon à exprimer en pourcentage la participation de chaque point.

c – Exemple numérique

L'exemple concerne toujours l'analyse des correspondances de la table 4.1 - 1. Les coordonnées sur le premier axe (tableau 4.3 - 2) montrent que la couleur des cheveux "blond" s'oppose à toutes les autres sur le premier axe, mais surtout à "brun". Le point "blond", avec une contribution de *71.7%* au premier axe et un cosinus carré de *0.99*, se trouve pratiquement sur cet axe et ne pourra donc pas caractériser les axes ultérieurs.

Tableau 4.3 – 2. Coordonnées, contributions, cosinus carrés pour l'analyse des correspondances de la table 1.3 - 1

COLONNES	COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
	1	2	3	1	2	3	1	2	3
CHEVEUX									
Ch.Brun	−.50	.21	−.06	22.2	37.9	21.6	.84	.15	.01
Ch.châtain	−.15	−.03	.05	5.1	2.3	44.3	.86	.04	.09
Ch.roux	−.13	−.32	−.08	1.0	55.1	31.9	.13	.81	.05
Ch.blond	.84	.07	−.02	71.7	4.7	2.2	.99	.01	.00
LIGNES	COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
	1	2	3	1	2	3	1	2	3
YEUX									
y.marron	−.49	.09	−.02	43.1	13.0	6.7	.97	.03	.00
y.noisette	−.21	−.17	.10	3.4	19.8	61.1	.54	.34	.12
y.vert	.16	−.34	−.09	1.4	55.9	31.9	.18	.77	.05
y.bleu	.55	.08	.00	52.1	11.2	.3	.98	.02	.00

Le second axe (dont on a vu qu'il correspondait à une valeur propre près de dix fois plus petite que le premier) est essentiellement construit par la couleur "roux" (*55.1 %*) qui s'oppose simultanément à "brun" et "blond". La couleur "roux" est le seul point bien représenté sur l'axe 2 (cosinus carré de *0.81*). Pour les points-lignes, le premier axe est construit presque exclusivement par les yeux "marrons" et "bleus" (contributions de *43.1%* et *52.1%*), points situés pratiquement sur l'axe (cosinus carrés de *0.97* et *0.98*), le second axe étant surtout lié aux yeux "verts". On note que la consultation des coordonnées pouvait faire penser que les yeux "noisettes" et "verts" jouaient un certain rôle dans la construction du premier axe.

La figure 4.3 - 8 qui utilise les deux premières coordonnées, montre le caractère suggestif de la représentation graphique simultanée des lignes et des colonnes. Elle permet d'interpréter les proximités ou distances entre points d'un même ensemble par leur association avec ceux de l'autre ensemble.

Pourquoi par exemple le point "ch.blond" est-il plus excentré que le point "y.bleu" sur ce premier axe très dominant ? Parce que les cheveux blonds sont beaucoup mieux caractérisés par les yeux bleus que l'inverse : d'après le tableau 4.1 - 3 (profils colonnes), *74%* des blonds ont les yeux bleus, alors que d'après le tableau 4.1 - 2 (profils lignes) *44%* des personnes ayant les yeux bleus ont des cheveux blonds.

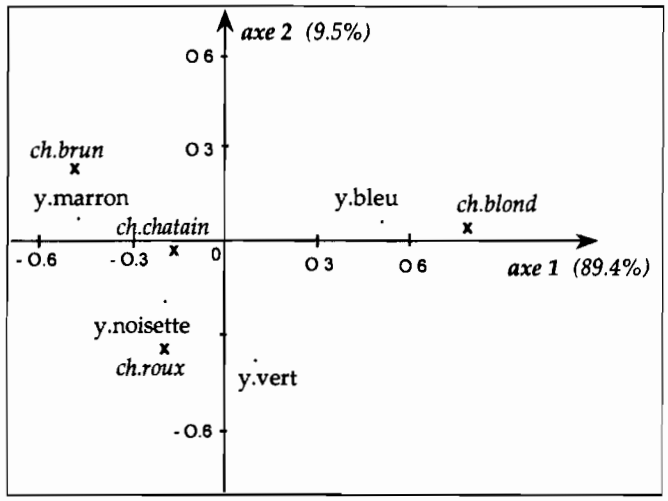


Figure 4.3 – 8. Premier plan factoriel pour l'analyse de la table 1.3 - 1

En d'autres termes, dans la relation quasi-barycentrique qui permet de positionner le point "ch.blond", le point "y.bleu" a un poids relatif de *0.74*, alors que dans la relation quasi-barycentrique qui permet de positionner le point "y.bleu", le point "blond" n'a qu'un poids relatif de *0.44*.

4.3.3 Eléments supplémentaires

On dispose, par exemple, de p_s colonnes supplémentaires qui concernent des modalités de variables nominales, analogues aux colonnes de la table de contingence. Il s'agit de situer ces nouveaux points-colonnes par rapport aux p points analysés. Soit k_{ij}^{+} la $i^{ème}$ coordonnée de la $j^{ème}$ colonne supplémentaire. Son profil est donné par :

$$\left\{\frac{k_{ij}^{+}}{k_{j}^{+}}; i = 1,2,\ldots,n\right\} \quad \text{avec} \quad k_{j}^{+} = \sum_{i=1}^{n} k_{ij}^{+}$$

On projette ce point j sur l'axe α en utilisant la même formule de transition [4.2 – 11] que pour les colonnes du tableau de fréquences :

$$\varphi_{\alpha j}^{+} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{k_{ij}^{+}}{k_{j}^{+}} \psi_{\alpha i}$$

Pour une modalité i d'une variable portée en ligne supplémentaire, on aura de façon analogue (formule de transition [4.2 – 10]) :

$$\psi_{\alpha i}^{+} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{k_{ij}^{+}}{k_{i.}^{+}} \varphi_{\alpha j}$$

A l'instar des éléments analysés, les modalités supplémentaires se calculent et s'interprètent comme des quasi-barycentres.

Remarques

1) Les éléments en supplémentaires, n'intervenant pas dans la construction du nuage, sont affectés d'un poids nul et leur contribution est donc nulle. En revanche, les cosinus carrés restent des aides à l'interprétation de ces éléments. Pour une vue d'ensemble sur le rôle et l'utilisation des variables supplémentaires en analyse des correspondances, cf. Cazes (1982).

2) La somme des cosinus carrés d'un élément supplémentaire sur l'ensemble des facteurs peut être inférieure à *1* alors que pour les éléments actifs elle est exactement égale à *1*. En effet, supposons $n>p$. Un point-colonne actif j est défini dans $\mathcal{R}^n$ mais il est situé, par l'analyse, dans l'espace factoriel à $p - 1$ dimensions. Il suffit de $p - 1$ coordonnées pour positionner cet élément. Un élément-colonne supplémentaire j^+ sera positionné dans l'espace à $p - 1$ dimensions construit par l'analyse alors qu'il appartient à $\mathcal{R}^n$. Les éléments supplémentaires ne sont donc pas entièrement contenus dans l'espace factoriel.

4.4 Méthodes et critères de validation

Nous prolongeons dans cette section les travaux évoqués dans la section 4.3 précédente dévolue à l'interprétation des résultats. Nous donnons tout d'abord des résultats sur les valeurs propres et les taux d'inertie comme paramètres caractérisant de façon globale les espaces de représentation (§ 4.4.1). Puis, à la suite du paragraphe 3.4.4 du chapitre 3 qui était consacré à la validation par *bootstrap* dans le cas des composantes principales, nous montrerons l'adaptation de cette technique au cas des correspondances (§ 4.4.2). Les principes sont les mêmes, seule change la nature statistique du tableau analysé.

4.4.1 Signification des valeurs propres et taux d'inertie

Pour toute analyse en axes principaux, qu'il s'agisse d'analyse en composantes principales ou d'analyse des correspondances, l'hypothèse d'indépendance des lignes et des colonnes d'un tableau est en général une hypothèse trop sévère pour être réaliste. Il est en effet extrêmement improbable qu'un tableau soumis à l'analyse puisse être aussi dépourvu de structure qu'une table de nombres au hasard. Bien qu'étant un cas extrême d'une portée pratique limitée, l'hypothèse d'indépendance va cependant nous permettre de définir des *seuils de signification* pour les valeurs propres et les pourcentages d'inertie, qui pourront donner des ordres de grandeur pour les utilisateurs.

La grande variété des tableaux analysables (tableaux de mesure, de classements, de comptage, etc.) rend extrêmement délicate l'interprétation de ces valeurs propres et des taux d'inertie correspondant, dont on sait qu'ils sont étroitement liés au codage des données.

Sous l'hypothèse d'indépendance des lignes et des colonnes du tableau analysé, les valeurs propres suivent des *lois paramétriques* dans le cas de l'analyse en composantes principales[1], des *lois non-paramétriques* dans le cas de l'analyse des rangs[2] et de l'analyse des correspondances des tableaux de contingence[3]. Dans ces situations favorables, il a été possible de procéder à des *tabulations approchées*, et de tracer des *abaques* qui les résument. Pour plusieurs méthodes d'analyse en axes principaux, l'utilisation des taux d'inertie (ou pourcentages de variance) comme outil d'évaluation globale de la qualité d'une représentation est délicate. Les taux d'inertie sont des mesures pessimistes de la qualité d'une représentation[4]. La variance brute initiale n'étant pas en général une mesure de référence adéquate, il est souvent injustifié de parler de *part d'information* à propos des *taux d'inertie*. Ces remarques concernent surtout l'analyse en composantes principales et l'analyse des correspondances multiples (chapitre suivant). L'analyse des correspondances des vraies tables de contingence fait exception, car, nous l'avons vu, la trace a dans ce cas une interprétation statistique.

a – Approximation de la distribution des valeurs propres

La distribution des valeurs propres en analyse des correspondances[5] sous l'hypothèse d'indépendance des lignes et des colonnes peut être approchée par celle des valeurs propres d'une matrice dont la loi est connue (matrice de Wishart). Nous avons vu au paragraphe 4.1.1 que l'hypothèse d'indépendance des lignes et des colonnes se traduit par la relation :

$$p_{ij} = p_{i.}p_{.j}$$

---

[1] Il est nécessaire de spécifier la forme analytique de la distribution des variables - loi normale - et d'estimer les paramètres correspondants.

[2] Cf. § 3.3.4; la loi de la matrice de corrélation des rangs sous l'hypothèse d'indépendance ne suppose que la continuité des distributions des variables. Des abaques approchés relatives à l'analyse des rangs figurent dans Lebart et Fénelon (1971).

[3] Comme dans le test du χ^2, test d'indépendance appliqué aux tables de contingence (voir section 4.3), la normalité résulte de la convergence de la loi multinomiale vers la loi normale.

[4] Contrairement, par exemple, aux coefficients de corrélation multiple qui sont des mesures optimistes de la qualité d'une régression.

[5] Cf. Lebart (1975 b, 1976), Corsten (1976), et dans le cas d'hypothèses plus générales O'Neill (1978, 1981). La loi des valeurs propres issues de l'analyse des correspondances a donné lieu à maintes publications erronées. Ainsi dans le traité classique de statistique de Kendall et Stuart (1961), les valeurs propres sont supposées suivre, comme l'inertie totale, des lois du χ^2. Lancaster (1963, 1969) a réfuté ce résultat en montrant que l'espérance mathématique de la première valeur propre est toujours supérieure aux valeurs découlant des assertions de Kendall et Stuart.

où p_{ij} désigne la probabilité correspondant à la case *(i,j)* et $p_{i.}$ et $p_{.j}$ les marges théoriques. p_{ij} est estimée par :

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

Ainsi k_{ij} est l'une des np composantes d'un vecteur multinomial, dont l'espérance mathématique $E(k_{ij})$ s'écrit :

$$E(k_{ij}) = k\, p_{i.} p_{.j} \text{ , avec } k = \sum_{i,j} k_{ij}$$

On fera une approximation analogue à celle qui est faite lors de l'établissement de la loi du χ^2 pour tester l'indépendance des lignes et des colonnes d'un tableau de contingence : k sera supposé suffisamment grand pour permettre l'utilisation de l'approximation normale de la loi multinomiale.

On considérera d'autre part que les marges observées $f_{i.}$ et $f_{.j}$ peuvent être substituées sans dommage aux marges théoriques $p_{i.}$ et $p_{.j}$ sans toutefois négliger les contraintes impliquées par cette substitution. Ces hypothèses permettront d'ailleurs de retrouver le test classique du χ^2 sur les tables de contingence.

Le détail des calculs figure dans l'annexe 4.6.2 de ce chapitre. Il y est établi que si λ_α est la $\alpha^{ème}$ valeur propre issue de l'analyse des correspondances d'un tableau **K** d'ordre (n, p), de somme totale k, alors la distribution de $k\lambda_\alpha$ est approximativement celle de la $\alpha^{ème}$ valeur propre d'une matrice de Wishart définie par les paramètres $W(p\text{-}1, n\text{-}1, \mathbf{I})$ [1].

b – Indépendance des taux d'inertie et de la trace (cf. annexe 4.6.3)

En analyse des correspondances, la trace mesure la dilatation générale du nuage de points-profils, alors que les taux d'inertie mesurent la forme du nuage en termes d'aplatissement et d'allongement. Ainsi, même si la trace ne permet pas de rejeter l'hypothèse d'indépendance (test habituel du χ^2), les premiers taux d'inertie pourront néanmoins être significativement élevés : l'analyse des correspondances pourra être utile même sur les tableaux que le χ^2 ne désigne pas comme étant très riches d'informations (nuage peu dilaté mais non-sphérique de points-profils).

Inversement, à une trace significativement élevée pourront correspondre des taux d'inertie non significatifs. Bien que l'hypothèse d'indépendance soit rejetée par le test du χ^2, l'analyse des correspondances n'est peut-être pas alors le meilleur outil pour décrire la dépendance entre les lignes et les colonnes de la table (nuage dilaté sphérique de points profils). Ces situations ont été schématisées par la figure 4.3 - 1 du paragraphe 4.3.1 – a : les taux d'inertie

---

[1] On trouvera une vérification expérimentale de la qualité de l'approximation montrant la concordance entre les lois théoriques des valeurs propres et celles qui résultent de l'approximation ci-dessus dans Lebart (1975 b, 1976).

significatifs ne concernent que la seconde colonne de cette figure (formes non-sphériques), alors que les χ^2 significatifs ne concernent que la seconde ligne de la figure (forte inertie correspondant à des nuages dilatés). Schématisons le modèle de l'analyse des correspondances (avec les relations et contraintes entre $\alpha, \beta, \varphi, \psi, \lambda$ qui sont les relations et contraintes usuelles de la formule de reconstitution des données) par la formule :

$$f_{ij} \approx \alpha_i \beta_j \left(1 + \sum_{h=1}^{m} \sqrt{\lambda_h}\ \varphi_h(i) \psi_h(j) \right)$$

Ce modèle peut rester pertinent même si le χ^2 ne permet pas de rejeter l'hypothèse d'indépendance, contrairement à la plupart des modélisations concernant les tables de contingence.

c – Exemples d'abaques et tables statistiques

Les tables statistiques établies par simulation et les abaques qui en résultent permettent d'apprécier le degré de signification de la plus grande valeur propre issue de l'analyse des correspondances de tableaux de contingence depuis la dimension (*6×6*) jusqu'à la dimension (*50×100*). La figure 4.4- 1 donne les valeurs médianes du pourcentage d'inertie relatif à la plus grande valeur propre pour les largeurs $p = 6, 8, 10, 20, 30, 40, 50$ [1].

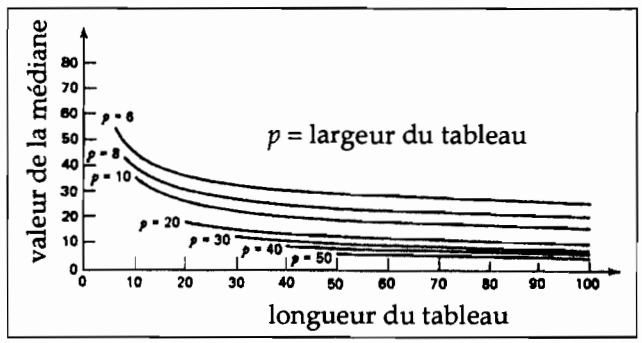


Figure 4.4- 1. Valeurs médianes du pourcentage d'inertie de la plus grande valeur propre

Les estimations des valeurs des taux d'inertie correspondant à la première valeur propre apparaissent sur la figure 4.4- 2 (pour un seuil de *0.05*). Les extrémités des courbes (points : *6×6, 8×8, 10×10*) ont été établies à l'aide de *1000* simulations (*100* pour les autres points), afin de préciser leur tracé. Ces figures schématiques ne donnent cependant que des ordres de grandeurs. Par exemple, on lit sur la figure 4.4- 2 que, pour un tableau *10×10*, la première valeur propre

[1] Des informations plus détaillées concernant la construction de ces abaques (notamment sur les modes de génération de tableaux pseudo-aléatoires) et des *tables approchées*, pour les tableaux dont les dimensions n'excèdent pas *50×100* relatives aux cinq premières valeurs propres sont données dans Lebart (*op cit.*).

peut atteindre ou dépasser *40%* de l'inertie (la loi des taux ne dépendant pas de l'effectif total du tableau) dans *5%* des cas, sous l'hypothèse d'indépendance des lignes et des colonnes de la table.

Il s'agit donc ici d'un test de sphéricité du nuage de points-profils, qui ne remplace pas un test sur les valeurs propres elles-mêmes (il faut alors tabuler $k\lambda_1$, car la loi de λ_1 seule dépend de k, effectif total de la table). Ce test donne néanmoins des ordres de grandeur ayant une certaine valeur pédagogique sur l'effet de fluctuations d'échantillonnage sur la forme de nuage de points-profils. En revanche, ce type de résultats étendu à l'ensemble des valeurs propres sous l'hypothèse d'indépendance ne peut aider à déterminer le nombre d'axes à retenir, car les valeurs propres ne sont pas indépendantes (même sous l'hypothèse d'indépendance des lignes et des colonnes du tableau, et *a fortiori* si cette hypothèse est rejetée). Il faudrait donc connaître la loi conditionnelle de la seconde valeur propre, ce qui ne peut donner lieu à des résultats généraux.

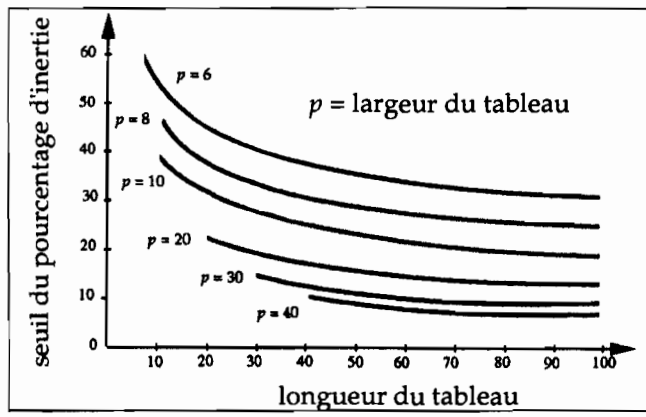


Figure 4.4- 2. Seuil (0,05 unilatéral) du pourcentage d'inertie de la plus grande valeur propre

d – Autres critères de choix statistiques, résultats asymptotiques

Dans le cas de l'analyse des correspondances des vraies tables de contingence, la loi des valeurs propres ne permet que de juger la signification du premier axe, puisque les lois conditionnelles des autres valeurs propres ne sont pas connues. Une procédure approchée (Malinvaud, 1987) peut être utilisée pour déterminer le rang à partir duquel les valeurs propres ne sont plus significativement différentes entre elles.

Revenons au "modèle" que représente la formule de reconstitution approchée avec m facteurs (cf. formule [4.2-12] du § 4.2) des fréquences relatives f_{ij} du tableau de contingence **K** d'ordre (n, p) de terme général k_{ij} et d'effectif total k.

$$f_{ij} \approx g_{ij} = \alpha_i \beta_j (1 + \sum_{h=1}^{m} \sqrt{\lambda_h}\ \varphi_h(i) \psi_h(j)) \qquad [4.4\text{- }1]$$

Les restrictions suivantes sont imposées aux différents paramètres du modèle (moyenne nulle, variance unité, et orthogonalité des facteurs) :

$$(a)\quad \sum_{i=1}^{n} \alpha_i \varphi_h(i) = \sum_{j=1}^{p} \beta_j \psi_h(j) = 0, \quad \text{pour} \quad h \le m$$

$$(b)\quad \sum_{i=1}^{n} \alpha_i \varphi_h^2(i) = \sum_{j=1}^{p} \beta_j \psi_h^2(j) = 1, \quad \text{pour} \quad h \le m$$

$$(c)\quad \sum_{i=1}^{n} \alpha_i \varphi_h(i) \varphi_{h'}(i) = \sum_{j=1}^{p} \beta_j \psi_h(j) \psi_{h'}(j) = 0, \quad \text{pour} \quad h \ne h'$$

$$(d)\quad \lambda_h \ge 0, \quad \text{pour} \quad h \le m$$

On note que $g_{ij} = f_{ij}$ si $m = p - 1$ (modèle dit *saturé*). Il ressort de la première ligne de contraintes que : $\alpha_i \beta_i = f_{i.} f_{.j}$.

Le tableau reconstitué dans le cas où $m = 0$ correspond à l'indépendance entre les lignes et les colonnes du tableau. Pour savoir si cette hypothèse est rejetée, on calcule la statistique du χ^2 usuelle (à $(n-1)(p-1)$ degrés de liberté) :

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

Plus généralement, pour $m \ge 1$ fixé, l'ajustement du modèle [4.4- 1] en rendant minimal le critère sous les contraintes précédentes :

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

fournit la reconstitution $\hat{g}_{ij}$ de f_{ij} à partir de m facteurs.

La statistique $X^2 = k \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{(f_{ij} - \hat{g}_{ij})^2}{\hat{g}_{ij}}$, mesurant l'écart entre le modèle saturé f_{ij} et le modèle estimé $\hat{g}_{ij}$, doit suivre une loi du χ^2 dont le nombre de degrés de liberté $d(m) = (n-m-1)(p-m-1)$ s'obtient en retranchant de la dimension de l'espace le nombre de paramètres indépendants qui est le nombre de paramètres estimés duquel on retranche le nombre de contraintes qui peuvent être écrites sous la forme de fonctions admettant des dérivées partielles du premier ordre continues. (cf. Cramer, 1946 ; Rao, 1973). Chaque nouveau facteur (de rang m) demande en effet l'estimation de $n + p + 1$ nouveaux paramètres (φ_h, ψ_h et λ_h), liés par 2 contraintes de centrage (a), 2 contraintes de normalisation (b), $(m-1)$ contraintes d'orthogonalité (c), d'où :

$$d(m) = (n-1)(p-1) - m(n+p-2-m), \quad \text{soit :} \quad d(m) = (n-m-1)\,(p-m-1)$$

Utilisée avec prudence, cette statistique est plus utile que celles déduites des études de distribution des valeurs propres sous l'hypothèse d'indépendance. Il faut cependant éviter de procéder à des approximations dans le calcul de la statistique X^2. On peut effet être tenté de remplacer le dénominateur $\hat{g}_{ij}$ par $f_{i.} f_{.j}$,

et d'approcher alors X^2 par la somme des valeurs propres de rang supérieur à m. Cette approximation conduit à comparer la $m^{ième}$ valeur propre λ_m à un χ^2 à $(n + p - 2m - 1)$ degrés de liberté. Ce résultat, proposé par plusieurs auteurs, a été réfuté par Lancaster (1963), qui a montré que la plus grande valeur propre sous l'hypothèse d'indépendance a une espérance supérieure à $(n + p - 3)$.

e - Régions de confiances analytiques

Bien que les techniques de *multidimensional scaling* [1] ne soient pas traitées dans cet ouvrage, il faut mentionner, pour leur intérêt méthodologique, les travaux de Ramsay (1978) (zones de confiances fondées sur la distribution des distances entre individus pour la méthode dite *MULTISCALE*). On mentionnera également les ellipses de confiance proposées par Saporta et Hatabian (1986), qui s'appliquent à toute catégorie de variable nominale supplémentaire.

Gifi (1981, 1990) a également proposé des ellipsoïdes de confiance fondés sur la méthode *delta* (cf. par exemple, Rao, 1973 ; Efron, 1982). Cette méthode généralise au cas multidimensionnel le résultat simple suivant :

Proposons-nous de calculer la variance et la loi asymptotique d'une (bonne) fonction $g(X)$ d'une variable aléatoire X de moyenne μ et de variance σ^2. A partir du développement de Taylor de g autour de μ : $g(t) \approx g(\mu) + (t - \mu)\, g'(\mu)$, on déduit immédiatement : $var\,[g(X)] \approx g'(\mu)^2 \sigma^2$.

Plus généralement, la méthode *delta* est fondée sur le résultat suivant :

Si l'on a une fonction $\mathbf{y}_n = \Phi(\mathbf{x}_n)$ d'une suite $\mathbf{x}_n$ de vecteurs aléatoires tels que $\sqrt{n}(\mathbf{x}_n - \mu)$ est asymptotiquement normal de moyenne nulle et de matrice des covariances Σ (Φ est supposée différentiable en μ), alors $\sqrt{n}\,(\mathbf{y}_n - \Phi(\mu))$ est aussi asymptotiquement normal de moyenne nulle et de matrice des covariances $\mathbf{V}(\sqrt{n}\,\mathbf{y}_n) = \partial\Phi(\mu)\,\Sigma\,\partial\Phi(\mu)$, où $\partial\Phi(\mu)$ est la matrice des dérivées partielles de Φ au point μ. Si les composantes de $\mathbf{y}_n$ sont les coordonnées d'un point sur deux axes factoriels et les composantes de $\mathbf{x}_n$ sont les éléments du tableau de données, la méthode *delta* permet d'estimer la matrice des covariances de $\mathbf{y}_n$, et donc de construire des zones de confiance ellipsoïdales autour du point correspondant. Des formules analogues à la formule [1.5.2 - 4] (§ 1.5.2.1 de l'annexe 2 du chapitre 1) permettent d'estimer $\partial\Phi(\mu)$. Dans le calcul de $\mathbf{V}(\sqrt{n}\,\mathbf{y}_n)$, les valeurs théoriques sont remplacées par leurs estimations empiriques. Comme les zones de confiance *bootstrap*, (§ 4.4.2) avec lesquelles la compatibilité empirique semble bonne, les zones déterminées par la méthode delta peuvent concerner les variables actives.

---

[1] Techniques de représentations de systèmes de distances entre points, développées autour des *Bell Laboratories* et de la revue *Psychometrika*, avec, à l'origine, des contributions de Shepard, Guttman, Kruskal, Carroll (cf. par exemple Kruskal et Wish, 1978 ; Schiffman *et al.*, 1981). Cf. également l'article de synthèse de Drouet d'Aubigny (1993).

4.4.2 Bootstrap pour l'analyse des correspondances

Comme pour l'analyse en composantes principales, le bootstrap est un outil privilégié pour étudier la stabilité des formes. Une application à l'exemple d'analyse des correspondances des sections précédentes nous montre la simplicité et l'efficacité de la méthode. Tout ce qui a été dit au chapitre 3 à propos de l'analyse en composantes principales peut être transposé ici : possibilité de bootstrap partiel ou total, et, pour le bootstrap total, existence de trois options plus ou moins sévères (type 1 : simple correction du signe des axes ; type 2 : correction, en plus, des interversions d'axes ; enfin type 3 : rotations procrustéennes). Ce qui va changer en analyse des correspondances, c'est le mode de construction des réplications. Il ne s'agit plus de tirer avec remise des lignes du tableau, mais des individus dont les effectifs emplissent les cases de la table de contingence.

a – Le principe des réplications

Reprenons les données du tableau 4.1 - 1. Une simulation bootstrap consiste à tirer avec remise les $k = 592$ personnes (chacune d'entre elle appartenant à une case (i,j) du tableau 4.1- 1). Cela revient à faire autant de tirages selon une loi multinomiale dont les probabilités de tirage sont : $p_{ij} = k_{ij}/k$. On peut vérifier (empiriquement) qu'il est équivalent, au niveau des résultats de la simulation, d'utiliser l'approximation normale de la loi multinomiale, c'est-à-dire de générer une réalisation d'une variable normale de moyenne k_{ij} et de variance $k_{ij}(1-k_{ij})/k$ (la valeur ainsi générée sera arrondie à l'entier supérieur)[1].

Le tableau 4.4- 1 donne un exemple de deux tableaux générés de cette façon. Pour cet exemple, on va générer *20* réplications, chiffre largement suffisant, on va le voir, pour donner une bonne idée de la stabilité des résultats.

Notons qu'une analyse des correspondances faite sur un seul tableau répliqué suffit à donner une forte présomption de stabilité. L'observation, au sens près des axes, du même *pattern* (de la même forme) signifie que la structure observée a résisté à la perturbation constituée par la simulation. Il est en effet improbable de retrouver par hasard un agencement complexe de points. C'est là une différence fondamentale avec la statistique uni-dimensionnelle, pour laquelle une réplication isolée n'est pas utilisable. Cependant, dans la plupart des cas, la structure est partiellement déformée et l'on souhaite pouvoir isoler ses éventuelles parties stables. C'est alors que la répétition des réplications est utile, pour limiter la subjectivité dans les appréciations.

---

[1] On parle dans ce cas de bootstrap paramétrique. Les deux cases d'effectif faible (<12) pour lesquelles une telle approximation est discutable ont en fait une influence quasi nulle sur les résultats. Cela ne serait pas le cas si une colonne (ou une ligne) entière avait des effectifs faibles. Pour la génération de variables pseudo-aléatoires normales, cf. Neave (1973), Brent (1974).

Tableau 4.4.1. Table originale (tableau 4.1-1), suivie de deux tables répliquées

		couleur des cheveux			
		Brun	**Châtain**	**Roux**	**Blond**
Original					
couleur des yeux	marron	*68*	*119*	*26*	*7*
	noisette	*15*	*54*	*14*	*10*
	vert	*5*	*29*	*14*	*16*
	bleu	*20*	*84*	*17*	*94*
Réplication 1					
couleur des yeux	marron	*79*	*120*	*23*	*9*
	noisette	*14*	*60*	*15*	*12*
	vert	*3*	*29*	*16*	*9*
	bleu	*21*	*82*	*20*	*110*
Réplication 2					
couleur des yeux	marron	*72*	*111*	*32*	*7*
	noisette	*14*	*47*	*13*	*14*
	vert	*5*	*30*	*15*	*19*
	bleu	*20*	*89*	*16*	*98*

b – Les zones de confiance

Il existe plusieurs façon de mettre à profit ces *20* réplications pour construire des intervalles de confiance.

► Projeter en éléments supplémentaires les lignes (et colonnes) simulées dans les plans factoriels issus de l'analyse de la table de contingence initiale (qui est pour ce modèle, l'espérance des matrices simulées). C'est le bootstrap partiel.

► Procéder à *20* analyses indépendantes correspond au bootstrap total, dont le principe a été défini dans la section 3.4.4.

Le bootstrap partiel est illustré par la figure 4.4 – 3, qui représente les ellipses de confiance dans le premier plan factoriel, ou plan (F_1, F_2), de l'analyse des correspondances de la table à *4* colonnes et *4* lignes originale. On voit que les ellipses de confiance des points correspondant à des colonnes homologues (couleur des cheveux : identificateurs soulignés) sont bien séparés, avec cependant une faible intersection des couleurs Roux et Châtain. Tous les points lignes (couleur des yeux) sont bien séparés.

Le bootstrap total est illustré par la figure 4.4-4 qui reprend le même plan principal que la figure 4.4-3. Il s'agit ici de la version la plus sévère de bootstrap total, puisque ni les interversions d'axes, ni les éventuelles rotations des axes ne sont corrigées. Bien que les ellipses aient des formes sensiblement différentes, les zones de confiance autour des points ont des amplitudes similaires. Nous avons ici un exemple de structure stable, puisque celle-ci est retrouvée, pratiquement sans interversion d'axes ni de rotations par *20* analyses indépendantes. Il n'est donc pas utile ici de procéder à des validations bootstrap

intermédiaires (type 2 et 3) puisqu'il y a convergence de l'option la plus sévère du bootstrap total (type 1) avec le bootstrap partiel.

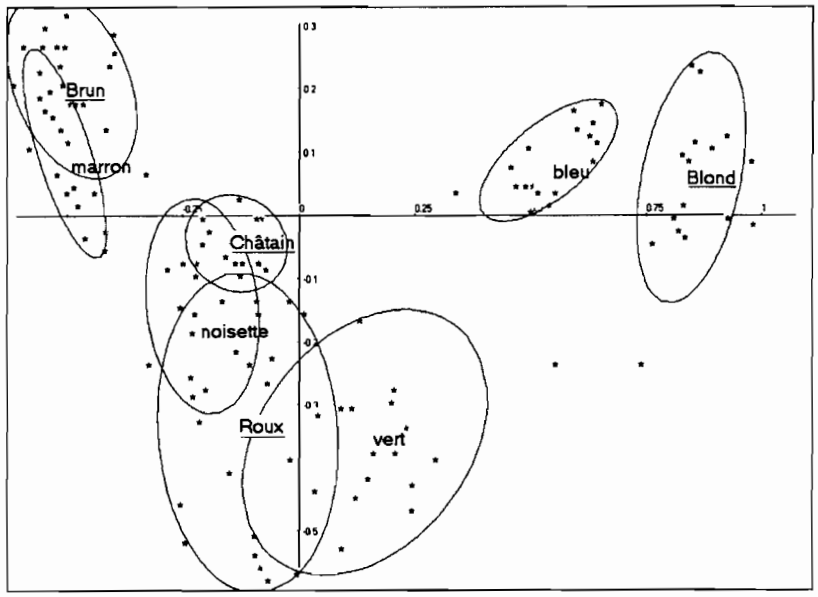


Figure 4.4.3. Zone de confiance (bootstrap partiel).
Plan principal de l'analyse des correspondances de la table 4.1.1.

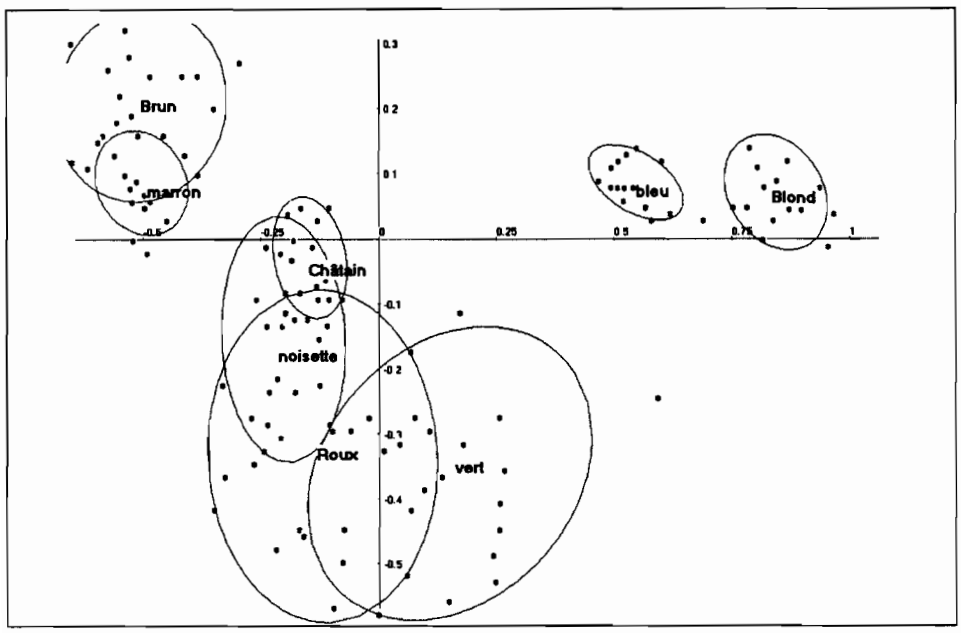


Figure 4.4.4. Zone de confiance : bootstrap total type 1 :
[simple correction des changements éventuels de signe des axes]

4.5 Exemple d'application

L'exemple considéré ici est extrait de l'*Enquête Budget-temps Multimédia 1991-1992* du CESP, comme l'exemple d'analyse en composantes principales traité au paragraphe 2.5.1. Il repose sur l'analyse d'un tableau de contingence qui croise *8* professions et catégories socioprofessionnelles (PCS) et *6* types de médias pour un échantillon de *12 388* "contacts média" relatifs à *4433* personnes interrogées. L'individu statistique sera pour nous le "contact média" et non la personne interrogée dans l'enquête. Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population enquêtée telles que le sexe, l'âge, le niveau d'instruction.

4.5.1 Données et premiers résultats

Nous disposons des tables de contingence suivantes (cf. tableau 4.5 – 1). Pour le premier blocs **K** de *8* lignes (lignes actives) on trouve, à l'intersection de la ligne *i* et de la colonne *j* le nombre k_{ij} d'individus appartenant à la catégorie *i* et ayant eu la veille (un jour de semaine) au moins un contact avec le type de média *j*.

Tableau 4.5 – 1. Tables de contingence croisant les types de contacts-média (colonnes) avec professions, sexe, âge, niveau d'éducation (lignes).

	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782
Sexe						
Homme	1630	1900	285	854	621	776
Femme	1667	2069	152	815	683	938
Age						
15-24 ans	660	713	69	216	234	360
25-34 ans	640	719	84	230	212	380
35-49 ans	888	1000	130	429	345	466
50-64 ans	617	774	84	391	262	263
65 ans ou +	491	761	70	402	251	245
Education						
Primaire	908	1307	73	642	360	435
Secondaire	869	1008	107	408	336	494
Techn. prof.	901	1035	80	140	311	504
Supérieur	619	612	177	209	298	281

Les blocs suivants (lignes supplémentaires) s'interprètent de façon analogue. Une personne interrogée pouvant avoir des contacts avec plusieurs médias, les sommes en ligne représentent des "nombres de contacts".

Tableau 4.5 – 2. Valeurs propres, pourcentages d'inertie pour la table K "Professions-Contacts média" (8 premières lignes de la table 4.5 - 1)

NUM.	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0139	62.20	62.20	**********
2	.0072	32.37	94.56	**********
3	.0008	3.70	98.26	**
4	.0003	1.36	99.63	*
5	.0001	.37	100.00	*
Trace	.0223			

Tableau 4.5 – 3. Poids relatifs (P.REL), Distances à l'origine (DIS), coordonnées, contributions et cosinus carrés des éléments sur les trois premiers axes

	FREQUENCES		COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
LIBELLES	P.REL	DIS	1	2	3	1	2	3	1	2	3
COLONNES ACTIVES											
Radio	26.61	.00	−.01	.02	−.05	.4	1.8	70.4	.08	.17	.75
Télévision	32.04	.00	.05	.00	.02	6.6	.0	10.5	.85	.00	.08
Quotid.natio	3.54	.29	−.54	−.01	.02	74.6	.0	1.8	.99	.00	.00
Quotid regio	13.46	.02	.11	−.11	.01	11.5	22.4	.4	.49	.49	.00
Presse Mag	10.52	.03	−.09	−.13	.02	6.8	25.6	4.5	.32	.62	.01
Pres.Mag. T.V.	13.84	.03	.01	.16	.03	.1	50.1	12.4	.00	.96	.03
LIGNES ACTIVES											
Agriculteur	2.86	.13	.17	−.31	−.07	5.7	38.0	17.9	.21	.74	.04
Petit patron	3.51	.03	.07	−.14	−.06	1.2	10.0	17.7	.15	.67	.14
Prof.Cadre Sup	5.62	.19	−.43	−.06	.00	75.0	2.9	.1	.98	.02	.00
Prof. interm	10.15	.01	−.11	.03	−.03	8.3	1.5	11.8	.80	.08	.07
Employé	14.98	.01	.02	.10	−.01	.3	18.9	.5	.03	.93	.00
Ouvrier qual	11.16	.01	.04	.10	−.02	1.5	15.9	5.1	.14	.74	.03
Ouvrier n−q	4.40	.02	.12	.09	−.04	4.4	5.5	8.4	.56	.36	.06
Inactif	47.32	.00	.03	−.03	.03	3.6	7.3	38.7	.37	.39	.24
LIGNES ILLUSTRATIVES (SUPPLEMENTAIRES)											
Homme	48.97	.01	−.05	−.02	−.01	.0	.0	.0	.48	.11	.02
Femme	51.05	.00	.05	.02	.01	.0	.0	.0	.49	.10	.02
15−24 ans	18.18	.02	−.02	.10	−.04	.0	.0	.0	.02	.56	.08
25−34 ans	18.28	.02	−.03	.12	−.01	.0	.0	.0	.05	.87	.01
35−49 ans	26.30	.00	−.03	.01	−.01	.0	.0	.0	.61	.10	.07
50−64 ans	19.30	.01	.02	−.10	.00	.0	.0	.0	.05	.80	.00
65 ans ou +	17.92	.03	.07	−.14	.07	.0	.0	.0	.14	.58	.16
Primaire	30.07	.03	.13	−.08	.02	.0	.0	.0	.63	.24	.02
Secondaire	26.01	.00	.00	.04	.00	.0	.0	.0	.00	.69	.00
Techn. prof.	23.98	.07	−.03	.18	−.04	.0	.0	.0	.01	.46	.02
Supérieur	17.73	.09	−.29	−.02	−.01	.0	.0	.0	.99	.00	.00

Il y a *12 388* contacts pour *4433* individus concernés. Les chiffres publiés ici ayant été arrondis après un redressement, les totaux relatifs aux différentes partitions de la population peuvent ne pas coïncider. On cherche à décrire les

éventuelles affinités entre les groupes socioprofessionnels et les différents types de médias. L'analyse des correspondances de la table **K** conduit aux valeurs propres consignées dans le tableau 4.5 – 2.

Le produit de la trace $t = 0.0223$ par l'effectif total $k = 12\ 388$ vaut : $kt = 276.25$.

Dans l'hypothèse d'indépendance des lignes et des colonnes de la table, kt serait une réalisation d'un χ^2 à 35 degrés de liberté (noté χ^2_{35}) [$35 = (8 - 1)(6-1)$]. Lorsque le nombre de degrés de liberté ν dépasse 30, on considère que la variable $u = \frac{\chi^2_\nu - \nu}{\sqrt{2\nu}}$ est une variable normale (de Laplace-Gauss) centrée réduite. Ici, $u = 28.8$ (28.8 écarts-types de la moyenne). L'hypothèse d'indépendance est évidemment rejetée.

Deux facteurs sont dominants et représentent près de 95% de l'inertie totale. Les coordonnées et les aides à l'interprétation correspondants figurent dans le tableau 4.5 – 3. Celui-ci donne également les coordonnées et les cosinus carrés des lignes supplémentaires.

4.5.2 Visualisation et interprétation

On note que l'élément "Quotidien national" dont la fréquence relative (colonne P.REL) est très faible (3.54%) a une distance au point moyen (colonne DIS) très élevée : le profil correspondant est donc atypique. Il contribue pour 74.6% à la construction du premier axe, qui en est très proche (cosinus carré : 0.99). Ce point contribue essentiellement à la construction de cet axe et ne participera pas à la construction des autres axes. Ce même premier axe est caractérisé par la ligne active "Prof.Cadre" (profession libérale, cadres supérieurs) et par la ligne supplémentaire "Supérieur" (niveau d'étude supérieur).

Le second axe sépare la "Presse Magazine de Télévision" (associée aux catégories employés et ouvriers, et aux classes d'âges plutôt jeunes) de la presse magazine (Presse TV exclue) et de la presse quotidienne régionale, toutes deux associées aux agriculteurs et aux petits patrons, et à des catégories d'âge plus élevées. Les figures 4.5 – 1 et 4.5 – 2 résument ce réseau d'associations.

Il est clair dans une analyse de ce type que le premier axe correspond à une interprétation ponctuelle : les contacts média avec la presse quotidienne nationale sont, de façon significative, surtout le fait de cadres supérieurs et/ou de personnes d'un haut niveau d'éducation.

Ce résultat n'est cependant pas d'emblée visible sur le tableau 4.5 – 1.

En revanche, les positions des points sur les deux figures donnent une interprétation plus nuancée du second axe : les professions salariées, de niveau d'éducation moyen, composées surtout de jeunes (contact média : Presse magazine TV), s'opposent aux petits patrons et agriculteurs, en moyenne

sensiblement plus âgés et moins instruits (contacts : presse magazine autre que TV, et presse quotidienne régionale).

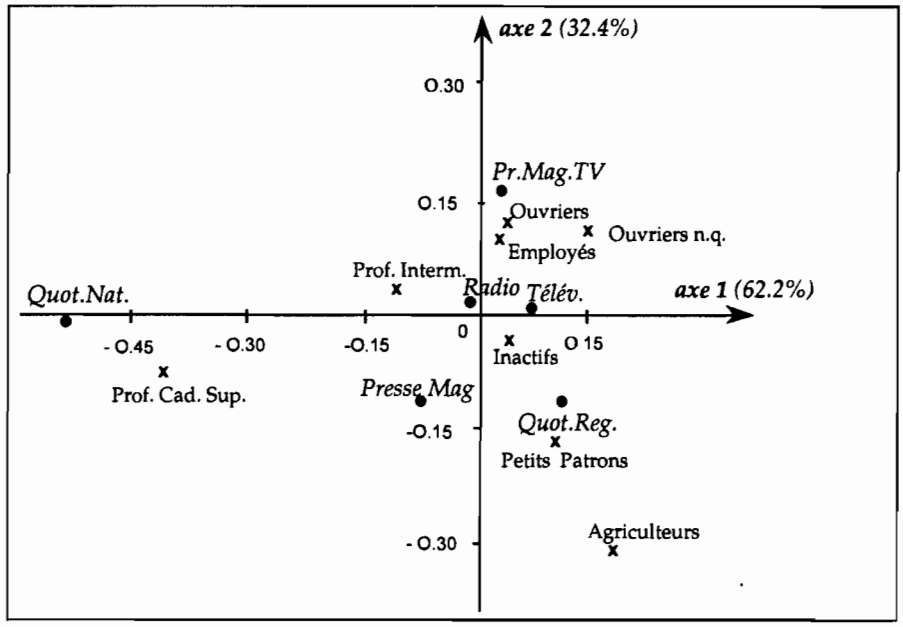


Figure 4.5 – 1. Variables actives dans le premier plan factoriel

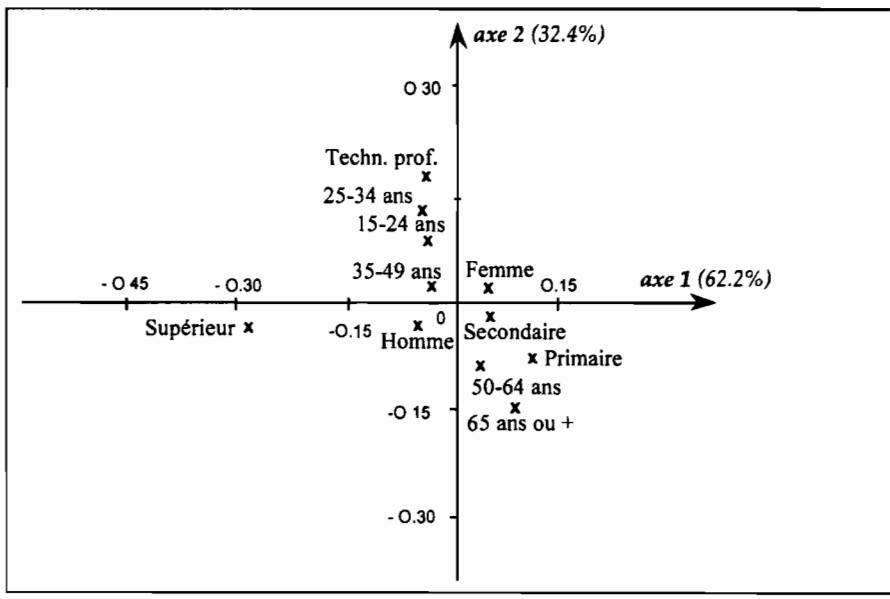


Figure 4.5 – 2. Variables supplémentaires ou illustratives dans le premier plan factoriel

Que se passe-t-il si l'on supprime, au sein des colonnes actives, la colonne "Quot. N." dont le rôle est prédominant, pour la positionner en élément supplémentaire ?

On a vu que cette colonne est presque située sur l'axe *1* (cosinus carré de *0.99*). Sa suppression enlèverait *74.6%* de l'inertie dans cette direction (valeur de la contribution), et donc l'inertie dans cette direction serait inférieure à celle du second axe actuel[1] sur lequel la colonne supprimée a d'ailleurs une contribution nulle.

Donc le nouveau premier axe d'inertie maximale sera très voisin de l'ancien second axe. Tous calculs faits, on trouve, après suppression de la colonne en question, une première valeur propre de *0.0074* (la seconde valeur propre valait *0.0072*) et des coordonnées sur ce nouveau premier axe qui diffèrent d'au plus de *0.01* de celles de l'ancien second axe.

Le nouveau second axe (sur lequel la colonne supplémentaire "Presse Quot." a une coordonnée de *0.54* et un cosinus carré de *0.88*) est très voisin de l'ancien premier axe.

Cet exemple aura illustré le positionnement de lignes supplémentaires et de colonnes supplémentaires, l'usage simultané des trois types d'aides à l'interprétation (valeurs propres, contributions, cosinus carrés) ainsi que le caractère itératif de l'analyse, qui fait penser à un "épluchage" progressif des nuages de points profils.

4.5.3 Validation par bootstrap

Reprenons le sous-tableau du tableau 4.5 - 1 correspondant aux seules lignes actives.

La simulation bootstrap classique consiste à tirer avec remise les $k = 12\,888$ contacts-média. On a vu que cela revient à faire autant de tirages selon une loi multinomiale dont les probabilités de tirage sont : $p_{ij} = k_{ij} / k$.

Le tableau 4.5- 4 donne un exemple de deux tableaux générés de cette façon. On va, pour cet exemple, générer *30* réplications, chiffre largement suffisant, on va le voir, pour donner une bonne idée de la stabilité des résultats.

La figure 4.5- 3 représente le premier plan factoriel, ou plan (F_1, F_2), de l'analyse des correspondances de la table à *6* colonnes et *8* lignes, avec positionnement des réplications des lignes, et ajustement par des ellipses des *8* nuages de réplications.

---

[1] *25.4* %(complément à *100* de *74.6* %) de *0.0139* (première valeur propre λ_1) est en effet très inférieur à *0.0072* (seconde valeur propre λ_2).

Tableau 4.5- 4. Exemple de deux réplications des valeurs du tableau 4.5- 1

			Réplication 1			
1	109	120	1	78	48	20
2	126	142	8	76	53	30
3	196	181	80	77	109	72
4	384	365	60	133	138	203
5	514	596	59	228	172	316
6	378	467	33	171	100	223
7	169	188	8	79	38	81
8	1519	1961	158	893	632	764
			Réplication 2			
1	83	138	3	79	62	19
2	142	142	8	82	50	26
3	198	163	63	68	114	85
4	359	367	73	155	132	196
5	503	561	56	266	173	294
6	395	432	25	171	104	220
7	149	179	16	74	50	83
8	1488	1919	182	852	611	794

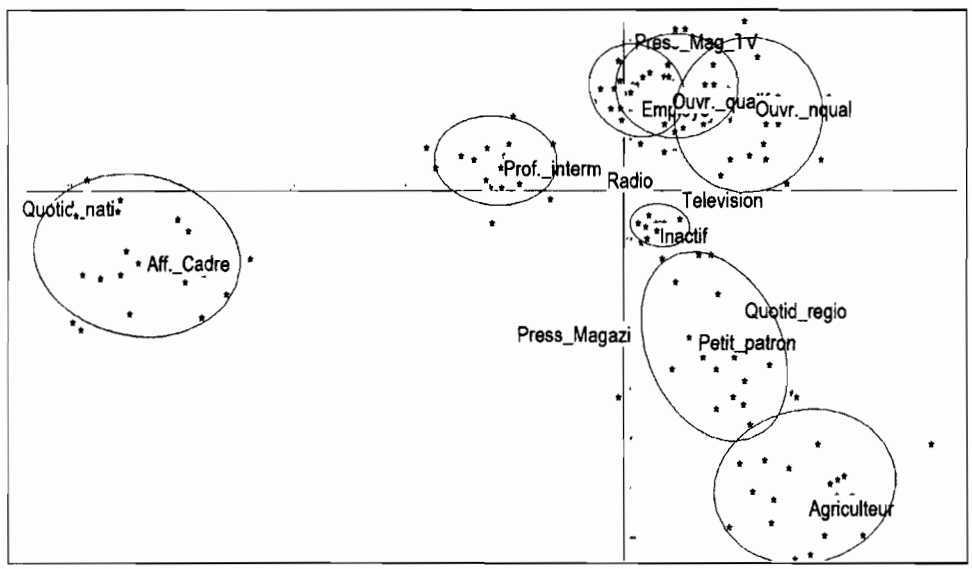


Figure 4.5 – 3. Zones de confiance "bootstrap partiel" pour les lignes actives Plan factoriel principal relatif au tableau 4.5 – 1

On voit que les ellipses de confiance des points correspondant à des lignes homologues (catégories socio-professionnelles) sont bien séparées, à l'exception des catégories 5 et 6 (employés et ouvriers qualifiés). Bien entendu, la même procédure peut être appliquée aux points-colonnes (contact-média).

La forme, mais aussi la taille des ellipses apportent une information supplémentaire par rapport à la figure 4.5- 1.

Ainsi, on peut affirmer que les ouvriers non qualifiés ont un comportement en continuité, mais cependant résolument distinct de celui des ouvriers qualifiés.

Même observation en ce qui concerne l'absence de solution de continuité entre agriculteurs et petits patrons.

La figure 4.5- 4 représente le second plan factoriel, ou plan (F_2, F_3), de la même table. On retrouve sur l'axe horizontal les distinctions observées sur l'axe vertical précédent, mais la confusion est totale sur l'axe *3* (vertical).

Seule la classe *8* (inactifs) occupe une position typée sur l'axe *3* en s'opposant aux autres classes.

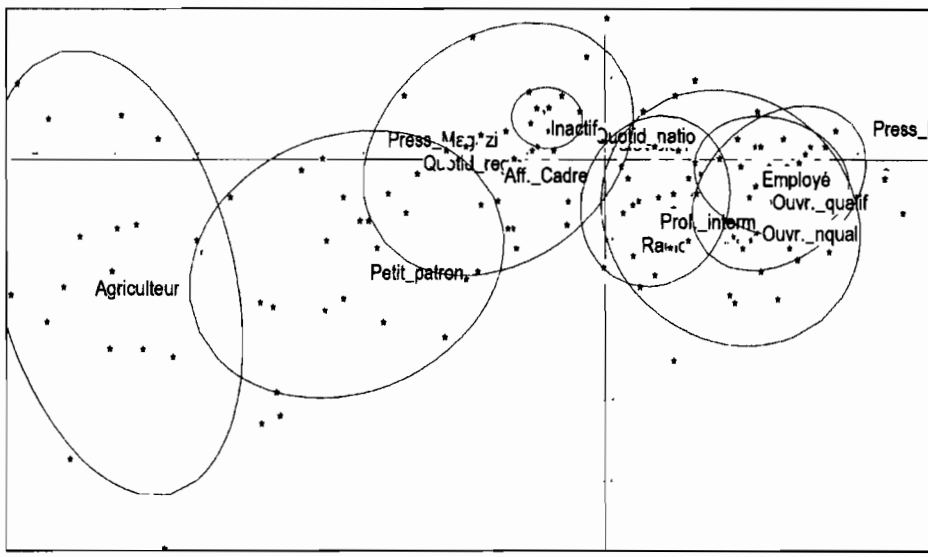


Figure 4.5 - 4
Zones de confiance "bootstrap partiel" pour les lignes actives
Plan factoriel (2,3) relatif au tableau 4.5 – 1

Nous ne publions pas les graphiques relatifs aux différents types de bootstrap total car ils sont très similaires à ceux du bootstrap partiel, confirmant ainsi la stabilité de la structure bi-dimensionnelle observée.

En conclusion, en même temps qu'un enrichissement de l'information sur le plan (F_1, F_2), on a un critère pour choisir le nombre d'axes de représentation, limité ici à 2.

Notons que le processus s'applique également aux variables supplémentaires. Ainsi, pour prendre un exemple, les lignes du tableau de contingence 4.4 - 1 (sexe, âge, niveau d'éducation) peuvent être répliquées en utilisant un schéma multinomial similaire à celui utilisé pour les variables actives, les projections de ces lignes simulées dans les plans factoriels définissant alors des zones de confiance.

4.6 Annexes techniques du chapitre 4

Cette annexe technique comprend trois volets, bien différents, dont la présence dans les sectons précédentes aurait pu encombrer ou ennuyer la lectrice ou le lecteur. La section 4.6.1 évoque deux simplifications dans la mise en oeuvre pratique des calculs : l'équivalence entre analyse par rapport à l'origine et l'analyse par rapport au centre de gravité (propriété typique de l'analyse des correspondances), et la possibilité de diagonaliser une matrice symétrique, opération plus avantageuse du point de vue numérique.

La section 4.6.2 donne des précisions sur l'approximation de la distribution des valeurs propres avec celles dérivées de la diagonalisation d'une matrice de Wishart.

La section 4.6.3 montre la propriété d'indépendance des pourcentages d'inertie des axes et de la trace (inertie totale), dans l'hypothèse d'indépendance des lignes et des colonnes de la table analysée.

4.6.1 Mise en œuvre pratique des calculs

La distance du χ^2 ne diffère en fait de la métrique euclidienne usuelle que par l'introduction d'une pondération. On peut se ramener à la métrique euclidienne usuelle par un changement de coordonnées initial. Les calculs en sont simplifiés et, notamment, la matrice à diagonaliser devient symétrique. Par ailleurs, l'analyse par rapport aux centres de gravité est équivalente à l'analyse par rapport à l'origine.

a – Analyse par rapport à l'origine ou au centre de gravité du nuage

Nous raisonnerons, pour fixer les idées, dans $\mathcal{R}^p$. Le centre de gravité G du nuage des profils-lignes a pour $j^{ième}$ composante :

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

L'analyse par rapport au centre de gravité revient à remplacer $\frac{f_{ij}}{f_{i.}}$ par $\frac{f_{ij}}{f_{i.}} - f_{.j}$

c'est-à-dire par $\frac{f_{ij} - f_i f_j}{f_{i.}}$.

Remarquons que le nuage est contenu dans un hyperplan $\mathcal{H}$ à p-1 dimensions défini pour tout i par la relation :

$$\sum_{j=1}^{p} \frac{f_{ij}}{f_i} = 1$$

Ce sous-espace contient le centre de gravité G et les axes factoriels de l'analyse par rapport à G. La somme des composantes de ces facteurs est nulle.

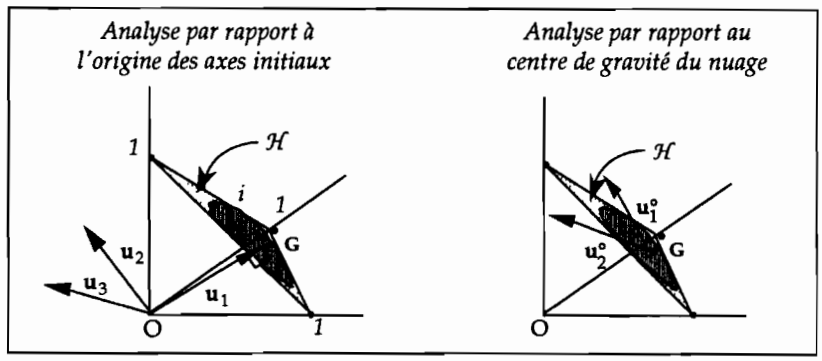


Figure 4.6 – 1. Analyse dans $\mathcal{R}^3$

Dans l'analyse par rapport à l'origine, la première direction $\mathbf{u}_1$ est l'axe joignant l'origine au centre de gravité du nuage orthogonalement à $\mathcal{H}$. L'inertie projetée sur cet axe vaut *1*, égale à la distance entre l'origine et le centre de gravité, puisque la projection des points du nuage sur cet axe est confondue avec le centre de gravité. Les p-1 axes suivants $(\mathbf{u}_2,...,\mathbf{u}_\alpha,...,\mathbf{u}_p)$ contenus dans $\mathcal{H}$ constituent une base définissant des directions de droites d'inertie maximum du nuage. Ils coïncident avec les p-1 premiers axes de l'analyse par rapport au centre de gravité $(\mathbf{u}_1^\circ,...,\mathbf{u}_\alpha^\circ,...,\mathbf{u}_{p-1}^\circ)$.

Le $p^{ième}$ axe correspond à $\mathbf{u}_1$ et n'indique aucune direction dans $\mathcal{H}$ puisqu'il n'est pas contenu dans $\mathcal{H}$. Son inertie (valeur propre) associée, est nulle.

$\mathbf{S}$ étant la matrice à diagonaliser du nuage non centré et $\mathbf{S}^\circ$ celle du nuage centré, on a les relations :

$$s^\circ_{jj'} = s_{jj'} - f_j$$

et pour $1 < \alpha < p - 1$:

$$\mathbf{u}_\alpha^\circ = \mathbf{u}_{\alpha+1} \qquad \text{et} \qquad \lambda_\alpha^\circ = \lambda_{\alpha+1}$$

$$\mathbf{u}_p^\circ = \mathbf{u}_1 \qquad \text{et} \qquad \lambda_p^\circ = 0 \quad \text{et} \quad \lambda_1 = 1$$

Ainsi dans $\mathcal{R}^p$ (et il en est de même dans $\mathcal{R}^n$), il est équivalent de réaliser l'analyse des correspondances sur le tableau de données centrées de terme général :

$$\frac{f_{ij}}{f_{i.}} - f_j$$

ou sur le tableau de données non centrées de terme général :

$$\frac{f_{ij}}{f_{i.}}$$

On peut donc diagonaliser la matrice **S** de l'analyse par rapport à l'origine[1], en prenant soin d'éliminer le premier vecteur propre reliant l'origine au centre de gravité du nuage et la valeur propre associée égale à *1*.

b – Symétrisation de la matrice à diagonaliser

La matrice à diagonaliser $\mathbf{S} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}$, dans $\mathcal{R}^p$, n'est pas en général symétrique. Son terme général s'écrit :

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}$$

Considérons la matrice $\hat{\mathbf{A}} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}$ symétrique et la matrice $\mathbf{D}_p^{-1}$ diagonale. On exprime alors **S** de la manière suivante :

$$\mathbf{S} = \hat{\mathbf{A}}\mathbf{D}_p^{-1/2}\mathbf{D}_p^{-1/2}$$

Partant de la relation $\mathbf{Su} = \lambda\mathbf{u}$, il vient :

$$\hat{\mathbf{A}}\mathbf{D}_p^{-1/2}\mathbf{D}_p^{-1/2}\mathbf{u} = \lambda\mathbf{u}$$

Prémultiplions les deux membres par $\mathbf{D}_p^{-1/2}$ et en posant $\mathbf{D}_p^{-1/2}\mathbf{u} = \mathbf{w}$, on obtient :

$$\mathbf{D}_p^{-1/2}\hat{\mathbf{A}}\mathbf{D}_p^{-1/2}\mathbf{w} = \lambda\mathbf{w}$$

La matrice **A** est symétrique :

$$\mathbf{A} = \mathbf{D}_p^{-1/2}\hat{\mathbf{A}}\mathbf{D}_p^{-1/2} = \mathbf{D}_p^{-1/2}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1/2}$$

et :

$$\mathbf{Aw} = \lambda\mathbf{w}$$

Les matrices **S** et **A** ont mêmes valeurs propres λ. Leurs vecteurs propres sont liés par la relation :

$$\mathbf{u} = \mathbf{D}_p^{-1/2}\mathbf{w}$$

---

[1] Compte tenu du critère d'ajustement, on considère l'inertie totale du nuage centré, égale à la trace $tr(\mathbf{S}^\circ)$ de $\mathbf{S}^\circ$ et l'on a : $tr(\mathbf{S}^\circ) = tr(\mathbf{S}) - 1$.

Il est plus facile de diagonaliser la matrice **A** de terme général :

$$a_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j} f_{.j'}}}$$

Remarque :

C'est la matrice à diagonaliser si l'on choisit de prendre comme coordonnées initiales du point i, les p quantités :

$$x_{ij} = \frac{f_{ij}}{f_{i.} \sqrt{f_{.j}}} \qquad (j = 1, \ldots, p)$$

Dans ce cas, la distance du χ^2 entre deux points i et i' devient, avec les nouvelles coordonnées, la distance euclidienne usuelle :

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.} \sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.} \sqrt{f_{.j}}} \right)^2$$

Cette transformation du tableau des fréquences relatives conduit à la diagonalisation d'une matrice symétrique.

Notons que les coordonnées du centre de gravité G sont alors :

$$G_j = \sqrt{f_{.j}}$$

et les coordonnées du point i après recentrage :

$$\frac{f_{ij}}{f_{i.} \sqrt{f_{.j}}} - \sqrt{f_{.j}} = \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} \sqrt{f_{.j}}}$$

4.6.2 Précisions sur l'approximation de la distribution des valeurs propres

On utilise les notations du paragraphe 4.4.2 (a).

Désignons par **h** le vecteur à np composantes tel que :

$$h_{ij} = \frac{\sqrt{k}(f_{ij} - f_{i.} f_{.j})}{\sqrt{f_{i.} f_{.j}}}$$

Ce vecteur de $\mathcal{R}^{np}$ a, sous les conditions précédentes, une distribution normale avec $E(h_{ij}) = 0$ pour tout i et j.

Sa matrice des covariances a pour terme général:

$$V_h(i\,j, i'j') = \delta_{(i\,j, i'j')} - \sqrt{f_{i.} f_{j} f_{i'.} f_{j'}}$$

où

$$\delta_{(ij,i'j')} = 1 \text{ si } i = i' \text{ et } j = j'$$
$$\delta_{(ij,i'j')} = 0 \text{ sinon}$$

Construisons une matrice orthogonale **A**, d'ordre (p,p), telle que sa première colonne ait pour $j^{ème}$ élément $\sqrt{f_{.j}}$ (pour tout $j \leq p$), les $p-1$ autres colonnes formant avec la première une base orthonormée de $\mathcal{R}^p$.

De la même façon, construisons une matrice orthogonale **B** d'ordre (n,n) telle que sa première ligne ait pour $i^{ème}$ élément $\sqrt{f_{i.}}$ (pour tout $i \leq n$), les $n-1$ autres lignes formant avec la première une base orthonormée de $\mathcal{R}^n$.

La matrice $\mathbf{B} \otimes \mathbf{A}'$ d'ordre (np,np), produit *direct* ou de *Kronecker* des matrices **B** et **A'**, est aussi orthogonale.

Pour tous $1 < i < n$, $1 < j < p$, $1 < r < n$ et $1 < s < p$, on a les relations :

$$\sum_j \sqrt{f_j} h_{ij} = 0 \; ; \sum_i \sqrt{f_i} h_{ij} = 0 \; ; \sum_m b_{rm} \sqrt{f_{m.}} = 0 \; ; \sum_k a_{ks} \sqrt{f_k} = 0 \; ;$$

De ces relations, on déduit que le vecteur **y** de $\mathcal{R}^{np}$ tel que:

$$\mathbf{y} = \mathbf{B} \otimes \mathbf{A}'\mathbf{h}$$

a seulement $(n-1)(p-1)$ composantes non nulles. On a:

$$y_{rs} = 0 \quad \text{si} \quad r = 1 \quad \text{ou si} \quad s = 1$$

La matrice des covariances de **y** est :

$$\mathbf{V}_y = (\mathbf{B} \otimes \mathbf{A}')\mathbf{V}_h(\mathbf{B}' \otimes \mathbf{A})$$

Pour tout couple de composantes non nulles, on a:

$$V_y(r\,s, r's') = \delta_{rr'}\delta_{ss'}$$

Soit **Y** la matrice d'ordre (n,p) définie par :

$$\mathbf{Y} = \mathbf{BHA}$$

où **H** est la matrice d'ordre (n,p) de terme général h_{ij}. La première ligne et la première colonne de **Y** sont nulles.

Les éléments de la sous-matrice $\hat{\mathbf{Y}}$ d'ordre $(n-1)(p-1)$, formée des éléments non nuls de **Y**, sont donc distribués indépendamment suivant la loi normale centrée réduite.

La matrice :

$$\mathbf{S} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$$

est donc distribuée suivant une loi de Wishart $W(p\text{-}1, n\text{-}1, \mathbf{I})$ de paramètres $(n-1)$ et $(p-1)$.

Or **S** a les mêmes valeurs propres non nulles que **Y'Y** c'est-à-dire que **A'H'HA** ; ce sont finalement les mêmes valeurs propres que **H'H**, puisque **A** est orthogonale.

Remarquons que ceci implique que $tr(\mathbf{H'H})$ est un χ^2 à $(n-1)(p-1)$ degrés de liberté. Or :

$$tr(\mathbf{H'H}) = k\sum_i\sum_j \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{j}}$$

Il s'agit du test usuel du χ^2 sur les tableaux de contingence.

La matrice symétrisée $\mathbf{S^*}$ que l'on diagonalise lors de l'analyse des correspondances du tableau **K**, est la matrice :

$$\mathbf{S^*} = \frac{1}{k}\mathbf{H'H}$$

Ainsi, si λ_α est la $\alpha^{ème}$ valeur propre issue de l'analyse des correspondances d'un tableau **K** d'ordre (n, p), de somme totale k, alors la distribution de $k\lambda_\alpha$ est approximativement celle de la $\alpha^{ème}$ valeur propre d'une matrice de Wishart définie par les paramètres $W(p-1,n-1,\mathbf{I})$.

4.6.3 Indépendance des taux d'inertie et de la trace

On a vu (cf annexe 2.6) que la densité $g(\Lambda)$ de la loi jointe des valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ d'une matrice de Wishart a la forme :

$$g(\Lambda) = D(n,p)\prod \lambda_k^{\frac{n-p-1}{2}} exp\left\{-\tfrac{1}{2}\sum\lambda_k\right\}\prod_{k<j}(\lambda_k - \lambda_j)$$

Si l'on pose :

$$\begin{cases} \lambda_k = z\,\tau_k \text{ , pour } k<p \\ \lambda_p = z\left(1-\tau_1-\tau_2-\ldots-\tau_{p-1}\right) \end{cases}$$

alors z est la trace de la matrice de Wishart : $z = \sum_k \lambda_k$

On trouve aisément une factorisation de la densité (le jacobien de cette transformation vaut z^{p-1}) :

$$g(\Lambda) = g_1(z)\, g_2(\tau_1,\ldots,\tau_{p-1})$$

la fonction $g_1(z)$ s'écrivant :

$$g_1(z) = \frac{1}{2\Gamma(\frac{np}{2})}(\frac{z}{2})^{\frac{np}{2}-1} exp(-\frac{z}{2})$$

où l'on reconnaît la densité de la loi du χ^2 à np degrés de liberté.

La factorisation des densités (et l'indépendance des domaines d'intégration) montrent que les pourcentages de variance $\tau_1, \tau_2, \ldots, \tau_{p-1}$ sont indépendants de la trace z .

Cette propriété a en fait été établie (dans le cadre de l'analyse en composantes principales non normée) par Bartlett (1951). Elle suppose évidemment vérifiée l'hypothèse d'indépendance et l'hypothèse paramétrique de multinormalité, ce qui limite sa portée pratique.

La propriété semble encore valable dans le cas de l'analyse des correspondances, pour laquelle la loi de Wishart est seulement une loi approchée (les simulations extensives entreprises pour construire les abaques ont permis de vérifier cette indépendance, que nous avions d'ailleurs conjecturée à partir de résultats empiriques)[1].

[1] Rappelons que la normalité est obtenue par convergence de la loi multinomiale vers la loi normale dans le cas de l'analyse des correspondances des tables de contingence. La propriété d'indépendance des taux d'inertie et de la trace a donc dans ce cas une portée plus générale. Elle requiert cependant l'hypothèse d'indépendance des lignes et des colonnes. La figure 4.3.1 de la section 4.3 a précisément montré les cas de trace (inertie totale) significative (cas 3 et 4) et les cas de taux d'inertie significatifs (cas 2 et 4). Le cas 1 étant le seul cas où aucune des deux quantités n'est significative. Le cas 2 est le cas de dépendance indécelable par le test classique du $\chi 2$ qui ne porte que sur la trace.

Chapitre 5

Analyse des correspondances multiples

L'analyse des correspondances peut se généraliser de plusieurs façons au cas où plus de deux ensembles sont mis en correspondance. Une des généralisations la plus utilisée est l'*analyse des correspondances multiples* qui permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié : les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs dizaine de milliers) ; les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions.

Il s'agit en fait d'une simple extension du domaine d'application de l'analyse des correspondances appliquée non plus à une table de contingence, mais à un *tableau disjonctif complet*. Les propriétés d'un tel tableau sont intéressantes, les procédures de calculs et les règles d'interprétation des représentations obtenues sont simples et spécifiques.

On peut faire remonter les principes de cette méthode à Guttman (1941), mais aussi à Burt (1950) ou à Hayashi (1956). D'autres types d'extension de l'analyse des correspondances ont été proposés par Benzécri (1973), Escofier-Cordier (1965), et par Masson (1974) qui s'appuie notamment sur les travaux de Carroll (1968), Horst (1961) et Kettenring (1971)[1].

---

[1] L'analyse des correspondances multiples a été développée également sur le nom d'*Homogeneity Analysis* par l'équipe de J. de Leeuw depuis 1973 (cf. Gifi, 1990) et sous le nom de *Dual Scaling* par Nishisato (1980). Une application de l'analyse des correspondances à un tableau disjonctif complet se trouve dans Nakache (1973). L'ensemble des résultats et propriétés de l'analyse des correspondances multiples présentés dans cette section figurent, avec des programmes et des exemples, dans Lebart et Tabard (1973). Le nom d'*analyse des correspondances multiples* figure pour la première fois dans Lebart (1975 a). Un exposé synthétique de ces diverses approches a été réalisée par Tenenhaus et Young (1985).

L'extension du domaine d'application de l'analyse des correspondances se fonde sur l'équivalence suivante : si pour n individus, on dispose des valeurs (réponses) prises par deux variables nominales ayant respectivement p_1 et p_2 modalités, il est alors équivalent de soumettre à l'analyse des correspondances le tableau de contingence (p_1, p_2) croisant les deux variables ou d'analyser le tableau binaire à n lignes et $(p_1+ p_2)$ colonnes décrivant les réponses. L'analyse de ce dernier tableau est plus coûteuse, mais plus intéressante, car elle se généralise immédiatement au cas de plus deux variables nominales.

5.1 Notations et définitions

Une partie généralement importante des fichiers d'enquête se compose de réponses à des questions mises sous forme *disjonctive complète*, pour lesquelles les diverses modalités de réponses s'excluent mutuellement et une modalité est obligatoirement choisie. Nous commencerons par étudier ce type de tableau.

5.1.1 Tableau disjonctif complet

Prenons l'exemple de la question :

Etes-vous ?

1- célibataire,	*2- marié(e) ou vivant maritalement,*	
3- veuf(ve),	*4- divorcé(e),*	*5- non réponse,*

cinq modalités de réponses (dont une non-réponse) sont possibles.

Une variable continue peut être transformée en variable nominale par le découpage en classes des valeurs de la variable. A la question "âge de l'enquêté", on prévoit *8* modalités de réponse :

1- moins de 25 ans;	*2- de 25 à 29 ans;*	*3- de 30 à 34 ans;*
4- de 35 à 39 ans;	*5- de 40 à 44 ans;*	*6- de 45 à 49 ans;*
7- de 50 ans et plus;	*8- non-réponse.*	

Si l'on désigne par s le nombre des questions posées à n individus, on dispose ainsi d'un tableau de données **R** ayant n lignes et s colonnes mis sous forme de codage condensé, illustré sur la figure 5.1 - 1 par un tableau pour lequel $s = 3$ et $n = 12$.

Le terme général r_{iq} désigne la modalité de la question q choisie par le sujet i. En notant p_q le nombre des modalités de réponses à une question q, on a : $r_{iq} \leq p_q$.

Mais un tel tableau n'est pas exploitable : les sommes en ligne et en colonne n'ont pas de sens. Il faut recoder les variables.

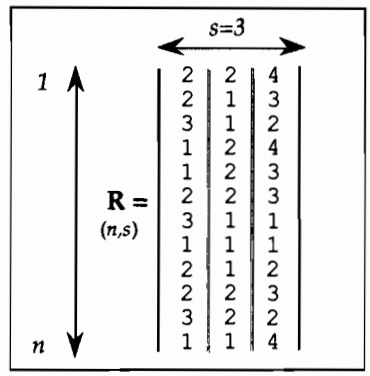


Figure 5.1 – 1. Tableau de données sous forme de codage condensé

a _ Hypercube de contingence

Pour disposer de toute l'information, on peut construire l'hypercube de contingence H croisant les s questions et dont les éléments constituent l'éventail des réponses possibles des sujets enquêtés. On dispose d'un ensemble-produit des modalités des s questions dont les éléments sont constitués des suites de s modalités, chacune étant prise dans une question différente.

Pour $s = 3$ questions ayant respectivement 3, 2 et 4 modalités, il existe 24 combinaisons possibles de réponses selon lesquelles sont réparties les individus. Dans le cas de deux questions, l'hypertable est le tableau de contingence. Pour un nombre important de questions, l'hypertable sera en général presque vide. Si l'on pose à 1000 individus 12 questions ayant chacune 10 modalités de réponse, le nombre de réponses possibles distinctes vaut 10^{12}. Au plus une case sur un milliard de l'hypertable ne sera pas vide.

b _ Le codage disjonctif

On désigne par I l'ensemble des n sujets ayant répondu au questionnaire et par p le nombre total des modalités des s questions. On a :

$$p = \sum_{q=1}^{s} p_q$$

On construit, à partir du tableau de données **R**, le tableau **Z** à n lignes et p colonnes décrivant les s réponses des n individus par un codage binaire. Le tableau **Z** est la juxtaposition de s sous-tableaux :

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_q, \ldots, \mathbf{Z}_s]$$

Le sous-tableau $\mathbf{Z}_q$, à n lignes et p_q colonnes, est tel que sa $i^{ème}$ ligne contient $p_q - 1$ fois la valeur 0 et une fois la valeur 1 dans la colonne correspondant à la modalité de la question q choisie par le sujet i. Autrement dit le tableau $\mathbf{Z}_q$ décrit la partition des n individus induite par les réponses à la question q.

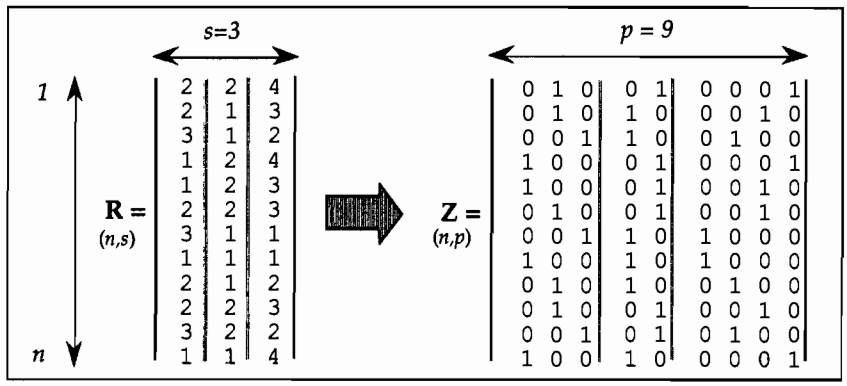


Figure 5.1 - .2. Construction du tableau disjonctif complet Z

Le tableau **Z** est appelé *tableau disjonctif complet* dont le terme général s'écrit :

$$z_{ij} = 1 \quad \text{ou} \quad z_{ij} = 0$$

selon que le sujet i a choisi la modalité j de la question q ou non.

Les marges en ligne du tableau disjonctif complet sont constantes et égales au nombre s de questions : $z_{i.} = \sum_{j=1}^{p} z_{ij} = s$

Les marges en colonne correspondent au nombre de sujets ayant choisi la modalité j de la question q :

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

On vérifie que, pour chaque sous-tableau $\mathbf{Z}_q$, l'effectif total est bien :

$$z_q = \sum_{j \in q} z_{.j} = n$$

La somme des marges donne l'effectif total z du tableau **Z** soit :

$$z = \sum_{i=1}^{n} \sum_{j=1}^{p} z_{ij} = ns$$

5.1.2 Tableau de contingence de Burt

L'ensemble des p_q modalités de réponse à une question permet de partitionner l'échantillon en au plus p_q classes. La donnée de deux questions mises sous forme disjonctive complète permet de réaliser deux partitions de l'ensemble des individus enquêtés et l'on obtient un tableau de contingence.

L'analyse du tableau croisant les deux partitions peut être généralisée au cas de s partitions, s étant un entier supérieur à 2.

On construit, à partir du tableau disjonctif complet **Z**, le tableau symétrique **B** d'ordre (p,p) qui rassemble les croisements deux à deux de toutes les variables :

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

B est appelé *tableau de contingence de Burt*[1] associé au tableau disjonctif complet **Z**.

Le terme général de **B** s'écrit : $b_{jj'} = \sum_{i=1}^{n} z_{ij} z_{ij'}$

B est une juxtaposition de tableaux de contingence.

Les marges sont pour tout $j \leq p$:

$$b_j = \sum_{j}^{p} b_{jj'} = s z_{.j}$$

et l'effectif total b vaut :

$$b = s^2 n$$

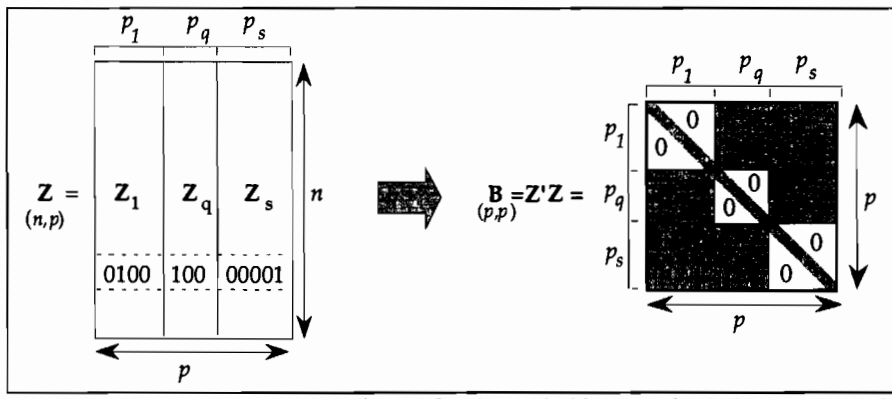


Figure 5.1 – 3. Construction du tableau des faces de l'hypercube (tableau de Burt) B à partir du tableau disjonctif complet Z

Le tableau **B** est formé de s^2 blocs où l'on distingue :

- le bloc $\mathbf{Z}'_q\mathbf{Z}_{q'}$ indicé par (q,q'), d'ordre $(p_q, p_{q'})$ qui n'est autre que la table de contingence croisant les réponses aux questions q et q'.
- le $q^{ième}$ bloc carré $\mathbf{Z}'_q\mathbf{Z}_q$ obtenu par le croisement d'une variable avec elle-même. C'est une matrice d'ordre (p_q, p_q), diagonale puisque deux modalités

[1] Sir Cyril Burt a été un incontestable innovateur au point de vue méthodologique (cf. son article précité de 1950, dans lequel il préconise le calcul de **B**, et sa diagonalisation après une normalisation qui correspond à celle de l'analyse des correspondances multiples). Sa réputation a cependant été ternie par des accusations de falsifications d'observations et de fraude scientifique.

d'une même question ne peuvent être choisies simultanément. Les termes diagonaux sont les effectifs des modalités de la question q.

Nous désignerons par **D** la matrice diagonale, d'ordre (p,p) ayant les mêmes éléments diagonaux que **B** ; ces éléments sont les effectifs correspondant à chacune des modalités (cf. figure 5.1 - 4) :

$$d_{jj} = b_{jj} = z_{.j}$$

$$d_{jj'} = 0 \quad \text{pour tout} \quad j' \neq j$$

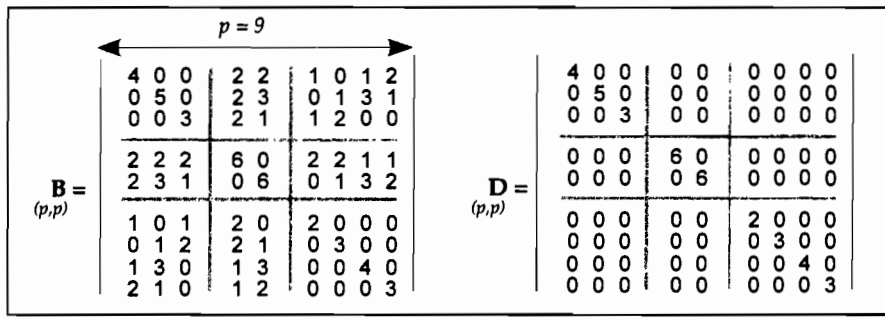


Figure 5.1 - 4. Tableau de Burt B et matrice diagonale D associée
(données des figures 5.1 - 1 et 5.1 - 2)

La matrice **D** peut être également considérée comme formée de s^2 blocs.

Seules les s matrices diagonales $\mathbf{D}_q = \mathbf{Z'}_q\mathbf{Z}_q$ $(q = 1,..s)$ constituant les blocs diagonaux de **B** sont des matrices non nulles.

5.2 Principes de base de l'analyse des correspondances multiples

L'analyse des correspondances multiples est l'analyse des correspondances d'un tableau disjonctif complet. Ses principes sont donc ceux de l'analyse des correspondances à savoir :

- mêmes transformations du tableau de données en profils-lignes et en profils-colonnes;
- même critère d'ajustement avec pondération des points par leurs profils marginaux;
- même distance, celle du χ^2.

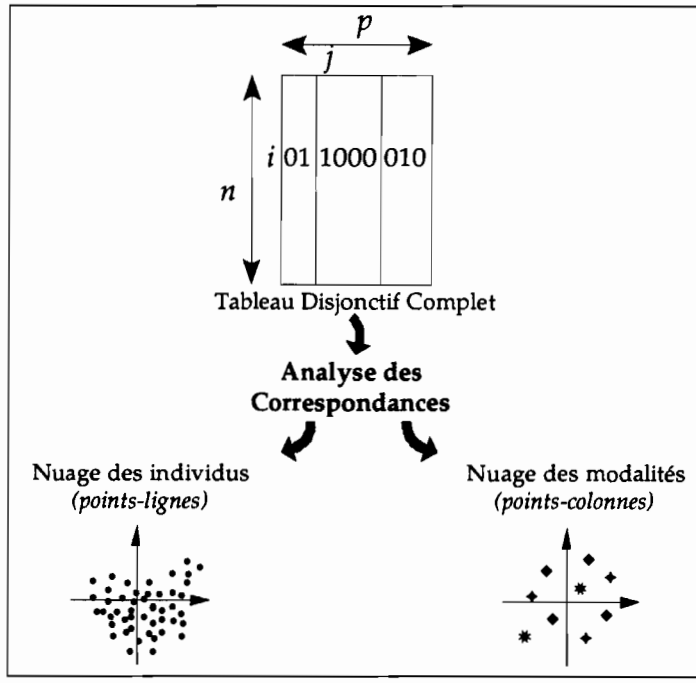


Figure 5.2 – 1. Analyse des correspondances multiples

5.2.1 Schéma général

a – Critère d'ajustement et distance du χ^2

Les individus sont tous affectés d'une masse identique égale à $m_i = \frac{1}{n}$ et chacune des modalités j est pondérée par sa fréquence $m_j = \frac{z_j}{ns}$.

La distance du χ^2 appliquée à un tableau disjonctif complet conserve un sens. En effet, dans $\mathcal{R}^n$, la distance entre modalités s'écrit :

$$d^2(j,j') = \sum_{i=1}^{n} n\left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}}\right)^2$$

Ainsi deux modalités choisies par les mêmes individus coïncident. Par ailleurs, les modalités de faible effectif sont éloignées des autres modalités.

Dans $\mathcal{R}^p$, la distance entre deux individus i et i' s'exprime par :

$$d^2(i,i') = \frac{1}{s}\sum_{j=1}^{p} \frac{n}{z_j}(z_{ij} - z_{i'j})^2$$

Deux individus sont proches s'ils ont choisi les mêmes modalités. Ils sont éloignés s'ils n'ont pas répondu de la même manière[1].

b – Axes factoriels et facteurs

En reprenant les résultats de l'analyse des correspondances et les notations adoptées (cf. § 4.2.1.b), on pose[2] :

$$\mathbf{F} = \frac{1}{ns}\mathbf{Z} \qquad \text{de terme général} \qquad f_{ij} = \frac{z_{ij}}{ns}$$

$$\mathbf{D}_p = \frac{1}{ns}\mathbf{D} \qquad \text{de terme général} \qquad f_{.j} = \delta_{ij}\frac{z_{.j}}{ns}$$

$$\mathbf{D}_n = \frac{1}{n}\mathbf{I}_n \qquad \text{de terme général} \qquad f_{i.} = \frac{\delta_{ij}}{n}$$

Pour trouver les axes factoriels $\mathbf{u}_\alpha$ on diagonalise la matrice :

$$\mathbf{S} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{Z}'\mathbf{Z}\,\mathbf{D}^{-1}$$

de terme général :

$$s_{jj'} = \frac{1}{s\,z_{.j'}}\sum_{i=1}^{n} z_{ij}z_{ij'}$$

(attention, s [sans indice] désigne le nombre de questions dans ce chapitre)

Dans $\mathcal{R}^p$, l'équation du $\alpha^{\text{ième}}$ axe factoriel $\mathbf{u}_\alpha$ est :

$$\frac{1}{s}\mathbf{Z}'\mathbf{Z}\,\mathbf{D}^{-1}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \qquad [5.2 - 1]$$

Nous désignons ici par facteur $\hat{\varphi}_\alpha$ l'opérateur projection sur l'axe factoriel $\mathbf{u}_\alpha$. L'équation du $\alpha^{\text{ième}}$ facteur $\hat{\varphi}_\alpha = \mathbf{D}^{-1}\mathbf{u}_\alpha$ s'écrit :

$$\frac{1}{s}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\,\hat{\varphi}_\alpha = \lambda_\alpha \hat{\varphi}_\alpha \qquad [5.2 - 2]$$

De même, l'équation du $\alpha^{\text{ième}}$ facteur $\hat{\psi}_\alpha$ dans $\mathcal{R}^n$ s'écrit :

$$\frac{1}{s}\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\,\hat{\psi}_\alpha = \lambda_\alpha \hat{\psi}_\alpha$$

Les facteurs φ_α et ψ_α (de norme λ_α) représentent les coordonnées des points-lignes et des points-colonnes sur l'axe factoriel α.

---

[1] On note qu'une modalité j intervient d'autant plus dans le calcul de la distance entre deux individus que sa masse est plus faible.

[2] $\mathbf{I}_n$ est la matrice identité d'ordre (n,n) et δ_{ij} est tel que :

$$\delta_{ij} = 1 \quad \text{si} \quad i = j \quad \text{et} \quad \delta_{ij} = 0 \quad \text{si} \quad i \neq j$$

Les relations de transition entre les facteurs φ_α et ψ_α sont :

$$\begin{cases} \varphi_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \mathbf{D}^{-1}\mathbf{Z}'\psi_\alpha \\ \psi_\alpha = \dfrac{1}{s\sqrt{\lambda_\alpha}} \mathbf{Z}\,\varphi_\alpha \end{cases}$$

c – Facteurs et relations quasi-barycentriques

La coordonnée factorielle de l'individu *i* sur l'axe α est donnée par :

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{z_{ij}}{z_{i.}} \varphi_{\alpha j}$$

c'est-à-dire :

$$\psi_{\alpha i} = \frac{1}{s\sqrt{\lambda_\alpha}} \sum_{j \in p(i)} \varphi_{\alpha j} \qquad [5.2 - 3]$$

où *p(i)* désigne l'ensemble des modalités choisies par l'individu *i*.

Au coefficient $\frac{1}{\sqrt{\lambda_\alpha}}$ près, l'individu *i* se trouve au point moyen du nuage des modalités qu'il a choisies (cf. figure 5.2-2).

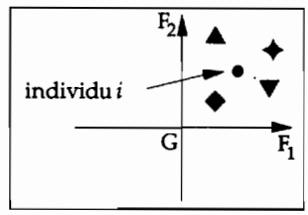


Figure 5.2 - 2. Projection d'un individu au point moyen des modalités choisies

De même, la coordonnée de la modalité *j* sur l'axe α est donnée par :

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{z_{ij}}{z_{.j}} \psi_{\alpha i}$$

c'est-à-dire :

$$\varphi_{\alpha j} = \frac{1}{z_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in I(j)} \psi_{\alpha i} \qquad [5.2 - 4]$$

où *I(j)* désigne l'ensemble des individus ayant choisi la modalité *j*.

Avant la dilatation sur l'axe α, la modalité *j* se trouve au point moyen du nuage des individus qui l'ont choisie comme réponse (cf. figure 5.2-3).

Dans le calcul des relations quasi-barycentriques, les individus ne sont pas pondérés. Il s'agit de simples calculs de moyennes arithmétiques de coordonnées.

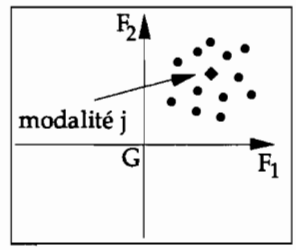


Figure 5.2 – 3. Projection d'une modalité au point moyen des individus concernés

d – Sous-nuage des modalités d'une même variable

Le nuage des modalités dans $\mathcal{R}^n$ peut être décomposé en s sous-nuages, le $q^{ème}$ correspondant à l'ensemble des p_q modalités de la variable q.

Ces sous-nuages ont même centre de gravité G qui est celui du nuage global. En effet, les coordonnées des points du sous-nuage relatif à la variable q sont les colonnes de $\mathbf{Z}_q\mathbf{D}_q^{-1}$ et les éléments diagonaux de $\frac{1}{n}\mathbf{D}_q$ sont les masses relatives des p_q points de ce sous-nuage. Puisque :

$$\sum_{j \in p_q} z_{ij} = 1$$

alors la $i^{ème}$ composante du centre de gravité du sous-nuage vaut :

$$G_{qi} = \sum_{j \in p_q} \frac{d_{jj}}{n} \frac{z_{ij}}{d_{jj}} = \frac{1}{n} = G_i$$

où il apparaît que G_{qi} ne dépend pas de q.

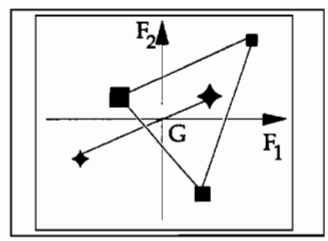


Figure 5.2 – 4. Composantes centrées

Les composantes φ_q des modalités d'une variable q (relatives aux facteurs non-triviaux φ) sont centrées puisque ces facteurs correspondent à une analyse du nuage après translation de l'origine en G. Les facteurs opposent les modalités d'une même variable (cf. figure 5.2-4).

Remarques

1) Si le tableau disjonctif n'est pas complet (c'est-à-dire si, pour au moins un individu, aucune modalité de réponse à une question n'a été choisie), les modalités d'une même variable ne sont plus centrées sur le centre de gravité du nuage global.

2) Le codage disjonctif complet permet de transformer une variable continue en une variable nominale dont les modalités sont des classes ordonnées. Il est alors utile de tracer la trajectoire qui relie les classes, trajectoire qui peut suggérer des liaisons non linéaires entre cette variable et les axes.

e – Support du nuage des modalités

Les coordonnées des modalités dans $\mathcal{R}^n$ sont les colonnes de $\mathbf{ZD}^{-1}$. Elles engendrent un sous-espace dont la dimension est le rang de $\mathbf{ZD}^{-1}$, donc le rang de $\mathbf{Z} = [\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_q,...,\mathbf{Z}_s]$.

Tous les sous-espaces engendrés par les $\mathbf{Z}_q$ ont en commun la première bissectrice notée Δ. Le rang maximum de $\mathbf{Z}$ est donc :

$$p_1 + (p_2 - 1) + ... + (p_s - 1) = p - s + 1$$

Le rang maximum de la matrice à diagonaliser $\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}$ sera donc $p - s + 1$. Mais dans l'analyse du nuage par rapport à l'origine O, la première bissectrice est vecteur propre correspondant à la valeur propre *1* (le nuage est contenu dans le sous-espace $\mathbf{D}^{-1}$-orthogonal à Δ).

Dans l'analyse par rapport au centre de gravité G, on trouvera donc $p - s$ valeurs propres non nulles. En choisissant une base dans le support du nuage, on pourra se ramener à la recherche d'éléments propres d'une matrice d'ordre $p - s$.

f – Meilleure représentation simultanée

La présentation de l'analyse des correspondances peut être formulée ici de façon particulière en raison du codage spécifique du tableau disjonctif complet.

Nous cherchons sur un même axe les coordonnées des n individus et des p modalités de façon que :

- la coordonnée d'un individu i soit la moyenne arithmétique des coordonnées des modalités qu'il a choisies (à une dilatation β près, que l'on s'efforcera de rendre minimale).
- la coordonnée d'une modalité j soit la moyenne arithmétique des coordonnées des individus qui l'ont choisie (à une même dilatation β près).

Bien entendu, on obtient les relations dite quasi-barycentriques issues de l'analyse du tableau disjonctif complet $\mathbf{Z}$ avec, pour le coefficient de dilatation β, la valeur minimale $\beta = \frac{1}{\sqrt{\lambda}}$:

$$\begin{cases} \varphi = \frac{1}{\sqrt{\lambda}} \mathbf{D}^{-1}\mathbf{Z}'\psi \\ \psi = \frac{1}{s\sqrt{\lambda}} \mathbf{Z}\varphi \end{cases}$$

La représentation simultanée des individus et des modalités est importante pour l'interprétation des résultats.

Cependant elle n'est pratiquement pas utilisée, d'une part pour des raisons d'encombrement graphique (on dispose souvent de plusieurs centaines voire de plusieurs milliers d'individus) et d'autre part parce que les individus sont, dans la plupart des applications, anonymes. Ils ne présentent de l'intérêt que par l'intermédiaire de leurs caractéristiques. On peut cependant vouloir projeter les individus sur un plan factoriel afin d'apprécier leur répartition et les zones de densité.

5.2.2 Autres propriétés

L'analyse des correspondances multiples présente cependant des propriétés particulières dues à la nature même du tableau disjonctif complet.

a – Inertie du nuage des modalités et conséquences pratiques

On rappelle que la distance du χ^2 dans $\mathcal{R}^n$ est la métrique $\mathbf{D}_n^{-1}$.

La distance entre la modalité j et le centre de gravité du nuage G, dont toutes les n coordonnées valent $\frac{1}{n}$, s'écrit :

$$d^2(j,G) = n\sum_{i=1}^{n}\left(\frac{z_{ij}}{z_{.j}} - \frac{1}{n}\right)^2 = \frac{n}{z_{.j}} - 1$$

La distance d'une modalité au centre de gravité est d'autant plus grande que l'effectif est plus faible.

- Inertie d'une modalité

L'inertie $I(j)$ de la modalité j vaut :

$$I(j) = m_j d^2(j,G)$$

avec :
$$m_j = \frac{z_j}{ns}$$

d'où :
$$I(j) = \frac{1}{s}\left(1 - \frac{z_j}{n}\right)$$

La part d'inertie due à une modalité de réponse est d'autant plus grande que l'effectif dans cette modalité est plus faible.

Le maximum $\frac{1}{s}$ serait atteint par une modalité d'effectif nul. En conséquence, on évite, au moment du codage, les modalités à faibles effectifs susceptibles de perturber les directions des premiers axes factoriels.

- Inertie d'une question

L'inertie de la question q, notée $I(q)$, vaut :

$$I(q) = \sum_{j=1}^{p_q} I(j) = \frac{1}{s}(p_q - 1)$$

Ainsi la part d'inertie due à une question est fonction croissante du nombre de modalités de réponse.

La part minimale $\frac{1}{s}$ correspond aux questions à deux modalités. D'où l'intérêt d'équilibrer le système des questions, c'est-à-dire le découpage des variables en modalités, si on veut faire jouer le même rôle à toutes les questions.

- Inertie totale

On en déduit que l'inertie totale I vaut :

$$I = \sum_q I(q) = \sum_{j=1}^{p} \frac{z_j}{ns} d^2(j,G)$$

d'où :

$$I = \frac{p}{s} - 1$$

En particulier, elle vaut *1* dans le cas où toutes les questions ont deux modalités de réponse (cas où $p=2s$). On verra au paragraphe 5.3.1 – a que dans ce cas, analyse des correspondances multiples et analyse en composantes principales donnent des résultats équivalents.

L'inertie totale dépend uniquement du nombre de variables et de modalités et non des liaisons entre les variables. C'est une quantité qui, dans le cadre de l'analyse des correspondances multiples (comme dans celui de l'analyse en composantes principales normée), n'a pas de signification statistique.

- Une fragilité de l'analyse des correspondances multiples

Les formules précédentes, notamment celle donnant l'inertie d'une modalité, mais aussi les formules de distances du Chi-2 (χ^2) sur données individuelles binaires nous montrent que les modalités d'effectifs très faibles donneront lieu à des distances et à des inerties parfois démesurément grandes, facteur d'instabilité des résultats.

C'est pourquoi des procédures de « robustification » sont proposées dans la plupart des logiciels.

La formule souvent retenue[1], consiste à éliminer à titre provisoire les modalités à très faibles effectifs des modalités actives, les répondants correspondants étant ventilés aléatoirement dans les autres modalités (au prorata de leurs effectifs), de façon à sauvegarder la structure disjonctive. Toutes les données originales y compris les modalités d'effectifs très faibles sont ensuite projetées comme des modalités supplémentaires. Les très faibles effectifs sont définis par l'utilisateur par un paramètre décrivant en pourcentage un seuil relatif pour l'effectif. Dans la plupart des applications, un seuil relatif de 1% ou 2% convient.

Il est également possible d'analyser par analyse des correspondances simple le tableau privé de ses modalités d'effectifs faibles, qui n'est plus alors un tableau disjonctif complet. Certaines propriétés de l'analyse des correspondances multiples sont alors perdues.

b – Règles d'interprétation

Dire qu'il existe des affinités entre réponses, c'est dire aussi qu'il existe des individus qui ont choisi simultanément toutes ou presque toutes ces réponses.

L'analyse des correspondances multiples met alors en évidence des types d'individus ayant des profils semblables quant aux attributs choisis pour les décrire. Compte tenu des distances entre les éléments du tableau disjonctif complet et des relations barycentriques particulières, on exprime :

- *la proximité entre individus en terme de ressemblances :*
  deux individus se ressemblent s'ils ont choisi globalement les mêmes modalités.
- *la proximité entre modalités de variables différentes en terme d'association :*
  ces modalités correspondent aux points moyens des individus qui les ont choisies et sont proches parce qu'elles concernent globalement les mêmes individus ou des individus semblables.
- *la proximité entre deux modalités d'une même variable en terme de ressemblance :*
  par construction, les modalités d'une même variable s'excluent. Si elles sont proches, cette proximité s'interprète en terme de ressemblance entre les groupes d'individus qui les ont choisies (vis-à-vis d'autres variables actives de l'analyse).

Les règles d'interprétation des résultats (coordonnées, contributions, cosinus carrés) concernant les éléments actifs d'une analyse des correspondances multiples sont sensiblement les mêmes que celles d'une analyse des correspondances simple (cf. § 4.3.2). On calcule la contribution et la qualité de représentation de chaque modalité et de chaque individu, si ceux-ci ne sont pas anonymes pour l'analyse.

---

[1] C'est le cas du logiciel SPAD (qu'il s'agisse de l'ancienne version académique du CESIA ou de la version commerciale actuelle de la société éponyme) ou du logiciel académique DTM.

Cependant, la notion de variable doit être prise en compte au moment de l'interprétation, ceci au travers de ses modalités. Compte tenu de la décomposition de l'inertie du nuage des modalités, on calcule la contribution d'une variable au facteur α en sommant les contributions de ses modalités sur ce facteur :

$$Cr_\alpha(q) = \sum_{j \in q} Cr_\alpha(j)$$

On repère ainsi, en plus des modalités responsables des axes factoriels, les variables qui ont participé à la définition du facteur. On obtient un indicateur de liaison entre la variable et le facteur [cf. Escofier, 1979 c].

En revanche, les règles d'interprétation des valeurs propres et des taux d'inertie sont différentes (on a vu que la trace n'avait plus d'interprétation statistique).

c – Principes du découpage en classes

Les variables continues, pour être actives dans une analyse des correspondances multiples, doivent être soit rendues nominales (découpées en classes), soit recodées selon deux colonnes numériques[1].

Lorsque l'on cherche ainsi à découper une variable en classes, on est confronté à plusieurs problèmes : combien de classes choisir et comment les choisir ? Où placer les bornes des classes d'une variable continue ? La consultation de la distribution de chaque variable (tris-à-plat et histogrammes) est indispensable pour effectuer ces choix.

Certains principes, déduits des propriétés de l'analyse des correspondances multiples peuvent être utilisés pour guider la phase de recodage : constituer des modalités d'effectifs semblables, découper les variables de manière à avoir un nombre comparable de modalités. Pour donner un ordre de grandeur, un découpage entre 4 à 8 modalités convient dans la plupart des applications.

Il s'agit par conséquent de trouver un compromis entre un découpage techniquement acceptable selon ces principes et un découpage qui exhibe au mieux l'information à retenir. On ne peut généralement pas avoir recours à des algorithmes aveugles pour élaborer un découpage satisfaisant[2]. On retiendra par exemple une modalité de faible effectif si celle-ci est importante pour l'étude. De même pour sélectionner les bornes des classes d'une variable continue, on respectera un ou plusieurs seuils naturels dans le contexte de l'étude, ou significatifs après examen de l'histogramme (le découpage en classes d'amplitudes égales est parfois inapproprié).

---

[1] Cf. le recodage préconisé par Escofier (1979 b) présenté au chapitre 8, § 8.3.5.c.

[2] L'algorithme de W. D. Fisher (1958) fournit une partition optimale exacte (critère variance inter/variance totale maximal), mais ce critère rend très mal compte des mélanges de distributions ayant des variances très inégales et ne sépare donc pas des classes qu'une inspection visuelle d'histogramme distinguerait sans hésiter.

Ces principes sont moins rigoureux pour une variable supplémentaire. Celles-ci n'intervenant pas dans la formation des facteurs ou des classes, on a parfois intérêt à effectuer un découpage fin pour les variables supplémentaires.

La transformation de variables continues en variables nominales occasionne une perte de l'information brute mais présente certains avantages : exploiter simultanément des variables nominales et continues en correspondances multiples ; valider a posteriori les données en permettant d'observer l'éventuelle contiguïté des classes voisines ; et surtout : mettre en évidence les éventuelles liaisons non linéaires entre variables continues.

Pour un exposé de synthèse sur les méthodes de codage, on consultera Cazes (1990), Grelet (1993). L'article précité de Cazes et les travaux de Gallego (1982), van Rijckevorsel (1987) portent en particulier sur l'utilisation du codage flou en analyse des correspondances.

5.3 Analyse du tableau de contingence de Burt

Le tableau de Burt **B**, tableau de correspondances multiples, obtenu à partir d'un tableau disjonctif complet, est un assemblage particulier des tableaux de contingence qui sont les faces de l'hypercube de contingence.

5.3.1 Equivalence avec l'analyse du tableau disjonctif complet

L'analyse des correspondances appliquée à un tableau disjonctif complet **Z** est équivalente à l'analyse du tableau de Burt **B** et produit les mêmes facteurs.

L'analyse des correspondances du tableau de Burt **B**, tableau symétrique d'ordre *(p,p)*, se ramène à l'analyse d'un nuage de p points-modalités dans $\mathcal{R}^p$. Les marges de ce tableau, en ligne comme en colonne, sont les éléments diagonaux de la matrice s **D**.

Compte tenu de l'équation [5.2 - 2] donnant le $\alpha^{\text{ième}}$ facteur φ_α de l'analyse du tableau disjonctif complet **Z**, la matrice à diagonaliser est :

$$\mathbf{S} = \frac{1}{s}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z} = \frac{1}{s}\mathbf{D}^{-1}\mathbf{B}$$

Pour l'analyse du tableau de **B** associé à **Z**, le tableau des fréquences relatives **F** s'écrit :

$$\mathbf{F} = \frac{1}{ns^2}\mathbf{B}$$

et

$$\mathbf{D}_p = \mathbf{D}_n = \frac{1}{ns}\mathbf{D}$$

On diagonalise la matrice :

$$\mathbf{S}^* = \frac{1}{s^2}\mathbf{D}^{-1}\mathbf{B}\ \mathbf{D}^{-1}\mathbf{B}$$

ce qui donne : $$\mathbf{S}^* = \mathbf{S}^2$$

En prémultipliant les deux membres de [5.2 - 2] par $\frac{1}{s}\mathbf{D}^{-1}\mathbf{B}$, on obtient :

$$\frac{1}{s^2}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}\varphi_\alpha = \lambda_\alpha^2\varphi_\alpha$$

Les facteurs des deux analyses sont donc colinéaires dans $\mathcal{R}^p$ mais les valeurs propres associées diffèrent. Celles issues de l'analyse de **B**, notées λ_B, sont le carré de celles issues de l'analyse de **Z** :

$$\lambda_B = \lambda^2 \qquad [5.3 - 1]$$

Les facteurs φ_α issus de l'analyse de **Z**, représentant les coordonnées factorielles des modalités, ont pour norme λ, alors que le facteur correspondant de l'analyse de **B**, noté $\varphi_{B\alpha}$, aura pour norme λ^2.

D'où la relation liant les deux systèmes de coordonnées factorielles :

$$\varphi_{B\alpha} = \varphi_\alpha\sqrt{\lambda_\alpha} \qquad [5.3 - 2]$$

5.3.2 Equivalences dans le cas de deux questions

Dans le cas de deux questions q_1 et q_2, le tableau disjonctif complet s'écrit :

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$$

et nous ramène directement à l'analyse du tableau de contingence.

Il est alors équivalent, au point de vue de la description des associations entre modalités, d'effectuer (cf. figure 5.3 – 1) :

- l'analyse des correspondances du tableau **Z** d'ordre (n,p);
- l'analyse des correspondances du tableau **B** d'ordre (p,p);
- l'analyse des correspondances du tableau $\mathbf{K} = \mathbf{Z}_1'\mathbf{Z}_2$ d'ordre (p_1, p_2).

L'équivalence entre l'analyse des correspondances du tableau disjonctif complet **Z** et celle du tableau des correspondances multiples **B** a été donnée dans le cas général de plusieurs questions.

Intéressons-nous maintenant à l'équivalence entre l'analyse des correspondances du tableau disjonctif complet $\mathbf{Z}=[\mathbf{Z}_1, \mathbf{Z}_2]$ d'ordre (n,p) et celle du tableau de contingence $\mathbf{K} = \mathbf{Z}_1'\mathbf{Z}_2$ d'ordre (p_1, p_2) avec $p = p_1 + p_2$.

Montrons que, pour tout couple de facteurs $(\psi_\alpha, \varphi_\alpha)$ relatifs à une même valeur propre μ_α issus de l'analyse du tableau de contingence $\mathbf{Z}_1'\mathbf{Z}_2$, il correspond un facteur Φ_α de l'analyse de $\mathbf{Z}$ (ou celle de $\mathbf{B}$), avec :

$$\Phi_\alpha = \begin{bmatrix} \psi_\alpha \\ \varphi_\alpha \end{bmatrix}$$

Rappelons que l'on note $\mathbf{D}_1 = \mathbf{Z}_1'\mathbf{Z}_1$ et $\mathbf{D}_2 = \mathbf{Z}_2'\mathbf{Z}_2$ et que :

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}$$

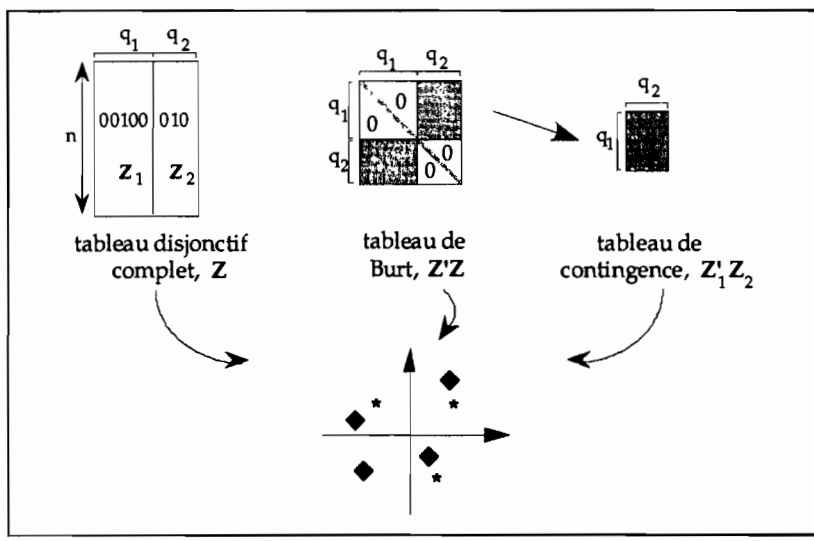


Figure 5.3 – 1. Equivalence des trois analyses des correspondances

Les éléments diagonaux de $\mathbf{D}_1$ et $\mathbf{D}_2$ sont respectivement les marges en ligne et en colonne du tableau $\mathbf{Z}_1'\mathbf{Z}_2$.

L'analyse de ce tableau nous conduit aux relations de transition de l'analyse des correspondances simple :

$$\begin{cases} \psi_\alpha = \dfrac{1}{\sqrt{\mu_\alpha}} \mathbf{D}_1^{-1}\, \mathbf{Z}_1'\mathbf{Z}_2\, \varphi_\alpha & [5.3-3] \\ \varphi_\alpha = \dfrac{1}{\sqrt{\mu_\alpha}} \mathbf{D}_2^{-1}\, \mathbf{Z}_2'\mathbf{Z}_1\, \psi_\alpha & [5.3-4] \end{cases}$$

On peut écrire ces relations sous la forme du système :

$$\begin{cases} \mathbf{D}_1^{-1}(\mathbf{D}_1\psi_\alpha + \mathbf{Z}_1'\mathbf{Z}_2\varphi_\alpha) = (1+\sqrt{\mu_\alpha})\psi_\alpha \\ \mathbf{D}_2^{-1}(\mathbf{D}_2\varphi_\alpha + \mathbf{Z}_2'\mathbf{Z}_1\psi_\alpha) = (1+\sqrt{\mu_\alpha})\varphi_\alpha \end{cases}$$

soit encore :

$$\begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{D}_1 & \mathbf{Z}_1'\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{D}_2 \end{bmatrix} \begin{bmatrix} \psi_\alpha \\ \varphi_\alpha \end{bmatrix} = (1+\sqrt{\mu_\alpha}) \begin{bmatrix} \psi_\alpha \\ \varphi_\alpha \end{bmatrix}$$

Cette équation s'écrit de façon plus condensée :

$$\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\,\mathbf{\Phi}_\alpha = (1+\sqrt{\mu_\alpha})\,\mathbf{\Phi}_\alpha \qquad [5.3-5]$$

Après multiplication des deux membres par $\frac{1}{s}$, soit ici $\frac{1}{2}$, il vient :

$$\frac{1}{s}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\,\mathbf{\Phi}_\alpha = (\frac{1+\sqrt{\mu_\alpha}}{2})\mathbf{\Phi}_\alpha$$

On y reconnaît la relation [5.2 - 2] avec :

$$\lambda_\alpha = \frac{1+\sqrt{\mu_\alpha}}{2}$$

Si μ_α est la $\alpha^{\text{ième}}$ plus grande valeur propre issue de l'analyse du tableau de contingence $\mathbf{Z}_1'\mathbf{Z}_2$, alors λ_α est la $\alpha^{\text{ième}}$ plus grande valeur propre issue de l'analyse de $\mathbf{Z}$.

Si par exemple $p_1 \leq p_2$, l'analyse de $\mathbf{Z}$ conduit à :

- p_1 facteurs du type $\begin{bmatrix} \psi_\alpha \\ \varphi_\alpha \end{bmatrix}$, correspondant à la valeur propre $\frac{1+\sqrt{\mu_\alpha}}{2}$
- p_1 facteurs du type $\begin{bmatrix} \psi_\alpha \\ -\varphi_\alpha \end{bmatrix}$, correspondant à la valeur propre $\frac{1-\sqrt{\mu_\alpha}}{2}$;
- p_2-p_1 facteurs du type[1] $\begin{bmatrix} 0 \\ \xi_\alpha \end{bmatrix}$, correspondant à la valeur propre $\frac{1}{2}$.

[1] Les axes ξ_α complètent la base des ψ_α dans $\mathcal{R}^p$

Les résultats relatifs aux trois analyses équivalentes sont rassemblés dans le tableau 5.3 - 1.

Tableau 5.3 – 1. Equivalence des analyses des trois tableaux dans le cas de deux questions

Tableau analysé	Dimension	Facteur	Valeur propre
$\mathbf{Z}_1'\mathbf{Z}_2$ tableau de contingence	(p_1, p_2)	ψ dans $\mathcal{R}^{p_1}$ φ dans $\mathcal{R}^{p_2}$	μ
$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ tableau disjonctif complet	(p,n) où $p = p_1 + p_2$.	$\Phi = \begin{bmatrix} \psi \\ \varphi \end{bmatrix}$	$\lambda = \frac{1+\sqrt{\mu}}{2}$
$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$ Tableau de Burt	(p,p)	$\Phi_B = \Phi\sqrt{\lambda}$	λ^2

Remarques :

1) Les analyses de correspondances appliquées à ces trois types de tableaux, reposant sur la même information brute, donnent les mêmes axes factoriels, mais avec des valeurs propres différentes, donc des taux d'inertie différents. Les relations existant entre les taux d'inertie nous montrent que ceux-ci seront toujours plus élevés pour l'analyse du tableau de contingence $\mathbf{Z}_1'\mathbf{Z}_2$ que pour l'analyse du tableau disjonctif complet **Z**.
Ainsi, la somme des valeurs propres non triviales issues de l'analyse de **Z** vaut :

$$\frac{p_1 + p_2}{2} - 1$$

Comme les valeurs propres sont inférieures ou égales à *1*, aucun facteur ne peut avoir un taux d'inertie supérieur en pourcentage à :

$$\frac{2 \times 100}{p_1 + p_2 - 2}$$

Prenons l'exemple du tableau de contingence croisant les *8* professions et les *6* médias (cf. § 1.3.8). Le premier facteur prend en compte *50%* de l'inertie totale. La remarque ci-dessus montre que l'analyse du tableau disjonctif correspondant ne peut pas donner un premier facteur expliquant plus de $\frac{200}{8+6-2} = 16{,}6\%$. Les taux d'inertie sont donc dépendants du codage préliminaire de l'information brute. Il faut donc éviter de les interpréter en termes "d'information". On reviendra sur ce point à propos d'un exemple à la section 5.5.2.

2) Dans l'analyse du tableau disjonctif complet **Z**, les points représentant les diverses modalités de réponses aux deux questions sont les éléments d'un même ensemble, l'ensemble des colonnes de **Z**. Au contraire dans l'analyse du tableau de contingence $\mathbf{Z}_1'\mathbf{Z}_2$, ils se scindent en points-lignes et en points-colonnes (cf. figure 5.3 - 2).

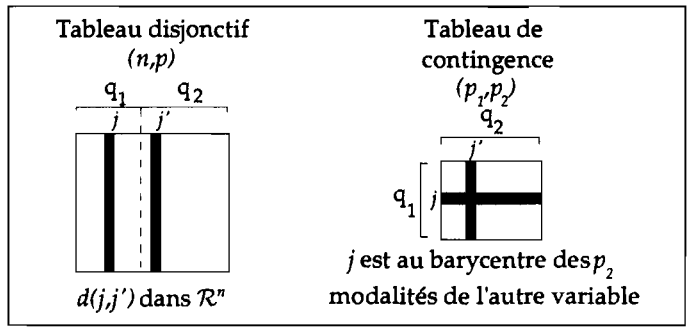


Figure 5.3 – 2. Proximité entre deux modalités de variables différentes

Le fait que les représentations obtenues dans l'espace des premiers facteurs soient identiques (à une dilatation près, due au fait que les valeurs propres ne sont pas les mêmes) montre que la représentation simultanée des points-lignes et des points-colonnes en analyse des correspondances n'est pas un simple artifice graphique.

L'interprétation de la position de deux modalités relatives à deux variables différentes dépend du tableau d'analyse. Dans le tableau disjonctif complet, cette position s'interprète en terme de distance. Dans le tableau de contingence, la distance entre une ligne et une colonne n'a pas de sens et une modalité est au "quasi-barycentre" des modalités de l'autre variable. L'analyse de ces deux tableaux fournit des représentations similaires.

5.3.3 Autres équivalences

a – Cas où toutes les questions ont deux modalités : équivalence avec l'analyse en composantes principales

Dans le cas où toutes les variables ont deux modalités, l'analyse des correspondances multiples se ramène à l'analyse en composantes principales des variables caractérisées par une seule de leurs modalités ($p - s = \frac{p}{2}$).

Les variables n'étant représentées que par une seule de leurs modalités, on obtient directement la matrice à diagonaliser qui n'est autre que la matrice des corrélations entre variables (Nakhlé, 1976). La démonstration de cette propriété, assez technique, figure dans l'annexe 5.7 de ce chapitre.

b - Sous-tableau d'un tableau de correspondances multiples

Dans le cas où l'ensemble des s questions est partitionné en au moins deux groupes s_1 et s_2 à l'intérieur desquels les questions sont indépendantes, on peut vouloir analyser la correspondance entre les deux groupes en effectuant

l'analyse des correspondances du sous-tableau du tableau de Burt $\mathbf{B}_{12}$ obtenu en croisant les deux sous-ensembles s_1 et s_2. L'analyse du tableau des correspondances multiples **B** permet d'étudier les liaisons entre toutes les questions.

L'analyse du sous-tableau $\mathbf{B}_{12}$ permet d'étudier les relations existant entre les éléments de s_1 et ceux de s_2 sans tenir compte des dépendances internes à s_1, ni des dépendances internes à s_2. Le groupe de questions s_1 est caractérisé par ses associations avec les questions de s_2 et réciproquement (cf. Leclerc, 1975).

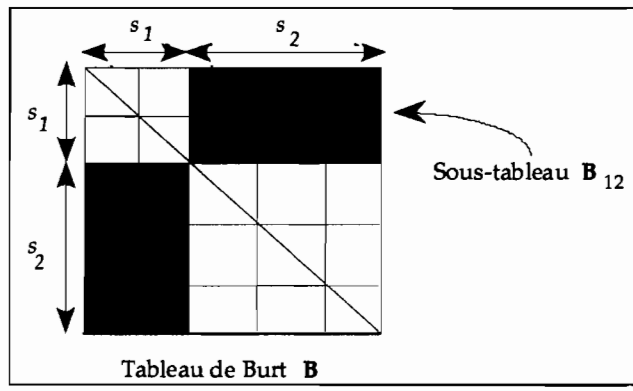


Figure 5.3 – 3. Sous-tableau $\mathbf{B}_{12}$ du tableau de contingence de Burt B

Lorsqu'un des groupes est réduit à une seule question q_0, le tableau de données est une bande du tableau des correspondances multiples croisant la variable q_0 avec un groupe de variables ne contenant pas q_0. C'est aussi le tableau des barycentres des groupes d'individus définis par les modalités de q_0. Nous verrons au paragraphe 7.4.2-f que l'analyse d'une bande d'un tableau de correspondances multiples constitue une méthode de discrimination appelée analyse discriminante barycentrique. Les résultats obtenus par l'analyse des correspondances du tableau de Burt **B** et celle de la tranche $\mathbf{B}_{12}$ sont en général différents (les nuages relatifs à ces tableaux ne sont pas dans le même espace).

Ce sont les objectifs de l'étude qui doivent guider le choix du tableau à analyser. Cependant, si les variables de chaque sous-ensemble sont indépendantes entre elles, les analyses réalisées à partir des tableau **B** et $\mathbf{B}_{12}$ sont équivalentes et celles de chaque sous-ensemble s_1 et s_2 ne présentent pas d'intérêt.

c – Cas où l'analyse multiple se ramène à une correspondance binaire

Les deux sous-ensembles s_1 et s_2 totalisent respectivement p_1 et p_2 modalités (avec $p_1 + p_2 = p$). Le cas d'une correspondance binaire s'est révélé particulièrement intéressant du point de vue des calculs à mettre en œuvre. En effet, l'analyse du tableau des correspondances multiples d'ordre (p,p) est équivalente à l'analyse des correspondances du tableau de contingence croisant

les modalités des deux questions, ce qui conduit à diagonaliser une matrice dont l'ordre est déterminé par le plus petit des nombres p_1 et p_2.

Nous retiendrons la propriété suivante. Si à l'intérieur des deux sous-ensembles s_1 et s_2 les questions sont indépendantes, l'analyse des s questions se ramène à celle d'une correspondance binaire, et donc à la diagonalisation d'une matrice d'ordre $Inf(p_1, p_2)$.

Nous dirons ici que deux questions q et q' sont indépendantes si la table de contingence correspondante vérifie la relation[1] :

$$\mathbf{Z}'_q\mathbf{Z}_{q'} = \frac{1}{n}\mathbf{d}_q\mathbf{d}_{q'}$$

où les vecteurs $\mathbf{d}_q$ et $\mathbf{d}_{q'}$ ont respectivement pour composantes les éléments diagonaux de $\mathbf{Z}'_q\mathbf{Z}_q$ et $\mathbf{Z}'_{q'}\mathbf{Z}_{q'}$ (c'est-à-dire les éléments diagonaux de $\mathbf{D}_q$ et $\mathbf{D}_{q'}$ par définition de ces matrices).

Ecrivons de nouveau la relation [5.3 - 6] en partitionnant Φ en deux blocs Φ_{s_1} et Φ_{s_2} ; on découpe également les matrices **B** et **D** en quatre blocs, de façon à faire apparaître la partition $s = s_1 \approx s_2$:

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}$$

On obtient les deux relations :

$$\begin{cases} \frac{1}{s}(\mathbf{D}_1^{-1}\mathbf{B}_{11}\Phi_{s_1} + \mathbf{D}_1^{-1}\mathbf{B}_{12}\Phi_{s_2}) = \lambda\Phi_{s_1} \\ \frac{1}{s}(\mathbf{D}_2^{-1}\mathbf{B}_{21}\Phi_{s_1} + \mathbf{D}_2^{-1}\mathbf{B}_{22}\Phi_{s_2}) = \lambda\Phi_{s_2} \end{cases}$$

Remarquons que les s_1 (respectivement s_2) blocs diagonaux de $\mathbf{D}_1^{-1}\mathbf{B}_{11}$ (respectivement $\mathbf{D}_2^{-1}\mathbf{B}_{22}$) sont des matrices unité dont les ordres correspondent aux cardinaux de chacune des questions .

On a d'autre part, pour $k \in \{1, 2\}$:

$$q \in s_k ; q' \in s_k ; q \neq q' \quad \Rightarrow \quad \mathbf{D}_q^{-1}\mathbf{Z}'_q\mathbf{Z}_{q'} = \frac{1}{n}\mathbf{D}_q^{-1}\mathbf{d}_q\mathbf{d}'_{q'}$$

En désignant par $\mathbf{e}_q$ un vecteur dont les q composantes valent 1 :

$$\mathbf{D}_q^{-1}\mathbf{Z}'_q\mathbf{Z}'_{q'} = \frac{1}{n}\mathbf{e}_q\mathbf{d}'_q$$

---

[1] Bien entendu, l'indépendance théorique entre les deux questions n'implique pas que cette relation soit exactement vérifiée sur l'échantillon.

Les relations $\mathbf{d}_q'\Phi_{q'} = 0$ (centrage des modalités relatives à chaque question) impliquent finalement :

$$\mathbf{D}_1^{-1}\mathbf{B}_{11}\Phi_{s_1} = \Phi_{s_1} \qquad \text{et} \qquad \mathbf{D}_2^{-1}\mathbf{B}_{22}\Phi_{s_2} = \Phi_{s_2}$$

Le système ci-dessus s'écrit alors :

$$\begin{cases} \mathbf{D}_1^{-1}\mathbf{B}_{12}\ \Phi_{s_2} = (\lambda s - 1)\ \Phi_{s_1} \\ \mathbf{D}_2^{-1}\mathbf{B}_{21}\ \Phi_{s_1} = (\lambda s - 1)\ \Phi_{s_2} \end{cases}$$

D'où par substitution :

$$\mathbf{D}_2^{-1}\mathbf{B}_{21}\mathbf{D}_1^{-1}\mathbf{B}_{12}\Phi_{s_2} = (\lambda s - 1)^2\Phi_{s_2}$$

Ainsi Φ_{s_2} est obtenu par diagonalisation d'une matrice d'ordre (s_2, s_2). On en déduit facilement Φ_{s_1}.

Remarquons que $\mathbf{B}_{12}$ est obtenu par juxtaposition des tableaux de contingence croisant l'ensemble des modalités des questions du premier groupe avec celles relatives au second groupe. Les marges du tableau $\mathbf{B}_{12}$ sont les éléments diagonaux de $s_2\mathbf{B}_1$ et $s_1\mathbf{B}_2$.

Les facteurs issus de l'analyse des correspondances directe du tableau $\mathbf{B}_{12}$ considéré comme un tableau de contingence vérifient la relation :

$$\frac{1}{s_1 s_2}\mathbf{D}_2^{-1}\mathbf{B}_{21}\mathbf{D}_1^{-1}\mathbf{B}_{12}\Psi = \lambda\Psi$$

Ils sont donc proportionnels aux facteurs trouvés précédemment[1].

5.3.4 Liens avec l'analyse canonique

L'analyse canonique contient comme cas particulier l'analyse des correspondances simple et peut se généraliser au cas de plus de deux variables nominales.

En reprenant les notations du présent chapitre, le tableau de données $\mathbf{R} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_q, \ldots, \mathbf{Z}_s]$ à n lignes et p colonnes est le tableau disjonctif complet juxtaposant s sous-tableaux. Chaque sous-tableau $\mathbf{Z}_q$ correspond à une question q totalisant p_q modalités de réponse et engendre, dans l'espace $\mathcal{R}^p$, un sous-espace V_{Z_q} à p_q dimensions[2].

---

[1] Ces propriétés concernant les sous-tableaux de tableaux de Burt ont été étudiées par A. Leclerc (1975), puis généralisées par P. Cazes (cf. Cazes, 1977, 1980, 1981).

[2] Rappelons que les s sous-espaces ont en commun au moins la première bissectrice. Le rang de R est donc au plus égal à $p - s + 1$.

a – Le cas de l'analyse des correspondances simples

L'analyse des correspondances du tableau de contingence croisant deux variables q et q' revient à étudier les positions relatives des sous-espaces V_{Z_q} et $V_{Z_{q'}}$. C'est l'analyse canonique du tableau $[\mathbf{Z}_q, \mathbf{Z}_{q'}]$.

Soit φ_q le vecteur dont les p_q composantes sont les coordonnées d'un point $\mathbf{m}_q$ de V_{Z_q} dans la base définie par les colonnes de $\mathbf{Z}_q$.

Les coordonnées de $\mathbf{m}_q$ dans $\mathcal{R}^n$ sont les composantes de $\mathbf{m}_q = \mathbf{Z}_q \varphi_q$.

Le carré de la distance de ce point $\mathbf{m}_q$ à l'origine, selon la norme euclidienne usuelle, n'est autre que :

$$\varphi'_q \mathbf{Z}'_q \mathbf{Z}_q \varphi_q = \varphi'_q \mathbf{D}_q \varphi_q$$

Les relations de double transition [4.3 - 3] et [4.3 - 4] s'écrivent ici (en omettant l'indice α de l'axe pour alléger les notations) :

$$\begin{cases} \varphi_q = \frac{1}{\sqrt{\lambda}} \mathbf{D}_q^{-1} \mathbf{Z}'_q \mathbf{Z}_{q'} \varphi_{q'} \\ \varphi_{q'} = \frac{1}{\sqrt{\lambda}} \mathbf{D}_{q'}^{-1} \mathbf{Z}'_{q'} \mathbf{Z}_q \varphi_q \end{cases}$$

On en déduit le système suivant :

$$\begin{cases} \mathbf{Z}_q \varphi_q = \frac{1}{\sqrt{\lambda}} \mathbf{Z}_q \mathbf{D}_q^{-1} \mathbf{Z}'_q \mathbf{Z}_{q'} \varphi_{q'} \\ \mathbf{Z}_{q'} \varphi_{q'} = \frac{1}{\sqrt{\lambda}} \mathbf{Z}_{q'} \mathbf{D}_{q'}^{-1} \mathbf{Z}'_{q'} \mathbf{Z}_q \varphi_q \end{cases}$$

soit :

$$\mathbf{m}_q = \frac{1}{\sqrt{\lambda}} \mathbf{P}_q \mathbf{m}_{q'} \qquad [5.3 - 5]$$

$$\mathbf{m}_{q'} = \frac{1}{\sqrt{\lambda}} \mathbf{P}_{q'} \mathbf{m}_q \qquad [5.3 - 6]$$

où :

$$\mathbf{P}_q = \mathbf{Z}_q (\mathbf{Z}'_q \mathbf{Z}_q)^{-1} \mathbf{Z}'_q \qquad \text{et} \qquad \mathbf{P}_{q'} = \mathbf{Z}_{q'} (\mathbf{Z}'_{q'} \mathbf{Z}_{q'})^{-1} \mathbf{Z}'_{q'}$$

Les matrices $\mathbf{P}_q$ et $\mathbf{P}_{q'}$ représentent respectivement les opérateurs projection sur les sous-espaces V_{Z_q} et $V_{Z_{q'}}$.

Les relations [5.3 - 5] et [5.3 - 6] expriment que la projection orthogonale de $\mathbf{m}_q$ sur $V_{Z_{q'}}$ est colinéaire à $\mathbf{m}_{q'}$ (et semblablement pour $\mathbf{m}_{q'}$ sur V_{Z_q}).

Présentée comme la recherche des plus petits angles entre deux sous-espaces V_{Z_q} et $V_{Z_{q'}}$, l'analyse canonique ne se généralise pas facilement au cas de plus de deux questions[1].

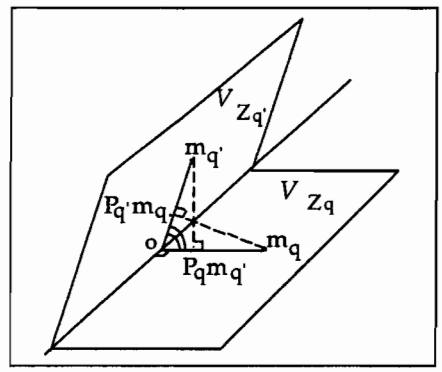


Figure 5.3 – 4. Projections sur V_{Z_q} et $V_{Z_{q'}}$

Mais une autre formulation va permettre de présenter l'analyse des correspondances multiples comme une analyse canonique généralisée particulière.

b – L'analyse des correspondances multiples

L'analyse canonique du tableau $[\mathbf{Z}_q, \mathbf{Z}_{q'}]$ peut aussi se formuler de la façon suivante :

> ▶ trouver deux points $\mathbf{m}_q$ et $\mathbf{m}_{q'}$ tels que la somme des carrés de leurs distances à l'origine soit constante :
>
> $$\varphi'_q \mathbf{D}_q \varphi_q + \varphi'_{q'} \mathbf{D}_{q'} \varphi_{q'} = 2n \qquad [5.3 - 7]$$
>
> et tels que la distance à l'origine du point $\mathbf{m} = \mathbf{m}_q + \mathbf{m}_{q'}$ soit maximale.

En effet, cette distance a pour carré :

$$\|\mathbf{m}\|^2 = \varphi'_q \mathbf{D}_q \varphi_q + \varphi'_{q'} \mathbf{D}_{q'} \varphi_{q'} + 2\varphi'_q \mathbf{Z}'_q \mathbf{Z}_{q'} \varphi_{q'}$$

soit :

$$\|\mathbf{m}\|^2 = 2n\left(1 + \frac{1}{n}\varphi'_q \mathbf{Z}'_q \mathbf{Z}_{q'} \varphi_{q'}\right)$$

---

[1] On reviendra sur ce lien entre analyse de correspondances et analyse canonique au chapitre 7, à propos de l'analyse factorielle discriminante, qui est elle aussi une analyse canonique particulière.

Rendre maximale $\|\mathbf{m}\|^2$ avec la contrainte [5.3 - 7], ou avec les deux contraintes :

$$\varphi'_q \mathbf{D}_q \varphi_q = \varphi'_{q'} \mathbf{D}_{q'} \varphi_{q'} = n$$

conduit au même résultat. En effet, les multiplicateurs de Lagrange relatifs à ces deux dernières contraintes sont égaux.

Avec la contrainte unique [5.3 - 7], le problème se généralise aisément au cas de plus de deux questions.

On désigne par $\varphi_1, \ldots, \varphi_{q'}, \ldots, \varphi_s$ respectivement les vecteurs des composantes de s points $\mathbf{m}_1, \ldots, \mathbf{m}_{q'}, \ldots, \mathbf{m}_s$ dans les bases $\mathbf{Z}_1, \ldots, \mathbf{Z}_{q'}, \ldots, \mathbf{Z}_s$ et soit $\mathbf{m} = \mathbf{m}_1 + \ldots + \mathbf{m}_q + \ldots + \mathbf{m}_s$.

On cherchera à rendre maximale la quantité :

$$\|\mathbf{m}\|^2 = \sum_{q \in s} \sum_{q' \in s} \varphi'_q \mathbf{Z}'_q \mathbf{Z}_{q'} \varphi_{q'}$$

avec la contrainte :

$$\sum_{q \in s} \varphi'_q \mathbf{D}_q \varphi_q = sn$$

Si Φ désigne le vecteur à p composantes défini par :

$$\Phi' = \{\varphi'_1, \ldots, \varphi'_q, \ldots, \varphi'_s\}$$

le problème revient à rendre maximal :

$$\Phi' \mathbf{B} \Phi$$

avec la contrainte :

$$\Phi' \mathbf{D} \Phi = sn$$

où l'on rappelle que $\mathbf{B}$ est le tableau de contingence de Burt obtenu à partir du tableau disjonctif complet.

Les facteurs Φ cherchés sont donc les vecteurs propres de $\mathbf{D}^{-1}\mathbf{B}$ relatifs aux plus grandes valeurs propres.

Il s'agit d'une généralisation simple de l'analyse canonique au cas de plus de deux ensembles : elle conduit à une diagonalisation de matrice symétrique, opération classique et maîtrisée.

Cette extension de l'analyse canonique sera présentée à nouveau dans un cadre plus général au chapitre 8.

Les autres méthodes (introduction de s contraintes au lieu d'une seule) demandent des algorithmes itératifs assez coûteux et ne conduisent pas à des règles d'interprétation simples.

5.4 Méthodes de validation

Tout comme ce fut le cas en analyse en composantes principales et en analyse des correspondances simples, nous sommes amenés à nous interroger sur la validité des facteurs et leur stabilité.

Nous nous focalisons sur deux méthodes de validation présentées dans les chapitres précédents. La première relève de procédures externes et est basée sur le positionnement de variables qualitatives ou quantitatives supplémentaires (cf § 3.4.2-b et 4.5.3-b). La seconde s'appuie sur les méthodes de bootstrap (cf §3.4.4. et § 4.4.2).

5.4.1 Validation externe : éléments supplémentaires

L'utilisation des éléments supplémentaires en analyse des correspondances multiples permet de prendre en compte toute l'information susceptible d'aider à comprendre ou à interpréter la typologie induite par les éléments actifs.

Ceci est particulièrement intéressant lorsque l'ensemble des variables se décompose en thème, c'est-à-dire en groupes de variables homogènes quant à leur contenu.

Dans l'analyse du tableau disjonctif complet, on fera intervenir des éléments supplémentaires pour :

- Enrichir l'interprétation des axes par des variables n'ayant pas participé à leur construction. On projettera alors dans l'espace des variables les centres de groupes d'individus définis par les modalités des variables supplémentaires.
- Adopter une optique de prévision en projetant les variables supplémentaires dans l'espace des individus. Celles-ci seront "expliquées" par les variables actives. On peut projeter des individus supplémentaires dans l'espace des variables, pour les situer par rapport aux individus actifs ou par rapport à des groupes d'individus actifs dans une optique de discrimination (cf. chapitre 7).

Suivant la nature des variables supplémentaires, nominales ou continues, on interprète différemment leur position sur les axes factoriels.

a – Valeurs-test pour les modalités supplémentaires

Tout comme pour l'analyse des correspondances simples, il n'est pas nécessaire de projeter en supplémentaire *toutes* les modalités d'une variable nominale.

La coordonnée factorielle $\varphi_{\alpha j}$ d'une modalité j sur un axe α (que cette modalité figure parmi les variables actives ou qu'elle soit supplémentaire) est le produit

par le coefficient $\frac{1}{\sqrt{\lambda_\alpha}}$ de la moyenne arithmétique des coordonnées $\psi_{\alpha i}$ des individus ayant choisi cette modalité j de réponse :

$$\varphi_{\alpha j} = \frac{1}{z_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in I(j)} \psi_{\alpha i}$$

où $I(j)$ est l'ensemble des individus ayant choisi la modalité j. Ceci suggère alors le test d'hypothèse suivant :

Supposons qu'une modalité supplémentaire j concerne n_j individus ($n_j = z_{.j}$). Si ces n_j individus sont tirés au hasard (hypothèse nulle H_0) parmi les n individus analysés (tirage supposé sans remise), la moyenne de n_j coordonnées tirées au hasard dans l'ensemble fini des n valeurs $\psi_{\alpha i}$ est une variable aléatoire $X_{\alpha j}$:

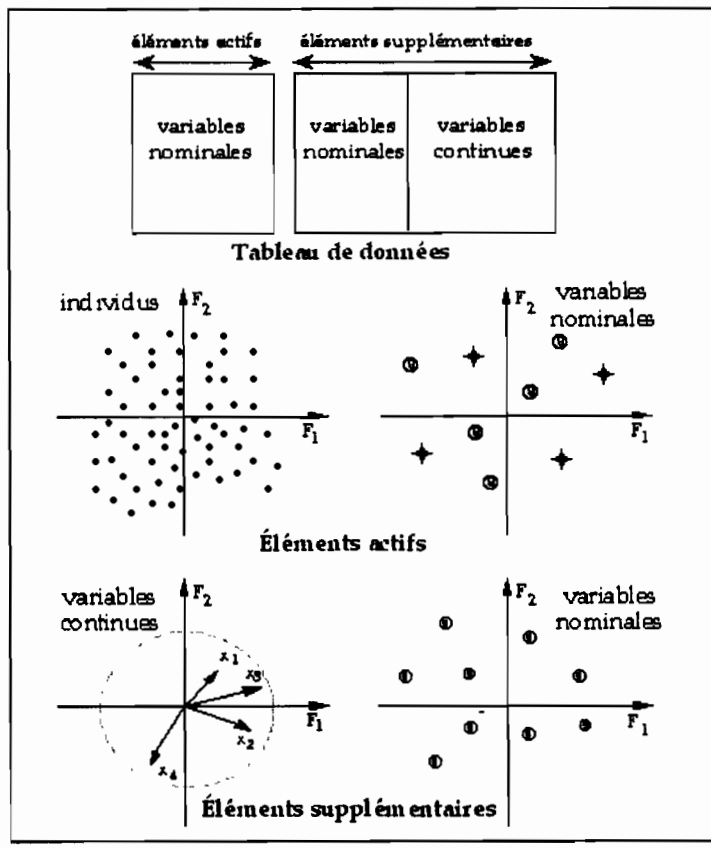


Figure 5.4 – 1. Représentation des variables supplémentaires en analyse des correspondances multiples

avec pour espérance :

$$E(X_{\alpha j}) = 0$$

et pour variance[1] :

$$Var(X_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{\lambda_\alpha}{n_j}$$

La coordonnée $\varphi_{\alpha j}$ de la modalité supplémentaire est liée à la variable aléatoire $X_{\alpha j}$ par la relation :

$$\varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} X_{\alpha j}$$

On a donc :

$$E(\varphi_{\alpha j}) = 0$$

et

$$Var(\varphi_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{1}{n_j}$$

La quantité $t_{\alpha j}$:

$$t_{\alpha j} = \sqrt{n_j \frac{n-1}{n-n_j}}\ \varphi_{\alpha j}$$

mesure *en nombre d'écart-types* la distance entre la modalité *j*, c'est-à-dire le quasi-barycentre des n_j individus, et l'origine sur l'axe factoriel α. On appelle cette quantité "*valeur-test*".

D'après le théorème de la limite centrale, sa distribution tend vers une loi de Laplace-Gauss centrée réduite. Ainsi, la position d'une modalité est intéressante dans une direction α donnée si le sous-nuage qu'elle constitue occupe une zone étroite dans cette direction et si cette zone est éloignée du centre de gravité du nuage.

La valeur-test est un critère qui permet d'apprécier rapidement si une modalité a une position "significative" sur un axe. On considère généralement comme occupant une "position significative" les modalités dont les valeurs-test sont supérieures à 2 en valeur absolue, correspondant approximativement au seuil 5%.

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples*, bien connu des statisticiens. Supposons que l'on projette *100* modalités supplémentaires qui soient vraiment tirées au hasard. Les valeurs-test attachées à ces modalités sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, en moyenne, sur *100* valeurs-test calculées, *5* seront en dehors de l'intervalle [*-1.96* , *+1.96*], et *5* dépasseront la valeur *1.65*

[1] Il s'agit de la formule classique donnant la variance d'une moyenne lors d'un tirage sans remise de n_j objets parmi n, en fonction de la variance totale λ_α.

(test unilatéral). Le seuil de 5% n'a de sens en fait que pour un seul test, et non pour des tests multiples. On résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère[1].

On note que les valeurs-test n'ont de sens que pour les modalités supplémentaires ou encore pour les modalités actives ayant des contributions absolues faibles, c'est-à-dire se comportant comme des modalités supplémentaires[2]. Lorsque l'on dispose d'un nombre important de modalités supplémentaires, les valeurs-test permettent de repérer rapidement les modalités utiles à l'interprétation d'un axe ou d'un plan factoriel.

b – Variables continues supplémentaires

Il est possible de positionner des variables continues en élément supplémentaire (sans transformation au préalable en variable nominale par découpage en classes). On calcule, comme dans l'analyse en composantes principales, le coefficient de corrélation de ces variables avec le facteur. Celui-ci fournit la coordonnée de la variable continue sur l'axe factoriel (cf. la schématisation de la figure 5.2 - 5). Les carrés des coefficients obtenus sont l'équivalent des cosinus carrés. La position d'une variable sur un plan définit donc la direction où se situent les fortes valeurs de la variable. Ceci est d'autant plus vrai que la variable est proche du cercle des corrélations (de rayon *1*) : il existe dans ce cas une liaison forte et linéaire entre la variable et les facteurs [3].

5.4.2 Validation interne : inertie et méthode de bootstrap

a – Taux d'inertie et information

L'utilisation des taux d'inertie (ou pourcentages de variance) comme outil d'évaluation globale de la qualité d'une représentation est très délicate.

Les taux d'inertie sont des mesures pessimistes de la qualité d'une représentation (contrairement, par exemple, aux coefficients de corrélation

---

[1] Plusieurs solutions ont été proposées pour ce faire : la méthode de *Bonferroni* recommande de diviser le seuil statistique par le nombre de tests (ici, le nombre de valeurs-test calculées). Cette réduction du seuil est souvent considérée comme étant trop sévère (Hochberg, 1988; Perneger, 1998 ; Saville 1990). Des études approfondies de ce problème se trouvent dans Hsu (1996), Westfall et Young (1993). Les valeurs-test permettent surtout de classer les modalités supplémentaires par ordre d'intérêt décroissant, ce qui constitue une aide précieuse à l'interprétation des facteurs.

[2] Les coordonnées sur un axe des individus correspondant à une modalité active ne peuvent être considérées comme tirées au hasard, puisque cette modalité a contribué à construire l'axe.

[3] La lecture de la trajectoire des classes d'une variable continue transformée en variable nominale apporte souvent plus de précision que la seule position de la variable considérée comme continue (détection éventuelle de liaisons non linéaires).

multiple qui sont des mesures optimistes de la qualité d'une régression). La variance brute initiale n'étant pas en général une mesure de référence adéquate, il est souvent injustifié de parler de *part d'information* à propos des *taux d'inertie.* Quelques contre-exemples montrent que ces coefficients ne sont pas aptes à caractériser de façon satisfaisante la qualité d'une représentation.

Nous avons vu à la première remarque du paragraphe 5.3.2 – a que, pour une même représentation, l'analyse de deux questions (ou variables) sous codage disjonctif peut donner des taux d'inertie considérablement plus faibles que l'analyse, pourtant équivalente, du tableau de contingence croisant les deux variables. En effet, le codage disjonctif, en introduisant une orthogonalité entre les colonnes (modalités) relatives à une même question, introduit une sorte de sphéricité artificielle du nuage de points-profils, que l'on retrouve dans la forme du spectre.

Comme cela sera évoqué dans l'exemple d'application (cf. § 5.5.2), Benzécri (1979) a proposé une formule de calcul de taux d'inertie à partir de pseudo-valeurs propres $\rho(\lambda)$ obtenues par la formule suivante :

$$\rho(\lambda) = \left(\frac{s}{s-1}\right)^2 \left(\lambda - \frac{1}{s}\right)^2 \qquad \text{pour } \lambda > \frac{1}{s}$$

où s représente le nombre de questions actives, λ représente la valeur propre issu de l'analyse des correspondances du tableau disjonctif complet, (λ^2 étant la valeur propre issue de l'analyse des correspondances du tableau de Burt).

Les valeurs propres issues du tableau de Burt dont la diagonale a été annulée sont précisément $\left(\lambda - \frac{1}{s}\right)^2$ et seulement celles qui vérifient $\lambda > \frac{1}{s}$ correspondent à des facteurs directs.

De plus, dans le cas $s = 2$, on retrouve les taux d'inertie de l'analyse des correspondances de la vraie table de contingence croisant les deux questions[1].

b – Bootstrap pour l'analyse des correspondances multiples

Nous avons vu au chapitre précédent une application du bootstrap à la validation des représentations issues de l'analyse des tables de contingence simples.

---

[1] Dans le cas de l'exemple numérique de la section 4.5, le taux correspondant à la première valeur propre (22.77%) devient alors 64%. Greenacre (cf. Greenacre et Blasius, 1994) propose une modification itérative du tableau de Burt qui conduit à des représentations très similaires, mais à des taux intermédiaires entre les taux bruts et les taux rectifiés (sous le nom de *Joint Correspondence Analysis*).

Dans le cas de l'analyse des correspondances multiples, une réplication bootstrap est obtenue en tirant avec remise les individus, lignes du tableau de données **R** ou de façon équivalente, lignes du tableau disjonctif associé **Z**.

Chaque réplication permet de construire un tableau de Burt, dont les lignes sont projetées en éléments supplémentaires dans les plans factoriels issus de l'analyse du tableau de Burt initial.

Les zones de confiance obtenues sont d'autant plus utiles ici pour choisir la dimension de l'espace de représentation que les valeurs propres et les taux d'inertie sont, on l'a vu, d'une interprétation difficile.

Les premières application du bootstrap pour évaluer la stabilité et pour construire des zones de confiance à partir d'une analyse des correspondances multiples (*homogeneity analysis* selon la terminologie de ces auteurs) sont celles de Gifi (1981), Meulman (1982), puis Greenacre (1984), Markus (1994).

En fait, le cas des correspondances multiples est en tout point analogue à celui des composantes principales, car ces deux méthodes ont en commun des tableaux dont une des dimensions est la dimension « individu ». Dans les deux cas, les réplications se font par tirage avec remise dans l'ensemble des individus.

En pratique, il s'agit dans les deux cas de générer, pour chaque réplication, un « poids bootstrap entier » $p(i)$ pour l'individu i, qui est le nombre d'apparition de l'individu i dans la réplication. La somme des $p(i)$ pour les n individus vaut n.

On se référera donc à la section 3.4.4 du chapitre 3 pour les différentes options de bootstrap possible : *bootstrap partiel*, *bootstrap total* de trois types (*type 1* : simple correction du signe des axes ; *type 2* : correction supplémentaire des éventuelles interversions d'axes ; *type 3* : rotation procustéennes pour rapprocher les réplications de l'échantillon initial.

► *Un trait spécifique de l'analyse des correspondances multiples :*

Le manque de robustesse de l'analyse des correspondances multiples en présence de modalités d'effectifs très faibles souligné en section 5.2.2-a va rendre les épreuves de *bootstrap total* particulièrement sévères lorsque la table initiale contient des effectifs faibles, effectifs qui pourront devenir très faibles dans certaines réplications.

L'exemple qui sera traité à la section suivante, parce qu'il porte sur un sous-échantillon de *105* individus (extraits d'une enquête annuelle comportant *2000* individus par an) illustrera cette fragilité vis-à-vis du bootstrap total.

5.5 Interprétation et validation à propos d'un exemple

L'exemple qui suit concerne un petit sous-échantillon (*105* individus, *9* questions) de l'enquête "Conditions de vie et aspirations des Français", enquête qui sera présentée et détaillée au paragraphe 6.4.4 du chapitre 6, à propos de la complémentarité entre analyse factorielle et classification.

5.5.1 Description des données

Le tableau 5.5 - 1 est le tableau de données proprement dit, en codage condensé (cf. paragraphe 5.1.2), à l'exception de la variable V2 (âge) qui est numérique.

Les libellés des questions figurent dans le tableau 5.5 - 2, les libellés des modalités correspondantes se retrouveront dans les listages de résultats plus bas.

Les *4* variables actives servent à calculer les distances et les axes, les *4* variables illustratives et la variable continue illustrative servent à interpréter *a posteriori* les axes et les proximités.

Les tableaux disjonctifs complets correspondant aux variables nominales ne sont pas présentés et ne sont jamais développés tels quels dans les calculs. Le tableau de Burt (tableau 5.3 - 3) est calculé directement à partir du codage condensé[1].

Il ne représente que la moitié inférieure du tableau de Burt relatif aux *4* questions actives. On trouve dans ce tableau les *6* tableaux de contingence croisant les *4* questions actives deux à deux. Sur la diagonale se trouvent les questions croisées avec elles-mêmes, et donc les effectifs correspondant à chaque modalité.

5.5.2 Eléments d'interprétation

On vérifie ensuite (tableau 5.5 - 4) qu'il y a *6* valeurs propres non nulles ($6 = p - s$), et on peut constater que les taux d'inertie correspondant à chaque valeur propre sont modestes, malgré la petite taille de cet exemple pédagogique.

---

[1] Cette procédure divise le nombre d'opérations par le coefficient $(p/s)^2$, s étant le nombre de questions actives et p le nombre total de modalités correspondantes. Dans le cas d'applications courantes ce gain est appréciable.

Tableau 5.5 – 1. Tableau de données R en codage condensé

n°	V1	V2	V3	V4	V5	V6	V7	V8	V9	n°	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	2	27	1	2	2	1	1	2	1	54	2	54	1	2	2	2	1	1	1
2	2	42	1	3	2	2	1	1	3	55	2	48	1	3	2	2	1	1	1
3	1	71	1	2	2	2	2	1	1	56	2	30	1	3	2	1	1	1	1
4	1	52	1	2	2	1	1	2	1	57	2	50	1	2	2	2	2	1	1
5	2	36	1	2	2	2	2	1	3	58	1	21	1	2	1	2	2	2	2
6	1	22	2	2	2	2	1	2	3	59	2	47	1	2	2	2	2	1	2
7	1	26	2	2	2	2	2	2	2	60	1	51	2	2	2	2	1	1	2
8	2	43	1	2	1	1	2	2	1	61	1	27	2	3	2	1	2	1	2
9	2	33	1	2	2	2	2	1	1	62	2	37	2	3	2	1	2	1	2
10	2	54	2	4	2	2	1	1	3	63	1	67	1	4	2	2	1	1	1
11	1	57	1	3	2	1	1	2	3	64	2	30	2	3	2	2	1	1	3
12	1	33	1	2	2	1	1	1	1	65	1	40	2	2	2	2	2	1	2
13	1	65	1	2	2	2	1	2	1	66	1	67	1	4	2	2	2	1	1
14	2	58	1	2	2	2	2	2	2	67	2	51	1	2	2	2	2	2	1
15	2	33	1	3	2	1	2	1	1	68	1	35	2	2	2	1	1	1	2
16	1	37	1	4	2	2	2	1	1	69	1	24	1	2	2	2	1	1	3
17	1	46	1	3	2	2	1	1	2	70	1	34	2	4	2	2	2	1	4
18	2	30	2	3	2	1	2	1	1	71	1	55	1	4	2	1	1	2	2
19	1	64	1	2	2	2	1	2	1	72	1	41	1	2	2	1	1	1	1
20	2	64	1	2	2	2	2	2	1	73	1	32	1	2	2	1	2	1	2
21	1	41	1	3	2	2	2	2	1	74	1	35	1	2	1	2	1	1	1
22	1	56	1	3	2	2	1	1	1	75	2	27	2	2	2	2	2	1	2
23	2	21	1	3	1	1	1	2	1	76	2	22	2	2	2	2	1	1	2
24	2	49	1	2	2	1	1	1	1	77	2	31	2	2	2	2	1	1	1
25	1	60	2	2	1	2	2	2	2	78	1	35	2	3	2	2	2	1	3
26	1	63	1	1	2	1	1	2	1	79	2	33	2	2	1	1	1	2	1
27	2	46	1	3	1	1	2	1	1	80	1	39	1	2	2	1	2	1	4
28	2	53	2	2	2	2	2	1	3	81	2	21	1	2	2	2	2	2	3
29	2	29	2	3	2	2	1	1	2	82	2	51	1	3	2	2	2	1	3
30	1	59	1	3	2	2	2	1	1	83	2	35	1	3	2	2	1	1	1
31	2	48	1	3	2	2	2	1	3	84	1	58	2	2	2	2	2	1	3
32	2	19	2	2	2	2	2	1	3	85	1	54	1	1	2	1	1	2	3
33	1	56	1	2	2	2	2	2	1	86	2	21	2	3	2	1	2	2	2
34	1	30	1	4	1	2	2	1	3	87	1	29	2	2	2	2	2	1	1
35	2	66	2	3	2	1	1	1	1	88	2	32	1	1	2	2	1	1	3
36	2	30	1	3	2	1	1	1	1	89	2	40	1	2	1	1	2	2	1
37	2	39	1	3	1	1	1	1	1	90	1	34	2	2	1	2	2	2	2
38	1	52	1	2	2	2	2	1	3	91	1	33	2	3	2	2	2	1	2
39	1	23	1	1	2	1	1	1	1	92	2	82	1	1	2	2	1	2	1
40	1	52	1	1	1	2	2	2	3	93	1	69	1	3	2	2	1	2	1
41	1	47	1	1	1	1	2	1	1	94	2	38	2	2	2	2	1	1	3
42	1	47	1	3	2	2	1	1	1	95	1	80	1	3	2	1	1	1	1
43	2	71	2	2	1	2	2	2	1	96	2	39	1	2	1	1	1	1	4
44	2	64	1	2	2	1	1	2	1	97	2	61	1	1	1	2	1	2	1
45	1	37	2	2	1	2	2	2	3	98	1	67	1	2	2	2	1	2	3
46	2	62	1	2	2	2	2	2	1	99	1	24	1	2	1	2	2	2	2
47	1	45	2	1	2	2	2	1	2	100	2	43	1	3	2	2	2	1	1
48	1	26	2	2	2	2	1	2	2	101	1	54	2	1	2	2	2	2	2
49	2	40	1	2	1	1	1	1	1	102	2	76	1	2	2	2	2	2	1
50	1	23	1	3	2	2	2	1	2	103	2	45	1	1	1	1	1	2	2
51	2	28	1	2	1	2	2	1	2	104	2	24	1	2	2	2	2	2	1
52	1	40	2	2	1	2	2	2	2	105	2	80	1	2	2	2	1	2	1
53	1	40	1	2	2	2	1	1	1										

Tableau 5.5 – 2. Description des libellés des 9 questions

4 questions actives **10 modalités associées**

-V3- La famille est le seul endroit où l'on se sent bien (2 modalités)
FA01 = oui, FA02 = non.

-V4- Les dépenses de logement sont pour vous une charge (4 modalités)
DL01 = négligeable, DL02 = sans gros problème,
DL03 = une lourde charge, DL04 = Une très lourde charge.

-V7- Avez-vous souffert récemment de mal au dos (2 modalités)
MA01 = oui, MA02 = non.

-V8-Vous imposez-vous régulièrement des restrictions (2 modalités)
RE01 = oui, RE02 = non.

4 questions illustratives **10 modalités associées**

-V1- Sexe de l'enquêté(e) (2 modalités)
MASC = masculin, FEMI = féminin.

-V5 Disposez-vous d'un magnétoscope (2 modalités)
MAG1 = oui, MAG2 = non.

-V6- Avez-vous souffert récemment de maux de tête (2 modalités)
MT01 = oui, MT02 = non.

-V9- Regardez-vous la télévision ? (4 modalités)
TV01 = tous les jours, TV02 = assez souvent,
TV03 = pas très souvent, TV04 = jamais.

1 variable continue illustrative

-V2- Age de l'enquêté(e) (continue)

Tableau 5.5 – 3. Tableau de Burt des $s = 4$ questions actives

	FA01	FA02	DL01	DL02	DL03	DL04	MA01	MA02	RE01	RE02
FA01	72	0								
FA02	0	33								
DL01	9	2	11	0	0	0				
DL02	37	20	0	57	0	0				
DL03	21	9	0	0	30	0				
DL04	5	2	0	0	0	7				
MA01	38	12	7	24	16	3	50	0		
MA02	34	21	4	33	14	4	0	55		
RE01	42	22	4	29	25	6	31	33	64	0
RE02	30	11	7	28	5	1	19	22	0	41
	FA01	FA02	DL01	DL02	DL03	DL04	MA01	MA02	RE01	RE02

Il s'agit là d'une propriété propre à cette méthode : les taux d'inertie sont toujours des mesures très pessimistes de l'information extraite, car le codage disjonctif induit une orthogonalité artificielle des colonnes du tableau. Plusieurs indicateurs de remplacement ont été proposés.

On peut considérer les carrés des valeurs propres, qui sont les valeurs propres de l'analyse des correspondances du tableau de Burt considéré comme tableau de données (cf. § 5.3.1) et qui fournissent des taux d'inertie un peu moins pessimistes.

Tableau 5.5 – 4. Valeurs propres et taux d'inertie

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.3416	22.77	22.77	**********
2	.3175	21.17	43.94	*********
3	.2520	16.80	60.74	*******
4	.2232	14.88	75.6	******
5	.2075	13.84	89.46	******
6	.1582	10.54	100.00	*****
Total	1.5000	100.00		

On peut également prendre en compte des fonctions particulières des valeurs propres comme mesures de l'inertie (Benzécri, 1979). On a vu que Benzécri a proposé la quantité $\rho(\lambda) = (\frac{s}{s-1})^2(\lambda - \frac{1}{s})^2$ qui est voisine de λ^2 si le nombre de questions s est grand, et qui correspond, dans le cas $s = 2$, à la valeur propre μ de l'analyse des correspondances de la table de contingence croisant les deux questions [dans ce cas, en effet, $\rho(\lambda) = \mu = (2\lambda - 1)^2$]. (voir aussi § 5.4.2).

Le tableau 5.5 - 5 fournit les indicateurs nécessaires pour interpréter les positions des modalités actives.

Les règles de lecture sont semblables à celles du tableau 4.4 - 3 relatif à l'analyse des correspondances simple. Seuls les calculs de contributions cumulées pour les modalités de chaque question ont été ajoutés. Leur interprétation est immédiate. Il est clair, par exemple, que les deux questions relatives aux dépenses de logement et aux restrictions définissent entièrement le premier axe.

Le tableau 5.5 - 6 donne les valeurs-test et les coordonnées des modalités supplémentaires sur les trois premiers axes.

On note que les seules coordonnées significatives sur le premier axe sont relatives à la possession d'un magnétoscope (valeurs-test de 2.8). Les mentions de maux de têtes et l'écoute de la télévision - toutes deux liées à l'âge - sont caractéristiques du deuxième axe.

Tableau 5.5 – 5. Coordonnées, contributions et cosinus carrés des modalités actives sur les axes 1 à 3

MODALITES		COORDONNEES			CONTRIBUTIONS			COSINUS CARRES		
IDEN - LIBELLE	P.REL DISTO	1	2	3	1	2	3	1	2	3
- la famille est le seul endroit ou l'on se sent bien										
FA01 - - oui -	17.14 .46	.14	-.42	.12	1.0	9.3	.9	.05	.38	.03
FA02 - - non -	7.86 2.18	-.31	.91	-.26	2.3	20.4	2.1	.05	.38	.03
				CUMUL =	3.3	29.7	3.0			
- les dépenses de logement sont pour vous une charge										
DL01 - négligeable	2.62 8.55	1.32	-1.32	.33	13.4	14.4	1.2	.20	.20	.01
DL02 - sans gros problème	13.57 .84	.41	.52	-.11	6.7	11.8	.6	.20	.33	.01
DL03 - une lourde charge	7.14 2.50	-1.00	-.50	-.72	21.1	5.7	14.8	.40	.10	.21
DL04 - très lourde charge	1.67 14.00	-1.11	-.05	3.45	6.0	.0	78.7	.09	.00	.85
				CUMUL =	47.2	31.9	95.2			
- avez-vous souffert récemment de mal au dos										
MA01 - - oui -	11.90 1.10	.03	-.73	-.14	.0	19.8	.9	.00	.48	.02
MA02 - - non -	13.10 .91	-.02	.66	.13	.0	18.0	.8	.00	.48	.02
				CUMUL =	.0	37.9	1.86			
- vous imposez-vous régulièrement des restrictions										
RE01 - - oui -	15.24 .64	-.66	-.06	.01	19.3	.2	.0	.68	.01	.00
RE02 - - non -	9.76 1.56	1.03	.10	-.01	30.2	.3	.0	.68	.01	.00
				CUMUL =	49.5	.5	.0			

Tableau 5.5 – 6. Coordonnées et valeurs-test des modalités illustratives sur les axes 1 à 3

MODALITES			VALEURS-TEST			COORDONNEES			
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	1	2	3	DISTO.
-sexe de l'enquêté(e)									
MASC - masculin	53	53.00	.5	.4	2.1	.05	.04	.21	.98
FEMI - féminin	52	52.00	-.5	-.4	-2.1	-.05	-.04	-.21	1.02
- disposez-vous d'un magnétoscope									
MAG1 - oui -	22	22.00	2.8	.7	.5	.54	.13	.09	3.77
MAG2 - non -	83	83.00	-2.8	-.7	-.5	-.14	-.03	-.02	.27
- avez-vous souffert récemment de maux de tête									
MT01 - oui -	33	33.00	.0	-3.1	-1.3	.01	-.45	-.19	2.18
MT02 - non -	72	72.00	.0	3.1	1.3	.00	.21	.09	.46
- regardez-vous la télévision ?									
TV01 - tous les jours	53	53.00	.7	-3.4	-.2	.07	-.33	-.02	.98
TV02 - assez souvent	27	27.00	.1	3.3	-.9	.02	.56	-.16	2.89
TV03 - pas très souvent	22	22.00	-.6	.3	.4	-.11	.07	.08	3.77
TV04 - jamais	3	3.00	-1.0	.7	1.9	-.56	.39	1.11	34.00

Tableau 55.7. Coordonnées (corrélations) de la variable continue illustrative sur les axes 1 à 3.

VARIABLE CONTINUE		CARACTERISTIQUES			CORRELATIONS		
(IDEN) LIBELLE COURT	EFFECTIF	MOYENNE	EC.TYPE	1	2	3
-(age) âge de l'enquêté(e)	105	43.89	15.50	.23	-.23	.15

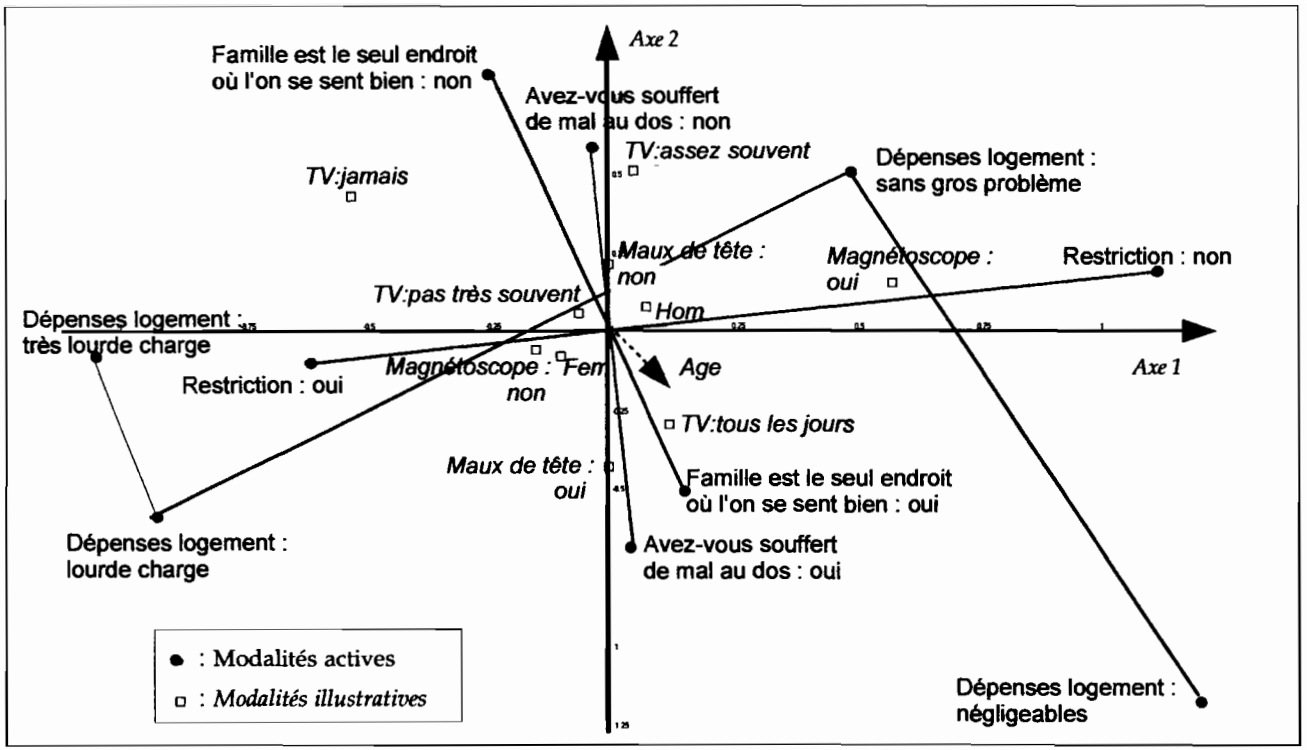


Figure 5.5 – 1. Position des modalités actives et illustratives sur le premier plan factoriel

Les modalités "consécutives" des questions actives sont jointes par des lignes polygonales. On vérifie que l'origine est bien un centre de gravité pour les modalités de chaque question, ce qui implique un alignement avec l'origine pour les questions à 2 modalités. Les variables "*restrictions*" (*ne s'impose pas de restriction*) et "*dépenses de logement*" (*négligeables, sans trop de problème*) déterminent le premier axe, illustré a posteriori par la position du point "*possession d'un magnétoscope*". La variable continue "*âge*" est repérée par ses coefficients de corrélation avec les axes (flèche en pointillé).

Le tableau 5.5 - 7 est relatif à la variable continue "âge". On y lit sa moyenne, son écart-type, et ses coefficients de corrélation avec les trois premiers axes.

La structure du nuage des modalités actives est décrite par le plan factoriel de la figure 5.5 - 1, qui résume donc les *6* tables de contingence.

Le petit nombre de questions et le faible nombre d'individus limitent l'intérêt des résultats, mais permettent en revanche de comprendre le mécanisme de la méthode.

Les deux questions les plus liées (*dépenses de logements* et *restrictions*) caractérisent le premier axe, la question relative aux dépenses de logement intervenant avec un poids double compte tenu du nombre de ses modalités (cf. § 5.2.2 - a). Les deux questions restantes, plus faiblement liées, caractérisant le deuxième axe.

La représentation simultanée des lignes et des colonnes liée à l'analyse des correspondances n'est pas utilisée sur la figure 5.5 - 1. Les *105* points-lignes correspondent à des individus anonymes ; seules leurs caractéristiques présentent de l'intérêt. Les individus n'interviennent donc que par le truchement des variables supplémentaires qui les caractérisent.

Les positions des modalités supplémentaires doivent être interprétées à partir de leurs valeur-tests.

Dans les études en vraie grandeur où ces modalités peuvent être très nombreuses, seules celles ayant des valeurs-test significatives sont portées sur les graphiques.

Ainsi, la variable *sexe* (valeurs-test *0.5* et *0.4* sur les axes *1* et *2*) pourrait ne pas figurer dans ce plan factoriel. De même, la modalité TV04, (*ne regarde jamais la télévision)* malgré sa position relativement excentrée à gauche, n'est pas non plus significative (valeur-test = *-1.0*) car elle ne concerne que *3* individus.

Remarquons que la seule phase du processus permettant de procéder à une inférence statistique est précisément le calcul des valeurs-test relatives aux modalités supplémentaires.

Malgré la taille modeste de l'échantillon et le petit nombre de variables, on peut rejeter l'hypothèse d'indépendance entre la possession d'un magnétoscope (point MAG2) et l'aisance financière telle qu'elle est décrite par les modalités (DL01, DL02) relatives aux dépenses de logement, et RE02 : pas de restriction.

La variable continue AGE est représentée comme un axe, en pointillé. Cette direction a une certaine cohérence, malgré la faible taille de l'échantillon (les individus plus âgés ont des idées plus traditionalistes sur la famille, sont plus souvent propriétaires de leur logements, plus fréquemment téléspectateurs).

5.5.3 Eléments de validation

Nous envisageons ici, comme ce fut le cas au chapitre 3 pour le deuxième exemple d'application de l'analyse en composantes principales, différents types de bootstrap pour valider la position des variables dans les sous-espaces factoriels.

a – Bootstrap partiel pour les variables actives

Avec ce type de bootstrap, le plan initial sert d'espace de référence pour accueillir les réplications, qui sont projetées comme des variables supplémentaires. Le bootstrap partiel n'a pas pour vocation de valider la stabilité de l'espace de départ qui n'est pas remis en question. Il donne une idée de la variablilité imputable aux réplications pour chaque point-modalité pris isolément.

La figure 5.5-2 nous montre que dans ce plan, les réponses « lourdes charges » et « très lourdes charges » pour les dépenses de logement sont indiscernables. La plupart des autres modalités sont relativement bien séparées. Conclusion : même si le pattern général, c'est-à-dire l'agencement des points-modalités, n'est pas quelque chose de stable, la variabilité due aux réplications reste limitée dans ce sous-espace particulier.

b – Bootstrap partiel pour les variables supplémentaires

Pour les variables supplémentaires, le bootstrap ne peut être que partiel. Il s'agit d'une validation externe, et donc d'un test statistique parfaitement légitime, ces variables n'ayant pas participé à la construction du sous-espace de référence, qui est le même que celui de la figure 5.5-2.

Sur la figure 5.5-3, on n'a projeté qu'une question supplémentaire : « Regardez-vous la télévision ? » représentée par ses quatre modalités de réponse : « tous les jours », « assez souvent », pas très souvent », « jamais ».

Le résultat trouvé est tout-à-fait cohérent avec les valeurs-tests de ces modalités qui figurent dans le tableau 5.5-6. On ne note rien de significatif le long du premier axe horizontal (les valeurs-tests des quatre modalités étaient toutes très inférieures à 2 en valeur absolue sur cet axe).

La modalité « jamais » (trois répondants, et, bien sûr, des valeurs-tests faibles) a une position indéterminée. La modalité « pas très souvent », bien que concernant 27 personnes, semble répartie aléatoirement autour de l'origine des axes dans ce plan (les valeurs-tests sur les axes étaient de *-0.6* et de *0.3*). Seuls les modalités « tous les jours » et « assez souvent » donnent lieu à des zones de confiance nettement disjointe le long du second axe. Le long de cet axe, les valeurs-tests s'opposaient de façon significative (*-3.4* et *3.3*).

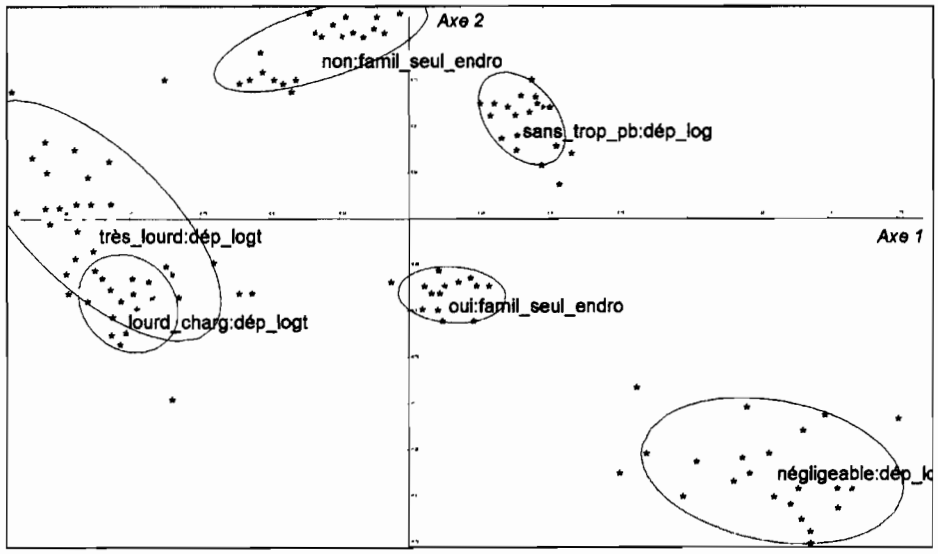


Figure 5.5 – 2. Bootstrap partiel – variables actives
Les réplications sont considérées comme des variables supplémentaires

Les valeurs-tests sont donc des outils de validation précieux, mais qui ne caractérisent que les axes un par un. C'est parce que l'opposition significative de cet exemple ne se fait que le long d'un axe (l'axe vertical) que les résultats des zones de confiance bootstrap semblent redondants avec ceux donnés par les valeurs-tests. L'excentricité des ellipses et la position des zones bootstrap peuvent être quelconques dans le plan.

c – Bootstrap total pour les variables actives

Rappelons que dans ce cas, chaque réplication donne lieu à une analyse des correspondances multiples séparée. On a noté au paragraphe 5.4.2.b ci-dessus que cette opération était périlleuse pour des petits échantillons comportant des effectifs de modalité faibles, car ceux-ci peuvent être encore plus faibles, voire nuls, lors de certains tirages avec remise, c'est-à-dire pour certaines réplications.

Pour des questions de place et de lisibilité, nous ne publierons ici que la validation utilisant le bootstrap total de type 3, c'est-à-dire après rotations procrustéennes des réplications. Une seule question active : « Les dépenses de logements sont pour vous... » ayant quatre modalités (« négligeables », « sans trop de problèmes », « une lourde charge », « une très lourde charge ») sera représentée.

Les bootstrap de type 1 (simples corrections du signes des axes) et le bootstrap de type 2 (corrections des interversions d'axes) donnent pour notre exemple et dans le plan (1, 2) des zones de confiance qui empiètent et se superposent assez largement, traduisant ainsi le manque de robustesse de l'analyse des correspondances multiples dans le contexte simplifié de cet exemple d'école.

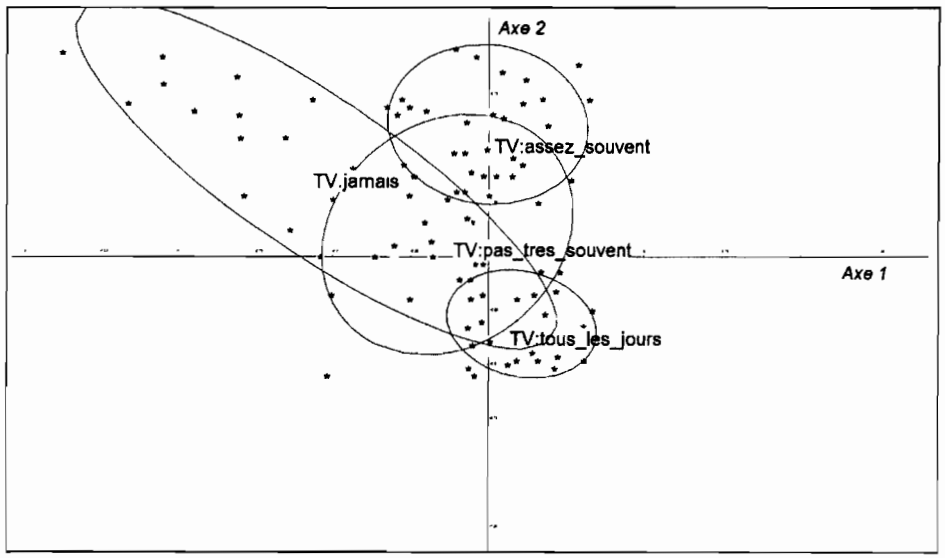


Figure 5.5 – 3. Bootstrap partiel - variables supplémentaires

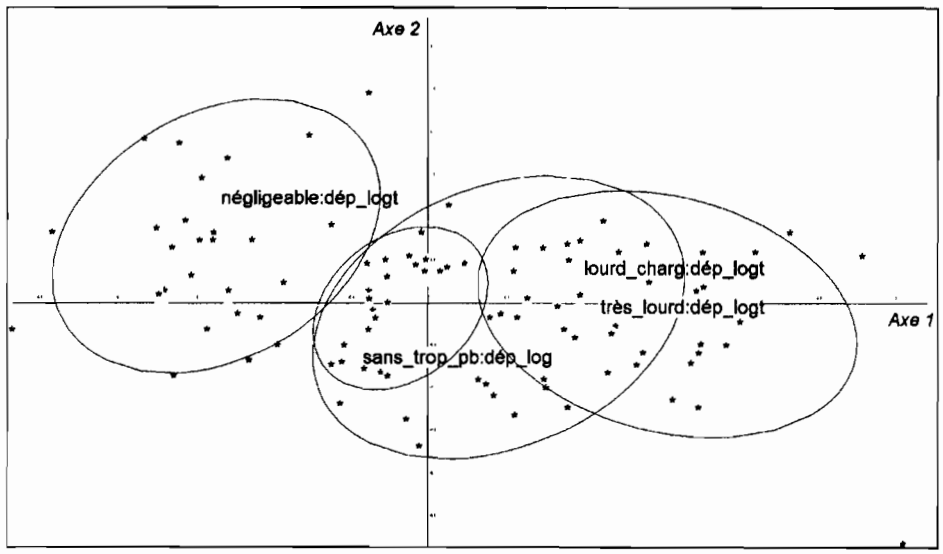


Figure 5.5 – 4. Bootstrap total (type 3) - variables actives
Quatre modalités de la question sur les dépenses de logement.

La figure 5.5-4 nous montre que le premier axe de l'analyse initiale traduit une opposition stable entre les modalités « négligeables » et les autres modalités, et parmi celles-ci, plus particulièrement la modalité « très lourdes charges ». Cette opposition n'est pas forcément observée sur le premier axe de l'analyse de

chaque réplication (puisque celles-ci sont soumises ensuite à des rotations procrustéennes), mais il existe, pour chaque réplication, une dimension pour laquelle ces modalités s'opposent.

En revanche, le second axe n'a pas pu être validé en tant que dimension stable par cette méthode : les zones de confiance ne se séparent pas. L'épreuve de validation externe (lien significatif avec une question supplémentaire : « le temps passé devant la télévision ») désigne cependant cette seconde dimension comme intéressante. Il n'y a pas de contradiction : la stabilité d'une dimension est une hypothèse beaucoup plus forte que l'existence d'un lien avec des variables externes. Une dimension peut ne dépendre que d'une variable, et de ce fait être liée à des variables externes. Mais elle peut disparaître dans les analyses de certaines réplications.

En conclusion, le fait de développer un exemple sur un petit échantillon ($n = 105$) nous a situé dans un domaine où l'analyse des correspondances multiples est fragile (peu stable), ce qui nous a permis de voir sous un verre grossissant les subtilités et les difficultés des procédures de validation.

5.6 Modèles log-linéaires et analyse des correspondances multiples

L'analyse des tables de contingences multiples peut également être effectuée par les modèles log-linéaires. Ceux-ci permettent alors d'étudier et de modéliser les liaisons entre plusieurs variables nominales en tenant compte de leurs éventuelles interactions. Ils relèvent d'une analyse exploratoire (non supervisée) car aucune variable ne joue le rôle privilégié de variable à prévoir. Mais ces modèles s'apparentent aussi, par leur démarche, à l'analyse de la variance (sélection de modèles sur la base de tests statistiques) et également à la régression logistique (chapitre 7).

Les modèles log-linéaires et logistiques donnent lieu à des publications nombreuses. Après les premiers travaux de Birch (1963) et Goodman (1970), il faut mentionner les ouvrages de base de Haberman (1974), Bishop *et al.* (1975), Fienberg (1980).

Après ces travaux de pionniers, Dobson (1983), Agresti (1990), Christensen (1990) rédigent des synthèses enrichies de contributions personnelles.

Goodman (1986, 1991) fait des rapprochements avec certains aspects de l'analyse des correspondances. On consultera également l'ouvrage collectif plus récent édité par Droesbeke *et al.* (2005).

5.6.1 Formulation du problème et principes de base

Présentons le problème à partir d'un exemple médical. Considérons un échantillon d'individus ayant été irradiés accidentellement. Ces individus sont caractérisés par un état (être *décédés ou non* à la suite de leucémie : variable nominale à 2 modalités), par la dose de radiations reçue mesurée en *Rad* (variable continue, ordonnée ensuite en 6 modalités) et par l'âge au moment des accidents (variable continue regroupée en 5 modalités).

Ces données se présentent sous forme d'un tableau de contingence **K** croisant ces trois variables de terme général k_{ijl}. On s'intéresse aux relations existant entre ces trois variables : sont-elles indépendantes ou non dans leur ensemble ou une variable est-elle indépendante conditionnellement à une ou aux deux autres ? Autrement dit, on cherche à connaître la structure des liaisons entre ces données en tenant compte des interactions éventuelle entre les 3 variables.

D'une manière générale, p variables nominales $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_p$ ayant respectivement $m_1, m_2, \ldots, m_p$ modalités, constituent un tableau de contingence multidimensionnel à p entrées comprenant $m_1 \times m_2 \times \ldots \times m_p$ cases. Le terme général $k_{ij\ldots p}$ de cet hypercube de contingence indique le nombre d'individus ayant répondu simultanément aux modalités $i,j,\ldots,p$ de $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ avec $1 < i < m_1, 1 < j < m_2, \ldots, 1 < p < m_p$.

L'effectif total d'individus observés est noté k avec :

$$k = \sum_{i,j,\ldots,p} k_{ij\ldots p}$$

Les hypothèses que nous formulons sur les liaisons entre ces p variables nous amènent à construire des tableaux de fréquences théoriques espérées **T** de terme général $t_{ij\ldots p}$. La confrontation des fréquences observées $k_{ij\ldots p}$ et des fréquences théoriques $t_{ij\ldots p}$ va permettre de tester ces hypothèses. On construira par conséquent autant de tableaux **T** (et donc de modèles log-linéaires) qu'il y a d'hypothèses à tester. Dans le cas d'un tableau de contingence à deux dimensions, on construit, sous l'hypothèse d'indépendance entre les deux variables, le tableau **T** tel que $t_{ij} = t_{i.}\, t_{.j}$. Le test du χ^2 permet de rejeter ou non cette hypothèse en confrontant le tableau théorique **T** au tableau des fréquences observées **K**. Ainsi les modèles log-linéaires peuvent être considérés comme une généralisation du test du χ^2 à un ensemble de p variables nominales ($p > 2$), la difficulté résidant alors dans le choix des modèles, c'est-à-dire des hypothèses concernant les liaisons entre les variables.

5.6.2 Ajustement d'un modèle log-linéaire

On suppose que la fréquence observée $k_{ij\ldots p}$ est la réalisation d'une variable aléatoire $x_{ij\ldots p}$ d'espérance mathématique inconnue $t_{ij\ldots p}$.

$$E(x_{ij\ldots p}) = t_{ij\ldots p}$$

Nous envisagerons successivement le cas du tableau de contingence à deux dimensions et celui à p entrées. Les notations étant lourdes dans le cas général, nous nous bornerons à $p = 3$ pour simplifier l'exposé.

a – Tableau de contingence à deux entrées

Intéressons-nous d'abord à la relation entre deux variables nominales, le *risque de décès* et la *dose de radiation reçue*, par exemple. Dans ce cas, deux hypothèses peuvent être formulées : y a-t-il indépendance ou non entre les deux variables ?

En supposant t_{ij} non nul, le modèle log-linéaire le plus complet décompose le logarithme népérien de l'espérance t_{ij} sous la forme :

$$log(t_{ij}) = \alpha_0 + \alpha_1(i) + \alpha_2(j) + \alpha_{12}(ij)$$

Par analogie avec l'analyse de la variance, $log(t_{ij})$ se décompose en une somme de coefficients α décrivant plusieurs effets :

- α_0, l'effet global;
- $\alpha_1(i)$, l'effet dû à la variable x_1,
- $\alpha_2(j)$, l'effet dû à la variable x_2,
- $\alpha_{12}(ij)$, l'effet dû à l'interaction entre les variables x_1 et x_2.

Afin d'avoir une solution unique, on impose les contraintes suivantes :

$$\sum_i \alpha_1(i) = \sum_j \alpha_2(j) = \sum_i \alpha_{12}(ij) = \sum_j \alpha_{12}(ij) = 0$$

Sous l'hypothèse d'indépendance des deux variables, la fréquence espérée s'exprime par $t_{ij} = t_{i.} t_{.j}$. Dans ce cas, tous les coefficients d'interaction $\alpha_{12}(ij)$ sont nuls. Le modèle log-linéaire correspondant à cette hypothèse s'écrit :

$$log(t_{ij}) = \alpha_0 + \alpha_1(i) + \alpha_2(j)$$

La nullité des interactions traduit l'hypothèse d'indépendance entre les deux variables. A partir des coefficients α_0, $\alpha_1(i)$ et $\alpha_2(j)$, on calcule le tableau des fréquences théoriques espérées noté **T**.

b – Tableau de contingence à p entrées

On généralise ces modèles au cas de plus de deux variables. Pour trois variables par exemple, le modèle qui prend en compte toutes les liaisons entre les variables est le suivant :

$$\begin{aligned} log(t_{ijl}) = {} & \alpha_0 + \alpha_1(i) + \alpha_2(j) + \alpha_3(l) \\ & + \alpha_{12}(ij) + \alpha_{13}(il) + \alpha_{23}(jl) + \alpha_{123}(ijl) \end{aligned} \qquad [5.6 - 1]$$

Ce modèle est appelé *modèle saturé*. Il contient tous les effets et toutes les interactions qu'il est possible de définir avec les variables disponibles.

Les coefficients α_0, $\alpha_1(i)$,..., $\alpha_{123}(ijl)$ traduisent des effets différents :

- α_0, l'effet global;
- $\alpha_1(i)$, $\alpha_2(j)$, $\alpha_3(l)$, les effets principaux;
- $\alpha_{12}(ij)$, $\alpha_{13}(ik)$, $\alpha_{23}(jl)$, les effets dus aux interactions deux à deux des variables;
- $\alpha_{123}(ijl)$, l'effet dû à l'interaction à trois variables;

On impose la nullité de la somme des coefficients du modèle faisant intervenir une modalité d'une variable (somme calculée sur l'ensemble des modalités de cette même variable).

Par exemple, pour la variable x_1 et pour tout $1 \leq i \leq m_1$, on a :

$$\sum_i a_1(i) = \sum_i a_{12}(ij) = \sum_i a_{13}(il) = \sum_i a_{123}(ijl) = 0$$

et il en est de même pour les autres variables.

Le modèle [5.6 - 1], comme tous les modèles saturés, permet de reconstituer exactement le tableau de fréquence **K**. Celui-ci présentant souvent un trop grand nombre de coefficients, on va rechercher un ou des modèles ayant moins de coefficients mais devant reconstituer le mieux possible le tableau **K** (principe de parcimonie). Ceci est réalisé en annulant certains termes du modèle saturé.

Si on arrive à une reconstitution correcte du tableau **K**, l'hypothèse de nullité des coefficients supprimés ne peut pas être rejetée. Ces modèles non saturés mettent alors en évidence les liaisons les plus significatives entre les variables.

Dans le cas de deux variables, l'hypothèse de nullité du terme d'interaction s'interprète en terme d'indépendance. Si cette hypothèse est rejetée, on incriminera une dépendance entre les deux variables. Lorsque l'on s'intéresse à plus de deux variables, l'interprétation est plus complexe :

- pour exprimer *l'indépendance mutuelle* entre toutes les variables x_1, x_2, x_3, on annule tous les termes d'interactions. Cela nous conduit au modèle :

$$log(t_{ijl}) = \alpha_0 + \alpha_1(i) + \alpha_2(j) + \alpha_3(l)$$

- pour exprimer *l'indépendance conditionnelle* de deux variables x_1 et x_2 par rapport à x_3, on annule tous les termes d'interaction contenant les indices relatifs aux variables x_1 et x_2 c'est-à-dire :

$$\alpha_{12}(ij) = \alpha_{123}(ijl) = 0$$

on en déduit le modèle suivant :

$$log(t_{ijl}) = \alpha_0 + \alpha_1(i) + \alpha_2(j) + \alpha_3(l) + \alpha_{13}(il) + \alpha_{23}(jl)$$

Chaque modèle log-linéaire met ainsi en évidence une liaison particulière entre les variables : la dépendance ou l'indépendance mutuelle des variables dans leur ensemble ou l'indépendance de certaines variables conditionnellement à une ou plusieurs autres. Pour des modèles à plus de trois variables, on trouvera des compléments sur les interactions, dans par exemple, Agresti (1990).

c – modèles hiérarchiques

Un modèle log-linéaire est dit hiérarchique si la condition suivante est vérifiée : quand un coefficient d'interaction est présent dans le modèle, les coefficients des variables mises en jeu et toutes les interactions d'ordre inférieur sont aussi dans le modèle.

Par exemple, si dans un modèle à 5 variables on trouve l'interaction x_{135}, alors le modèle, pour être hiérarchique, doit contenir au moins x_1, x_3 et x_5 ainsi que les interactions d'ordre inférieur x_{13}, x_{15} et x_{35}.

Parmi les modèles log-linéaires possibles dans le cas d'un tableau de contingence à deux variables, certains modèles sont hiérarchiques :

- $log(t_{ij}) = \alpha_0 + \alpha_1(i) + \alpha_2(j) + \alpha_{12}(ij)$
- $log(t_{ij}) = \alpha_0 + \alpha_1(i) + \alpha_2(j)$

et d'autres ne le sont pas :

- $log(t_{ij}) = \alpha_0 + \alpha_1(i) + \alpha_{12}(ij)$;
- $log(t_{ij}) = \alpha_0 + \alpha_2(j) + \alpha_{12}(ij)$;
- $log(t_{ij}) = \alpha_0 + \alpha_{12}(ij)$

Traditionnellement et pour des raisons de simplicité d'interprétation, on se limite aux modèles hiérarchiques.

5.6.3 Estimation et tests d'ajustement du modèle

On se donne un modèle traduisant une hypothèse exprimée par la nullité de certains coefficients α.

On cherche ainsi à estimer les fréquences théoriques pour construire puis confronter le tableau $\hat{\mathbf{T}}$ des estimations au tableau **K** des fréquences observées.

Cette confrontation est réalisée par des tests d'ajustement. Ils permettent de rejeter ou non l'hypothèse sur les liaisons exprimée par le modèle.

Cette démarche est proche de celle de la régression logistique ce qui conduit parfois à présenter les modèles log-linéaires parmi les techniques statistiques explicatives.

a – Estimation des paramètres

Les fréquences théoriques espérées t_{ijl} sont en général estimées par la méthode du maximum de vraisemblance. Elle consiste à rechercher les paramètres qui maximisent la fonction de vraisemblance $\mathcal{L}(k_{ijl}, t_{ijl})$.

Pour cela, on suppose que les variables aléatoires x_{ijl} suivent soit une loi de Poisson, soit une loi multinomiale[1].

On montre alors (cf. par exemple Haberman, 1974) que maximiser $\mathcal{L}(k_{ijl}, t_{ijl})$ revient à maximiser :

$$\sum_{i,j,l} k_{ijl} \log(t_{ijl})$$

On calcule les estimations $\hat{t}_{ijl}$ des fréquences espérées t_{ijl} données par le modèle. On peut utiliser la méthode de régression pondérée de Grizzle et *al.* (1969) ou celle des algorithmes itératifs (méthode de Newton-Raphson ou méthode des moindres carrés itératifs) qui est la méthode la plus répandue, utilisée pour tous les modèles linéaires généralisés, dont les modèles log-linéaires sont des cas particuliers[2].

b – Tests d'ajustement

Pour comparer le tableau des fréquences estimées $\hat{\mathbf{T}}$ avec le tableau des fréquences observées **K**, deux tests (voisins) sont généralement utilisés :

- le test du χ^2 de Karl Pearson :

$$\chi^2 = \sum_{i,j,l} \frac{(k_{ijl} - \hat{t}_{ijl})^2}{\hat{t}_{ijl}}$$

- le test du rapport de vraisemblance[3] :

$$G^2 = -2 \sum_{i,j,l} k_{ijl} \log \frac{\hat{t}_{ijl}}{k_{ijl}}$$

Les statistiques χ^2 et G^2 suivent une distribution du χ^2 à m degrés de liberté où m est le nombre de cases du tableau auquel on soustrait le nombre de coefficients estimés.

Pour l'une et l'autre de ces statistiques, les valeurs augmentent avec le nombre de variables introduites dans le modèle.

---

[1] Ce sont des hypothèses assez naturelles dans le cas des tables de contingence multidimensionnelles. Brièvement dit, la loi de Poisson correspond au cas où l'effectif total k n'est pas fixé ou borné a priori.

[2] Cf. Haberman (1974), Nelder et Wedderburn, (1972), McCullagh et Nelder (1989), Christensen, (1990).

[3] G^2 est aussi une mesure de proximité entre les distributions de fréquence $\hat{\mathbf{T}}$ et **K** selon la théorie de l'information développée en particulier par Kullback et Leibler (1951), Kullback (1959). En fait la première formule (χ^2) correspond au premier terme non nul du développement limité de G^2, en écrivant : $G^2 = -2 \sum_{i,j,l} k_{ijl} \log\left(1 + \frac{\hat{t}_{ijl} - k_{ijl}}{k_{ijl}}\right)$.

Plus ces statistiques sont voisines de zéro, meilleur est l'ajustement. Elles sont nulles pour le modèle saturé. On recherche le modèle le plus simple (peu de paramètres) et qui reste acceptable (bon ajustement).

c – Choix du modèle

Le choix du modèle log-linéaire est d'autant plus difficile que le nombre de variables est élevé. La méthode dite "combinatoire" est une des méthodes possibles pour obtenir un "bon" modèle.

A partir du modèle saturé, on construit des modèles plus simples en retirant un à un les termes d'interaction.

La statistique G^2 croît progressivement et l'on peut arrêter la procédure lorsqu'elle augmente plus rapidement (cf. figure 5.6.1). On retiendra alors le modèle correspondant et l'on en déduira les liaisons importantes entre les variables[1].

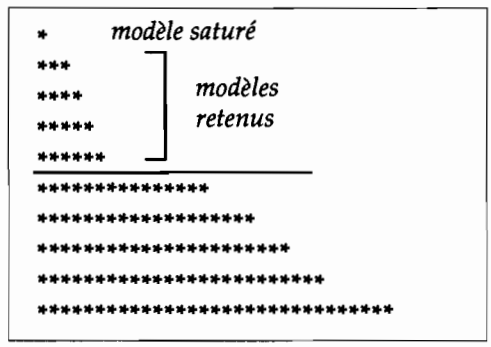


Figure 5.6 – 1. Histogramme de G^2 et recherche du palier de croissance

Cette méthode combinatoire est applicable aux modèles mettant en jeu un petit nombre de variables. Pour fixer les idées, avec 4 variables, il y a *167* modèles hiérarchiques possibles.

Il existe un nombre considérable de travaux sur ce problème de sélection de modèles (problème qui se pose également dans le cas de la régression, mais de façon moins complexe).

La multiplication des tests pose des problèmes de *comparaisons multiples* spécifiques (Gabriel, 1969; Aitkin, 1979).

On peut restreindre la recherche aux *modèles graphiques* (sous-ensemble des modèles hiérarchiques) et à l'intérieur de ceux-ci aux *modèles décomposables.*

---

[1] On note que l'estimation du critère d'Akaike (1973), fonction de la statistique G^2, est souvent utilisé pour sélectionner un modèle et mesurer sa qualité. Elle offre l'avantage d'être obtenue sans étudier l'ensemble des modèles possibles. Ce critère équivaut asymptotiquement à la validation croisée (Stone, 1977).

Whittaker (1990) fait une présentation générale des modèles graphiques et une revue des problèmes de sélection des modèles log-linéaires graphiques[1].

5.6.4 lien avec l'analyse des correspondances

Le modèle log-linéaire et l'analyse des correspondances multiples ne répondent pas aux mêmes préoccupations et ne fournissent pas des résultats de même nature. Ce sont en fait des techniques complémentaires.

D'assez nombreux travaux ont porté sur la comparaison des différentes approches dans des contextes d'application divers, parfois sensiblement éloignés des contextes réels.

On ne mentionnera ici qu'un petit nombre de publications sur ce thème en suivant un ordre chronologique : Daudin et Trécourt (1980) sont parmi les premiers à faire une comparaison sur une table de contingence à 6 entrées ($21 \times 2 \times 2 \times 2 \times 2 \times 2$) entre une des analyses des correspondances possibles et le modèle log-linéaire.

Escoufier (1982), Lauro et Decarli (1982) proposent également des rapprochements entre utilisations des méthodes. Leclerc et *al.* (1985) comparent sur un exemple approfondi l'analyse des correspondances et la régression logistique.

Van der Heijden et de Leeuw (1985), Van der Heijden (1987), puis Van der Heijden et al. (1989) proposent une méthodologie de l'utilisation simultanée de l'analyse des correspondances et des modèles log-linéaires en préconisant de décrire par des analyses des correspondances les résidus des modèles log-linéaires.

D'autres comparaisons et applications se trouvent dans Worsley (1987) et plus généralement dans le numéro spécial 35 -3 (1987) de la Revue de Statistique Appliquée, animé par le L.S.P. de l'Université Paul Sabatier. Cf. également Hudon (1990), Tenenhaus et al. (1993).

Gilula (1986), Gilula et Ritov (1990), Goodman (1986, 1991) étudient les performances de l'analyse des correspondances et des modèles log-linéaires dans le contexte d'utilisation des modèles qu'ils ont eux-mêmes développés pour les tables de contingences multiples ou à modalités ordonnées (approche confirmatoire pour des tables de dimensions très réduites).

---

[1] Les références de base sur les modèles graphiques sont Wermuth (1976), et Darroch *et al.* (1980). Pour une synthèse récente, voir Wermuth et Cox (1992). On pourra consulter Fine (in : Droesbeke *et al.*, 1992), de Falguerolles et Jmel (1993).

a – Des champs d'application différents

Bien que s'appliquant aux mêmes types de variables, les variables nominales, ces deux méthodes ont des problématiques et des champs d'application différents.

► Le *modèle log-linéaire* s'applique avec profit lorsque l'on dispose de peu de variables (rarement plus de cinq variables surtout si elles ont beaucoup de modalités) avec cependant beaucoup d'individus, pour que les cellules de l'hypertable de contingence obtenue en croisant les variables ne soient pas vides. Le nombre des sous-modèles explicitant les liaisons entre les variables augmente beaucoup plus vite que le nombre de variables. On augmente alors le nombre de coefficients à tester et donc les chances de trouver des effectifs nuls, ce qui rend les résultats plus instables. De ce fait, le modèle log-linéaire est bien adapté lorsque le problème posé permet de procéder à une sélection préalable des variables et de formuler les hypothèses nulles.

► L'*analyse des correspondances binaires* (sur vraies tables de contingence, que l'on appelle parfois tables de contingence binaire ou à double entrées) s'applique avec profit lorsque les deux partitions mises en correspondances (colonnes et lignes actives) sont relativement importantes : par exemple, tables de contingence croisant *95* départements métropolitains et *12* causes de décès, tables croisant *373* communes de la région parisienne et *29* catégories socio-professionnelles. Pour des petites tables de contingence, la fonction de l'analyse des correspondances est surtout iconographique, illustrative.

► L'*analyse des correspondances multiples* (sur tableaux disjonctifs complets) est utile pour les tableaux de type "sous-fichiers d'enquête" : en général une à plusieurs dizaines de variables nominales, de *200* à *20 000* individus. Il n'est pas rare que l'hypertable de contingence soit à *99%* vide[1].

Qu'il s'agisse de correspondances binaires ou multiples, la dichotomie entre *variables actives et illustratives* est fondamentale. C'est elle qui permet de confronter une information illimitée au sous-espace des variables actives, dont la description ne constitue qu'une phase préliminaire.

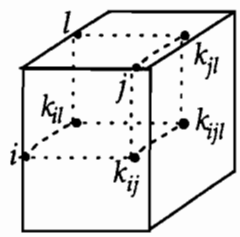
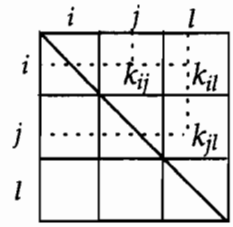
Enfin, ces méthodes ne permettent que de décrire des tableaux. Et une table de contingence multiple permet de construire différents types de tableaux.

Si l'on s'intéresse aux interactions d'ordre élevé entre certaines variables, on construira de nouvelles variables en croisant ces variables et en considérant selon les cas la nouvelle variable comme active ou supplémentaire.

---

[1] Ainsi, pour une petite batterie de *10* questions à *4* modalités posées à *1000* répondants, l'hypertable présente 4^{10} ($\approx 10^9$) cases; moins d'une case sur *1000* sera non-vide.

Tableau 5.6 – 1. Vocations spécifiques des deux approches

Modèle log-linéaire	Correspondances multiples
- Description des interactions entre plus de deux variables dans un cadre inférentiel.	- Description des liaisons entre les variables prises deux à deux sous forme essentiellement graphique.
- Des hypothèses sur les liaisons doivent être formulées au préalable.	- N'impose aucune hypothèse sur les liaisons, mais impose une certaine homogénéité de l'ensembles des variables actives.
- Est limité à peu de variables (en pratique moins de 5) .	- N'est pas limitée dans le nombre de variables
- Met en jeu toutes les cases d'un hypercube de contingence : 	- Met seulement en jeu les faces de l'hypercube représentées par le tableau de Burt : 
- Méthode par essence confirmatoire, utilisée pour explorer l'univers des modèles. On cherche celui ou ceux qui s'adaptent le mieux aux observations.	- Méthode descriptive et exploratoire de la structure intrinsèque des données.
- Les individus n'apparaissent pas.	- Les individus peuvent jouer un rôle central. L'analyse sert souvent à produire des typologies d'individus.
- La notion de variable supplémentaire n'est pas directement pertinente.	- La notion de variable supplémentaire est fondamentale.

C'est le problème sous-jacent qui permet de guider la démarche : choix des tableaux à décrire dans un cas, choix des modèles à sélectionner et à éprouver dans l'autre. Rappelons également que l'usage simultané de la classification et des analyses en axes principaux fait partie intégrante de la démarche exploratoire.

Le tableau 5.6 - 1 résume ces différences d'objectifs et d'applications dans le cas de l'analyse des correspondances multiples.

Certains travaux de confrontation entre méthodes perdent de leur portée en raison de la méconnaissance des vocations (essentiellement attestée par une expérience pratique) de chacune des approches.

Il est vrai que le *paradoxe pédagogique* inhérent à l'analyse des données - *comment prouver sur un modèle réduit l'efficacité de méthodes qui ne sont utiles et profitables que sur de grands tableaux* - ne facilite pas la tâche d'explication de la vocation réelle de ces méthodes.

Il faut reconnaître cependant que si l'analyse des correspondances est bien utile dans le cas des grandes tables de contingences à deux entrées et dans le cas des grands tableaux disjonctifs complets, elle est beaucoup plus délicate à utiliser dans le cas intermédiaire des petites tables de contingence multidimensionnelles.

Pour ce type de tableau aux facettes peu nombreuses, l'intérieur de la table de contingence (croisements de plus de deux variables), s'il contient des effectifs suffisants, est intéressant à décrire de façon détaillée. Une analyse des correspondances multiples sur un tableau comportant trois ou quatre variables nominales donne des résultats assez grossiers, d'une stabilité douteuse.

Il existe en la matière des savoir-faire, sans qu'une méthodologie rigoureuse se soit imposée définitivement : on peut juxtaposer des tranches en ligne ou en colonnes (cf. par exemple van der Heijden (1987) pour le cas des données longitudinales) ; juxtaposer des tableaux obtenus par croisements des variables initiales ; positionner en éléments supplémentaires les croisements de variables deux à deux dans les plans factoriels d'une analyse des correspondances multiples ; dans certains cas, réaliser une analyse factorielle multiple (cf. chapitre 8).

D'autres approches seront évoquées plus loin. C'est à propos de ce type d'applications que l'on pourra parler de complémentarité entre les méthodes.

b – Liens théoriques entre l'analyse des correspondances et les modèles log-linéaires

L'analyse des correspondances analyse l'écart entre un tableau de fréquence f_{ij} et un tableau modèle $f_{i.}f_{.j}$ correspondant à l'hypothèse d'indépendance. Lorsque cet écart est significatif[1], elle décrit de façon suggestive les associations privilégiées entre lignes et colonnes responsables de cet écart.

Ce principe d'analyse est manifestement insuffisant pour les tables de contingence à plus de deux entrées. Certes, l'analyse des correspondances multiples constitue *une* généralisation possible de cette démarche, réalisant une sorte de compromis entre tous les croisements des variables prises deux à deux.

Cette généralisation est opératoire lorsque le nombre et la nature des variables nominales exclut une étude méthodique de leurs interactions : on a alors à

---

[1] Le classique χ^2 permet d'alerter l'utilisateur sur la signification de cet écart, mais les premières valeurs propres de l'analyse des correspondances, ainsi que les taux d'inertie correspondants, peuvent également mesurer des écarts que le χ^2 ne décèle pas; cf. § 4.4.1.

traiter un tableau (individus × variables), comme en analyse en composantes principales.

Mais il n'existe pas d'analogue du théorème d'Eckart et Young dans le cas des tableaux tridimensionnels[1]. Il ne peut donc exister dans ce cas de démarche exploratoire aussi bien assise que dans le cas des tableaux à double entrée.

La démarche proposée par van der Heijden et de Leeuw (1985) puis développée par van der Heijden (1987), qui s'apparente aux analyses partielles évoquées au chapitre 8, va effectivement dans le sens d'une utilisation synergique des deux approches : utiliser le modèle log-linéaire pour éliminer l'effet complexe de certaines variables et utiliser l'analyse des correspondances pour décrire les résidus que le modèle log-linéaire ne permet pas d'expliquer.

Elle rejoint une généralisation de l'analyse des correspondances introduite par Escofier (1984) qui permet d'introduire des modèles moins restrictifs. L'analyse factorielle des correspondances se généralise à un modèle différent du modèle d'indépendance en supposant que les marges du tableau de référence sont distinctes de celles du tableau étudié.

Les liens théoriques entre l'analyse des correspondances et les modèles log-linéaires sont très ténus, même dans des contextes relativement simples. Après Escoufier (1982), Worsley (1987), van der Heijden *et al.* (1989), écrivons ce que pourrait être un *modèle* de l'analyse des correspondances dans le cas d'une approximation bi-dimensionnelle de la loi f_{ij}.

La formule de reconstitution des données en analyse des correspondances (cf. § 4.2.2-e) peut s'écrire, en retenant deux axes :

$$f_{ij} \approx f_{i.}\, f_{.j} \left\{ 1 + \sum_{h=1}^{2} \sqrt{\lambda_h}\ \varphi_h(i)\psi_h(j) \right\}$$

ce qui suggère un modèle de la forme :

$$f_{ij} \approx e_{ij} = c\, p_i q_j \left(1 + r_{1i} s_{1j} + r_{2i} s_{2j} \right)$$

où les coefficients inconnus, assujettis aux mêmes contraintes que leurs homologues de la formule de reconstitution, sont déterminés de façon à rendre minimale une distance entre f_{ij} et e_{ij}[2].

---

[1] Ce que l'on peut exprimer dans les termes suivants : il existe une décomposition hiérarchique unique d'un élément du produit tensoriel de deux espaces euclidiens en une somme de produits tensoriels de vecteurs appartenant à chacun des deux espaces. Mais une telle décomposition n'est pas unique dans le cas de d'un élément du produit tensoriel de plus de deux espaces euclidiens (cf. Benzécri, 1973; Tome 2B, n°6 [RED.TENS.]).

[2] Distance du χ^2, critère de Kullback-Leibler, ou encore critère de la déviance, très utilisé pour les modèles logistiques (cf. par exemple Celeux et Nakache, 1994 ; Droesbeke et al. , 2005).

Le modèle peut s'écrire, si les valeurs propres λ_1 et λ_2 sont petites par rapport à *1*, ce qui est le cas au voisinage de l'indépendance :

$$log\ f_{ij} \approx log\ e_{ij} = a_0 + a_i + b_j + r_{1i}s_{1j} + r_{2i}s_{2j}$$

alors qu'un modèle log-linéaire saturé s'écrit :

$$log\ e_{ij} = a'_0 + a'_i + b'_j + u_{ij}$$

Ainsi, l'analyse des correspondances suggère de décomposer le terme d'interaction u_{ij} sous forme simplement multiplicative dans le cas d'un seul facteur, et plus généralement sous forme de matrice de rang q dans le cas où l'on retient q facteurs.

Il est vrai que dans le cas d'une table de contingence à double entrée, le modèle log-linéaire non-saturé est trivial (hypothèse d'indépendance) et le modèle saturé aussi (ajustement parfait).

D'où les tentatives de donner au terme d'interaction des formes plus simples, avec en particulier les modèles dit *RC*, puis *multifactor* de Goodman (cf. Goodman, 1986).

L'analyse des correspondances, qui revient à une décomposition aux valeurs singulières de la matrice normée (que l'on peut appeler matrice d'interaction) :

$$\frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}}$$

répond à une même préoccupation[1].

Le cas des tables de contingences multiples est beaucoup plus complexe, et dans les configurations où le modèle log-linéaire peut être appliqué (peu de variables, beaucoup d'individus, des idées a priori sur le rôle de telle ou telle variable) l'approche "analyse de résidus" mentionnée plus haut paraît bien appropriée.

c – Difficultés de l'articulation exploration-inférence

Lorsque l'on est en situation trop exploratoire pour pouvoir formuler des hypothèses, ou lorsque le nombre de variables est trop élevé par rapport au nombre des individus pour pouvoir construire un modèle pertinent, on a recours à l'analyse des correspondances multiples.

Son utilisation permet d'une part de déceler, dans un premier temps, les liaisons intéressantes entre certaines variables, et d'autre part de sélectionner et réduire les variables et leurs modalités. Rappelons que l'on travaille sur les "faces de

---

[1] Elle effectue cette décomposition dans un cadre géométrique euclidien simple, en produisant des visualisations assorties de règles d'interprétation.

l'hypercube" c'est-à-dire sur les cumuls de fréquences correspondant à des effectifs importants.

On pourrait penser tester les liaisons par des modèles log-linéaires afin de préciser et de mesurer le niveau et l'intensité de celles-ci (l'intérieur de l'hypercube, lorsque le nombre d'individus le permet). Cette démarche demande cependant une certaine prudence. Ce serait en effet une erreur de raisonnement (malheureusement répandue chez les praticiens) de penser que l'on peut tester sur des données un modèle suggéré par les mêmes données.

Comme l'a spécifié Cox (1977) dans un remarquable article de synthèse sur les tests de signification, l'articulation *exploratoire - confirmatoire* pose des problèmes d'une grande complexité, analogues à ceux que nous avons rencontrés dans la section précédente à propos de l'analyse discriminante : tester une fonction discriminante sur l'échantillon d'apprentissage donne une idée trop optimiste de son pouvoir de prédiction. Dans les deux cas en effet, les échantillons, *et donc les fluctuations qui leurs sont propres,* sont sollicités soit pour construire une fonction ou une règle de classement (cas de l'analyse discriminante) soit pour choisir un modèle (cas d'une analyse des correspondances multiples préalable à un modèle log-linéaire).

La difficulté est accentuée par l'effet "comparaisons multiples" que l'on peut craindre dans la mesure où l'analyse des correspondances multiples peut traiter simultanément plusieurs dizaines, voire des centaines de variables.

Même lorsque le tableau contenant p variables nominales est généré selon un modèle stipulant l'indépendance totale entre les p variables, un certain nombre de paires de variables (parmi les $p(p-1)/2$ paires possibles) peut donner lieu à des liaisons significatives selon les valeurs usuelles des seuils, et ceci d'autant plus facilement que p est grand.

Un modèle restreint à cette sélection de variables pourrait effectivement confirmer une structure qui ne serait en fait qu'un artefact.

Il existe au moins deux types de solutions pragmatiques pour contourner ces difficultés : travailler sur un échantillon supplémentaire (échantillon-test, validation croisée) comme dans le cas de la discrimination; travailler avec des seuils de signification plus sévères au niveau de la lecture des modèles (comme dans le cas de comparaisons multiples).

Remarquons que la démarche "analyse des correspondances des résidus d'un modèle log-linéaire" mentionnée plus haut, qui correspond à une articulation en sens inverse : *Inférence -Exploration,* ne prête pas le flanc à ces critiques. Elle correspond à une situation méthodologique plus particulière, pour laquelle les modèles log-linéaires pouvaient être utilisés d'emblée.

5.7 Annexe technique du chapitre 5

► *Equivalence entre analyse en composantes principales et analyse des correspondances multiples dans le cas où toutes les questions* ***ont*** *deux modalités :*

Rappelons que d'après la formule [5.2 - 2] donnée en section 5.2.1:

$$\frac{1}{s}\mathbf{D}^{-1}\mathbf{B}\Phi = \lambda\Phi \qquad [5.7 - 1]$$

Explicitons cette relation où **D** désigne la matrice diagonale ayant les mêmes éléments diagonaux que **B** et où l et j désignent deux modalités :

$$\frac{1}{s}\sum_{j\in p}\frac{b_{lj}}{b_{ll}}\Phi_j = \lambda\Phi_l \qquad [5.7 - 2]$$

L'ensemble p des p modalités est partitionné en deux sous-ensembles p^1 et p^2 formés respectivement des premières et des deuxièmes modalités de chacune des s questions :

$$p = p^1 \cup p^2$$

Pour tout $q \in s$:

$$p_q = \{j_q^1, j_q^2\}$$

avec $j_q^1 \in p^1$ et $j_q^2 \in p^2$. Notons les relations, pour tout $q \in s$:

$$b_{lj_q^1} + b_{lj_q^2} = b_{ll} \quad \text{pour tout } l \in p$$

Cette relation exprime que ceux qui ont choisi la réponse l et l'une *ou* l'autre des deux modalités de la question J_q sont simplement ceux qui ont choisi la réponse l.

$$b_{j_q^1 j_q^1} + b_{j_q^2 j_q^2} = n \quad \text{et} \quad b_{j_q^1 j_q^1}\Phi_{j_q^1} = -b_{j_q^2 j_q^2}\Phi_{j_q^2}$$

La première relation exprime que tous les individus doivent choisir au moins une modalité de réponse pour chaque question, et la seconde traduit le fait que les coordonnées sont centrées pour chaque question.

Il suffit donc de restreindre la sommation de la relation [5.7 - 2] au seul ensemble p^1, dont l'élément courant sera désormais noté j :

$$\frac{1}{sb_{ll}}\sum_{j\in p^1}\left(b_{lj} - \frac{(b_{ll} - b_{lj})b_{jj}}{n - b_{jj}}\right)\Phi_j = \lambda\Phi_l$$

Ce qui peut s'écrire :

$$\sum_{j\in p^1}\frac{n\,b_{lj} - b_{ll}\,b_{jj}}{s\,(n - b_{ll})\,b_{ll}}\,\Phi_j = \lambda\,\Phi_l \qquad [5.7 - 3]$$

Calculons les moments empiriques centrés du second ordre des s variables caractérisées par leurs premières modalités :

$$Cov(l,j) = \frac{1}{n}(b_{lj} - \frac{b_{ll}b_{jj}}{n})$$

$$Var(j) = \frac{1}{n}(b_{jj} - \frac{b_{jj}^2}{n})$$

Le terme général de la matrice des corrélations des s variables s'écrit :

$$Cor(l,j) = \frac{nb_{lj} - b_{ll}b_{jj}}{\sqrt{(n - b_{jj})\, b_{jj}\ (n - b_{ll})\, b_{ll}}}$$

Il est clair que si (Φ, λ) est la solution de l'équation [5.7 - 3] alors (Φ^*, λ^*) est la solution de :

$$\sum_{j \in p^1} Cor(l,j)\Phi_j^* = \lambda^*\Phi_l^*$$

avec :

$$\Phi_j = \Phi_j^* \frac{\sqrt{n - b_{jj}}}{\sqrt{b_{jj}}}$$

et :

$$\lambda^* = \lambda s$$

Les facteurs et les valeurs propres d'une analyse des correspondances multiples de s variables à deux modalités ($p = 2s$) sont bien reliés par une relation simple à ceux d'une analyse en composantes principales normées effectuée sur les premières (ou les secondes) modalités de chacune des s questions (sélection de s colonnes du tableau disjonctif complet, une par question).

Chapitre 6

Méthodes de classification

Les techniques de classification automatique sont destinées à produire des groupements (ou : classes) d'objets ou d'individus à partir d'un certain nombre de variables ou de caractères. On distingue ces techniques des méthodes de *classement* pour lesquelles il s'agit d'affecter des objets à des classes préalablement identifiées[1].

Les méthodes de classification sont des méthodes dites d'*apprentissage non supervisé* alors que les méthodes de classement relèvent de l'*apprentissage supervisé*.

La classification est une branche de l'analyse des données qui a donné lieu à des publications nombreuses et diversifiées. Elle s'est beaucoup développée, ces dernières années, pour répondre au besoin d'extraire de façon automatique l'information cachée ou d'identifier des groupes ou classes à partir d'importantes masses d'information de plus en plus spécifiques. C'est l'outil de la statistique exploratoire multidimensionnelle sans doute la plus utilisée pour la fouille de données. Les ouvrages spécialisés (notamment, en langue française, le tome 1 du traité d'analyse des données de Benzécri, 1973) contiennent d'importantes considérations historiques et de rigoureux développements formels sur la notion de classification. L'ouvrage de base, historique, est celui de Sokal et Sneath (1963). Les premiers manuels publiés furent ceux de Lerman (1970), Anderberg (1973), Benzécri (1973), Hartigan (1975), Lerman (1981) et Gordon (1981) auxquels nous ne pouvons que renvoyer le lecteur pour des préalables fondamentaux. A l'intention des praticiens, l'ouvrage de Nakache et

---

[1] La méthode de classement la plus utilisée est l'analyse discriminante qui fait l'objet du chapitre 7.

Confais (2004) présente un large éventail de méthodes de classification des plus classiques aux plus récentes[1]. Nous nous bornerons ici aux principes de base des méthodes les plus utilisées en insistant sur le *lien* et la *complémentarité* avec les méthodes en axes principaux des cinq premiers chapitres.

Les circonstances d'utilisation sont sensiblement les mêmes que celles des méthodes d'analyse factorielle présentées aux chapitres précédents : l'utilisateur se trouve face à un tableau rectangulaire de valeurs numériques. L'objet est de produire des groupements de lignes ou de colonnes de ce tableau. Ce peut être un tableau de valeurs numériques continues (valeur de la variable j pour l'individu i, à l'intersection de la ligne i et de la colonne j du tableau), un tableau de contingence (croisant deux partitions d'une même population), ou encore un tableau de présence-absence (valeurs *0* ou *1* selon que tel individu ou objet possède tel caractère ou attribut). Dans certaines applications, l'utilisateur peut disposer d'un tableau carré symétrique de similarités ou de distances.

Le recours aux techniques de classification automatique est sous-tendu par quelques idées générales concernant le champ d'observation. On suppose que certains regroupements doivent exister, ou au contraire on exige que certains regroupements soient effectués. Autrement dit, on ne se satisfait pas d'une visualisation plane et continue des associations statistiques et l'on manifeste, implicitement ou explicitement, un intérêt pour la mise en évidence de *classes* d'individus ou de caractères.

Les représentations synthétiques se manifestent soit sous la forme de *partitions* des ensembles étudiés (lignes ou colonnes du tableau analysé), soit sous la forme de *hiérarchie de partitions* que nous définirons de façon plus précise ultérieurement. Quelquefois, il s'agira *d'arbres* au sens de la théorie des graphes, arbres dont les sommets sont les objets à classer. Enfin on pourra rechercher des *classes empiétantes* ou simplement mettre en évidence des *zones à forte densité*, laissant de nombreux individus ou caractères non classés.

A une même famille de résultats correspond parfois des démarches et des interprétations différentes. Il peut s'agir de découvrir une partition ayant une existence réelle (cette existence étant conjecturée avant l'analyse statistique ou étant révélée à l'issue des calculs) ou l'on veut au contraire utiliser les partitions produites comme des outils ou des intermédiaires de calculs permettant une exploration des données.

Cette dernière démarche généralise en quelque sorte la construction d'histogrammes de la statistique unidimensionnelle : en vue d'une étude plus aisée, les observations sont regroupées par paquets homogènes, même si la

[1] Une des premières synthèses historiques sur le sujet est celle de Cormack (1971). Une synthèse de travaux plus récents en classification hiérarchique a été faite par Gordon (1987). Cf. également les manuels généraux de Jambu et Lebeaux (1978), Chandon et Pinson (1981), Murtagh (1985), Roux (1985), Kaufman et Rousseeuw (1990)

construction de ces paquets implique un découpage quelque peu arbitraire d'un ensemble continu.

Pour l'essentiel, les techniques de classification font appel à une démarche algorithmique et non aux calculs formalisés usuels. Alors que les valeurs des composantes des axes factoriels, par exemple, sont la solution d'une équation pouvant s'écrire sous une forme très condensée (même si sa résolution est complexe), la définition des classes ne se fera qu'à partir d'une formulation algorithmique: une série d'opérations est définie de façon récursive et répétitive. Il en découle que la mise en œuvre de la plupart des techniques de classification ne nécessite que des notions mathématiques relativement élémentaires.

Il existe plusieurs familles d'algorithmes de classification : les *méthodes de partitionnement* comme les méthodes d'agrégation autour de centres mobiles; les *algorithmes ascendants* (ou encore agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux, et qui fournissent une hiérarchie de partitions des objets; enfin les *algorithmes descendants* (ou encore divisifs) qui procèdent par dichotomies successives de l'ensemble des objets, et qui peuvent encore fournir une hiérarchie de partitions.

On se limitera ici aux deux premières techniques de classification : les méthodes de partitionnement (section 6.1) et les classifications ascendantes hiérarchiques (section 6.2). Ces techniques ont chacune leurs avantages et peuvent être utilisées conjointement. Il est ainsi possible d'envisager une stratégie de classification basée sur un *algorithme mixte*, particulièrement adapté au partitionnement d'ensembles de données comprenant des milliers d'individus à classer (§ 6.3.1).

Un des avantages des méthodes de classification est de donner lieu à des éléments (les classes) souvent plus faciles à décrire automatiquement que les axes factoriels. Les outils de description seront évoqués au paragraphe 6.3.2.

La pratique montre que l'utilisateur a intérêt à utiliser de façon conjointe les méthodes factorielles et les méthodes de classification.

Les aspects théoriques et pratiques de la complémentarité entre ces deux familles de méthodes exploratoires seront abordés au paragraphe 6.4.

Enfin, la dernière section (6.5) rend compte de la validation des méthodes de classification en évoquant brièvement les travaux relatifs au nombre et à la signification des classes.

6.1 Méthodes de partitionnement

Il s'agit pour l'essentiel des techniques *d'agrégation autour de centres mobiles*, et des *cartes auto-organisées (Self Organising Maps)* appelées encore *cartes de Kohonen*. Ces méthodes sont particulièrement intéressantes dans le cas des grands tableaux car elles sont peu coûteuses en temps calcul et peu gourmandes en espace mémoire.

6.1.1 Agrégation autour des centres mobiles

Bien qu'elle ne fasse appel qu'à un formalisme limité et que son efficacité soit dans une large mesure attestée par les seuls résultats expérimentaux, la méthode d'*agrégation autour de centres mobiles* est probablement la technique de partitionnement la mieux adaptée actuellement aux vastes recueils de données ainsi que la plus utilisée pour ce type d'application. Produisant des partitions des ensembles étudiés, elle est utile aussi bien comme technique de description et d'analyse que comme technique de réduction, généralement en association avec des analyses factorielles et d'autres méthodes de classification.

L'algorithme peut être imputé principalement à Forgy (1965), bien que de nombreux travaux (parfois antérieurs : Thorndike, 1953), le plus souvent postérieurs (MacQueen, 1967; Ball and Hall, 1967) aient été menés parallèlement et indépendamment pour introduire des variantes ou des généralisations. Cette méthode peut être considérée comme un cas particulier de techniques connues sous le nom de *nuées dynamiques* étudiées dans un cadre formel par Diday (1971).

Elle est particulièrement intéressante pour les gros fichiers numériques car les données sont traitées en *lecture directe* : le tableau des données, conservé sur une mémoire auxiliaire (disque) est lu plusieurs fois de façon séquentielle, sans jamais encombrer de zones importantes dans la mémoire vive de l'ordinateur. La lecture directe permet également d'utiliser au mieux les particularités du codage des données, ce qui réduit le temps de calcul dans le cas des codages disjonctifs.

a _ Bases théoriques de l'algorithme

Soit un ensemble I de n individus à partitionner, caractérisés par p caractères ou variables.

On suppose que l'espace $\mathcal{R}^p$ supportant les n points-individus est muni d'une distance appropriée notée d (souvent distance euclidienne usuelle ou distance du χ^2). On désire constituer au maximum q classes. Les étapes de l'algorithme sont illustrées par la figure 6.1 - 1.

Étape *0* : On détermine q centres provisoires de classes (par exemple, par tirage pseudo-aléatoire sans remise de q individus dans la population à classifier, selon une préconisation de MacQueen). Les q centres :

$$\left\{C_1^0,\ldots,C_k^0,\ldots,C_q^0\right\}$$

induisent une première partition P^0 de l'ensemble des individus I en q classes :

$$\left\{I_1^0,\ldots,I_k^0,\ldots,I_q^0\right\}$$

Ainsi l'individu i appartient à la classe I_k^0 s'il est plus proche de C_k^0 que de tous les autres centres[1].

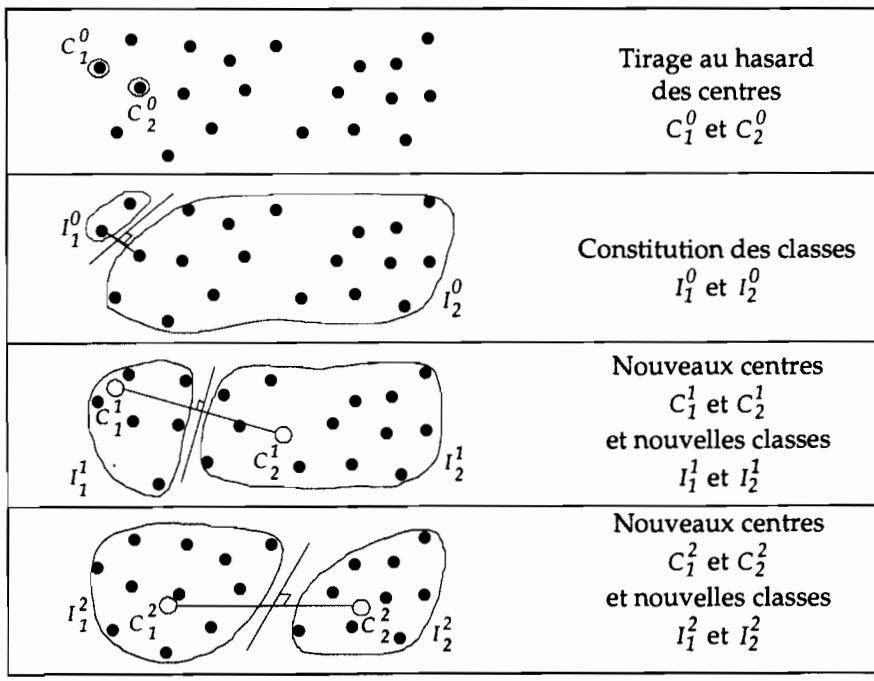


Figure 6.1 – 1 : Etapes de l'algorithme des centres mobiles

Étape *1*: On détermine q nouveaux centres de classes :

$$\left\{C_1^1,\ldots,C_k^1,\ldots,C_q^1\right\}$$

en prenant les centres de gravité des classes qui viennent d'être obtenues :

$$\left\{I_1^0,\ldots,I_k^0,\ldots,I_q^0\right\}$$

[1] Les classes sont alors délimitées dans l'espace par les cloisons polyédrales convexes formées par les plans médiateurs des segments joignant tous les couples de centres.

Ces nouveaux centres induisent une nouvelle partition P^1 de I construite selon la même règle que pour P^0.

La partition P^1 est formée des classes notées :

$$\left\{I_1^1,\ldots,I_k^1,\ldots,I_q^1\right\}$$

Étape *m*: On détermine *q* nouveaux centres de classes :

$$\left\{C_1^m,\ldots,C_k^m,\ldots,C_q^m\right\}$$

en prenant les centres de gravité des classes qui ont été obtenues lors de l'étape précédente,

$$\left\{I_1^{m-1},\ldots,I_k^{m-1},\ldots,I_q^{m-1}\right\}$$

Ces nouveaux centres induisent une nouvelle partition P^m de l'ensemble I formée des classes :

$$\left\{I_1^m,\ldots,I_k^m,\ldots,I_q^m\right\}$$

Le processus se stabilise nécessairement (voir paragraphe suivant) et l'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit lorsqu'un critère convenablement choisi (par exemple, la mesure de la variance intra-classes) cesse de décroître de façon sensible, soit encore parce qu'un nombre maximal d'itérations a été fixé *a priori.*

Généralement, la partition obtenue finalement dépend du choix initial des centres.

b _ Justification élémentaire de l'algorithme

On va montrer que la variance intra-classes ne peut que décroître (ou rester stationnaire) entre l'étape *m* et l'étape *m + 1*. Des *règles d'affectation*[1] permettent de faire en sorte que cette décroissance soit stricte et donc de conclure à la convergence de l'algorithme puisque l'ensemble de départ I est fini[2].

Supposons que les *n* individus de l'ensemble à classer I soient munis de masses relatives p_i (leur somme vaut *1*) et soit $d^2(i,C_k^m)$ le carré de la distance entre l'individu *i* et le centre de la classe *k* à l'étape *m*. Nous nous intéressons à la quantité *critère* :

$$v(m)=\sum_{k=1}^{q}\left\{\sum_{i\in I_k^m} p_i d^2(i,C_k^m)\right\}$$

---

[1] Ces règles sont des conventions de programmation propres à chaque variante ou spécification de l'algorithme.

[2] Bien entendu ce n'est pas la convergence, mais la vitesse de convergence qui justifierait en pratique l'utilisation de la méthode.

Rappelons qu'à l'étape m, la classe I_k^m est formée des individus plus proches de C_k^m que de tous les autres centres (ces centres étant des centres de gravité des classes I_k^{m-1} de l'étape précédente).

La variance intra-classes à l'étape m est la quantité :

$$V(m) = \sum_{k=1}^{q} \left\{ \sum_{i \in I_k^m} p_i d^2(i, C_k^{m+1}) \right\}$$

où C_k^{m+1} est le centre de gravité de la classe I_k^m. A l'étape $m+1$, la quantité critère s'écrit :

$$v(m+1) = \sum_{k=1}^{q} \left\{ \sum_{i \in I_k^{m+1}} p_i d^2(i, C_k^{m+1}) \right\}$$

On va montrer que :

$$v(m) \geq V(m) \geq v(m+1)$$

ce qui établira la décroissance simultanée du critère et de la variance intra-classes. En notant p_k la somme des p_i pour $i \in I_k^m$, remarquons tout d'abord d'après le théorème de Huygens :

$$v(m) = V(m) + \sum_{k=1}^{q} p_k d^2(C_k^{m+1}, C_k^m)$$

ce qui établit la première partie de l'inégalité.

La seconde partie découle du fait qu'entre les accolades qui apparaissent dans les définitions de $V(m)$ et $v(m+1)$, seules changent les affectations des points aux centres. Puisque I_k^{m+1} est l'ensemble des points plus proches de C_k^{m+1} que de tous les autres centres, les distances n'ont pu que décroître (ou rester inchangées) au cours de cette réaffectation.

c _ Techniques connexes

Il existe de nombreux algorithmes dont le principe général est voisin de l'algorithme d'agrégation autour de centres mobiles mais qui en diffèrent cependant sur certains points[1].

Ainsi, dans la technique des *nuées dynamiques* (Diday, 1972, 1974), les classes ne sont pas caractérisées par un centre de gravité, mais par un certain nombre d'individus à classer, dénommés "étalons", qui constituent alors un "noyau" ayant pour certaines utilisations un meilleur pouvoir descriptif que des centres ponctuels. Ce formalisme a permis plusieurs généralisations de la méthode.

[1] Pour des informations plus détaillées sur les techniques d'agrégation autour des centres mobiles, cf. les ouvrages de Benzécri (1973) et Anderberg (1973).

La méthode dite des *k-means (k-moyennes)* introduite par MacQueen (1967) commence effectivement par un tirage pseudo-aléatoire de centres ponctuels. Cependant la règle de calcul des nouveaux centres n'est pas la même. On n'attend pas d'avoir procédé à la réaffectation de tous les individus pour modifier la position des centres : chaque réaffectation d'individus entraîne une modification de la position du centre correspondant[1].

En une seule itération, cette procédure peut ainsi donner une partition de bonne qualité. Mais celle-ci dépendra de l'ordre des individus sur le fichier, ce qui n'est pas le cas pour la technique exposée précédemment[2].

d _ Formes fortes et groupements stables

Les algorithmes d'agrégation autour de centres mobiles convergent vers des *optima locaux*. Le problème de la recherche d'une partition *optimale* en q classes (en prenant comme critère la variance intra-classes, qu'il faut alors rendre minimale sur l'ensemble des partitions possibles en q classes) n'a pas jusqu'à présent donné lieu à un algorithme satisfaisant[3]. Les partitions obtenues dépendent en général des premiers centres choisis.

La procédure de recherche de *groupements stables* (ou encore *formes fortes)*, suggérée pour l'essentiel par E. Diday (1972), permet de remédier au moins partiellement à cet inconvénient. Elle a surtout l'avantage de nuancer les résultats souvent trop frustes que l'on obtient dans le cadre rigide d'une seule partition, en mettant en évidence les zones à forte densité du nuage des points-individus. Cette technique consiste à effectuer plusieurs partitions à partir de plusieurs ensembles différents de centres, et à retenir comme *groupements stables* les ensembles d'individus qui ont toujours été affectés à une même classe dans chacune des partitions (cf. figure 6.1 - 2).

Supposons que l'on effectue s partitions $\{P_1, P_2, ..., P_s\}$ en q classes chacune. Dans la *partition-produit*, la classe indexée par $\{k_1, k_2, ..., k_s\}$ contient les individus ayant appartenu à la classe k_1 de P_1, puis à la classe k_2 de P_2, etc., enfin à la classe k_s de P_s. Les classes contenant plus d'un individu de la partition-produit constitueront les groupements stables.

---

[1] On parle parfois d'algorithme en ligne (*on line*) pour ce type de modification en cours de lecture, alors que la méthode exposée plus haut procède par paquet (*batch*).

[2] D'autres méthodes diffèrent par le choix initial des *centres* (individus équidistants pour Thorndike (1953), par l'introduction de *seuils* ou de *protections* destinés à modifier éventuellement le nombre des classes. Ainsi la technique proposée sous le nom Isodata par Ball et Hall (1965) met en jeu plusieurs paramètres destinés à piloter l'élaboration de la partition.

[3] Dans le cas où les individus ne sont décrits que par un seul paramètre, le calcul d'une partition optimale exacte est possible car il existe une relation d'ordre entre les individus, ce qui limite considérablement l'éventail des partitions à examiner (cf. W.D. Fisher, 1958).

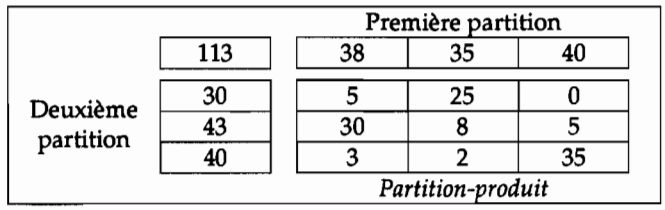


Figure 6.1 – 2 : Groupements stables dans la partition-produit

En pratique, le nombre de groupements stables ayant un effectif notable sera très inférieur à q^s. Sur les *38* individus de la classe *1* de la partition *1*, on en retrouve *30* dans la classe 2 de la partition 2.

Pour fixer les idées, on obtient sur *1000* individus une première partition en *6* classes autour de centres mobiles (*15* itérations ont été nécessaires pour assurer une stabilité des groupes). On répète deux fois cette procédure. Le tableau 6.1 - 1 donne les effectifs des *6* classes des *3* partitions de base successives.

Tableau 6.1 – 1 Trois partitions de base en 6 classes

	1	2	3	4	5	6
Partition 1	127	188	229	245	151	60
Partition 2	232	182	213	149	114	110
Partition 3	44	198	325	99	130	204

Ces *3* partitions sont, à l'étape suivante, croisées entre elles et l'on obtient $6^3 = 216$ classes. Les individus de chacune de ces *216* classes sont ceux qui ont toujours été regroupés ensemble dans les *3* partitions de base. Ils constituent les groupements stables. En fait seulement *50* groupes ne sont pas vides et seulement *10* ont plus de *15* individus. La distribution de ces individus est donnée dans le tableau 6.1 - 2.

Tableau 6.1 – 2. Groupements stables rangés par effectifs décroissants

Groupes 1 à 10	168	118	114	107	88	83	78	26	22	16
Groupes 11 à 20	15	14	12	12	12	11	10	7	7	7
Groupes 21 à 30	6	6	4	4	4	4	3	3	3	3
Groupes 31 à 40	3	3	3	2	2	2	2	2	2	2
Groupes 41 à 50	1	1	1	1	1	1	1	1	1	1

Remarque

La recherche des groupements stables constitue une exploration des zones de fortes densité dans l'espace, mais ne fournit pas une partition utilisable en pratique, car le nombre de classes est en général trop élevé, et corrélativement les effectifs de certaines classes sont trop faibles (cf. les 50 groupements du tableau

6.1 - 2). De façon pragmatique, on peut utiliser les premiers groupements stables pour définir une partition de la façon suivante : le nombre de classes pourra être suggéré par le nombre de groupements d'effectifs notables : ainsi, les 7 premiers groupements du tableau 6.1 - 2 ont des effectifs importants (il y a de plus un écart important entre 78 et 26). Les classes seront obtenues par réaffectation des individus restants aux groupements retenus les plus proches (affectation des individus des groupements 8 à 50 autour des centres des 7 premiers groupements pour notre exemple). Mais nous verrons que les méthodes mixtes à la section 6.3 permettent de perfectionner cette démarche.

6.1.2 Cartes auto-organisées

Les cartes auto-organisées[1] (ou SOM, *Self-Organising Maps*), introduites en 1981 par Teuvo Kohonen (1989) sont des partitions dont les classes sont présentées sur une grille rectangulaire ou octogonale. La position des classes sur la grille reflète, autant que faire se peut, les proximités entre classes.

L'algorithme associé est analogue à celui des centres mobiles présenté ci-dessus, mais s'en distingue par la conservation de la topologie de l'espace des objets à classer.

Ces cartes sont souvent efficaces en reconnaissance des formes selon des techniques non linéaires et constituent de bons outils de visualisation. Elles donnent lieu à plusieurs applications relevant par exemple de l'analyse de textes, des diagnostics médicaux et industriels, des contrôles de processus, de robotique et font l'objet de nombreuses publications[2].

a _ Principe

Une carte auto-organisée de Kohonen est un réseau de neurones traduit le plus souvent sous forme d'une grille rectangulaire[3] (parfois hexagonale) aux mailles déformables : les neurones ou encore les cellules de la grille sont en fait les classes.

La carte s'efforce ainsi de conserver la topologie des objets, et ceux qui se ressemblent seront proches sur la carte. L'exemple de la figure 6.1.3 montre une grille carrée de 16 neurones (cases de la grille).

---

[1] Elles font également partie des méthodes dites *neuronales* (cf. chapitre 7).

[2] Citons l'ouvrage de Blayo et Verleysen (1996) et les travaux du laboratoire SAMOS de l'université de Paris I (http://samos-univ-paris1.fr), et Thiria et al. (1997).

[3] On peut aussi considérer une carte unidimensionnelle, appelée ficelle, ou une carte à trois dimensions définissant un cylindre.

13	14	15	16
9	10	11	12
5	6	7	8
1	2	3	4

Figure 6.1.3 : Exemple de grille rectangulaire (4 x 4)

Il s'agit de classer un ensemble de n individus ou objets, chacun décrit par p variables. On considère par conséquent, comme pour les méthodes précédentes, un nuage de points représenté par une matrice **X** d'ordre (n,p).

Pour des raisons historiques et disciplinaires, on utilisera le vocabulaire des réseaux de neurones, mais il suffit de savoir que dans ce contexte, *neurone* et *classe* sont synonymes. On a donc affaire à un réseau de k neurones, les entrées sont les n individus à classer et un neurone $\mathbf{u}_j$ est défini dans l'espace $\mathcal{R}^p$ par un vecteur poids ou encore un vecteur-code, $\mathbf{C}_j$ $(j=1,\ldots,k)$.

Au départ, il est tiré au hasard parmi les objets à classer. On se fixe k, le nombre de neurones qui n'est autre que le nombre de classes souhaité. On définit également une distance d entre objets et neurones (classes) et un système de voisinages entre neurones qui rend compte de la topologie du réseau.

A chaque neurone $\mathbf{u}_j$, est attribué un voisinage $V_r(\mathbf{u}_j)$ de rayon r formé de l'ensemble des neurones situés sur le réseau à une distance inférieure ou égale à r. Ces voisinages peuvent être choisis de diverses manières mais en général on les suppose directement contigus sur la grille rectangulaire[1].

A chaque itération, un objet $\mathbf{x}_i$ est tiré aléatoirement et est classé selon la méthode du plus proche voisin : $\mathbf{x}_i$ appartient à la classe $\mathbf{u}_j$ si et seulement si le vecteur poids $\mathbf{C}_j$ est le plus proche de $\mathbf{x}_i$ parmi tous les autres vecteurs codes. On détermine ainsi le « neurone gagnant », u_{k_0}. Son vecteur poids est mis à jour ainsi que ceux de ses voisins ce qui permet de rapprocher le « neurone gagnant » et ses voisins de l'objet à classer. C'est le code de la classe gagnante qui est donc modifiée, déformant à chaque étape la carte des neurones.

La conservation de la topologie permet de positionner toutes les classes les unes par rapport aux autres et la représentation sous forme d'une carte rend l'interprétation facile.

[1] Ce voisinage représente alors *quatre* ou *huit* voisins pour un « neurone » non situé au bord de la grille, selon la définition plus ou moins stricte de l'adjacence : quatre si l'on considère comme adjacentes seulement les cases ayant un côté commun, huit si l'on prend en compte les cases ayant un sommet commun.

b _ L'algorithme de Kohonen

L'algorithme d'apprentissage pour classer n points est itératif. On dispose d'un nuage de n points, définis dans $\mathcal{R}^p$ et pondérés par une loi de probabilité P=(p_1, p_2, ..., p_n). On initialise de façon aléatoire les vecteurs codes $\mathbf{C}_j$ (j=1,..,k) des k neurones $\mathbf{u}_j$, vecteurs définis dans le même espace des objets à classer.

A l'itération t, on choisit au hasard et suivant la loi de probabilité P un objet x_i (i=1,...,n). On cherche le neurone u_{k_0} de vecteur code $\mathbf{C}_{k_0}(t)$ le plus proche au sens de la distance d fixée a priori. On rapproche alors de l'objet x_i le neurone gagnant u_{k_0} et ses voisins en modifiant les vecteurs codes de la façon suivante :

$$\mathbf{C}_k(t)=\mathbf{C}_k(t\text{-}1) + h(t)\,(\mathbf{x}_i - \mathbf{C}_k(t\text{-}1))$$

où $\mathbf{C}_k(t\text{-}1)$ est le vecteur code associé à la classe k pour l'étape précédente, $\mathbf{x}_i$ est l'objet présenté à l'étape t, $h(t)$ un paramètre d'adaptation positif et inférieur à 1 qui contrôle la vitesse d'apprentissage. Cette expression n'intervient que pour la classe « gagnante » et ses voisines. Si $\mathbf{x}_i \notin V_r(u_{k_0})$, $\mathbf{C}_k(t)=\mathbf{C}_k(t\text{-}1)$.

A l'étape suivante $t+1$, on tire au hasard un autre objet à classer. On itère ainsi jusqu'à la convergence des neurones c'est-à-dire jusqu'à la stabilité de leurs vecteurs-codes traduite par : $|\mathbf{C}_k(t+1)-\mathbf{C}_k(t)| < \varepsilon$.

Remarque :

On montre que pour une meilleure classification, il est utile de faire décroître le paramètre $h(t)$ et le nombre de voisins au cours du processus. En effet, en pratique et pour avoir des résultats de convergence satisfaisants, le paramètre h ne doit être ni trop petit (le modèle ne s'adapte pas assez aux données et l'apprentissage est lent), ni trop grand (l'apprentissage est rapide mais il y a risque de non convergence et d'instabilité)[1].

Comme l'algorithme des centres mobiles, cet algorithme est adapté aux données importantes et difficiles à stocker en mémoire vive.

c _ Comparaison avec les centres mobiles

L'algorithme ne diffère de celui des centres mobiles que par la mise à jour des classes voisines. Pour les deux familles de méthodes, cette mise à jour peut se faire à chaque lecture d'individu (*on line*), ou après une lecture complète du jeu de données (*batch*).

Les cartes auto-organisées ont l'avantage de montrer une organisation des classes dans un format d'édition pratique (voir l'exemple de la figure 6.1-4).

---

[1] Contrairement au cas de l'agrégation autour de centres mobiles, la convergence n'est pas facile à démontrer. Cottrell et Fort (1987) ont démontré cette convergence dans le cas d'un réseau unidimensionnel (dénommé : *fil* ou *ficelle*).

Si l'on veut visualiser les proximités entre classes après une agrégation autour de centres mobiles classique, il est possible de projeter les centres des classes (points moyens ou centres de gravité) comme des variables supplémentaires dans les plans factoriels d'une analyse en axes principaux réalisée sur les mêmes variables que la classification (cf. l'exemple d'application du paragraphe 6.4.4 de ce chapitre). Il est également possible de procéder à une analyse en axes principaux du tableau (k, p) décrivant les k classes par les p coordonnées de leurs centres. Mais il s'agit alors de plusieurs visualisations bi-dimensionnelles (plans factoriels). La carte de Kohonen a l'avantage de fournir une représentation unique et non linéaire : la grille peut dans certains cas s'adapter (comme un filet) à des formes de nuage de points dans $\mathcal{R}^p$. On peut objecter que la contrainte de voisinage produit des classes de moins bonne qualité qu'une agrégation autour de centres mobiles sans contraintes. Ceci serait vrai si cette dernière méthode produisait des optima globaux, ce qui n'est pas le cas en général. Disons en bref que les cartes auto-organisées réalisent un compromis intéressant entre classification et visualisation[1].

d _ Application au jeu de données sémiométriques

Nous reprenons l'exemple des enquêtes sémiométriques décrites au chapitre 3 : un ensemble de $p = 70$ mots notés par $n = 300$ individus . Les visualisations des corrélations entre mots par analyse en composantes principales ont été examinées au paragraphe 3.5.2.

La carte auto-organisée de Kohonen (figure 6.1 - 4) regroupe les mots en classes. La grille imposée est la grille carrée (4, 4) de la figure 6.1-3. Elle conserve autant que faire se peut la topologie initiale de l'espace des mots. En ce sens, elle concurrence l'analyse en composantes principales. Ainsi, par exemple, les mots *sensuel, nudité, désir, charnel* appartiennent à une même classe (la classe 4 selon la numérotation de la figure 6.1-3). La classe 3, (*vitesse, sublime, gratuit*) voisine, contient des mots corrélés avec les premiers, mais à un moindre degré. Les cartes auto-organisées permettent une visualisation plane d'un nuage de points différente de celle obtenue par une analyse factorielle ; la représentation, non linéaire, n'est pas le résultat d'une projection. D'autre part, une seule carte est censée représenter le nuage de points alors qu'on consulte plusieurs plans factoriels.

La figure 6.1.5 représente la projection des centres de classes de la carte de Kohonen dans le plan factoriel (2, 3). La carte est repliée, car elle prend en compte d'autres dimensions. L'axe horizontal de la figure 6.1-5 est bien différencié le long de la diagonale de la carte 6.1-4 allant du haut-gauche [classe 13] vers le bas-droit [classe 4], l'axe 3 est plus dispersé sur la carte.

---

[1] Pour des présentations améliorées des cartes auto-organisées, cf. Kleiweg (1996), Cottrell et Rousset (1997), Rouset et Guinot (2002).

âme tradition sacré noble justice absolu Dieu	partir paix attachement	pureté féconder enfance consoler campagne	sommet montagne fleuve escalader eau douceur animal
modération interdire	vide rompre muraille	écrire livre	île nager lune bleu
rigide raison métallique	science réfléchir fusil certitude	créateur armure	évasion rêver inconnu feu danger aventurier
utilitaire prudence produire perfection matériel honneur commander	élégance précieux or maîtriser inventeur hériter confiance admirer	vitesse sublime gratuit	sensuel nudité désir charnel

Figure 6.1 - 4 : Proximités entre mots décrites par une carte de Kohonen

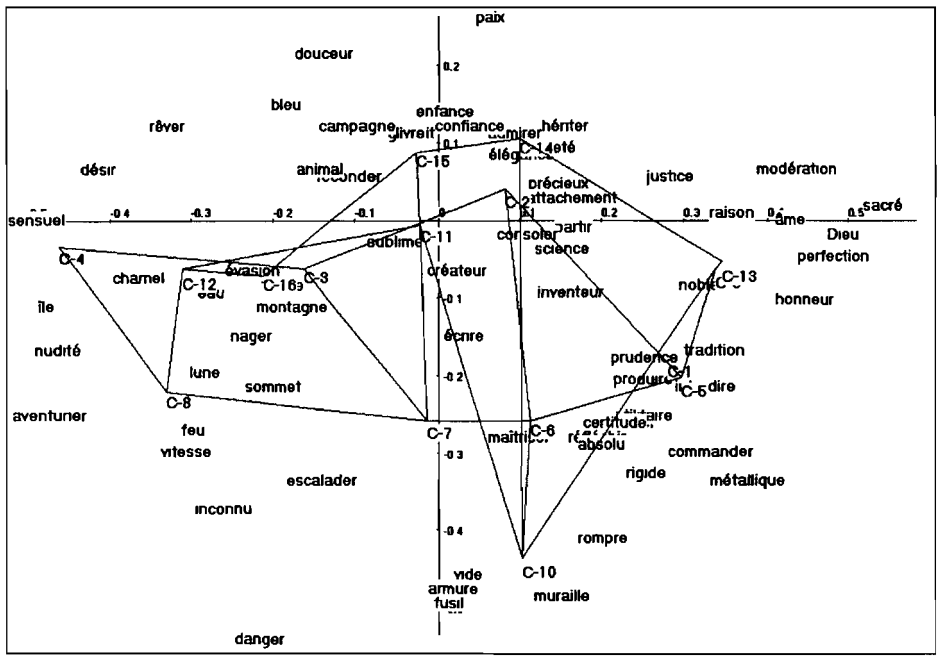


Figure 6.1 - 5 : Projection de la carte de Kohonen dans le plan factoriel (2,3)

Cependant la lecture de l'ensemble des plans factoriels, c'est-à-dire la prise en compte successive des axes pris deux à deux, nuance et affine la description des axes et leurs interprétations. De plus, certains axes ou certaines directions peuvent avoir une interprétation en tant que « variables latentes ». Un même mot peut être corrélé avec plusieurs axes révélant ainsi plusieurs dimensions sémantiques.

Ces différentes dimensions peuvent être appréhendées avec la carte de Kohonen mais l'analyse en composantes principales permet, au prix d'un effort de lecture plus grand, une description plus approfondie. L'analyse factorielle est donc plus complexe, mais plus riche.

En revanche, si l'on prend en compte une « économie de la visualisation » qui intègre le temps passé à déchiffrer des graphiques, à mettre en forme ces graphiques en vue d'un rapport ou d'une publication, à former les praticiens aux règles d'interprétation, les cartes auto-organisées constituent un outil particulièrement performant.

6.2 Classification hiérarchique

L'autre grande méthode de classification se fait par agglomération progressive, de façon ascendante des éléments deux à deux. Nous nous restreindrons ici à la classification ascendante hiérarchique dont nous présenterons plusieurs critères d'agrégation.

Nous envisagerons d'une part la technique "du saut minimal" (single linkage) équivalente, d'un certain point de vue, à la recherche de l'arbre de longueur minimale, et d'autre part la technique d'agrégation "selon la variance", intéressante par la compatibilité de ses résultats avec certaines analyses factorielles.

Les principes généraux communs aux diverses techniques de classification ascendante hiérarchique sont également extrêmement simples. Il est difficile de leur trouver une paternité car ces principes relèvent plus du bon sens que d'une théorie formalisée. Les exposés les plus systématiques et les plus anciens sont peut-être ceux de Sokal et Sneath (1963), puis de Lance et Williams (1967). Pour une revue synthétique, nous renvoyons à Gordon (1987, réédité en 1999)[1] .

[1] On trouvera également des exposés sur la classification hiérarchique dans pratiquement tous les ouvrages cités dans le paragraphe introductif de ce chapitre (incluant plusieurs manuels en langue française).

6.2.1 Principe

Le principe de l'algorithme consiste à créer, à chaque étape, une partition obtenue en agrégeant deux à deux les éléments les plus proches.

On désignera alors par *élément* à la fois les individus ou objets à classer eux-mêmes et les regroupements d'individus générés par l'algorithme. Il y a différentes manières de considérer le nouveau couple d'éléments agrégés, d'où un nombre important de variantes de cette technique.

L'algorithme ne fournit pas une partition en q classes d'un ensemble de n objets mais une *hiérarchie de partitions*, se présentant sous la forme *d'arbres* appelés également *dendrogrammes* et contenant $n - 1$ partitions. L'intérêt de ces arbres est qu'ils peuvent donner une idée du nombre de classes existant effectivement dans la population.

Chaque coupure d'un arbre fournit une partition, ayant d'autant moins de classes et des classes d'autant moins homogènes que l'on coupe plus haut dans la hiérarchie.

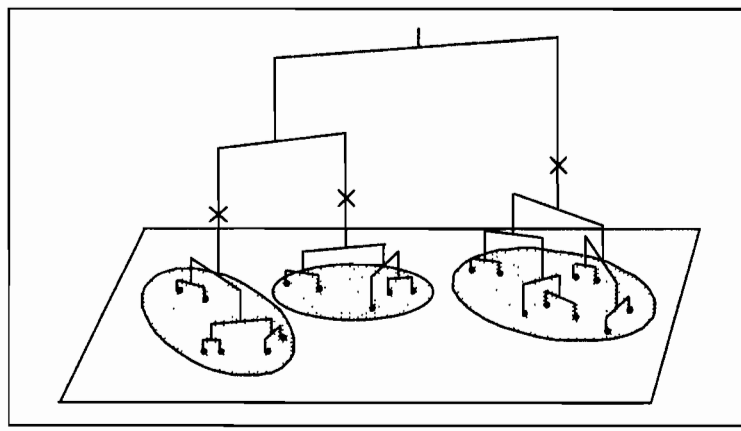


Figure 6.2 – 1. Dendrogramme ou arbre hiérarchique

a _ Distances entre éléments et entre groupes

On suppose au départ que l'ensemble des individus à classer est muni d'une *distance*[1]. Ceci ne suppose donc pas que les distances soient toutes calculées au départ : il faut pouvoir les calculer ou les recalculer à partir des coordonnées des points-individus, celles-ci devant être accessibles rapidement. On construit alors une première matrice de distances entre tous les individus.

---

[1] Il s'agira parfois simplement d'une mesure de dissimilarité. Dans ce cas, l'inégalité triangulaire $d(x,y) \leq d(x,z) + d(y,z)$ n'est pas exigée.

Une fois constitué un groupe d'individus, il convient de se demander ensuite sur quelle base on peut calculer une distance entre un individu et un groupe et par la suite une distance entre deux groupes.

Ceci revient à définir une stratégie de regroupements des éléments, c'est-à-dire se fixer des *règles de calcul des distances entre groupements* disjoints d'individus, appelées *critères d'agrégation.* Cette distance entre groupements pourra en général se calculer directement à partir des distances des différents éléments impliqués dans le regroupement.

Par exemple, si x, y, z sont trois objets, et si les objets x et y sont regroupés en un seul élément noté h, on peut définir la distance de ce groupement à z par la plus petite distance des divers éléments de h à z :

$$d(h,z) = Min \{d(x,z), d(y,z)\}$$

Cette distance s'appelle le *saut minimal (single linkage)* (Sneath, 1957 ; Johnson, 1967) et constitue un critère d'agrégation.

On peut également définir la distance du *saut maximal* (ou diamètre) en prenant la plus grande distance des divers éléments de h à z :

$$d(h,z) = Max \{d(x,z), d(y,z)\}$$

Une autre règle simple et fréquemment employée est celle de la *distance moyenne ;* pour deux objets x et y regroupés en h :

$$d(h,z) = \frac{\{d(x,z) + d(y,z)\}}{2}$$

Plus généralement, si x et y désignent des sous-ensembles disjoints de l'ensemble des objets, ayant respectivement n_x et n_y éléments, h est alors un sous-ensemble formé de $n_x + n_y$ éléments et on définit :

$$d(h,z) = \frac{\{n_x d(x,z) + n_y d(y,z)\}}{n_x + n_y}$$

b _ Algorithme de classification

L'algorithme fondamental de classification ascendante hiérarchique se déroule de la façon suivante :

- Étape 1 : il y a n éléments à classer (qui sont les n individus);
- Étape 2 : on construit la matrice de distances entre les n éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à n-1 classes;
- Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement (n-1) éléments à

classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec *n-2* classes et qui englobe la première;

- Étape *m* : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

Nous illustrons cette procédure en prenant comme objets à classer cinq points (figure 6.2 - 2).

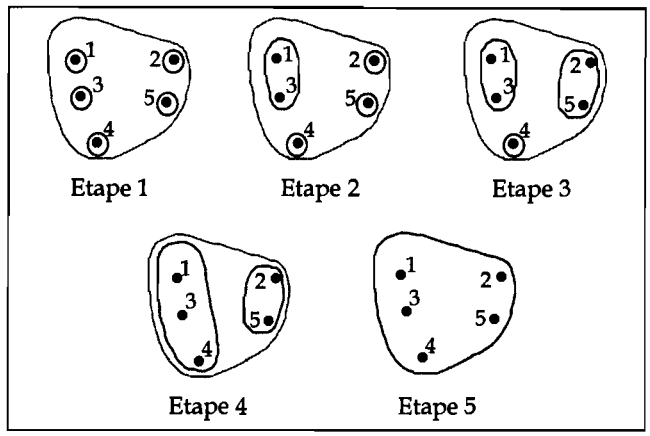


Figure 6.2 – 2. Agglomération progressive de 5 points

Les regroupements successifs peuvent être représentés par un arbre ou dendrogramme, comme le montre la figure 6.2 - 3 où l'on a porté en ordonnée les valeurs des indices ou encore distances correspondant aux différents niveaux d'agrégation.

c _ Eléments de vocabulaire

Quelques remarques vont nous permettre d'introduire les notions et la terminologie habituellement utilisées en classification ascendante hiérarchique. Le fonctionnement de l'algorithme nous montre que les distances (avec ces règles de calcul) n'interviennent que par les *inégalités* qui existent entre elles. Le même arbre (à une dilatation près des ordonnées) aurait été obtenu à partir d'un simple classement des couples d'objets dans l'ordre des distances croissantes. Un tel classement s'appelle une *ordonnance* (une *préordonnance* s'il y a des distances égales). Dans ce cas on tracera conventionnellement l'arbre avec des niveaux équidistants.

La famille H des parties de l'ensemble I des objets construite à partir d'algorithmes ascendants forme ce que l'on appelle une *hiérarchie.*

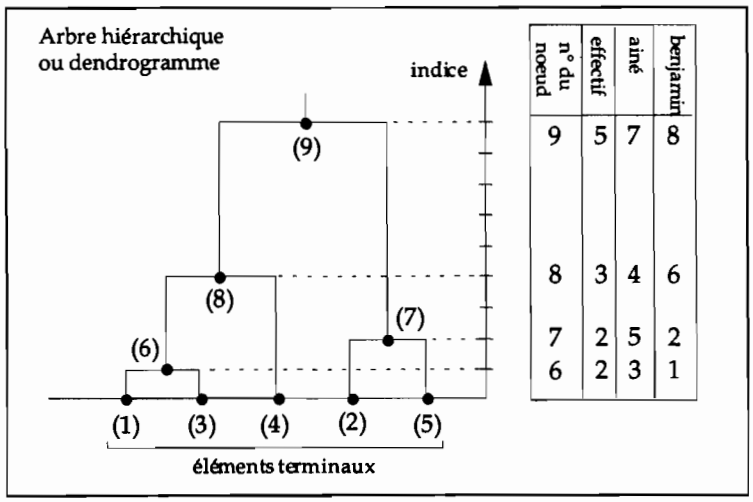
Cette famille a pour propriété de contenir l'ensemble tout entier ($I \in H$) ainsi que chacun des objets pris isolément ($i \in I \Rightarrow \{i\} \in H$).

Cette famille a pour propriété de contenir l'ensemble tout entier ($I \in H$) ainsi que chacun des objets pris isolément ($i \in I \Rightarrow \{i\} \in H$).

Les autres couples de parties h, h' de H sont alors soit disjointes ($h \cap h' = \varnothing$), soit incluses l'une dans l'autre ($h \subset h'$). En effet lors du fonctionnement de l'algorithme, chaque fois qu'une classe se forme à partir d'éléments disjoints, elle est elle-même considérée comme un nouvel élément, donc strictement incluse dans une classe ultérieure (cf. figure 6.2 - 2).

Les objets ou individus (1, 2, 3, 4, 5) sont les *éléments terminaux* de l'arbre (ou de la hiérarchie). Les classes 6, 7, 8, 9 sont les nœuds de l'arbre : ce sont des classes issues de regroupements de deux éléments (terminaux ou non) numérotés à la suite des éléments terminaux et dont chacune détermine une nouvelle partition.

On appelle arbitrairement *aîné* et *benjamin*, les deux éléments groupés constituant un nœud (cf. figure 6.2 - 3).



n° du noeud	effectif	aîné	benjamin
9	5	7	8
8	3	4	6
7	2	5	2
6	2	3	1

Figure 6.2 – 3 : Arbre hiérarchique et éléments de vocabulaire

On a une *hiérarchie indicée* si à toute partie h de la hiérarchie est associée une valeur numérique $v(h) \geq 0$ compatible avec la relation d'inclusion au sens suivant:

$$\text{si } h \subseteq h' \text{ alors } v(h) \leq v(h')$$

La hiérarchie de la figure 6.2 - 3 est indicée de façon naturelle par les valeurs des distances correspondant à chaque étape d'agrégation (ces distances sont portées en ordonnées). L'indice est la distance déterminant le regroupement.

En "coupant" l'arbre de la figure 6.2 - 3 par une droite horizontale, on obtient une partition, d'autant plus fine que la section est proche des éléments terminaux.

Si par exemple l'indice est supérieur à 4 et inférieur à 9, on obtient une partition en deux classes {1,3,4} et {2,5}. S'il vaut 3, on obtient trois classes {1,3}, {4} et {2,5}. Une hiérarchie permet donc de fournir une chaîne de n partitions ayant de 1 à n classes.

6.2.2 Classification ascendante selon le saut minimal et arbre de longueur minimale

Ce mode de classification hiérarchique, présenté lors de l'illustration du paragraphe précédent, est particulièrement simple à mettre en œuvre et possède des propriétés intéressantes que nous allons énoncer et étudier.

a _ Définition d'une ultramétrique

Nous allons montrer que la notion de hiérarchie est étroitement liée à une certaine classe de distances entre individus, que l'on appelle les *distances ultramétriques*. Pour la hiérarchie produite par l'algorithme du saut minimal, on montrera que la distance ultramétrique correspondante est, dans un certain sens, la plus proche de la distance initiale. Ce sera l'*ultramétrique inférieure maximale*, appelée encore *sous-dominante*. On montrera ensuite que l'application de cette méthode est pratiquement équivalente à la résolution d'un problème classique de recherche opérationnelle : la mise en évidence de l'*arbre de longueur minimale* sur un graphe.

Rappelons qu'un ensemble E est muni d'une *métrique* ou *distance d*, si d est une application à valeurs positives ou nulles obéissant aux conditions suivantes :

1. $d(x,y) = 0$ si et seulement si $x = y$.
2. $d(x,y) = d(y,x)$ (symétrie)
3. $d(x,y) \leq d(x,z) + d(y,z)$ (inégalité triangulaire)

Cette distance sera dite *ultramétrique* si elle vérifie la condition suivante, plus forte que l'inégalité triangulaire :

4. $d(x,y) \leq Max\ \{\ d(x,z), d(y,z)\ \}$

b _ Équivalence entre ultramétrique et hiérarchie indicée

Il est équivalent de munir un ensemble fini E d'une ultramétrique ou de définir une hiérarchie indicée de parties de cet ensemble.

Montrons tout d'abord que toute hiérarchie indicée permet de définir une distance entre éléments ayant les propriétés requises. On prendra comme distance $d(x,y)$ la valeur de l'indice correspondant à la plus petite partie contenant à la fois x et y.

En remplissant ainsi le tableau des valeurs de d correspondant à la hiérarchie de la figure 6.2 - 3, on obtient la matrice des distances du tableau 6.2 - 1.

On peut noter que l'inégalité 4 ci-dessus est vérifiée par toutes les distances de ce tableau. Ainsi par exemple : $d(1,2) \leq Max \{ d(1,5), d(2,5) \}$

Tableau 6.2 1. Matrice des distances

	(1)	(2)	(3)	(4)	(5)
(1)	0	9	1	4	9
(2)	9	0	9	9	2
(3)	1	9	0	4	9
(4)	4	9	4	0	9
(5)	9	2	9	9	0

Montrons plus généralement que l'on a toujours :

$$d(x,y) \leq Max \{ d(x,z) + d(y,z) \}$$

Rappelons que deux parties de la hiérarchie H sont soit disjointes, soit liées par une relation d'inclusion.

Appelons $h(x, z)$ la plus petite partie de H contenant x et z (dont l'indice est par conséquent $d(x, z)$). Puisque $h(x, z)$ et $h(y, z)$ ne sont pas disjointes, on a par exemple $h(x, z) \subset h(y, z)$. Et x, y, z étant tous trois contenus dans $h(y, z)$, on a obligatoirement :

$$h(x, y) \subset h(y, z) \quad \text{d'où} \quad d(x,y) \leq d(y,z)$$

ce qui établit l'inégalité.

Réciproquement, à toute ultramétrique d on peut faire correspondre une hiérarchie indicée dont d est l'indice associé. Il suffit d'appliquer l'algorithme du saut minimal au tableau des distances correspondant. On s'aperçoit alors qu'il est inutile de procéder au calcul des distances à chaque étape : il suffira de rayer l'un des deux éléments agrégés.

En effet, si x et y sont agrégés en t, il faut en principe calculer les distances au nouvel élément t (cf. figure 6.2 - 4). Or on a obligatoirement, pour tout élément z non encore agrégé, $d(z,x) \geq d(x,y)$ et $d(z,y) \geq d(x,y)$, sinon (z,x) ou (z,y) auraient été agrégés à la place de (x,y).

Pour une ultramétrique, cela implique à la fois $d(z,x) \geq d(z,y)$, et $d(z,y) \geq d(z,x)$ c'est-à-dire $d(z,x) = d(z,y)$, ce que l'on exprime de façon imagée en disant que, pour une ultramétrique, tous les triangles sont isocèles, avec le plus petit coté pour base (figure 6.2 - 4).

Il est en effet facile de montrer que si une distance est ultramétrique, tous les triangles sont isocèles.

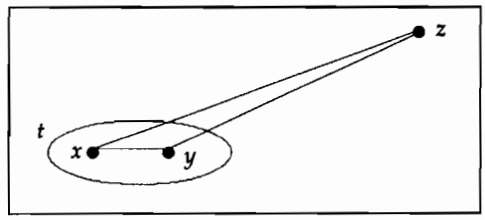


Figure 6.2 – 4. Agrégation de x et y en un nouvel élément t

On a les inégalités :

$$d(z, x) \leq Max\ \{\ d(x,y), d(y,z)\ \} \quad \text{donc} \quad d(z, x) \leq d(y, z)$$

De la même façon :

$$d(y, z) \leq Max\ \{\ d(x,y),\ d(z, x)\ \} \quad \text{donc} \quad d(y, z) \leq d(z, x)$$

Il s'ensuit que :

$$d(z, x) = d(y, z)$$

Le calcul des distances de z à t est finalement inutile puisque les deux distances mises en cause sont égales. Ceci nous montre comment l'algorithme du saut minimal a opéré sur la matrice des distances : il a transformé la métrique initiale en ultramétrique en diminuant certaines distances à chaque étape.

c _ L'ultramétrique sous dominante

Le passage d'une métrique à une ultramétrique (ou, de façon équivalente, à une hiérarchie) s'est effectué par diminution des valeurs de certaines distances. On peut se poser la question suivante : existe-t-il une ultramétrique plus proche (en un sens à préciser) de la métrique ?

On peut donner l'élément de réponse suivant.

On dira qu'une métrique d_1 est inférieure[1] à une métrique d_2 si, pour tout x et tout y :

$$d_1(x\ y) \leq d_2(x, y)$$

La plus grande ultramétrique inférieure à une métrique d, au sens précédent, est appelée ultramétrique inférieure maximale ou sous-dominante. C'est l'ultramétrique sous-dominante qui est fournie par l'algorithme du saut minimal.

Pour le démontrer nous allons successivement :

1. définir, à partir d'une distance d, une nouvelle distance dite du plus petit saut maximal;

---

[1] Cette définition permet de munir l'ensemble des métriques définies sur un ensemble E d'une relation d'ordre partiel.

2. montrer que cette distance est une ultramétrique;
3. montrer que cette ultramétrique est la sous-dominante;
4. montrer enfin que cette distance correspond à l'ultramétrique fournie par l'algorithme du saut minimal.

1. *La distance du plus petit saut maximal :*
Soit un ensemble E muni d'une distance d. Soit x et y deux éléments de E. Le couple (x,y) sera appelé *arête* de longueur $d(x,y)$ du *graphe complet*[1] dont les sommets sont les éléments de E. Toujours en utilisant le vocabulaire de la théorie des graphes, on appelle *chemin* de x à y une succession d'arêtes de types (x, t_1), (t_1, t_2), $(t_2, t_3,)$ …, (t_{k-1}, t_k), (t_k, y), où $t_1, \ldots, t_k$ sont des éléments de E. Étant donné un chemin de x à y, on appelle *saut maximal* la longueur de la plus grande arête du chemin de x à y.

A tout chemin joignant x à y correspond un saut maximal. L'ensemble des sommets étant fini, il existe un *plus petit saut maximal* sur l'ensemble des chemins allant de x à y ; nous le noterons $d^*(x,y)$.

2. *Le plus petit saut maximal entre x et y est une ultramétrique :*
Il est clair que les deux premiers axiomes d'une distance sont vérifiés par d^*. Pour vérifier que cette distance est une ultramétrique, considérons trois éléments quelconques x, y, z de E (figure 6.2 - 5). Le plus petit saut maximal de x à y, en s'astreignant à passer par z est $Max\ \{d^*(x, z), d^*(z, y)\}$. Le plus petit saut maximal de x à y sans la contrainte de passer par z ne peut qu'être inférieur ou égal à cette quantité, d'où :

$$d^*(x,y) \leq Max\ \{d^*(x,z), d^*(y,z)\}$$

et d^* est donc bien une ultramétrique.

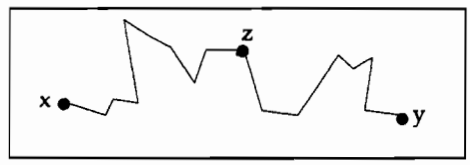


Figure 6.2 – 5 : Chemin de x à y contenant z

3. *La distance d^* est la sous-dominante :*

Pour montrer que d^* est la sous-dominante, on montrera que d^* est inférieure à d, et que d^* est supérieure à toute ultramétrique inférieure à d.

Tout d'abord, il est clair que l'arête (x,y) est un chemin particulier allant de x à y, donc $d^*(x,y) \leq d(x,y)$ et d^* est inférieure à d.

---

[1] L'appellation *graphe complet* est due au fait que tout couple de sommets est joint par une arête.

Soit maintenant d_1 une ultramétrique inférieure à d. On a évidemment pour tout triplet x_1, x_2, x_3 :

$$d_1(x_1,x_3) \leq Max\ \{d_1(x_1, x_2), d_1(x_2, x_3)\}$$

En appliquant de façon successive cette inégalité à un chemin :

$$(x_1, x_2),\ (x_2, x_3), ..., (x_{p-1}, x_p)$$

on obtient :

$$d_1(x_1,x_p) \leq \underset{j<p}{Max}\{d_1(x_j,x_{j+1})\}$$

Puisque $d_1 \leq d$, on a :

$$d_1(x_1,x_p) \leq \underset{j<p}{Max}\{d(x_j,x_{j+1})\}$$

Cette inégalité est valable pour tout chemin joignant x_1 à x_p. Pour l'un au moins d'entre eux, on a par définition de d^* :

$$\underset{j<p}{Max}\{d(x_j,x_{j+1})\} = d^*(x_j,x_{j+1})$$

Cette dernière relation établit l'inégalité annoncée.

4. *La distance ultramétrique d_u produite par l'algorithme du saut minimal n'est autre que la distance d^* du plus petit saut maximal :*

   Soit $d_u(x, y)$ la valeur de la distance à l'étape où les points x et y sont réunis pour la première fois. Auparavant ces deux points étaient dans des classes distinctes (éventuellement réduites aux points eux-mêmes).

   Le mode de calcul des distances à chaque agrégation assure que $d_u(x, y)$ est la plus petite distance entre deux éléments appartenant chacun à une classe : les distances à l'intérieur des classes sont inférieures à $d_u(x, y)$ puisque l'agrégation est antérieure ; les distances avec des éléments n'appartenant pas aux deux classes sont supérieures puisque ceux-ci seront agrégés à une étape ultérieure.

   Les chemins joignant x et y auront donc des arêtes internes aux deux classes, de longueur inférieure à $d_u(x, y)$ et des arêtes externes nécessairement supérieures ou égales à $d_u(x, y)$. Ainsi $d_u(x, y)$ est bien le plus petit saut maximal $d^*(x, y)$.

d _ Arbre de longueur minimale : définition et généralités

L'ensemble des n objets à classer peut être considéré comme un ensemble de points d'un espace. Cette représentation est classique si les objets sont décrits par une série de p variables : on a n points dans l'espace $\mathcal{R}^p$. On peut alors calculer une distance pour chaque paire de points.

Plus généralement, si l'on ne dispose que des valeurs d'un indice de dissimilarité (ne vérifiant pas obligatoirement tous les axiomes d'une distance), on peut représenter les objets par des points (d'un plan par exemple), chaque

couple d'objets étant joint par une ligne continue, à laquelle est attachée la valeur de l'indice de dissimilarité.

On représente ainsi l'ensemble des objets et des valeurs de l'indice par un *graphe complet valué*[1]. Mais si le nombre d'objets dépasse quelques unités, ce type de représentation devient inextricable. On cherchera alors à extraire de ce graphe un *graphe partiel* (ayant les mêmes sommets, mais moins d'arêtes) plus aisé à représenter, et permettant néanmoins de bien résumer les valeurs de l'indice.

Parmi tous les graphes partiels, ceux qui ont une structure d'*arbre*[2] sont particulièrement intéressants, car ils peuvent faire l'objet d'une représentation plane.

Un arbre est un *graphe connexe* (il existe un chemin reliant tout couple de sommets) *sans cycle* (un cycle est un chemin partant et aboutissant au même point sans emprunter deux fois la même arête). On peut définir de façon équivalente un arbre à n sommets soit comme un graphe sans cycle ayant $n - 1$ arêtes, soit comme un graphe connexe ayant $n - 1$ arêtes[3].

La *longueur* d'un arbre sera la somme des "longueurs" (valeurs de l'indice) de ses arêtes. Parmi tous les graphes partiels qui sont des arbres, l'*arbre de longueur minimale* a retenu depuis longtemps l'attention des statisticiens en raison de ses bonnes qualités descriptives, qui ne sont pas étrangères à sa parenté avec les classifications hiérarchiques. Si l'on désire par exemple déceler rapidement sans ordinateur les traits de structure que peut cacher une matrice de corrélations relative à une trentaine de variables, c'est probablement la plus aisée des procédures à mettre en œuvre.

Nous allons tout d'abord présenter les algorithmes de recherche de l'arbre de longueur minimale, puis nous montrerons les équivalences avec la classification selon le saut minimal. Nous supposerons que toutes les arêtes du graphe ont des longueurs différentes (valeurs de l'indice ou de la distance) car dans ces conditions l'arbre cherché est unique et ceci simplifie l'exposé des algorithmes.

e _ Arbre de longueur minimale : algorithme de Kruskal (1956)

On range les $n(n - 1)/2$ arêtes dans l'ordre des valeurs croissantes de l'indice. On part des deux premières arêtes, puis on sélectionne successivement toutes les arêtes qui ne font pas de cycle avec les arêtes déjà choisies. On interrompt la

---

[1] Les objets à classer sont alors les nœuds du graphe (non orienté); les lignes continues joignant les paires de points sont les arêtes; et les indices, les valuations de ces arêtes.

[2] On ne confondra pas un tel arbre, entendu au sens de la théorie des graphes, et dont les sommets sont les objets à classer, avec l'arbre des parties d'un ensemble (dendrogramme) produit par les techniques de classification hiérarchique, dont les sommets sont des parties (à l'exception des éléments terminaux qui sont les objets à classer eux-mêmes).

[3] On trouvera la démonstration de ces propriétés dans les manuels classiques tels que ceux de Berge (1963, 1973).

procédure dès que l'on a $n - 1$ arêtes. De cette façon, on est sûr d'avoir obtenu un arbre (graphe sans cycle ayant $n - 1$ arêtes).

Montrons en effet que si V_k dénote le graphe obtenu à l'étape k, après avoir sélectionné les arêtes v_1, v_2, ..., v_k, alors V_{n-1} est de longueur minimale. Supposons qu'il existe un arbre distinct U, de longueur minimale (figure 6.2 - 6).

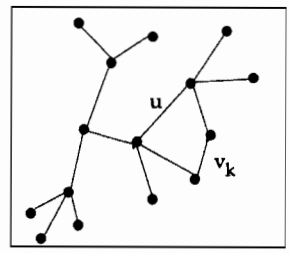


Figure 6.2 – 6 : Représentation de l'arbre U

Soit v_k la première arête sélectionnée dans la construction de V_{n-1} et qui n'appartienne pas à U (les arêtes de V_{k-1} sont donc également des arêtes de U).

En ajoutant cette arête à U on crée nécessairement un cycle (car U est connexe) et un seul (car U est sans cycle). Il existe donc une arête u de ce cycle qui n'appartient pas à V_{n-1} (puisque V_{n-1} n'a pas de cycle). Alors l'arbre U* obtenu à partir de U en ajoutant v_k et en supprimant u est plus court que U. En effet, le graphe obtenu en ajoutant u à V_{k-1} est sans cycle (c'est une partie de U); donc u est plus long que v_k, par définition de v_k, et par conséquent U* est plus court que U. Mais ceci contredit la définition de U. Donc V_{n-1} est bien de longueur minimale.

f _ Arbre de longueur minimale : algorithme de Prim (1957)

On part d'un objet quelconque (sommet du graphe). L'étape *1* consiste à chercher l'objet v_1 le plus proche, c'est-à-dire l'arête la plus courte. L'étape k consiste à adjoindre au recueil d'arêtes déjà constitué V_{k-1} la plus courte arête v_k qui touche un des sommets de V_{k-1}. Il y a $n-1$ étapes. Cet algorithme est plus rapide que le précédent. L'arbre obtenu est de longueur minimale car V_k est à tout moment un arbre de longueur minimale sur les k sommets concernés.

g _ Arbre de longueur minimale : algorithme de Florek (1951)[1]

A la première étape, on joint chaque sommet à son voisin le plus proche. Cela revient à prendre la plus petite distance dans chaque ligne du tableau des distances. Cette opération rapide produit une forêt F_1 (famille d'arbres, c'est-à-

[1] Cet algorithme, à l'origine de la *Wroclaw taxonomy* développée par Florek, est plus ancien et plus performant que ceux de Kruskal et Prim, mais il est assez rarement cité. Cf. aussi : Graham et Hell (1985) pour une histoire (passionnante) de cet algorithme.

dire simplement : graphe sans cycle). A l'étape k, chaque arbre de la forêt F_{k-1} (chaque composante connexe du graphe sans cycle) est joint à son plus proche voisin en prenant comme distance entre arbres la plus petite distance entre un sommet quelconque de l'un et un sommet quelconque de l'autre. Le processus s'arrête dès que le graphe F_k est connexe.

Cet algorithme est plus rapide à mettre en œuvre manuellement sur des tableaux de distances assez grands. En général, il n'y a que 2 ou 3 étapes.

Montrons que l'on obtient un arbre, ce qui se ramène à prouver que la première étape fournit bien une forêt.

Il n'y a pas de sommet isolé car chaque sommet admet effectivement un plus proche voisin.

Montrons par l'absurde que l'on ne peut pas créer de cycle.

Supposons qu'il en existe un et orientons les arêtes de chaque sommet vers son plus proche voisin. Si les arêtes du cycle sont toutes orientées dans le même sens, le résultat est absurde, car celles-ci seraient nécessairement de plus en plus courtes.

Sinon la figure serait également absurde, car deux arêtes partiraient d'un même sommet, alors que chaque sommet n'a qu'un seul plus proche voisin.

Il reste à montrer que cet arbre est de longueur minimale. Notons que toute arête tracée à la première étape appartient à l'arbre de longueur minimale V. En effet, s'il n'en était pas ainsi, il existerait y, plus proche voisin de x, tel que l'arête (x,y) n'appartienne pas à V. En ajoutant cette arête à V, on crée un cycle. En supprimant l'autre arête du cycle issue de x, on obtient un nouvel arbre plus court que V, ce qui contredit la définition de V.

De la même façon, toute arête tracée à l'étape k appartient à V, sachant que la forêt F_{k-1} est une partie de V. Le raisonnement est en tout point analogue au précédent.

h_ Exemple d'application

La figure 6.2-7 montre, à titre pédagogique, un exemple de la représentation de cet arbre dans le plan factoriel principal (plan des deux premiers axes principaux) d'une analyse en composantes principales des données sémiométriques présentées au paragraphe 3.5.2-b.

La distance est calculée dans le plan, c'est-à-dire dans l'espace des deux premiers axes factoriels.

Cet arbre pourrait être tracé à la main, car les distances utilisées pour le calculer sont celles lisibles directement dans le plan du graphique.

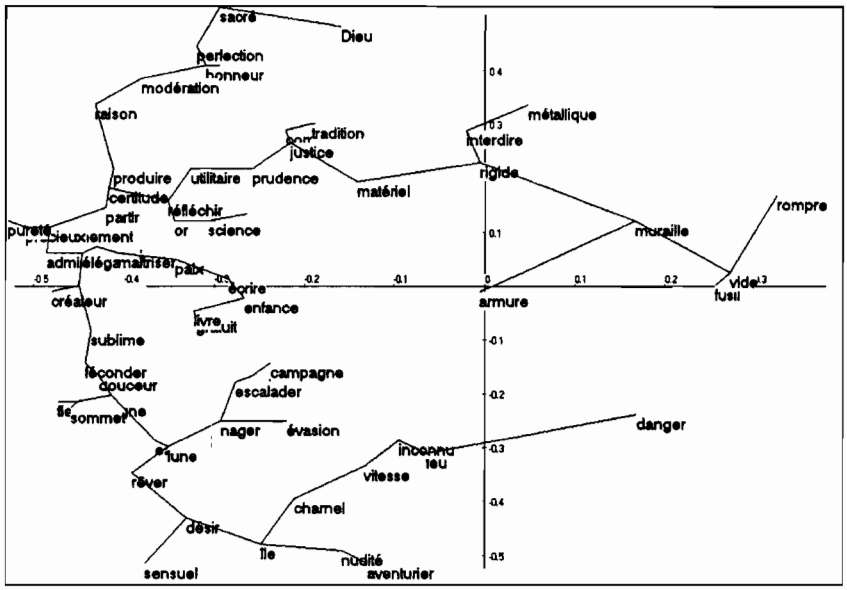


Figure 6.2 – 7. Représentation d'un arbre de longueur minimale dans le plan factoriel principal des données sémiométriques (arbre calculé avec 2 axes principaux)

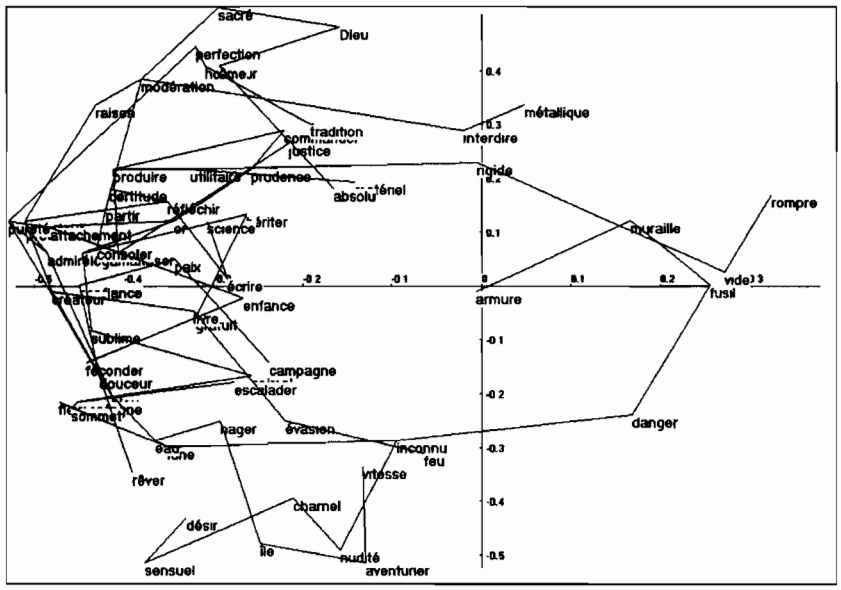


Figure 6.2 – 8. Représentation d'un arbre de longueur minimale dans le même plan principal (arbre calculé cette fois dans l'espace des 12 premiers axes principaux)

La figure 6.2-8 montre, dans le même plan principal, un arbre de longueur minimale à partir des distances entre mots calculées cette fois dans l'espace des 12 premiers axes factoriels. Cet arbre apporte donc une information non visible dans le plan et constitue un complément pour la représentation plane.

Il est intéressant de constater, par exemple, que les points *vide* et *fusil*, proches sur le plan (1,2) mais aussi faiblement corrélés au plan, sont en fait éloignés dans l'espace $\mathcal{R}^{12}$, *fusil* étant plus proche du mot *danger*.

De la même façon, dans l'espace à 12 dimensions (figure 6.2-8), l'arbre relie d'une part les mots : *désir, sensuel, charnel, nudité*, et dans une branche très distincte : *nage, île, aventurier, vitesse*. Ces deux groupes étaient entremêlés dans l'espace à deux dimensions, et indûment reliés sur l'arbre de longueur minimale construit dans cet espace (figure 6.2-7).

i _ Lien entre l'arbre et le saut minimal (Gower et Ross, 1969)

Soit V un arbre de longueur minimale construit à partir du tableau des distances entre n objets. V étant connexe et n'ayant pas de cycle, il existe un chemin et un seul joignant deux sommets x et y. Appelons $d_v(x, y)$ la longueur de la plus grande arête rencontrée sur ce chemin. Nous allons montrer que $d_v(x, y)$ n'est autre que $d^*(x, y)$, la distance ultramétrique du plus petit saut maximal entre x et y.

En effet, soit v la plus grande arête rencontrée entre x et y. La suppression de v entraîne la division de V en deux composantes connexes séparées. S'il existe un chemin (n'empruntant pas obligatoirement des arêtes de V) de x à y dont la plus grande arête est plus courte que v, il existe une arête u distincte de v, et plus courte qui joint les deux composantes connexes. Le fait de remplacer v par u donnerait un arbre de longueur inférieure à celle de V, ce qui contredit la définition de V. Ainsi $d_v(x, y)$, longueur de v, est bien le plus petit saut maximal.

Le raisonnement fournit un mode de construction de la hiérarchie associée au saut minimal, à partir de l'arbre de longueur minimale V. Cette construction, descendante, s'opère de la façon suivante. On rompt la plus grande arête de V; on obtient ainsi les deux groupes les plus éloignés, l'indice correspondant à leur fusion étant la longueur de cette arête. On rompt ensuite successivement les arêtes par ordre de grandeur décroissantes, ce qui fait descendre dans la hiérarchie jusqu'aux éléments terminaux qui sont les objets eux-mêmes. La dernière arête rompue correspond aux deux objets agrégés en premier dans l'algorithme ascendant.

On peut représenter simultanément la hiérarchie et l'arbre de longueur minimale en perspective comme le montre la figure 6.2 - 9. Cette représentation qualifiée de « squelette arborescent » a été proposée par Benzécri et Jambu, (1976). Quelques informations complémentaires sont apportées à la représentation de la figure 6.2 - 3. En particulier les positions relatives des points sont mieux respectées.

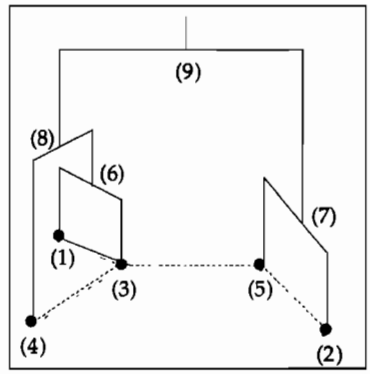


Figure 6.2 – 9. Représentation simultanée de la hiérarchie et de l'arbre de longueur minimale

Nous avons vu à propos de l'exemple que pour le praticien de l'analyse factorielle, il sera souvent intéressant de porter l'arbre de longueur minimale sur les plans factoriels de façon à remédier, dans une certaine mesure, aux possibles déformations imputables à l'opération de projection.

6.2.3 Critère d'agrégation selon la variance

Les techniques de classification selon le saut minimal ont l'avantage de conduire à des calculs simples (pas de recalcul numérique des distances) et possèdent des propriétés mathématiques intéressantes.

Pour certaines applications les résultats sont cependant critiquables. En particulier, le saut minimal a le défaut de produire des "effets de chaîne".

Ainsi pour le nuage de points représenté par la figure 6.2 - 10 les groupes A et B ne seront pas facilement discernables dans l'arbre hiérarchique; de plus, les quelques sommets qui les relient seront agrégés au niveau le plus bas.

D'autres critères d'agrégation donnent éventuellement des résultats plus fiables, par exemple la distance moyenne (cf. également Wishart, 1969).

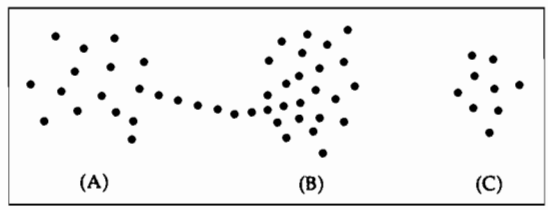


Figure 6.2 – 10 : "Effets de chaîne"

Les techniques d'agrégation selon la variance cherchent à optimiser, à chaque étape, selon des critères liés à des calculs d'inertie, la partition obtenue par agrégation de deux éléments. Cette technique est particulièrement aisée à mettre en œuvre lorsque l'agrégation est effectuée après une analyse factorielle, les objets à classer étant repérés par leurs coordonnées sur les premiers axes factoriels.

a _ Notations et principe

Nous considérons ici les n objets à classer comme un nuage de points (le nuage des individus) d'un espace à p dimensions (espace des variables).

Chaque point $\mathbf{x}_i$ (vecteur à p composantes) est muni d'une masse m_i . On note m la masse totale du nuage :

$$m = \sum_{i=1}^{n} m_i$$

Le carré de la distance entre les points $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ est notée :

$$\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 = d^2(\mathbf{x}_i, \mathbf{x}_{i'})$$

L'inertie totale I du nuage est la quantité :

$$I = \sum_{i}^{n} m_i \|\mathbf{x}_i - \mathbf{g}\|^2$$

où $\mathbf{g}$ désigne le centre de gravité du nuage :

$$\mathbf{g} = \frac{1}{m} \sum_{i}^{n} m_i \mathbf{x}_i$$

S'il existe une partition de l'ensemble des éléments en s classes, la $q^{ième}$ classe a pour masse :

$$m_q = \sum_{i \in q} m_i$$

et pour centre de gravité :

$$\mathbf{g}_q = \frac{1}{m_q} \sum_{i \in q} m_i \mathbf{x}_i$$

La relation de Huygens fournit une décomposition de la quantité I en inerties intra-classes et inter-classes suivant la formule :

$$I = \sum_{q} m_q \|\mathbf{g}_q - \mathbf{g}\|^2 + \sum_{q} \sum_{i \in q} m_i \|\mathbf{x}_i - \mathbf{g}_q\|^2 \qquad [6.2 - 1]$$

La qualité globale d'une partition est liée à l'homogénéité à l'intérieur des classes (et donc à l'écartement entre les classes). I étant une quantité constante, il s'agit par conséquent de minimiser la quantité relative à l'inertie intra-classes :

$$I_{intra} = \sum_{q} \sum_{i \in q} m_i \|\mathbf{x}_i - \mathbf{g}_q\|^2$$

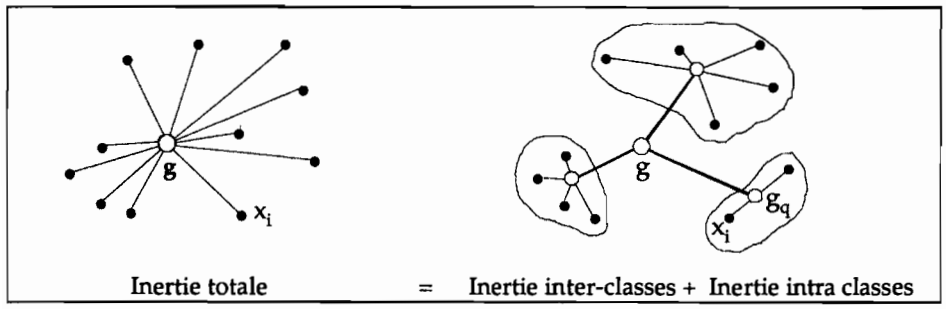


Figure 6.2 – 11 : Décomposition de l'inertie selon la relation de Huygens

soit encore à maximiser celle relative à l'inertie inter-classes :

$$I_{inter} = \sum_q m_q \|\mathbf{g}_q - \mathbf{g}\|^2$$

A l'étape initiale, l'inertie intra-classes est nulle et l'inertie inter-classes est égale à l'inertie totale du nuage puisque chaque élément terminal constitue à ce niveau une classe.

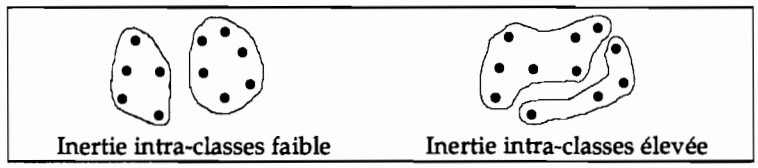


Figure 6.2 – 12 : Qualité globale d'une partition

A l'étape finale, c'est l'inertie inter-classes qui est nulle et l'inertie intra-classes est équivalente à l'inertie totale puisque l'on dispose à ce niveau d'une partition en une seule classe (cf. l'étape 5 de la figure 6.2 - 2). Par conséquent, au fur et à mesure que l'on effectue des regroupements, l'inertie intra-classes augmente et l'inertie inter-classes diminue. Le principe de l'algorithme d'agrégation selon la variance consiste à rechercher à chaque étape une partition telle que la variance interne de chaque classe soit minimale et par conséquent la variance entre les classes soit maximale.

b _ Perte d'inertie par agrégation de deux éléments : le critère de Ward généralisé

Faire varier le moins possible l'inertie intra-classes à chaque étape d'agrégation revient à rendre minimale la perte d'inertie inter-classes résultant de l'agrégation de deux éléments (objets à classer ou classes).

Soit $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ deux éléments de masses m_i et $m_{i'}$ appartenant à une partition P_s à s classes, que l'on agrège en un seul élément $\mathbf{x}$ de masse $m_t = m_i + m_{i'}$, produisant la partition P_{s-1} à $s-1$ classes, avec :

$$\mathbf{x} = \frac{m_i \mathbf{x}_i + m_{i'} \mathbf{x}_{i'}}{m_i + m_{i'}}$$

$\mathbf{x}$ est le centre de gravité de $\mathbf{x}_i$ et $\mathbf{x}_{i'}$.

On peut décomposer l'inertie $I_{ii'}$ de $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ par rapport à $\mathbf{g}$ suivant la relation de Huygens :

$$I_{ii'} = m_i \|\mathbf{x}_i - \mathbf{g}\|^2 + m_{i'} \|\mathbf{x}_{i'} - \mathbf{g}\|^2 = m_i \|\mathbf{x}_i - \mathbf{x}\|^2 + m_{i'} \|\mathbf{x}_{i'} - \mathbf{x}\|^2 + \mathrm{m}_t \|\mathbf{x} - \mathbf{g}\|^2$$

Seul le dernier terme subsiste si $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ sont remplacés par leur centre de gravité $\mathbf{x}$. La perte d'inertie inter-classes $\Delta I_{ii'}$ due au passage de la partition à s classes à la partition à $s - 1$ classes équivaut à :

$$\Delta_s = \Delta I_{ii'} = I_{inter(P_s)} - I_{inter(P_{s-1})}$$

et vaut donc :

$$\Delta I_{ii'} = m_i \|\mathbf{x}_i - \mathbf{x}\|^2 + m_{i'} \|\mathbf{x}_{i'} - \mathbf{x}\|^2$$

En remplaçant $\mathbf{x}$ par sa valeur en fonction de $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ il vient, tous calculs faits :

$$\Delta I_{ii'} = \frac{m_i m_{i'}}{m_i + m_{i'}} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 = \frac{m_i m_{i'}}{m_i + m_{i'}} d^2(\mathbf{x}_i, \mathbf{x}_{i'})$$

La stratégie d'agrégation fondée sur le critère de la perte d'inertie minimale, dit critère de Ward généralisé, est donc la suivante : au lieu de chercher les deux éléments les plus proches, on cherchera les éléments $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ correspondant à $\Delta I_{ii'}$ minimale. Ainsi à chaque étape l'inertie inter-classes diminue de la quantité $\Delta I_{ii'}$ (et l'inertie intra-classes augmente de cette même quantité). Ceci revient à considérer les $\Delta I_{ii'}$ comme de nouveaux indices de dissimilarités[1] appelés aussi "*indices de niveau*".

On vérifie que la somme des indices de niveau dans la hiérarchie est égale à l'inertie totale du nuage I :

$$\sum_{s=2}^{n} \Delta_s = \sum_{s=2}^{n} I_{inter(P_s)} - I_{inter(P_{s-1})} = I \qquad [6.2 - 2]$$

Si l'on travaille sur les coordonnées des points, on effectuera les calculs des centres de gravité ($\mathbf{x}$ pour $\mathbf{x}_i$ et $\mathbf{x}_{i'}$). Par contre si l'on travaille sur les distances, il est commode de pouvoir calculer les nouvelles distances à partir des anciennes (comme cela était le cas pour les techniques précédentes). Le carré des distances entre un point quelconque $\mathbf{z}$ et le centre de classe $\mathbf{x}$ s'écrit, en fonction des distances à $\mathbf{x}_i$ et $\mathbf{x}_{i'}$:

$$d^2(\mathbf{x}, \mathbf{z}) = \frac{1}{m_i + m_{i'}} \left(m_i d^2(\mathbf{x}_i, \mathbf{z}) + m_{i'} d^2(\mathbf{x}_{i'}, \mathbf{z}) - \frac{m_i m_{i'}}{m_i + m_{i'}} d^2(\mathbf{x}_i, \mathbf{x}_{i'}) \right)$$

---

[1] Par cette transformation de la matrice des distances, les points les plus légers seront plus facilement agrégés.

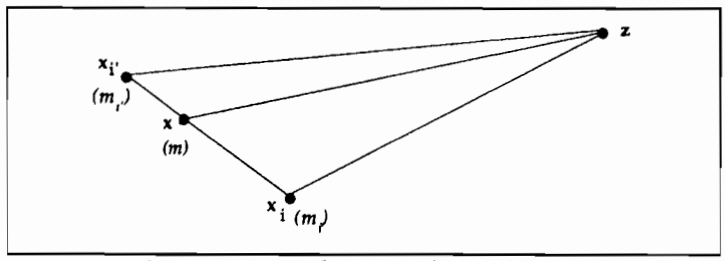


Figure 6.2 – 13 : Théorème de la médiane

Cette formule (théorème de la médiane) s'établit en décomposant l'inertie du doublet ($\mathbf{x}_i$, $\mathbf{x}_{i'}$) par rapport à $\mathbf{z}$ en inertie par rapport à $\mathbf{x}$, et en inertie de $\mathbf{x}$ par rapport à $\mathbf{z}$:

$$m_i \|\mathbf{x}_i - \mathbf{z}\|^2 + m_{i'} \|\mathbf{x}_{i'} - \mathbf{z}\|^2 = (m_i + m_{i'}) \|\mathbf{x} - \mathbf{z}\|^2 + \frac{m_i m_{i'}}{m_i + m_{i'}} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2$$

L'expression de $d^2(x, z)$ s'en déduit immédiatement. On réitère le processus sur les éléments restants et le nouvel élément construit par agrégation[1].

6.2.4 Algorithme de recherche en chaîne des voisins réciproques

La principale difficulté dans la construction d'un arbre hiérarchique est le nombre important d'opérations. A chaque étape de l'algorithme est construit un nœud regroupant deux éléments, ce qui nécessite des calculs et des comparaisons de distances entre tous les éléments restant à classer. Le nombre d'opérations à effectuer est de l'ordre de n^3 s'il y a n objets à classer.

Les nouveaux algorithmes réunissent à chaque étape non plus deux éléments mais plusieurs couples d'éléments. Ceci réduit considérablement le nombre des opérations qui passe de n^3 à n^2 permettant ainsi la classification de plusieurs milliers d'objets en un temps raisonnable.

Ces algorithmes utilisent le concept de voisins réciproques introduits par McQuitty (1966) : deux éléments $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ sont voisins réciproques si $\mathbf{x}_i$ est le plus proche voisin de $\mathbf{x}_{i'}$ et si $\mathbf{x}_{i'}$ est le plus proche voisin de $\mathbf{x}_i$.

---

[1] Il existe des variantes de cette méthode qui font appel à des formules de calcul légèrement différentes. On peut par exemple rechercher les classes ayant une inertie interne minimale; on peut aussi utiliser le critère de la variance interne minimale, en désignant par variance l'inertie divisée par la masse. On trouvera des précisions sur ces techniques dans Benzécri (1973).

Ils utilisent également la propriété d'une agrégation hiérarchique selon laquelle, à une étape donnée, deux éléments agrégés pour constituer un nœud sont nécessairement des voisins réciproques (sinon, ils ne constitueraient pas la paire à distance minimale). Enfin ils utilisent la propriété plus forte (valable seulement si le critère d'agrégation vérifie le critère de la médiane, explicité plus loin) selon laquelle tous les voisins réciproques, à une étape donnée, seront ultérieurement des nœuds de la hiérarchie[1]. A chaque étape de l'algorithme, au lieu d'agréger seulement les deux plus proches voisins, il y a donc autant de nœuds créés qu'il y a de voisins réciproques. A l'étape finale, tous les éléments sont regroupés en une seule classe et l'arbre est construit. Le problème de l'algorithme est alors ramené à une recherche efficace des voisins réciproques. Nous allons décrire l'algorithme de cette recherche qui s'effectue en chaîne (Benzécri, 1982c).

a _ Algorithme

Le principe des voisins réciproques peut s'énoncer de la manière suivante : si x_i est plus proche voisin de $x_{i'}$ ($x_i \rightarrow x_{i'}$) et si $x_{i'}$ est plus proche voisin de x_i ($x_{i'} \rightarrow x_i$) alors x_i et $x_{i'}$ sont voisins réciproques ($x_i \leftrightarrow x_{i'}$)

Etape 1 : on part d'un objet quelconque x_1 et on cherche son plus proche voisin, noté x_2 puis le plus proche voisin de x_2, noté x_3, etc.. On crée ainsi une chaîne d'éléments successifs:

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \ldots \rightarrow x_{i-2} \rightarrow x_{i-1} \rightarrow x_i \rightarrow \ldots$$

Une telle chaîne s'arrête nécessairement lorsque deux éléments successifs sont voisins réciproques :

$$\ldots \rightarrow x_i \rightarrow \ldots \rightarrow x_{k-1} \leftrightarrow x_k$$

La chaîne s'arrêtera ici sur l'élément x_k si x_{k-1} est aussi le plus proche voisin de x_k. x_{k-1} et x_k sont voisins réciproques et sont agrégés pour former un nœud.

Etape 2 : si $k = 2$ alors la chaîne commence avec un élément qui possède un voisin réciproque:

$$x_1 \leftrightarrow x_2$$

Nous choisissons un nouvel élément à partir duquel une chaîne est construite et qui s'arrête sur de nouveaux voisins réciproques dont l'agrégation fournit un nœud.

Etape 3 : si $k > 2$, on continue la recherche des voisins réciproques par extension de la chaîne commençant à l'élément x_{k-2}. L'algorithme se termine lorsque $n - 1$ nœuds ont été créés.

---

[1] Le critère de la médiane assure qu'ils resteront toujours voisins réciproques.

b _ Critère de la médiane

Afin de pouvoir utiliser cet algorithme, la chaîne doit pouvoir être prolongée au delà de $\mathbf{x}_{k-2}$ lorsque les voisins réciproques $\mathbf{x}_{k-1}$ et $\mathbf{x}_k$ ont été agrégés en $\mathbf{x}_{k-2}$. Il est indispensable que cette agrégation ne détruise pas la relation du voisin le plus proche qui existait au préalable entre $\mathbf{x}_{i-1}$ et $\mathbf{x}_i$ avec $i = 2, 3, ..., k-2$. Cette propriété est assurée si le critère d'agrégation utilisé pour construire l'arbre ne crée pas une inversion.

Il n'y a pas inversion si le nœud **n**, créé par agrégation de **a** et **b**, ne peut être plus près d'un quelconque autre élément **c** que ne le sont l'élément **a** ou l'élément **b**. Cette condition[1] dite de "la médiane" s'écrit:

$$\text{si } d(\mathbf{a}, \mathbf{b}) < inf\{ d(\mathbf{a}, \mathbf{c}), d((\mathbf{b}, \mathbf{c}) \} \text{ alors } inf\{d(\mathbf{a}, \mathbf{c}), d(\mathbf{b}, \mathbf{c})\} < d(\mathbf{n}, \mathbf{c})$$

Cette propriété est vérifiée par plusieurs critères d'agrégation[2] :

- Saut minimal : $d(\mathbf{a},\mathbf{b}) = inf\{ d(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \in \mathbf{a}, \mathbf{v} \in \mathbf{b} \}$
- Saut maximal : $d(\mathbf{a},\mathbf{b}) = sup\{d(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \in \mathbf{a}, \mathbf{v} \in \mathbf{b} \}$
- Distance moyenne : $d(\mathbf{a},\mathbf{b}) = \frac{1}{m_a m_b}\{\sum_{u \in a}\sum_{v \in b} m_u m_v d(\mathbf{u},\mathbf{v})\}$
- Critère de Ward : $d(\mathbf{a},\mathbf{b}) = \frac{m_a m_b}{m_a + m_b} d(\mathbf{g}_a, \mathbf{g}_b)$

où $\mathbf{g}_a$ et $\mathbf{g}_b$ sont les centres de gravité des groupes **a** et **b**.

6.2.5 Exemple d'application 1

Le premier exemple d'application reprend celui de l'*Enquête Budget-temps Multimédia 1991-1992* du CESP développé à la section 4.5. Il comprend deux classifications hiérarchiques effectuées sur les lignes et les colonnes de la table de contingence 4.5 - 1. Les distances entre éléments sont les distances du χ^2 entre points-profils et l'agrégation se fait en utilisant le critère de Ward généralisé (cf § 6.2.3 – b). Seuls les éléments actifs de l'analyse des correspondances ont été retenus : il s'agit d'une table (*8,6*) croisant *8* catégories socioprofessionnelle et *6* types de médias, l'unité statistique étant le "contact média". Comme ce fut le cas pour l'analyse des correspondances de cette même table, la fonction de ce traitement n'est pas la réduction d'un tableau de

---

[1] Cette condition a été présentée par Bruynooghe (1978) sous le nom d'axiome de réductibilité. Elle permet en effet la mise d'un oeuvre d'un autre algorithme, dit des *voisinages réductibles*, qui permet d'accélérer l'algorithme de base de la classification hiérarchique par l'utilisation de seuils de distances.

[2] On désignera ici à la fois par **a** (ou **b**) un élément ou un nœud à une certaine étape de l'agrégation, et l'ensemble des éléments constituant ce nœud.

données trop grand et complexe mais une présentation pédagogique des différentes étapes de calcul.

a _ Classification des lignes (professions)

Les principales étapes de la classification des lignes sont résumées sur la figure 6.2 - 14, qu'il faut lire de la façon suivante : la première colonne (NUM) donne les numéros des nœuds, qui sont donc des nouveaux éléments à classer et prennent la suite des 8 éléments à classer[1].

```
CLASSIFICATION HIERARCHIQUE : DESCRIPTION DES 7 NOEUDS (de 9 à 15)
 NUM. AINE  BENJ   EFF. POIDS INDICE     HISTOGRAMME DES INDICES DE NIVEAU
  9    6     7      2    1927 .00024    *
 10    9     5      3    3783 .00038    **
 11    2     1      2     789 .00064    ****
 12   10     4      4    5041 .00208    ********
 13    8    11      3    6651 .00276    *************
 14   12    13      7   11692 .00493    ***********************
 15    3    14      8   12388 .01125    ****************************************
SOMME  DES  INDICES       =     .02228
```

Figure 6.2 – 14 : Description des étapes de la classification hiérarchique
(lignes actives de la table de contingence 4.5 - 1, section 4.5)

On lit ainsi sur la première ligne que le nœud n°9 est formé des éléments terminaux 6 et 7, il est donc formé de 2 éléments (colonne : EFF.) dont le poids total (colonne POIDS) est de *1927*. La valeur de l'indice d'agrégation correspondant est de *0.00024*. Les valeurs croissantes de l'indice seront illustrées par une esquisse d'histogramme à droite des colonnes numériques[2]. On vérifie que la somme des indices est égale à la somme des valeurs propres issues de l'analyse des correspondances de la même table (tableau 4.5 - 2 du §4.5.1).

Le dendrogramme de la figure 6.2 - 15 donne en fait la même information, présentée de façon plus suggestive, car la composition des nœuds à partir des éléments terminaux est maintenant lisible. On note la grande homogénéité des ouvriers (N.Q. et Qual.) et employés (indice très bas), les agriculteurs, petits patrons et inactifs constituant un deuxième groupe moins homogène, alors que les professions intermédiaires occupent une position médiane. Enfin les cadres supérieurs et professions libérales ne se rattachent à l'ensemble des autres catégories que beaucoup plus tard.

---

[1] La terminologie Aîné et Benjamin (deuxième et troisième colonnes) s'applique aux deux éléments qui sont agrégés à une étape donnée (c'est-à-dire les plus proches à cette étape au sens de l'indice d'agrégation retenu).

[2] Comme l'indiquait la figure 6.2 - 3, ces histogrammes peuvent donner une idée du nombre de classes d'une bonne partition, qui correspond à un saut important de l'indice.

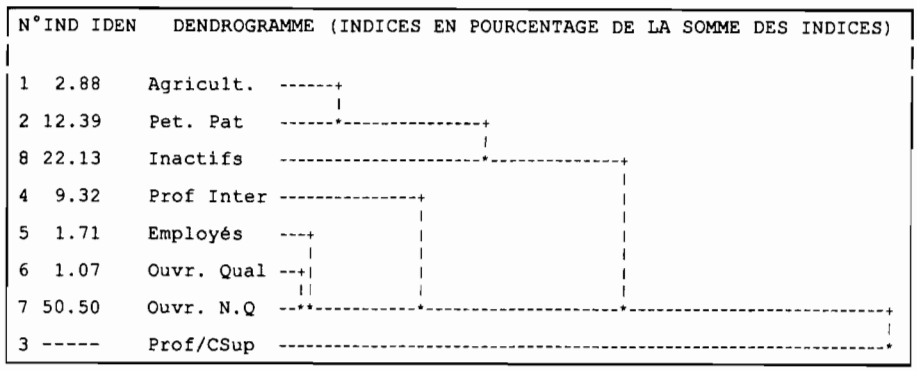


Figure 6.2 – 15 : Dendrogramme pour les huit catégories (lignes)
(lignes actives de la table de contingence 4.5 - 1, section 4.5)

On retrouve donc les regroupements visibles sur la figure 4.5 - 1 du paragraphe 4.5.2)[1]. Notons ici que le plus grand indice correspond au premier facteur de l'analyse de la section 4.5 (opposition des cadres supérieurs et de l'ensemble des catégories), et que le second plus grand indice correspond au second facteur (opposition entre les deux groupes ouvriers/employés et agriculteurs/petits patrons). Cette correspondance entre nœuds et facteurs n'est pas générale, mais fréquente [2].

b _ Classification des colonnes (médias)

La méthode d'agrégation est la même et conduit évidemment à la même somme des indices (inertie totale). Les règles de lectures des figures 6.2 - 16 et 6.2 - 17 sont les mêmes que précédemment.

Les deux plus grands indices correspondent encore aux principales oppositions visibles sur les deux premiers facteurs de l'analyse des correspondances.

CLASSIFICATION HIERARCHIQUE : DESCRIPTION DES 5 NOEUDS (de 7 à 11)

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
7	2	1	2	7266	.00135	**
8	4	7	3	8933	.00251	******
9	5	8	4	10236	.00323	**********
10	6	9	5	11950	.00439	**********
11	3	10	6	12388	.01079	**********

SOMME DES INDICES = .02228

Figure 6.2 – 16 : Description des étapes de la classification hiérarchique
(colonnes de la table de contingence 4.5 - 1, §4.5.1)

[1] La complémentarité entre les deux approches sera développée en section 6.4.

[2] On note également que les deux plus grands indices (*0.0112, 0.0049*) sont ici inférieurs aux deux plus grandes valeurs propres (*0.0139, 0.0072*). La section 6.4 précisera quelques relations et inégalités entre ces grandeurs.

La structure observable sur le dendrogramme est celle d'un effet de chaîne, ou de classe absorbante : l'agrégation se fait en ajoutant un élément terminal à la classe de l'étape précédente.

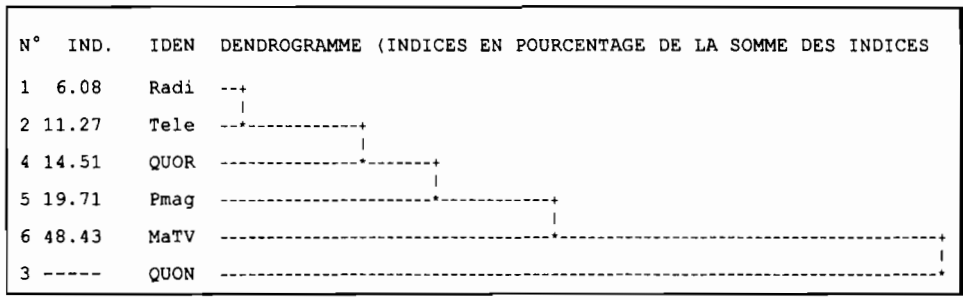


Figure 6.2 – 17 : Dendrogramme pour les six colonnes
(colonnes de la table de contingence 4.5 - 1, §4.5.1)

Il ne s'agit pas d'un artefact de la méthode[1]. Cela traduit plutôt les diffusions très inégales des différents médias considérés.

Notons que si la classification apporte (dans le cas de tableaux en vraie grandeur) certaines informations supplémentaires par rapport à l'analyse des correspondances (les distances sont ici calculées dans tout l'espace), l'absence de représentation simultanée des lignes et des colonnes limite cependant les possibilités d'interprétation.

6.2.6 Exemple d'application 2

Cet exemple d'application s'appuie sur l'enquête sémiométrique présentée au paragraphe 3.5.2 du chapitre 3, déjà illustrée par des analyses en composantes principales et des cartes de Kohonen dans le présent chapitre. Nous proposons ici d'appliquer une classification ascendante hiérarchique sur les 70 mots présentés sous forme d'un tableau de notes[2] et munis de la distance euclidienne.
Nous nous intéressons encore aux proximités entre les mots, les seules à être suggestives et interprétables (les répondants sont anonymes). Pour des questions d'encombrement, nous ne décrirons pas la totalité de l'arbre hiérarchique, mais seulement les dernières coupures de l'arbre.

---

[1] Contrairement à l'agrégation suivant le saut minimal, le critère de Ward généralisé ne provoque pas facilement d'effets de chaîne.

[2] Une classification ascendante hiérarchique concerne le plus souvent les lignes (individus, observations) du tableau de données. Pour réaliser une classification sur les colonnes (variables), le tableau doit être transposé.

Le principe de l'algorithme est de regrouper les mots deux par deux par agglomération progressive fournissant ainsi une hiérarchie de *partitions* des mots. Nous retenons ici les partitions emboîtées en 3, 6 et 9 classes sélectionnées à partir des sauts observés sur l'histogramme des indices de niveaux (seul le bas de cet histogramme, correspondant à l'agrégation des dernières classes, est représenté sur la figure 6.2-18).

```
num. eff. poids indic  histogramme des indices

127   6   18  .00411  ***
128   8   24  .00470  ****
129  14   42  .00511  ****
130   7   21  .00591  ******
131  10   30  .00612  ******
132  16   48  .01210  **************
133  23   69  .01267  ***************
134  27   81  .01490  ******************
135  23   69  .01996  ***************************
136  31   93  .02153  *****************************
137  32   96  .02408  ********************************
138  38  114  .03460  ***********************************************
139  70  210  .04673  ************************************************************

    somme des indices   =     .27337
```

Figure 6.2 – 18 : Etapes de la classification hiérarchique sur les données sémiométriques

La dernière barre de cet histogramme correspond à l'agrégation des deux dernières classes, donc les deux dernières barres (nettement détachées) correspondent à trois classes. Les cinq dernières barres à six classes, et les huit dernières à neuf classes.

La figure 6.2-19 schématise le dendrogramme hiérarchique des 9 classes, alors que le tableau 6.2-2 donne la composition des 9 classes.

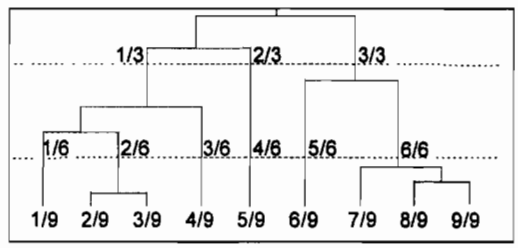


Figure 6.2 - 19 : Dendrogramme de la classification hiérarchique sur les données sémiométriques

On note par exemple que la classe 2/3 regroupe les mots isolés de l'axe 1 de l'analyse en composantes principales qui sont en fait les mots mal notés : *inconnu, danger, rompre, vide, armure, fusil, muraille*. C'est donc une classe qui s'agrège directement à un niveau assez haut.

On retrouve par ailleurs des regroupements déjà observés sur le plans factoriels du chapitre 3 et dans les cartes auto-organisées du présent chapitre.

Tableau 6.2 - 2 : Classification hiérarchique en 3, 6 et 9 classes emboîtées sur les mots (données sémiométriques)

1/3	1/6	1/9	absolu, tradition, noble, Dieu, sacré,âme, justice
	2/6	2/9	modération, raison, attachement, réfléchir, inventeur, science, créateur, commander, certitude, maîtriser
		3/9	métallique, interdire, rigide, matériel, utilitaire, produire
	3/6	4/9	honneur, perfection, prudence, élégance, précieux, or, hériter, gratuit
2/3	4/6	5/9	inconnu, danger, rompre, vide, armure, muraille, fusil
3/3	5/6	6/9	vitesse, nudité, charnel, désir, sensuel
	6/6	7/9	écrire, livre, évasion, feu
		8/9	nager, île, aventurier, fleuve, eau, lune, escalader, sommet, montagne
		9/9	féconder, animal, campagne, enfance, sublime, rêver, bleu, partir, consoler, pureté, admirer, paix, confiance, douceur

6.3 Classification mixte, description statistique des classes

Les algorithmes de classification sont plus ou moins bien adaptés à la gestion d'un nombre important d'objets à classer. Les méthodes de partitionnement (agrégation autour des centres mobiles ou cartes auto-organisées) offrent des avantages incontestables puisqu'elles permettent d'obtenir une partition sur un ensemble volumineux de données à un faible coût, mais elles présentent l'inconvénient de fixer a priori le nombre de classes et de produire des partitions dépendant des premiers centres choisis. Au contraire, la classification hiérarchique est une famille d'algorithmes que l'on peut qualifier de "déterministes" (i.e. qui donnent toujours les mêmes résultats à partir des mêmes données). Par contre si ces algorithmes donnent des indications sur le nombre de classes à retenir ils sont mal adaptés aux vastes recueils de données. Aussi on procède souvent à une classification mixte qui cumule les avantages des deux types de classification.

6.3.1 Stratégie de classification mixte

La classification autour des centres mobiles peut en fait être utilisée comme auxiliaire d'autres méthodes de classification. En fournissant des partitions de vastes ensembles de données, elle permet de réduire la dimension de l'ensemble des éléments à classer en opérant des regroupements préalables.

De ce fait, un algorithme de classification qui paraît actuellement bien adapté au partitionnement d'un ensemble comprenant des milliers ou des dizaines de milliers d'individus est un *algorithme mixte*. L'idée repose sur la combinaison des deux techniques de classification présentées précédemment. Cette idée a été mise en œuvre spontanément par de nombreux praticiens ; elle se trouve, par exemple, sous le nom de *hybrid clustering* dans Wong (1982).

a _ Les étapes de l'algorithme

L'algorithme de *classification mixte* procède en trois phases : l'ensemble des éléments à classer subit un partitionnement initial (centres mobiles) de façon à obtenir quelques dizaines, voire quelques centaines de groupes homogènes ; on procède ensuite à une agrégation hiérarchique de ces groupes, dont le dendrogramme suggérera éventuellement le nombre de classes finales à retenir ; et enfin, on optimise (encore par la technique des centres mobiles appliquée à partir des centres de classe déjà trouvés) la ou les partitions correspondant aux coupures choisies de l'arbre. La figure 6.3 - 1 schématise les différentes étapes de l'algorithme de classification mixte.

1 - Partitionnement initial

Cette première étape vise à obtenir, rapidement et à un faible coût, une partition des n objets en k classes homogènes, où k est largement plus élevé que le nombre s de classes désiré dans la population, et largement plus petit que n. Nous utilisons, pour ce partitionnement initial en quelques dizaines de classes, un algorithme de partitionnement. Ce sera, par exemple, l'algorithme de l'agrégation autour des centres mobiles.

2 - Agrégation hiérarchique des classes obtenues

La seconde étape consiste à effectuer une classification ascendante hiérarchique où les éléments terminaux de l'arbre sont les k classes de la partition initiale. Quelques uns de ces groupements peuvent être proches les uns des autres. Ils correspondent à un groupe "réel" qui aurait été coupé artificiellement par l'étape précédente. D'autre part, la procédure crée, en général, plusieurs petits groupes ne contenant parfois qu'un seul élément. Le but de l'étape d'agrégation hiérarchique est de reconstituer les classes qui ont été fragmentées et d'agréger des éléments apparemment dispersés autour de leurs centres d'origine. L'arbre correspondant est construit selon le critère de Ward qui tient compte des masses au moment des choix des éléments à agréger.

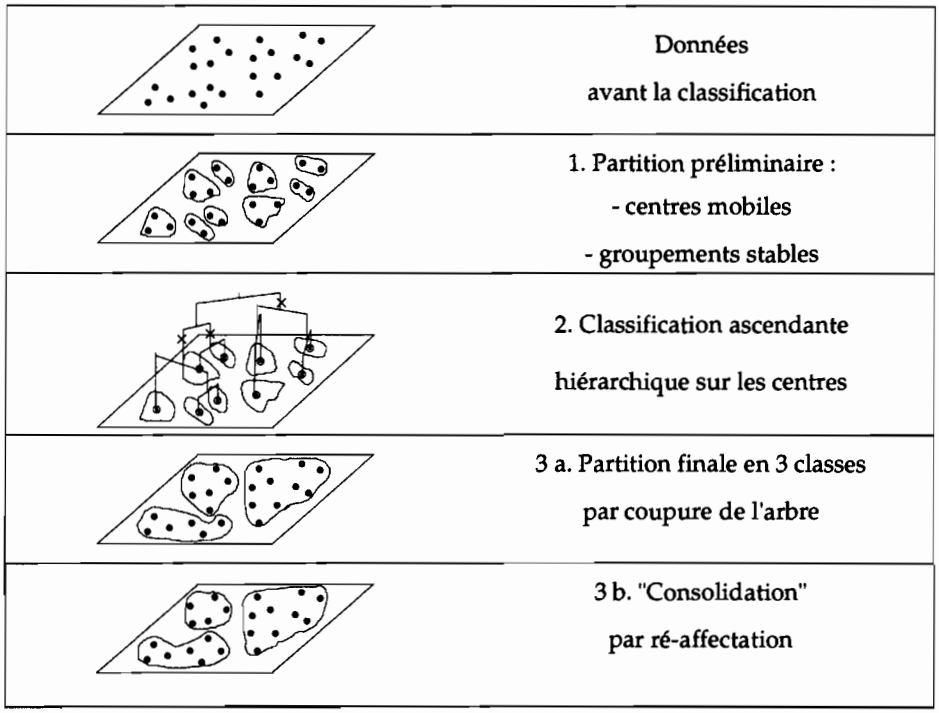


Figure 6.3 – 1 : Schématisation de la classification mixte

*3 - **Partitions finales***

La partition finale de la population est définie par coupure de l'arbre de la classification ascendante hiérarchique. L'homogénéité des classes obtenues peut être optimisée par réaffectations.

b _ Choix du nombre de classes par coupure de l'arbre

Le choix du niveau de la coupure, et ainsi du nombre de classes de la partition, peut être facilité par une inspection visuelle de l'arbre (cf. figures 6.3 - 1 et 6.3 - 2) : la coupure doit être faite après les agrégations correspondant à des valeurs peu élevées de l'indice, qui regroupent les éléments les plus proches les uns des autres, et avant les agrégations correspondant à des valeurs élevées de l'indice, qui dissocient les groupes bien distincts dans la population.

En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure sont nécessairement éloignés, ce qui est la définition d'une bonne partition.

En pratique, la situation n'est pas aussi clairement définie que le montre la figure 6.3 - 2. L'utilisateur pourra choisir entre deux ou trois niveaux de coupure possibles et donc entre deux ou trois partitions finales.

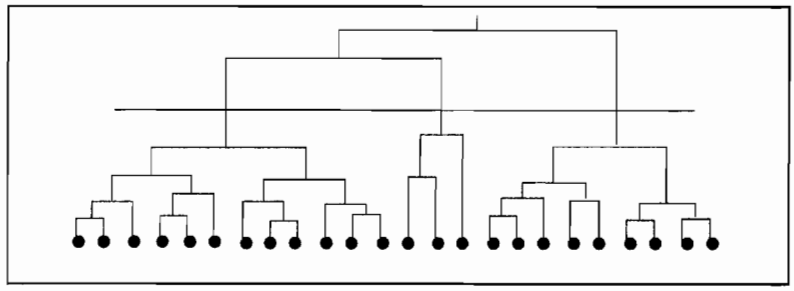


Figure 6.3 – 2 : Coupure visuelle de l'arbre

La coupure de l'arbre peut être facilitée par l'examen de l'histogramme des indices croissants de niveau et l'on coupera au niveau pour lequel cet histogramme marque un palier important. Toute barre de cet histogramme indique la valeur de l'indice d'une agrégation c'est-à-dire la perte d'inertie obtenue en passant d'une partition en s classes à la partition en $s - 1$ classes.

La situation idéale est montrée par la figure 6.3 – 3 (a) où l'on observe un palier évident entre le $4^{ème}$ et le $5^{ème}$ indices suggérant ainsi une bonne partition en cinq classes. La figure 6.3 – 3 (b) est typique de la situation où il est difficile de décider d'un nombre "réel" de groupes dans la population. Mais une telle partition, en s classes par exemple, n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a pas la propriété de donner à chaque étape une partition optimale.

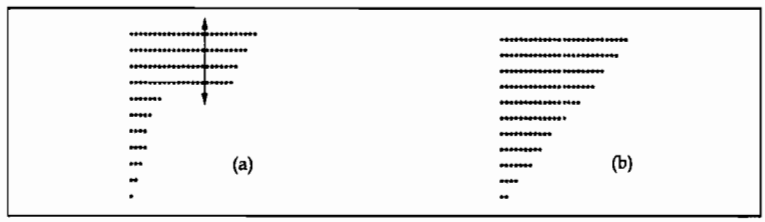


Figure 6.3 – 3 : Histogrammes des indices de niveau

c _ Procédure de consolidation

Pour améliorer la partition obtenue, on utilise de nouveau une procédure d'agrégation autour des centres mobiles dont on sait qu'elle ne peut qu'augmenter l'inertie entre les classes à chaque itération. Cette procédure de consolidation a pour effet d'optimiser, par réaffectation, la partition obtenue par coupure de l'arbre hiérarchique. Malgré la relative complexité de la procédure, on ne peut toujours pas être assuré d'avoir trouvé la "meilleure

partition en k classes" mais on s'en approche vraisemblablement dans beaucoup de situations courantes.

6.3.2 Description statistique des classes

La description automatique des classes constitue en pratique une indispensable étape de toute procédure de classification.

Les aides à l'interprétation des classes sont généralement fondées sur des comparaisons de moyennes ou de pourcentages à l'intérieur des classes avec les moyennes ou les pourcentages obtenus sur l'ensemble des éléments à classer. Pour sélectionner les variables continues ou les modalités des variables nominales les plus caractéristiques de chaque classe, on mesure l'écart entre les valeurs relatives à la classe et les valeurs globales. Ces statistiques peuvent être converties en un critère appelé *valeur-test* permettant d'opérer un tri sur les variables, et de désigner ainsi les variables les plus caractéristiques (cf. Morineau, 1984).

Parmi les variables figurent également celles qui n'ont pas contribué à la construction des classes mais qui peuvent participer à leur description sur le même principe que les variables supplémentaires dans une analyse factorielle.

Ces variables permettent *a posteriori* d'identifier et de caractériser les regroupements établis à partir des variables actives.

a _ Valeurs-test pour les variables continues

Pour caractériser une classe par les variables continues, on compare $\overline{X}_k$, la moyenne d'une variable X dans la classe k, à la moyenne générale $\overline{X}$ et on évalue l'écart en tenant compte de la variance $s_k^2(X)$ de cette variable dans la classe. La valeur-test est ici simplement la quantité :

$$t_k(X) = \frac{\overline{X}_k - \overline{X}}{s_k(X)}$$

avec :

$$s_k^2(X) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}$$

où $s^2(X)$ est la variance empirique de la variable X. On reconnaît en $s_k^2(X)$ la variance d'une moyenne dans le cas d'un tirage sans remise des n_k éléments concernés.

Interprétation en termes de probabilités (variables supplémentaires)

Sous l'hypothèse "nulle" d'un tirage au hasard sans remise des n_k individus de la classe k, la variable $\overline{X}_k$ représentant la moyenne dans la classe a pour espérance $\overline{X}$ et pour variance théorique $s_k^2(X)$.

La valeur-test $t_k(X)$ suit donc approximativement une loi normale centrée et réduite (théorème de la limite centrale). Elle évalue la distance entre la moyenne dans la classe et la moyenne générale en nombre d'écarts-types d'une loi normale.

Il va de soi que cette interprétation n'a de sens que pour une variable X supplémentaire, c'est-à-dire n'ayant pas participé à la construction des classes.

On ne peut en effet stipuler une indépendance entre les classes d'une partition et une des variables ayant servi à définir cette partition.

On calcule ensuite la probabilité que la variable dépasse la valeur absolue de la différence observée. Plus la valeur-test est forte (plus la probabilité est faible), plus l'hypothèse d'avoir les n_k valeurs de la variable X tirées au hasard parmi les valeurs possibles est discutable.

Dans ce cas, la moyenne dans la classe diffère de la moyenne générale, et la variable est caractéristique de la classe. Le classement des variables par probabilités de dépassement croissantes est le même que le classement par valeurs-test décroissantes. Du point de vue de la désignation des variables les plus caractéristiques, les deux informations sont équivalentes.

Extension aux variables actives

S'il n'est pas licite d'interpréter de façon probabiliste les valeurs-test calculées sur les variables actives, il est possible de les utiliser pour obtenir un *classement* de celles-ci en vue de caractériser chaque classe. Les valeurs absolues des valeurs-test constituent alors de simples mesures de similarité entre variables et classes.

b _ Valeurs-test pour les variables nominales

Une modalité (ou catégorie) d'une variable nominale est considérée comme caractéristique de la classe si son abondance dans la classe est jugée significativement supérieure à ce qu'on peut attendre compte tenu de sa présence dans la population.

En notant n_{kj} le nombre d'individus ayant la modalité j parmi les n_k individus de la classe k, n_j le nombre d'individus ayant la modalité j et n l'effectif total, l'abondance de la modalité j est définie, en premier lieu, en comparant son pourcentage dans la k ème classe :

$\frac{n_{kj}}{n_k}$ à son pourcentage dans la population $\frac{n_j}{n}$.

La valeur-test prend en compte tous les éléments du tableau 6.3 - 1.

Tableau 6.3 – 1 : Modalités de variables nominales et classes d'individus

	classe k	autres classes	population
modalité j	n_{kj}	*	n_j
autres modalités	*	*	*
population	n_k	*	n

Sous l'hypothèse "nulle"[1] où les n_k individus de la classe k sont tirés au hasard sans remise parmi la population des n individus, le pourcentage d'individus de la classe k ayant la modalité j d'une part, et le pourcentage d'individus ayant la modalité j dans la population d'autre part, devraient coïncider aux fluctuations aléatoires près :

$$\frac{n_{kj}}{n_k} \approx \frac{n_j}{n}$$

C'est l'hypothèse d'indépendance sous laquelle le nombre N d'individus de la classe k ayant la modalité j est une variable aléatoire qui suit une loi hypergéométrique dont les trois paramètres apparaissent dans les marges du tableau 2.3 - 1. On calcule donc la probabilité d'obtenir une valeur N supérieure à n_{kj} :

$$p_k(j) = Prob(N \geq n_{kj})$$

Plus cette probabilité[2] $p_k(j)$ est faible, plus l'hypothèse d'un tirage au hasard est difficile à accepter. On se sert de cette probabilité pour ranger les modalités caractéristiques de la classe (la plus caractéristique correspondant à la plus petite probabilité).

Cette probabilité est souvent très faible. Il est commode de lui substituer la valeur $t_k(N)$ de la variable normale correspondant à la même probabilité. C'est la *valeur-test*. Elle mesure l'écart entre la proportion dans la classe et la proportion générale, en nombre d'écarts-types d'une loi normale. La valeur-test, pour une modalité d'une variable nominale, est donc un critère statistique associé à la comparaison des effectifs dans le cadre d'une loi hypergéométrique.

---

[1] Comme dans le cas des variables continues, cette hypothèse nulle n'a de sens que pour des variables nominales supplémentaires. Mais les valeurs-test que l'on va calculer pourront encore jouer le rôle d'indices de similarités entre modalités actives et classes et donc servir à ranger ces modalités par ordre d'intérêt pour chaque classe.

[2] Si l'on désigne par C_a^b le nombre de parties distinctes de b éléments que l'on peut extraire d'un ensemble de a éléments, la probabilité $Prob\ (N = x)$ s'écrit ici :

$Prob(N = x) = \dfrac{C_{n_j}^{x} C_{n-n_j}^{n_k - x}}{C_n^{n_k}}$ et la probabilité $p_k(j)$ vaut alors : $p_k(j) = \sum_{x=n_{kj}}^{x=n_k} Prob(N = x)$.

Notons qu'une estimation approchée de la valeur-test peut être obtenue de façon plus simple en prenant en compte l'espérance de N :

$$E(N) = n_k \frac{n_j}{n}$$

et la variance de X

$$s_k^2(N) = n_k \frac{n-n_k}{n-1} \frac{n_j}{n}\left(1-\frac{n_j}{n}\right),$$

et en calculant la quantité :

$$t_k(N) = \frac{N - E(N)}{s_k(N)}$$

qui donne directement la variable centrée, réduite et normale si l'on peut appliquer l'approximation normale de la loi hypergéométrique Cette approximation est suffisante dans les applications qui ne mettent pas en jeu des effectifs faibles.

c _ Variables caractéristiques d'une classe

La valeur-test revient à effectuer un changement de mesure en transformant la probabilité d'une distribution quelconque en nombre d'écart-types d'une loi normale centrée réduite. Que ce soit pour la recherche des variables continues ou des modalités des variables nominales caractéristiques d'une classe, la valeur absolue de la valeur-test est l'analogue de la valeur absolue d'une variable normale centrée réduite.

Les variables sont d'autant plus intéressantes que les valeurs-test associées sont fortes en valeur absolue. On peut alors ranger ces variables suivant les valeurs-test décroissantes et ne retenir que les éléments les plus significatifs, ce qui permet de caractériser très rapidement les classes.

En sélectionnant, pour chaque classe, les variables les plus caractéristiques, et en calculant leur moyenne ou leur pourcentage dans la classe, on constitue ainsi le "profil-type" de la classe.

Rappelons que la valeur-test ne correspond à un vrai test d'hypothèse (Ici on a l'hypothèse qu'une variable continue ou une modalité d'une variable nominale est indépendante de la partition).que si la variable à laquelle elle est associée est supplémentaire.

Mentionnons alors, comme cela a été fait à propos de l'analyse en composantes principales, que le fait de calculer simultanément plusieurs valeurs-test met l'utilisateur dans une situation de "comparaisons multiples", qui impose de prendre des seuils de signification plus sévères que ceux mis en œuvre lors d'un test unique.

6.4 Complémentarité entre analyse factorielle et classification

Les méthodes factorielles (notamment l'analyse des correspondances multiples) sont particulièrement bien adaptées à l'exploration de grands tableaux de données individuelles tels que ceux produits par les enquêtes. Mais elles ne suffisent pas toujours à fournir une vue satisfaisante de l'ensemble des données. Non seulement les visualisations ne véhiculent qu'une partie de l'information, mais elles sont parfois elles-mêmes trop complexes pour être interprétées facilement. Dans ces circonstances, les techniques de classification peuvent compléter et nuancer les résultats des analyses factorielles. La complémentarité entre analyse factorielle et classification concerne la compréhension de la structure des données et celle des aides pratiques dans la phase d'interprétation des résultats. Dans une première partie (§6.4.1), on justifiera cette utilisation conjointe du point de vue de l'utilisateur confronté à un ensemble complexe de données. Puis on examinera au paragraphe 6.4.2 quelques aspects techniques et théoriques de cette complémentarité.

6.4.1 Utilisation conjointe des axes principaux et de la classification

Face à de très grands tableaux de données, il est indispensable de disposer d'une vue d'ensemble de la base d'information. De ce point de vue, les analyses en axes principaux ou méthodes factorielles sont certainement les techniques exploratoires les mieux adaptées.

a _ Nécessité... et insuffisance des méthodes factorielles

Les représentations graphiques issues des méthodes factorielles présentent certains inconvénients, dont certains sont d'ailleurs interdépendants :

1) Difficultés d'interprétation
Il est toujours difficile d'interpréter les axes ou plans factoriels au delà du plan principal. Le plan (3,4), engendré par les axes factoriels 3 et 4, décrit des proximités qui sont des termes correctifs par rapport aux proximités principales observées sur les deux premiers axes. L'interprétation de ces proximités est donc assez délicate.

2) Compression excessive et déformations
Les visualisations sont limitées à deux, ou en général à très peu de dimensions, alors que le nombre d'axes "significatifs" peut être bien supérieur. Cette

compression excessive de l'espace peut entraîner des distorsions fâcheuses et des superpositions de points occupant des positions distinctes dans l'espace.

3) Manque de robustesse
Les visualisations peuvent manquer de robustesse. Un point-profil aberrant peut notablement influencer le premier facteur et par là toutes les dimensions suivantes, puisque ces dimensions sont reliées au premier axe à travers la contrainte d'orthogonalité des axes.

4) Graphiques factoriels inextricables
Les visualisations peuvent concerner des centaines de points et donner lieu à des graphiques chargés ou illisibles.

Pour remédier à ces lacunes, montrons, point par point, quels peuvent être les apports d'une classification menée simultanément.

- Difficultés d'interprétation et compression excessive des données (points 1 et 2) :
On complète l'analyse factorielle par une classification réalisée sur l'espace tout entier ou sur un sous-espace défini par les premiers facteurs les plus significatifs. Les classes prennent en compte la dimension réelle du nuage de points. Elles corrigent donc certaines déformations dues à l'opération de projection. Une classe peut aussi être typique d'un axe de rang élevé et aider à l'interprétation de ce sous-espace particulier difficilement observable autrement.

- Robustesse imparfaite (point 3) :
La plupart des algorithmes de classification, et particulièrement les algorithmes d'agglomération, sont localement robustes au sens où les parties basses des dendrogrammes produits (nœuds correspondant aux plus petites distances) sont indépendantes des éventuels points marginaux isolés.

- Allégement et description automatique des sorties graphiques (point 4) :
Lorsqu'il y a trop de points-individus sur un plan factoriel, il paraît utile de procéder à des regroupements d'individus en familles homogènes. Il faut donc à ce stade faire appel aux capacités de gestion et de calcul de l'ordinateur pour compléter, aider et clarifier la présentation des résultats. Les classes peuvent être utilisées pour aider l'interprétation des plans factoriels en identifiant des zones bien décrites. Il est en effet plus facile de décrire des classes qu'un espace continu, même à deux dimensions. Comme les algorithmes utilisés pour ces regroupements fonctionnent de la même façon que les points soient situés dans un espace à deux ou à dix dimensions, on allège les sorties graphiques tout en améliorant la qualité de la représentation (points 1 et 2 ci-dessus).

Mais les méthodes factorielles sont nécessaires, malgré leurs insuffisances : la faculté descriptive des axes, les descriptions sous forme de continuum géométrique restent irremplaçables. La classification ne réussit pas toujours à montrer l'importance de certaines tendances ou de facteurs latents continus. Pour observer l'organisation spatiale des classes, le positionnement des classes sur les axes factoriels s'avère indispensable.

La classification peut évidemment aider à découvrir l'existence de groupes d'individus. L'analyse factorielle peut mettre en avant des facteurs latents inattendus.

La découverte de tels phénomènes ou dimensions cachées est l'objectif le plus ambitieux de ces deux familles de méthodes. Leur utilisation complémentaire est souvent indispensable pour atteindre cet objectif.

b _ Mise en œuvre pratique dans le cas de la classification mixte

Pour décrire un ensemble de données de grande taille, principale circonstance dans laquelle l'usage complémentaire des techniques factorielles et de classification est utile, la mise en œuvre conjointe de ces techniques s'opère de la façon suivante.

Etape 1 : **L'analyse factorielle**

L'analyse factorielle est utilisée comme une étape préalable à la classification pour deux raisons : pour son pouvoir de description, présenté dans les chapitres précédents, et pour son pouvoir de filtrage, qui permettra éventuellement de travailler sur des coordonnées factorielles moins nombreuses que les variables de départ.

Etape 2 : **Classification à partir des facteurs**

Il est équivalent d'effectuer une classification des individus sur un ensemble de p variables ou sur l'ensemble des p facteurs. Mais on peut aussi ne prendre en compte qu'un sous-espace factoriel de dimension q $(q < p)$ et réaliser une classification sur les q premiers axes. Le fait d'abandonner les derniers facteurs revient à effectuer un "lissage" des données, ce qui en général améliore la partition en produisant des classes plus homogènes. Les distances entre points sont calculées dans l'espace des premiers axes factoriels avec la distance euclidienne usuelle. Le calcul est simple et la classification peut être menée sur des grands ensembles d'individus[1].

Etape 3 : **Description automatique des classes**

Une fois les individus regroupés en classes, on a vu (§ 6.3.2) qu'il est facile d'obtenir une description automatique de ces classes. On calcule, pour les variables numériques comme pour les variables nominales, des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales. Les valeurs-test permettent de les ranger par ordre d'intérêt.

---

[1] Une technique de classification hiérarchique tel que l'algorithme des voisins réciproques (et particulièrement l'algorithme de recherche en chaîne) peut être réalisée sans garder la matrice des distances en mémoire vive. Les distances entre couples de points sont recalculées à la demande dans l'espace réduit des q premiers facteurs. La place en mémoire de la matrice (n,q) construite à partir des q principales coordonnées des n observations est souvent inférieure à celle du tableau des $n(n-1)/2$ distances.

Etape 4 : **Positionnement des classes dans le plan factoriel**

La division en classes opère un découpage plus ou moins arbitraire d'un espace continu. L'analyse en axes principaux préalable permet alors de visualiser les positions relatives des classes dans l'espace et peut mettre en évidence certaines "trajectoires" masquées par la discontinuité des classes. Il est intéressant de projeter les centres de gravité des classes au sein des variables ou des modalités actives sur le 1er plan factoriel (figure 6.4 - 1).

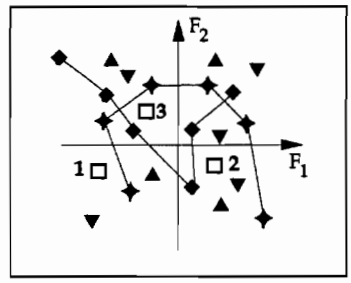


Figure 6.4 – 1. Positions relatives des classes dans l'espace factoriel

Le support visuel permet d'apprécier les distances entre les classes. Par ailleurs, la position des individus repérés par leurs numéros de classe permet de représenter la densité et la dispersion des classes (cf. figure 6.4 - 2).

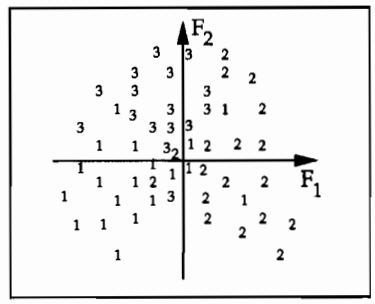


Figure 6.4 – 2. Densité et dispersion des classes dans l'espace factoriel

L'utilisation conjointe de l'analyse factorielle et de la classification permet de se prononcer non seulement sur la réalité des classes, mais également sur leurs positions relatives, leur forme, leur densité et leur dispersion. Les deux techniques se valident mutuellement.

c _ Autres travaux sur la complémentarité

A propos des liens entre les méthodes d'analyse par axes principaux et les méthodes de classification, il faudrait évoquer des méthodes que l'on peut qualifier d'hybrides, c'est-à-dire qui produisent simultanément des axes et des classes. Ainsi, le lien existant entre le haut de l'arbre et les premiers axes factoriels peut suggérer d'utiliser ceux-ci pour construire un arbre à partir des plus grands indices (classification descendante ou divisive, cf., par exemple

Reinert, 1986). On peut également chercher des axes principaux susceptibles de représenter au mieux une classification (Art et *al.*, 1982 ; Gnanadesikan et *al.*, 1982). Certaines de ces méthodes (projections révélatrices, analyses de contiguïté) seront brièvement présentées au chapitre 8. Dans un autre esprit, van Buuren et Heiser (1989), pour classer des individus décrits par des variables nominales, cherchent simultanément des classes et un codage des variables qui optimise un critère de qualité de la classification.

6.4.2 Aspects techniques et théoriques de la complémentarité

On a vu que la complémentarité entre l'analyse des correspondances et la classification ascendante hiérarchique présente des avantages pratiques pour l'utilisateur. On examinera dans ce paragraphe certains aspects plus techniques de cette complémentarité.

a _ Classification des lignes ou colonnes d'un tableau de contingence

La classification ascendante hiérarchique agrège des groupes d'éléments suivant différents critères d'agrégation. Parmi ceux-ci, le critère de Ward généralisé apparaît compatible avec l'analyse des correspondances puisqu'il est fondé sur une notion d'inertie similaire. On a montré en particulier (cf. § 6.2.3) que la somme des valeurs propres (inertie totale du nuage) est égale la somme des indices de niveau. Aussi, il y a une certaine cohérence à utiliser le critère d'inertie de Ward sur un tableau de coordonnées factorielles elles-mêmes issues d'un calcul d'inertie. Si l'arbre de la classification est construit sur les q premiers axes factoriels, on vérifiera que la somme des indices de niveau est égale à la somme des q premières plus grandes valeurs propres retenues.

Une propriété de l'analyse des correspondances assure une bonne compatibilité avec la classification : l'équivalence distributionnelle (cf. § 4.2.1) qui garantit une invariance par regroupement des éléments ayant des profils semblables. Agréger les lignes et les colonnes d'un tableau de contingence est naturel car il s'agit de remplacer des classes par des classes au lieu de remplacer des individus par des groupes d'individus ou des variables par des groupes de variables[1].

b _ Un exemple de coïncidence entre les deux approches

Considérons la table de contingence $\mathbf{K}_{IJ}$ (tableau 6.4 - 1). Nous allons montrer qu'une analyse des correspondances et une classification hiérarchique utilisant

[1] La classification des éléments d'une table de contingence fondée sur le regroupement de catégories homogènes a été abordée par Benzécri (1973), Jambu et Lebeaux (1978), Govaert (1984), Cazes (1986), Gilula (1986), Escoufier (1988), Greenacre (1988).

le critère d'agrégation de Ward (cf. § 6.2.3.b) donnent des résultats équivalents pour cette table.

Tableau 6.4 – 1. Table de Contingence K_{IJ}

	COL7	COL2	COL3	COL4	COL5	COL6	COL1	COL8
LIG1	2	18	12	12	2	2	30	2
LIG4	2	12	21	27	2	2	12	2
LIG5	14	2	2	2	24	20	2	14
LIG2	2	30	12	12	2	2	18	2
LIG6	14	2	2	2	20	24	2	14
LIG7	23	2	2	2	14	14	2	21
LIG3	2	12	27	21	2	2	12	2
LIG8	21	2	2	2	14	14	2	23

En fait, un réarrangement des lignes et des colonnes montre que cette table n'est pas anodine. Elle contient de forts traits structuraux (tableau 6.4 - 2). Elle est symétrique et semble formée de blocs et de sous-blocs particuliers. Ce réarrangement est en fait un sous-produit de l'analyse des correspondances.

Tableau 6.4 – 2 : Table de Contingence K_{IJ} réordonnée

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8
LIG1	30	18	12	12	2	2	2	2
LIG2	18	30	12	12	2	2	2	2
LIG3	12	12	27	21	2	2	2	2
LIG4	12	12	21	27	2	2	2	2
LIG5	2	2	2	2	24	20	14	14
LIG6	2	2	2	2	20	24	14	14
LIG7	2	2	2	2	14	14	23	21
LIG8	2	2	2	2	14	14	21	23

Cette table de contingence fait en fait partie d'une famille plus large de tableaux décrits dans Benzécri (1973, vol. 2, chapitre 11) qui seront évoqués plus bas. Une classification ascendante hiérarchique utilisant le critère de Ward produit le dendrogramme représenté sur la figure 6.4 - 3, où les indices de niveaux figurent entre parenthèses près des nœuds correspondants.

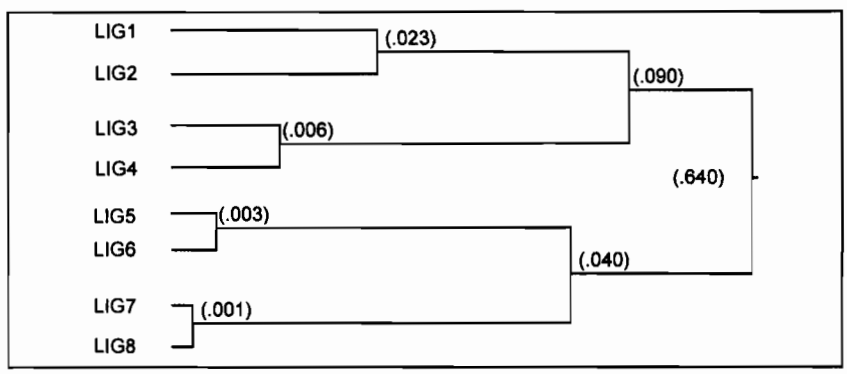


Figure 6.4 – 3 : Esquisse du dendrogramme décrivant la classification hiérarchique de la table de contingence (*8,8*) K_{IJ}

Les valeurs propres issues de l'analyse des correspondances de K_{IJ} figurent dans le tableau 6.4 - 3. Elles coïncident avec les indices d'agrégation.

Tableau 6.4 – 3 : Valeurs propres issues de l'analyse des correspondances de K_{IJ}

λ1 =	.640	(80.0 % de la trace)
λ2 =	.090	(11.0 %)
λ3 =	.040	(5.0 %)
λ4 =	.023	(3.0 %)
λ5 =	.006	(.7 %)
λ6 =	.003	(.4 %)
λ7 =	.001	(.1 %)

Le tableau 6.4 - 4 donne les coordonnées factorielles des points lignes (qui sont les mêmes que celles des points colonnes au signe près, puisque la matrice de départ est symétrique).

La façon dont sont organisés ces vecteurs propres permet de comprendre le processus de construction de la table de contingence : on part des facteurs structurés de cette façon et on utilise la formule de reconstitution des données.

Chaque vecteur oppose deux blocs. Il est orthogonal au vecteur précédent et les coordonnées sont égales à l'intérieur de chaque bloc. Tous les vecteurs sont centrés et orthogonaux à la première bissectrice.

La figure 6.4 - 4 donne la représentation des points-profils dans le plan des deux premiers axes factoriels.

Tableau 6.4 – 4 : Coordonnées factorielles issues de l'analyse des correspondances de K_{IJ}

Axes	1	2	3	4	5	6	7
LIGNE1	-.80	.42	0.00	.30	0.00	0.00	0.00
LIGNE2	-.80	.42	0.00	-.30	0.00	0.00	0.00
LIGNE3	-.80	-.42	0.00	0.00	-.15	0.00	0.00
LIGNE4	-.80	-.42	0.00	0.00	.15	0.00	0.00
LIGNE5	.80	0.00	-.28	0.00	0.00	.10	0.00
LIGNE6	.80	0.00	-.28	0.00	0.00	-.10	0.00
LIGNE7	.80	0.00	.28	0.00	0.00	0.00	.06
LIGNE8	.80	0.00	.28	0.00	0.00	0.00	-.06

On constate que cette figure bi-dimensionnelle permet de distinguer les deux grands blocs (axe 1), puis, à l'intérieur de l'un d'eux, deux sous-blocs (axe 2) , mais qu'elle est moins riche d'information que la figure 6.4 - 3, elle aussi bidimensionnelle.

La figure 6.4 - 3 (dendrogramme) a l'avantage de montrer simultanément tous les blocs et tous les niveaux de la hiérarchie.

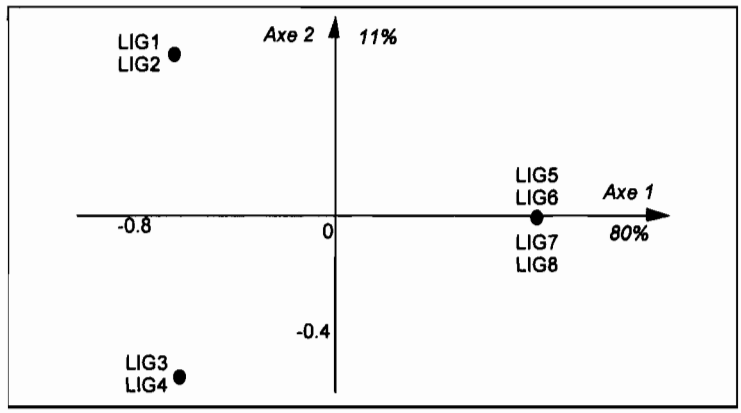


Figure 6.4 – 4 .Premier plan factoriel de l'analyse de K_{IJ}

On décrira brièvement ces tableaux de *correspondances hiérarchiques* dans l'annexe technique 6.6.1 de ce chapitre, en renvoyant à Benzécri (1973, *op. cit.*) et Cazes (1984, 1986 a) pour une présentation systématique et des généralisations de ces notions.

6.4.3 Valeurs propres et indices de niveau

Hormis des cas très particuliers, comme ceux constitués par les correspondances hiérarchiques étudiées au paragraphe précédent, les relations entre analyse des correspondances et classification opérées sur une même table de contingence sont difficiles à étudier. Dans le cas de la classification hiérarchique utilisant le critère de Ward, on peut mettre en évidence certaines inégalités et étudier certaines structures particulières.

a _ Quelques inégalités

Notons tout d'abord que pour une table de contingence quelconque (si l'on excepte les tables symétriques), la classification hiérarchique donnera des indices différents selon que l'on agrège les lignes et les colonnes, alors que l'analyse des correspondances ne fournit qu'une série de valeurs propres.

La plus grande valeur propre issue de l'analyse des correspondances est supérieure ou égale au plus grand indice d'agrégation (lignes ou colonnes) donné par la classification.

Cet indice est en effet une mesure de variance externe (dite variance "inter", par opposition à la variance "intra", mesurant la dispersion à l'intérieur des groupes) entre les deux derniers groupes agrégés. Cette variance externe est inférieure à la variance totale mesurée sur la droite qui joint les centres de gravités des deux groupes, elle-même inférieure à la meilleure variance totale

possible sur une droite quelconque, ce qui est la définition de la plus grande valeur propre[1].

Plus généralement, Benzécri et Cazes (1978) ont montré que la *somme des r plus grandes valeurs propres est supérieure ou égale à la somme des r plus grands indices d'agrégation.* Enfin, ces auteurs ont donné un intéressant contre-exemple montrant qu'il n'existe pas de borne inférieure positive pour le quotient entre le plus grand indice d'agrégation et la plus grande valeur propre : on peut trouver des distributions de densité telles que le plus grand indice soit une fraction arbitrairement petite de la plus grande valeur propre.

b _ Le cas des tables de contingence structurées par blocs

Cette structure évoquée en section 4.3.1 est aisément reconnue par l'analyse des correspondances car *k* blocs engendrent *k* valeurs propres égales à *1* (y compris la valeur propre triviale, qui correspond au cas usuel d'un seul bloc).

Cette structure n'est cependant pas systématiquement reconnue par la classification hiérarchique utilisant le critère de Ward, comme l'ont montré par un contre-exemple Kharchaf et Rousseau (1988, 1989).

c _ Lien entre valeurs propres et indices

Ces inégalités et contre-exemples ne donnent que peu d'information sur les liaisons entre valeurs propres et indices, et les liaisons fonctionnelles du paragraphe 6.4.2 ne concernent que des cas d'école. Les liaisons stochastiques entre indices et valeurs propres (dans le cas d'une famille de tables de contingence aléatoires) sont certainement trop complexes pour faire l'objet d'une étude analytique.

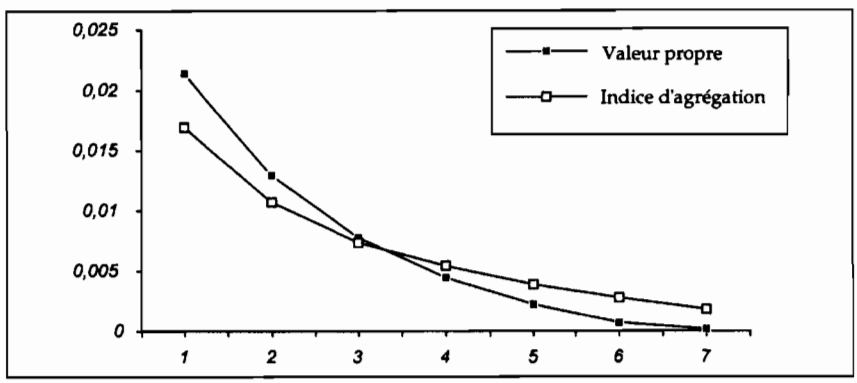


Figure 6.4 - 5 : Exemple de séquences moyennes de valeurs propres et des indices

[1] Notons bien, sur la figure 6.4 - 4 précédente, le cas de coïncidence pour lequel les variances "intra" sur l'axe sont nulles, et pour lequel le meilleur axe factoriel est précisément celui qui relie les deux centres de classes.

Une étude empirique portant sur des simulations de milliers de tables aléatoires (Lebart et Mirkin, 1992) montre que les plus grandes valeurs propres et les plus grands indices d'agrégation sont fortement corrélés. La figure 6.4 – 5, calculée à partir de 1000 tables de contingences pseudo-aléatoires (8, 8) met en évidence l'intervalle de variation plus réduit des indices dans l'hypothèse d'indépendance des lignes et des colonnes.

6.4.4 La complémentarité en pratique : un exemple

Cet exemple d'enchaînement résume certaines étapes d'une application "en vraie grandeur". Il est extrait de traitements de l'enquête sur les conditions de vie et aspirations des Français[1] L'objectif poursuivi ici est double : donner une description d'ensemble des principales attitudes et opinions relevées dans le système d'enquêtes précité ; montrer dans quel cadre factuel s'inscrivent les attitudes et opinions.

Le fichier partiel correspondant à cette application comprend *14* variables nominales actives et en fait plusieurs centaines de variables nominales supplémentaires. Les *14 000* individus correspondent à *7* vagues de *2000* individus (de 1978 à 1984), chaque vague étant représentative de la population de résidants métropolitains âgés de *18* ans ou plus. Ces 14 questions actives pour décrire les perceptions des conditions de vie et du cadre de se décomposent en : deux questions sur la perception de l'évolution des conditions de vie, trois questions sur le thème «Famille», trois questions sur l'environnement physique et technologique, trois questions sur la santé et l'institution médicale, une question sur l'attitude vis-à-vis des équipements collectifs, deux questions sur la justice et la société Un des intérêts de cet exemple est que les structures observées pourront être validées par les échantillons indépendants annuels. Il s'agit d'une situation exceptionnellement favorable pour éprouver la stabilité des résultats d'une analyse exploratoire.

a _ Les étapes

L'enchaînement de méthodes décrit ici est une formulation plus détaillée de la procédure d'utilisation conjointe des méthodes factorielles et de la classification exposée au § 6.4.1. Cette procédure est présentée du point de vue du praticien.

- ***Etape 1 : Analyse factorielle*** . Elle comprend les trois phases suivantes :

- ***Choix d'un thème actif***
Choisir un thème, c'est-à-dire une batterie homogène de variables actives, c'est adopter un point de vue particulier pour la description. On peut décrire les individus à partir d'un thème particulier de l'enquête par exemple les habitudes de

[1] Cf. Lebart et Houzel (1981), Babeau et Lebart (1984), Lebart (1987 b) pour des informations générales sur cette enquête.

consommation, les durées d'activité (budgets-temps), les contacts-médias, les déplacements, etc. Ici, le thème choisi est : la perception des conditions de vie et du cadre de vie (cf. encadré ci-dessus).

- ***Description graphique de la population***
Les graphiques résultant des analyses factorielles (ici : correspondances multiples) fournissent une description de l'échantillon des individus interrogés. La proximité entre individus est fonction de la similitude des réponses aux questions du thème actif.

- ***Positionnement des éléments illustratifs sur les plans factoriels***
On s'intéresse aux questions ne faisant pas partie du thème actif pour aider à interpréter les proximités entre individus. Lorsque la lecture des résultats est gênée par l'abondance des éléments illustratifs, les seuls éléments pertinents pour l'interprétation seront sélectionnés par leurs valeurs-test. Ceci permet d'envisager des explorations systématiques, avec de nombreux croisements de variables.

Comme au § 6.4.1 b, les trois phases suivantes sont :

- ***Etape 2 : Partition de l'ensemble des individus***

- ***Etape 3 : Descriptions statistiques du contenu de chaque classe***

- ***Etape 4 : Positionnement des centres des classes en éléments supplémentaires dans les plans factoriels***

Cet enchaînement est souvent utilisé sous le nom de *thémascope*. C'est donc un outil qui permet de décrire un thème (actif), multidimensionnel par nature, en utilisant la conjonction des deux techniques disponibles (réduction de dimension d'une part, regroupement d'autre part). Il situe ensuite ce thème dans le contexte global de l'enquête, grâce aux techniques de projection de variables supplémentaires sur les plans factoriels et de description automatique des classes. La sélection automatique des éléments les plus significatifs sur les plans factoriels et lors de la description des classes fournit au lecteur une information filtrée et lisible.

b _ L'espace des variables actives (Figure 6.4 - 6)

La figure 6.4 - 6 est l'esquisse du premier plan factoriel d'une analyse des correspondances multiples du tableau (*14 000, 60*). Les *14* réponses aux questions actives (*60* modalités) répartissent les individus interrogés de façon continue dans l'espace. Il est possible de découper ce continuum en grandes zones de la façon la moins arbitraire possible ; les cloisons entoureront ainsi les régions de forte densité et seront disposées de façon à ce que la dispersion des individus soit minimale à l'intérieur des zones. C'est l'arbre hiérarchique de la figure 6.4.7 qui est schématiquement tracé sur le plan factoriel (coupure correspondant à *8* classes). Pour limiter le nombre de graphiques, le résultat de l'*étape 4* figure d'emblée sur la figure.

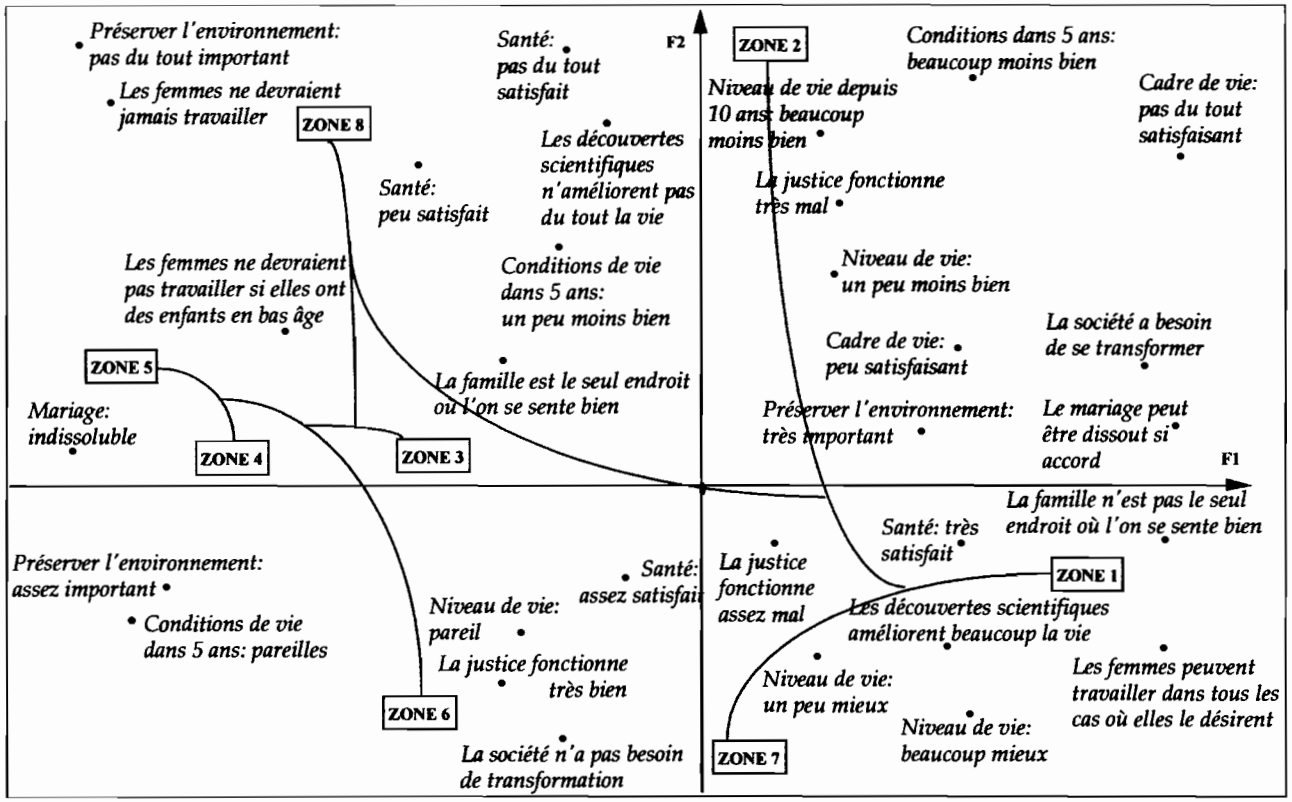


Figure 6.4- 6

Visualisation plane de l'espace des opinions et positionnement des zones

On représente ici les proximités statistiques existant entre une trentaine de modalités de réponses aux questions actives choisies parmi les plus caractéristiques. Les centres des zones sont positionnés comme des modalités supplémentaires.

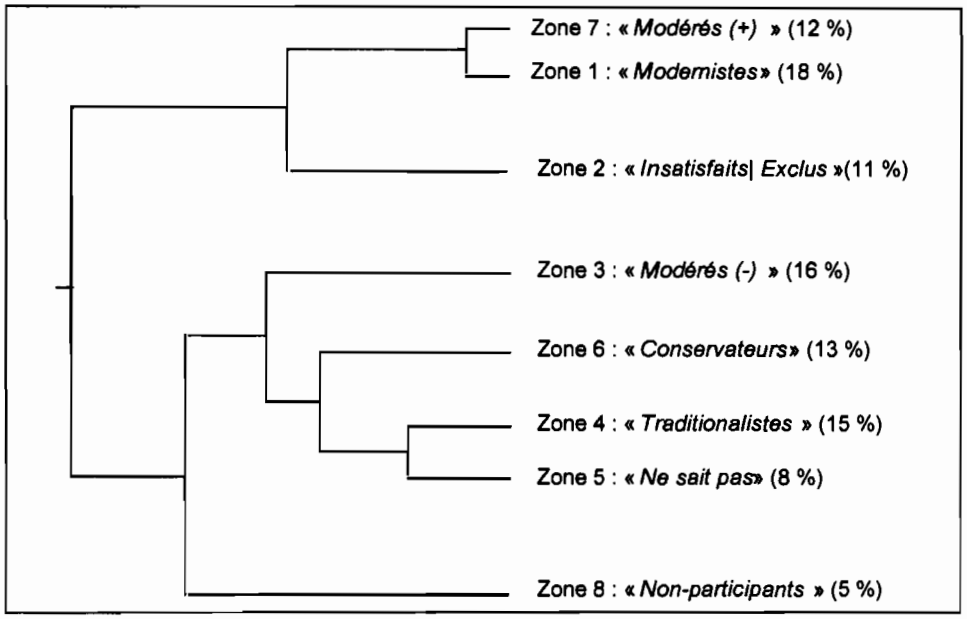


Figure 6.4 – 7. Classification hiérarchique des 14 000 individus en 8 zones

Guide de lecture du dendrogramme : *L'algorithme de classification mixte de la section 6.3 permet de mettre en évidence huit zones* [1]*, positionnées en éléments supplémentaires sur le plan factoriel de la figure 6.4 - 6, et comme éléments terminaux du dendrogramme de la figure 6.4 - 7. Cette figure complète la figure 6.4 -6 . Ainsi, contrairement à ce que l'on observe sur la figure 6.4 - 6 qui ne donne qu'une approximation plane de l'espace, et donc qui déforme les distances, la zone 2 est, d'après le dendrogramme, plus proche des zones 1 et 7 que de la zone 8.*

c _ Exemples de description automatique de trois classes

On va maintenant illustrer la description automatique des classes (cf. § 6.3.2) en caractérisant de façon plus détaillée trois classes (ou zones) sélectionnées parmi les huit précédentes. On distinguera successivement les opinions et perceptions (éléments actifs, et pour certains d'entre eux, supplémentaires), puis les caractéristiques de base (éléments toujours supplémentaires dans cette analyse). Chaque pourcentage interne à la zone sera suivi, entre parenthèses, du pourcentage moyen dans l'ensemble de la population. Les valeurs-test (cf. §6.3.2) qui ont permis de sélectionner et de classer ces variables caractéristiques sont des fonctions de l'écart entre ces deux pourcentages.

---

[1] On parle de zones et non de classes ou de groupes pour rappeler qu'il s'agit de portions d'espace et non d'entités sociologiques ou de catégories ayant une existence indépendante de la batterie des questions actives utilisées ici. Les libellés de ces zones sont purement mnémotechniques.

– **Description de la zone 1 (Modernistes) [droite de la figure 6.4 - 6]**
Cette zone stable représentant en moyenne 18% des personnes interrogées se distinguent par une certaine distance vis-à-vis de la famille traditionnelle.

Variables actives
- 87% pensent que «la famille n'est pas le seul endroit où l'on se sent bien et détendu» (ce pourcentage n'est que de 35% pour l'ensemble de la population)
- 84% déclarent «le mariage peut être dissout sur simple accord» (35%)
- 83% estiment : «les femmes devraient travailler quand elles le désirent» (37%)
- 86% jugent que «préserver l'environnement est très important» (65%)

Variables supplémentaires (signalétique) : jeunes, instruits, parisiens
- 52% n'ont jamais eu d'enfant (28%)
- 32% habitent la région parisienne (15%)
- 78% ont moins de 40 ans (47%)
- 67% sont des locataires (51%)
- 20% sont diplômes d'université ou de grande école (8%)

Autres variables supplémentaires : spécificités de comportement
- 31% se couchent après 23 h (13%), 35% fréquentent un cinéma (17%)
- 57% participent aux activités d'au moins une association (44%)

– **Description de la zone 2 (Insatisfaits / exclus) [haut de la figure 6.4 - 6]**
Cette zone est probablement la seule à mériter le statut de «classe» au sens statistique du terme dans la mesure où elle réapparaît chaque année (de 1978 à 1985) avec un effectif remarquablement constant qui oscille entre 9% et 13%.

Opinions et perceptions : niveau et cadre de vie non satisfaisants
- 69% pensent « le niveau de vie personnel :beaucoup moins bien» (13%)
- 62% estiment que leurs «conditions de vie vont beaucoup se détériorer au cours des cinq prochaines années» (12%)
- 61% considèrent que «la justice fonctionne très mal» (26%)
- 85% déclarent «s'imposer régulièrement des restrictions» (61%)
- 17% ne sont «pas du tout satisfaits de leur cadre de vie quotidien» (5%) ;
- 90% pensent que «la société a besoin de se transformer» (74%)

Variables supplémentaires (signalétique) : des ressources faibles [1]
- 38% souffrent d'un handicap, d'une infirmité ou d'une maladie chronique (26%)

---

[1] Cette zone n'a pas de caractéristiques socio-démographiques aussi typées que la zone 1. Elle constitue avant tout une classe de personnes aux ressources faibles, au niveau de vie bas, qui subissent des tensions où font face à des difficultés variées. On a affaire ici typiquement à une «classe polythétique», c'est-à-dire une classe qui peut être définie non par une combinaison fixe d'attributs, mais par la possession d'un certain nombre d'attributs dans une liste : il y a dans ce cas cumul de handicaps d'origines variées.

- 38% n'ont aucun élément de patrimoine (27%), 53% sont locataires (44%)
- 15% sont chômeurs (en 1983 et 84) (6%)
- 22% habitent en HLM ou ILN (16%) 9% sont séparés ou divorcés (5%)

Autres variables supplémentaires :

- 55% ont déclaré «avoir souffert de nervosité au cours des quatre dernières semaines» (37%). 28% ont dit avoir souffert d'«état dépressif» (15%),
- 38% d'«insomnie» (25%), 49% de «mal au dos» (38%),
- 45% s'estiment «beaucoup inquiets de l'éventualité du chômage» (25%).

– **Description de la zone 5 (réponses "ne-sait-pas") [gauche de la fig. 6.4-6]**

Cette zone *a priori* peu intéressante du point de vue des opinions exprimées joue cependant un rôle méthodologique important. Alors que les refus ou les dissimulations entachent la qualité des enquêtes socio-économiques usuelles, les réponses du type «ne sait pas» viennent s'ajouter aux défections précédentes dans le cas des mesures de perceptions ou d'opinions.

Variables actives

- 65% répondent NSP (pour «ne sait pas») à la question «la société a-t-elle besoin de se transformer ?» (9%)
- 53% répondent NSP à la question sur «le fonctionnement de la justice» (7%) ; 8% refusent de répondre à cette question (2%)

Variables supplémentaires (signalétique) : femmes âgées peu instruites

- 67% sont des femmes (53%), 46% n'ont aucun diplôme (26%)
- 43% habitent des communes de moins de *2 000* habitants (29%)
- 75% n'appartiennent à aucune association (56%)[1].

d _ Projection de variables signalétiques (en supplémentaires) sur le plan principal de la figure 6.4 - 6 (figure 6.4 - 8)

Les descriptions zones par zones donnent déjà une idée de l'«ancrage factuel» des perceptions, mais un positionnement direct des caractéristiques de base a le mérite de montrer à quel point l'espace des perceptions est un continuum [2].

---

[1] Le fait qu'il s'agisse surtout de femmes âgées peu instruites habitant en milieu rural, alors que les questions «non répondues» sont les plus « politiques » (les transformations de la société, la justice) confirme les travaux de méthodologie d'enquête (cf. Michelat et Simon, 1985).

[2] L'étude complète comporte une description beaucoup plus détaillée de l'ensemble des classes, une étude de l'évolution des trajectoires des points-modalités et des classes dans les plans factoriels au cours du temps (cf. Lebart, 1986; 1988).

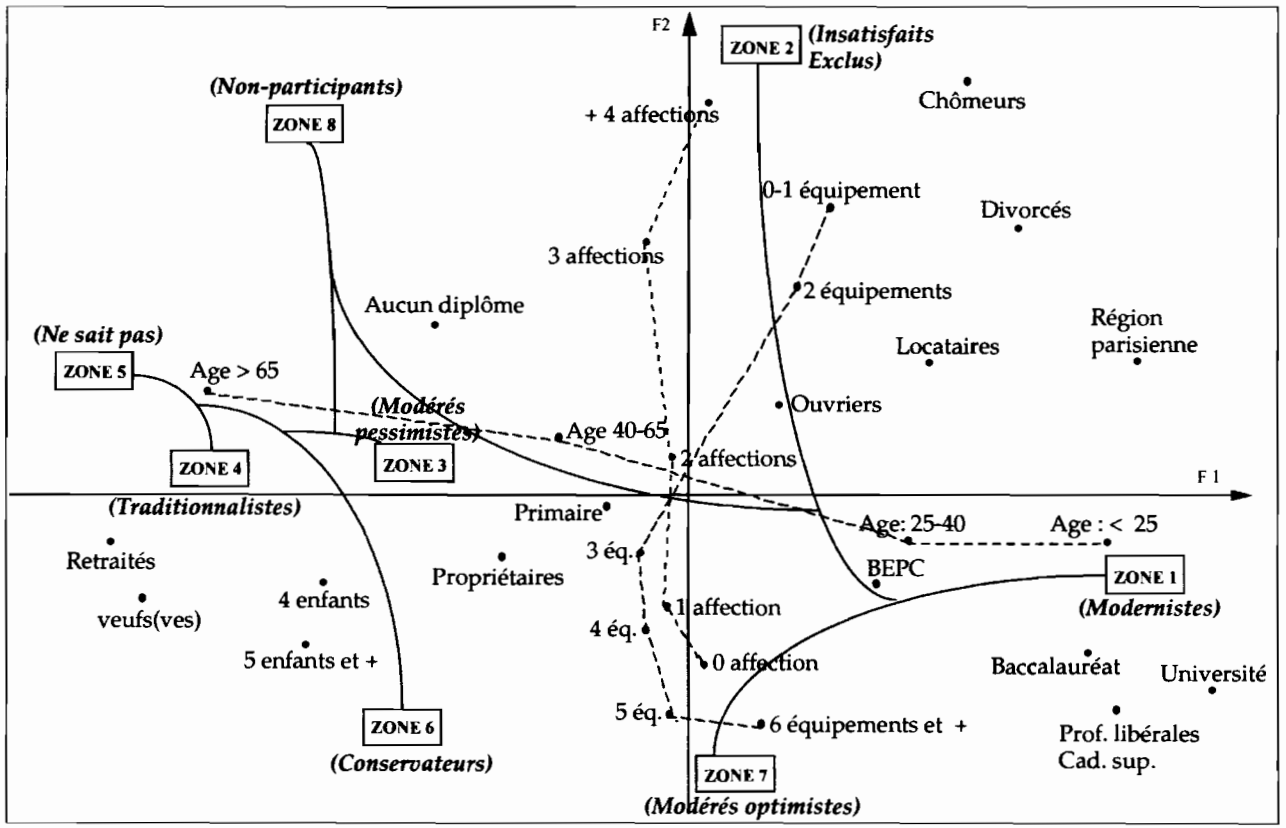


Figure 6.4 - 8 : Projection de quelques caractéristiques (en supplémentaires) sur le plan principal (cf. figure 6.4 - 6).
De gauche à droite, glissement des *traditionalistes* vers les *modernistes*, et de bas en haut, des *conservateurs* vers les *instatisfaits/exclus*

Les modalités des différentes variables s'ordonnent en effet régulièrement dans le plan de la figure 6.4 - 8.

Il n'y a pas de discontinuité entre les «traditionalistes» âgés, ruraux, peu instruits situés dans la partie gauche de la figure et les «modernistes» jeunes, instruits, urbains, situés à l'extrémité droite de l'axe horizontal. Il y a de même une continuité entre les «conservateurs» et les «modérés +» d'âge moyen situés dans la partie basse de la figure 6.4 - 8 et les insatisfaits dans la partie haute. *Le nombre d'équipements et d'éléments de patrimoine* jalonne cette direction verticale, tout comme *le nombre d'affections déclarées* (petites affections au cours des quatre dernières semaines), indicateur dont les liens avec l'insatisfaction sont connus.

6.5 Validation des classifications

Dès les premières tentatives de classification s'est posé le problème du nombre de classes à retenir *en vue d'une utilisation particulière*. Déjà, sous cette formulation pragmatique, le problème est moins ambitieux que celui de savoir combien de classes existent réellement dans le corpus de données soumis à l'analyse.

La classification peut en effet être utilisée simplement pour explorer les données, généralisant au cas multidimensionnel l'histogramme qui permet de schématiser une distribution numérique unidimensionnelle[1]. On peut aussi espérer découvrir des classes existantes, dans les cas les plus favorables.

Les questions sont aussi simples que les réponses sont complexes : Existe-t-il des classes ? Si oui, combien? On évoquera brièvement quelques travaux réalisés à propos de l'existence et de la détermination du nombre des classes. La méthodologie de la validation est analogue à celle déjà rencontrée à propos des méthodes factorielles.

Nous avons choisi de présenter, dans ce chapitre dévolu à la classification, les deux familles d'algorithmes sans doute les plus utilisées, celle des centres mobiles et celle de la classification hiérarchique. Nous resterons dans ce périmètre méthodologique pour notre survol des outils de validation. Mais les méthodes et algorithmes de classification sont nombreux[2] et la plupart font

---

[1] Il s'agit en fait de l'utilisation la plus courante dans le cas des traitements de fichiers d'enquêtes. On parle alors parfois de « dissection » des données plutôt que de classification.

[2] Voir Nakache et Confais, 2002, pour une liste plus détaillée des méthodes.

intervenir un grand nombre de critères et de métriques répondant *a priori* à une problématique précise.

6.5.1 Cadre général

Comme dans le cas des méthodes factorielles, trois approches des procédures de validation peuvent être envisagées : l'approche inférentielle classique, l'approche de type simulation/re-échantillonnage, la validation externe.

a _ Cadre inférentiel général

Il sera possible de tester des hypothèses nulles (analogues de l'hypothèse d'indépendance pour les méthodes factorielles) qui sera selon les cas une hypothèse d'homogénéité ou d'uniformité spatiale de la distribution multidimensionnelle des observations à classer. Toutefois, comme dans le cas des méthodes factorielles, ce type de test, tout en fournissant des repères et un cadre conceptuel intéressant, sera de peu d'utilité pratique, car l'hypothèse d'absence de structure, trop sévère, sera la plupart du temps rejetée. S'il est facile de définir une absence de classe, il est beaucoup plus délicat de définir ce qu'est une classe en dehors de tout contexte pratique.

b _ Validation empirique, calculs de stabilité

Des procédures empiriques, en général variables selon les domaines d'application ou la nature du tableau des données, seront assez largement utilisées. Enfin, des calculs de stabilité, utilisant des méthodes de simulation ou de ré-échantillonnage, permettront d'éprouver la qualité de résultats et de porter une appréciation sur la réalité des classes produites par les algorithmes.

c _ Importance des critères externes

Le rôle des critères externes (connaissances a priori, identification ou caractérisation des classes à partir de variables supplémentaires) sera souvent primordial dans la pratique. Ainsi, une classe mal différenciée, mais identifiée par une catégorie de variable nominale supplémentaire pertinente pour l'utilisateur deviendra, dans bien des cas, digne d'être retenue.

Il existe cependant une différence fondamentale avec les méthodes factorielles : il n'y a pas en classification l'équivalent du théorème d'Eckart et Young (décomposition aux valeurs singulières), et donc pas de paramètres aussi intrinsèques que les valeurs propres[1]. Il existe en revanche une riche flore d'algorithmes dont l'utilisation simultanée sur un même tableau constitue d'ailleurs une épreuve pragmatique de stabilité de structures observées.

---

[1] On a vu au paragraphe 6.4.3, les relations qui peuvent exister dans certains cas entre valeurs propres et indices de niveaux relatifs à une même table de contingence.

Commençons par mentionner quelques travaux de synthèse sur le sujet. Une contribution de Bock (1994) sur les problèmes et l'avenir des méthodes de classification comprend une brève mais dense revue des problèmes de validation. D'autres revues intéressantes sont celles de Gordon (1987) (limitée à la classification hiérarchique), de Hartigan (1985), de Bock (1985), de Perruchet (1983), de Dubes et Jain (1979). Enfin on trouvera plus bas plusieurs références de contributions consacrées à des comparaisons de méthodes.

6.5.2 L'hypothèse d'absence de structure, les modèles

Il existe de nombreux travaux sur ce thème et, à de rares exceptions près, ils ne concernent que les méthodes de classification utilisées de façon isolée.

Dans cet ouvrage où nous considérons les méthodes factorielles et les méthodes de classification comme complémentaires (et devant être utilisées simultanément), on peut donc préconiser sans hésiter, au moins dans un premier temps, les tests d'indépendance ou de sphéricité déjà évoqués à propos des méthodes factorielles. Il est en effet extrêmement improbable que des variations de densité à l'intérieur d'un nuage de points ne se répercutent pas sur une ou plusieurs valeurs propres d'une analyse en axes principaux. On peut objecter qu'un ellipsoïde peut être allongé, mais parfaitement homogène. Dans ce cas, une coupure en deux de son grand axe produit deux classes qui, même si elles ne sont pas séparées par une zone de faible densité, ne peuvent être considérées comme le fruit du hasard. Il s'agit en fait de la meilleure coupure en deux classes de l'échantillon. On voit qu'il faudrait préciser ce que l'on entend par classe. En fait, il y a presque autant de définitions des classes que de critères de classification utilisés pour les obtenir.

Parmi les nombreuses méthodes de classification, les modèles probabilistes permettent de formaliser l'incertitude sur l'appartenance aux classes et de répondre en partie à la question du nombre de classes et de la validation de la structure des classes.

a _ Modèles de mélanges

Le modèle théorique de base le plus répandu est le modèle des mélanges de distributions qui repose sur une approche probabiliste. Ce modèle répond à une grande diversité de situations comme celles des éléments aberrants ou manquants, ou encore des populations hétérogènes.

Le principe est de considérer les données comme étant un échantillon issu d'une population et de s'appuyer sur la distribution de probabilité de cette population pour définir une classification.

L'observation x_i ($i \leq n$) est alors une réalisation d'une variable aléatoire $\mathbf{x}$ de densité $f(x)$:

$$f(x) = \sum_{k=1}^{q} p_k f_k(x), \text{ avec pour tout k, } \quad 0 < p_k < 1 \quad \text{et} \quad \sum_{k=1}^{q} p_k = 1$$

Dans cette formule, $f_k(x)$ est la densité de la classe k (dont la forme doit être spécifiée; par exemple : densité d'une loi normale de moyenne μ_k et de matrice des covariances Σ_k).

On note que le nombre de classes q est supposé connu. Le terme p_k est la probabilité pour une observation d'appartenir à la classe k. Dans ces conditions, l'hypothèse d'absence de structure peut être celle de l'identité des diverses composantes $f_k(x)$ de la densité $f(x)$.

Généralement, on suppose que chaque observation x_i appartient à une seule classe k caractérisée par les paramètres θ_k (moyenne, variance, ...) de la distribution $f_k(x_i, \theta_k)$ et on peut donc estimer les probabilités *Pr(k/x)* d'affectation au k^{ième} modèle.

La classification automatique s'effectue alors en utilisant le cadre suivant :

1. Un individu i est aléatoirement tiré de la population.
2. L'individu est affecté à l'une des q classes, notée k ($1 \leq k \leq q$), à laquelle on associe une distribution $f_k(x_i, \theta_k)$, où x_i est la donnée de l'individu i et θ_k sont les paramètres de la distribution de probabilité.
3. chaque x_i suit la loi de probabilité associée à la classe à laquelle il appartient.

La distribution de probabilité des individus est obtenue par :

$$f(x_i, \theta) = \sum_{k=1}^{q} p_k f_k(x_i, \theta_k)$$

où $\theta = (\theta_1, \ldots, \theta_k, \ldots \theta_q)$ sont les paramètres du modèle qui décrivent au mieux un ensemble d'observations X. Afin d'estimer ces paramètres, on cherche à maximiser la vraisemblance du modèle du mélange :

$$V(x, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{q} p_k f_k(x_i, \theta_k)$$

et l'on a recours pour cela à l'algorithme EM (Expectation-Maximisation) présenté brièvement en annexe de ce chapitre. De nombreux travaux ont été publiés sur l'estimation des mélanges de densités, dont on trouvera une synthèse dans Celeux (1992).

Parmi les premiers travaux sur ce thème, citons ceux de Day (1969), de Bock (1974, 1977) et des ouvrages comme ceux de Bock (1996), McLachlan et Peel (2000)], Govaert (2003). Cette formalisation donne lieu à beaucoup de travaux théoriques intéressants (cf. l'ouvrage de Everitt et Hand, 1981), mais peu d'entre

eux débouchent sur des procédures utilisables en pratique pour valider les classifications ou déterminer le nombre de classes.

b _ Modèles de partitions fixes

Ces modèles de référence supposent l'existence d'une partition inconnue en q classes (I_1, I_2,..., I_q) d'effectifs respectifs (n_1, n_2,..., n_q) avec $\sum_{k=1}^{q} n_k = n$.

A chacune des q classes I_k est associée une densité $f_k(\mathbf{x})$. Dans le cas où les densités $f_k(\mathbf{x})$ sont celles de lois normales sphériques de même matrice des covariances $\sigma^2\mathbf{I}$ et de moyennes μ_k, la partition qui réalise le maximum de vraisemblance est celle qui minimise le critère :

$$cr(q) = \sum_{k=1}^{q} \sum_{i \in I_k} \left\| \mathbf{x}_i - \bar{\mathbf{x}}_k \right\|^2 = \sum_{k=1}^{q} \sum_{i \in I_k} d^2(i, C_k)$$

où l'on a noté, comme au § 6.1.2, C_k le centre de gravité de la classe I_k, de composantes $\bar{\mathbf{x}}_k$. On reconnaît le critère utilisé dans l'agrégation autour de centres mobiles. La partition optimale exacte est actuellement impossible à déterminer, mais la méthode des centres mobiles, on l'a vu, conduit rapidement à un optimum local. Ce critère permet donc, dans le cadre fourni par ce modèle, d'évaluer la qualité d'une partition.

c _ Autre modèles

Une autre modélisation directe de l'hypothèse nulle est l'hypothèse d'homogénéité spatiale, développée par Dubes et Zeng (1987). Ces auteurs s'inspirent des tests de répartition spatiale aléatoire et de processus de Poisson généralisés (cf. par exemple Ripley, 1981) pour explorer, par des simulations extensives, les possibilités de ces épreuves de validation dans le domaine de la classification.

Aux critères qui permettraient de détecter l'existence d'une partition, on peut préférer les critères plus pragmatiques et modestes qui permettraient de comparer deux partitions (ou éventuellement d'améliorer une partition).

Parmi les critères les plus utilisés, citons le critère dit "critère F", quotient de la variance totale inter-classes par la variance totale intra-classes (traces des matrices E et D, matrices utilisées pour calculer les fonctions linéaires discriminantes au chapitre 7), le critère dit "critère de Wilks", quotient des déterminants des deux matrices des covariances précédentes[1].

---

[1] On a pu établir la loi asymptotique des maxima de ces deux critères (maxima calculés sur toutes les partitions possibles) sous l'hypothèse nulle de distributions uniformes ou unimodales (cf. Bock, 1989).

6.5.3 Nombre de classes à retenir

On présentera dans ce paragraphe les méthodes visant à déterminer, par des procédures empiriques (souvent inspirées par les modèles évoqués plus haut) le nombre de classes, sans faire intervenir d'information externe. On examinera tout d'abord le cas de la classification mixte. Puis on évoquera brièvement quelques travaux réalisés à propos de l'existence et de la détermination du nombre des classes.

a _ Cas de la classification mixte

Cette stratégie de classification adaptée à l'analyse exploratoire de grands tableaux (plusieurs milliers d'individus, plusieurs centaines de variables ou modalités) comporte des possibilités de contrôle et de validation dans son processus même de calcul. Nous reprenons les étapes de l'utilisation conjointe de l'analyse factorielle et de la classification mixte du paragraphe 6.4.

1- La première étape de cet enchaînement de méthodes est une analyse en axes principaux, qui permet d'éprouver l'hypothèse d'absence éventuelle de structure, et donne une idée de l'éventuelle concentration du nuage de points à classer dans un sous-espace. Cette étape produit un système de coordonnées euclidiennes que les variables de départ soient numériques (analyse en composantes principales), fréquentielles (analyse des correspondances) ou nominales (analyse des correspondances multiples). On peut alors choisir de garder tous les axes correspondant à des valeurs propres non nulles, ou de tronquer le support de façon à réaliser un filtrage. La possibilité de moduler le nombre d'axes permettra d'éprouver la stabilité des résultats de la classification qui va suivre.

2- La seconde étape est la classification mixte proprement dit.

- Lors de la première phase (partition préliminaire par les centres mobiles destinées à réduire la dimension du problème), la possibilité de calculer des groupements stables (ou formes fortes) constitue une première épreuve de validation, fondée sur l'initialisation aléatoire de la méthode des centres mobiles.

- La seconde phase (classification hiérarchique sur données agrégées utilisant le critère de Ward généralisé, adaptée aux classifications de données regroupées) produit un dendrogramme et un histogramme des indices de niveau (schématisés sur les figures 6.3 - 1 et 6.3 - 2, par exemple) qui permettent d'apprécier les sauts importants de l'indice et donc de proposer, sur inspection visuelle, une coupure de l'arbre hiérarchique, à laquelle correspondra le nombre de classes retenu[1]. Si le critère d'inspection

[1] La consolidation de la coupure par réaffectation des individus (centres mobiles) donne une information importante sur la qualité des résultats. Si l'évolution de la variance inter-classes (par exemple) est trop importante au cours de la consolidation, cela met en question la qualité

visuelle a été retenu ici, c'est par absence de consensus sur les autres critères, nombreux, qui ont été proposés dans la littérature[1].

b _ Cas général

Nous sommes vraiment ici, plus encore que dans le cas de la validation des méthodes factorielles, dans le domaine de la statistique expérimentale. Même si les développements théoriques sont parfois importants, il reste indispensable de tester empiriquement l'adéquation des résultats à la réalité, par simulation et bootstrap ou/et par essai sur des jeux de données-test. On mentionnera les premiers travaux de simulation de Gower et Banfield (1975), qui étudient empiriquement, à partir de plusieurs critères, la distorsion entre la métrique initiale et l'ultramétrique produite par agrégation hiérarchique.

Matusita et Ohsumi (1980) proposent un critère dit d'affinité pour comparer plusieurs partitions dans le cadre d'un algorithme à centres mobiles. Milligan et Cooper (1985) ont étudié et comparé plus de 30 tests et critères par simulation. Wong (1985), Jain et Moreau (1987) utilisent systématiquement le bootstrap pour étudier la stabilité des résultats et en déduire le nombre de classes stables. Hardy (1994) compare 7 critères appliqués à des résultats de classifications issues de six méthodes différentes, chaque couple classification-critère étant appliqué à 4 jeux de données artificiels différents choisis en raison de leurs aptitudes à représenter des situations typiques distinctes.

Rasson et Kubushishi (1994) proposent un nouveau test (Gap test), fondé sur des processus de Poisson stationnaires, qui utilise les éventuelles zones vides entre classes. Testé sur des jeux de données simulées ou classiques, il est efficace pour reconnaître les classes isolées.

c _ Les critères externes

Comme le souligne Bock (1994), il ne faudrait pas exagérer la pertinence et l'importance de la notion de nombre de classes d'une classification, car une classification n'est jamais une fin en soi. C'est beaucoup plus souvent d'une dissection dont on a besoin, selon la terminologie de Kendall (1966) qui considère qu'un découpage de la réalité multidimensionnelle est toujours utile, même si les classes ne sont pas bien séparées, même si tous les individus ne sont pas classés.

---

de la coupure de l'arbre, qui se révèle loin d'un optimum local. Cela doit inciter à la prudence dans le maniement de la partition obtenue.

[1] Mollière (1986, 1989) propose également dans le cadre d'une stratégie d'agrégation mixte d'utiliser le coefficient CCC (Cubic Clustering Criterion) proposé par Sarle (1983) qui est une fonction de R^2 (rapport de la variance interclasses à la variance totale) déterminée empiriquement. Ce coefficient CCC a été considéré comme satisfaisant à l'issue des simulations de Milligan et Cooper (1985).

Que signifie alors un critère global de qualité, qui pourrait nous faire rejeter des traits structuraux importants ? Et quels modèles théoriques pourraient rendre compte d'une situation aussi complexe ?

William et Lance (1965) pensent qu'une classification "ne peut pas être vraie ou fausse, ni probable ou improbable, mais seulement profitable ou non profitable". Cette notion de profitabilité ne peut qu'être externe au tableau de données. Elle est liée au contexte et aux objectifs de la recherche ou de l'étude, aux méta-données (meta-data), c'est-à-dire à l'information sur l'information.

Les procédures de description automatique des classes (cf. § 6.3.2) à partir des variables actives ayant créé la partition, mais aussi à partir de toute l'information externe disponible (ayant le statut de variables supplémentaires, numériques ou nominales) sont des procédures de validation potentielles.

Elles nous disent que telle portion connexe de l'espace engendré par les variables actives présente de l'intérêt vis-à-vis d'autres informations présentes dans la base de données.

6.6 Recherche non supervisée de règles d'associations

Les méthodes de classification constituent, avec les méthodes de visualisations des chapitres 3, 4, et 5 un des outils importants de la fouille de données (*Data Mining*).

Ces méthodes font parties de la famille des méthodes d'apprentissage non-supervisé, au contraire de la régression (chapitre 2) et des méthodes d'analyse discriminante (chapitre 7) qui sont, elles, des méthodes d'apprentissage supervisé. Ces dernières impliquent en effet, on l'a vu, une variable à expliquer (quantitative ou qualitative), et cette variable joue un rôle de « professeur » pendant la phase d'apprentissage du modèle.

La fouille de données, dévolue aux analyses secondaires de très grandes bases de données, est apparue en tant que discipline après le succès des premiers algorithmes de recherches de règles d'association, mis en œuvre essentiellement par des informaticiens.

Ce n'est que plus tard que la fouille de données a intégré dans sa panoplie les outils de la statistique exploratoire (méthodes non-supervisées : analyses en axes principaux et classification) et de la statistique plus décisionnelle (régression, discrimination). Sur les rencontres du *Data Mining* et de la statistique, on pourra consulter les analyses de Hand (1998), de Saporta (2001),

et la contribution de Hébrail et Lechevallier (chapitre 11 de l'ouvrage édité par: Govaert , 2003).

Nous évoquerons dans cette section l'algorithme d'extraction de règles le plus populaire, dit algorithme *Apriori* (Agrawal *et al.*, 1994, 1995), qui est typiquement une méthode non-supervisée, puis nous évoquerons les techniques d'analyse statistique implicative qui s'intéressent à la détection et à la description des relations non-symétriques d'implication.

6.6.1 Algorithme *Apriori* pour la recherche de règles

L'exemple typique de recherche d'association est celui de l'étude du panier de la ménagère (*market basket analysis*). Les tickets de caisse des grandes surfaces génèrent des bases de données géantes (des centaines de milliers de transactions par jour, relatives à des milliers de produits), avec, lorsque les clients utilisent des cartes de paiement spécialisées ou des cartes fidélité, des informations externes relatives au possesseur de la carte.

On imagine aisément qu'il est possible de tirer parti de cette montagne d'information à des fins de marketing, de publicité, de communication, ou plus généralement de gestion.

On a donc au départ un tableau binaire T(n, p) (croisant n *transactions* [tickets, caddies] en ligne, avec p *items* [produits, objets] en colonnes).

On n'étudie dans un premier temps que la présence ou l'absence d'un item dans la transaction.

Sous cette forme, et avec les ordres de grandeur réels ($n = 10^6$, $p = 10^4$) les données sortent de l'épure des principaux logiciels d'analyse des données[1].

L'ensemble des items d'une transaction i s'appelle un *itemset*. C'est l'ensemble des colonnes qui ont un « *1* » pour la ligne i.

a _ Les étapes de l'algorithme

Chaque étape sera une lecture séquentielle des données sur disque (compte tenu de leur volume).

- La première étape de *Apriori* consiste précisément à faire en une lecture ce que l'on ferait également en analyse exploratoire des données : procéder à un « tri-à-plat » des données, et ne retenir que les items ayant une fréquence (un

[1] Pour une exploration ou une visualisation, cela n'a pas grand sens de travailler directement sur de pareils tableaux. Pour les analyses en axes principaux, on peut en effet se demander pourquoi extraire des vecteurs propres à partir de 10^6 lignes, alors qu'en prenant au hasard 10^4 lignes dans la même base, on obtient les mêmes premiers vecteurs propres (une dizaine par exemple), qui suffisent largement pour une visualisation déjà très complexe. Pour une classification de clients à des fins de gestion, il en est autrement. Nous reviendrons sur ce point.

support, dans la terminologie du *Data Mining*) supérieure à *s*, seuil fixé à l'avance (colonnes ayant plus de *s* « *1* »). On ignore définitivement les autres items[1].

[Exemple : (s = *10*) ; Saumon fumé (*34*) ; Bière (*2345*) ; Pain (*6789*) ; Blini (*13*) ; etc]

- La seconde étape consiste, toujours en une lecture, à calculer le support (fréquence) de tous les couples d'items (apparaissant dans une même transaction) formés à partir des survivants de l'étape 1, et à ne retenir que les couples apparaissant plus de *s* fois.

[Exemple : (itemsets incluant les items précédents); (Saumon fumé ∩ Bière) (*14*); (Saumon fumé ∩ Pain) (*21*); (Bière ∩ Pain) (*1181*); (Saumon fumé ∩ Blini) (*12*); (Bière ∩ Blini)(*12*)]

- La $k^{ème}$ étape consiste, en une lecture, à combiner les items survivants de l'étape *k-1* (apparaissant plus de s fois) avec les items sélectionnés à l'étape 1. on ne garde que les *k*-uples apparaissant plus de s fois.

[Exemple pour k = 3 : (s = *10*); (Saumon fumé ∩ Bière ∩ Blini) (*11*); (Saumon fumé ∩ Bière ∩ Pain) (*12*) ;]

La convergence intervient par épuisement des possibilités d'association, et ce d'autant plus rapidement que *s* est grand et que *n* n'est pas trop grand.

b _ Support, Confiance, Confiance attendue, *Lift*

L'algorithme produit donc une liste d'itemsets plus fréquents que *s*.

Chaque itemset *m*, formé de *l(m)* éléments (*l(m)* ≥ 2) peut être décomposé en deux parties disjointes X et Y telles que $m = \{ X \cup Y \}$ de $2^{l(m)-1} - 1$ façons.

On note alors pour chaque décomposition : X ⇒ Y, X est *l'antécédent*, Y est le *conséquent*. La notation $\mathcal{S}$ désigne un support relatif (i.e. : divisé par *n*).

Le support de la règle X ⇒ Y, $\mathcal{S}(X \Rightarrow Y)$ est le support relatif de l'itemset *m* :

$$\mathcal{S}(X \Rightarrow Y) = l(m)/n$$

La confiance de la règle X ⇒ Y , $\mathcal{C}(X \Rightarrow Y)$ est le support de la règle divisé par le support de l'antécédent :

$$\mathcal{C}(X \Rightarrow Y) = \mathcal{S}(X \Rightarrow Y)/\mathcal{S}(X).$$

C'est la fréquence conditionnelle d'observation de Y connaissant X.

La confiance attendue $\mathcal{C}a(X \Rightarrow Y)$ de la règle (X ⇒ Y) est le support relatif $\mathcal{S}(Y)$ de Y.

C'est une estimation de la probabilité *a priori* d'avoir l'itemset Y.

---

[1] La distribution statistique des items étant souvent très dissymétrique, cette élimination des items rares a en général pour effet de réduire considérablement le nombre des items.

Le *Lift* $\mathcal{L}(X \Rightarrow Y)$ de la règle $(X \Rightarrow Y)$ est le quotient de la confiance divisée par la confiance attendue :

$$\mathcal{L}(X \Rightarrow Y) = \mathcal{C}(X \Rightarrow Y) \,/\, \mathcal{C}a\,(X \Rightarrow Y)$$

[Exemples (pour n = *10 000* transactions)
Le support de la règle (Saumon fumé ∩ Bière⇒ Blini) *est 11/10 000 = 0.0011*
La confiance de cette règle est : *11/14 = 0.786*
La confiance attendue de cette règle est : *13/10 000 = 0.0013*
Le *Lift* est donc très exceptionnel : $\mathcal{L} = 0.786 / 0.0013 \approx 604$
Notons que la règle plus simple : (Saumon fumé ⇒ Blini) a un support de *0.0012*, une confiance de *0.35*, une confiance attendue de *0.0013*, et un *Lift* d'environ *271*. La règle : (Bière ⇒ Blini) a un support de *0.0012*, une confiance de *12/2345*, une confiance attendue de *0.0013*, et un *Lift* ≈ *4*.]

Les résultats sont parfois triés en fonction du *Lift* mais le nombre de règles produites restant souvent considérable, il existe de nombreuses propositions de paramètres définissant le « caractère intéressant » d'une règle (*interestingness*)[1].

Il existe par ailleurs de nombreuses variantes et améliorations de cet algorithme.

c_ Règles et visualisation

Dans l'esprit de cet ouvrage et de la démarche qu'il s'attache à décrire, la première chose à faire serait de procéder à une analyse exploratoire du corpus, en utilisant l'enchaînement d'analyse factorielle et de classification (*thémascope*) proposé dans la section 6.4 de ce chapitre.

La première étape est alors la même que celle de l'algorithme *Apriori*, c'est-à-dire une sélection des colonnes du tableau binaire $T(n, p)$ en fonctions de leur totaux. Au lieu de fixer un seuil de fréquence s, on prendra, par exemple, les *1000* colonnes correspondant aux items les plus fréquents, ce qui définit *a posteriori* un seuil s_0. Pour les lignes (transactions), on utilisera un procédure de tirage aléatoire pour garder, par exemple, *20 000* lignes (les deux nouvelles dimensions, $p' = 1000$ et $n' = 20\,000$ sont compatibles avec plusieurs logiciels courants d'analyse exploratoire).

Une analyse des correspondances de la table réduite nous décrira l'ensemble des principales associations entre items dans les transactions. Il n'est pas

---

[1] Il existe de nombreux travaux sur le caractère intéressant des règles. Citons en particulier l'ouvrage de Hilderman et Hamilton (2001), et les travaux de Bayardo et Agrawal (1999), l'étude du caractère surprenant (*surprisingness*) d'une règle de Freitas (1998). Mentionnons également les travaux de Lallich *et al.*(2004) concernant les problèmes statistique de comparaisons multiples posés par la multiplicité des règles, de Morineau et Rakotomalala (2006). Vaillant *et al.* (2004a, 2004b) ont procédé à des études et des comparaisons expérimentales (incluant des classifications) de différentes mesures du caractère potentiellement intéressant des règles.

nécessaire de procéder à une validation bootstrap dans ce cas puisqu'il suffit de retirer un second échantillon de lignes (de transactions) pour vérifier la stabilité de la structure observée.

Une telle structure stable, donc extrapolable à la totalité de la base initiale, n'a aucune raison d'être ignorée. Une carte de Kohonen sur les *1000* items caractérisés par les principales coordonnées factorielles permet également de visualiser de façon vivante les principales associations entre items. Enfin, une classification avec description automatique des classes permet de repérer des classes de transactions fréquentes dont le profil peut être intéressant.

Les *itemsets* correspondant aux règles les plus intéressantes peuvent être projetés en éléments supplémentaires. Le coût d'une telle visualisation est infime par rapport aux coûts de la recherche de règle. L'existence d'une structure d'association globale pour les items les plus fréquents ne peut qu'aider à organiser des résultats pléthoriques et à faciliter leur interprétation. La description de graphes de règles fait aussi partie des préoccupations de l'analyse statistique implicative.

6.6.2 Méthodes d'analyse statistique implicative

L'analyse statistique implicative[1] (ASI), est une méthode d'analyse exploratoire de données qui vise l'extraction (ou la fouille) de règles d'association non symétriques pour modéliser des relations de quasi-implication du type *"si l'on observe a alors on devrait observer b"*.

Ces règles sont représentées sous forme de graphe orienté, et une série d'indices sont élaborés pour en évaluer la pertinence. Elles s'appliquent à des variables nominales ordinales ou des fréquences qui peuvent être ramenées à des variables binaires, ce que l'on va considérer par la suite.

a – Extraction de règles, indices d'implication et graphe orienté

On considère un ensemble I de n individus et M=$\{a_1, a_2, a_3, b_1, b_2, c_1, c_2, c_3, c_4,.\}$ l'ensemble des p modalités ou attributs ou items[2].

Dans la construction de règles, l'ensemble M est ramené à des items de variables binaires : M = $\{a, \bar{a}, b, \bar{b}, c, \bar{c},...\}$ où a (respectivement b et c) est une modalité et $\bar{a}$ (respectivement $\bar{b}$ et $\bar{c}$) est le complémentaire de a (respectivement b et c) [3].

---

[1] Cf.Gras et Lahrer (1992), Gras *et al.*(2001).

[2] Nous retrouvons ici une situation identique à celle de l'analyse des correspondances multiples (cf. chapitre 5).

[3] Par exemple si $a=a_1$ alors $\bar{a} = \{a_2, a_3\}$.

Ainsi la règle "a → b" est vraie "si pour tout individu i qui vérifie a alors il vérifie presque b". a est la prémisse et b est la conclusion de la règle. Le mot règle n'a pas le même sens qu'au paragraphe 6.6.1.

Une telle règle correspond au schéma suivant (cf fig. 6.6.1) où A (resp. B) est le sous-ensemble de I des n_a (resp. n_b) individus qui vérifient a (resp.b):

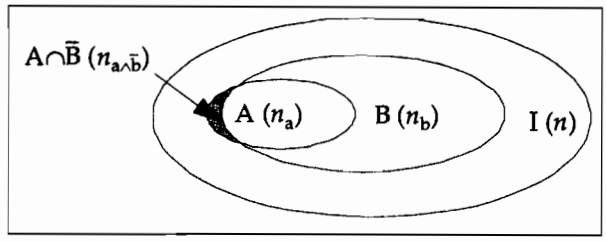


Figure 6.6-1 : Schéma associé à la règle "a → b"

Elle est associée à une table de contingence (cf tab. 6.6.1.) :

Tableau 6.6-1 : Table de contingence associée à la règle "a → b"

	b	$\bar{b}$	
a	$n_{a\wedge b}$	$n_{a\wedge\bar{b}}$	n_a
$\bar{a}$	$n_{\bar{a}\wedge b}$	$n_{\bar{a}\wedge\bar{b}}$	$n_{\bar{a}}$
	n_b	$n_{\bar{b}}$	n

Comme pour la recherche de règles *Apriori* (cf. paragraphe précédent), on définit le support et la confiance d'une règle, mais avec les notations suivantes :

$$support(a \rightarrow b) = \frac{n_{ab}}{n} \quad \text{et} \quad confiance(a \rightarrow b) = \frac{n_{ab}}{n_a}$$

Les algorithmes pour la recherche de règles d'association s'appuient encore sur des seuils minimaux du support et de la confiance.

b – Mesure et évaluation de règles

Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Gras, dans la continuité des travaux de Lerman (1970, 1981) sur la similarité, propose plusieurs mesures d'évaluation de la qualité d'une règle à partir de ses contre-exemples $n_{a\wedge\bar{b}}$ [1].

[1] En analyse de données classique c'est l'écart entre l'effectif observé et le théorique qui est pris en compte. En analyse statistique implicative, c'est l'écart entre le contre-exemple observé, circonstance dans laquelle l'implication est mise en défaut, et le théorique qui est étudié.

L'*indice d'implication* q d'une règle est un indicateur de *non-implication* de a sur b :

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$$

Soient X et Y deux parties quelconques de I, tirées de façon aléatoire et indépendante, et de mêmes cardinaux respectivement que A et B. La variable aléatoire $\mathrm{Card}(X \cap \bar{Y})$ est le nombre de contre-exemple dans ce tirage (cf. fig. 6.6.2).

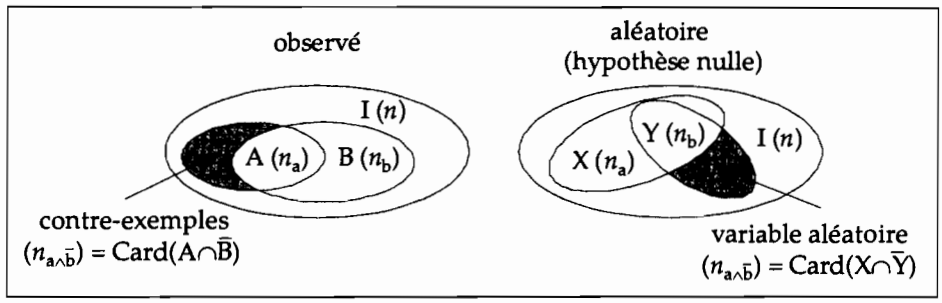


figure 6.6.2 : Observations et tirage aléatoire

La règle "a → b" sera admissible si $\mathrm{Card}(A \cap \bar{B})$ est particulièrement petit par rapport à $\mathrm{Card}(X \cap \bar{Y})$ c'est-à-dire si $\Pr\left[\mathrm{Card}(X \cap \bar{Y}) \leq \mathrm{Card}(A \cap \bar{B})\right]$ est faible.

L'indice d'intensité d'implication permet alors d'évaluer la pertinence des tendances implicatives associées et s'exprime de la façon suivante :

$$\varphi(a,b) = 1 - \Pr(\mathrm{Card}(X \cap \bar{Y}) \leq \mathrm{Card}(A \cap \bar{B})) \text{ si } n_b \neq n$$

$$\text{et } \varphi(a,b) = 0 \text{ sinon}$$

La règle "a → b" est retenue pour un seuil α donné si $\varphi(a,b) \geq 1 - \alpha$.

Cependant lorsque la population croît, l'intensité d'implication n'est plus assez discriminante. On utilise alors un indice d'inclusion de A dans B qui intègre la réalisation d'un faible nombre de contre-exemples, d'une part à la règle "a → b" et d'autre part à la règle "non b → n on a" (cf. Gras *et al.*, 2003).

De nombreux autres indices ont été élaborés pour évaluer une règle. On trouvera une liste complète dans Gras *et al.* (2004).

c – Graphes de règles

Le nombre de règles est très vite important et il convient alors d'organiser les ensembles de règles. La représentation d'un ensemble de règles est par la suite donnée par un graphe orienté et pondéré par les intensités d'implication (cf. fig

6.6.3). Si l'on considère par exemple 6 variables a, b, c, d, e, f, répondant aux règles suivantes :

"c → a"; "c → e"; "a → e"; "a → d"; "e → b"; "d → b"; "e → f'

L'ensemble de ces règles est traduit par le graphe suivant :

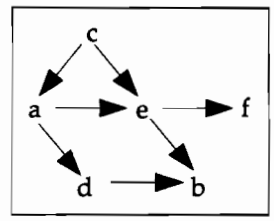


Figure 6.6.3. Exemple de règles et de graphe associé

Le graphe peut être plus réduit (respectivement élargi) selon que le seuil α de l'intensité d'implication est élevé (respectivement abaissé).

Lorsque les règles sont nombreuses et complexes comme cela est souvent le cas, Gras, Kuntz et Briand (2003) proposent de structurer des ensembles de règles en définissant des R-Règles (règles de règles, ou méta-règles) et des indices de cohésion, pouvant conduire à des hiérarchies orientées de règles.

Concluons ce très bref survol en évoquant d'autres directions de recherche, pour la mise en évidence de règles, fondées sur les méthodes de segmentation, ou plus généralement sur les arbres d'induction (Zighed et Rakotomalala, 2000).

6.7 Annexe technique du chapitre 6

Cette annexe technique comprend deux parties contenant respectivement un complément sur les *correspondances hiérarchiques* évoquées au paragraphe 6.4.2, et des précisions sur l'*algorithme EM*, qui intervient en particulier dans l'estimation des modèles de mélanges (paragraphe 6.5.2).

6.7.1 Les correspondances hiérarchiques

D'une manière générale, dans une *hiérarchie binaire* H sur un ensemble I à n éléments chaque élément non terminal $h \in H$ peut être partitionné de façon unique en deux éléments $a(h)$ et $b(h)$:

$$h = a(h) \cup b(h) \qquad \text{avec } a(h) \in H \text{ et } b(h) \in H$$

On suppose cette hiérarchie indicée (cf. § 6.2.2). On suppose également que l'indice $\lambda(h)$ prend ses valeurs dans $[0,1]$ et qu'il est nul pour les éléments terminaux. Chaque élément $i \in I$ est d'autre part muni d'une masse p_i strictement positive avec :

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

Pour chaque nœud h de la hiérarchie, on peut associer une fonction sur I à valeurs réelles f_h, de moyenne nulle, c'est-à-dire telle que :

$$\sum_{i=1}^{n} p_i f_h(i) = 0$$

Cette fonction est nulle en dehors de h ($i \notin h \Rightarrow f_h(i) = 0$) et constante sur chacun des deux nœuds $a(h)$ et $b(h)$ qui constituent h.

Ces constantes sont définies par les formules suivantes, en notant p_h, p_a et p_b les masses respectives des éléments h, $a(h)$ et $b(h)$:

$$f_h(i) = \sqrt{\frac{p_b}{p_h p_a}} \qquad \text{pour } i \in a(h)$$

$$f_h(i) = -\sqrt{\frac{p_a}{p_h p_b}} \qquad \text{pour } i \in b(h)$$

Si l'on munit l'espace des fonctions f_h du produit scalaire :

$$< f_h, f_{h'} > = \sum_{i=1}^{n} p_i f_h(i) f_{h'}(i)$$

On vérifie facilement que les fonctions f_h sont de norme (ou de variance) 1 et que les n-1 fonctions correspondant aux nœuds de la hiérarchie constituent une base orthonormée de l'ensemble des fonctions sur I.

La formule de reconstitution des données en analyse des correspondances (cf. § 4.2.2) permet alors de générer un tableau symétrique **C** de terme général $c_{ii'}$:

$$c_{ii'} = p_i p_{i'} \left(1 + \sum_{h=1}^{n-1} \sqrt{\lambda_h}\, f_h(i) f_h(i') \right)$$

les n-1 nœuds repérés par h étant supposés numérotés par ordre d'indices d'agrégation λ_h décroissants.

La table de contingence $\mathbf{K}_{IJ}$ de la section 6.4.2 a été générée[1] de cette façon.

---

[1] On trouvera la preuve de la non-négativité des termes $c_{ii'}$ dans Benzécri (1973, Tome IIB, Chapitre 11).

6.7.2 L'algorithme EM

Si l'on considère un échantillon $X = (x_1,...,x_i,...,x_n)$ d'individus suivant une loi $f(X,\theta)$ paramétrée par $\theta = (\theta_1, ..., \theta_k, ..., \theta_q)$, on cherche à déterminer ce paramètre maximisant la vraisemblance d'un modèle de mélange donnée par :

$$V(X,\theta) = \prod_{i=1}^{n} \sum_{k=1}^{q} p_k f_k(x_i,\theta_k)$$

Notons que tous les mélanges de lois ne sont pas identifiables[1].

Pour faciliter les calculs on utilise la log-vraisemblance :

$$L(X,p,\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{q} p_k f_k(x_i,\theta_k)\right)$$

La méthode EM (*Expectation – Maximisation*) , (cf. Dempster et al., 1977 ; Bishop, 1995 ; Celeux, 1992 ; Govaert, 2003) est un des algorithmes d'optimisation les plus employés.

Au départ, la forme analytique de la fonction f est spécifiée (en général de la famille des lois exponentielles : Normale, Poisson, exponentielle), et les valeurs de départ des q composantes des vecteurs p (p^0) et θ (θ^0) sont fixées.

La méthode d'optimisation EM est itérative en deux étapes :

- ► dans l'étape E (itération r) on estime les probabilités conditionnelles suivantes, pour tout individu à classer i ($i \leq n$) et pour toute classe k ($k \leq q$) :

$$t_k^r(x_i) = \frac{p_k^r f(x_i,\theta_k^r)}{\sum_{j=1}^{q} p_j^r f(x_i,\theta_j^r)}$$

$t_k^r(x_i)$ est la probabilité conditionnelle d'appartenance de x_i au composant k à la r-ème itération, connaissant les paramètres.

► dans l'étape M (itération $r+1$) :

On calcule $p_k^{r+1} = \frac{1}{n}\sum_{i=1}^{n} t_k^r(x_i)$, et on détermine une approximation du modèle en maximisant la log-vraisemblance, ce qui conduit à résoudre les équations suivantes (le vecteur θ_k est supposé avoir p composantes) :

---

[1] On montre que les mélanges de lois normales, de lois exponentielles, de lois de Poisson, sont identifiables. Ce n'est pas le cas des lois binomiales et des lois uniformes. Une loi uniforme u sur [0, 1] peut être décomposée d'une infinité de façon comme un mélange de deux lois uniformes : ainsi, quel que soit $0 < p < 1$, on peut avoir $u[0, 1] = pu[0, p] + (1-p)u[p, 1]$, u[x,y] désignant une loi uniforme entre x et y.

$$\sum_i t_k^r(x_i) \frac{\partial \log f_k(x_i, \theta_k)}{\partial \theta_k^\alpha} = 0 \text{ , pour tout } k \leq q \text{ et pour tout } \alpha \leq p$$

Evidemment, la résolution de ces équations dépend de la forme analytique de la loi f.

Le résultat est étonnamment simple dans le cas de la loi normale. Si f_k est une loi normale de vecteur moyen m_k et de matrice des covariance S_k, l'estimation de m_k à l'itération r+1 s'écrit :

$$m_k^{r+1} = \frac{\sum_i t_k^r(x_i) x_i}{\sum_i t_k^r(x_i)}$$

et l'estimation de la matrice des covariances :

$$S_k^{r+1} = \frac{\sum_i t_k^r(x_i)(x_i - m_k^{r+1})(x_i - m_k^{r+1})'}{\sum_i t_k^r(x_i)}$$

Autrement dit, il s'agit d'estimations classiques où les probabilités conditionnelles calculées à l'étape r interviennent comme simples coefficients de pondération des observations.

La procédure itérative est arrêtée quand la convergence de la vraisemblance est atteinte (pour les propriétés de convergence voir Wu 1983, Jordan et Xu 1996, McLachlan et Krishnan 1997)[1].

[1] La procédure EM dans le cas de la classification a été étendue en procédure SEM (S pour : *stochastic*) par Celeux et Diebolt (1985), et CEM (C pour classification) (Celeux et Govaert, 1992)

Chapitre 7

Analyse discriminante, classification supervisée

On désigne sous le nom d'*analyse discriminante* une famille de techniques destinées à classer (affecter à des classes préexistantes) des individus caractérisés par un certain nombre de variables numériques ou nominales. Le principe de la démarche remonte aux travaux de Fisher (1936) ou, de façon moins directe, à ceux de Mahalanobis (1936). Elle est une des techniques d'analyse multidimensionnelle les plus utilisées en pratique (Credit-scoring, diagnostic automatique, contrôle de qualité, prévision de risques, reconnaissance des formes). En théorie de l'apprentissage et dans d'autres domaines d'application, cette famille de techniques porte souvent le nom de « classification supervisée », dénomination qui indique qu'il s'agit bien de classement, et non de classification (*clustering*), d'où, par souci de communication interdisciplinaire, le nom du présent chapitre.

L'*analyse factorielle discriminante* ou *analyse linéaire discriminante*, est une méthode à la fois descriptive et prédictive, qui donne lieu, comme les méthodes factorielles présentées au chapitre 1, à des calculs d'axes principaux. Elle peut être considérée comme une extension de la régression multiple (cf chapitre 2) dans le cas où la variable à expliquer est nominale et constitue la variable de partition. Ces deux techniques constituent d'ailleurs des cas particuliers de l'analyse canonique (cf. chapitre 2). L'analyse linéaire discriminante reste une méthode de référence face au flux permanent de méthodes concurrentes.

La première section de ce chapitre est consacrée à l'analyse linéaire discriminante, parfois appelée analyse discriminante de Fisher, ou analyse canonique discriminante, ou encore analyse factorielle discriminante, ces diverses dénominations reflétant l'importance historique et théorique de la méthode. Puis nous étudierons dans une seconde section les liens de l'analyse linéaire discriminante avec l'analyse canonique, la régression, l'analyse des correspondances. La troisième section sera dévolue aux règles de classement.

Les procédures de régularisation (destinées à résoudre certaines difficultés de calcul et à rendre les résultats plus robustes) font l'objet de la quatrième section. Les sections 5 et 6 sont consacrées respectivement à la régression logistique et aux méthodes de segmentation. Enfin dans la septième et dernière section, on évoquera les méthodes de discrimination neuronales liées à la théorie de l'apprentissage. Nous ne présenterons pas toutes les techniques d'analyse discriminante qui donnent lieu à une littérature très étendue. Nous renvoyons le lecteur à des ouvrages spécifiques sur la question, notamment l'ouvrage de Tomassone et *al.* (1988) et les ouvrages édités par Celeux (1990) (discrimination à partir de variables continues) et Celeux et Nakache (1994) (discrimination à partir de variables qualitatives), de Bardos (2001) (ouvrage orienté vers les applications dans le secteur bancaire)[1].

7.1 Analyse linéaire discriminante

7.1.1 Formulation du problème et notations

On dispose de n individus ou observations décrits par un ensemble de p variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ et répartis en q classes définies a priori par la variable $\mathbf{y}$ nominale à q modalités[2]. L'analyse discriminante se propose dans un premier temps de séparer au mieux les q classes à l'aide des p variables explicatives. Dans un deuxième temps, elle cherche à résoudre le problème de l'affectation d'individus nouveaux, caractérisés par les p variables, à certaines classes déjà identifiées sur l'échantillon des n individus (appelé *échantillon d'apprentissage).* On distingue par conséquent deux démarches successives, d'ordre descriptif puis décisionnel :

- chercher des fonctions linéaires discriminantes sur l'échantillon d'apprentissage de taille n qui sont les combinaisons linéaires des variables explicatives $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ dont les valeurs séparent au mieux les q classes.
- connaître la classe d'affectation de n' nouveaux individus décrits par les variables explicatives $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$. Il s'agit ici d'un problème de *classement*

---

[1] Signalons dans la littérature de langue anglaise l'ouvrage de synthèse (riche de plus de *1200* références) de McLachlan (1992) et les articles, également de synthèse, de Lachenbruch et Goldstein (1979), de Gnanadesikan (1989); parmi les manuels classiques généralistes qui traitent de l'analyse discriminante, Anderson (1958), Cacoullos (1973), Krishnaiah et Kanal (1982) ; parmi les manuels plus spécialisés, Goldstein et Dillon (1978), Hand (1981). Dans le domaine des méthodes statistiques de la reconnaissance des formes, outre l'ouvrage précité de McLachlan, les ouvrages de base sont Fukunaga (1972), Duda et Hart (1973), Devijver et Kittler (1982). Agrawala (1977) contient des réimpressions de références historiques.

[2] Dans ce chapitre, le vecteur $\mathbf{y}$ a des composantes entières donnant les numéros des classes, et $\mathbf{Y}$ désigne le tableau disjonctif d'ordre (n,q) correspondant.

dans des classes préexistantes, par opposition au problème de *classification* (traité au chapitre 6) qui consiste à construire des classes les plus homogènes possibles dans un échantillon.

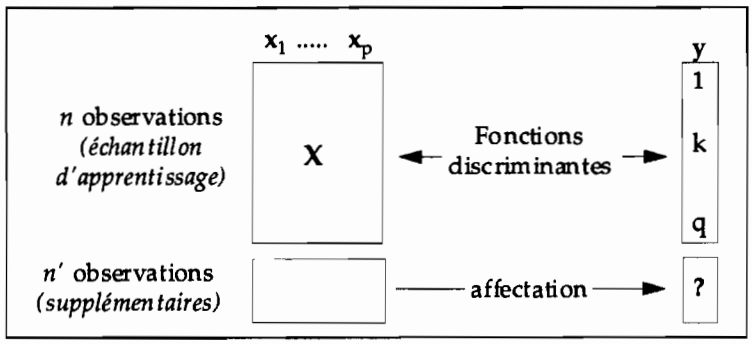


Figure 7.1 – 1. Principe de l'analyse discriminante

Considérons pour fixer les idées le tableau de données *(200, 30)* qui contient, pour $n = 200$ malades, les valeurs de $p = 30$ variables issues d'analyses biologiques et d'examens cliniques. Il existe par ailleurs une partition de ces *200* malades selon $q = 3$ catégories de diagnostics réalisés après des interventions beaucoup plus coûteuses que les *30* mesures précédentes. On se pose la question suivante : étant donné des patients supplémentaires (en nombre n') sur lequel on réalise les *30* analyses et examens, peut-on *prévoir* leurs catégories de diagnostic ? La question répond ici à un besoin pratique[1] : est-ce que des mesures nombreuses mais d'accès facile peuvent contenir une information sur un phénomène ou un état plus difficile à identifier ?

Soit le tableau des données **X** à n lignes (individus ou observations) et p colonnes (variables), de terme général x_{ij}. Les n individus sont partitionnés en q classes. Chaque classe k caractérise un sous-nuage I_k de n_k individus i avec :

$$\sum_{k=1}^{q} n_k = n$$

Par $\bar{x}_{kj}$ on désigne la moyenne de la variable x_j dans la classe k. C'est la $j^{ème}$ coordonnée du centre de gravité G_k du sous-nuage I_k :

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} x_{ij} = G_{kj}$$

[1] Les exemples les plus classiques d'analyse discriminante appartiennent sans doute au domaine médical (aide au diagnostic, aide à la décision en matière d'intervention) mais de nombreuses applications se développent dans le domaine du scoring bancaire (prévision de l'éventuelle défaillance d'un débiteur), du contrôle de qualité (prévision de qualité d'un produit en agro-industrie à partir de mesures externes) et surtout de la reconnaissance des formes (reconnaissances de caractères manuscrits ou d'images-radar, etc.).

La moyenne de la variable x_j sur l'ensemble des individus qui correspond à la $j^{ème}$ coordonnée du centre de gravité G du nuage des individus vaut :

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \sum_{k=1}^{q} \frac{n_k}{n}\bar{x}_{kj} = G_j$$

7.1.2 Fonctions linéaires discriminantes

L'analyse factorielle discriminante consiste à rechercher les combinaisons linéaires de p variables explicatives (x_1, x_2, ..., x_p), généralement continues, qui permettent de séparer au mieux les q classes.

La première combinaison linéaire sera celle dont la variance entre les classes (inter-classes) est maximale, afin d'exalter les différences entre les classes, et dont la variance à l'intérieur des classes (intra-classes) minimale pour que l'étendue dans les classes soit délimitée. Puis, parmi les combinaisons linéaires non corrélées à la première, on recherchera celle qui discrimine le mieux les classes, etc.

Ces combinaisons linéaires seront les *fonctions linéaires discriminantes*.

Désignons par *a(i)* la valeur, pour l'individu *i*, d'une combinaison linéaire **a** des p variables préalablement centrées :

$$a(i) = \sum_{j=1}^{p} a_j(x_{ij} - \bar{x}_j)$$

La variance *var*(**a**) de la nouvelle variable synthétique *a(i)* vaut, puisque *a(i)* est centrée :

$$var(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n} a^2(i) = \frac{1}{n}\sum_{i=1}^{n}[\sum_{j=1}^{p} a_j(x_{ij} - \bar{x}_j)]^2$$

$$var(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{j'=1}^{p} a_j a_{j'}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

En intervertissant les sommations et en posant :

$$t_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) = cov(x_j, x_{j'})$$

la variance de la combinaison des variables **a** peut s'écrire :

$$var(\mathbf{a}) = \sum_{j=1}^{p}\sum_{j'=1}^{p} a_j a_{j'} cov(x_j, x_{j'}) = \mathbf{a}'\mathbf{T}\mathbf{a}$$

où **a** désigne le vecteur dont les p composantes sont a_1, ..., a_p et **T** désigne la matrice des covariances des p variables, de terme général $t_{jj'}$.

Nous allons montrer que la variance de **a** se décompose en variance intra-classes et en variance inter-classes, ce qui correspond à une décomposition analogue de la matrice des covariances **T**.

a – Décomposition de la matrice de covariance

La covariance totale entre deux variables $\mathbf{x}_j$ et $\mathbf{x}_{j'}$ s'écrit :

$$cov(x_j, x_{j'}) = \frac{1}{n}\sum_{k=1}^{q}\left[\sum_{i \in I_k}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})\right] = t_{jj'}$$

Comme en *analyse de la variance*, nous allons décomposer $cov(x_j, x_{j'})$ en somme de covariances *intra-classes* (à l'intérieur des classes) et covariances *inter-classes* (entre les classes). Pour cela nous partirons de l'identité, pour *i, j, k* :

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j)$$

La somme entre crochets dans la formule de la covariance se décompose alors en quatre termes, dont deux sont nuls.
En effet, par définition de $\bar{x}_{kj}$:

$$\sum_{i \in I_k}(x_{ij} - \bar{x}_{kj})(x_{kj'} - \bar{x}_{j'}) = (x_{kj'} - \bar{x}_{j'})\sum_{i \in I_k}(x_{ij} - \bar{x}_{kj}) = 0$$

de façon analogue, les sommes ci-dessous s'annulent :

$$\sum_{i \in I_k}(x_{kj} - \bar{x}_j)(x_{ij'} - \bar{x}_{kj'}) = 0$$

Il reste la formule dite *formule de décomposition de Huyghens* (ou équation d'analyse de la variance) :

$$t_{jj'} = d_{jj'} + e_{jj'}$$

avec :

$$d_{jj'} = \frac{1}{n}\sum_{k=1}^{q}\sum_{i \in I_k}(x_{ij} - \bar{x}_{kj})(x_{ij'} - \bar{x}_{kj'})$$

$$e_{jj'} = \sum_{k+1}^{q}\frac{n_k}{n}(\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

Ces p^2 relations se notent sous forme matricielle[1].:

$$\mathbf{T} = \mathbf{D} + \mathbf{E} \qquad [7.1 - 1]$$

Ainsi, la variance d'une combinaison linéaire **a** des variables se décompose d'après la relation [3.3-1] en variance interne et variance externe :

$$\mathbf{a'Ta} = \mathbf{a'Da} + \mathbf{a'Ea} \qquad [7.1 - 2]$$

Rappelons que, parmi toutes les combinaisons linéaires des variables, on cherche celles qui ont une variance intra-classe minimale et une variance inter-classes maximale. En projection sur l'axe discriminant **a,** chaque sous-nuage doit être, si possible, bien regroupé et bien séparé des autres sous-nuages. Il s'agit donc de chercher **a** tel que le quotient **a'Ea/a'Da** soit maximal (ou **a'Da/a'Ea** minimal).

---

[1] La matrice des covariances Totale T se décompose en une matrice d'inertie intra-classes D (*Dans* les classes) et une matrice d'inertie inter-classes E (*Entre* les classes).

D'après la relation [7.1-2] il est équivalent de minimiser **a'Da/a'Ea** ou de rendre maximal $f(\mathbf{a})$ tel que :

$$f(\mathbf{a}) = \frac{\mathbf{a'Ea}}{\mathbf{a'Ta}}$$

b – Calcul des fonctions linéaires discriminantes

La fonction $f(\mathbf{a})$ à maximiser est le rapport de la variance inter-classes à la variance totale. Cette fonction étant homogène de degré *0* en **a** (invariante si **a** est changé en $\xi\,\mathbf{a}$, ξ étant un scalaire quelconque non nul), il est équivalent de chercher le maximum de la forme quadratique **a'Ea** sous la *contrainte* : **a'Ta** = 1. Ceci conduit à la relation[1] :

$$\mathbf{Ea} = \lambda\mathbf{Ta} \qquad [7.1 - 3]$$

Lorsque la matrice des covariances **T** est inversible, on obtient :

$$\mathbf{T^{-1}Ea} = \lambda\mathbf{a}$$

a est vecteur propre de $\mathbf{T^{-1}E}$ relatif à la plus grande valeur propre λ.

En prémultipliant les deux membres de [7.1 - 3] par le vecteur **a'** on constate que **a'Ea**, le maximum cherché, n'est autre que λ. La plus grande valeur propre λ, quotient de la variance *externe* par la variance *totale*, est inférieure à *1* d'après la relation [7.1 - 1]. On l'appelle *pouvoir discriminant* de la fonction **a**.

Remarque

En rendant maximum le quotient **b'Eb/b'Db** les combinaisons linéaires discriminantes **b** seraient alors les vecteurs propres de la matrice $\mathbf{D^{-1}E}$ où la matrice $\mathbf{D^{-1}}$ définit la *métrique de Mahalanobis* (cf. section 7.2.4 et annexe 7.8). La valeur propre μ correspondant, solution de $\mathbf{D^{-1}Eb} = \mu\mathbf{b}$ est reliée à λ par la formule :

$$\mu = \frac{\lambda}{1-\lambda}$$

On a évidemment $\mu \geq \lambda$, puisque la variance interne est toujours inférieure à la variance totale. Le vecteur **b** est comme **a** solution de l'équation [7.1 - 3] mais doit respecter la contrainte **b'Db** = *1*. Les vecteurs **a** et **b** sont liés par la relation[2] :

$$\mathbf{a} = \left(\sqrt{1-\lambda}\right)\mathbf{b}$$

c – Diagonalisation d'une matrice symétrique

La matrice $\mathbf{T^{-1}E}$ n'est pas symétrique. Mais il est possible de se ramener à la diagonalisation d'une matrice *(q,q)* symétrique. (Rappelons que *p* est le nombre

---

[1] Comme en analyse générale (chapitre 1) ou en analyse canonique (chapitre 2), nous sommes conduits à annuler le vecteur des dérivées partielles du *lagrangien* $\mathcal{L} = \mathbf{a'Ea} - \lambda(\mathbf{a'Ta} - 1)$ par rapport à **a**, ce qui donne la relation : $2\mathbf{Ea} - 2\lambda\mathbf{Ta} = 0$, d'où finalement $\mathbf{Ea} = \lambda\mathbf{Ta}$.

[2] Posant $\mathbf{a} = \xi\,\mathbf{b}$, les deux relations **a'Ea** = λ et **b'Eb** = μ conduisent à la relation $\xi^2\,\mathbf{b'Eb} = \lambda$, d'où: $\xi^2\mu = \lambda$ et $\xi = \sqrt{1-\lambda}$

de variables et q le nombre de classes, avec, dans la plupart des applications $q < p$). En effet la matrice **E**, de terme général :

$$e_{jj'} = \sum_{k=1}^{q} \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

est le produit d'une matrice **C** à p lignes et q colonnes par sa transposée; cette matrice **C** a pour terme général :

$$c_{jk} = \sqrt{\frac{n_k}{n}} (\bar{x}_{kj} - \bar{x}_j) \qquad [7.1 - 4]$$

Avec la décomposition $\mathbf{E} = \mathbf{CC}'$, la relation [7.1 - 3] s'écrit :

$$\mathbf{CC'a} = \lambda \mathbf{Ta}$$

Posons :

$$\mathbf{a} = \mathbf{T}^{-1}\mathbf{Cw} \qquad [7.1 - 5]$$

cette relation s'écrit alors :

$$\mathbf{CC'T}^{-1}\mathbf{Cw} = \lambda \mathbf{Cw} \qquad [7.1 - 6]$$

Il est clair que tout vecteur propre **w** relatif à une valeur propre λ (différente de *0*) de la matrice symétrique $\mathbf{C'T}^{-1}\mathbf{C}$ d'ordre (q, q) vérifie également [7.1-6]. Le vecteur **a** et le scalaire λ vérifient alors la relation [7.1 - 3]. Il suffit en pratique d'effectuer la diagonalisation de cette matrice symétrique, puis d'en déduire **a** par la transformation [7.1 - 5]. De plus cette matrice symétrique d'ordre (q, q) sera en général plus petite que la matrice non-symétrique $\mathbf{T}^{-1}\mathbf{E}$ d'ordre (p, p).

7.2 Lien avec d'autres méthodes

Lorsque la variable **y** ne prend que deux valeurs, chacune caractérisant une classe, des simplifications apparaissent. L'analyse discriminante est alors un cas particulier de la régression multiple. Elle est également un cas particulier de l'analyse canonique lorsque l'un des deux ensembles de variables est formé par les indicatrices d'une partition. Lorsque les deux ensembles sont formés de variables indicatrices, on retrouve l'analyse des correspondances (cf. chapitre 4), qui est ainsi une double analyse discriminante. On peut également présenter la méthode comme une analyse en axes principaux du nuage des points moyens dans une métrique particulière.

7.2.1 Cas de deux classes : équivalence avec la régression multiple

On repérera les deux classes par les indices *1* et *2*. La matrice des covariances **E** entre classes a pour terme général :

$$e_{jj'} = \frac{n_1}{n}(\bar{x}_{1j} - \bar{x}_j)(\bar{x}_{1j'} - \bar{x}_{j'}) + \frac{n_2}{n}(\bar{x}_{2j} - \bar{x}_j)(\bar{x}_{2j'} - \bar{x}_{j'})$$

avec :

$$\bar{x}_j = \frac{n_1}{n}\bar{x}_{1j} + \frac{n_2}{n}\bar{x}_{2j}$$

En remplaçant $\bar{x}_j$ par sa valeur et en tenant compte du fait que $n_1 + n_2 = n$, on trouve :

$$e_{jj'} = \frac{n_1 n_2}{n^2}(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1j'} - \bar{x}_{2j'})$$

La matrice symétrique **E** d'ordre *(p,p)* et de rang *1*, peut être considérée comme le produit d'une matrice colonne **c** par sa transposée :

$$\mathbf{E} = \mathbf{cc}'$$

avec :

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

La relation [7.1 - 3] s'écrit alors :

$$\mathbf{T}^{-1}\mathbf{cc}'\mathbf{a} = \lambda\mathbf{a}$$

Prémultiplions les deux membres par $\mathbf{c}'$:

$$[\mathbf{c}'\mathbf{T}^{-1}\mathbf{c}]\mathbf{c}'\mathbf{a} = \lambda\mathbf{c}'\mathbf{a}$$

La quantité entre crochets est un scalaire, égal par conséquent à λ qui est ici une valeur propre unique car **E** est de rang *1*.

Cette valeur propre vaut donc : $\lambda = \mathbf{c}'\mathbf{T}^{-1}\mathbf{c}$

λ est appelée *distance généralisée* entre les deux classes ou encore *"Distance de Mahalanobis"*. Le vecteur propre correspondant :

$$\mathbf{a} = \mathbf{T}^{-1}\mathbf{c}$$

est l'unique fonction discriminante.

Considérons un vecteur **w** à n composantes, défini par :

$$w_i = \begin{cases} \sqrt{n_2/n_1} & \text{si l'individu } i \text{ est membre de la classe } 1 \\ -\sqrt{n_1/n_2} & \text{si l'individu } i \text{ est membre de la classe } 2 \end{cases}$$

La régression multiple expliquant **w** par les colonnes de **X** conduit au vecteur de coefficients noté ici **b** :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}, \qquad \text{avec :} \quad \frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{T}$$

On vérifie que :

$$\frac{1}{n}\mathbf{X}'\mathbf{w} = \mathbf{c}, \quad \text{d'où:} \quad \mathbf{b} = \mathbf{T}^{-1}\mathbf{c}$$

Le vecteur des *coefficients de régression* **b** coïncide par conséquent avec le vecteur des composantes de la *fonction discriminante* **a** calculé précédemment.

7.2.2 Lien avec l'analyse canonique

Comme en analyse des correspondances multiples, la variable nominale à q classes sera représentée par un codage disjonctif complet. On construit ainsi une matrice **Y** à n lignes et q colonnes de terme général y_{ik} valant *1* si l'individu i appartient à la classe k ou *0* sinon. Autrement dit, nous ajoutons aux variables initiales **X** des variables *artificielles* **Y** qui indiquent l'appartenance aux classes.

Les p colonnes des variables observées du sous-tableau **X** (cf figure 7.2-1) seront centrées et notées $\hat{\mathbf{X}}$. Nous poserons : $\hat{x}_{ij} = x_{ij} - \overline{x}_j$

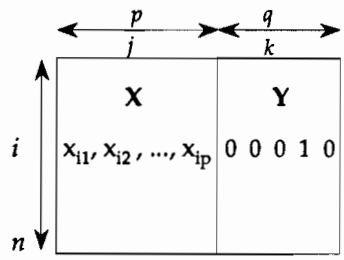


Figure 7.2 – 1. Tableau de données [X,Y]

Notons qu'à la différence de l'analyse canonique, les colonnes de **Y** ne sont pas centrées : la somme des éléments de la $k^{ème}$ colonne vaut n_k.

L'analyse canonique du tableau $[\hat{\mathbf{X}}, \mathbf{Y}]$ conduit à chercher le vecteur propre **a** de la matrice **N** (formule [2.1 - 4] du chapitre 2) :

$$\mathbf{N} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\hat{\mathbf{X}}$$

Explicitons les différents éléments de la matrice **N** en tenant compte de la nature particulière des colonnes de **Y** :

- la matrice $\frac{1}{n}\hat{\mathbf{X}}'\hat{\mathbf{X}}$ n'est autre que la matrice des covariances empiriques désignée précédemment par **T**.
- la matrice $\mathbf{D} = \mathbf{Y}'\mathbf{Y}$ est diagonale et son $k^{ème}$ élément diagonal vaut n_k, effectif de la $k^{ème}$ classe[1].
- la matrice à p lignes et q colonnes $\mathbf{H} = \hat{\mathbf{X}}'\mathbf{Y}$ a pour terme général :

$$h_{jk} = \sum_{i=1}^{n} \hat{x}_{ij} y_{ik} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j) y_{ik} = \sum_{i \in I_k} (x_{ij} - \overline{x}_j) = n_k (\overline{x}_{kj} - \overline{x}_j)$$

---

[1] En effet, on a la relation $\sum_{i=1}^{n} y_{ik} y_{ik'} = \delta_{kk'} n_k$ car l'individu i appartient soit à la classe k, soit à la classe k'; $\delta_{kk'}$=*1* si k=k' et vaut *0* sinon. Pour k=k', il y aura autant de termes non nuls dans la somme que d'individus dans la classe k.

En vertu de la relation [7.1 - 4], on peut écrire :

$$h_{jk} = \sqrt{n n_k} c_{jk}$$

soit :

$$\mathbf{H} = \hat{\mathbf{X}}'\mathbf{Y} = \sqrt{n}\,\mathbf{C}\,(\mathbf{Y}'\mathbf{Y})^{1/2}$$

Ces dernières remarques nous permettent d'écrire :

$$\hat{\mathbf{X}}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\hat{\mathbf{X}} = n\,\mathbf{C}\mathbf{C}' = n\,\mathbf{E}$$

puisque :

$$(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} = \frac{1}{n}\mathbf{T}^{-1}$$

la matrice **N** devient finalement $\mathbf{N} = \mathbf{T}^{-1}\mathbf{E}$ et le vecteur **a** cherché vérifie bien la relation [7.1 - 3] :

$$\mathbf{Ea} = \lambda\,\mathbf{Ta}$$

Nous pouvons également noter que l'on a, pour les deux types d'analyse, la même contrainte de normalisation : $\mathbf{a'Ta} = 1$

Il y a donc coïncidence entre variable canonique et fonction discriminante. L'analyse discriminante apparaît ainsi comme un cas particulier de l'analyse canonique (sans centrage préalable des variables indicatrices) lorsque l'un des deux ensembles est constitué de vecteurs booléens décrivant la partition de l'ensemble des individus.

7.2.3 Lien avec l'analyse des correspondances

Lorsque le sous-tableau **X** décrit lui aussi une partition en p classes, les résultats du paragraphe précédent montrent immédiatement que l'analyse des correspondances est un cas particulier de l'analyse factorielle discriminante.

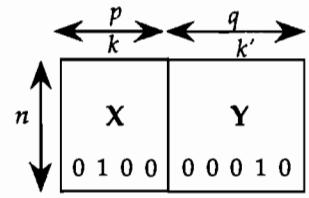


Figure 7.2 – 2. Tableau de données [X,Y]

Les deux sous-tableaux **X** d'ordre (n,p) et **Y** d'ordre (n,q) de la matrice des données **[X,Y]** sont formés de variables indicatrices et jouent maintenant des rôles analogues. Dans ce cas, les matrices **X'X** et **Y'Y** sont diagonales et ont pour $k^{ème}$ élément les effectifs de la classe k de chacune des partitions ; la matrice **X'Y** n'est autre que le tableau de contingence d'ordre (p,q) croisant les deux partitions P_X et P_Y.

Conformément aux conventions adoptées en analyse des correspondances, on notera (n est ici l'effectif global alors qu'il était désigné par k à la section 4.1) :

- $f_{k.}$, le $k^{ème}$ élément diagonal de la matrice $\frac{1}{n}\mathbf{X'X}$ $(=\mathbf{D}_p)$, $(k \leq p)$
- $f_{k'}$, le $k'^{ème}$ élément diagonal de la matrice $\frac{1}{n}\mathbf{Y'Y}$ $(=\mathbf{D}_q)$, $(k' \leq q)$
- $f_{kk'}$, l'élément générique de la matrice $\frac{1}{n}\mathbf{X'Y}$ $(=\mathbf{F})$, d'ordre (p,q)

Rappelons les formules établies au paragraphe 2.1.2 reliant les variables canoniques :

$$\mathbf{a} = \frac{1}{\lambda}(\mathbf{X'X})^{-1}\mathbf{X'Yb} \quad \text{et} \quad \mathbf{b} = \frac{1}{\lambda}(\mathbf{Y'Y})^{-1}\mathbf{Y'Xa}$$

Leurs composantes s'écrivent :

$$a_k = \frac{1}{\lambda}\sum_{k'=1}^{q}\frac{f_{kk'}}{f_{k.}}b_{k'} \quad \text{et} \quad b_{k'} = \frac{1}{\lambda}\sum_{k=1}^{p}\frac{f_{kk'}}{f_{.k'}}a_k$$

On reconnaît, sous cette forme, les relations barycentriques de l'analyse des correspondances [4.2 - 10] et [4.2 - 11] reliant les coordonnées des deux nuages sur un même axe factoriel. Cette identité suffit à établir qu'une analyse des correspondances est une analyse canonique particulière où les tableaux **X** et **Y** contiennent les variables indicatrices de deux partitions[1].

Les sous-espaces V_X et V_Y ont maintenant en commun la première bissectrice[2] de $\mathcal{R}^n$; leur plus petit angle est donc nul. Son cosinus ($=1$) est la valeur propre triviale déjà rencontrée en analyse des correspondances lorsque l'analyse est faite par rapport à l'origine et non par rapport au centre de gravité.

On a alors $\lambda = 1$, $a_i = 1$ et $b_j = 1$, pour tout i et tout j dans les relations écrites ci-dessus. Le fait de centrer le tableau **X** revient à projeter les points-colonnes sur le sous-espace orthogonal à la première bissectrice. Cette opération ne modifie donc pas les variables canoniques non triviales.

L'analyse des correspondances apparaît comme une *double analyse discriminante* car chacun des blocs dans **[X,Y]** décrit une partition et aucun d'entre eux n'est privilégié. Les fonctions linéaires discriminantes coïncident avec les facteurs de l'analyse des correspondances[3] du tableau de contingence d'ordre (p,q) croisant les deux partitions.

---

[1] La première racine canonique λ^2 est l'homologue de la première valeur propre, notée λ précédemment pour l'analyse des correspondances.

[2] La somme des colonnes de **X** et la somme des colonnes de **Y** constituent le vecteur dont toutes les composantes valent *1*.

[3] Cette présentation permet de montrer directement que les valeurs propres de l'analyse des correspondances, étant des *coefficients de corrélation canonique* (ou des pouvoirs discriminants) sont inférieures ou égales à *1*. On interprète ainsi les valeurs propres de l'analyse des correspondances en terme de *pouvoir discriminant* des facteurs vis-à-vis des partitions étudiées.

7.2.4 Une analyse avec une métrique particulière

L'analyse factorielle discriminante peut être considérée comme une analyse générale du nuage des q centres de gravité des classes k munis des masses n_k/n et avec la métrique $\mathbf{T}^{-1}$ ou la métrique $\mathbf{D}^{-1}$ dite de *Mahalanobis*.

Le nombre d'axes discriminants est égal à q - *1* dans le cas où $n > p > q$. Il suffit en effet de se reporter au paragraphe 7.2.1.c précédent où est intervenu pour la première fois le tableau **C** des moyennes centrées.

L'analyse générale de ce tableau **C** avec la métrique $\mathbf{T}^{-1}$, selon les résultats du paragraphe 1.3.1 (analyse générale avec une métrique quelconque : ici, $\mathbf{X} = \mathbf{C}$, $\mathbf{M} = \mathbf{T}^{-1}$ et $\mathbf{N} = \mathbf{I}$) conduit, pour trouver l'axe factoriel **u**, à la relation :

$$\mathbf{C'CT^{-1}u} = \lambda\mathbf{u}$$

Posant $\mathbf{T^{-1}u} = \mathbf{a}$, où **a** est le facteur (opérateur projection) sur l'axe factoriel **u** :

$$\mathbf{C'Ca} = \lambda\mathbf{Ta}$$

De la même façon, avec la métrique $\mathbf{D}^{-1}$, on obtient :

$$\mathbf{C'Ca} = \lambda\mathbf{Da}$$

Choisir la métrique $\mathbf{D}^{-1}$ pour analyser le nuage des points-moyens, c'est considérer comme équidistantes du centre j (par exemple) des zones équiprobables (au sens des ellipsoïdes de densité) d'équation :

$$(\mathbf{x} - \bar{\mathbf{x}}_j)'\mathbf{D}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_j) = \textit{constante}$$

Avec cette métrique, la distance représente une *vraisemblance d'appartenance*

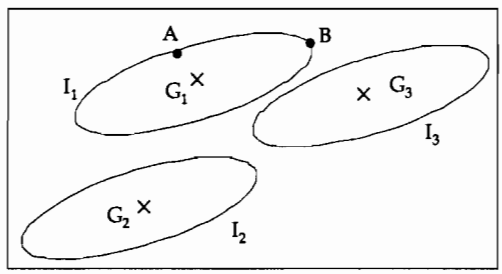


Figure 7.2 – 3. Illustration de la métrique $\mathbf{D}^{-1}$

Ainsi, sur la figure 7.2 - 3, où sont représentées trois classes ayant mêmes ellipsoïdes de densité, les points A et B sont équidistants (selon la métrique $\mathbf{D}^{-1}$) du centre de classe G_1. Avec la métrique euclidienne usuelle, B serait affecté plutôt à la classe *3* qu'à la classe *1*. On voit donc l'intérêt de cette métrique qui n'a cependant de sens que si les ellipsoïdes de densité sont les mêmes à l'intérieur de chaque classe. C'est précisément ce qui caractérise l'analyse discriminante linéaire, par opposition à l'analyse discriminante quadratique, qui autorise des densités de formes différentes, et donc des métriques différentes pour chaque classe (cf. annexe 7.8 de ce chapitre).

7.3 Règles de classement

Une fois trouvées les fonctions discriminantes qui séparent au mieux les individus répartis en q classes, on veut trouver la classe d'affectation d'un nouvel individu, pour lequel on connaît les valeurs des variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$. Une règle simple et géométrique d'affectation est de choisir la classe dont le centre de gravité est le plus proche du point-individu. La métrique généralement utilisée dans les applications les plus courantes est celle de *Mahalanobis globale* ($\mathbf{D}^{-1}$), ou *locale* ($\mathbf{D}_k^{-1}$, où $\mathbf{D}_k$ est la matrice des covariances internes au groupe I_k). Cette approche géométrique ne prend pas en compte les probabilités *a priori* des différentes classes, parfois très inégales dans certaines applications (prévision de défaillance, ou diagnostic d'un événement rare). Le modèle bayésien d'affectation permet d'enrichir ce point de vue.

7.3.1 Le modèle bayésien d'affectation

Lors de l'apprentissage, nous savons que l'individu i appartient au groupe I_k (appartenance codée par la valeur : $y_i = k$) et nous calculons une estimation de la probabilité $P(\mathbf{x}_i \mid I_k)$ (probabilité de $\mathbf{x}_i$ sachant que I_k est réalisé). Au moment de l'affectation d'un individu nouveau noté $\mathbf{x}$, on peut calculer les différents $P(\mathbf{x} \mid I_k)$ pour $k = 1, 2, \ldots, q$. Il paraît raisonnable d'affecter $\mathbf{x}$ à la classe I_k pour laquelle $P(\mathbf{x} \mid I_k)$ est maximale.

Cependant, ce ne sont pas les probabilités $P(\mathbf{x} \mid I_k)$ qu'il faudrait connaître mais les probabilités $P(I_k \mid \mathbf{x})$, c'est-à-dire la probabilité du groupe I_k sachant que $\mathbf{x}$ est réalisé. Le théorème de Bayes[1] permet de procéder à cette *inversion des probabilités*. Il exprime $P(I_k \mid \mathbf{x})$ en fonction de $P(\mathbf{x} \mid I_k)$, $P(I_k)$ et $P(\mathbf{x})$:

$$P(I_k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid I_k) P(I_k)}{P(\mathbf{x})}$$

$P(I_k)$ est la probabilité *a priori* du groupe k. $P(\mathbf{x})$ s'exprime en fonction de $P(\mathbf{x} \mid I_k)$ et de $P(I_k)$; d'où la formulation classique du théorème de Bayes :

$$P(I_k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid I_k) P(I_k)}{\sum_{k=1}^{q} P(\mathbf{x} \mid I_k) P(I_k)}$$

[1] Pour un exposé de l'approche bayésienne qui fournit un cadre conceptuel spécifique à l'estimation et à la décision statistique, voir Droesbeke *et al.* (2002), Robert (2006).

Le dénominateur est le même pour toutes les classes. La classe d'affectation de x sera celle pour laquelle le produit $P(\mathbf{x}|\mathrm{I}_k)\times P(\mathrm{I}_k)\times P(\mathrm{I}_k)$ est maximal. Si les probabilités *a priori* $P(\mathrm{I}_k)$ des classes sont égales pour toutes les valeurs de *k*, les classements selon $P(\mathrm{I}_k|\mathbf{x})$ et $P(\mathbf{x}|\mathrm{I}_k)$ sont identiques.

Pour tester l'efficacité des règles d'affectation, on mesure les erreurs de classement par des méthodes de rééchantillonnage, notamment la validation croisée ou le bootstrap (cf. chapitres 1.4.2 et 3.4.4). Comme dans le cas du modèle linéaire, le choix des variables explicatives est une opération délicate. L'étude de la stabilité des fonctions discriminantes est difficile. Les règles d'affectation ainsi que l'estimation des taux d'erreur de classement dépendent souvent de la taille de l'échantillon d'apprentissage.

7.3.2 Le modèle bayésien dans le cas normal

Notons $f_k(\mathbf{x})$ la densité de probabilité de x connaissant I_k dans le cas multinormal, μ_k et Σ_k désignant respectivement la moyenne et la matrice des covariances théoriques à l'intérieur du groupe I_k :

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}\left|\Sigma_k\right|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}$$

L'affectation se fera selon la règle :

$$\text{choisir } \hat{k} \text{ tel que } f_{\hat{k}}(\mathbf{x})P(\mathrm{I}_{\hat{k}}) = \max_{k\le q}\{f_k(\mathbf{x})P(\mathrm{I}_k)\}$$

ce qui est équivalent à trouver le minimum sur *k* de la fonction $sc_k(\mathbf{x})$ appelée *score discriminant* :

$$sc_k(\mathbf{x}) = (\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k) + Log\left|\Sigma_k\right| - 2LogP(\mathrm{I}_k) \qquad [7.3 - 1]$$

Dans le cas où les distributions dans chaque classe ont même matrice des covariances (cas illustré par la figure 7.2-3 précédente), la densité s'écrit :

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}\left|\Sigma\right|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma^{-1}(\mathbf{x}-\mu_k)\right\}$$

Il suffit alors de prendre pour score discriminant :

$$sc_k(\mathbf{x}) = (\mathbf{x}-\mu_k)'\Sigma^{-1}(\mathbf{x}-\mu_k) - 2LogP(\mathrm{I}_k) \qquad [7.3 - 2]$$

Si de plus les probabilités *a priori* $P(\mathrm{I}_k)$ sont égales, le score discriminant coïncide avec la distance de Mahalanobis :

$$sc_k(\mathbf{x}) = (\mathbf{x}-\mu_k)'\Sigma^{-1}(\mathbf{x}-\mu_k) \qquad [7.3 - 3]$$

et la règle bayésienne d'affectation devient la recherche du centre le plus proche selon cette distance.

Le score discriminant donné par la formule [7.3 - 1] correspond à *l'analyse discriminante quadratique.* Les cloisons interclasses données par l'équation $sc_k(\mathbf{x}) = sc_{k'}(\mathbf{x})$, $(k \neq k')$, sont en effet des hyperquadriques.

Les scores discriminants donnés par les formules [7.3 - 2] ou [7.3 - 3] correspondent à *l'analyse discriminante linéaire.* Dans l'équation $sc_k(\mathbf{x}) = sc_{k'}(\mathbf{x})$, $(k \neq k')$, les termes du second degré en $\mathbf{x}$ disparaissent et les cloisons interclasses sont dans ce cas des hyperplans. Ces hyperplans ont une équation de la forme :

$$\mathbf{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_k) = \textit{constante}$$

Notons que le calcul suppose connus les paramètres théoriques $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$. Ils suggèrent de substituer en pratique les estimations empiriques aux paramètres théoriques. Cette substitution est également encouragée par l'approche descriptive développée au début de cette section, dans laquelle les distances de Mahalanobis sont apparues de façon naturelle, en cherchant à maximiser le rapport variance externe sur variance interne, sans recours à l'hypothèse de normalité.

Les scores discriminants utilisés en pratique[1], lorsque l'hypothèse de normalité est plausible, sont donc ceux présentés ici avec utilisation des estimations empiriques des paramètres.

7.3.3 Autres règles d'affectation

Il existe d'autres méthodes de discrimination que celles apparentées à l'analyse factorielle discriminante ou au modèle multinormal. Elles impliquent d'autres règles d'affectations.

Citons, parmi les méthodes les plus utilisées[2] : les méthodes d'estimation non-paramétriques de la densité, connues également sous le nom de méthodes des *noyaux (de Rosenblatt ou de Parzen),* et les méthodes d'affectation (également non-paramétriques) utilisant les *m plus proches voisins.*

a – Estimation de la densité par noyaux

Une méthode simple de discrimination consisterait à diviser l'espace multidimensionnel de l'échantillon d'apprentissage en cellules de volumes comparables v_r puis de compter, à l'intérieur de chaque classe k, $(k \leq q)$, les n_{rk} observations contenues dans chaque cellule r.

---

[1] Il n'est cependant pas aisé de démontrer l'optimalité de cette démarche intuitive, sauf dans des contextes asymptotiques assez particuliers (cf. Anderson, 1958 ; Friedman, 1989).

[2] D'autres techniques de discrimination seront évoquées plus loin (méthodes neuronales, régression logistique).

La fréquence n_{rk}/n_k est une estimation de la probabilité qu'une observation de la catégorie k appartienne à la cellule v_r. La règle de Bayes permet alors d'affecter une observation supplémentaire $\mathbf{x}$ à une catégorie k, après avoir déterminé la cellule v_r qui la contient. Cette méthode est malheureusement impossible à mettre en oeuvre car le nombre de cellules devient vite prohibitif dans un espace à p dimensions et les échantillons n'ont pas une taille suffisante pour permettre une estimation de fréquence à l'intérieur de chaque cellule.

On peut, pour la classe k, entourer d'une cellule chaque point observé, de façon à décrire la densité dans l'espace $\mathcal{R}^p$. Si le point à affecter $\mathbf{x}$ tombe à l'intersection de trois cellules de la classe k par exemple et en dehors des cellules relatives aux autres classes, cela signifiera qu'il est dans une zone de forte densité pour la classe k et donc qu'il a plus de chance d'appartenir à cette classe qu'aux autres. Cette idée, présentée ici de façon intuitive, est celle des noyaux de Rosenblatt (1956).

Au lieu d'entourer les points de cellules de volumes fixes, on peut les entourer d'une sorte de halo, une zone de densité qui décroît lorsqu'on s'éloigne du point, de façon à procéder à un lissage de cette densité dans l'espace multidimensionnel. C'est la méthode d'estimation directe de la densité par noyaux à laquelle on attache le nom de Parzen (1962).

La méthode des noyaux consiste à estimer la densité de probabilité à l'intérieur de la classe k dans l'espace $\mathcal{R}^p$ par une formule du type :

$$f_k(x) = \frac{1}{h^p n_k} \sum_{i=1}^{n_k} K\left(\frac{x - x_i}{h}\right) \qquad [7.3 - 4]$$

La fonction $K(\mathbf{z})$ doit vérifier les relations $K(\mathbf{z}) \geq 0$, et $\int K(\mathbf{z})d\mathbf{z} = 1$. Elle pourra être choisie parmi les densités de probabilité usuelles. On note que l'on a bien dans ces conditions :

$$\int f_k(\mathbf{x})d\mathbf{x} = 1$$

On utilise souvent la densité de la loi normale sphérique :

$$K(\mathbf{z}) = (2\pi)^{-p/2} exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\}$$

Le paramètre h qui intervient dans la formule [7.3 - 4] est la *dimension de la fenêtre*. Dans le cas des noyaux normaux sphériques, il correspond à l'écart-type de la densité locale autour de chaque point. Si h est petit, le lissage risque d'être mauvais; si h est trop grand, il risque d'être excessif. Le choix de la dimension de la fenêtre est une des difficultés de ces méthodes d'estimation directe de densité[1].

---

[1] Cf. Tomassone et al. (1988), Silverman (1986), Delecroix (1983), Hand (1982). Le paramètre h, supposé unique dans la formule [7.3 - 4] peut avoir, dans des modèles plus généraux, une valeur différente pour chacune des p variables et également pour chacune des q classes.

b – Règle des *m* plus proches voisins (Fix et Hodges, 1951)

Cette technique, utilisée surtout en reconnaissance des formes, résout d'une autre façon le problème des cellules à densité trop faible : on étend le voisinage autour du point **x** jusqu'à ce qu'il contienne *m* points de l'échantillon d'apprentissage. On affecte **x** à la classe la plus représentée dans ce voisinage.

Cette méthode est particulièrement simple à mettre en œuvre, surtout dans un processus d'apprentissage progressif, car il n'y a pas de fonctions complexes à recalculer pour prendre en compte les nouveaux individus qui enrichissent l'échantillon d'apprentissage. Elle nécessite cependant des effectifs importants, des calculs d'affectation coûteux (pour les exigences de la reconnaissance des formes, où le classement s'effectue souvent en temps réel) si les paramètres *m* ou *p* sont grands[1].

7.3.4 Qualité des règles de classement

Il existe un cadre inférentiel paramétrique, apparenté à *l'analyse multidimensionnelle de la variance*, qui permet de tester l'hétérogénéité des classes (test de l'égalité des moyennes μ_k, test de l'égalité des matrices de covariances internes $\mathbf{D}_k$). Ces tests (mentionnés dans la plupart des manuels de référence cités ici au chapitre 2) dont la robustesse est difficile à établir, sont moins utilisés depuis l'avènement des méthodes non-paramétriques de rééchantillonnage qui seront évoquées au paragraphe 1.4.2. On esquissera ici, pour les besoins des développements qui suivront, la méthode dite de *validation croisée*.

Dans cette méthode, la mesure de la qualité d'une discrimination se fait à partir des pourcentages de bien classés (ou de mal classés) dans chaque classe, et du pourcentage global de bien classés. Cette mesure peut également, dans certaines applications, faire intervenir des coûts de mauvais classement.

On peut calculer un pourcentage de bien classés sur l'échantillon d'apprentissage, ce qui donnera une idée optimiste de la qualité de la discrimination. Ce pourcentage de bien classés augmente avec le nombre de paramètres du modèle, et peut être excellent si le nombre de paramètres est considérable, sans pour cela assurer que le modèle permet de réaliser une prévision correcte. Le pourcentage de mal classés dans ces conditions est appelé le *taux d'erreur apparent* ou encore le *taux d'erreur par resubstitution*.

---

[1] Il existe des ponts théoriques entre la méthode des *m* plus proches voisins et l'estimation directe de densité dans le cas de variables binaires (cf. Fix et Hodges, 1951; Aitchison et Aitken, 1976). Il est également possible de travailler avec des noyaux adaptatifs, en faisant varier la dimension de la fenêtre *h* ou en tenant compte des distances des *m* plus proches voisins. Pour une discussion de ces diverses variantes, voir McLachlan (1992). Sur les divers algorithmes de *m* plus proches voisins utilisés en reconnaissance des formes, cf. Dubuisson (1990). Sur les problèmes posés par des probabilités *a priori* inégales, cf. Chateau (1994).

La méthode des échantillons-tests recommande d'effectuer la discrimination sur une partie seulement de l'échantillon d'apprentissage (disons *80%*) et de tester les règles de discrimination sur les *20%* non utilisés. On peut faire remonter cette pratique à Highleyman (1962), mais elle a probablement dû être utilisée antérieurement, tant son principe relève du bon sens. Elle a été prônée notamment par Romeder (1973).

On peut améliorer le calcul du taux d'erreur en divisant l'échantillon d'apprentissage en m parties égales, en calculant la règle sur un échantillon partiel formé de $m-1$ parties, et le taux d'erreur sur la partie restante, ce qui peut être fait de m façons différentes. Ceci permet donc de calculer un taux d'erreur moyen sur un échantillon aussi important que l'échantillon d'apprentissage.

Plus m est proche de n, plus on se rapproche de la situation réelle de classement. La validation croisée correspond au cas $m = n$, autrement dit, au cas pour lequel on effectue n discriminations en excluant à chaque fois une observation. Attribuée à Lachenbruch et Mickey, 1968, cette méthode *(cross-validation)* aurait été utilisée dès 1964 par des chercheurs russes, selon Toussaint (1974). Ses propriétés ont été étudiées par Stone (1974) et Geisser (1975). Une revue est faite par Hand(1986). Cette méthode est évidemment coûteuse en calcul mais on peut parfois mettre en œuvre des algorithmes évitant des recalculs complets des fonctions discriminantes. La minimisation du taux d'erreur par validation croisée peut être utilisée comme critère pour calculer les paramètres de certains modèles de discrimination.

7.4 Régularisation en analyse discriminante

Comme la régression multiple (dont elle est un cas particulier dans le cas où la variable nominale à prédire n'a que deux catégories, cf. § 7.2.1), l'analyse factorielle discriminante nécessite l'inversion d'une matrice des covariances des prédicteurs (la matrice totale **T** ou la matrice intraclasse **D**).

Dans le cas de l'analyse discriminante quadratique, le calcul des *distances de Mahalanobis locales* demande d'inverser les matrices de covariances internes à chaque classe $\mathbf{D}_k$ (dont **D** est une moyenne pondérée).

Ces matrices **D** ou **T**, et surtout les matrices $\mathbf{D}_k$, calculées sur un effectif n_k plus petit que n, peuvent être mal conditionnées ou même singulières.

C'est systématiquement le cas en analyse discriminante qualitative lorsque les prédicteurs sont des variables nominales codées sous forme disjonctive comme en analyse des correspondances multiples ou en analyse de la variance (cf. § 7.5). On présentera brièvement ci-dessous une méthode de régularisation proposée par Friedman et la méthode de régularisation par axes principaux déjà proposée pour la régression (§ 2.2.5). Cette méthode a l'avantage de fournir

une description préalable de l'espace des prédicteurs et des possibilités ultérieures de filtrage et de sélection de l'information.

7.4.1 Analyse régularisée

Dans cette méthode de régularisation proposée par Friedman (1989), une nouvelle estimation $\mathbf{D}_k(\lambda,\gamma)$ est calculée pour chaque matrice des covariances locales $\mathbf{D}_k$, qui devient une moyenne pondérée des matrices des covariances globales et locales (rôle du poids λ) et de la matrice unité (rôle du poids γ) :

$$\mathbf{D}_k(\lambda,\gamma) = (1-\gamma)\mathbf{D}_k(\lambda) + \frac{\gamma}{p} tr[\mathbf{D}_k(\lambda)]\mathbf{I}$$

avec :

$$\mathbf{D}_k(\lambda) = \frac{(1-\lambda)\mathbf{D}_k + \lambda\mathbf{D}}{(1-\lambda)n_k + \lambda n}$$

Le scalaire $tr[\mathbf{D}_k(\lambda)]$ est la trace de la matrice $\mathbf{D}_k(\lambda)$.

La détermination des paramètres λ et γ se fait en optimisant les pourcentages de bien classés obtenus par validation croisée. Ces techniques donnent des résultats intéressants dans le cas de tableaux de données petits ou moyens, lorsque le problème initial est *mal posé* ($n \leq p$) ou *pauvrement posé* ($n > p$, mais encore comparable à p)[1]. Dans le cas de grandes matrices clairsemées cependant, l'échelle du phénomène crée de nouveaux problèmes. Il est alors nécessaire de comprendre ce qui se passe dans les espaces de dimension élevée. Est-il vraiment nécessaire de garder tous les axes principaux ? Est-il possible de filtrer l'information de base caractérisée parfois par un haut niveau de bruit ? L'analyse par axes principaux répond à ces préoccupations.

7.4.2 Analyse régularisée par axes principaux

Du point de vue numérique, la diagonalisation est une opération plus sûre que l'inversion des matrices. La théorie de la perturbation nous apprend que la stabilité des vecteurs propres est une fonction croissante des différences entre valeurs propres consécutives[2]. Dans ce contexte, s'il est nécessaire d'éliminer les dimensions correspondant à des valeurs propres nulles, il peut être aussi avantageux d'éliminer les dimensions correspondant aux petites valeurs propres, qui sont très sensibles aux perturbations du tableau de données[3].

---

[1] Voir aussi Callant (1991) pour une technique d'estimation des paramètres λ et γ.

[2] Cf. la section 1.4 du chapitre 1. Cf. aussi, par exemple, Wilkinson (1965), Kato (1966) et les travaux de Escofier et Le Roux (1972).

[3] Cf. Les travaux de Wold (1976). Benzécri (1977 a) recommande que les analyses discriminantes soient réalisées sur les axes d'une analyse factorielle préalable.

a – Axes principaux de l'échantillon total

La technique de réduction qui sera utilisée durant la première étape dépend de la nature et des propriétés statistiques des données de base[1]. Une simple décomposition aux valeurs singulières suffit pour une régularisation numérique, si l'on ne désire pas de description de l'espace des prédicteurs.

Les nouvelles coordonnées de l'individu i sur l'axe principal r issu de l'analyse de l'échantillon total sont désignées par z_{ri} ,

$$z_{ri} = \mathbf{u}_r'(\mathbf{x}_i - \bar{\mathbf{x}})$$

où ici $\mathbf{u}_r$ est le $r^{ème}$ vecteur propre normalisé de **T** matrice des covariances totales correspondant à la valeur propre α_r ; $\mathbf{u}_r$ est aussi la $r^{ème}$ colonne de la matrice **U** d'ordre (p, r_{max}) (où r_{max} est le nombre de valeurs propres retenues).

La distance euclidienne usuelle dans $\mathcal{R}^p$ de tout point i au point-moyen G_k de la classe k (le point i peut ne pas appartenir à la classe k ni à l'échantillon d'apprentissage) peut s'écrire :

$$d^2(i,G_k) = \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \qquad [7.4 - 1]$$

si $r_{max} = p'$ (p' désignant le rang de la matrice de données **X**), cette même distance s'écrit, pour la nouvelle base :

$$d^2(i,G_k) = \sum_{r=1}^{r_{max}} (z_{ir} - \bar{z}_{kr})^2 \qquad [7.4 - 2]$$

avec $\bar{z}_{kr} = u_r'(\bar{x}_k - \bar{x})$.

La distance de tout point i au centre G_k de la classe k dans la métrique $\mathbf{T}^{-1}$ (intervenant en analyse discriminante linéaire, cf. 7.2.4) est telle que :

$$D^2(i,G_k) = \sum_{r=1}^{r_{max}} \frac{(z_{ir} - \bar{z}_{kr})^2}{\alpha_r} \qquad [7.4 - 3]$$

On a toujours $r_{max} \leq p'$. La distance $D^2(i,G_k)$ est dite *régularisée* si $r_{max} < p'$ ou si $r_{max} = p'$ avec $p' < Min\ (n,p)$.

b – Axes principaux de l'échantillon projeté

Si l'on substitue à la matrice de données **X**, de terme général x_{ij}, la matrice $\hat{\mathbf{X}}$ de terme général $\hat{x}_{ij} = x_{ij} - \bar{x}_{kj}$ où k est l'indice de la classe I_k à laquelle appartient l'observation i et où $\bar{x}_{kj}$ désigne la moyenne de la variable j dans cette classe, on est conduit à diagonaliser la matrice **D** (au lieu de **T**). Les valeurs propres de

---

[1] Analyse en composantes principales dans le cas où les prédicteurs sont des variables continues, situation retenue au cours des développements qui précèdent ; mais cette réduction pourra aussi être une analyse des correspondances dans le cas de fréquences ou des correspondances multiples dans le cas de variables nominales.

$\mathbf{D}$ sont notées $\hat{\alpha}_r$ et les coordonnées des observations sur les nouveaux axes principaux $\hat{\mathbf{u}}_r$ sont notées $\hat{z}_{ir}$ [1].

La distance de tout point i au centre G_k de la classe k dans la métrique $\mathbf{D}^{-1}$ (*distance de Mahalanobis globale*) est telle que :

$$\hat{D}^2(i,G_k) = \sum_{r=1}^{r_{max}} \frac{(\hat{z}_{ir} - \bar{\hat{z}}_{kr})^2}{\hat{\alpha}_r} \qquad [7.4 - 4]$$

$\hat{D}^2(i,G_k)$ est *régularisée* si $r_{max} = p''$ (où p'' désigne le rang de la matrice transformée $\hat{\mathbf{X}}$) quand $p''<Min\ (n,p)$ ou si $r_{max} < p''$.

c – Axes principaux dans les groupes

Pour chaque classe I_k, les matrices de covariances d'ordre (r_{max}, r_{max}) sont calculées séparément. On les exprimera ici à partir des coordonnées de l'analyse globale précédente.

Les nouvelles coordonnées de l'individu i sur l'axe principal s de l'analyse réalisée à l'intérieur de la classe I_k (il s'agit donc dans ce cas d'une simple analyse en composantes principales non normée) sont[2] :

$$w_{ski} = \mathbf{v}'_{sk}(\mathbf{z}_i - \bar{\mathbf{z}}_k)$$

où $\mathbf{v}_{sk}$ est le $s^{ème}$ vecteur propre normalisé de $\mathbf{U'D_kU}$ correspondant à la valeur propre β_{sk} (β_{sk} est également valeur propre de $\mathbf{D}_k$).

Avec ces coordonnées, on peut évidemment retrouver les distances usuelles, calculées cette fois dans chacune des q nouvelles bases (pour tout point i et tout point-moyen G_k), lorsque le nombre $s_{max}(k)$ d'axes retenus à ce stade pour la classe k, vérifie : $s_{max}(k) = r_{max}$.

$$d^2(i,G_k) = \sum_{s=1}^{s_{max}(k)} (w_{ski} - \bar{w}_{ks})^2 \qquad [7.4 - 5]$$

avec : $$\bar{w}_{ks} = \mathbf{v}'_{ks}(\bar{\mathbf{z}}_k - \bar{\mathbf{z}})$$

La *distance de Mahalanobis locale* (intervenant en *analyse discriminante quadratique*) peut s'écrire :

$$\mathcal{D}^2(i,G_k) = \sum_{s=1}^{s_{max}(k)} \frac{(w_{ski} - \bar{w}_{ks})^2}{\beta_{sk}} \qquad [7.4 - 6]$$

---

[1] Comme l'opération de centrage global, cette opération correspond à une projection P. Si $\mathbf{Y}$ désigne le tableau disjonctif complet d'ordre (n,q) décrivant la partition à prédire, l'opérateur projection s'écrit : $\mathbf{P} = \mathbf{I} - \mathbf{Y}(\mathbf{Y'Y})^{-1}\mathbf{Y'}$. On peut parler dans ces conditions d'*analyse interne* ou *conditionnelle* : comme en analyse de la variance, on a *éliminé* la dispersion due aux classes en supposant que celles-ci avaient un effet additif.

[2] Cette formule de projection sur l'axe s est évidemment valable pour des points n'appartenant pas à la catégorie k (*points supplémentaires ou illustratifs*).

Une telle distance peut être "régularisée" à deux niveaux :

- une première fois si $r_{max} < p'$ (p' désigne le rang du tableau de donnée) ;
- de nouveau si $s_{max}(k) < r_{max}$.

On a noté que, si $s_{max}(k) = r_{max} = p$, les distances données par les formules [74 - 1], [7.4 - 2] et par les q formules [7.4 - 5] (il y a q bases orthonormées différentes donc q formules différentes) sont toutes égales.

d – Exemple numérique d'application

L'exemple qui suit concerne les effets de la dimension des sous-espaces sur les pourcentages de bien-classés, à la fois dans les échantillons d'apprentissage et dans les échantillons-tests.

Le jeu de données utilisé est un tableau binaire clairsemé de dimensions (*634, 83*) contenant *4039* cases non-nulles[1].

L'ensemble des *634* lignes (répondants) peut être réparti en $q = 3$ classes d'âge. Le problème est de savoir dans quelle mesure ces classes d'âge peuvent être prédites à partir des réponses. Notre critère d'évaluation de la discrimination est le pourcentage de succès (bien classés), qui sera calculé systématiquement à la fois pour l'échantillon d'apprentissage et pour un échantillon-test qui comprend le tiers (*211* individus) de l'échantillon global.

La première étape est un changement d'axes par analyse des correspondances. La séquence des valeurs propres, visible sur la figure 7.4-1, est assez typique des tableaux clairsemés : la décroissance des valeurs propres est très lente, presque linéaire après l'axe *15*

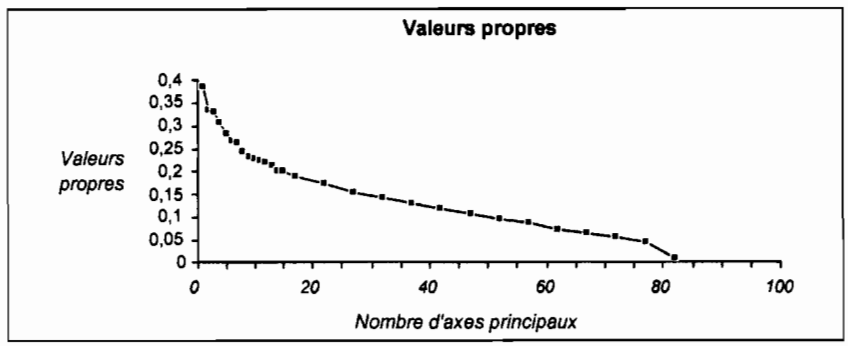


Figure 7.4 – 1. Séquence des valeurs propres de la première analyse

. Les *15* premières valeurs propres correspondent à *37%* de la trace. Chacun des axes restant correspond approximativement à *1%* de la trace. La figure7.4 - 2

[1] Il s'agit pour cet exemple de *4039* occurrences de $p = 83$ mots utilisés dans $n = 634$ réponses à une question ouverte dans une enquête (cf. Lebart, 1992).

montre les trajectoires des pourcentages de succès obtenus pour chacune des trois distances précédentes : *Distance euclidienne usuelle* (formule [3.3 - 12]), *distance de Mahalanobis globale* (formule [7.4 - 4]), *distance de Mahalanobis locale* (formule [7.4 - 6]).

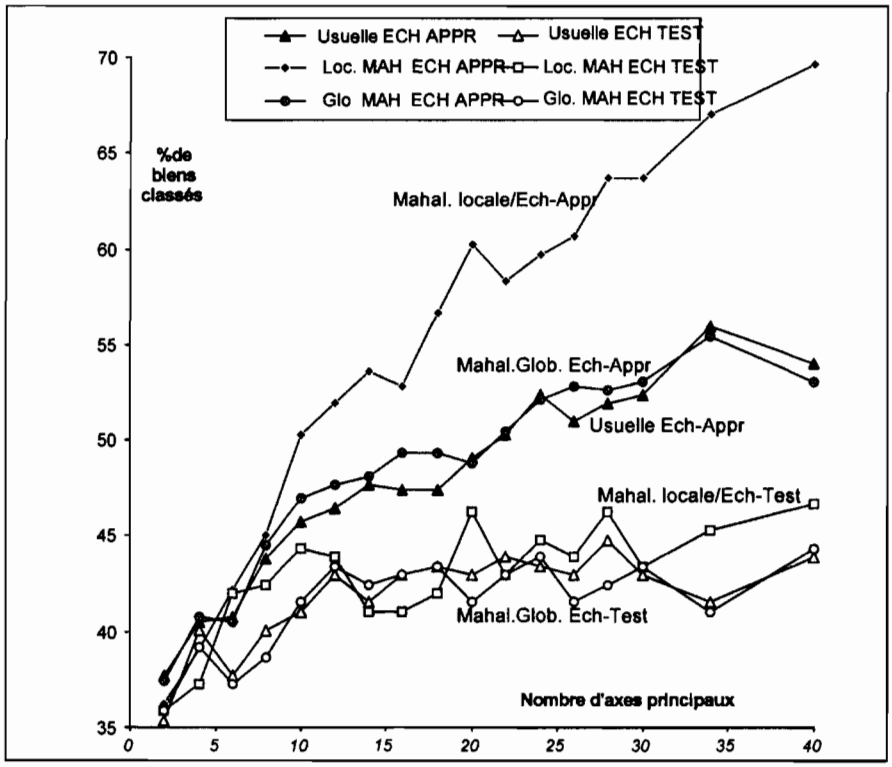


Figure 7.4 – 2. Trajectoires des pourcentages de bien classés en fonction du nombre d'axes principaux (axe des abscisses) selon trois distances et selon le type d'échantillon (test ou apprentissage)

On note que les taux correspondant aux échantillons d'apprentissage augmentent continûment avec le nombre d'axes alors que les taux correspondant aux échantillons-tests sont pratiquement stabilisés dès l'axe *15* (avec les notations ci-dessus, on peut choisir sans inconvénient $s_{max}(k) = r_{max} = 15$ alors que $p = 83$ et $p' = 82$).

Parmi les trajectoires des échantillons d'apprentissage, le pourcentage de bien classés correspondant à la *distance de Mahalanobis locale* croît fortement et atteint un niveau de *70%* de succès pour *40* axes. Une telle distance dépendant d'un nombre de paramètres plus important que les deux autres, s'adapte

souplement aux données d'apprentissage[1], sans fournir d'amélioration notable sur les échantillons-tests.

Pour cet exemple, la *distance de Mahalanobis globale* a des performances très voisines de la distance euclidienne usuelle. Les performances sont légèrement supérieures pour l'échantillon d'apprentissage.

Cet exemple met bien en évidence la puissance du filtrage réalisé par l'analyse factorielle préalable. La plupart des traits structuraux susceptibles de donner lieu à une prévision figurent dans l'espace à *15* dimensions des premiers axes.

e – Analyse discriminante sur variables qualitatives

L'analyse factorielle discriminante que nous venons de présenter s'applique à un ensemble de n individus répartis en q classes définies *a priori* par la variable nominale **y** et décrits par p variables $(x_1, x_2, \ldots, x_p)$ continues. Lorsque les p variables explicatives sont nominales, le calcul des fonctions linéaires discriminantes ne peut plus être appliqué, en raison des singularités de la matrice **X**, mais la régularisation par axes principaux permettra de lever cette difficulté.

Comme pour tout traitement de variables nominales, on procède au codage disjonctif complet des p variables explicatives. L'analyse factorielle discriminante qualitative consiste alors en une analyse factorielle discriminante classique sur les indicatrices des variables explicatives.

La matrice des variables explicatives $\mathbf{X}=\left[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_p\right]$ n'est pas inversible puisqu'il existe p relations linéaires entre les colonnes du tableau disjonctif complet. On peut alors, comme pour l'analyse de la variance, supprimer une modalité de chaque variable nominale ce qui ne modifie pas le sous-espace des variables explicatives V_X. Ceci ne suffit pas à assurer que la matrice réduite est bien conditionnée.

La régularisation par axes principaux revient dans ce cas à réaliser une analyse discriminante classique sur les facteurs de l'analyse des correspondances multiples[2].

On procède alors en effectuant :

- une analyse des correspondances sur le tableau disjonctif complet ; les p variables nominales sont donc remplacées par h variables continues qui sont les h facteurs de l'analyse des correspondances multiples.
- une analyse factorielle discriminante sur les h variables continues dont les valeurs sont les coordonnées sur les axes factoriels de l'analyse des correspondances multiples.

---

[1] Plus le nombre de paramètres augmente, plus l'apprentissage se rapproche de ce que l'on appelle en intelligence artificielle *l'apprentissage par coeur*, c'est-à-dire une adaptation trompeuse à une situation donnée, sans possibilité de généralisation.

[2] Enchaînement connu en particulier sous le nom de méthode DISQUAL (Saporta, 1977).

Compte tenu du nombre généralement important de facteurs de l'analyse des correspondances multiples, on retiendra les facteurs les plus discriminants et qui ne figurent pas toujours parmi les premiers[1].

f – Analyse discriminante barycentrique

L'analyse discriminante barycentrique revient simplement à faire l'analyse des correspondances du tableau croisant la variable à expliquer **y** avec les variables explicatives $(x_1, x_2, \ldots, x_p)$ (empilement de tables de contingences) : les lignes sont constituées par les modalités de **y** et les colonnes par la juxtaposition des modalités de $(x_1, x_2, \ldots, x_p)$.

Il s'agit en fait d'une bande du tableau de Burt (cf. section 5.3.3) qui permet de décrire les liaisons existant entre la variable à expliquer et l'ensemble des variables explicatives (cf. Saporta, 1975 a ; Leclerc, 1976).

En plaçant en éléments supplémentaires de nouveaux individus caractérisés par les variables explicatives, on réalise une réaffectation similaire à celle l'analyse discriminante (cf. Nakache *et al.* ,1977).

Dans le cas où les variables explicatives sont indépendantes deux à deux, l'analyse discriminante barycentrique est équivalente à l'analyse factorielle discriminante qualitative (puisque l'analyse d'une bande du tableau de Burt est alors équivalente à l'analyse du tableau complet). Dans le cas général, elle est, en théorie, moins performante puisque, comme nous l'avons vu dans § 5.3.3.b, elle ne tient pas compte des liaisons entre les variables explicatives. Elle est cependant largement utilisée en raison de sa simplicité et sa robustesse (cf. Carlier, in : Celeux et Nakache, 1994).

g – Note sur le "scoring"

Fréquemment utilisée par les organismes bancaires cherchant à prévoir la défaillance éventuelle d'un client (individu ou entreprise), la méthode dite de "scoring" permet une mise en forme simple des résultats d'une analyse discriminante généralement à deux groupes. Elle n'est pas à proprement parler une méthode de discrimination sur variables nominales ; mais elle utilise les résultats d'analyses discriminantes sur variables nominales ou continues pour construire une *fonction de score* [Cf. dans le cas d'analyses appliquées à la détection de défaillances d'entreprises (à partir de sélection de variables continues) : Bardos (1989, 2001)]. On dispose ainsi d'un instrument de décision accessible pour affecter un individu dans un groupe. Dans le cas de deux groupes, on obtient une seule fonction discriminante : la combinaison linéaire des variables qui

---

[1] Que ce soit pour l'analyse factorielle discriminante qualitative et, nous allons le voir, pour l'analyse discriminante barycentrique, il est conseillé de procéder au préalable à une *première sélection des variables nominales explicatives* en croisant par exemple chacune d'entre elles avec la partition à expliquer **y**, en calculant les χ^2 correspondants, et gardant celles qui correspondent aux χ^2 les plus significatifs.

sépare au mieux les deux groupes d'individus. Un individu est affecté à l'un des groupes si la fonction prend pour lui une valeur supérieure à un certain seuil.

Cette fonction discriminante est ensuite transformée en un système équivalent de coefficients attribués aux modalités des variables nominales ou aux éventuelles variables continues (en général après une sélection sévère). Cette transformation fournit la fonction score dont les coefficients constituent des notes attachées aux modalités ou aux variables. Pour chaque individu, on calcule le score c'est-à-dire la somme des notes associées aux prédicteurs. On affectera alors cet individu à un groupe si son score est supérieur à un certain seuil. L'introduction d'une tolérance d'erreur de classement permet en fait de définir trois zones de décisions sur la fonction score : la zone des scores élevés, celle des scores faibles et une zone d'indécision pour laquelle un individu n'est pas automatiquement classé.

7.5 Régression logistique

La régression logistique, comme l'analyse linéaire discriminante, cherche à décrire la liaison entre une variable nominale **y** (variable à expliquer) et un ensemble de p variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$. On veut également connaître l'effet d'une variable sur la variable à expliquer en tenant compte des liaisons qu'elle entretient avec les autres variables du modèle. Le modèle utilisé dans une régression logistique est étroitement lié au modèle log-linéaire, bien que la problématique soit différente. Le plus souvent la variable à expliquer est dichotomique et les variables explicatives sont continues. Les n individus caractérisés par l'ensemble des p variables sont partitionnés en deux groupes définis par les modalités de la variable **y**.

Le modèle logistique a été proposé originellement par Cornfield (1962). Etudié notamment par Cox (1972), il a été situé dans le cadre du modèle linéaire généralisé (§ 2.2.8 du chapitre 2) par Nelder et Wedderburn (1972). Une revue de ses applications en analyse discriminante est faite par Anderson (1982). Cf. également Hosmer et Lemeshow (1989), Devaud (1985). L'ouvrage collectif édité par Celeux et Nakache (1994) et le manuel de Nakache et Confais (2003) présentent les contributions du modèle logistique à la discrimination.

7.5.1 Le modèle logistique

Pour reprendre l'exemple du paragraphe 5.6.1 du chapitre 5, on désire étudier par exemple l'influence de la dose de radiation reçue et de l'âge des individus au moment des accidents sur le risque de décès par leucémie.

On suppose que la probabilité qu'un individu appartienne au groupe I_1 ($y = 1$) dépend des valeurs des variables ($x_1, x_2, \ldots, x_p$) observées sur cet individu.

On note $\mathbf{x}$ le vecteur dont les p composantes sont les valeurs des variables explicatives. Le modèle logistique se propose de fournir une estimation de cette probabilité notée $\pi(\mathbf{x})$:

$$\pi(\mathbf{x}) = P(I_1|\mathbf{x}) = P(y=1|\mathbf{x}).$$

Le théorème de Bayes (§ 7.3.1) nous permet d'écrire dans le cas de deux groupes I_1 et I_2 :

$$P(I_1|\mathbf{x}) = \frac{P(\mathbf{x}|I_1)P(I_1)}{P(\mathbf{x}|I_1)P(I_1)+P(\mathbf{x}|I_2)P(I_2)}$$

qui s'écrit encore :

$$P(I_1|\mathbf{x}) = \frac{\dfrac{P(\mathbf{x}|I_1)P(I_1)}{P(\mathbf{x}|I_2)P(I_2)}}{1+\dfrac{P(\mathbf{x}|I_1)P(I_1)}{P(\mathbf{x}|I_2)P(I_2)}} \qquad [7.5 - 1]$$

Cette formule ne fait intervenir que les quotients des deux probabilités conditionnelles de l'observation $\mathbf{x}$.

Dans le cas multinormal avec matrices des covariances $\boldsymbol{\Sigma}$ égales dans les deux groupes, chacune des deux probabilités conditionnelles s'écrit, pour $k = 1, 2$:

$$P(\mathbf{x}|I_k) = (2\pi)^{-p/2}\left|\boldsymbol{\Sigma}\right|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

Le quotient des probabilités pondérées fait disparaître les termes du second degré en $\mathbf{x}$ et s'écrit comme l'exponentielle d'une forme linéaire en $\mathbf{x}$ avec terme constant (fonction affine de $\mathbf{x}$) :

$$\frac{P(\mathbf{x}|I_1)P(I_1)}{P(\mathbf{x}|I_2)P(I_2)} = exp\{\boldsymbol{\beta}'\mathbf{x}+b\}$$

Pour alléger les notations, le vecteur $\mathbf{x}$ désignera désormais un vecteur à $p+1$ composantes (avec $x_0 = 1$ et les autres composantes égales à celles de l'ancien $\mathbf{x}$) et le nouveau vecteur de coefficients sera désigné par $\boldsymbol{\alpha}$, de sorte que $\boldsymbol{\beta}'\mathbf{x}+b$ s'écrit maintenant $\boldsymbol{\alpha}'\mathbf{x}$.

Ceci permet de réécrire la formule [7.5 - 1] et conduit à l'expression du *modèle logistique* :

$$\pi(\mathbf{x}) = \frac{exp\{\boldsymbol{\alpha}'\mathbf{x}\}}{1+exp\{\boldsymbol{\alpha}'\mathbf{x}\}} = \frac{exp\left\{\sum_{j=0}^{p}\alpha_j x_j\right\}}{1+exp\left\{\sum_{j=0}^{p}\alpha_j x_j\right\}}, \qquad [7.5 - 2]$$

où les α_j, composantes du vecteur $\boldsymbol{\alpha}$, sont les coefficients inconnus du modèle. Il s'agit d'un modèle qui ne fait pas intervenir de termes d'interaction entre les variables explicatives.

On peut écrire [7.5 - 2] sous la forme :

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = exp\{\boldsymbol{\alpha}'\mathbf{x}\} \qquad [7.5 - 3]$$

ou encore :

$$log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \boldsymbol{\alpha}'\mathbf{x} = \sum_{j=0}^{p} \alpha_j x_j \qquad [7.5 - 4]$$

La fonction :

$$F(\pi(\mathbf{x})) = log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$$

est appelée fonction *Logit*.

Remarques

1) Les modalités de la variable nominale seront codées *0* ou *1*. Comme pour l'analyse de la variance, on élimine une des modalités de chaque variable nominale. Le coefficient associé est égal à *0* et cette modalité est appelée traditionnellement "situation de référence" : on mesure en fait les différences avec la ou les autres modalités de la même variable.

2) Le modèle logistique, ou de régression logistique, ou de discrimination logistique, s'applique à une famille de distributions de x plus générale que la loi multinormale avec matrices de covariances égales qui nous a servi à l'introduire. Il suffit, on l'a vu, que le quotient des probabilités conditionnelles s'exprime comme l'exponentielle d'une fonction affine de x. Ceci est le cas de la plupart des distributions de la famille exponentielle (cf. § 2.2.8) dans certaines conditions (Anderson, 1982).

7.5.2 Estimation et tests des coefficients

a_ Procédure d'estimation

Pour estimer les coefficients α_j du modèle, on utilise le plus souvent la méthode du maximum de vraisemblance.

Les n observations $(y_i, \mathbf{x}_i)$ [où $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$] sont indépendantes et les y_i sont des variables de Bernoulli.

La vraisemblance $\mathfrak{L}(\boldsymbol{\alpha},\mathbf{y}_i)$ pour une observation s'écrit :

$$\mathfrak{L}(\boldsymbol{\alpha},y_i) = \pi(\mathbf{x}_i)^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}$$

et pour l'ensemble des observations, on a :

$$\mathcal{L}(\alpha, \mathbf{y}) = \prod_{i=1}^{n} \mathcal{L}(\alpha, y_i) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i}$$

La procédure d'estimation revient à rechercher la valeur $\hat{\alpha}$ de α qui maximise le logarithme de la vraisemblance :

$$log[\mathcal{L}(\alpha, \mathbf{y})] = \sum_i \left[y_i \, log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} + log[1 - \pi(\mathbf{x}_i)] \right]$$

soit encore en utilisant les formules [7.5 - 3] et [7.5 - 4]:

$$log[\mathcal{L}(\alpha, \mathbf{y})] = \sum_i y_i \, \alpha' \mathbf{x}_i - \sum_i log[1 + exp(\alpha' \mathbf{x}_i)]$$

Pour apprécier l'éventuelle *non-influence* d'une variable ou d'une modalité $\mathbf{x}_j$ sur la variable **y**, on teste l'hypothèse nulle H_0 : $\alpha_j = 0$.

On considère alors la statistique de Student :

$$t = \frac{\hat{\alpha}_j}{\sqrt{Var(\hat{\alpha}_j)}}$$

où $\hat{\alpha}_j$ est la $j^{ème}$ composante de l'estimateur $\hat{\alpha}$ et $Var(\hat{\alpha}_j)$ est la variance estimée associée à cette composante[1].

Pour tester l'influence d'une variable nominale à q modalités, on procède à un test de nullité des q coefficients αj affectés à ses modalités. D'une manière générale, l'hypothèse H_0 stipulant une éventuelle non-influence d'un ensemble de q variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ sur **y**, s'exprime par la nullité des q coefficients associés :

(H_0) : $\alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$

Notons $\hat{\alpha}_0$ l'estimateur des α_j sous l'hypothèse H_0 et $\hat{\alpha}$ l'estimateur des coefficients du modèle alternatif.

On teste l'hypothèse nulle en calculant le rapport de vraisemblance R:

$$R = \mathcal{L}(\hat{\alpha}_0, \mathbf{y}) / \mathcal{L}(\hat{\alpha}, \mathbf{y})$$

On démontre que $\Lambda = -2 \, log \, R$ suit une distribution du χ^2 à q degrés de liberté sous des hypothèses de travail convenables. Si l'hypothèse nulle est rejetée, on en déduit qu'au moins une des q variables (ou une modalité de la variable nominale) influe sur la variable **y**.

b – Comparaison de deux modèles

Considérons deux modèles emboîtés : le modèle *1* à p variables explicatives et le modèle *2* à $p + q$ variables explicatives comportant entre autres celles du

---

[1] On peut également tester la significativité du coefficient αj à partir de la *statistique de Wald* qui est le carré de celle de Student, et qui suit approximativement une loi du χ^2 à 1 degré de liberté.

modèle *1*. Choisir le modèle *1*, c'est supposer nuls les q coefficients existant dans le modèle *2* et non dans le modèle *1* .

En référence au test de nullité d'un ensemble de coefficients , on retiendra le modèle *1* si l'hypothèse de nullité des q coefficients n'est pas rejetée, c'est-à-dire si la statistique du rapport de vraisemblance Λ est inférieure à la valeur critique du χ^2 à q degrés de liberté. En pratique, le choix du modèle logistique repose sur la comparaison de modèles emboîtés. On adopte une procédure pas à pas en commençant par prendre en compte le modèle comportant le plus de variables explicatives que l'on compare à un modèle restreint comprenant un sous-ensemble des prédicteurs. On procédera généralement par élimination progressive des variables ne modifiant pas de manière significative la vraisemblance jusqu'à avoir un modèle ne pouvant plus être réduit. Cette procédure n'assure cependant qu'un optimum local.

c – Modèle avec interaction

Un fois établi le modèle logistique réduit, certains utilisateurs proposent, pour affiner les résultats, d'introduire des termes d'interaction entre les prédicteurs. Pour cela, on ajoute certains produits des x_j.

Par exemple pour un modèle à deux variables explicatives, le modèle s'écrira :

$$F(\pi(\mathbf{x})) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{12} x_1 x_2$$

La notion d'interaction d'ordre élevé est complexe. Un terme d'interaction d'ordre *2* en modèle logistique correspond au terme d'interaction d'ordre *3* en modèle log-linéaire.

7.6 Segmentation

Les méthodes de segmentation cherchent à résoudre les problèmes de discrimination et de régression en segmentant de façon progressive l'échantillon pour obtenir un *arbre de décision binaire*. La voie a été ouverte par Sonquist et Morgan (1964) et Morgan et Messenger (1973) avec la méthode dite AID (*Automatic Interaction Detection*, cf. Bouroche et Tenenhaus, 1970). De nombreuses contributions ont suivi, mais les travaux de Breiman, Friedman, Olshen et Stone (1984) ont renouvelé l'approche et suscité un regain d'intérêt pour la segmentation. Leur méthode, connue sous le nom de CART (*Classification And Regression Tree*), diffère de l'AID par le mode de construction de l'arbre et la technique d'*élagage* conduisant à un sous-arbre exploitable ayant des propriétés satisfaisantes.

La segmentation par la méthode CART vient donc concurrencer les méthodes plus classiques que sont la régression multiple, l'analyse discriminante et la régression logistique.

Elle présente des avantages importants dont le premier est sans doute la lisibilité des règles d'affectation, l'interprétation des résultats étant directe et intuitive. Par ailleurs la technique est non-paramétrique et peu contrainte par la nature des données. On peut en effet utiliser en même temps comme variables explicatives, des variables continues, ordinales et nominales sans codage préalable.

De plus, la technique fournit d'office la sélection des variables à utiliser en tenant compte d'éventuelles interactions. Elle est robuste vis-à-vis de données erronées ou de valeurs aberrantes et gère les données manquantes aussi bien dans la construction de l'arbre et l'estimation de son risque que dans l'application de la règle à un nouveau sujet. Enfin c'est le même principe, la même méthode, le même algorithme qui sont mis en œuvre pour analyser une variable nominale (discrimination) et une variable continue (régression) [1].

Cependant, les règles d'affectation pourront paraître parfois "abruptes" et trop sensibles à de légères perturbations des données. Il apparaîtra parfois difficile de décider quel est l'arbre "optimal". On peut également regretter l'absence d'une fonction globale mettant en jeu l'ensemble des variables (fonction linéaire discriminante ou équation de régression) qui prive l'utilisateur d'une représentation géométrique sous forme de configurations de points dans l'espace.

7.6.1 Formulation du problème, principe et vocabulaire

Comme en régression (linéaire ou logistique) et en discrimination, on est en présence d'un tableau de données contenant une variable privilégiée $\mathbf{y}$ "à expliquer" par les autres variables du tableau $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

Il s'agit d'une part de sélectionner parmi les variables explicatives celles qui sont les plus discriminantes pour la variable nominale $\mathbf{y}$ (ou celles qui sont le plus liées au phénomène décrit par la variable continue $\mathbf{y}$), et d'autre part de construire une règle de décision permettant d'affecter un nouvel individu à l'une des k classes (cas de la discrimination) ou de lui affecter une valeur y (cas de la régression).

La méthode de segmentation consiste à rechercher d'abord la variable $\mathbf{x}_i$ qui explique le mieux la variable $\mathbf{y}$. Cette variable définit une première division de l'échantillon en deux sous-ensembles, appelés *segments*. Puis on réitère cette

[1] On pourra se reporter pour des éléments théoriques à l'ouvrage cité de Breiman *et al.*, et pour une présentation pratique à l'article de Guegen et Nakache (1988) et aux ouvrages Celeux (1990), Celeux et Nakache (1994), Zighed et Rakotomalala (2000), Nakache et Confais (2003).

procédure à l'intérieur de chacun de ces deux segments en recherchant la deuxième meilleure variable, et ainsi de suite [1].

On construit ainsi un *arbre de décision binaire* par divisions successives de l'échantillon en deux sous-ensembles (figure 7.6 - 1) où l'on distingue :

- les *segments intermédiaires* ou *nœuds* qui engendrent deux segments descendants immédiats,
- les *segments terminaux* qui ne sont plus divisés,
- une *branche* d'un segment t qui comprend tous les segments descendant de t, t n'étant pas inclus dans la branche,
- *l'arbre binaire complet* noté A_{max} pour lequel chaque segment terminal contient un seul individu,
- un *sous-arbre* A qui est obtenu à partir de A_{max} par *élagage* d'une ou de plusieurs branches.

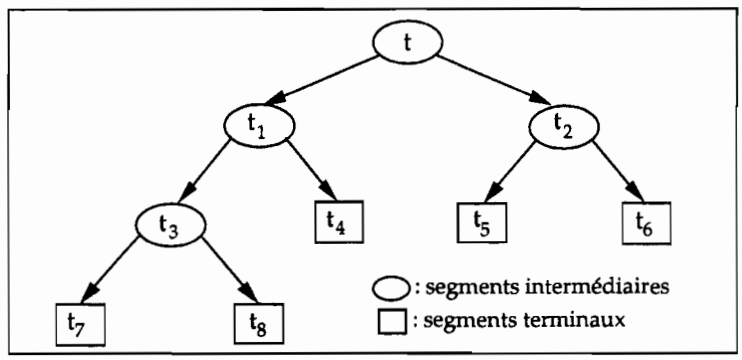


Figure 7.6 - 1
Arbre de décision binaire

Par ailleurs, la méthode CART, contrairement aux autres méthodes de segmentation, n'impose aucune règle (fondée sur un seuil) d'arrêt de division des segments. Elle fournit, à partir de l'arbre binaire complet, la séquence des sous-arbres obtenue en utilisant une *procédure d'élagage*. Celle-ci est basée sur la suppression successive des branches les moins informatives en terme de discrimination entre les classes ou en terme d'explication de la variable **y**.

Au cours de la phase d'élagage, la méthode sélectionne un sous-arbre "optimal" en se fondant sur l'estimation de l'erreur théorique d'affectation ou de prévision à l'aide, soit d'un *échantillon-test* (technique présentée ci-après) quand l'échantillon est suffisamment important, soit de la *validation croisée*.

---

[1] Notons que cette méthode, contrairement aux autres méthodes multidimensionnelles, ne considère pas simultanément l'ensemble des variables explicatives mais les examine une par une. Cependant, les liaisons entre variables explicatives sont prises en compte aux différentes étapes de la construction de l'arbre.

7.6.2 Construction d'un arbre de décision binaire

L'idée de base est d'effectuer la division d'un nœud de telle sorte que les deux segments descendants soient plus homogènes que le nœud parent et qu'ils soient les plus différents possible entre eux vis-à-vis de la variable **y**.

Cette procédure nécessite de définir un critère permettant de sélectionner la "meilleure" division d'un nœud. Le critère de la régression différera de celui de la discrimination, mais le principe de construction reste le même dans les deux cas.

Les différentes phases de construction de l'arbre sont les suivantes :

1- établir pour chaque nœud l'ensemble des divisions admissibles.

2- définir un critère permettant de sélectionner la "meilleure" division d'un nœud.

3- définir une règle permettant de déclarer un nœud comme terminal ou intermédiaire.

4- affecter chaque nœud terminal à l'un des groupes (cas de la discrimination), ou affecter une valeur à y pour chaque nœud terminal (cas de la régression).

5- estimer le risque d'erreur de classement (cas de la discrimination) ou de prévision (cas de la régression) associé à l'arbre.

a _ Algorithme général de segmentation

Les variables explicatives peuvent être de nature quelconque. Dans un premier temps, considérons le cas des variables continues. Les étapes de l'algorithme sont les suivantes :

1 - Au départ, on dispose d'un seul segment contenant l'ensemble des individus.

2 - A la première étape, la procédure de construction de l'arbre examine une par une toutes les variables explicatives.

Pour une variable x_j donnée, elle passe alors en revue toutes les divisions possibles de la forme $x_j < \alpha$ où α est une valeur quelconque contenue dans l'étendue de la variable x_j considérée.

Chaque division scinde l'échantillon en segments descendants : le segment de gauche t_g contient les sujets vérifiant $x_j < \alpha$ et le segment de droite t_d contient les autres ($x_j \geq \alpha$). De toutes les divisions d_j^m possibles de x_j, où m représente la $m^{ième}$ division (soit encore la $m^{ième}$ valeur classée de x_j), la procédure sélectionne la "meilleure" d_j^*, au sens d'un critère de division qui sera précisé.

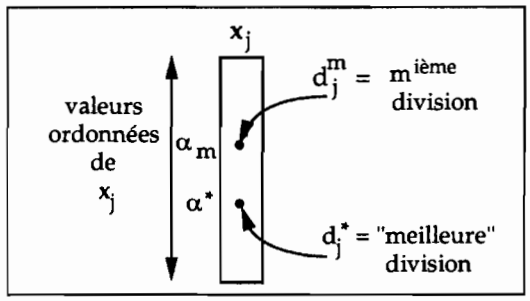


Figure 7.6 – 2. Divisions possibles pour la variable x_j

On obtient ainsi, pour chacune des *p* variables, la meilleure division et l'on retiendra finalement, parmi ces *p* divisions, celle, notée d*, qui fournit les deux segments les plus "typés" vis-à-vis de **y** .

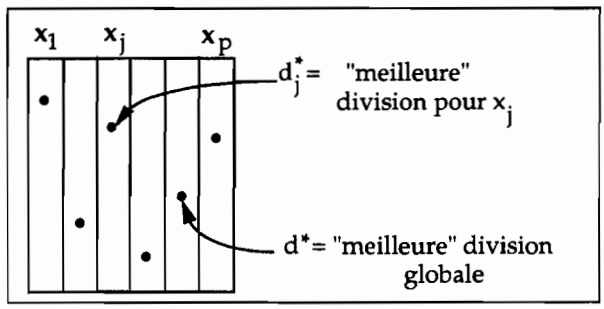


Figure 7.6 – 3. Meilleures divisions pour l'ensemble des variables

3 - A l'étape suivante, on applique la même procédure à chacun des deux segments descendants obtenus. Les variables explicatives peuvent être différentes selon les segments.

4 - On arrête la procédure lorsque tous les segments sont déclarés terminaux : soit parce qu'ils ne nécessitent plus de divisions soit parce que leur taille est inférieure à un effectif fixé.

Pour un nouvel individu, on définit une règle d'affectation simple en le faisant *descendre* dans l'arbre.

Si, parmi les variables explicatives, certaines sont nominales, elles sont prises en compte de la manière suivante :

- une variable à deux modalités ne peut fournir qu'une seule division,
- une variable à *k* modalités ordonnées fournit *k - 1* divisions,
- une variable à *k* modalités non ordonnées fournit $2^{k-1} - 1$ divisions; toutes les divisions correspondant aux différents sous-ensembles de modalités sont examinées.

Tableau 7.6 – 1. Divisions possibles d'un segment par une variable nominale

	t_g	t_d
var. binaire	(a_1)	(a_2)
variable	(b_1)	(b_2, b_3, b_4)
ordonnée	(b_1, b_2)	(b_3, b_4)
(ordinale)	(b_1, b_2, b_3)	(b_4)
variable	(c_1)	(c_2, c_3)
non	(c_2)	(c_1, c_3)
ordonnée	(c_3)	(c_1, c_2)

Par exemple, à partir d'une variable **a** à deux modalités, d'une variable **b** à 4 modalités ordonnées et d'une variable **c** à 3 modalités non ordonnées, les divisions possibles d'un nœud en deux segments descendants t_g (celui de gauche) et t_d (celui de droite) sont les suivantes (cf. tableau 7.6-1)[1] :

b – Cas de la régression

Lorsque la variable à expliquer **y** est continue, le critère de sélection de la "meilleure" division d'un nœud est fondé sur la variance de **y** dans les segments descendants qui doit être plus faible que dans le nœud parent.

- Critère de la variance résiduelle minimale

Pour toute division d_j^m d'un nœud t par une variable x_j, on calcule la moyenne pondérée des variances de **y** à l'intérieur de chacun de ses segments descendants t_g et t_d, c'est-à-dire la variance résiduelle du nœud t :

$$var(d_j^m, t) = \left(\frac{n_g}{n_t} s_g^2\right) + \left(\frac{n_d}{n_t} s_d^2\right)$$

où n_g, n_d, n_t sont respectivement les effectifs des segments t_g, t_d, t et s_g^2, s_d^2 sont les variances de la variable continue **y** à l'intérieur des segments t_g et t_d [2].

On retient la "meilleure" division d_j^* réalisée par la variable x_j qui correspond à la variance résiduelle minimale :

$$var(d_j^*, t) = \min_{m \in d_j} \{var(d_j^m, t)\}$$

où d_j est l'ensemble des divisions de la variable x_j.

---

[1] Remarquons que la segmentation effectue simultanément un découpage sur la population observée et sur les valeurs des variables explicatives.

[2] Il s'agit de la variance interne ou *intra* (non expliquée par la coupure).

Parmi toutes les meilleures divisions d_j^* obtenues à partir des p variables explicatives, la meilleure division (globale) du nœud t est effectuée à l'aide de la variable qui assure :

$$var(d^*, t) = \min_{j=1,\ldots,p} \{var(d_j^*, t)\}$$

- Les étapes de l'algorithme

Considérons un ensemble d'individus sur lesquels on relève les informations concernant une variable continue **y** et $p = 8$ variables explicatives $\mathbf{x}_1, \ldots, \mathbf{x}_8$. On suppose que les valeurs de **y** ont pour moyenne $m = 10$ et pour variance $s^2 = 60$.

On commence par examiner la variable continue $\mathbf{x}_1$.

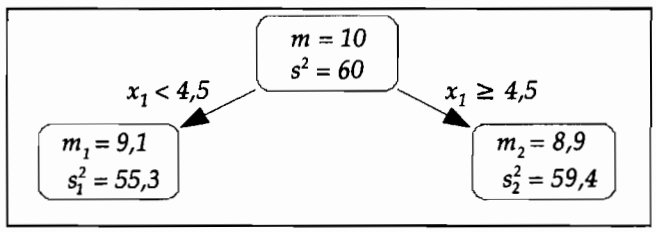


Figure 7.6 – 4. Régression : meilleure division pour la variable $\mathbf{x}_1$

On retient la valeur de coupure qui minimise la variance à l'intérieur des deux segments descendants, par exemple la division associée à la valeur 4,5 (cf. figure 7.6 - 4)

Mais cette meilleure division obtenue avec $\mathbf{x}_1$ n'est peut-être pas la plus efficace en terme de réduction de la variance. Il faut étudier les autres variables. On recherche, de la même manière, la meilleure division de l'échantillon pour chacune des $p - 1 = 7$ autres variables. On choisira alors la division qui présente la plus faible moyenne pondérée des variances de **y** à l'intérieur des deux segments descendants, par exemple la variable continue $\mathbf{x}_5$ pour la valeur $\alpha = 7,2$.

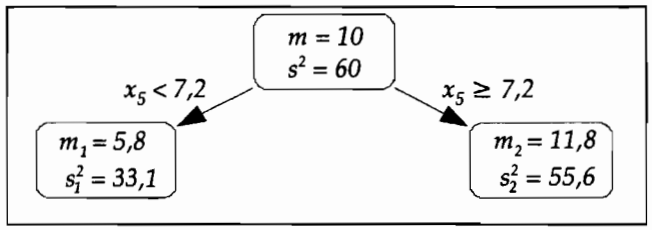


Figure 7.6 – 5. Régression : meilleure division pour toutes les variables

On réitère cette procédure à l'intérieur de chacun des deux segments obtenus t_1 et t_2. Pour le segment t_1, ce sera par exemple la variable nominale $\mathbf{x}_7$ à deux modalités ; la meilleure division sera obtenue pour les valeurs $\mathbf{x}_7$=*1* (segment t_3), et $\mathbf{x}_7$=2 (segment t_4).

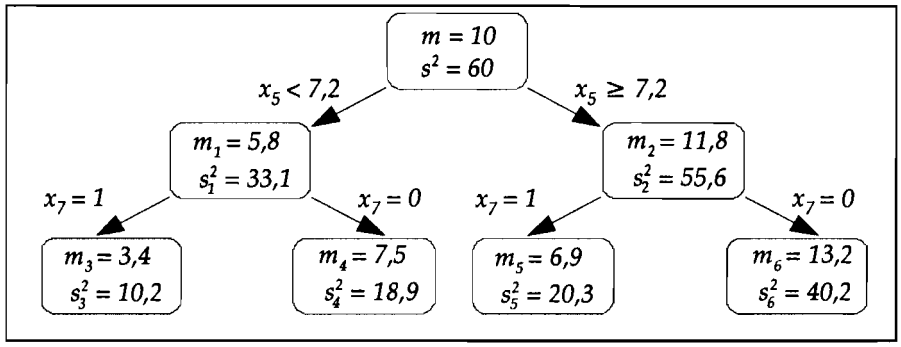


Figure 7.6 – 6. Régression : Arbre à deux niveaux

On sélectionnera la variable x_2 à deux modalités, pour le segment t_2. On aboutit ainsi à l'arbre à deux niveaux représenté sur la figure 7.6 - 6. (Sur cette figure, l'indice bas des variances est celui des segments correspondants : s_t^2 correspond au segment t_i). On pourrait arrêter là la procédure de division et produire l'arbre de prédiction à 4 segments terminaux.

- Règle d'affectation

Considérons alors un nouvel individu i dont on cherche à prévoir la valeur de y_i. Il tombera dans un de ces 4 segments terminaux après avoir parcouru un chemin de l'arbre suivant les valeurs qu'il présente pour x_5, x_7 et x_2. La valeur affectée à y_i sera la moyenne dans le segment et l'écart-type correspondra à celui du segment.

- Erreur Apparente de Prévision (EAP) associée à un arbre A

Si certaines variances des segments sont encore importantes, on peut continuer les divisions dans le but de réduire davantage les variances des segments terminaux. Ainsi on associe à chaque segment terminal t de l'arbre A l'erreur R_t suivante :

$$R_t = \frac{n_t}{n} \times s_t^2$$

où n est le nombre total d'individus, n_t est le nombre d'individus du segment t, s_t^2 est la variance de la variable y à l'intérieur du segment t c'est-à-dire :

$$s_t^2 = \frac{1}{n_t} \sum_i (y_i - \bar{y}_t)^2$$

avec $\bar{y}_t$, la moyenne des valeurs y_i des individus du segment t.

L'*Erreur Apparente de Prévision (EAP)* associée à l'arbre A vaut :

$$EAP(A) = \sum_{t \in A} R_t \qquad [7.6 - 1]$$

et correspond à la moyenne pondérée des variances de y dans chacun des segments terminaux de l'arbre A.

Le rapport $EAP(A)/s^2$ est l'équivalent de l'expression $(1 - R^2)$ de la régression linéaire multiple[1] et représente le pourcentage de la variance totale non expliquée par les variables $x_1, x_2, \ldots, x_p$. Plus on divise, plus les variances décroissent pour être finalement nulles quand chaque segment terminal contient un seul individu. Au grand arbre complet noté A_{max} ainsi obtenu est affectée une Erreur Apparente de Prévision nulle.

c – Cas de la discrimination

Lorsque la variable **y** est nominale et répartit les individus en *k* classes, la sélection d'une division doit être telle que les segments descendants soient plus "purs" que le nœud parent. Autrement dit, il faut que le mélange des classes soit moins important dans les segments descendants que dans le nœud parent.

- Critère de la pureté maximale

A chaque segment t est donc associée une mesure de l'*impureté i(t)* définie par :

$$i(t) = \sum_{r}^{k} \sum_{s}^{k} P(r|t)P(s|t)$$

avec $r \neq s$ et où $P(r|t)$ et $P(s|t)$ sont les proportions d'individus dans les classes c_r et c_s dans le segment[2] t.

Un segment est *pur* s'il ne contient que des individus d'une seule classe, dans un tel cas : $i(t) = 0$. Plus le mélange des classes dans le segment t est important, plus l'impureté $i(t)$ est élevée.

Chaque division d_j^m du nœud t par la variable x_j entraîne une réduction de l'impureté qui s'exprime par :

$$\Delta_j^m = i(t) - p_g i(t_g) - p_d i(t_d)$$

où p_g et p_d sont les proportions d'individus du nœud t respectivement dans les segments descendants t_g et t_d (la fonction $i(t)$ étant concave, l'impureté moyenne ne peut que décroître par division d'un nœud).

Par conséquent pour chaque variable x_j, la meilleure division d_j^* est telle que la réduction de l'impureté Δ_j^* est maximale : $\Delta_j^* = \max_{m \in d_j} \{\Delta_j^m\}$

---

[1] Dans la régression linéaire multiple, on suppose que la variance de la réponse **y** conditionnellement aux covariables (variables explicatives) est constante, ce qui n'est pas le cas pour la régression par arbre.

[2] La fonction $i(t)$ est l'*indice de diversité de Gini* (cf. Goodman et Kruskal, 1954). On aurait pu également utiliser l'*entropie de Shannon* : $i(t) = -\sum_{r}^{k} P(r|t) \log P(s|t)$.

où d_j est l'ensemble des divisions de la variable x_j. Sur l'ensemble des p variables, la division du nœud t est effectuée à l'aide de la variable qui assure :

$$\Delta^* = \max_{j=1,\ldots,p} \{\Delta_j^*\}$$

- Les étapes de l'algorithme

Considérons maintenant *300* individus répartis en 3 classes c_1, c_2, c_3 de même taille et sur lesquels *10* mesures quantitatives ont été relevées. On procède comme dans le cas de la régression par segmentation en examinant toutes les variables. Pour la variable x_1, on aboutit par exemple à la meilleure division (qui n'est pas nécessairement la plus discriminante) observable sur la figure 7.6 - 7.

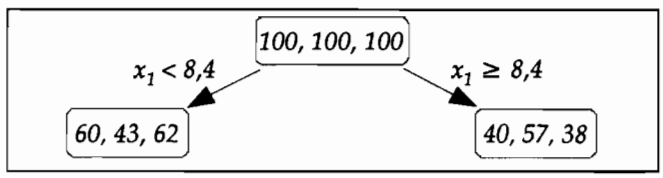


Figure 7.6 – 7. Discrimination : meilleure division pour la variable x_1

On retient finalement, parmi toutes les variables, celle qui produit la meilleure *"meilleure division"*, par exemple la variable continue x_8 pour $\alpha = 3{,}5$.

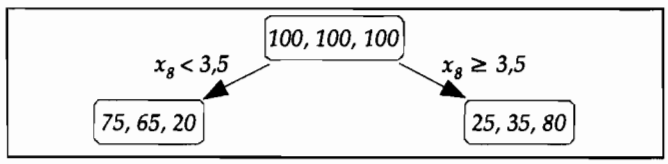


Figure 7.6 – 8. Discrimination : meilleure division pour toutes les variables

On obtient ainsi la meilleure séparation entre les 3 classes, ce qui se traduit par le schéma de la figure 7.6 - 8. On applique cette même procédure aux deux segments descendants obtenus.

- Règle d'affectation

Si on considère le segment terminal t de taille n_t, il contient $n_1(t)$ sujets appartenant à la classe *1*, ..., $n_r(t)$ sujets de la classe *r*, ..., $n_k(t)$ sujets de la classe *k*. Chaque segment terminal est affecté à la classe qui y est la mieux représentée.

Par exemple, les segments *1* et *4* de la figure 7.6 - 9 sont affectés à la classe 2. Un nouvel individu qui *descend* dans l'arbre arrive dans un segment terminal et sera affecté à la classe correspondante.

- Taux d'Erreur Apparente de classement

A tout segment terminal t de l'arbre A associé à une classe c_s correspond une erreur de classement de la forme :

$$R(s|t) = \sum_{r=1}^{k} P(r|t)$$

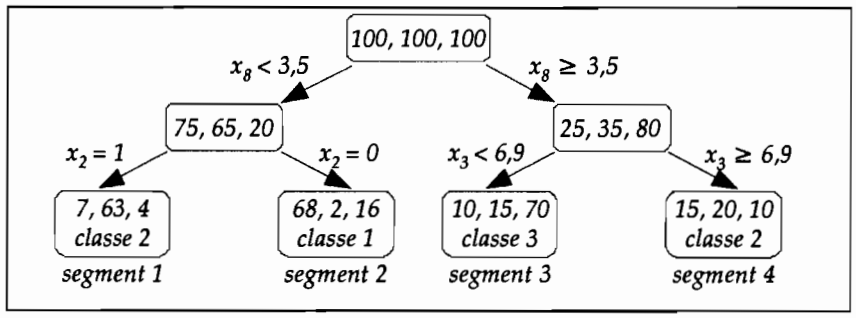


Figure 7.6 – 9. Discrimination : Arbre à deux niveaux

avec $r \neq s$ et où $P(r|t) = \frac{n_r(t)}{n_t}$ est la proportion d'individus du segment t affectés à la classe c_s et qui appartiennent à la classe c_r.

Le *Taux d'Erreur Apparent de classement* (*TEA*) associé à l'arbre vaut :

$$TEA(A) = \sum_{t \in A} \frac{n_t}{n} R(s|t) = \sum_{t \in A} \sum_{r=1}^{k} \frac{n_r(t)}{n} \qquad [7.6 - 2]$$

avec $r \neq s$.

Il représente la proportion d'individus mal classés dans l'ensemble des segments terminaux.

Ainsi, l'arbre de la figure 7.6 - 9 ne fournit pas une bonne règle de décision en terme d'erreur de classement. En effet, un sujet qui parcourt l'arbre et qui tombe dans le segment *1* est affecté à la classe *2* avec une erreur de classement de *14,9 %*; celui qui tombe dans le segment *4* est affecté à la classe *2* avec une erreur de classement de *55,5 %*.

Le Taux d'Erreur Apparent de classement associé à l'arbre est la moyenne des erreurs de classement dans les différents segments terminaux, soit :

$$TEA = \frac{(74 \times 14,9\% + 86 \times 20,9\% + 95 \times 26,3\% + 45 \times 55,5\%)}{300} = 26,3\%$$

On a sans doute intérêt à continuer à diviser les segments.

La question est de savoir à quel moment il faut arrêter la procédure de division.

7.6.3 Sélection du "meilleur sous-arbre"

Par "meilleur" sous-arbre, on entend un arbre qui contient le moins de segments terminaux et dont l'erreur apparente de prévision ou de classement est la plus petite possible, tout en fournissant une estimation correcte de l'erreur théorique.

Un sous-arbre ayant peu de segments terminaux entraîne une erreur apparente qui, bien que reflétant l'erreur théorique, est trop importante.

En effet, si l'arbre est trop petit, on peut être conduit à perdre de bonnes divisions et à ne pas utiliser toute l'information contenue dans l'échantillon.

Inversement, à un arbre trop grand (avec de nombreuses divisions) est associée une erreur apparente faible mais qui donne une estimation trop optimiste de l'erreur théorique. C'est donc entre ces deux extrêmes que doit être choisi le "meilleur" sous-arbre.

La méthode proposée par Breiman et *al.* (*op. cit.*) est fondée sur l'utilisation d'un échantillon-test et présente un double avantage :

- déterminer le "meilleur" sous-arbre sans employer de tests statistiques pour définir une règle d'arrêt de la procédure de division,
- obtenir une estimation précise de l'erreur théorique de prévision ou de classement.

a – Procédures de sélection

Il est nécessaire de diviser l'échantillon de base en deux parties, l'*échantillon d'apprentissage* (par exemple constitué par les 2/3 de l'échantillon de base) et l'*échantillon-test* (le tiers restant). La recherche du "meilleur" sous-arbre A* se fait alors de la façon suivante :

- A partir de *l'échantillon d'apprentissage*, on construit l'arbre complet A_{max} ou un arbre tel que chaque segment terminal contienne peu d'individus.
- Puis l'opération d'*élagage* de l'arbre A_{max} consiste à construire une séquence optimale de sous-arbres emboîtés $\{A_H, \ldots, A_h, \ldots, A_1\}$ où A_H coïncide avec A_{max}, A_h est le sous-arbre ayant h segments terminaux et A_1 est l'échantillon total. Chaque sous-arbre A_h de cette séquence est optimal au sens suivant : son Erreur Apparente (EA) est minimale parmi les sous-arbres ayant le même nombre de segments terminaux[1].
- Si S_h est l'ensemble des sous-arbres de A_{max} ayant h segments terminaux alors :

$$EA(A_h) = \min_{A \in S_h} \{EA(A)\}$$

- A partir de *l'échantillon-test*, on sélectionne, parmi les sous-arbres de la séquence optimale, le meilleur sous-arbre A*. C'est celui qui présente la plus petite erreur théorique (*ET*) :

---

[1] En fait, des algorithmes appropriés permettent de choisir une séquence sous-optimale, mais accessible par le calcul (cf. Breiman *et al.*, 1984; Celeux et Lechevallier, dans : Celeux, 1990).

$$ET(A^*) = \min_{1 \le h \le H} \{ET(A_h)\}$$

- Les individus de l'échantillon-test parcourent chacun des sous-arbres de la séquence optimale et tombent dans un segment terminal, ce qui entraîne une estimation de l'erreur théorique pour chaque sous-arbre. En pratique, l'estimation de l'erreur théorique décroît rapidement à mesure que le nombre de segments terminaux des sous-arbres augmente, puis elle passe par un palier et croît ensuite lentement. Le sous-arbre A* sélectionné comme optimal est le plus petit sous-arbre associé à l'estimation la plus petite de l'erreur théorique.

b – Estimation de l'Erreur Théorique de Prévision

L'estimation de l'*Erreur Théorique de Prévision* pour un sous-arbre A de la séquence optimale, $\hat{ETP}(A)$, est calculée sur l'échantillon-test suivant la formule utilisée pour l'*Erreur Apparente de Prévision* [7.6 - 1] :

$$\hat{ETP}(A) = \sum_{t \in A} \tilde{R}_t$$

avec $\tilde{R}_t = \frac{\tilde{n}_t}{\tilde{n}} \times \tilde{s}_t^2$ et où $\tilde{n}$ est la taille de l'échantillon-test, $\tilde{n}_t$ est le nombre d'individus de l'échantillon-test qui appartiennent au segment t et $\tilde{s}_t^2$ est la variance de la variable **y** à l'intérieur du segment t.

c – Estimation du Taux d'Erreur Théorique de classement

Les appellations de *Taux d'Erreur Apparent* ou *Théorique de Classement* n'ont de sens que dans le cas le plus simple c'est-à-dire si les probabilités *a priori* des classes sont estimées par les fréquences des classes dans l'échantillon et si les coûts de mauvaise classification sont tous égaux.

Dans le cas général, on utilise un *Coût d'Erreur Apparent* ou *Théorique* pour lesquels les formules de calcul sont plus complexes.

- Cas le plus simple

L'estimation du *Taux d'Erreur Théorique de classement* se calcule comme le *Taux d'Erreur Apparent* [7.6 - 2] à partir de l'échantillon-test. Elle est égale à la proportion $\tilde{p}_t$ d'individus mal classés par le sous-arbre A dans l'échantillon-test (cf. formule [7.6 - 2]).

$$\hat{TEA}(A) = \sum_{t \in A} \sum_{r=1}^{k} \frac{\tilde{n}_r(t)}{\tilde{n}} = \tilde{p}_t$$

avec $r \neq s$, où $\tilde{n}$ est l'effectif de l'échantillon-test et $\tilde{n}_r(t)$ est le nombre d'individus de l'échantillon-test affectés à la classe c_s et qui appartiennent à la classe c_r dans le segment terminal t.

Il est possible de fournir un intervalle de confiance associé à cette proportion $\tilde{p}_t$ à partir de l'estimation de la variance de cette proportion :

$$\hat{Var}(\tilde{p}_t) = \frac{\tilde{p}_t(1-\tilde{p}_t)}{\tilde{n}}$$

- *Cas général*

La règle de décision la plus générale est celle qui tient compte des probabilités *a priori* π_r $(r = 1,2,\ldots, k)$ des k classes à discriminer et des *coûts de mauvais classement* notés $C(r|s)$ où $r \neq s = 1, 2,\ldots,k$.

$C(r|s)$ désigne le coût[1] entraîné par l'affectation d'un individu à la classe c_s alors qu'il appartient à la classe c_r. La règle générale d'affectation d'un segment terminal t à une classe est fondée sur le *coût moyen d'erreur de classement* (appelé aussi risque d'erreur).

Si $n_r(t)$ désigne le nombre d'individus de la classe c_r du segment t et n_r l'effectif total de la classe c_r, on a :

$$P(r|t) = \frac{p_r \dfrac{n_r(t)}{n_r}}{P(t)}$$

où $P(t) = \sum_{r=1}^{k} p_r \dfrac{n_r(t)}{n_r}$ est la probabilité d'"aboutir" au segment t.

Le coût moyen d'erreur de classement $R(s|t)$ entraîné par l'affectation du segment t à la classe c_s est égal à :

$$R(s|t) = \sum_{r=1}^{k} C(r|s)P(r|t)$$

Ainsi le segment terminal t est affecté à la classe c_j si :

$$R(l|t) = \min_{s=1,\ldots,k} \{R(s|t)\}$$

Remarque

Si la probabilité π_r d'appartenance *a priori* à la classe c_r est égale à la proportion d'individus de cette classe dans l'échantillon :

$$\pi_r = \frac{n_r}{n}$$

alors $P(t)$ tel que :

$$P(t) = \sum_{r=1}^{k} p_r \frac{n_r(t)}{n_r}$$

est simplement la proportion d'individus composant le segment terminal t.

---

[1] Les différents coûts $C(s|s)$ sont nuls et en général $C(r|s) \neq C(s|r)$.

7.6.4 Divisions équi-réductrices et équi-divisantes

La *meilleure* division d* d'un nœud est celle qui assure la plus grande réduction de la variance résiduelle ou de l'impureté en passant du nœud à ses segments descendants. Cette notion de maximum absolu est très stricte. Il peut exister en effet des divisions presque aussi bonnes, pouvant jouer un rôle important au niveau des interprétations.

Par extension, on définit, à côté de d*, deux autres types de divisions :

- les divisions *équi-réductrices* qui assurent, après d*, les plus fortes réductions de l'impureté ou les plus faibles variances résiduelles. Elles permettent d'intervenir sur le choix de la "meilleure" variable explicative.

- les divisions *équi-divisantes* qui fournissent les répartitions les plus proches de la meilleure division d*. Elles permettent de gérer l'existence de données manquantes dans l'affectation d'un nouvel individu à une classe ou à une valeur de **y**.

a – Divisions équi-réductrices

La procédure de division d'un nœud fournit les premières *meilleures divisions* d'un nœud pour lesquelles la réduction de la variance résiduelle ou de l'impureté Δ_j^* est élevée (cf. 3.5.2.b et c). Si la meilleure division d* du nœud t est obtenue à partir de la variable $\mathbf{x}^*$, on définit la première division équi-réductrice d_i^* effectuée sur la variable $\mathbf{x}_i$ ($\mathbf{x}_i \neq \mathbf{x}^*$) avec $i = 1,\ldots,p$. C'est celle qui correspond à une réduction des segments descendants la plus proche de celle de la *meilleure division* d*.

En d'autres termes, c'est la deuxième meilleure division du nœud t. On définit par extension les $2^{ème}$, $3^{ème}$,..., divisions équi-réductrices[1].

b – Divisions équi-divisantes

Les divisions équi-divisantes (ou: suppléantes) permettent de classer un nouvel individu présentant une donnée manquante pour la variable définissant la division. L'idée est la suivante : on cherche une variable qui remplace au mieux la variable divisant le nœud, c'est-à-dire qui assure presque la même séparation des individus. De la même manière, on peut définir la seconde, troisième,..., meilleure division équi-divisante. Ainsi, si la valeur de $\mathbf{x}_j$ est manquante pour un nouvel individu, on l'affectera à un des segments descendants en utilisant la meilleure division équi-divisante de d*. Et si la valeur de la variable associée à la meilleure division équi-divisante est manquante, on aura recours à la deuxième meilleure division équi-divisante, etc.

---

[1] Les divisions équi-réductrices sont parfois appelées concurrentes. Il est possible ainsi d'intervenir sur le choix des variables associées aux "meilleures" divisions suivant la perception personnelle qu'a l'utilisateur du problème. En effet, à la variable produisant la "meilleure" division, on peut préférer une autre variable que l'on sait plus pertinente pour l'étude.

7.6.5 Lien avec les méthodes de classement

Segmentation, discrimination, classement, classification ou classification supervisée, régression linéaire multiple, régression logistique, régression pas-à-pas, ..., le vocabulaire ne manque pas pour désigner, suivant le domaine d'application, des opérations qui sont souvent proches. On va, dans ce paragraphe, brièvement situer la segmentation parmi les autres outils.

La segmentation, bien que travaillant par divisions de l'échantillon en classes, est plus proche des techniques de régression pas à pas (qu'il s'agisse de régression linéaire multiple ou de régression logistique) et de discrimination pas-à-pas que des méthodes non-supervisées de classification. En effet, il ne s'agit pas de faire apparaître des classes, mais de chercher les groupes d'individus les plus "explicatifs" des modalités d'une variable qualitative particulière (ou des valeurs d'une variable continue). Le principe est, on l'a vu, de chercher la dichotomie (induite à chaque pas par *une* des variables) la plus "liée" à la variable privilégiée.

La segmentation n'est pas vraiment multidimensionnelle au sens géométrique du terme (on ne calcule pas de distances dans $\mathcal{R}^p$ ni dans $\mathcal{R}^n$ comme pour les méthodes factorielles ou de classification), mais on utilise les variables explicatives conditionnellement les unes par rapport aux autres. On peut donc parfois atteindre des effets d'interaction assez difficiles à saisir par d'autres méthodes, sans prétendre d'ailleurs les atteindre tous. La parenté avec les méthodes descriptives reste forte, dans la mesure où les aspects "contrôle des opérations par l'utilisateur", "transparence du fonctionnement", voire "ergonomie des résultats" occupent une position de premier plan. L'arbre de décision binaire est lisible par tout utilisateur. Autre avantage déjà évoqué dans l'introduction de cette section, la mixité des variables qu'accepte la procédure : nominales, ordinales, continues peuvent être mélangées au niveau des variables explicatives, et peuvent constituer la variable à expliquer. La validation par une méthode de rééchantillonnage (limité aux échantillons-test dans l'exposé qui précède) est elle-même une des techniques de validation les plus transparentes pour l'utilisateur.

Pour conclure, on doit cependant reconnaître quelques défauts à la segmentation par arbre binaire, qui rendent son *utilisation exclusive* insuffisante.

L'aspect séquentiel est redoutable, car les covariations qui servent à sélectionner les variables ne mesurent pas un lien causal et une variable peut en cacher une autre, beaucoup plus fondamentale, qui n'a plus aucune chance d'apparaître dans la suite du processus. Les divisions de réserve (équi-réductrices et équi-divisantes) sont là pour pallier partiellement cet inconvénient. Mais l'arbre binaire perd alors une partie de sa séduisante simplicité. L'absence de visualisation globale, propice à une réflexion critique sur le recueil de données et à une observation simultanée de l'ensemble des covariations, est également une faiblesse par rapport aux méthodes factorielles.

Enfin, il se peut que la nature du phénomène étudié fasse que des combinaisons linéaires (après éventuel recodage) soient optimales pour prévoir la variable étudiée (ou son *logit* ou toute autre fonction). Dans ce cas, la segmentation progressive sera surclassée. Ces quelques critiques ne portent cependant que sur l'usage exclusif de la segmentation par arbre binaire. Une démarche impliquant plusieurs points de vue (visualisation préalable des variables explicatives avec positionnement *a posteriori* de la variable à expliquer, régression ou discrimination) permet d'éviter la plupart des écueils mentionnés.

7.7 Discrimination et réseaux de neurones

Cette section ne constitue qu'un brève survol, accompagné de quelques références, destiné à aider le lecteur statisticien désireux d'aborder les techniques neuronales de discrimination. Développées au milieu des années quatre-vingt, les méthodes neuronales (ou réseaux neuronaux ou encore neuro-mimétiques) ont renouvelé et stimulé la discipline connue sous le nom de *reconnaissance de formes* qui recouvre beaucoup d'applications industrielles (notamment des applications en temps réel) des méthodes de discrimination.

Fondées au départ sur des analogies biologiques et sur un effort de modélisation des mécanismes de perception visuelle et auditive, ces méthodes ont acquis depuis une certaine autonomie. Les relations avec la statistique ont été frileuses en raison de différences d'approches et de vocabulaires[1] . Mais des ponts ont été jetés et les années récentes ont vu la parution d'une série d'articles de revue ou de synthèse qui ont prouvé la complémentarité des points de vue et l'enrichissement mutuel à attendre des contacts et échanges entre statisticiens et neuromiméticiens[2].

Schématiquement, disons que les statisticiens peuvent compléter la panoplie des modèles qui leur sont familiers avec les modèles essentiellement *non-linéaires* et *à seuils* qui sont attachés aux réseaux de neurones. La structure de ces réseaux permet d'autre part des calculs parallèles indispensables pour une implémentation matérielle directe de ces méthodes et des utilisations en temps réel, domaine peu abordé par les statisticiens. Inversement, l'essentiel de ce qui concerne l'inférence ou la validation des démarches et des résultats est à mettre au crédit des approches statistiques. Ces aspects sont reconnus comme indispensables dès qu'il s'agit de comparer des modèles, d'évaluer des risques,

---

[1] Ce sont des informaticiens en milieu industriel qui sont à l'origine de ces méthodes.

[2] Citons en particulier les articles de synthèse de Ripley (1993, 1994) et de Cheng et Titterington (1994), les ouvrages de Bishop (1995) et de Thiria *et al.* (1997).

de calculer des taux d'erreurs, préoccupations caractéristiques d'une discipline arrivée à maturité. On évoquera seulement dans cette note bibliographique le modèle neuronal le plus répandu dans le cadre de la discrimination qui est le *perceptron multi-couche,* puis on dira quelques mots des méthodes non-supervisées, enfin on évoquera les séparateurs à vastes marges (SVM).

7.7.1 Schéma et modèle du perceptron multi-couches

Le contexte est le même que celui qui a été défini au début de cette section. On dispose d'une variable qualitative y à q modalités (ou catégories) que l'on doit prédire à partir de p variables $(x_1, x_2, \ldots, x_p)$ prédictrices. On dispose par ailleurs de n individus ou observations (échantillon d'apprentissage) décrits par les p variables $(x_1, x_2, \ldots, x_p)$ et pour lesquels ont connaît la classe d'affectation notée ici y_k $(k \leq q)$.

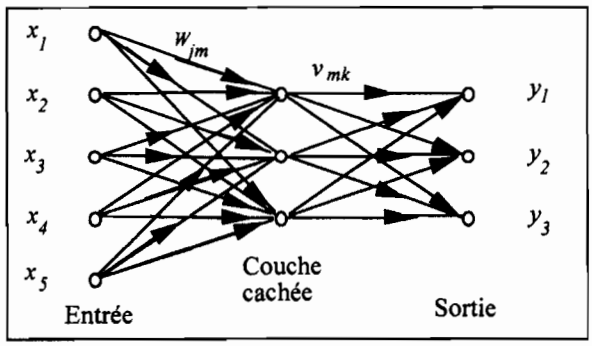


Figure 7.7 – 1. Perceptron à une couche cachée

La figure 7.7 - 1 se commente de la façon suivante en utilisant le vocabulaire et les concepts de l'approche neuronale : la couche d'entrée est formée de $p = 5$ *entrées,* auxquelles seront appliquées des coefficients appelés les *poids synaptiques* w_{jm}. La *couche cachée* comprend $c = 3$ *neurones* qui seront chacun *activés* par une intégration (en général fonction monotone de la somme) des p signaux en provenance de la couche d'entrée. La même opération a lieu pour les $q = 3$ éléments de la couche de sortie mettant en jeu des poids synaptiques v_{mk}.

En termes de modèle analytique, on écrira :

$$y_k = \Phi_o\left\{a_k + \sum_{m=1}^{c} v_{mk}\ \Phi\left(a_m + \sum_{j=1}^{p} w_{jm} x_j\right)\right\} \qquad [7.7 - 1]$$

Dans cette formule, la fonction Φ est dans la plupart des applications la fonction logistique abordée au paragraphe 7.5.2. Elle s'écrit :

$$\Phi(z) = \frac{exp\{z\}}{1 + exp\{z\}}$$

La fonction Φ_0 peut être selon les cas linéaire, logistique, ou à seuil (par exemple : $\Phi_0(z) = 0$ si $z \leq 0$ et $\Phi_0(z) = 1$ si $z > 0$).

On voit que la figure 7.7 - 1 est utile pour visualiser l'enchaînement de fonctions correspondant aux étapes du traitement. La lecture de droite à gauche de la figure correspond bien sûr à une lecture de gauche à droite de la formule [7.7 - 1]. Il y a $\{c\,(p+1) + q\,(c+1)\}$ paramètres à estimer.

L'équation [7.7 - 1] correspond à une observation *(i)*. On a en réalité n équations de ce type, faisant chacune intervenir q valeurs y_{ik} (valeurs *0* ou *1* s'il s'agit d'appartenance à une classe d'une partition en q classes) et p valeurs x_{ij}.

7.7.2 Modèles supervisés

L'estimation des paramètres se fait en minimisant une fonction de perte f, qui peut simplement être la somme des carrés des écarts entre les valeurs calculées $\tilde{y}_{ik}$ et les valeurs observées y_{ik} dans l'échantillon d'apprentissage[1].

Remarquons que pour une sortie binaire (deux classes possibles pour y qui peut alors être un scalaire prenant les valeurs *0* ou *1*) et un perceptron sans couche cachée, on retrouve la régression logistique évoquée au paragraphe 7.5.2.

La formule [7.7 - 1] s'écrit alors :

$$y = \Phi_o\left\{\Phi\left(a_m + \sum_{j=1}^{p} w_{jm} x_j\right)\right\} \qquad [7.7 - 2]$$

Ici, la fonction Φ_0 peut être une fonction à seuil, qui convertit la probabilité donnée par le modèle logistique proprement dit (à l'intérieur des accolades) en l'une des deux valeurs *0* ou *1*.

Si l'on réduit les deux fonctions Φ_0 et Φ à la fonction identique $\Phi(x) = x$, on retrouve la régression multiple (cf. section 2.2) et l'analyse discriminante à deux groupes (cf. paragraphe 7.2.1) qui en sont des cas particuliers.

Cet exemple très simple du perceptron multi-couches montre donc que les généralisations les plus évidentes par rapport aux modèles explicatifs usuels de la statistique concernent la présence éventuelle des fonctions Φ_0 et Φ et l'existence d'une ou plusieurs couches cachées qui autorisent des interventions non-linéaires des paramètres [2].

---

[1] L'estimation numérique se fait par une *descente de gradient* dite de *back-propagation*. (cf. Werbos, 1974, 1990; Rumelhart et al., 1986). Pour un programme de calcul, cf. Proriol (1991).

[2] Notons que dans un modèle comme celui de la formule [7.7 - 1], il n'est pas nécessaire de retenir toutes les flèches entre deux couches consécutives (certain poids synaptiques peuvent être nuls *a priori*) et l'on réduit le nombre de paramètres à estimer).

Reprenons l'exemple du perceptron multicouche, pour lequel nous supposerons les fonctions Φ_0 et Φ linéaires ou, sans perte de généralité dans ce dernier cas, égales à la fonction identique. Nous supposerons de plus que les variables sont des variables numériques centrées, et que les termes constants sont nuls.

Un cadre général (cf. Baldi et Hornik, 1989) permet de traiter simultanément les cas supervisés et non-supervisés. On explicite maintenant l'indice i de l'observation i (appelé "exemple" par les neuromiméticiens), et le résidu e_{ik} non expliqué par le modèle. La formule [7.7-1] s'écrit :

$$y_{ik} = \sum_{m=1}^{c} v_{mk} \left(\sum_{j=1}^{p} w_{jm} x_{ij} \right) + e_{ik} \qquad [7.7 - 3]$$

On peut encore écrire [7.7 - 3] sous la forme matricielle: $\mathbf{Y} = \mathbf{XWV} + \mathbf{E}$

Le rang de la matrice **WV**, d'ordre (p, q), est au plus égal au plus petit des trois nombres q, c, p. Si la taille c de la couche cachée n'introduit pas de restriction sur le rang de **WV**, on peut poser **A=WV**.

On est alors dans le cadre de la régression multiple simultanée comportant plusieurs variables endogènes, et la minimisation de *f=trace* (**E'E**) équivaut à faire q régressions multiples (cf. chapitre 2). On a alors : $\mathbf{A} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$.

Si la taille c de la couche cachée induit une contrainte de rang ($c < min(p, q)$), on peut introduire la contrainte $\mathbf{VV'} = \mathbf{I_c}$ (Identité d'ordre c), et la minimisation de *f=trace* (**E'E**) conduit à l'estimation: $\mathbf{W} = (\mathbf{X'X})^{-1}\mathbf{X'YV'}$, les colonnes de **V'** étant les vecteurs propres unitaires de la matrice **M** telle que [1] :

$$\mathbf{M} = \mathbf{Y'X}\,(\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Notons que si l'on travaille avec le critère *f=trace* ($\mathbf{E'(Y'Y)^{-1}E}$), avec la nouvelle contrainte sur **V** : $\mathbf{V(Y'Y)^{-1}V'} = \mathbf{I_c}$, le perceptron à une couche cachée réalise une analyse canonique du tableau à deux blocs (**Y**,**X**) (cf. chapitre 2). Si **Y** est un tableau disjonctif complet décrivant une partition des observations, les résultats du § 7.2. nous montrent que le perceptron réalise une analyse discriminante, résultat établi pour la première fois par Gallinari *et al.* (1988).

7.7.3 Modèles non-supervisés ou auto-organisés

Alors que les modèles supervisés (pour lesquels on dispose d'un échantillon d'apprentissage permettant d'estimer les paramètres) correspondent tout à fait à la démarche de la régression et de l'analyse discriminante, les modèles non-

[1] Cf. Lebart (1996). On notera la parenté de cette approche avec les *analyses partielles et projetées* (cf. chapitre 8) ou analyses sur variables instrumentales (cf. Rao,1964; Sabatier *et al.*, 1989).

supervisés ou auto-organisés sont le pendant des méthodes purement exploratoires. Dans les modèles non-supervisés dits d'auto-association, on ne connaît pas **y**, (il n'y a pas de "professeur") et on utilise l'artifice qui consiste à remplacer **y** par **x** (Baldi et Hornik, 1989). On désignera par **z** la valeur commune de **x** et **y**.

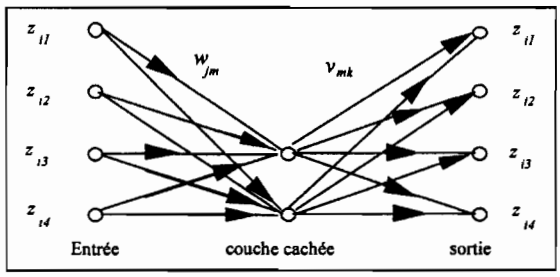


Figure 7.7 – 2. Réseau auto-associatif réalisant une compression du signal

Ceci semble une trivialité, et est effectivement une trivialité si la couche cachée possède autant d'éléments que **z** ($c = p$) et s'il n'y a pas de contraintes sur **A** = **WV** (auquel cas on a la solution évidente **A** = **I**). Mais si la couche cachée est notablement plus réduite que les couches d'entrée et de sortie, (c << p), elle forme un étranglement et le réseau réalise une *compression* du signal d'entrée.

On s'efforce donc de réduire le plus possible la déformation moyenne de **z** après intervention du réseau, qui n'est autre ici qu'une projection sur un sous-espace de dimension c inférieure à p. La solution est fournie par l'analyse en composantes principales du tableau **Z** (qui est aussi une décomposition aux valeurs singulières, puisque nous avons supposé les variables centrées) dont les n lignes sont les vecteurs **x**. Il suffit, pour le démontrer, de substituer **X** et **Y** par **Z** dans les formules du cas supervisé. Ainsi, par exemple, avec un seul neurone dans la couche cachée, la matrice **WV** est de rang *1*, ce qui conduira au premier axe de l'analyse en composantes principales de **Z**. L'*auto-organisation*, notion étudiée et formalisée par Kohonen (1989), qui est un des pionniers de l'approche neuronale, est donc rendue possible par la structure interne du réseau (cf. les cartes auto-organisées, chapitre 6, § 6.1.2). D'autres travaux sont relatifs aux algorithmes à lecture directe, comme l'algorithme de diagonalisation par approximation stochastique proposé par Benzécri (1969 b), antérieurement aux approches neuronales [1].

Ces algorithmes peuvent en effet être interprétés en terme d'apprentissage et d'auto-organisation. Un algorithme identique à une normalisation près a été proposé indépendamment par Oja et Karhunen (1981), puis amélioré par la suite par ces auteurs et d'autres neuromiméticiens. Ce domaine, qui a des applications potentielles importantes en compression d'image, a depuis été très

[1] On trouvera une étude plus numérique de la convergence de l'algorithme dans Lebart (1974), et le programme de calcul correspondant dans Lebart *et al.* (1977).

développé. Sur les liens entre réseaux neuronaux et analyse en composantes principales, cf. Oja (1982), Bourlard et Kamp (1988), Sirat (1991), Oja (1992). Lebart (1997) a montré que l'analyse des correspondances d'une table de contingence **K** pouvait être obtenue à partir de trois réseaux de neurones différents : un perceptron à une couche cachée supervisé (les matrices **X** et **Y** sont alors des tableaux disjonctifs complets, avec **K** = **Y'X**); un perceptron à une couche cachée non-supervisé (la matrice **Z** a pour terme général $z_{ij} = (k_{ij} - k_{i.}k_{.j})/\sqrt{k_{i.}k_{.j}}$); enfin comme un réseau linéaire adaptatif.

7.7.4 SVM : « Séparateurs à vastes marges » ou : « *Support Vector Machines* »

Cette technique de discrimination est essentiellement issue de la théorie de l'apprentissage (Learning theory) à qui est attaché principalement le nom de Vapnik (Vapnik, 1995) .

Il y a deux idées principales à la base des SVM[1] qui ont été développées principalement pour discriminer entre deux classes :

1) Dans l'espace $\mathcal{R}^p$ dont les p axes sont les prédicteurs (supposés continus dans un premier temps), chercher une cloison plane qui sépare au mieux les deux sous-populations pour l'échantillon d'apprentissage, mais au mieux « localement », autrement dit en accordant une importance particulière aux points frontaliers ou litigieux.
   a) Si les deux sous-populations à discriminer sont effectivement séparables dans l'espace, on cherchera un plan séparateur à vastes marges, c'est-à-dire un plan tel que les distances des points les plus proches de ce plan soient les plus grandes possibles.
   b) Sinon, on introduit des *pénalités* pour les points que l'on ne parviendrait pas à classer d'un même côté du plan.
2) En fait, la définition locale de la séparabilité des deux sous-populations est trop sévère pour une surface plane. On effectue des transformations non linéaires (par exemple on ajoute des fonctions polynomiales des variables de départ) qui ont pour effet d'augmenter la dimension p de l'espace de départ pour effectuer une séparation (qui elle, sera planaire) dans ce nouvel espace.

C'est la plupart du temps l'ensemble de ces deux opérations et des algorithmes (plus coûteux que ceux qui ont été présentés dans ce chapitre) qui leurs sont attachés que l'on désigne par SVM.

---

[1] Il semble que la « traduction/adaptation » du sigle SVM en Français (Séparateurs à vastes marges) soit due à Cornuéjols (2002). Cette nouvelle signification du sigle SVM est en fait plus adaptée et suggestive que l'originale. Cf. aussi : Cornuéjols et Miclet (2002).

Remarque :

Avant d'esquisser de façon un peu plus détaillée le fonctionnement des SVM, on peut remarquer qu'en discrimination classique, qu'il s'agisse d'analyse factorielle discriminante de Fisher, de segmentation, de régression logistique, il est toujours possible (quand le nombre d'observations le permet) d'augmenter la taille de l'espace de départ pour délinéariser le problème (en veillant à éviter un sur-apprentissage, à l'aide d'échantillons-test par exemple).

a_Hyperplan séparateur (rappels élémentaires)

Soit **u** un vecteur unitaire de $\mathcal{R}^p$. Comme précédemment, on désignera également par **u** la matrice colonne associée, et par **u'** sa transposée, et l'on exprimera que **u** est unitaire par la relation **u'u** = *1*. L'équation [7.7 - 4] est celle du plan passant par l'origine et orthogonal à **u**.

$$\mathbf{u'x} = 0 \qquad \left(\sum_j^p u_j x_j = 0 \right) \qquad [7.7 - 4]$$

L'équation, $\mathbf{u'x} + u_0 = 0$, qui fait intervenir la constante réelle u_0, est l'équation d'un plan orthogonal à **u**, situé à la distance $|u_0|$ de l'origine.

Cette distance est en effet la valeur absolue du produit scalaire de tout point **x** du plan avec **u**. Le produit scalaire <**u**, **x**> de tout point **x** de $\mathcal{R}^p$ par **u** s'écrit **u'x**, et donc, si **x** appartient au plan d'équation : $\mathbf{u'x} + u_0 = 0$, $< \mathbf{u}, \mathbf{x} > = - u_0$.

($u_0 < 0$ si le plan coupe la droite portée par **u** du même côté de l'origine que l'extrémité de **u**, et $u_0 > 0$ sinon).

La distance (signée) d'un point quelconque $\mathbf{x}_i$ de $\mathcal{R}^p$ au plan d'équation $\mathbf{u'x} + u_0 = 0$ est, de même, le produit scalaire d de **u** avec $(\mathbf{x}_i - \mathbf{x})$, x étant un point quelconque du plan, donc $d = \mathbf{u'}(\mathbf{x}_i - \mathbf{x})$, d'où $d = \mathbf{u'x}_i + u_0$. Notons que d change de signe lorsque l'on *traverse* le plan.

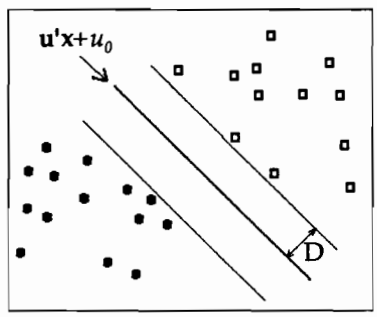


Figure 7.7.3. Plan séparateur pour deux groupes séparables

b_Cas de deux groupes séparables

Si l'on appelle **y** la variable binaire à prédire, et que l'on attribue à $\mathbf{y}_i$ les valeurs *-1* ou *+1* selon l'appartenance aux deux groupes, la fonction $d(i) = y_i (\mathbf{u}'\mathbf{x}_i + u_0)$ aura un signe constant, par exemple positif, si $(\mathbf{u}'\mathbf{x}_i + u_0) = 0$ est un plan séparant les deux groupes.

La recherche de ce plan revient à résoudre le problème d'optimisation suivant, en appelant D la demi-marge du séparateur.

Problème 1 :

> Chercher **u** et u_0 (c'est-à-dire le plan) tels que D soit maximum, avec les contraintes : $\mathbf{u}'\mathbf{u} = 1$, et, pour tout $i \leq n$: $y_i (\mathbf{u}'\mathbf{x}_i + u_0) \geq D$.

On peut se ramener à un problème plus classique en reprenant l'équation d'un plan sans la contrainte de normalisation : Si l'équation de l'hyperplan séparateur s'écrit $\mathbf{v}'\mathbf{x} + v_0$ sans contrainte sur le vecteur **v**, alors la distance d'un point $\mathbf{x}_i$ à ce plan s'écrit, en faisant intervenir la norme $\|\mathbf{v}\|$ de **v**. ($\|\mathbf{v}\| = (\mathbf{v}'\mathbf{v})^{1/2}$)

$$d(i) = (\mathbf{v}'\mathbf{x} + v_0) / \|\mathbf{v}\|$$

Les inégalités précédentes s'écrivent alors :

$$y_i (\mathbf{v}'\mathbf{x} + v_0) / \|\mathbf{v}\| \geq D,$$

ou encore : :

$$y_i (\mathbf{v}'\mathbf{x} + v_0) \geq D \|\mathbf{v}\|$$

La multiplication de **v** et de v_0 par une constante positive ne modifiant pas ces inégalités, on peut prendre $\|\mathbf{v}\| = 1/D$, et donc avoir à résoudre le système :

> Minimiser $\|\mathbf{v}\|$ avec les contraintes : $y_i (\mathbf{v}'\mathbf{x} + v_0) \geq 1$ pour tout $i \leq n$.

c_Cas de deux groupes non séparables

Dans ce cas, on convient de modifier le problème 1 de la façon suivante :

Problème 2 :

> Chercher **u** et u_0 (c'est-à-dire le plan) tels que D doit maximum, avec les contraintes : $\mathbf{u}'\mathbf{u} = 1$, et, pour tout $i \leq n$: $y_i (\mathbf{u}'\mathbf{x}_i + u_0) \geq D(1 - \varepsilon_i)$.
>
> Avec $\varepsilon_i \geq 0$ pour tout $i \leq n$, et $\sum_i \varepsilon_i \leq E$, E étant une constante positive caractérisant l'intensité de l'empiétement.

On remarque que si $\varepsilon_i \leq 1$, le point i ne respecte plus les marges, mais reste situé du bon côté du plan. Si $\varepsilon_i > 1$ alors le point est « mal classé ».

Le problème 2 peut donner lieu à la même reformulation que le problème 1 (suppression de la contrainte de normalisation). Il s'écrit alors, avec les mêmes notations :

Minimiser $\|v\|$ avec les contraintes : $y_i (\mathbf{v}'\mathbf{x} + v_0) \geq 1 - \varepsilon_i$ pour tout $i \leq n$

Et avec : $\varepsilon_i \geq 0$ pour tout $i \leq n$, et $\sum_i \varepsilon_i \leq E$

La fixation de la valeur de la constante E relève du doigté des praticiens des SVM : elle constitue un bilan global des empiétements autorisés au delà des marges, et des mauvais classements.

Pour la solution numérique de ce problème d'optimisation quadratique convexe, nous renvoyons à Vapnik (1995, 1998), Hastie et al. (2001), Cristianini et Shawe-Taylor (2000).

La solution a la forme : $\hat{\mathbf{v}} = \sum_{i=1}^{n} a_i y_i \mathbf{x}_i$

Dans cette formule, seuls les coefficients a_i correspondant aux contraintes effectivement atteintes sont non nuls. Autrement dit, l'hyperplan trouvé ne dépendra que des points situés dans son voisinage. Ces points sont les « support vectors ». Notons alors que la fonction d'affectation aux classes : $f(x) = \text{signe}(\mathbf{v}'\mathbf{x} + v_0)$ s'écrit :

$$f(x) = signe(\hat{\mathbf{v}}'\mathbf{x} + v_0) = signe(\sum_{i=1}^{n} a_i y_i \mathbf{x}_i'\mathbf{x} + v_0)$$

d_ Extension des descripteurs

Si l'on substitue au vecteur de descripteur $\mathbf{x}$ de $\mathcal{R}^p$ le vecteur $\mathbf{h(x)}$ de $\mathcal{R}^m$ (m pouvant être beaucoup plus grand que p), de façon à rendre plus linéairement séparables les deux groupes dans $\mathcal{R}^m$, la fonction d'affectation, notée maintenant g(x), s'exprime de façon analogue, en faisant intervenir le produit scalaire $\langle \mathbf{h}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}) \rangle$:

$$f(x) = signe(\hat{\mathbf{v}}'\mathbf{h}(\mathbf{x}) + v_0) = signe(\sum_{i=1}^{n} a_i y_i \langle \mathbf{h}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}) \rangle + v_0)$$

Or on peut construire des vecteurs $\mathbf{h(x)}$ dont les composantes sont des fonctions polynomiales de celles de $\mathbf{x}$, et dont le produit scalaire peut prendre une forme simple.

On peut même ne pas spécifier la transformation $\mathbf{h(x)}$ et n'utiliser que les produits scalaires, puisqu'ils sont les seuls à intervenir.

Cette propriété fondamentale des SVM fait qu'on peut manipuler facilement des espaces de représentation ayant de bonnes propriété en matière de séparation (sur l'échantillon d'apprentissage, à vérifier sur un échantillon-test). Le lecteur trouvera plus d'informations dans les ouvrages référencés ci-dessus.

7.7.5 Les modèles statistiques et les réseaux de neurones

On complétera cet aperçu par un résumé de l'intervention de Tibshirani lors d'une discussion sur la synthèse de Cheng et Titterington (1994, *op.cit.*). Cette intervention commence par une remarque sur la statistique et les réseaux de neurones, qui fait allusion aux *boîtes noires* que sont les réseaux de neurones.

> "Les statisticiens ont tendance à travailler avec des modèles plus interprétables car, pour eux, mesurer l'effet des variables est plus important que la prédiction".

Tibshirani répond ensuite à deux questions :

- ***Que peut apprendre un statisticien d'un neuro-miméticien ?***

1 "On devrait moins se soucier de l'optimalité statistique que de trouver des méthodes qui fonctionnent, spécialement sur les grands ensembles de données.

2 On devrait plus attaquer les problèmes réels auxquels se consacrent les neuro-miméticiens : reconnaissance de l'écriture et de la parole, prédiction des structures de l'ADN. Comme le dit John Tukey *: il vaut mieux avoir une solution approchée d'un problème réel que la solution exacte d'un problème trop simplifié.*

3 Les modèles à très nombreux paramètres peuvent être utiles pour la prédiction, spécialement pour les grands tableaux de données et les données bruitées.

4 Modéliser des combinaisons linéaires des variables d'entrées est très utile, car cela prend en compte des traits structurels et réduit la dimension.

5 Des algorithmes itératifs comme la descente de gradient (avec taux d'erreurs) peuvent éviter des ajustements trop complaisants.

6 Nous (statisticiens) devrions mieux nous vendre..."

- ***Que peut apprendre un neuromiméticien d'un statisticien ?***

1 "Il devrait plus s'intéresser à l'optimalité statistique, ou au moins, aux propriétés statistiques des méthodes.

2 Il devrait faire plus d'efforts pour comparer ses méthodes à des méthodes statistiques plus simples. Il serait alors surpris de voir que la régression fait souvent aussi bien qu'un perceptron multi-couches. Il ne devrait jamais utiliser un modèle compliqué alors qu'un modèle simple suffit."

Ces remarques n'épargnent pas les statisticiens, qui ont devant eux une profusion d'idées nouvelles et un vaste chantier ouvert. Ceux d'entre eux qui se consacrent à l'analyse exploratoire des grands tableaux se sentent cependant moins concernés par les deux premières critiques de Tibshirani.

Outre les trois articles de synthèse précités, on mentionnera, toujours pour un lectorat de statisticien : l'ouvrage de base de Hertz *et al.* (1991), l'article plus théorique de Amari (1990), sur les fondements mathématiques des méthodes. Mentionnons également l'article de Hornik (1994) décrivant, à l'intention des statisticiens, le perceptron multicouche et les algorithmes d'analyses en composantes principales par apprentissage comme deux intersections

importantes entre les deux disciplines. En Français, on consultera les ouvrages généralistes de Bourret *et al.* (1991) de Milgram (1993), Hérault et Jutten (1994), Blayo et Verleysen (1996). Pour des présentations faisant le lien avec l'approche "analyse des données", on se référera à Lelu (1991), à Chabanon et Dubuisson (1991) et à l'ouvrage collectif "Statistique et méthodes neuronales", édité par Thiria *et al.* (1997). On trouvera dans ce dernier ouvrage divers points de vue et exposés sur la théorie de l'apprentissage (*statistical learning*) développée autour des travaux, déjà cités, de Vapnik (1995, 1998) et présentée dans Hastie *et al.* (2001).

7.8 Annexe technique du chapitre 7

Nous allons voir dans cette annexe que la théorie de l'information de Shannon-Wiener, telle qu'elle a été énoncée dans le contexte de la statistique par Kullback (1959) dans un ouvrage fondamental et trop peu connu, introduit naturellement les distances intervenant en analyse linéaire discriminante. Cette théorie montre aussi l'insuffisance des taux d'inertie comme mesure du degré de "non-sphéricité" d'un nuage.

7.8.1 Distances entre distributions

On utilisera pour le calcul la notion de *divergence* de Jeffreys (1946), qui permet de mesurer la distance entre deux hypothèses H_1 et H_2 dans le cas d'une réalisation d'un vecteur $\mathbf{x}$ issu de l'un des deux schémas relatifs à des lois normales dans $\mathbb{R}^p$:

(H_1) *Hypothèse d'indépendance :*

$$\begin{cases} \text{Moyenne théorique} = \mu_1 \\ \text{Matrice des covariances théorique} = \sigma^2 \mathbf{I} \end{cases}$$

(H_2) *Cas général :*

$$\begin{cases} \text{Moyenne théorique} = \mu_2 \\ \text{Matrice des covariances théoriques} = \boldsymbol{\Sigma} \quad \text{(supposée ici régulière)} \end{cases}$$

La divergence va permettre d'exprimer la distance entre les hypothèses H_1 et H_2 en fonction des valeurs propres de $\boldsymbol{\Sigma}$ et l'on s'apercevra qu'elle met en cause les petites valeurs propres alors que l'analyse en composantes principales ne retient que les grandes.

Rappelons que l'on définit, pour deux hypothèses H_1 et H_2 pouvant donner lieu à la réalisation d'un vecteur $\mathbf{x}$, la divergence $J(H_1, H_2)$ comme la différence :

$$J(H_1,H_2) = \int log \frac{P(H_1|\mathbf{x})}{P(H_2|\mathbf{x})}\, dv_1(\mathbf{x}) - \int log \frac{P(H_2|\mathbf{x})}{P(H_1|\mathbf{x})}\, dv_2(\mathbf{x})$$

v_1 et v_2 étant les mesures associées aux hypothèses H_1 et H_2 ; et $P(H_i|\mathbf{x})$ ($i = 1, 2$) étant la probabilité conditionnelle que H_i soit vraie connaissant $\mathbf{x}$.

Dans le cas de densités continues $f_1(\mathbf{x})$ et $f_2(\mathbf{x})$, on a :

$$J(H_1,H_2) = \int (f_1(\mathbf{x}) - f_2(\mathbf{x})) log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x}$$

La densité de probabilité du vecteur $\mathbf{x}$ s'écrit, pour une matrice des covariances théoriques Σ_i et un vecteur μ_i :

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma_i|}} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)'\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\}$$

d'où :

$$log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{1}{2} log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} tr\left(\Sigma_1^{-1}(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)'\right) + \frac{1}{2} tr\left(\Sigma_2^{-1}(\mathbf{x}-\mu_2)(\mathbf{x}-\mu_2)'\right)$$

Remplaçant cette valeur dans la formule donnant $J(H_1, H_2)$, on voit que le premier terme de $J(H_1, H_2)$ n'est autre que $I_{(1;2)}$, l'information moyenne apportée par l'échantillon $\mathbf{x}$ sous l'hypothèse H_1, en vue de discriminer en faveur de H_1 contre H_2 (cf. Kullback, 1959).

On écrira ce premier terme en posant : $\mathbf{x} - \mu_2 = \mathbf{x} - \mu_1 + \mu_1 - \mu_2$. Il vient :

$$\begin{aligned} I_{(1;2)} &= \int f_1(\mathbf{x}) log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} \\ &= \frac{1}{2} log \frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2} tr\left(\Sigma_1(\Sigma_2^{-1} - \Sigma_1^{-1})\right) + \frac{1}{2} tr\left(\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)'\right) \end{aligned}$$

et $J(H_1,H_2)$ s'écrit donc :

$$\begin{aligned} J(H_1,H_2) &= I_{(1;2)} + I_{(2;1)} \\ &= \frac{1}{2} tr\left((\Sigma_1 - \Sigma_2)(\Sigma_2^{-1} - \Sigma_1^{-1})\right) + \frac{1}{2} tr\left((\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1-\mu_2)(\mu_1-\mu_2)'\right) \end{aligned}$$

7.8.2 Distance de Mahalanobis et information

Dans le cas où les moyennes diffèrent ($\mu_1 \neq \mu_2$) et où les matrices de covariances théoriques sont égales : $\Sigma_1 = \Sigma_2$, la formule précédente montre que la divergence de Jeffrey $J(H_1, H_2)$ est proportionnelle à la *distance de Mahalanobis*, ou distance généralisée entre les deux populations théoriques 1 et 2 (cf. § 7.2.4).

On s'intéresse au cas pour lequel :

$$\mu_1 = \mu_2, \quad \Sigma_1 = \mathbf{I} \quad \text{et} \quad \Sigma_2 = \Sigma.$$

On notera en abrégé $J(H_1, H_2) = J(\mathbf{I}, \Sigma)$ avec :

$$J(\mathbf{I}, \Sigma) = \frac{1}{2} tr\left((\mathbf{I} - \Sigma)(\Sigma^{-1} - \mathbf{I})\right) = \frac{1}{2} tr(\Sigma + \Sigma^{-1}) - p$$

Soit, en faisant apparaître les valeurs propres λ_α de Σ :

$$J(\mathbf{I}, \Sigma) = \frac{1}{2}\left(\sum_{\alpha=1}^{p} \lambda_\alpha + \sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha}\right) - p$$

Si les inerties totales théoriques sont égales sous les hypothèses H_1 et H_2, on a la relation :

$$\sum_{\alpha=1}^{p} \lambda_\alpha = p$$

Le seul terme variable dans $J(\mathbf{I}, \Sigma)$ est donc le terme :

$$\sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha}$$

On voit que la divergence entre les deux hypothèses sera particulièrement grande dans le cas où certaines valeurs propres de Σ sont voisines de *0*.

Dans le cadre de cette formalisation de la théorie de l'information, une valeur propre de Σ infiniment petite jouera un rôle beaucoup plus déterminant que deux valeurs propres expliquant, par exemple, *80%* de l'inertie totale, alors que c'est dans le sous-espace des deux facteurs correspondants que l'on observera les principaux traits structuraux.

En fait, comme un *filtre* dans un processus de communication, la représentation des données dans l'espace des premiers axes factoriels a pour effet d'augmenter la *valeur pratique de l'information* au prix d'une perte d'information brute qui peut être considérable. Or cette notion de valeur pratique (Brillouin, 1959) est étrangère à la théorie classique de l'information. Comme le suggère Thom (1974), on gagnerait souvent à remplacer le mot *information* par le mot *forme* (ici à peu près équivalent au mot anglais *pattern*) lors d'un processus d'observation. Ceci est encore plus vrai dans le cas de ces observations particulières que sont les *visualisations exploratoires de données* qui occupent une place centrale dans cet ouvrage.

Chapitre 8

Analyses de données structurées

Les méthodes d'analyse de données structurées, présentées dans ce chapitre occupent une position intermédiaire entre les outils purement exploratoires des chapitres 3 à 6 (axes principaux et classification) et les méthodes à vocation plus explicative présentées dans les chapitres 2 et 7 (régression et discrimination).

Les méthodes exploratoires telles que nous les présentons posent un modèle très général qui distingue, pour chaque application, deux familles d'éléments : les éléments actifs (variables ou individus, ligne ou colonnes) qui servent à établir des espaces de visualisation complétés par des classifications, et les éléments supplémentaires, qui jouent un rôle passif, et interviennent *a posteriori* pour illustrer, identifier, caractériser les représentations obtenues à partir des éléments actifs.

En général, le tableau des éléments actifs est amorphe et homogène : il ne doit pas exister de structure *a priori* (dépendance fonctionnelle, relations comptables, etc.) entre les variables et les individus, et les distances entre éléments doivent avoir un sens pour l'utilisateur.

Or, il est fréquent que le tableau des données actives soit déjà structuré. C'est le cas par exemple des données géographiques ou temporelles où la structure intervient au niveau des observations (individus voisins ou consécutifs). Il peut exister des groupes d'individus ou des groupes de variables connus *a priori*. Le tableau peut ne pas se ramener de façon univoque à la forme rectangulaire (tables de contingences multiples, séries chronologiques de tableaux).

Il est souvent possible d'aborder ces problèmes dans le cadre du modèle exploratoire de base, mais la tentation est forte, dans le cas où les applications se présentent de façon répétitive, de proposer des variantes adaptées aux types de tableaux ou de structures rencontrés. Il reste que l'on doit envisager une économie de l'analyse des données, en ce sens que la panoplie des méthodes disponibles ne peut s'accroître indéfiniment, sous peine de voir le rendement

de ces méthodes décroître. Faut-il, pour un utilisateur dont la recherche statistique n'est pas l'activité principale, investir dans une méthode complexe qui ne servira qu'une fois? Vaut-il mieux utiliser une méthode de description un peu grossière, mais parfaitement dominée conceptuellement en raison d'expériences accumulées, qu'une méthode plus subtile dont les résultats laissent perplexes ? Le temps disponible, les possibilités de formation, les budgets d'acquisition de logiciels ne sont pas des ressources inépuisables.

A propos des méthodes de classification pour lesquelles il estime le nombre de publications à près de mille par an, Cormack (1971) remarquait déjà que "lorsque la technique (de classification) échoue, la réaction de l'auteur est de modifier la technique, au lieu d'utiliser une technique plus *standard* ou de remettre en question tout le traitement".

Cette attitude comporte un certain danger. Si la panoplie des techniques est très étendue, le risque d'adéquation purement accidentelle de la technique aux données est augmenté. Ce problème est récurrent lorsqu'il s'agit d'articuler exploration et inférence, et se rapproche du problème plus classique des comparaisons multiples, déjà évoqué à propos de la description des axes et des classes par les valeurs-test (cf. § 5.4.1 du chapitre 5 et § 6.3.2 du chapitre 6).

Un défi auquel est confrontée la statistique multidimensionnelle est précisément la gestion de cette diversification, nécessaire pour la recherche, mais source de difficultés au niveau des applications en vraie grandeur.

Précisons, dans ce contexte méthodologique, quelles sont les méthodes d'analyses de données structurées qui feront l'objet de ce chapitre.

- Les méthodes d'*analyses partielles ou projetées* (section 8.1) concernent les situations pour lesquelles les individus ou observations (lignes d'un tableau **X** d'ordre (n, p)) peuvent être décrits par p variables (colonnes de **X**) mais peuvent aussi être dépendants de q variables : colonnes d'un tableau **Z** d'ordre (n, q) dont on désirerait, dans la mesure du possible, soit prendre en compte, soit éliminer l'effet.

- Les techniques d'*analyses locales*, mettant en jeu des *structures de graphes* (section 8.2) sont appropriées lorsqu'il existe des informations *a priori* ou externes sur les couples d'individus ou d'observations (existence d'une relation binaire symétrique ou structure de graphe non orienté décrivant des proximités temporelles ou géographiques).

- Enfin les méthodes de traitement de *tableaux multiples* ou de *groupes de variables* (section 8.3), qui correspondent à une famille quasi-illimitée de techniques, seront évoquées au travers d'une sélection des approches qui nous paraissent les plus utiles en pratique : analyse procrustéenne, méthode STATIS, analyse factorielle multiple, analyse canonique généralisée.

8.1 Analyses partielles et projetées

Ces méthodes se proposent d'analyser les associations existant entre des variables et des individus, non seulement après élimination d'effets de niveaux ou d'échelle (cas de l'opération de standardisation : centrage et division par l'écart-type), mais également après avoir tenu compte de l'influence éventuelle de "variables exogènes". A l'origine et au centre de ces techniques se trouve l'analyse en composantes principales partielle ou sur variables instrumentales selon la terminologie de Rao (1964).

8.1.1 Définition du coefficient de corrélation partielle

Deux variables aléatoires X_1 et X_2 sont supposées dépendre d'une même variable aléatoire Z. On dispose d'un échantillon de chacune de ces variables. On peut mesurer directement le coefficient de corrélation $r(\mathbf{x}_1, \mathbf{x}_2)$ sur deux échantillons de taille n représentés dans $\mathcal{R}^n$ par les vecteurs à n composantes $\mathbf{x}_1$ et $\mathbf{x}_2$. Mais nous voulons en fait connaître la liaison existant entre $\mathbf{x}_1$ et $\mathbf{x}_2$ en éliminant l'effet de la variable Z dont les n observations sont les composantes du vecteur $\mathbf{z}$.

Pour prendre un exemple élémentaire classique[1], X_1 est la plus grande dimension d'un œuf, X_2 la plus petite et Z son poids. Sur un échantillon de $n = 100$ œufs, on trouvera un coefficient $r(\mathbf{x}_1, \mathbf{x}_2)$ fortement positif, car il existe de gros œufs, pour lesquels X_1 et X_2 ont des valeurs élevées, et des petits, pour lesquels ces valeurs sont faibles. Par contre, si le poids Z est fixé, la liaison observée sera inverse car, à poids égal, les œufs sont plus ou moins sphériques.

Comment mesurer cette liaison entre X_1 et X_2 "*à Z constant*"? Une première méthode consiste à regrouper les observations en classes à l'intérieur desquelles les valeurs de Z sont peu différentes. On calcule alors le coefficient de corrélation entre X_1 et X_2 dans chaque classe et l'on fait, par exemple, une moyenne pondérée de ces coefficients, pour avoir une idée d'ensemble de la liaison. Cette méthode est excellente et il est conseillé de l'employer chaque fois que la taille des échantillons permet une division en classes d'effectifs suffisants.

Une autre méthode va nous permettre de calculer la liaison entre X_1 et X_2 "*à Z constant*" de façon simple, même lorsque les échantillons sont petits (mais au prix d'une hypothèse sur la linéarité des liaisons). Ce coefficient de corrélation entre X_1 et X_2 "*à Z constant*" s'appellera le coefficient de *corrélation partielle* entre

---

[1] Cf. Darmois (1957).

X_1 et X_2, et on le notera $\rho(X_1, X_2|Z)$. Son calcul repose sur l'hypothèse que l'effet de Z sur les variables X_1 et X_2 se manifeste par des relations du type [1] :

$$\begin{cases} X_1 = \alpha_1 Z + \varepsilon_1 \\ X_2 = \alpha_2 Z + \varepsilon_2 \end{cases}$$

Une fois ôtée l'influence de la variable Z, les variables aléatoires X_1 et X_2 deviennent $X_1 - \alpha_1 Z = \varepsilon_1$ et $X_2 - \alpha_2 Z = \varepsilon_2$. Le coefficient de *corrélation partielle* théorique $\rho(X_1, X_2|Z)$ est par définition le coefficient de corrélation usuel entre ε_1 et ε_2 :

$$\rho(X_1, X_2|Z) = \frac{cov(\varepsilon_1, \varepsilon_2)}{\sqrt{var(\varepsilon_1) var(\varepsilon_2)}}$$

On définit de façon analogue une matrice des covariances partielles $\mathbf{V}(\mathbf{X}|\mathbf{Z})$ et une matrice des corrélations partielles $\mathbf{C}(\mathbf{X}|\mathbf{Z})$ entre p variables $X_1, X_2, \ldots, X_p$, lorsque q variables $Z_1, Z_2, \ldots, Z_q$ sont supposées fixées. On a alors le système suivant :

$$\begin{cases} X_1 = \alpha_{11} Z_1 + \alpha_{12} Z_2 + \ldots + \alpha_{1q} Z_q + \varepsilon_1 \\ X_2 = \alpha_{21} Z_1 + \alpha_{22} Z_2 + \ldots + \alpha_{2q} Z_q + \varepsilon_2 \\ \qquad\qquad \ldots\ldots \\ X_p = \alpha_{p1} Z_1 + \alpha_{p2} Z_2 + \ldots + \alpha_{pq} Z_q + \varepsilon_p \end{cases}$$

$\mathbf{V}(\mathbf{X}|\mathbf{Z})$ et $\mathbf{C}(\mathbf{X}|\mathbf{Z})$ sont respectivement les matrices des covariances et des corrélations théoriques entre les variables résiduelles : $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$.

8.1.2 Calcul des covariances et corrélations partielles

a – Cas de deux variables

Pour les n observations des trois variables X_1, X_2, Z, qui sont les composantes, supposées ici centrées, des 3 vecteurs $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{z}$, ces relations d'ajustement s'écrivent, avec les notations de la section 3.2 (mais la lettre $\mathbf{x}$ désigne maintenant des variables endogènes ou expliquées) :

$$\begin{cases} \mathbf{x}_1 = a_1 \mathbf{z} + e_1 \\ \mathbf{x}_2 = a_2 \mathbf{z} + e_2 \end{cases}$$

où a_1 et a_2 sont respectivement les estimations de α_1 et α_2 par la méthode des moindres carrés alors que e_1 et e_2 sont les résidus observés.

---

[1] Comme pour tout modèle linéaire, les variables entre lesquelles existe une relation linéaire peuvent être des variables transformées construites à partir des variables réellement observées. Le caractère linéaire de la relation n'est donc pas une contrainte importante.

La covariance partielle expérimentale s'écrit :

$$Cov(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z}) = \frac{1}{n}\mathbf{e}_1'\mathbf{e}_2 = \frac{1}{n}(\mathbf{x}_1 - a_1\mathbf{z})'(\mathbf{x}_2 - a_2\mathbf{z})$$

soit :

$$Cov(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z}) = \frac{1}{n}\{\mathbf{x}_1'\mathbf{x}_2 - a_1\mathbf{z}'\mathbf{x}_2 - a_2\mathbf{x}_1'\mathbf{z} + a_1a_2\,\mathbf{z}'\mathbf{z}\}$$

On remplace les coefficients de régression par leur valeur $a_1 = \mathbf{x}_1\mathbf{z}/\mathbf{z}'\mathbf{z}$ et $a_2 = \mathbf{x}_2\mathbf{z}/\mathbf{z}'\mathbf{z}$ et l'on obtient après simplification :

$$Cov(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z}) = \frac{1}{n}\left\{\mathbf{x}_1'\mathbf{x}_2 - \frac{(\mathbf{x}_1'\mathbf{z})(\mathbf{x}_2'\mathbf{z})}{\mathbf{z}'\mathbf{z}}\right\}$$

expression que l'on peut écrire:

$$Cov(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z}) = Cov(\mathbf{x}_1, \mathbf{x}_2) - \frac{Cov(\mathbf{x}_1, \mathbf{z})Cov(\mathbf{x}_2, \mathbf{z})}{Var(\mathbf{z})} \qquad [8.1 - 1]$$

Les variances résiduelles se calculent de façon analogue et l'on a pour $\mathbf{e}_1$:

$$\frac{1}{n}\mathbf{e}_1'\mathbf{e}_1 = Var(\mathbf{x}_1) - \frac{Var^2(\mathbf{x}_1, \mathbf{z})}{Var(\mathbf{z})} = \left(1 - r^2(\mathbf{x}_1, \mathbf{z})\right)Var(\mathbf{x}_1)$$

Le coefficient de corrélation partielle $r(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z})$ s'écrit alors, en faisant apparaître les coefficients de corrélation usuels :

$$r(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{z}) = \frac{r(\mathbf{x}_1, \mathbf{x}_2) - r(\mathbf{x}_1, \mathbf{z})r(\mathbf{x}_2, \mathbf{z})}{\sqrt{\left(1 - r^2(\mathbf{x}_1, \mathbf{z})\right)\left(1 - r^2(\mathbf{x}_2, \mathbf{z})\right)}}$$

b – Cas de *p* variables (X) et de *q* variables (Z)

Nous disposons maintenant de p vecteurs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ auxquels correspondent p points dans $\mathcal{R}^n$. On peut mesurer la covariance (ou la corrélation) entre ces variables après élimination de l'effet de q autres variables représentées dans $\mathcal{R}^n$ par les vecteurs $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q$. On désignera par $\mathbf{X}$ la matrice (n, p) et par $\mathbf{Z}$ la matrice (n, q) qui rassemblent en colonne ces divers vecteurs.

Pour la $k^{ième}$ variable, l'ajustement des moindres carrés entre $\mathbf{x}_k$ et les variables exogènes $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q$ s'écrit :

$$\mathbf{x}_k = a_{k1}\,\mathbf{z}_1 + a_{k2}\,\mathbf{z}_2 + \ldots + a_{kq}\,\mathbf{z}_q + \mathbf{e}_k$$

où $\mathbf{e}_k$ est le vecteur résiduel. Nous appellerons $\mathbf{a}_k$ le vecteur-colonne de ces q coefficients. Après avoir effectué les p ajustements similaires concernant $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ nous rassemblons dans la matrice $\mathbf{A}$ de dimension (q, p) les p vecteurs-colonnes de coefficients $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)$ et dans la matrice $\mathbf{E}$ de dimension (n, p) les p vecteurs résiduels $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$. Le système des ajustements s'écrit alors de façon synthétique :

$$\underset{(n,p)}{\mathbf{X}} = \underset{(n,q)}{\mathbf{Z}}\,\underset{(q,p)}{\mathbf{A}} + \underset{(n,p)}{\mathbf{E}}$$

Dans la matrice **A**, la $k^{ième}$ colonne est :

$$\mathbf{a}_k = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_k .$$

Il est donc possible d'écrire **A** sous la forme :

$$\mathbf{A} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \qquad [8.1 - 2]$$

Avec ces notations, la matrice (p, p) qui définit les covariances partielles expérimentales sur les **X** "*à Z constant*" s'écrira :

$$\mathbf{V}(\mathbf{X}|\mathbf{Z}) = \frac{1}{n}\mathbf{E}'\mathbf{E} = \frac{1}{n}(\mathbf{X} - \mathbf{Z}\mathbf{A})'(\mathbf{X} - \mathbf{Z}\mathbf{A})$$

$$= \frac{1}{n}(\mathbf{X}'\mathbf{X} - \mathbf{A}'\mathbf{Z}'\mathbf{X} - \mathbf{X}'\mathbf{Z}\mathbf{A} + \mathbf{A}'\mathbf{Z}'\mathbf{Z}\mathbf{A})$$

En remplaçant **A** par son expression [8.1 - 2] et après simplification :

$$\mathbf{V}(\mathbf{X}|\mathbf{Z}) = \frac{1}{n}\{\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\} \qquad [8.1 - 3]$$

Imaginons que soient rassemblés dans un tableau **T** à n lignes et $p + q$ colonnes les tableaux centrés **X** et **Z** :

$$\mathbf{T} = [\mathbf{X}, \mathbf{Z}]$$

Alors la matrice des covariances entre les colonnes de **T** peut être partitionnée en quatre sous-matrices de covariances :

$$\mathbf{V}(\mathbf{T}) = \begin{bmatrix} \mathbf{V}_{xx} & \mathbf{V}_{zx} \\ \mathbf{V}_{xz} & \mathbf{V}_{zz} \end{bmatrix}$$

avec :

$$\mathbf{V}_{xx} = \frac{1}{n}\mathbf{X}'\mathbf{X} \qquad \mathbf{V}_{zz} = \frac{1}{n}\mathbf{Z}'\mathbf{Z} \qquad \mathbf{V}_{zx} = \mathbf{V}'_{xz} = \frac{1}{n}\mathbf{Z}'\mathbf{X}$$

Alors la relation [8.1 - 3] prend la forme :

$$\mathbf{V}(\mathbf{X}|\mathbf{Z}) = \mathbf{V}_{xx} - \mathbf{V}_{zx}\mathbf{V}_{zz}^{-1}\mathbf{V}_{xz} \qquad [8.1 - 4]$$

où elle apparaît comme une généralisation, pour $q \geq 1$, de [8.1 -1].

La matrice des *corrélations partielles* se calcule aisément à partir de la matrice des covariances partielles $\mathbf{V}(\mathbf{X}|\mathbf{Z})$ comme une matrice des corrélations ordinaires à partir d'une matrice des covariances.

8.1.3 Analyse du nuage résiduel ou analyse partielle

L'analyse du tableau **X** lorsque les variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q$ sont fixées, se ramène donc à *l'analyse générale* (cf. chapitre 1) du tableau des écarts **E**. Ainsi les points représentant les variables dans $\mathcal{R}^n$ auront pour coordonnées (à une homothétie

près), sur l'axe factoriel α, les composantes du $\alpha^{\text{ième}}$ vecteur-propre $\mathbf{u}_\alpha$ de la *matrice des covariances partielles* (pour une analyse *normée*, on utiliserait la matrice des *corrélations partielles* qui s'en déduit) :

$$\mathbf{V}(\mathbf{X}|\mathbf{Z}) = \frac{1}{n}\mathbf{E}'\mathbf{E}$$

Poursuivant l'interprétation géométrique de l'ajustement des moindres carrés, on peut remarquer que :

$$n\mathbf{V}(\mathbf{X}|\mathbf{Z}) = \mathbf{X}'\left(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\mathbf{X} = \mathbf{X}'\left(\mathbf{I} - \mathbf{P}_\mathbf{Z}\right)\mathbf{X} = \mathbf{X}'\mathbf{P}_\mathbf{Z}^\star\mathbf{X}$$

où $\mathbf{P}_\mathbf{Z}^\star$ $(= \mathbf{I} - \mathbf{P}_\mathbf{Z})$ est une matrice (n, n) symétrique et idempotente, analogue à la matrice $\mathbf{P}_\mathbf{X}$ définie par la formule [2.2 - 3] au paragraphe 2.2.2. Ici $\mathbf{P}_\mathbf{Z}^\star$ effectue la projection de tout vecteur de $\mathcal{R}^n$ sur le sous-espace à $(n-q-1)$ dimensions, orthogonal au sous-espace engendré par $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q)$. C'est cette projection que l'on analyse lorsqu'on opère la transformation des données $\mathbf{E} = \mathbf{P}_\mathbf{Z}^\star\mathbf{X}$.

Ainsi, dans l'hypothèse où les régressions traduisent effectivement l'effet des variables que l'on désire fixer, il est possible d'étudier *a posteriori* les liaisons et les associations existant entre des variables et des observations, *"toutes choses égales par ailleurs"*.

Dans certains cas, on peut au contraire (cf. paragraphe 8.1.4) être intéressé par la projection du nuage sur le sous-espace engendré par $\mathbf{Z}$, le tableau analysé étant alors le tableau $\mathbf{F} = \mathbf{P}_\mathbf{Z}\mathbf{X}$. On réservera le nom d'*analyse projetée* à l'analyse de $\mathbf{F}$.

8.1.4 Autres analyses partielles ou projetées

Il existe plusieurs variantes de méthodes impliquant des projections sur des sous-espaces. Une vue générale ainsi que des extensions de ce type d'approche sont données par Sabatier (1984, 1987).

On a vu que l'analyse canonique (section 2.1) part d'une situation analogue, c'est-à-dire d'un tableau de la forme $\mathbf{R} = (\mathbf{X}, \mathbf{Z})$, mais cherche le plus petit angle entre les sous-espaces engendrés par les colonnes de $\mathbf{X}$ et de $\mathbf{Z}$ dans $\mathcal{R}^n$. Ceci a conduit à diagonaliser une matrice du type :

$$\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}$$

où $\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X} = (\mathbf{P}_\mathbf{Z}\mathbf{X})'\mathbf{P}_\mathbf{Z}\mathbf{X}$ est proportionnel à la matrice d'inertie du nuage projeté sur le sous-espace engendré par les colonnes de $\mathbf{Z}$.

Dans l'équation $\mathbf{Su} = \lambda\mathbf{u}$, posons $\mathbf{u} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}$. On obtient :

$$\mathbf{X}'\mathbf{P}_Z\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v} = \lambda\mathbf{v}$$

ce qui montre que $\mathbf{v}$ est bien un axe principal de l'analyse du nuage projeté[1] avec la métrique $(\mathbf{X}'\mathbf{X})^{-1}$ (cf. § 1.3).

On a également vu que l'analyse discriminante est un cas particulier de ce type d'analyse lorsque $\mathbf{Z}$ est le tableau de codage disjonctif d'une variable nominale.

a – Analyse canonique des correspondances

Une technique voisine, qui aurait pu avoir sa place dans les sections consacrées à l'analyse canonique ou à l'analyse discriminante, est l'analyse canonique des correspondances, proposée par Ter Braak (1986, 1987), étudiée et appliquée par Chessel *et al.* (1987), Lebreton *et al.* (1988), et étendue par Ter Braak (1988) à l'analyse canonique partielle des correspondances.

On a fait allusion au début de ce chapitre aux dangers d'une prolifération indéfinie de méthodes spécifiques, en reconnaissant cependant que si des situations typiques ou des structures typiques de tableaux se présentent avec une certaine fréquence, il est loisible de forger des instruments *ad hoc*.

En écologie précisément, les observations se présentent souvent sous la forme d'un tableau $\mathbf{R} = (\mathbf{X}, \mathbf{Z})$ où, pour n sites (lignes de $\mathbf{X}$ et de $\mathbf{Z}$), on dispose d'un tableau numérique $\mathbf{X}$ (qui peut aussi être une autre table de contingence ou un tableau disjonctif complet) décrivant les sites (variables géologiques, climatiques, pétrochimiques, botaniques, etc.) et d'une table de contingence (ou parfois de présence-absence) $\mathbf{Z}$ donnant le nombre ou la présence de q espèces animales ou végétales sur les n sites.

Si l'on appelle $\mathbf{D}_n$ et $\mathbf{D}_q$ les matrices diagonales d'ordres (n, n), et (q, q) contenant les marges de la table $\mathbf{Z}$, on munira les n lignes de $\mathbf{X}$ de masses proportionnelles à la diagonale de $\mathbf{D}_n$ (en particulier pour centrer les p colonnes de $\mathbf{X}$). On notera encore $\mathbf{X}$ dans la suite la matrice centrée de cette façon. L'analyse canonique des correspondances revient à diagonaliser :

$$\mathbf{S} = (\mathbf{X}'\mathbf{D}_n\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}\mathbf{D}_q^{-1})\mathbf{D}_q(\mathbf{D}_q^{-1}\mathbf{Z}'\mathbf{X})$$

Si le tableau $\mathbf{Z}$ est un tableau disjonctif complet (une seule espèce et un seul spécimen par site), $\mathbf{Z}'\mathbf{Z} = \mathbf{D}_q$ et la matrice $\mathbf{D}_n$ est une matrice scalaire ; l'analyse canonique des correspondances est alors simplement l'analyse discriminante visant à prédire les espèces à partir des caractéristiques des sites [2].

---

[1] On vérifie que $\mathbf{v}$ est bien de norme *1* pour la métrique $(\mathbf{X}'\mathbf{X})^{-1}$ puisque $\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} = 1$.

[2] Comme le remarquent Lebreton *et al.* (1988), on peut se ramener aux calculs d'une analyse discriminante dans le cas général en multipliant les lignes de $\mathbf{Z}$ de façon à ne laisser qu'un spécimen d'une seule espèce par ligne et en répétant de façon similaire les lignes de $\mathbf{X}$. Cette dilatation de Z supprime les cooccurrences d'espèces à l'intérieur d'un même site.

La matrice $\mathbf{A} = \mathbf{X}'\mathbf{Z}\mathbf{D}_q^{-1}$ d'ordre (p, q) contient les moyennes des variables par espèces.

Comme on vient de le voir à propos de l'analyse canonique, il s'agit ici d'une analyse en axes principaux de $\mathbf{A}$ dans la métrique définie par $(\mathbf{X}'\mathbf{D}_n\mathbf{X})^{-1}$, inverse de la matrice des covariances totales pondérées des variables-colonnes de $\mathbf{X}$.

Réécrivons une matrice du type de $\mathbf{S}$ dans le cas où $\mathbf{D}_n$ est une matrice scalaire (nombre constant d'espèces par site) et en posant $\mathbf{Y} = \mathbf{Z}\mathbf{D}_q^{-1/2}$:

$$\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}\,\mathbf{Y}'\mathbf{X})$$

Remarquons que si le vecteur $\mathbf{u}$ est vecteur propre de $\mathbf{S}$ relatif à la valeur propre λ, alors :

$$\mathbf{v} = \mathbf{Y}'\mathbf{X}\mathbf{u}$$

est vecteur propre de :

$$\mathbf{S}_1 = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\,\mathbf{P}_X\,\mathbf{Y} = (\mathbf{P}_X\mathbf{Y})'(\mathbf{P}_X\,\mathbf{Y})$$

relatif à la même valeur propre λ.

Or $\mathbf{S}_1$ correspond à l'analyse en axes principaux de la projection de la table de contingence (normalisée) $\mathbf{Y}$ sur le sous-espace engendré par les colonnes du tableau $\mathbf{X}$ dans l'espace $\mathcal{R}^n$.

L'analyse canonique des correspondances peut donc être considérée comme une analyse partielle particulière. Elle diffère de l'analyse canonique en ce sens qu'elle traite de façon dissymétrique les deux tableaux $\mathbf{X}$ et $\mathbf{Z}$ (elle ne fait jamais intervenir la matrice $(\mathbf{Z}'\mathbf{Z})^{-1}$, c'est-à-dire finalement la structure interne du tableau $\mathbf{Z}$, indépendamment de $\mathbf{X}$).

b – Analyse non-symétrique des correspondances

On a vu plus haut que, en présence d'un tableau de données $\mathbf{R} = (\mathbf{X}, \mathbf{Z})$, comprenant deux groupes de variables, l'analyse canonique conduisait à diagonaliser la matrice :

$$\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

alors que l'analyse du nuage des lignes de $\mathbf{X}$ projeté sur le sous-espace engendré par les colonnes de $\mathbf{Z}$ conduit à diagonaliser :

$$\mathbf{S}_1 = \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'\,\mathbf{P}_Z\,\mathbf{X}$$

Si les matrices $\mathbf{X}$ et $\mathbf{Z}$ sont des tableaux disjonctifs complets, la diagonalisation de $\mathbf{S}$ est celle impliquée dans l'analyse des correspondances de la table de contingence $\mathbf{C} = \mathbf{X}'\mathbf{Z}$.

La diagonalisation de $\mathbf{S}_1$ correspond (à un centrage près) à l'*analyse non-symétrique des correspondances* de cette même table **C**, introduite et développée par Lauro et D'Ambra (1984) pour traiter les situations où les variables lignes et colonnes jouent des rôles dissymétriques.

Cette méthode a connu des développements parallèles à ceux de l'analyse des correspondances : généralisations au cas multiple, liens avec les modèles log-linéaires, études de validation et de stabilité (pour une vue générale de ces travaux, cf. Balbi, 1994).

c – Régression PLS (*Partial Least Squares*)

On esquissera dans ce paragraphe la régression PLS (dans le cas d'une variable expliquée), qui constitue un des compromis possibles entre la régression multiple et la régression sur composantes principales, toutes deux étudiées précédemment.

La régression multiple a le défaut de mettre les variables en compétition, et donc d'être très sensible aux colinéarités.

La régression sur composantes principales n'a pas cet inconvénient, mais peut produire des résultats peu clair pour l'utilisateur, dans la mesure où les composantes principales sont calculées sans référence aucune à la variable expliquée.

On dispose ici encore d'un ensemble de n observations de $p+1$ variables $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. On veut expliquer ou prévoir $\mathbf{y}$ à l'aide des variables explicatives, ou prédicteurs, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, lesquels peuvent maintenant être plus nombreux que les observations.

On va dans un premier temps calculer la combinaison linéaire p_1 telle que :

$$p_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p$$

Les coefficients w_{1j} étant donnés par la formule $\mathrm{w}_{1j} = \dfrac{\mathrm{cov}(\mathrm{x}_j, \mathrm{y})}{\sqrt{\sum_{j=1,p} \mathrm{cov}^2(\mathrm{x}_j, \mathrm{y})}}$

Chaque variable contribue ainsi au prorata de sa covariance avec y. Puis on explique y par p_1 selon le modèle de régression simple: $y = c_1p_1 + e_1$, et l'on obtient ainsi la première composante PLS.

$$y_1 = c_1p_1 + e_1 = c_1w_{11}x_1 + c_1w_{12}x_2 + \ldots + c_1w_{1p}x_p + e_1$$

Si nécessaire, on explique le résidu e_1 à partir des résidus des régressions de p_1 par chacun des x_j, (tout en étant orthogonal à y_1). Le résidu est remplacé par sa nouvelle valeur, et la composante y_1 est ainsi améliorée.

On trouvera dans Tenenhaus (1998) un exposé complet de ces techniques, (qui se généralisent au cas de plusieurs variables expliquées et de plusieurs groupes de variables) ainsi que les références bibliographiques de base.

8.2 Structures de graphe, analyses locales

La nature ou l'origine du recueil de données suggèrent souvent une structure *a priori* de l'ensemble des individus ou observations, avant toute analyse statistique. On peut voir sur la figure 8.2 - 1 des représentations qui correspondent à trois structures distinctes de l'ensemble des observations

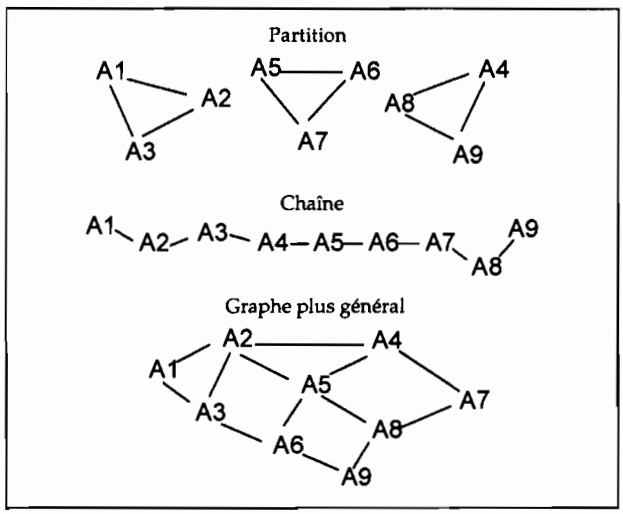


Figure 8.2 – 1. Graphes correspondant à trois types de structures courantes : Partition, chaîne (chronologie), graphe non orienté.

La structure de partition, qui correspond à un graphe formé de cliques disjointes, peut être décrite par une simple variable nominale, et entre donc dans le cas des analyses partielles présentées plus haut. Elle fera cependant l'objet d'un traitement particulier qui fait intervenir les matrices de covariances intra-classes et inter-classes, comme en analyse factorielle discriminante. La structure de chaîne correspond le plus souvent à des observations consécutives dans le temps, alors que la structure plus générale de graphe non orienté peut schématiser un système d'observations géographiques, pour lequel il existe une certaine dépendance entre observations contiguës. Ces structures ne peuvent pas être prises en compte par des variables nominales car elles concernent des couples d'observations.

8.2.1 Variance locale et covariance locale d'une variable

La décomposition de la variance en *variance entre classes* et *variance dans les classes* n'est plus possible dans le cas d'une structure de graphe. On peut faire intervenir une autre décomposition, fondée sur la propriété de la covariance

empirique entre deux variables x et y d'être également une covariance entre tous les couples d'observations[1] :

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}(x_i - x_{i'})(y_i - y_{i'})$$

On a évidemment l'expression de la variance si $x = y$:

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}(x_i - x_{i'})^2 \qquad [8.2-1]$$

S'il existe une relation binaire symétrique sur l'ensemble I des individus, définie par une partie symétrique R de l'ensemble produit $I \times I$ (R est l'ensemble des couples contigus (i, i')), on peut écrire, dans le cas de la variance :

$$var(x) = \frac{1}{2n(n-1)}\left\{\sum_{(i,i')\in R}(x_i - x_{i'})^2 + \sum_{(i,i')\notin R}(x_i - x_{i'})^2\right\}$$

On a séparé, dans le dénominateur de la variance, les contributions des couples contigus (ou adjacents sur le graphe) et des autres couples. On peut donc faire apparaître de cette façon deux composantes de la variance.

a – Matrice de contiguïté

Un couple de sommets adjacents du graphe est relié par une *arête*. Le nombre des arêtes attachées à un même sommet i est appelé le *degré* de ce sommet. Ce nombre est noté m_i. Le nombre d'arêtes du graphe s'écrit alors :

$$m_a = \frac{1}{2}\sum_{i=1}^{n} m_i$$

Si tous les sommets sont reliés par une arête, le graphe est dit *complet*. Un tel graphe possède $n(n-1)/2$ arêtes (on ne distingue pas l'arête (i, i') de l'arête (i', i)). On construit une matrice carrée **M**, d'ordre (n, n), dite *matrice de contiguïté*, ou matrice associée au graphe telle que $m_{ii'} = 1$ si i est voisin de i' et $m_{ii'} = 0$ sinon[2]. Notons qu'avec les notations précédentes :

$$m_i = \sum_{i'=1}^{n} m_{ii'}$$

On voit que cette matrice est symétrique du fait de la symétrie de la relation de contiguïté. On adoptera la convention selon laquelle une observation n'est pas contiguë à elle-même, ce qui implique que les termes m_{ii} situés sur la diagonale

---

[1] La covariance empirique sera calculée ici en divisant la somme des produits par $(n-1)$ (au lieu de n). On obtient ainsi une estimation sans biais de la covariance théorique.

[2] On peut également travailler sur des structures de contiguïté qui incluent des proximités à distance $1, 2, \ldots, n$ les matrices de contiguïté correspondantes étant construites à partir des puissances booléennes de la matrice **M** (cf. Lebart, 1969-a). Nous nous limiterons ici aux structures de contiguïté pour lesquelles deux parties sont immédiatement contiguës.

principale de la matrice **M** sont tous nuls. On peut réécrire, dans la dernière formule donnant la variance, le terme faisant intervenir les couples contigus sous la forme :

$$\sum_{(i,i')\in R} (x_i - x_{i'})^2 = \sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}(x_i - x_{i'})^2$$

On appelle variance locale $v_l(x)$ d'une variable x la demi-moyenne des carrés des accroissements correspondant à des observations contiguës. Posant :

$$m = \sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}$$

on a :

$$v_l(x) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}(x_i - x_{i'})^2 \qquad [8.2 - 2]$$

La variance totale $var(x)$ donnée par la formule [8.2 - 1] est donc la variance locale correspondant au graphe complet.

b – Coefficient de contiguïté de Geary (1954)

Il est clair que si la variable x est indépendante de la structure de graphe, la variance locale est une estimation de la variance totale. Si les valeurs voisines de x sont corrélées positivement, alors la variance locale sous-estime la variance totale. Le coefficient de contiguïté $c(x)$ est défini comme le rapport de la variance locale à la variance totale.

$$c(x) = v_l(x) / var(x)$$

Sous l'hypothèse selon laquelle les valeurs x_i sont des réalisations de variables aléatoires normales indépendantes, on peut calculer les quatre premiers moments du coefficient $c(x)$ en fonction de la trace des puissances de la matrice **M** associée au graphe[1]. On voit ainsi que pour le graphe des départements français (pour lequel deux sommets-départements sont joints par une arête s'ils ont une frontière commune) la distribution de $c(x)$ est très proche d'une distribution normale.

c – Nouvelle définition de la variance locale

Une modification va être faite sur la définition du coefficient $c(y)$ pour rendre la variance locale compatible avec la variance "intra" ("*within*") quand le graphe décrit une partition des observations (i.e. une série de cliques [sous-graphes complets] disconnectées). Cette modification a été proposée par Mom (1988), et indépendamment par Escofier (1989).

On note par **N** la (n,n) matrice diagonale ayant le degré de chaque sommet i comme élément diagonal n_i (n_i dénote ici n_{ii}).

---

[1] Pour un exposé plus complet cf. Lebart (1969a). Pour d'autres applications de la notion de contiguïté, cf. Aluja Banet et Lebart. (1984).

On a : $n_i = \sum_k m_{ik}$

La variance locale sera redéfinie comme:

$$v^*(x) = (1/n)\sum (x_i - m_i^*)^2$$

Dans cette dernière formule, la *moyenne locale* est définie comme :

$$m_i^* = (1/n_i)\sum_k m_{ik} x_k$$

C'est la moyenne des valeurs adjacentes au sommet *i*.

Notons que si le graphe *G* est régulier (i.e. si n_i est constant) :

$$v^*(y) = v_I(y)$$

On peut redéfinir le coefficient de contiguïté $c(x)$:

$$c(x) = v^*(x) / var(x) \qquad [8.2 - 3]$$

d – Bornes pour c(x)

On rappelle dans cette section que les vecteurs propres calculés à partir de l'analyse des correspondances d'une matrice **M** associée au graphe *G* ont des propriétés optimales vis-à-vis du coefficient de contiguïté.

Pour une variable centrée réduite *x*, le coefficient $c(x)$ s'écrit (**I** désignant la matrice unité et **N** la matrice diagonale définie plus haut):

$$c(x) = \mathbf{x}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\,\mathbf{x} / \mathbf{x}'\mathbf{x}$$

Donc, comme en analyse générale, le minimum de $c(x)$, que l'on notera μ, est la plus petite racine de:

$$(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\,\psi = \mu\,\psi$$

Si le graphe est régulier (si tous les sommets sont adjacents à un même nombre *a* d'arêtes) la matrice **N** s'écrit : **N** = a**I**, donc, dans ce cas, $\mathbf{N}^{-1}\mathbf{M}$ est symétrique, et l'équation précédente peut s'écrire: $(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})^2\,\psi = \mu\,\psi$

ce qui implique : $(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\,\psi = \sqrt{\mu}\,\psi$ et donc : $\mathbf{N}^{-1}\mathbf{M}\,\psi = (1 - \sqrt{\mu})\,\psi$

Notons que les *formules de transition* correspondant à l'analyse des correspondances de la matrice **M** s'écrivent, pour le premier facteur:

$$\mathbf{N}^{-1}\mathbf{M}\,\varphi = \varepsilon(\varphi)\sqrt{\lambda}\,\varphi$$

si $\varepsilon(\varphi) = +1$, le facteur φ est dit direct, alors que si $\varepsilon(\varphi) = -1$, le facteur est dit inverse. Un facteur inverse correspond en fait à une valeur propre négative de la matrice de données symétrique initiale **M** (matrice associée au graphe).

Puisque $c(x)$ est positif, la valeur minimale μ correspond à la valeur maximum de λ, notée λ_{max}, pour un facteur direct ($\varepsilon(\varphi) = +1$). Donc, la borne inférieure de $c(x)$ est:

$$Min\,[\,c(x)\,] = (1 - \sqrt{\lambda_{max}})^2$$

Ce minimum est atteint quand ψ est le premier facteur direct φ obtenu à partir de l'analyse des correspondances de la matrice **M**. Alors, la séquence des

premiers facteurs φ_r correspond à une séquence de variables N-orthogonales ayant la propriété de contiguïté extrémale. Cette propriété explique la bonne qualité de la description des graphes par l'analyse des correspondances de leur matrice associée[1].

e – Analyse des correspondances des matrices associées M

Dans certains cas lors de l'analyse des correspondances de la matrice associée à un graphe symétrique, un calcul analytique exact peut être fait sans recours à l'ordinateur. Il est alors intéressant d'étudier analytiquement les variations des représentations en fonction des différents codages de la matrice associée.

Examinons par exemple le cas de l'analyse d'un cycle simple. La matrice **M** n'a que deux éléments non nuls (égaux à *1*) par ligne et par colonne.

Désignons par n le nombre de sommets du graphe. Pour $n = 5$, on a la situation représentée par la figure 8.2-2.

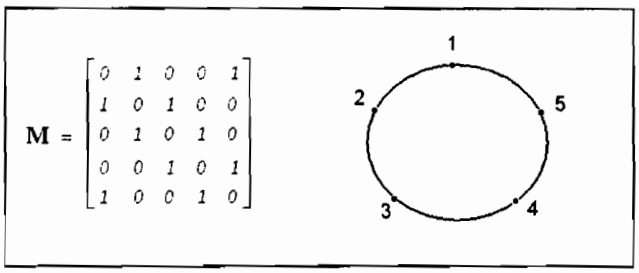


Figure 8.2 – 2. Exemple de cycle et de sa matrice associée

La relation $\mathbf{N}^{-1}\mathbf{M}\ \varphi = \varepsilon(\varphi)\sqrt{\lambda}\ \varphi$ s'écrit encore pour $1 < j < n$:

$$\frac{1}{2}(\varphi(j-1)+\varphi(j+1)) = \varepsilon(\varphi)\sqrt{\lambda}\ \varphi(j)$$

Les solutions de ce type classique d'équation aux différences finies sont, compte tenu des conditions aux limites :

$$\varphi_\alpha(j) = cos(\frac{2j\alpha\pi}{n}) \quad \text{et} \quad \psi_\alpha(j) = sin(\frac{2j\alpha\pi}{n})$$

---

[1] On trouvera dans Benzécri (1973, Tome II B, n°10 : "Sur l'analyse de la correspondance définie par un graphe") des exemples donnant lieu à des résolutions numériques ou analytiques de description de graphes particuliers (cartes géographiques, réseaux à mailles carrées, produits tensoriels de réseaux, etc.). On observe en particulier dans ces cas des "effets Guttman à plusieurs dimensions", ce qui se traduit par des vecteurs propres de rangs élevés dont les composantes sont des fonctions polynomiales de celles des premiers vecteurs propres. Benzécri note l'analogie avec les fonctions propres de l'opérateur de Laplace. Ces résultats sont retrouvés puis étendus dans une littérature plus récente [cf. Mohar (1991, 1997), Chung (1997), Koren et al. (2002)] qui porte sur les propriétés spectrales de la matrice L = N-M (L est le *Laplacien de* M).

Ce sont les *j*ièmes composantes des deux facteurs associés à la valeur propre double :

$$\lambda_\alpha = cos^2(\frac{2\alpha\pi}{n})$$

On obtient dans le plan des deux premiers facteurs l'équation paramétrique d'un cercle et donc une reconstitution satisfaisante de la structure dont le tableau **M** représente un codage particulier.

La trace de la matrice à diagonaliser s'écrit :

$$tr\frac{1}{4}\mathbf{M}^2 = \frac{n}{2}$$

Le taux d'inertie correspondant à l'axe α est donc :

$$\tau_\alpha = \frac{2}{n}cos^2(\frac{2\alpha\pi}{n})$$

Le résultat, en apparence paradoxal, est le suivant : le taux d'inertie du sous-espace qui "restitue" la structure initiale peut être rendu aussi petit que l'on veut, pourvu de choisir un cycle assez long : si $n = 10^3$, alors $\tau_1 \approx 2\times10^{-3}$.

Ces analyses de graphes servent de contre-exemple à la qualification des taux d'inertie comme « pourcentages d'information ». Elles confirment les faiblesses de ces taux d'inertie dans certaines situations.

8.2.2 Analyse locale

Généralisons les résultats précédents au cas de plusieurs variables [1]. Si **X** désigne la matrice d'ordre (n,p) de terme général $x_{jj'}$ (n observations de p variables), la matrice des covariances locales **V*** s'écrit :

$$\mathbf{V^*} = (1/n)\ \mathbf{X'(I - N^{-1}M)\ X} \qquad [8.2\text{-}4]$$

Si le graphe est formé de cliques disjointes (structure de partition), la matrice **V*** est proportionnelle à la matrice **D** de variance intra-classes (chapitre 7), qu'elle généralise dans ce cas.

On peut définir une matrice des corrélations locales, de terme général :

$$c^*(x_j, x_{j'}) = \frac{cov^*(x_j, x_{j'})}{\sqrt{v^*(x_j)v^*(x_{j'})}}$$

La diagonalisation de cette matrice nous fournit, comme en analyse en composantes principales, une image des liaisons existant au niveau local qu'il

---

[1] Alors que le coefficient de contiguïté de Geary est l'analogue, dans le cas d'un ensemble fini, d'un point du *variogramme* (correspondant à la distance "1" dans le cas isotropique) utilisé en géostatistique (Matheron, 1963), la matrice des covariances locales est l'analogue, dans le cas fini ou discret, de la matrice de *codispersion intrinsèque* qui intervient dans la théorie des variables régionalisées (Matheron, 1965).

peut être intéressant de confronter aux liaisons globales (ainsi, dans le cas de données géographiques, l'opposition entre grandes régions très différentes peut masquer des covariations que l'analyse de la matrice des corrélations locales restitue). Cette diagonalisation (Analyse en composantes principales locales) qui fournit une description des corrélations locales, peut être comparée aux résultats d'une analyse en composantes principales classique ne prenant pas en compte la structure de graphe. Les comparaisons entre matrices des covariances locales et globales peuvent se faire par analyses "Procrustéennes" (cf. section 8.3).

8.2.3 Analyse de contiguïté et projections révélatrices

a – Analyse de contiguïté

La variance locale $v^*(u)$ d'une combinaison linéaire $u(i)$ $(i = 1, 2, \ldots, n)$ des p variables s'écrit en fonction de la matrice de contiguïté, avec les notations habituelles :

$$v^*(u) = \mathbf{u'V^*u}$$

Si $\mathbf{V}$ désigne la matrice des covariances totales, le coefficient de contiguïté de la combinaison linéaire $u(i)$ s'écrit comme le quotient des deux formes quadratiques :

$$c(u) = \frac{\mathbf{u'V^*u}}{\mathbf{u'Vu}} \qquad [8.2 - 5]$$

La recherche des combinaisons linéaires de contiguïté minimale (analyse de contiguïté) constitue une généralisation de l'analyse factorielle discriminante, qui se réduit à celle-ci lorsque le graphe est formé de cliques disjointes.

L'analyse de contiguïté est beaucoup moins utilisée que l'analyse discriminante qui a le mérite de rapprocher des données complexes et une structure très simple (la structure de partition)[1].

Elle peut être utilisée dans le cadre d'une démarche s'apparentant aux techniques dites de *projections révélatrices* (cf. Caussinus, 1992, et la section 3.3.6 dévolue à l'analyse en composantes indépendantes) qui, très schématiquement, cherchent des directions "intéressantes" plutôt que des dimensions principales au sens des moindres carrés [2]. Il existe autant de variantes de la méthode qu'il existe de façons de définir l'intérêt d'une projection.

---

[1] Pour des programmes de calcul et des applications de l'analyse de contiguïté, cf. Lebart et Tabard (1973).

[2] L'expression "Projection révélatrice" est la traduction, par des auteurs français (Y. Escoufier, H. Caussinus) de l'expression "projection pursuit" (cf. Friedman et Tukey, 1974; Friedman , 1987; Jones et Sibson, 1987).

b – Représentation de groupes par projection

Si l'on veut déterminer une projection qui sépare le mieux possible des groupes existant dans l'ensemble des observations (sans connaître *a priori* ces groupes, sinon l'analyse factorielle discriminante classique répond à la question), on peut procéder de la façon suivante. On part d'un tableau de données **X** d'ordre *(n,p)* pour lequel on n'a aucune information externe. On définit une relation de contiguïté sur l'ensemble des lignes de **X** (il s'agit ici d'une contiguïté *a posteriori*) à partir d'un seuil de distance d_0. Parmi les *n(n-1)* couples d'observations (lignes de **X**) dans l'espace $\mathcal{R}^p$, les couples d'observations (i, i') tels que $d(i, i') \leq d_0$ sont déclarés "contigus". On définit donc la matrice de contiguïté **M** par les relations :

$$m_{ii'} = 1 \quad \text{si} \quad d(i, i') \leq d_0 \quad \text{et} \quad m_{ii'} = 0 \text{ sinon}$$

Une seconde façon de définir une relation de contiguïté *a posteriori* est de considérer comme contigus le pourcentage s_0 ($s_0 = 10$ par exemple) des couples les plus proches au sens de $d(i, i')$, ce qui permet de définir un seuil d_0 après le calcul des $n(n-1)/s_0$ plus petites distances.

Une troisième façon utilise les k plus proches voisins : sont considérés comme contigus à la ligne i de **X** les k lignes les plus proches au sens de la distance $d(i, i')$. Cette méthode permet d'obtenir un graphe régulier, (avec les notations précédentes : $m_i = k$) mais peut rattacher artificiellement des points isolés ou des petits groupes de points, au graphe d'ensemble qui est nécessairement connexe.

Une fois déterminée la matrice **M**, l'analyse de contiguïté, qui calcule les combinaisons linéaires réalisant les minima de $c(\mathbf{u})$ donné par la formule [8.2 - 5], va produire une représentation qui respectera au mieux la structure de graphe et donc les plus fortes proximités entre points. En revanche, les distances moyennes ou grandes joueront un rôle moins important, ce qui a pour effet de "déplier" une éventuelle structure continue (cf. figure 8.2 - 2).

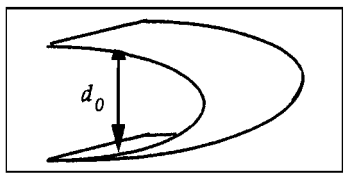


Figure 8.2 – 2. Exemple de dépliage d'une structure par analyse de contiguïté
Si le seuil est inférieur à la distance d_0, aucune arête ne joindra les deux plis; le graphe épousera donc la forme de la surface, qui sera dépliée dans les premiers plans de l'analyse.

On peut imaginer qu'au lieu de sélectionner les arêtes les plus courtes, on garde toutes les arêtes, que l'on pondère par une fonction décroissante de la distance (le graphe de contiguïté devenant un graphe complet valué). On rejoint alors une série de travaux sur ce sujet plus proches des approches

classiques de directions révélatrices. Les premiers travaux sur ces thèmes sont ceux de Art *et al.* (1982), de Gnanadesikan *et al.* (1982). Ils ont été suivi des travaux de Yenyukov (1988), Caussinus et Ruiz (1990, 2003) [1].

c – Liens avec les analyses partielles

Comme indiqué au paragraphe *d* ci-dessus, on peut définir (au moins de trois façons différentes) une matrice de contiguïté **M** d'ordre (n, n) à partir d'un tableau de données que l'on appellera maintenant **Z** d'ordre (n, q). Si l'on désire étudier des corrélations partielles entre les p colonnes d'une matrice **X** d'ordre (n,p) "*à Z constant*", on peut calculer la matrice des covariances partielles par la formule [8.1- 3], mais on peut également calculer la matrice des covariances locales données par la formule [8.2 - 4] où **M** est une matrice de contiguïté issue de **Z** (et **N** la matrice diagonale des degrés de **M**). Cette mesure de covariance partielle à l'avantage d'être non-linéaire (vis-à-vis des colonnes à fixer). Elle a l'inconvénient d'exiger des calculs de distances entre les lignes de **Z** et dépend donc des échelles des mesures ou des poids des colonnes de **Z**, ce qui n'est évidemment pas le cas pour des covariances calculées sur des résidus obtenus par régression multiple.

8.2.4 Extensions, généralisations, applications

Plusieurs variantes ou généralisations sont possibles autour de la notion de contiguïté. Déjà, à l'origine de ces travaux, les coefficients de Geary (1954) et de Moran (1948, 1954) constituaient deux mesures possibles (et très voisines) du degré de contiguïté [2].

Citons brièvement, sans être exhaustif, quelques extensions ou applications : Le Foll (1982) introduit une pondération des sommets du graphe (les arêtes sont alors valuées par les produits des masses des sommets adjacents); Le Foll et Burtschy (1983) confrontent l'analyse locale et l'analyse des correspondances classique pour décrire des tableaux d'échanges; Carlier (1985) étudie les évolutions de tables de contingence par plusieurs méthodes dont l'analyse locale; Sabatier (1987) situe l'analyse locale dans un formalisme qui intègre les analyses partielles. Les travaux de Mom (1988) et d'Escofier (1989) ont été cités précédemment. Dans l'analyse lissée, chaque point-individu i (ligne i de **X**) est remplacé par le barycentre de ses voisins sur le graphe. Ceci revient, avec nos

---

[1] L'approche par analyse de contiguïté permet de mettre en évidence les deux structures qui sont confrontées: la structure locale, traduite sous forme de graphe (qui peut lui même être décrit par analyse des correspondances; cf. § 3.7-2 ci-dessus), la structure globale (analyse en composantes principales de **X**), et le compromis entre les deux structures, décrit par l'analyse de contiguïté (cf. Burtschy et Lebart, 1991 ; Lebart, 2000).

[2] On pourra consulter les ouvrages de Cliff et Ord (1981), Ripley (1981), pour une vue plus large de la panoplie des outils disponibles.

notations (qui ne concernent que le cas où les sommets ont des poids *a priori* identiques, mais peuvent avoir des degrés différents) à remplacer $\mathbf{X}$ par $\mathbf{N^{-1}MX}$. Ce lissage a pour effet d'éliminer les variations locales.

Cazes et Moreau (1991), Moreau (1992), Moreau *et al.* (2000) considèrent le cas d'une double structure de graphe, présente à la fois sur les lignes et les colonnes d'une table de contingence. Faraj (1993) utilise l'analyse locale comme une analyse partielle pour fixer l'effet de plusieurs variables nominales. Mentionnons enfin une synthèse de travaux sur ces thèmes par Méot *et al.* (1993) [1].

8.2.5 Cas particuliers : Structure de partition

Il est fréquent que l'ensemble des individus ou observations soit partitionné en q classes connues *a priori* et jouant un rôle privilégié par rapport aux variables-colonnes du tableau de données $\mathbf{X}$ d'ordre (n, p). Cette situation a été rencontrée en analyse factorielle discriminante (chapitre 7) : il s'agissait alors de prédire l'appartenance d'un individu à une classe à partir des valeurs des variables pour cet individu. Selon la formule de Huygens (formule [6.2 - 1] du § 6.2.3.), l'inertie totale du nuage se décompose en inertie inter-classes (variabilité entre les classes) et inertie intra-classes (variabilité à l'intérieur des classes)[2] :

$$I = I_{inter} + I_{intra}$$

A cette décomposition est associée deux analyses : l'analyse *inter-classes* qui décrit les positions relatives des classes et ignore les individus, et l'analyse *intra-classes* qui s'attache à décrire les différences de comportement à l'intérieur des classes ce qui revient à éliminer l'effet dû à la structure de partition.

a – Analyse inter-classes

L'analyse inter-classes est simplement l'analyse du tableau agrégé d'ordre (q, p). On a vu que l'analyse factorielle discriminante est une analyse inter-classes particulière (§7.5.1)[3]. Dans le cas où les variables sont nominales, on réalise l'analyse des correspondances du tableau des centres de gravité (ou tableau des

---

[1] Il faudrait citer, dans ce survol des utilisations de la notion de contiguïté, les méthodes de classification faisant appel aux contraintes de contiguïté. Une revue en est faite par Gordon et Finden (1985).

[2] Dans le cas de variables continues, il s'agit plus spécifiquement de la décomposition de la matrice de covariance (ou de corrélation si les variables sont réduites) en variance inter-classes (variance des moyennes des classes) et en variance intra-classes (variance de chaque classe autour de sa moyenne) donnée par la formule [7.1 - 1].

[3] Elle peut en effet être décrite comme une analyse en axes principaux des points-moyens de chacune des classes dans la métrique définie par l'inverse de la matrice des covariances "intra-classes".

barycentres) des q groupes d'individus, obtenu en croisant les classes de la partition avec les modalités des autres variables[1]. L'analyse inter-classes correspond dans ce cas à l'analyse discriminante barycentrique (cf. § 7.5.1). L'analyse inter-classes est clairement un cas particulier de l'analyse lissée précitée lorsque le graphe est associé à une partition.

b – Analyse intra-classes

L'analyse intra-classes permet d'étudier les différences de comportement à l'intérieur des classes en analysant la dispersion des individus à l'intérieur de leurs classes d'appartenance (cf. Benzécri, 1983; Cazes, 1986-a; Benali et Escofier, 1990). Chaque individu est représenté par un point dont les coordonnées expriment l'écart entre ses propres coordonnées et celles du centre de gravité de sa classe. L'inertie inter-classes est ainsi éliminée. On ne cherche donc plus à savoir de quelle manière un individu se différencie de l'ensemble du nuage mais comment il se différencie de l'ensemble des individus appartenant à la même classe. On s'affranchit ainsi de l'influence de la variable de partition en étudiant les liaisons entre les variables à analyser, conditionnellement à la variable définissant la partition.

L'analyse intra-classes est un cas particulier de l'analyse des différences locales (graphe associé à une partition) et également un cas particulier de l'analyse partielle (cf. section 8.1) lorsque la variable exogène **z** est nominale.

Escofier (1987) introduit une méthode d'analyse intra-classes dans le cas de variables nominales, appelée *analyse des correspondances multiples conditionnelles*, qui est en fait un cas particulier de la généralisation de l'analyse des correspondances proposée également par Escofier (1984). L'influence de la variable de partition est éliminée ; le nuage des individus est recentré par classe, et le nuage des modalités est projeté sur l'orthogonal du sous-espace engendré par les modalités de la variable de partition[2].

Une extension de l'analyse des correspondances multiples conditionnelles, est étudiée par Piron (1990, 1992) lorsque les variables sont des fréquences. Dans ce cas, la structure induite sur les individus relève d'une série de partitions emboîtées (structure fréquente dans les relevés géographiques).

Pour le cas de doubles partitions (partition Q sur les lignes et partition S sur les colonnes d'une table de contingence) Cazes (1986-a et 1986-b), et Cazes, Chessel et Doledec (1988) proposent l'*analyse des correspondances internes* qui consiste à réaliser l'analyse intra-classes en considérant un double centrage dans l'espace des lignes et dans celui des colonnes. On projette d'une part le nuage des

---

[1] Il s'agit en fait d'une bande d'un tableau de Burt (cf.§ 5.3.3).

[2] L'analyse des correspondances multiples conditionnelle conserve toutes les propriétés de l'analyse des correspondances. Elle est implémentée dans le logiciel SPAD sous forme de procédure.

points-lignes sur l'orthogonal du sous-espace engendré par les modalités de la variable de partition Q, et d'autre part le nuage des points-colonnes sur l'orthogonal du sous-espace engendré par les modalités de la variable de partition S.

8.3 Tableaux multiples, groupes de variables

L'analyse des tableaux multiples est un très vaste domaine de recherche que l'on ne fera qu'effleurer dans cette section, en se limitant à quelques situations spécifiques, proches de la démarche exploratoire.

Le théorème d'Eckart et Young (décomposition aux valeurs singulières étudiée au chapitre 1) qui est à la base des méthodes factorielles, n'admet pas de généralisation au sens suivant : il n'existe pas de décomposition optimale unique d'un tableau à trois entrées (empilement de q tableaux $\mathbf{X}_k$, chacun d'ordre (n,p)) en tableaux de rangs *1*.

En revanche, il existe des modèles particuliers, qui varient selon les disciplines et la nature des tableaux, pour aborder ce type de données.

8.3.1 Quelques travaux de référence

Commençons par évoquer quelques travaux de référence sur le thème des tableaux à plusieurs dimensions[1].

Les premiers travaux sur ce thème sont ceux de Tucker (1964, 1966) puis ceux de Harshman (1970), tous les deux dans le cadre de l'analyse factorielle classique. Montrons brièvement quelles sont les relations qui sont à la base de ces modèles.

L'un des modèles de Tucker, dit TUCKALS-3 (Kroonenberg et de Leeuw, 1980), s'applique à une séquence de matrices symétriques d'ordre (p, p) $\mathbf{S}_1,\ldots, \mathbf{S}_q$ (qui sont par exemple des matrices de distances entre individus). Il conduit à la relation (s_{ijk} désignant une estimation, par le modèle, de l'élément (i, j) de la matrice $\mathbf{S}_k$) :

---

[1] On trouvera une synthèse et une classification des principales démarches dans l'ouvrage de Kroonenberg (1983) qui a prolongé les travaux de Tucker. On pourra aussi consulter la revue comparative de Carlier *et al.* (1988), qui fait d'ailleurs partie d'un recueil entièrement consacré à ce thème (Coppi et Bolasco, 1989). Une revue se trouve également dans Kiers (1989). Sur le thème plus circonscrit des évolutions de tables de contingence, cf. Carlier (1985), van der Heijden (1987).

$$s_{ijk} = \sum_{u=1}^{p} \sum_{v=1}^{p} \sum_{t=1}^{r} a_{iu} a_{jv} b_{kt} c_{uvt}$$

Le modèle dit PARAFAC, de Harshman, donne lieu à une relation analogue, mais plus simple.

Pour une série de matrices $\mathbf{X}_k$ d'ordre (n, p), le terme général x_{ijk} peut s'écrire :

$$x_{ijk} = \sum_{t=1}^{r} a_{it} b_{jt} c_{kt}$$

Ces formules peuvent être vues comme des généralisations possibles de la formule de reconstitution de données[1].

Une autre méthode très utilisée dans le contexte des méthodes de *multidimensional scaling* est la méthode INDSCAL de Carroll et Chang (1970) qui est un cas particulier de la méthode PARAFAC de Harshman.

Ces exemples laissent imaginer le nombre de modèles et de variantes possibles.

Les quatre paragraphes de cette section seront tous consacrés à une structure de tableaux multiples très particulière, mais fréquente en pratique : il s'agit d'un tableau $\mathbf{X}$ d'ordre (n,p) tel que :

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_q)$$

Les différents blocs n'ont pas forcément le même nombre de colonnes et cette structure est par conséquent plus générale qu'un tableau à trois entrées.

Selon les cas, les lignes seront des individus ou observations, les colonnes de chaque bloc des variables. Les blocs peuvent correspondre à des instants ou des contextes différents pour les mêmes variables, ou à des groupes de variables différents.

La section 8.1 a abordé le cas de l'analyse d'un tableau de données de type $\mathbf{R} = (\mathbf{X}, \mathbf{Z})$ dans laquelle les deux ensembles de colonnes (colonnes de $\mathbf{X}$ et de $\mathbf{Z}$) jouaient des rôles dissymétriques. Il existe des circonstances dans lesquelles les rôles sont parfaitement symétriques. C'est le cas notamment des méthodes d'analyses procrustéennes orthogonales qui visent à comparer deux structures de distances sur les mêmes objets, ceux-ci étant décrits successivement par deux ensembles différents de variables (§ 8.3.2).

La méthode STATIS (§ 8.3.3) et l'analyse factorielle multiple (§ 8.3.4) sont proches à bien des égards dans leurs procédures mais se différencient dans les options de traitements. Elles procèdent en trois étapes : la comparaison globale des tableaux, la représentation du nuage moyen et la représentation simultanée des tableaux.

---

[1] Cf. par exemple: Hayashi et Hayashi (1982) pour un algorithme d'estimation des coefficients du modèle.

Brièvement évoquée à propos de l'analyse canonique, l'analyse canonique généralisée (on désigne sous ce nom l'une des généralisations possibles de l'analyse canonique, en fait la plus mentionnée et utilisée) sera présentée dans un cadre plus général (§ 8.3.5). Cette méthode, assez délicate à utiliser directement en pratique, fournit un cadre théorique simple commun aux principales méthodes factorielles exploratoires et aux méthodes explicatives de base des chapitres 2 et 7, qu'elle contient toutes comme cas particulier.

8.3.2 Analyses procrustéennes[1]

Les méthodes d'analyse procrustéennes tentent de répondre à une préoccupation fréquente en statistique multidimensionnelle : n individus ou observations sont décrits d'une part par p variables (colonnes de **X**), d'autre part par q autres variables (colonnes de **Z**). Comment comparer les deux nuages d'individus, les deux systèmes de distances entre individus ?

C'est Tucker (1958) qui proposa à l'origine une telle méthode pour comparer deux batteries de tests passés sur les mêmes individus[2]. La technique a ensuite été étudiée par Cliff (1966), Schönemann (1968), Schönemann et Carroll (1970), puis généralisée par Gower (1975, 1984)[3].

a – Analyse procrustéenne orthogonale

Fixons **X** par exemple (les rôles de **X** et **Z** sont symétriques) et supposons $p = q$. Ceci n'est pas une restriction car, si par exemple $p > q$, on peut toujours compléter le tableau **Z** par $p - q$ colonnes nulles. Si les lignes de **Z**, d'ordre (n,p), subissent toutes un même déplacement (translation et rotation dans $\mathcal{R}^p$), **Z** est transformé en **ZB** + **T**, où **T** est une matrice d'ordre (n,p) dont les colonnes peuvent être différentes, mais constantes (translation) et où **B** (p, p) est une matrice orthogonale (rotation ou symétrie par rapport à l'origine).

On cherchera à rendre minimale la somme des carrés s des écarts entre **X** et (**ZB** + **T**), qui peut s'écrire[4] :

---

[1] Procruste est un aubergiste de la mythologie grecque qui raccourcissait ou allongeait ses clients (**X**, par exemple) pour les ajuster à la longueur de son lit (**Z**). Thésée mettra fin à ses jours en lui infligeant le même supplice. Cf. l'ouvrage de synthèse complet sur ces méthodes : Gower et Dijksterhuis (2004).

[2] On peut de la même façon comparer un même ensemble de variables sur deux ensembles d'individus différents. C'est le cas si l'on veut comparer deux matrices des corrélations (une matrice des corrélations globales, par exemple, à confronter à une matrice des corrélations locales).

[3] Cf. également Lafosse (1985), et pour des travaux récents de cet auteur, Kissita *et al.* (2004).

[4] Rappelons que $trace(\mathbf{A}'\mathbf{A}) = \sum_{i,j} a_{ij}^2$; que *trace* **A** = *trace* **A**'; et que, lorsque les opérations sont possibles, *trace* (**A**+**C**) = *trace* **A** + *trace* **C** ; *trace* **AC** = *trace* **CA**.

$$s = trace\ (\mathbf{X} - \mathbf{ZB} - \mathbf{T})'\ (\mathbf{X} - \mathbf{ZB} - \mathbf{T})$$

Le critère s s'écrit encore, si les tableaux $\mathbf{X}$ et $\mathbf{Z}$ sont centrés en colonnes :

$$s = trace\ (\mathbf{X} - \mathbf{ZB})'\ (\mathbf{X} - \mathbf{ZB}) + \mathbf{T}'\mathbf{T} \qquad [8.3 - 1]$$

La recherche d'un minimum pour s implique $\mathbf{T} = \mathbf{0}$ (aucune translation n'est requise quand les tableaux sont centrés).

Développant l'expression du critère s et en tenant compte du fait que :

$$trace\ \mathbf{B}'\mathbf{Z}'\mathbf{ZB} = trace\ \mathbf{Z}'\mathbf{ZBB}' = trace\ \mathbf{Z}'\mathbf{Z}$$

il vient :

$$s = trace\ (\mathbf{X}'\mathbf{X} + \mathbf{Z}'\mathbf{Z} - 2\ \mathbf{B}'\mathbf{Z}'\mathbf{X})$$

Rendre minimal le critère s revient à rendre maximal $trace\ (\mathbf{B}'\mathbf{Z}'\mathbf{X})$.

Ecrivons la formule de reconstitution des données (cf. § 1.2.5, formule [1.2-5]) issue de l'analyse générale (décomposition aux valeurs singulières) du tableau $\mathbf{Z}'\mathbf{X}$:

$$\mathbf{Z}'\mathbf{X} = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\, \mathbf{v}_\alpha \mathbf{u}'_\alpha$$

d'où :

$$trace(\mathbf{B}'\mathbf{Z}'\mathbf{X}) = trace\left(\sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\, \mathbf{B}'\mathbf{v}_\alpha \mathbf{u}'_\alpha \right) = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\, (\mathbf{u}'_\alpha \mathbf{B}'\mathbf{v}_\alpha)$$

$\mathbf{B}$ étant orthogonal et $\mathbf{v}_\alpha$ unitaire, $\mathbf{B}'\mathbf{v}_\alpha$ est unitaire et on aura toujours $\mathbf{u}'_\alpha \mathbf{B}'\mathbf{v}_\alpha \leq 1$. On aura $\mathbf{u}'_\alpha \mathbf{B}'\mathbf{v}_\alpha = 1$ si et seulement si $\mathbf{B}'\mathbf{v}_\alpha = \mathbf{u}_\alpha$.

D'où la relation $\mathbf{B}'\mathbf{V} = \mathbf{U}$ et la solution cherchée : $\mathbf{B} = \mathbf{VU}'$

L'analyse procrustéenne orthogonale implique donc la décomposition aux valeurs singulières de $\mathbf{Z}'\mathbf{X}$ et donc la diagonalisation de la matrice $\mathbf{X}'\mathbf{ZZ}'\mathbf{X}$.

► *Autre présentation de l'analyse procrustéenne orthogonale*

On peut donner une autre présentation de cette méthode, en procédant de façon hiérarchique, par extraction progressive d'axes procrustéens. La méthode est analogue à l'analyse canonique, aux contraintes de normalisation près.

Les tableaux $\mathbf{X}$ et $\mathbf{Z}$ étant centrés, elle consiste à chercher deux combinaisons linéaires $\mathbf{Xu}$ et $\mathbf{Zv}$, à coefficients normés ($\mathbf{u}'\mathbf{u} = 1$, $\mathbf{v}'\mathbf{v} = 1$), de covariances maximales, c'est-à-dire telles $\mathbf{v}'\mathbf{Z}'\mathbf{Xu}$ soit maximale.

Une démonstration en tout point analogue à celle du paragraphe 2.1.2 (comme dans le cas de l'analyse canonique, les deux multiplicateurs de Lagrange sont égaux à une même valeur λ) nous montre alors que $\mathbf{u}$ et $\mathbf{v}$ sont solutions de :

$$\mathbf{X}'\mathbf{ZZ}'\mathbf{Xu} = \lambda^2 \mathbf{u} \qquad \text{et} \qquad \mathbf{Z}'\mathbf{XX}'\mathbf{Zv} = \lambda^2 \mathbf{v}$$

qui sont bien les équations de l'analyse générale du tableau $\mathbf{Z}'\mathbf{X}$.

En extrayant les différents axes (avec des contraintes d'orthogonalité usuelles), et en notant **U** et **V** les matrices orthogonales contenant en colonnes les vecteurs $\mathbf{u}_\alpha$ et $\mathbf{v}_\alpha$ correspondant aux différents axes indexés par α, on aura rendu maximal le critère : *trace* (**V'Z'XU**) (les éléments diagonaux de cette matrice sont en effet les covariances maximales trouvées). Remarquons que, jusqu'ici, on n'a pas supposé $p = q$ dans cette présentation.

Or rendre maximale cette trace revient à rendre minimal le critère lorsque $p = q$ (**U** et **V** étant deux matrices orthogonales) :

$$s_1 = trace(\mathbf{XU - ZV})'(\mathbf{XU - ZV})$$

On peut écrire cette quantité :

$$s_1 = trace\ \mathbf{U}(\mathbf{XU - ZV})'(\mathbf{XU - ZV})\mathbf{U}'$$

Finalement :

$$s_1 = trace\ (\mathbf{X - ZVU}')'(\mathbf{X - ZVU}')$$

qui coïncide avec le critère s de la première approche pour $\mathbf{B} = \mathbf{VU}'$ (formule [8.3-1], avec $\mathbf{T} = \mathbf{0}$)

b – Analyse procrustéenne sans contrainte[1]

On cherchera à rendre minimale la somme des carrés s des écarts entre **X** et (**ZA** + **T**), ce qui revient à rendre minimal (si les tableaux **X** et **Z** sont centrés en colonnes), sans contrainte sur la matrice **A**, le critère :

$$s = trace\ (\mathbf{X - ZA})'(\mathbf{X - ZA})$$

On trouve après un calcul analogue à celui du calcul des coefficients de régression multiple[2] :

$$\mathbf{A} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

C'est la matrice des coefficients d'une régression simultanée, qui revient à effectuer séparément p régressions indépendantes pour chacune des p colonnes de **X**. Dans ce cas, une analyse des résidus **X - ZA** (analyse partielle, cf. section 8.1) nous renseignera sur les éventuels traits structuraux de **X** non-expliqués par **Z**. Notons que l'analyse procrustéenne sans contrainte fait intervenir de façon dissymétrique les tableaux **X** et **Z**.

Il existe de nombreuses autres variantes des analyses procrustéennes (impliquant des dilatations, des axes obliques) pour lesquelles on pourra consulter les références citées.

---

[1] C'est l'approche initiale de Hurley et Cattell (1962) qui sont d'ailleurs à l'origine du nom de cette méthode.

[2] Le problème a été résolu au § 8.1.2.

c – Formulaire de quelques méthodes d'analyse impliquant deux groupes de variables

Récapitulons quelques unes des méthodes d'analyse de tableaux du type $\mathbf{R} = (\mathbf{X}, \mathbf{Z})$, en donnant le formulaire des matrices à diagonaliser ou des matrices de coefficients :

$(\mathbf{X'X})^{-1}\,\mathbf{X'Z}\,(\mathbf{Z'Z})^{-1}\,\mathbf{Z'X}$	ou	$(\mathbf{Z'Z})^{-1}\,\mathbf{Z'X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'Z}$	*(Analyse canonique)*
$\mathbf{X'Z}\,(\mathbf{Z'Z})^{-1}\,\mathbf{Z'X}$	et	$\mathbf{Z'X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'Z}$	*(Analyses projetées)*
$\mathbf{X'}(\mathbf{I} - \mathbf{Z}\,(\mathbf{Z'Z})^{-1}\,\mathbf{Z'})\,\mathbf{X}$	et	$\mathbf{Z'}\,(\mathbf{I} - \mathbf{X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'})\,\mathbf{Z}$	*(Analyses partielles)*
$\mathbf{X'Z\ Z'X}$	ou	$\mathbf{Z'X\ X'Z}$	*(Analyse procrustéenne orthogonale)*
$(\mathbf{Z'Z})^{-1}\,\mathbf{Z'X}$	et	$(\mathbf{X'X})^{-1}\,\mathbf{X'Z}$	*(Analyse procrustéenne sans contrainte)*

Dans les cas où $\mathbf{X}$ et $\mathbf{Z}$ sont des tableaux de variables numériques, celles-ci sont centrées. Si l'on excepte les cas de l'analyse canonique et de l'analyse procrustéenne sans contrainte (ou régression multiple simultanée), il est en général souhaitable de réduire les variables. Notons également que les analyses projetées et l'analyse procrustéenne orthogonale sont équivalentes à des analyses en composantes principales lorsque $\mathbf{X} = \mathbf{Z}$.

8.3.3 Méthode STATIS

La méthode STATIS[1] a été proposée par l'équipe d'Escoufier (1980, 1985 a)[2] pour permettre l'analyse conjointe de plusieurs tableaux de données. Elle s'applique à des tableaux de mesures dans lesquels un ensemble d'individus est décrit par plusieurs groupes de variables ou pour lequel les mêmes variables sont observées sur plusieurs groupes d'individus. L'objet est de comparer les tableaux, puis de décrire l'éventuelle structure commune aux différents tableaux, enfin d'appréhender les différences entre tableaux. Nous présentons seulement les principes de la méthode STATIS sans entrer dans les détails, renvoyant le lecteur à l'ouvrage de Lavit (1988). Nous nous plaçons dans le cadre de q tableaux de mesures de variables centrées-réduites observées sur les mêmes individus de poids égaux à 1.

a – Notations

On note n le nombre d'individus ; p le nombre total de variables (supposées ici centrées-réduites) tous groupes confondus ; p_k le nombre des variables du $k^{ième}$ groupe ; $\mathbf{X}$ le tableau complet de terme général x_{ij} valeur de l'individu i pour la

[1] Le sigle STATIS signifie "Structuration des Tableaux À Trois Indices de la Statistique.

[2] cf. L'Hermier des Plantes (1976), Caillez et Pagès. (1976).

variable j ; $\mathbf{X}_k$ le sous-tableau de $\mathbf{X}$ associé au groupe k ; q représente le nombre de groupes.

L'individu i correspond à une ligne du tableau $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k, ..., \mathbf{X}_q)$; à cet individu, dit "moyen", sont associés q individus dits "partiels", notés i^k, correspondant aux lignes des divers tableaux $\mathbf{X}_k$.

b – Comparaison globale entre les tableaux : l'interstructure

On s'intéresse ici aux relations entre les q tableaux $\mathbf{X}_k$ d'ordre (n, p_k). On considère les matrices de produits scalaires entre les individus $\mathbf{X}_k\mathbf{X}_k'$ (ou $\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k'$ si l'on introduit une métrique particulière à chaque tableau $\mathbf{Q}_k$, mais dans cet exposé schématique, $\mathbf{Q}_k = \mathbf{I}$) de dimension (n,n) et l'on cherche à décrire les distances entre ces matrices. On considère pour cela chaque matrice $\mathbf{X}_k\mathbf{X}_k'$, notée $\mathbf{W}_k$, comme un point dans l'espace $\mathcal{R}^{n^2}$ obtenu en empilant les colonnes de cette matrice. On définit ainsi un nuage de q points-tableaux dans $\mathcal{R}^{n^2}$ et le tableau associé $\mathbf{W}_{n^2}$ de dimension (n^2,q).

L'analyse générale du tableau $\mathbf{W}_{n^2}$, qui revient à diagonaliser la matrice $\mathbf{S}$ d'ordre (q, q) de terme général $s_{kk'} = trace\ (\mathbf{W}_k\mathbf{W}_{k'})$, permet de représenter les q points-tableaux dans un espace de faible dimension et de comparer globalement les tableaux entre eux. Si tous les tableaux sont voisins, ils seront concentrés près d'un point dans l'espace, et le premier axe joindra l'origine à ce point. On pourrait au contraire voir les tableaux s'échelonner le long de cet axe et mesurer ainsi sur l'axe une sorte d'adéquation du tableau au modèle moyen.

Si le nombre p_k de variables du tableau k n'est pas constant, on a intérêt à normer les termes de $\mathbf{S}$ en analysant la matrice $\hat{\mathbf{S}}$ de terme général $\hat{s}_{kk'}$, qui n'est autre que le coefficient R_v de Robert et Escoufier (1976) :

$$\hat{s}_{kk'} = \frac{trace\ \mathbf{W}_k\mathbf{W}_{k'}}{\sqrt{trace\ \mathbf{W}_k^2\ trace\ \mathbf{W}_{k'}^2}}$$

Remarque :

Dans le cas où l'on dispose d'un ensemble de variables observées sur q groupes d'individus, on considère les matrices de covariances (ou de corrélations si les variables sont réduites) de dimension (p,p). On calculera alors, à partir d'un nuage de q points-tableaux dans l'espace $\mathcal{R}^{p^2}$, le tableau $\mathbf{W}_{p^2}$ de dimension (p^2, q).

c – Le nuage moyen ou compromis : l' intrastructure

On cherche à construire un nuage moyen qui soit un compromis des q nuages correspondants aux tableaux $\mathbf{X}_k$. Le compromis peut être calculé de différentes façons, en fonction de la nature des données et des connaissances a priori. Ce peut être une simple moyenne pondérée $\mathbf{C}_1$ des tableaux $\mathbf{X}_k$, lorsqu'il s'agit par

exemple de l'évolution d'un tableau impliquant les mêmes individus et les mêmes variables :

$$\mathbf{C}_1 = \sum_{k=1}^{q} \alpha_k \mathbf{X}_k$$

Si le nombre des variables p_k varie avec k, le compromis pourra toujours être calculé au niveau des produits scalaires (éventuellement normés) :

$$\mathbf{W}_1 = \sum_{k=1}^{q} \alpha_k \mathbf{X}_k \mathbf{X}_k' = \sum_{k=1}^{q} \alpha_k \mathbf{W}_k$$

Les promoteurs de cette stratégie d'analyse recommandent de prendre comme poids α_k la coordonnée du tableau k sur le premier axe de l'analyse de l'interstructure : un tableau aura ainsi un poids d'autant plus élevé qu'il est représentatif de la tendance moyenne.

L'analyse du compromis revient ensuite à effectuer l'analyse en composantes principales ou l'analyse générale du tableau $\mathbf{C}_1$ ou $\mathbf{W}_1$ selon le cas. Elle permet donc de dégager la structure du nuage des individus commune aux tableaux.

d – Représentation simultanée des nuages partiels : les trajectoires

L'analyse de l'interstructure met en évidence les écarts entre les tableaux. L'intrastructure est décrite par le ou les compromis. Il reste à décrire les écarts par rapport au compromis, au niveau des variables et des individus. Si le tableau compromis est du type $\mathbf{C}_1$, il est aisé de représenter en éléments supplémentaires, à partir des tableaux $\mathbf{X}_k$, les trajectoires d'individus (un individu i est représenté par les q points i_k) et, de façon similaire, les trajectoires de variables.

Dans le cas d'un compromis de type $\mathbf{W}_1$, on peut toujours représenter les trajectoires d'individus (lignes des tableaux $\mathbf{W}_k$).

8.3.4 Analyse factorielle multiple

L'analyse factorielle multiple (Escofier et Pagès, 1983), traite des tableaux dans lesquels un ensemble d'individus est décrit par plusieurs groupes de variables. Les variables peuvent être continues, nominales et même, sous certaines conditions, de type fréquence. Toutefois, à l'intérieur d'un groupe, elles doivent être de même type. Nous nous contentons ici d'esquisser les principales caractéristiques de la méthode, en nous plaçant dans le cas particulier de variables continues centrées-réduites de poids *1*. Nous renvoyons le lecteur désireux d'approfondir l'analyse factorielle multiple à l'ouvrage de Escofier et Pagès (1988), qui traite aussi le cas des tables de contingences multiples (cf. aussi dans ce cas Zarraga et Goitisolo, 2002). Les notations de base sont les mêmes que pour la méthode STATIS.

a – Une analyse en composantes principales pondérée

Le fait de vouloir introduire plusieurs groupes de variables en tant qu'éléments actifs dans une même analyse factorielle impose d'équilibrer leur influence *a priori* dans cette analyse. Une analyse simultanée de plusieurs groupes dont les premiers facteurs seraient engendrés par un seul d'entre eux ne présenterait en effet que peu d'intérêt.

En analyse factorielle multiple, chaque variable du groupe k est pondérée par $1/\sqrt{\lambda_1^k}$ où λ_1^k est la première valeur propre de l'analyse en composantes principales effectuées sur les variables de ce groupe k. A l'intérieur d'un groupe, toutes les variables ont le même poids : la structure de chaque groupe est respectée. Géométriquement, cela revient à rendre égale à 1 l'inertie axiale maximum de chacun des k sous-nuages. Du fait de cette pondération, aucun groupe ne peut engendrer à lui seul le premier axe ; en revanche, un groupe multidimensionnel contribue à un plus grand nombre d'axes qu'un groupe unidimensionnel.

Le principe de l'analyse factorielle multiple repose sur une analyse en composantes principales du tableau complet $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_q)$, les variables étant ainsi pondérées. Cette analyse permet d'équilibrer le rôle des groupes de variables et fournit une représentation des individus et des variables qui s'interprète selon les règles usuelles de l'analyse en composantes principales. Au-delà de cette analyse en composantes principales pondérée, la prise en considération de groupes de variables augmente les possibilités d'interprétation des résultats.

Le $\alpha^{\text{ième}}$ facteur de l'analyse factorielle multiple de $\mathbf{X}$ est noté ψ_α dans $\mathcal{R}^p$ et φ_α dans $\mathcal{R}^n$; il est associé à la valeur propre λ_α ; la $\alpha^{\text{ième}}$ valeur propre de l'analyse en composantes principales séparée de $\mathbf{X}_k$ est notée λ_α^k.

b – Recherche de facteurs communs (intrastructures)

Au groupe de variables k correspond dans $\mathcal{R}^n$ un sous-espace V_k à k dimensions ; un facteur commun est une dimension commune à ces sous-espaces. Cette idée est présente dans les analyses canoniques et *multicanoniques* (cas de plus de deux groupes). Mais ces analyses considèrent chaque nuage k uniquement au travers du sous-espace qu'il engendre, sans prendre en compte la répartition de l'inertie dans ce sous-espace. Comparée à ces méthodes, l'analyse factorielle multiple recherche des facteurs à la fois communs aux groupes de variables et représentant des directions de forte inertie de ces groupes.

Du fait de la pondération des variables, l'analyse factorielle multiple peut être interprétée comme une analyse multicanonique. En effet, dans ce cas l'inertie projetée des variables du groupe k sur la direction $\mathbf{z}$ constitue une mesure de

liaison entre la variable $\mathbf{z}$ et le groupe de variables k. Cette mesure, notée $L(\mathbf{z}, k)$, possède les propriétés suivantes :

- $0 \leq L(\mathbf{z}, k) \leq 1$.
- $L(\mathbf{z}, k) = 0 \Leftrightarrow \mathbf{z}$ est non corrélée avec chaque variable du groupe k.
- $L(\mathbf{z}, k) = 1 \Leftrightarrow \mathbf{z}$ est la première composante principale de k.

Le critère satisfait par la $\alpha^{\text{ième}}$ composante principale (notée $\mathbf{z}_\alpha$) de l'analyse factorielle multiple peut s'écrire, compte tenu des contraintes d'orthogonalité avec les α-1 premières composantes principales :

$$Max\left(\sum_k L(\mathbf{z}_\alpha, k)\right)$$

Du point de vue de ce critère, les composantes principales de l'analyse factorielle multiple composent la suite de variables orthogonales les plus liées aux groupes de variables. Ce sont les facteurs communs à ces groupes.

c – Représentation des groupes de variables (interstructure)

La mise en évidence de facteurs communs est une voie commode pour analyser les liaisons entre groupes de variables. On peut chercher à visualiser globalement ces liaisons par un graphique dans lequel chaque groupe est représenté par un point.

Au groupe de variables k on peut associer, comme dans la méthode STATIS, la matrice $\mathbf{W}_k = \mathbf{X}_k\mathbf{X}_k'$ des produits scalaires entre individus. Toutes ces matrices sont d'ordre (n,n). Ce sont des éléments de l'espace $\mathcal{R}^{n^2}$; ces éléments constituent le nuage des k groupes de variables. L'analyse factorielle multiple fait intervenir d'autres éléments de $\mathcal{R}^{n^2}$: les matrices de produits scalaires associées à chaque composante principale normée $\mathbf{z}_\alpha$; ces éléments, que l'on peut écrire $\mathbf{z}_\alpha\mathbf{z}_\alpha'$ forment une base orthonormée d'un sous-espace de $\mathcal{R}^{n^2}$. C'est sur cette base que l'on projettera les k points-groupes, pour visualiser leurs proximités. Cette représentation possède quelques propriétés remarquables. En particulier la projection de $\mathbf{W}_k = \mathbf{X}_k\mathbf{X}_k'$ sur $\mathbf{z}_\alpha\,\mathbf{z}_\alpha'$ est égale à $L(\mathbf{z}, k)$. Il est ainsi possible d'interpréter axe par axe les proximités entre les points-groupes.

d – Représentations superposées des nuages partiels des groupes actifs (trajectoires)

A chaque groupe de variables est associé un nuage partiel d'individus. La comparaison directe des représentations issues des analyses en composantes principales séparées des $\mathbf{X}_k$ ne répond pas directement à cet objectif car ces analyses, étant effectuées séparément, ne tiennent pas compte d'éventuelles structures communes. Il faudrait en fait une analyse procrustéenne généralisée pour résoudre ce problème.

En analyse factorielle multiple on projette les nuages partiels sur les axes principaux du nuage total. Bien qu'ils n'interviennent pas directement dans la construction des axes, les nuages partiels ne sont pas véritablement considérés comme supplémentaires puisque leurs données sont incluses dans le nuage total analysé. Il en résulte deux propriétés utiles lors de l'interprétation :

$$\psi_{\alpha i} = \frac{1}{p}\sum_{k}\psi_{\alpha i^k}$$

le point "moyen" i est au centre de gravité $\psi_{\alpha i}$ des points "partiels" $\psi_{\alpha i^k}$ qui lui sont homologues.

$$\psi_{\alpha i^k} = \frac{1}{\lambda_1^k}\frac{1}{\sqrt{\lambda_\alpha}}\sum_{j\in J_k} x_{ij}\varphi_{\alpha j}$$

Cette relation n'est autre que la restriction au groupe k de l'une des relations usuelles de transition. L'individu partiel i^k apparaît du côté des variables pour lesquelles il a de fortes valeurs (les x_{ij} sont des valeurs centrées-réduites) et à l'opposé de celles pour lesquelles il a de faibles valeurs.

► ***Cas particuliers***

Lorsque chaque groupe ne comporte qu'une seule variable quantitative, l'analyse factorielle multiple se confond avec une analyse en composantes principales. Lorsque chaque groupe ne comporte qu'une seule variable qualitative, l'analyse factorielle multiple se confond avec une analyse des correspondances multiples.

8.3.5 Analyse canonique généralisée

L'analyse canonique généralisée [1] est une méthode d'analyse de tableaux **X** d'ordre (n, p) qui peuvent s'écrire, comme aux paragraphes précédents, sous la forme :

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_k, \ldots \mathbf{X}_q)$$

On note encore n le nombre d'individus ; p le nombre total de variables, tous groupes confondus ; p_k le nombre des variables du $k^{ième}$ groupe ; q le nombre de groupes.

---

[1] L'analyse canonique généralisée a été présentée dans Horst (1961), où elle figure au troisième rang parmi quatre généralisations possibles de l'analyse canonique. Elle a été reprise ou développée par Carroll (1968) dont le nom est souvent attaché à la méthode, Kettenring (1971), Saporta (1975 a), Masson (1974). Casin et Turlot (1986) ont montré qu'elle pouvait être considérée comme une analyse discriminante particulière, et en déduisent des règles d'interprétation nouvelles. Ballif (1986) a développé sous le nom d'AMDG (Analyse multivariée descriptive généralisée) toute une méthodologie de traitement de données, pour laquelle la notion de variable est remplacée par celle plus large de "SEHO" (sous-espace homogène d'observables) et où l'analyse canonique généralisée joue un rôle central.

L'analyse canonique généralisée contient comme cas particulier une grande partie des méthodes descriptives et explicatives qui ont été présentées.

Si $q = 2$, l'analyse de $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ coïncide avec l'analyse canonique des deux groupes. On a vu qu'à ce titre, elle contient comme cas particulier l'analyse discriminante (cas où l'un des deux blocs est un tableau disjonctif complet) et donc l'analyse des correspondances des tables de contingence ($\mathbf{X}_1$ et $\mathbf{X}_2$ sont tous deux disjonctifs complets).

Toujours si $q = 2$, en tant qu'analyse canonique classique, elle contient également la régression multiple (si par exemple $\mathbf{X}_1$ n'a qu'une seule colonne), et donc l'analyse de la variance et de la covariance ($\mathbf{X}_2$ disjonctif complet ou mixte, après régularisation).

Si $q \geq 2$, et si chaque bloc $\mathbf{X}_k$ est un tableau disjonctif complet, l'analyse canonique généralisée n'est autre que l'analyse des correspondances multiples de $\mathbf{X}$. Enfin, toujours si $q \geq 2$, si chaque bloc $\mathbf{X}_k$ n'est formé que d'une seule colonne ($p_k = 1$ pour tout k), elle n'est autre que l'analyse en composantes principales normée de $\mathbf{X}$.

a – Formulation générale

L'analyse canonique généralisée a déjà été présentée au § 2.1. dans le cas particulier où les blocs sont des tableaux disjonctifs complets. Il convient de donner ici une formulation plus générale, qui puisse englober toutes les méthodes précitées.

Dans l'espace $\mathcal{R}^n$, où les p variables (colonnes de $\mathbf{X}$) sont des points, on désigne par V_k le sous-espace engendré par les colonnes de $\mathbf{X}_k$.

La projection $\mathbf{y}_k$ d'une variable $\mathbf{y}$ quelconque (point de $\mathcal{R}^n$) sur le sous-espace V_k s'écrit, si les colonnes de $\mathbf{X}_k$ sont linéairement indépendantes [1] :

$$\mathbf{y}_k = \mathbf{X}_k (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{y} = \mathbf{P}_k \mathbf{y} \qquad [8.3 - 2]$$

Remarquons que si $\mathcal{R}^n$ était muni d'un produit scalaire associé à une matrice diagonale $\mathbf{M}$, la formule précédente s'écrirait sous la forme plus générale :

$$\mathbf{y}_k = \mathbf{X}_k (\mathbf{X}_k' \mathbf{M} \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{M}\, \mathbf{y} = \mathbf{Q}_k \mathbf{y}$$

où l'opérateur idempotent de projection $\mathbf{Q}_k$ sur V_k vaut[2]:

$$\mathbf{Q}_k = \mathbf{X}_k (\mathbf{X}_k' \mathbf{M} \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{M}$$

Ce cadre plus général alourdirait les notations sans changer la substance de l'exposé, qui se poursuivra donc avec $\mathbf{M} = \mathbf{I}$, comme dans la formule [8.3 - 2].

---

[1] Si les p_k colonnes de $\mathbf{X}_k$ ne sont pas linéairement indépendantes, il suffit de les remplacer par les r_k colonnes de $\mathbf{V}$ correspondant à des valeurs propres non nulles dans la décomposition aux valeurs singulières de $\mathbf{X}$ qui s'écrit : $\mathbf{X} = \mathbf{V}\, \Lambda^{1/2} \mathbf{U}$.

[2] Alors que $\mathbf{P}_k$ est symétrique, l'opérateur-projection $\mathbf{Q}_k$ est $\mathbf{M}$-symétrique, c'est-à-dire que l'on a la relation : $\mathbf{M}\mathbf{Q}_k = \mathbf{Q}_k' \mathbf{M}$.

Le carré du cosinus de $\mathbf{y}$ avec V_k (et donc de $\mathbf{y}$ avec $\mathbf{P}_k\mathbf{y}$) que l'on notera $R^2(\mathbf{y},k)$ s'écrit :

$$R^2(\mathbf{y},k) = \frac{\mathbf{y}'\mathbf{P}_k\mathbf{y}}{\mathbf{y}'\mathbf{y}} = \frac{\mathbf{y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y}}{\mathbf{y}'\mathbf{y}} \qquad [8.3 - 3]$$

On définit le premier axe de l'analyse canonique généralisée comme un vecteur $\mathbf{y}$ tel que la quantité s :

$$s = \sum_{k=1}^{q} R^2(\mathbf{y}, k)$$

soit maximale.

Notons que si les $\mathbf{X}_k$ sont centrés, le coefficient $R^2(\mathbf{y}, k)$ est le carré du coefficient de corrélation multiple $R^2(\mathbf{y}, k)$ entre $\mathbf{y}$ et $\mathbf{X}_k$.

Chaque cosinus carré $R^2(\mathbf{y}, k)$ est une mesure de proximité entre le vecteur $\mathbf{y}$ et le sous-espace V_k engendré par les colonnes de $\mathbf{X}_k$.

La maximisation du critère s fait en sorte que le vecteur $\mathbf{y}$ soit le plus près possible de l'ensemble des groupes de variables .

Il s'agit donc de rendre maximale la somme :

$$s = \sum_{k=1}^{q} \mathbf{y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y}$$

avec la contrainte : $\mathbf{y}'\mathbf{y} = 1$

Le vecteur $\mathbf{y}$ de $\mathcal{R}^n$ sera donc le vecteur propre correspondant à la plus grande valeur propre λ de la matrice $\mathbf{S}$ d'ordre (n, n) :

$$\mathbf{S} = \sum_{k=1}^{q} \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k' \qquad [8.3 - 4]$$

Les axes suivants s'obtiennent en rendant maximal le même critère s, avec la même contrainte de norme, et des contraintes d'orthogonalité par rapport à l'ensemble des axes précédents.

b – Propriétés de l'Analyse Canonique Généralisée

On va montrer successivement que l'analyse canonique (et donc tout l'éventail des méthodes qui en sont des cas particuliers), l'analyse en composantes principales normée et l'analyse des correspondances multiples sont des cas particuliers de l'Analyse canonique généralisée.

- ***Pour $q = 2$, l'analyse canonique généralisée est une analyse canonique classique.***

L'équation donnant $\mathbf{y}$ s'écrit, pour $q = 2$:

$$\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\mathbf{y} + \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\,\mathbf{y} = \lambda\,\mathbf{y} \qquad [8.3 - 5]$$

Posons[1] : $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\mathbf{y} = \mathbf{a}$ et $(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\,\mathbf{y} = \mathbf{b}$.

La relation [8.3 - 5] devient simplement :

$$\mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{b} = \lambda\,\mathbf{y} \qquad [8.3\text{-}6]$$

Prémultiplions ensuite les deux membres de la relation [8.3 - 6] par $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, il reste :

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b} = (\lambda\text{-}1)\mathbf{a} \qquad [8.3 - 7]$$

On obtient de la même façon, en prémultipliant les deux membres de la relation [8.3 -6] par $(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$:

$$(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{a} = (\lambda\text{-}1)\mathbf{b} \qquad [8.3 - 8]$$

On obtient finalement, par substitution :

$$(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b} = (\lambda - 1)^2\mathbf{b}$$

La matrice à diagonaliser n'est autre que celle donnée par la formule [2.1 - 4] du paragraphe 2.1.2. On note également la relation entre valeurs propres : $\beta = \lambda - 1$.

- ***Si $q \geq 2$ et si les blocs ne comportent chacun qu'une colonne (centrée), l'analyse canonique généralisée est une analyse en composantes principales normée.***

Dans ce cas, on a $p_k = 1$ pour tout k, et donc $p = q$. On peut maintenant réécrire la formule [8.3 - 4], les $\mathbf{X}_k$ étant des vecteurs notés $\mathbf{x}_k$:

$$\mathbf{S} = \sum_{k=1}^{q} \mathbf{x}_k(\mathbf{x}_k'\mathbf{x}_k)^{-1}\mathbf{x}_k' = \sum_{k=1}^{q} \frac{1}{n\,s_k^2}\mathbf{x}_k\mathbf{x}_k' \qquad [8.3 - 9]$$

où :

$$s_k^2 = \frac{1}{n}\mathbf{x}_k'\mathbf{x}_k$$

est la variance empirique de la variable k.

Si l'on considère la matrice $\mathbf{T}$ des variables centrées réduites dont la $k^{ième}$ colonne vaut $\mathbf{t}_k = \frac{1}{s_k}\mathbf{x}_k$, la matrice $\mathbf{S}$ prend la forme $\mathbf{S} = \frac{1}{n}\mathbf{T}\mathbf{T}'$.

La relation $\mathbf{S}\mathbf{y} = \lambda\,\mathbf{y}$ s'écrit alors, en posant $\mathbf{T}'\mathbf{y} = \mathbf{u}$ et en prémultipliant ses deux membres par $\mathbf{T}'$:

$$\frac{1}{n}\mathbf{T}'\mathbf{T}\,\mathbf{u} = \lambda\,\mathbf{u}$$

soit finalement :

---

[1] On note que $\mathbf{a}$ et $\mathbf{b}$ sont les vecteurs de coefficients de régression de $\mathbf{y}$ expliqué respectivement par $\mathbf{X}_1$ et $\mathbf{X}_2$.

$$\mathbf{Cu} = \lambda\,\mathbf{u}$$

où **C** est la matrice des corrélations d'ordre (p,p) associée au tableau **X** initial.

Cette présentation a le mérite d'enrichir l'interprétation de l'analyse en composantes principales normée, qui peut être définie comme la recherche d'une variable artificielle (**y**) qui rend maximale la somme de ses corrélations avec toutes les variables actives.

- ***Pour $q \geq 2$, quand les blocs sont des tableaux disjonctifs complets, l'analyse canonique généralisée est une analyse des correspondances multiples.***

Pour retrouver (partiellement) les notations de la section 1.4, changeons les **X** en **Z**. Posons donc $\mathbf{Z} = \mathbf{X}$ et $\mathbf{Z}_k = \mathbf{X}_k$ et posons également : $\mathbf{D}_k = \mathbf{Z}_k'\mathbf{Z}_k$.

$\mathbf{D}_k$ est la matrice diagonale d'ordre (p_k,p_k) correspondant aux marges (sommes des colonnes) du tableau $\mathbf{Z}_k$. Enfin appelons **D** la matrice diagonale d'ordre (p,p) dont les q blocs diagonaux sont les $\mathbf{D}_k$.

L'équation $\mathbf{Sy} = \lambda\,\mathbf{y}$ s'écrit :

$$\sum_{k=1}^{q} \mathbf{Z}_k\,\mathbf{D}_k^{-1}\mathbf{Z}_k'\mathbf{y} = \lambda\,\mathbf{y} \qquad [8.3 - 10]$$

Posons, pour tout entier positif $h \leq q$, $\mathbf{Z}_h'\mathbf{y} = \mathbf{u}_h$, ce qui revient également à écrire $\mathbf{Zy} = \mathbf{u}$, **u** étant un vecteur à p composantes tel que :

$$\mathbf{u}' = (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_h', \ldots, \mathbf{u}_q')$$

Prémultipliant les deux membres de [8.3 - 10] par $\mathbf{Z}_h'\mathbf{y}$, on peut alors écrire, pour h = 1,..,q :

$$\sum_{k=1}^{q} \mathbf{Z}_h'\mathbf{Z}_k\,\mathbf{D}_k^{-1}\mathbf{u}_k = \lambda\,\mathbf{u}_h \qquad [8.3 - 11]$$

Ces q équations ne sont autres qu'une écriture par bloc de la relation matricielle :

$$\mathbf{Z}'\,\mathbf{Z}\,\mathbf{D}^{-1}\mathbf{u} = \lambda\,\mathbf{u}$$

Cette formule est à rapprocher des formules de la section 5.2.1 du chapitre 5, où le paramètre s est ici noté q (nombre de tableaux $\mathbf{X}_k$). Avec les notations du présent paragraphe, l'équation de l'analyse des correspondances multiples s'écrit :

$$\frac{1}{q}\mathbf{Z}'\,\mathbf{Z}\,\mathbf{D}^{-1}\mathbf{u} = \lambda'\,\mathbf{u}$$

d'où : $$\lambda = q\,\lambda'$$

La valeur propre issue de l'analyse canonique généralisée est q fois plus grande que celle issue de l'analyse des correspondances multiples du même tableau global **Z**.

- ***Pour $q \geq 2$ dans le cas général, l'analyse canonique généralisée est une analyse générale du tableau X dans une métrique que l'on peut qualifier de "Mahalanobis par bloc"***

Le raisonnement tenu à propos de l'analyse des correspondances multiples (sous-paragraphe précédent ci-dessus) s'applique dans le cas où $\mathbf{X}_k$ est centré, mais quelconque.

La formule [8.3 - 11] prend alors la forme :

$$\sum_{k=1}^{q} \mathbf{X}_h' \mathbf{X}_k \, (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{u}_k = \lambda \, \mathbf{u}_h \qquad [8.3 - 12]$$

Si l'on appelle **D** la matrice diagonale par bloc d'ordre (p,p) (**D** a q^2 blocs dont q blocs diagonaux) dont le $k^{ième}$ bloc diagonal est :

$$\mathbf{D}_{kk} = (\mathbf{X}_k' \mathbf{X}_k)^{-1}$$

$\mathbf{D}_{kk}$ est la matrice associée à la distance de Mahalanobis interne au groupe k (cf. section 7.2 du chapitre 7).

Les q formules [8.3 - 12] (pour $h = 1, ..., q$), s'écrivent :

$$\mathbf{X}' \mathbf{X} \mathbf{D}^{-1} \mathbf{u} = \lambda \, \mathbf{u}$$

Ce qui établit le résultat annoncé (cf. section 1.3.1 du chapitre 1).

c – Utilisation en pratique de l'analyse canonique généralisée

L'analyse canonique généralisée peut s'utiliser comme analyse de compromis dans des approches de type STATIS ou analyse factorielle multiple. Elle n'utilise cependant que les sous-espaces correspondant à chaque groupe, et non la structure interne des nuages dans ces sous-espaces. Ceci peut entraîner les mêmes difficultés d'interprétation que l'analyse canonique.

La figure 8.3 - 1 (cf. Escofier et Pagès, 1988) met ainsi en évidence une faiblesse classique du coefficient de corrélation multiple. Elle montre deux vecteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ contenus dans le sous-espace V_k, et un vecteur $\mathbf{y}$, proche du sous espace V_k, donc proche de sa propre projection $\mathbf{P}_k\mathbf{y}$ sur V_k. $R^2(\mathbf{y}, k)$ est donc voisin de 1, alors que $\mathbf{y}$ est presque orthogonal à $\mathbf{x}_1$ et $\mathbf{x}_2$.

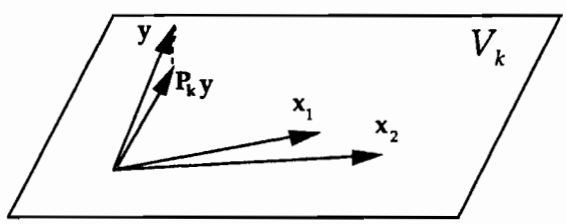


Figure 8.3 – 1. Exemple montrant les insuffisances du coefficient $R^2(\mathbf{y}, \mathbf{k})$

C'est ce type de difficulté qui a conduit ces auteurs à proposer, pour l'analyse globale de $\mathbf{X}$, une métrique diagonale par blocs (le $k^{ième}$ bloc $\mathbf{D}_{kk}$ étant lui-même diagonal et tel que $\mathbf{D}_{kk} = (1/\lambda_k)\,\mathbf{I}$, au lieu de $\mathbf{D}_{kk} = (\mathbf{X}_k'\mathbf{X}_k)^{-1}$ dans le cas de l'analyse canonique généralisée).

Les cas particuliers pour lesquels l'analyse canonique généralisée, dans le cas $q \geq 2$, rejoint des méthodes dont l'interprétation est aisée, sont précisément ceux qui excluent une situation telle que celle de la figure précédente (mauvaise base du sous-espace V_k).

En analyse en composantes principales, les V_k n'ont qu'une dimension, donc $R^2(\mathbf{y}, k)$ est un carré de coefficient de corrélation classique avec la variable $\mathbf{x}_k$ correspondante.

En analyse des correspondances multiples, le codage disjonctif complet fait que chaque $\mathbf{X}_k$ est une base orthogonale du sous-espace V_k correspondant.

On pourrait penser qu'une généralisation qui n'est utile que dans des cas particuliers n'a pas d'intérêt pour le praticien. On peut en fait aménager l'analyse canonique généralisée en la "régularisant" (cf. les sections 3. 3 et 7.4), c'est-à-dire en remplaçant chaque tableau $\mathbf{X}_k$ par le tableau des axes issus d'une analyse en axes principaux de $\mathbf{X}_k$ (qui aura moins de p_k colonnes s'il y a des colinéarités, ou des quasi-colinéarités, c'est-à-dire des valeurs propres faibles).

Ceci rejoint, en d'autres termes, la démarche de Ballif (*op. cit.*) qui conçoit l'analyse canonique généralisée (désignée, on l'a vu, par AMDG) comme une synthèse d'analyses (c'est-à-dire de sous-espaces stables) plutôt que de tableaux. Le principal intérêt de la méthode est alors de pouvoir traiter simultanément des tableaux très hétérogènes[1].

Notons que Escofier (1979 *b*) avait abordé directement ce problème dans un cas particulier en considérant (sans nommer l'analyse canonique généralisée) un tableau mixte $\mathbf{X}$ (qualitatif-quantitatif) contenant deux sortes de blocs : soit des variables continues isolées $\mathbf{x}_k$, soit des tableaux disjonctifs complets. Cet auteur a établi un résultat que l'on peut exprimer de cette façon : en remplaçant chaque colonne $\mathbf{x}_k$ de terme général x_{ik} par un bloc de deux colonnes de termes généraux $(1 - x_{ik})/2$ et $(1 + x_{ik})/2$, il est équivalent de procéder à l'analyse des correspondances de $\mathbf{X}$ ou à l'analyse canonique généralisée de $\mathbf{X}$, formé des nouveaux blocs.

D'autres propriétés de l'analyse canonique généralisée sont présentées dans les articles cités en début de paragraphe.

---

[1] Les blocs V_k formés de plusieurs variables nominales sont prétraités par analyse des correspondances multiples, les blocs formés de plusieurs variables continues par analyse en composantes principales, les blocs formés de tables de contingence par analyse des correspondances simple.

Bibliographie

Agrawal R., Mannila H., Srikant H., Toivonen H., Verkamo A. I., (1995) – Fast discovery of associations rules, IN : *Advances in knowledge Discovery and Data Mining,*. AAAI / MIT Press, Cambridge, Mass. USA.

Agrawal R., Srikant H.(1994) – Fast algorithms for mining associations rules, In : *Proceedings of the 20th International Conference on Very Large Data Bases,*. Santiago, Chile, 487-499..

Agrawala A.K. (Ed.) (1977) - *Machine Recognition of Patterns.* IEEE Press, New York.

Agresti A. (1990) - *Categorical Data Analysis.* J. Wiley, Chichester.

Agresti A. (1992) - A survey of exact inference for contingency tables. *Statistical Science,* 7, 1, p 131-177.

Aitchison J. (1983) - Principal component analysis of compositional data. *Biometrika,* 70, (1), p 57-65.

Aitchison J., Aitken C. G. G. (1976) - Multivariate binary discrimination by the kernel method. *Biometrika,* 63, p 413-420.

Aitkin M. A. (1979) - A simultaneous test procedure for contingency tables. *Appl. Statist.,* 28, p 233-242.

Akaike H. (1973) - Information theory and an extension of the maximum likelihood principle. In : *Second Internat. Symp. on Information Theory,* Petrov B.N., Czaki F., eds., Akademiai Kiado, Budapest, p 267-281.

Aluja Banet T., Lebart L. (1984) - Local and partial principal component analysis and correspondence analysis. In : *COMPSTAT, Proceedings in Computational Statistics,* Physica Verlag, Vienna, p 113-118.

Amari S. (1990) - Mathematical foundation of neurocomputing. *Proc of the IEEE,* 78, 9.

Anastassakos I., D'Aubigny G. (1984) - L'utilisation de tests de sphéricité pour la recherche de la dimension de l'espace latent en analyse factorielle classique et en analyse en composantes principales. *Revue Statist. Appl.,* 32, (2), p 45-57.

Anderberg M.R. (1973) - *Cluster Analysis for Applications.* Academic Press, New York.

Anderson J.A. (1982) - Logistic Discrimination. in : *Handbook of Statistics,* 2, Krishnaiah P.R. and Kanal L. (Eds) North Holland, Amsterdam, p 169-191.

Anderson T. W. (1951) - The asymptotic distribution of certain characteristic roots and vectors. *Proc. of the 2nd Berkeley Symp. on Math. Statist. and Prob.,* p 103-130, Univ. of California Press.

Anderson T. W. (1963) - Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34, p 122-148.

Anderson T. W., Rubin H. (1956) - Statistical Inference in factor analysis. *Proc. of the 3rd Berkeley Symp. on Math. Statist.*, 5, p 111-150.

Anderson T.W. (1958) - *An Introduction to Multivariate Statistical Analysis* (Second edition : 1984). J. Wiley, New York.

Andrews D. F. (1972) - Plots of High-dimensional data. *Biometrics*, 28, p 125-136.

Arabie P. (1978) - Constructing blockmodels : how and why. *J. of Math. Psychology*, 17, (1), p 21-63.

Arabie P. (1991) - Was Euclid an unnecessarily sophisticated psychologist?. *Psychometrika*, 56, p 567-587.

Ardilly P. (ed) (2004) – *Echantillonnage et méthodes d'enquêtes*. Dunod, Paris.

Art D., Gnanadesikan R, Kettenring J.R. (1982) - Data based metrics for cluster analysis. *Utilitas Mathematica*, 21 A, p 75-99.

Atkinson A.C. (1981) - Likelihood ratios, posterior odds and information criteria. *J. Econometrics*, 16, p 15-20.

Atkinson A.C. (1985) - *Plots, Transformation and Regression : an Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford.

Babeau A., Lebart L. (1984) - Les conditions de vie et aspirations des Français. *Futuribles*, 1, p 37-53.

Bailey R.A. (1981) - A unified approach to Design of Experiments. *J. Royal Statist. Soc.* (A), 144(2), p 214-233.

Balbi S. (1994) - *L'Analisi Multidimensionale dei dati negli anni'90*. Dipartimento di Matematica e Statistica. (Univ. Federico II), Rocco Curto Editore, Napoli.

Baldi P., Hornik K. (1989) - Neural networks and principal component analysis : learning from examples without local minima. *Neural Networks*, 2, p 52-58.

Ball G.H., Hall D.J. (1965) - *ISODATA, A Novel Method of Data Analysis and Pattern Classification*. AD 699616, Stanford Research Institute, Menlo Park, California.

Ball G.H., Hall D.J. (1967) A clustering technique for summarizing multivariate data. *Behavioral Sciences* , 12, p 153-155 .

Ballif J.-F. (1986) - *Analyse multivariée : un modèle descriptif général*. Univ. de Lausanne, Peter Lang, Berne.

Bardos M. (1989) - Trois méthodes d'analyse discriminante. *Cahiers Economiques et Monétaires*. 33, p 151-190.

Bardos M. (2001) - *Analyse discriminante*, Dunod, Paris.

Barnett V. (1976) - The ordering of multivariate data. *J. Royal Statist. Soc.* (A), 139, p 318-354.

Bartlett M.S. (1950) - Tests of significance in factor analysis. *British J. Psych. (Stat. Section)*, 3, p 77-85.

Bartlett M.S. (1951) - The effect of standardization on χ^2 approximation in factor analysis (with an appendix by W. Lederman). *Biometrika*, 38, p 337-344.

Bastin C., Benzécri J.-P., Bourgarit C., Cazes P. (1980) – *Pratique de l'analyse des données*. Dunod, Paris.

Bayardo R.J. , Agrawal D. (1999) Mining the most interesting rules. *Proceedings of the 5th In. Conf. on Knowledge Discovery and Data Mining(KDD99)*, San Diego, Cal. 145,154.

Beltrami E. (1873) - Sulle funzioni bilineari. *Giorn. Math. Battaglin.* 11, p 98-106.

Benali H., Escofier B. (1987) - Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et modalités à faibles effectifs. *Revue Statist. Appl.*, 35, n°1, p 41-52.

Benali H., Escofier B. (1990) - Analyse factorielle lissée et analyse des différences locales. *Revue Statist. Appl.* 38, 2, p 55-76.

Benasseni J. (1986a) - Stabilité de l'analyse en composantes principales par rapport à une perturbation des données. *Revue Statist. Appl.*, 35, 3, p 49-64.

Benasseni J. (1986b) - Stabilité en ACP par rapport aux erreurs de mesure. In : *Data Analysis and Informatics*, 4, Diday E. *et al.* (eds), North-Holland, Amsterdam, p 523-533.

Benzécri J.-P. (1969a) - Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, p 35-74.

Benzécri J.-P. (1969b) - Approximation stochastique dans une algèbre normée non commutative. *Bull. Soc. Math. France*, 97, p 225-241.

Benzécri J.-P. (1973) - *L'Analyse des Données.* Tome 1: *La Taxinomie.* Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.

Benzécri J.-P. (1974) - La place de l'*a priori*. In : *Organum- Encyclopaedia Universalis.* Paris.

Benzécri J.-P. (1977 a) - Analyse discriminante et analyse factorielle. *Les Cahiers de l'Analyse des Données*, 4, p 369-406.

Benzécri J.-P. (1977 b) - Choix des unités et des poids dans un tableau en vue d'une analyse des correspondances. *Cahiers de l'Analyse des Données*, 2, p 333-352.

Benzécri J.-P. (1979) - Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 3, p 377-378 .

Benzécri J.-P. (1982 a) - *Histoire et préhistoire de l'analyse des données.* Dunod, Paris.

Benzécri J.-P. (1982 b) - Sur la généralisation du tableau de Burt et son analyse par bandes. *Cahiers de l'Analyse des Données*, 7, p 33-43.

Benzécri J.-P. (1992) - *Correspondence Analysis Handbook.* Marcel Dekker, New York.

Benzécri J.-P., Cazes P. (1978) - Problème sur la classification. *Les Cahiers de l'Analyse des Données*, 3, 1, p 95-101.

Benzécri J.-P., Jambu M. (1976) - Agrégation suivant le saut minimum et arbre de longueur minimum. *Les Cahiers de l'Analyse des Données*, 1, p 441-452.

Benzécri, J.-P. (1982 c) - Construction d'une classification ascendante hiérarchique par la recherche en chaîne de voisins réciproques. *Cahiers d'Analyse des Données*, 7, p 209-218.

Benzécri, J.-P. (1983) - Analyse d'inertie intraclasse par l'analyse d'un tableau de correspondance. *Les Cahiers d'Analyse des Données*, 8, p 351-358.

Benzécri, J.-P., Lebeaux M.-O., and Jambu M. (1980) - Aides a l'interpretation en classification automatique. *Les Cahiers de l'Analyse des Données*, 5, p 101-123.

Beran R., Srivastava M.S. (1985) - Bootstrap test and confidence region for functions of a covariance matrix. *Ann. of Statist.*, 13, p 95-115.

Berge C. (1963) - *Théorie des graphes et ses applications.* Dunod, Paris.

Berge C. (1973) - *Graphs and Hypergraphs.* North Holland, Amsterdam.

Berk R.H. (1972) - Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.*, 43. p 193-204.

Berry M., Linoff G. (1997) - *Data Mining, techniques appliquées au marketing, à la vente et aux services clients*, InterEditions, Masson, Paris.

Bertin J . (1973) - Article : "Graphique (représentation -) ". In : *Encyclopaedia Universalis.*

Bertrand P., Diday E. (1990) - Une généralisation des arbres hiérarchiques : les représentations pyramidales. *Revue Statist. Appl.*, 38, (3), p 53-78.

Besley D. A., Kuh E., Welsh R. E. (1980) - *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*, J. Wiley, New York.

Besse P., Ferré L. (1993) - Sur l'usage de la validation croisée en analyse en composantes principales. *Revue Statist. Appl.*, 41, (1), p 71-76.

Birch M. W. (1963) - Maximum likelihood in three-way contingency tables. *J. Royal Statist. Soc.* (B), 25, p 220-233.

Bishop C. M. (1995) – *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop Y., Fienberg S. E., Holland P. (1975) - *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

Blasius J., Greenacre M., (1998) - *Visualization of Categorical Data*. Academic Press, San Diego.

Blayo F., Verleysen M. (1996) - *Les réseaux de neurones artificiels*. PUF, Paris.

Bock H. H. (1974) - *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung van Daten (Cluster Analysis)*. Vandenhoeck & Ruprecht, Götingen.

Bock H. H. (1977) - On tests concerning the existence of a classification. In : *First International Symposium on Data Analysis and Informatics*. INRIA, Rocquencourt, p 449-464.

Bock H. H. (1979) - Simultaneous clustering of objects and variables. In : *Analyse des données et informatique*. European C.C. Courses, INRIA, p 187-203.

Bock H. H. (1985) - On some significance tests in cluster analysis. *J. of Classification*, 2, p 77-108.

Bock H. H. (1989) - Probabilistic aspects in cluster analysis. In : *Conceptual and numerical analysis of data*. Opitz O. (ed.), Springer-Verlag, Berlin, Heidelberg.

Bock H. H. (1994) - Classification and clustering : Problems for the future. In : *New Approaches in Classification and Data Analysis*, Diday E. et al. (eds), Springer Verlag, Berlin, p 3-24.

Bock H. H., Diday E. (2000) – *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Berlin, Heidelberg.

Boeswillwald E. (1992) - L'expérience du CESP en matière de qualité des mesures d'audience. In : *La qualité de l'information dans les enquêtes*, (ASU), Dunod, Paris, p 313-341.

Bourlard H., Kamp Y. (1988) - Auto-association by multi-layers perceptrons and singular value decomposition. *Biological Cybernetics*, 59, p 291-294.

Bouroche J.-M., Saporta G. (1980) - *L'analyse des données*. coll."Que sais-je", n°1854, PUF, Paris .

Bouroche J.-M., Tenenhaus M. (1970) - Quelques méthodes de segmentation. *RAIRO*, 5, 2, p 29-42.

Bourret P., Reggia J., Samuelides M. (1991) - *Réseaux Neuronaux*. Teknea, Toulouse.

Box G. E. P., Cox D. R., (1982) - An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.*, 77, p 209-210.

Breiman L., Friedman J. H., Ohlsen R. A., Stone C. J. (1984) - *Classification and Regression Trees*. Wadsworth, Belmont.

Brent R.P. (1974) - A gaussian pseudo-random number generator. *Com. ACM*, 17, p 704-706.

Brillouin L. (1959) - *La science et la théorie de l'information*. Masson, Paris.

Brossier G., Dussaix A.-M. (eds) (1999) - *Enquêtes et sondages*. Dunod, Paris.

Bruynooghe M. (1978) - Classification ascendante hiérarchique des grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles. *Les Cahiers de l'Analyse des Données*, 3, p 7-33.

Bry X. (1995) – *Analyses factorielles simples*. Economica, Paris.

Bry X. (1996) – *Analyses factorielles multiples*. Economica, Paris.

Burt C. (1950) - The factorial analysis of qualitative data. *British J. of Statist. psychol.* 3, 3, p 166-185.

Burtschy B., Lebart L. (1991) - Contiguity analysis and projection pursuit. In : *Applied Stochastic Models and Data Analysis*, R. Gutierrez and M.J.M. Valderrama, Eds, World scientific, Singapore, p 117-128.

Cacoullos T. (Ed.) (1973) - *Discriminant Analysis and Applications*. Academic Press, New York.

Caillez F., Pagès J.P. (1976) - *Introduction à l'Analyse des Données*. S.M.A.S.H., Paris.

Callant C. M. (1991) - *Technique de Lissage et de Régularisation en Analyse Discriminante*. Thèse. Université Paris IX, Dauphine, (Publ. INRIA TU177), Paris.

Caraux G. (1984) - Réorganisation et représentation visuelle d'une matrice de données numériques : un algorithme itératif. *Revue de Statist. Appl.*, 32, p 5-24.

Cardoso J.-F. , Comon P. (1996) – Independent component analysis, a survey of some algebraic methods. In: *Proc. ISCAS'96*, vol. 2, p 93-96.

Cardoso J.-F.(1989) – Source separation using higher order moments. In: *Proc. ICASSP'89*, p 2109-2112.

Cardoso J.-F.(1998) – Blind signal separation: statistical principles. In: *Proceedings of the IEEE*, vol. 10, p 2009-2025.

Carlier A. (1985) Analyse des évolutions sur tables de contingences, quelques aspects opérationnels. In : *Data Analysis and Informatics*, Diday E. et al. (eds), North Holland, Amsterdam, p 421-428.

Carlier A., Lavit C., Pagès M., Pernin M.-O., Turlot J.-C. (1988) - A comparative review of methods which handles a set of indexed data tables. In : *Multiway Data Analysis*, Coppi R., Bolasco S. (eds), North Holland, Amsterdam, p 85-102.

Carroll J. D. (1968) - Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psychological Assoc.* p 227-228.

Carroll J. D., Chang J. J. (1970) - Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young' decomposition. *Psychometrika*, 35, p 283-319.

Carroll J. D., Pruzansky S., and Green P. F. (1977) - Estimation of the parameters of Lazarsfeld's Latent Class Model by application of canonical decomposition CANDECOMP to multi-way contingency tables. *AT&T Bell Laboratories*, unpublished paper.

Casin P., Turlot J.-C. (1986) - Une présentation de l'analyse canonique généralisée dans l'espace des individus. *Revue Statist. Appl.* 35, (3), p 65-75.

Cattell R.B. (1966) - The scree test for the number of factors. *Mult. Behavioral Research*, 1, p 245-276.

Caussinus H. (1992) - Projections révélatrices. In : *Modèles pour l'analyse des données multidimensionnelles*. J.J. Droesbeke, B. Fichet, P.Tassi, eds, Economica, Paris.

Caussinus H., Ruiz A. (1990) - Interesting projections of multidimensional data by means of generalized principal component analysis. In :*COMPSTAT 90*, Physica Verlag, Heidelberg, p 121-126.

Cazes P. (1977) - Etude des propriétés extrêmales des sous-facteurs issus d'un sous-tableau d'un tableau de Burt. *Les Cahiers de l'Analyse des Données*, 2, p 143-160.

Cazes P. (1980) - Analyse de certains tableaux rectangulaires décomposés en blocs. *Les Cahiers de l'Analyse des Données*, 5, p 145-161, et p 387-403.

Cazes P. (1981) - Analyse de certains tableaux rectangulaires décomposés en blocs : Codage simultané de variables qualitatives et quantitatives. *Les Cahiers de l'Analyse des Données*, 6, p 9-18.

Cazes P. (1982) - Note sur les éléments supplémentaires en analyse des correspondances. *Les Cahiers de l'Analyse des Données*, 1, p 9-23; 2, p 133-154.

Cazes P. (1984) - Correspondance hiérarchiques et ensembles associés. *Cahiers du B.U.R.O.*, n° 43-44, Université Pierre et Marie Curie, p 43-142.

Cazes P. (1986 a) - Une généralisation des correspondances multiples et des correspondances hiérarchiques. *Cahiers du B.U.R.O.*, 46-47, Université Pierre et Marie Curie, p 37-64.

Cazes P. (1986 b) - Correspondance entre deux ensembles et partition de ces deux ensembles. *Les Cahiers de l'Analyse des Données*, 11, p 335-340.

Cazes P. (1990) - Codage d'une variable continue en vue de l'analyse des correspondances. *Revue Statist. Appl.*, 38, 3, p 35-51.

Cazes P., Chessel D., Doledec S. (1988) - L'analyse des correspondances interne d'un tableau partitionné : son usage en hydrobiologie. *Revue Statist. Appl.* 36, (1), p 39-54.

Cazes P., Moreau J. (1991) - Contingency table in which the rows and columns have a graph structure. In : E.Diday, Y.Lechevallier (Eds) *Symbolic-Numeric Data Analysis and Learning*, Nova Science Publishers. New York, p 271-280.

Celeux G. (1992) - Résultats asymptotiques et validation en classification. In : *Modèles pour l'analyse des données multidimensionnelles*. J.J. Droesbeke, B. Fichet, P.Tassi, eds, Economica, Paris.

Celeux G. (ed) (1990) - *Analyse discriminante sur variables continues*. INRIA, Roquencourt.

Celeux G. et Govaert G. (1992) – A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. and Data Analysis*. Vol. 14, 3, 315 – 332..

Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H.(1989) - *Classification automatique des données: environnement statistique et informatique*. Dunod, Paris.

Celeux G., Diebolt J. (1985) – The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Stat. Quarterly.* 2, 1, 73-82.

Celeux G., Hébrail G., Mkhadri A., Suchard M. (1991). Reduction of a large scale and ill-conditioned statistical problem on textual data. In : *Applied Stochastic Models and Data Analysis, Proceedings of the 5th Symposium.* Gutierrez R. and Valderrama M.J. Eds,World Scientific, p 129-137.

Celeux G., Nakache J.-P. (eds) (1994) - *Analyse discriminante sur variables qualitatives.* Polytechnica, Paris.

Chabanon C., Dubuisson B. (1991) - Méthodes non probabilistes. In : *Analyse discriminante sur variables continues,* Celeux G. (ed.), INRIA, Paris.

Chandon J.-L., Pinson S. (1981) - *Analyse typologique : Théorie et applications.* Masson, Paris.

Chateau F. (1994) - Probabilités a priori inégales dans la règle des k plus proches voisins. *Actes des XXVIèmes Journées de Statistiques* (Neuchâtel), p 195-198.

Chateau F., Lebart L. (1996) - Assessing sample variabilitiy in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. In : *COMPSTAT96,* A. Prats (ed), Physica Verlag, Heidelberg, p 205-210.

Chatterjee S., Price B. (1991) - *Regression Analysis by Examples.* J. Wiley, New York.

Cheng B., Titterington D.M. (1994) - Neural networks: a review from a statistical perspective. *Statistical Science,* 9, n°1, p 2-54.

Chernoff H. (1973) - The use of faces to represent points in *k*-dimensional space graphically. *J. Amer. Statist. Assoc.,* 68, p 361-368.

Chessel D., Lebreton J.-D., Yoccoz N. (1987) - Propriétés de l'analyse canonique de correspondances ; une illustration en hydrobiologie. *Revue de Statist. Appl.* , 35, (4), p 55-72.

Choudary Hanumara R., Thompson W.A. (1968) - Percentage points of the extreme roots of a Wishart matrix. *Biometrika,* 55, p 505-512.

Christensen R. (1990) - *Log-Linear Models.* Springer-Verlag, New York.

Chung F.R.K., *Spectral Graph Theory.* CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, 1997.

Clemm D.S., Krishnaiah P.R., Waikar V.B. (1973) - Tables of the extreme roots of a Wishart matrix. *J. of Statist. Comput. and Simul.* 2, p 65-92.

Cliff A.D. and Ord J.K. (1981) - *Spatial Processes : Models and Applications.* Pion, London.

Cliff N. (1966) - Orthogonal rotation to congruence. *Psychometrika,* 31, p 33-42.

Cochran W.G., Cox G.M. (1957) - *Experimental Design* (2nd ed.). J. Wiley, New York.

Cohen A. (1980) - On the graphical display of the significant components in two-ways contingency tables. *Comm. in Statistics, Theory.Meth.,* A9 (10), p 1025-1041.

Cohen J. (1967) - *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, New York.

Comon P. (1994) – Independent Component Analysis : a new concept ? *Signal Processing,* vol. 36, p 287-314

Cook R.D., Weisberg S. (1982) - *Residuals and Influence in Regression.* Chapman and Hall, London.

Cook R.D., Weisberg S. (1994) - *An Introduction to Regression Graphics.* J. Wiley, New York.

Coppi R., Bolasco S. (eds) (1989) - *The Analysis of Multiway Data Matrices.* North Holland, Amsterdam.

Cormack R.M. (1971) - A review of classification. *J. of Royal Statist. Society,* Serie A, 134, Part. 3, p 321-367.

Cornejuols A. (2002) – Une nouvelle méthode d'apprentissage: Les SVM. Séparateurs à vastes marges. *Bulletin de l'AFIA,* vol. 51, p 14-23.

Cornejuols A., Miclet L. (2002) – *L'apprentissage artificiel, Méthodes et concepts.* Eyrolles, Paris.

Cornfield J. (1962) - Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function approach. *Fed. Amer. Socs. Exper. Biol. Proc. Suppl.,* 11, p 58-61.

Corsten L. C. A. (1976) - Matrix approximation, a key to application of multivariate methods. In : *Proc. 9th Int. Biometric Conf.,* 1, p 61-77, Raleigh, North Carolina.

Cottrell M., Fort J.-C. (1987) - Etude d'un algorithme d'auto-organisation. *Ann. de l'Inst. Henri Poincaré,* 23, p 1-20.

Cottrell M., Rousset P. (1997) - The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data, *in: Biological and Artificial Computation : From Neuroscience to Technology,* J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), Springer, p. 861-871.

Cottrell M., Ibbou S., Letrémy P., Rousset P. (2003) – Cartes auto organisées pou r l'analyse exploratoire de données et la visuallisation, *Journal de la Soc. Française de Stat. vol. 144, 4, p 67 - 106.*

Cox D. R. (1972) - *Analyse des données binaires.* Dunod, Paris.

Cox D. R. (1977) - The role of significance tests. *Scandinavian Journal of Statist.,* 4, p 49-70.

Cox D.R. (1958) - *Planning of Experiments.* J. Wiley, New York.

Craddock J.M., Flood C.R. (1970) - The distribution of the □2 statistic in small contingency tables. *Appl. Statist.,* 19, p 173-181.

Cramer H. (1946) - *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Cristianini N., Shawe-Taylor J. (2000) - *Support Vector Machines, and Other Kernel Methods,* Cambridge U. Press, Cambridge.

Critchley F. (1985) - Influence in principal component analysis. *Biometrika,* 72, p 626-636.

Dagnelie P. (1981) - *Principes d'expérimentation.* Les Presse Agronomiques de Gembloux, Gembloux.

Darmois G. (1957) - *Statistique et applications.* Armand Colin, Paris.

Darroch J. N., Lauritzen S. L., Speed T. P. (1980) - Markov field and log-linear interaction models for contingency tables. *Ann. of Statist.,* 8, 522-539.

Daudin J.-J., Duby C., Trécourt P. (1988) - Stability of principal components studied by the bootstrap method. *Statistics,* 19, p 241-258.

Daudin J.-J., Trécourt P. (1980) - Analyse factorielle des correspondances et modèle log-linéaire : Comparaison des deux méthodes sur un exemple. *Revue Statist. Appl.* 28, n° 1, p 5-24.

Davis A. W. (1977) - Asymptotic theory for principal component analysis : the non-normal case. *Australian J. of Statist.,* 19, p 206-212.

Davis C., Kahan W. M. (1970) - The rotation of eigenvectors by a perturbation. *Journal of SIAM (Numerical Analysis)*, 7, p 1-46.

Day N. E. (1969) Estimating the component of a mixture of normal distribution. *Biometrika*, 56, p 463-474.

Delecroix M. (1983) - *Histogrammes et estimation de densité*. P.U.F., Paris.

Deming W. E., Stephan F. F. (1940) - On a least squares adjustment of a sampled frequency table when the expected marginal total are known. *Ann. Math. Statist.*, 11, p 427-444.

Dempster A.P. (1971) - An overview of multivariate data analysis. *J. Mult. Analysis*, 1, p 316-346.

Deroo M., Dussaix A.-M. (1980) - *Pratique et analyse des enquêtes par sondage*. P.U.F., Paris.

Devaud J.-M. (1985) - Discrimination et description sur variables qualitatives : un exemple comparatif sur données réelles. *Revue Statist. Appl.* 33, n° 2, p 5-18.

Devijver P., Kittler J. (1982) - *Pattern Recognition : A Statistical Approach*. Prentice Hall, New York.

Deville J.-C., Malinvaud E. (1983) - Data analysis in official socio-economic statistics. *J. Royal Statist. Soc.*, A, 146, part 4.

Deville J.-C., Särndal C.-E. (1992) - Calibration estimator in Survey Sampling. *J.A.S.A.*, 87, 418, p 376-382.

Diaconis P., Efron B. (1983) - Computer intensive methods in statistics. *Scientific American*, 248, (May), p 116-130.

Diday E. (1971) - La méthode des nuées dynamiques. *Revue Statist. Appl.* 19, n° 2, p 19-34.

Diday E. (1972) - Optimisation en classification automatique et reconnaissance des formes. *Revue Française de Recherche Opérationnelle*, 3, p 61-96.

Diday E. (1974) - Classification automatique séquentielle pour grands tableaux. *Revue Fr. Inf. Rech. Opér.* 9, (Mars 1975), p 1-29.

Diday E. (1992) - From data to knowledge : Probabilist objects for a symbolic data analysis. In : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 193-214.

Diday E. , Lemaire J.L., Pouget J., Testu F. (1982) - *Eléments d'Analyse des Données*. Dunod, Paris.

Dobson A. (1983) - *An Introduction to Statistical Modelling*. Chapman and Hall, New York.

Dodge Y. (ed.) (1987) - *Statistical Data Analysis Based on the L1-Norm and Related Methods*. North Holland, Amsterdam.

Domenges D., Volle M. (1979) - Analyse factorielle sphérique: Une exploration. *Annales de l'INSEE*, n° 35.

Draper N. R., Smith H. (1981) - *Applied Regression Analysis (2nd ed)*. J. Wiley, New York.

Droesbeke J.-J., Fichet B., Tassi P. (ed.) (1987) - *Les sondages*. Economica, Paris.

Droesbeke J.-J., Fichet B., Tassi P. (ed.) (1992) - *Modèles pour l'analyse des données multidimensionnelles*. Economica, Paris.

Droesbeke J.-J., Fine J., Saporta G. (eds) (2002) *Méthodes bayésienne en statistique*, Technip, Paris.

Droesbeke J.-J., Lebart L. (eds) (2001) *Enquêtes, modèles et applications*, Dunod, Paris.

Droesbeke J.-J., Lejeune M., Saporta G. (eds.) (2005) - *Modèles statistiques pour données qualitatives*, Editions. Technip, Paris.

Droesbeke J.-J., Tassi P. (1990) - *Histoire de la statistique*. Que-sais-je? PUF, Paris.

Drouet d'Aubigny G. (1993) - Analyse des proximités et programmes de codage multidimensionnel. *La Revue de Modulad*, INRIA, Rocquencourt, 12, p 1-32.

Dubes R. C., Zeng G. (1987) - A test for spatial homogeneity in cluster analysis. *J. of Classification*, 4, p 33-56.

Dubes R. C., Jain A. K. (1979) - Validity studies in clustering methodology. *Pattern Recognition*, 11, p 235-254.

Dubuisson B. (1990) - *Diagnostic et reconnaissance des formes*. Hermès, Paris.

Duda R.O., Hart P.E. (1973) - *Pattern Classification and Scene Analysis*. J. Wiley, New York.

Dugué D. (1958) - *Traité de statistique théorique et appliquée*. Masson, Paris.

Eastment H. T., Krzanowski W., J. (1982) - Cross validatory choice of the number of components of a principal component analysis. *Technometrics*, 24, p 73-77.

Eckart C., Young G. (1936) - The approximation of one matrix by another of lower rank. *Psychometrika*, l, p 211-218.

Eckart C., Young G. (1939) - A principal axis transformation for non- hermitian matrices. *Bull. Amer. Math. Assoc.*, 45, p 118-121.

Edwards A. W. F., Cavalli-Sforza L. L. (1965) - A method for cluster analysis. *Biometrics*, 21, p 362-375.

Efron B. (1965) - The convex hull of a random set of points. *Biometrika*, 52, p 331-343.

Efron B. (1979) - Bootstrap methods : another look at the Jackknife. *Ann. Statist.*, 7, p 1-26.

Efron B., Tibshirani R. J. (1993) - *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Efron B. (1982) - *The Jacknife, the Bootstrap et other Resampling Plans*. SIAM, Philadelphia.

Engelman L., Hartigan J. A. (1969) - Percentage points of a test for clusters. *J. Amer. Statist. Assoc.*, 64, p 1647-1648.

Escofier B. (1978) - Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statist. Appl.*, 26, p 29-37.

Escofier B. (1979 a) - *Stabilité et approximation en analyse factorielle*. Thèse d'Etat, Université Pierre et Marie Curie, Paris.

Escofier B. (1979 b) - Traitement simultané de variables qualitatives et quantitatives. *Les Cahiers de l'Analyse des Données*, 4, (2), p 137-146.

Escofier B. (1979 c) - Une représentation des variables dans l'analyse des correspondances multiples. *Revue de Statist. Appl.*, 27, p 37-47.

Escofier B. (1984) - Analyse factorielle en référence à un modèle: application à l'analyse des tableaux d'échanges. *Revue de Statist. Appl.*, 32, 25-36.

Escofier B. (1987) - Analyse des correspondances multiples conditionnelles. In : *Data Analysis and Informatics*, Diday E. (ed.), North Holland, Amsterdam, p 13-22.

Escofier B. (1989) - Multiple correspondence analysis and neighboring relation. In : *Data Analysis, Learning Symbolic and Numeric Knowledge*. Diday E. (ed.), Nova Science Publishers, New York, p 55-62.

Escofier B. (2003) - *Analyse des correspondances*, Presses Univ. de Rennes, Rennes.

Escofier B. [Cordier B.] (1965) - *l'Analyse des correspondances.* Thèse, Faculté des Sciences de Rennes ; publiée en 1969 dans les *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, n°13.

Escofier B., Le Roux B. (1972) - Etude de trois problèmes de stabilité en analyse factorielle. *Publication de l'Institut Statistique de l'Université de Paris*, 11, p 1-48

Escofier B., Pagès J. (1983) - Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractéristation des vins rouges du Val de Loire. *Revue Statist. Appl.* 31, p 43-59.

Escofier B., Pagès J. (1984) - Analyses factorielles simples et multiples. *Cahiers du BURO*, 2, ISUP, Paris.

Escofier B., Pagès J. (1988) - *Analyses factorielles multiples.* Dunod, Paris.

Escoufier Y. (1970) - Echantillonnage dans une population de variables aléatoires réelles. *Publication de l'Institut Statistique de l'Université de Paris*, 19, Fasc 4, p 1-47.

Escoufier Y. (1980) - L'analyse conjointe de plusieurs matrices de données. In: *Biométrie et Temps*, Jolivet et al. (eds), p 59-76.

Escoufier Y. (1982) - L'Analyse des correspondances simples et multiples. *Metron*, 1-2, p 53-78.

Escoufier Y. (1985 a) - Objectifs et procédures de l'analyse conjointe de plusieurs tableaux. *Statist. et Anal. des Données.* 10, (1), p 1-10.

Escoufier Y. (1985 b) - L'Analyse des correspondances, ses propriétés, ses extensions. *Bull. of the Int. Statist. Inst.*, 4, p 28-2.

Escoufier, Y. (1988) - Beyond correspondence analysis. In: *Classification and Related Methods of Data Analysis*, H.H.Bock , Ed., North Holland, p 505-514.

Everitt B. S., Hand D. J. (1981) - *Finite Mixture Distributions.* Chapman and Hall, London.

Falguerolles (de) A., Jmel S. (1993) - Un modèle graphique pour la sélection de variables qualitatives. *Revue de Statist. Appl.* 41, p 23-41.

Falissard B. (1995) - Déploiement d'une matrice de corrélation sur la sphère unité de R^3. *Revue de Statist. Appl.*, 43, (2) p 35-48.

Falissard B. (1996) - *Comprendre et utiliser les statistiques dans les sciences de la vie.* Masson, Paris.

Faraj A. (1993) - Analyse de contiguïté : une analyse discriminante généralisée à plusieurs variables qualitatives. *Revue Statist. Appl.* 41, (3), p 73-84.

Farebrother R. W. (1987) - The historical development of the L_1 and L_0 estimation procedures. in: *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Dodge Y. ed., North Holland, Amsterdam, p 37-64.

Fichet B. (1987) - The role played by L_1 in data analysis. in: *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Dodge Y. ed., North Holland, Amsterdam, p 185-194.

Fichet B. (1988) - L_p space in Data Analysis. In : Classification and Related Methods of Data analysis. Boch H. H. (ed.), North-Holland, Amsterdam, p 439-444.

Fienberg S.E. (1980) - *The Analysis of Cross-classified Categorical Data.* MIT Press, Cambridge, Mass.

Fine J. (1992) - Modèles graphiques d'associations. In : ASU, (Droesbeke J.-J., Fichet B., Tassi P., ed.), *Modèles pour l'analyse des données multidimensionnelle*, Economica, Paris.

Fine J. (1993) - Problèmes d'indétermination en analyse en facteurs et analyse en composantes principales optimale. *Revue de Statist. Appl.*, 41, (4), p 45-72.

Fisher R. A. (1915) - Frequency distribution of the value of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, p 507-521.

Fisher R. A. (1935) - *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fisher R. A. (1936) - The use of multiple measurements in taxonomic problems. *Ann. of Eugenics*, 7, p 179-188.

Fisher R.A. (1939) - The sampling distribution of some statistics obtained from non linear equations. *Ann. Eugen.*, 7, p 179-188.

Fisher R.A. (1940) - The precision of discriminant functions. *Ann. Eugen.*, 10, p 422-429.

Fisher R.A., Yates F. (1949) - *Statistical Tables for Biological, Agricultural and Medical Research*. Hafner Publishing Company.

Fisher W.D. (1958) - On grouping for maximum homogeneity. *J. of Amer. Statist. Assoc.*, 53, p 789-798.

Fix E., Hodges J. L. (1951) - Discriminatory analysis - nonparametric discrimination : consistency properties. Report of the U.S.A.F. School of Aviation Medicine. In : Agrawala (1977).

Florek K. (1951) - Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, 2, p 282-285.

Flury B. (1988) - Common principal components and related multivariate models. J. Wiley, New York.

Forgy E. W. (1965) - Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometric Society Meetings*, Riverside, California (Abstract in : *Biometrics* 21, 3, p 768).

Fortin M. (1975) - Sur un algorithme pour l'analyse des données et la reconnaissance des formes. *Revue de Statist. appl.*, 23, p 37-46.

Fourgeaud C., Lenclud B. (1978) - *Econométrie*. P.U.F., Paris.

Francisco C. A., Finch M. D. (1980) - A comparison of methods used for determining the number of factors to retain in factor analysis. *Technometrics*, 22, p 105-110.

Freitas A. A. (1998) – On objective measures of rule surprisingness. *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD98)*, Nantes, 1-9.

Friedman J. H. (1989) - Regularized discriminant analysis. *J. of Amer. Statist. Assoc.*, 84, p 165-175.

Friedman J. H.(1987) - Exploratory projection pursuit. *J. of Amer. Statist. Assoc.*, 82, (397), p 249-266.

Friedman J. H., and Tukey J.W. (1974) - A Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, Ser. C, 23, p 881-889.

Fukunaga K. (1972) - *Statistical Pattern Recognition*. Academic Press, Boston.

Furnas G. W., Deerwester S., Dumais S. T., Landauer T. K., Harshman R. A., Streeter L. A., Lochbaum K. E. (1988) - Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. In : Information Retrieval*, p 465-480.

Furnival G. M. (1971) - All possible regressions with less computation, *Technometrics*, 13, p 403-408.

Furnival G. M., Wilson R.W. (1974) - Regressions by leaps and bounds, *Technometrics*, 16, p 499-511.

Gabriel K.R. (1969) - Simultaneous test procedures: some theory of multiple comparisons. *Ann. Math. Statist.*, 40, 1, p 224-250.

Gabriel K.R. (1971) - The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 3, p 453-467.

Gallego F. J. (1982) - Codage flou en analyse des correspondances, *Les Cahiers de l'Analyse des Données*, 7, n° 4, p 413-430.

Gallinari P., Thiria S., Fogelman-Soulié F. (1988) - Multilayer perceptrons and data analysis, *International Conference on neural Networks, IEEE,*, 1, p 391-399.

Garnett J.-C. (1919 - General ability, cleverness and purpose. *British J. of Psych.*, 9, p 345-366.

Geary R.C. (1954) - The contiguity ratio and statistical mapping.*The Incorporated Statistician*, 5, 3, p 115-145.

Geisser S. (1975) - The Predictive sample reuse method with applications. *J. of Amer. Statist. Assoc.* 70, p 320-328.

Gifi A. (1981) - *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden.

Gifi A. (1990) - *Non Linear Multivariate Analysis*, J. Wiley, Chichester.

Gilula Z., Ritov Y. (1990) - Inferential ordinal correspondence analysis : motivation, derivation and limitations. *Inter. Statist. Review*, 58, p 99-108.

Gilula, Z. (1986) - Grouping and association in contingency tables: an exploratory canonical correlation approach, *J. of Amer. Statist. Assoc.*, 81, p 773-779.

Girshick M.A. (1939) - On the sampling theory of roots of determinantal equations. *Ann. Math . Statist.*, 1, 10, p 203-224.

Gnanadesikan R. (1989) - Discriminant analysis and clustering, panel of experts. *Statistical Science*, 4, n°1, p 34-69.

Gnanadesikan R., Kettenring J.R., Landwehr J.M. (1982) - Projection plots for displaying clusters, In : *Statistics and Probability, Essays in Honor of C.R. Rao*, G. Kallianpur, P.R. Krishnaiah, J.K.Ghosh, eds, North-Holland.

Goldstein M., Dillon W. R. (1978) - *Discrete Discriminant Analysis*, J. Wiley, Chichester.

Good P. (1994) - *PermutationTest - A practical Guide to Resampling Method for Testing Hypotheses*. Springer Verlag, New York.

Goodman L.A. (1970) - The multivariate analysis of qualitative data: interaction among multiple classifications. *J. of Amer. Statist. Assoc.*, 65, p 226-256.

Goodman L.A. (1986) - Some useful extensions of the usual correspondence analysis approach and the usual log-linear approach in the analysis of contingency tables, *International Statist. Review*, 54, p 243-270.

Goodman L.A. (1991) - Measures, models, and graphical displays in the analysis of cross-classified data (with Discussion), *J. of Amer. Statist. Assoc.*, 86, 416, p 1085-1138.

Goodman L.A., Kruskal W.H. (1954) - Measures of association for cross classification. *J. of Amer. Statist. Assoc.*, 49, p 732-764.

Gordon A. D. (1979) - On the assessment and comparison of classification. In : *Analyse des données et informatique. Cours de la C.E.E.*, Tomassone R. (ed.), INRIA, Rocquencourt. p 149-160.

Gordon A. D. (1981) - *Classification : Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, London.

Gordon A.D. (1987) - A review of hierarchical classification, *J.R.Statist.Soc.*, A, 150, Part2, p 119-137.

Gordon A.D., Finden C.R. (1985) - Classification of spatially located data. *Comp. Statist. Quarterly*. 2, p 315-328.

Govaert G. (1977) - Algorithme de classification d'un tableau de contingence. In: *Premières Journées Internationales Analyse des Données et Informatique (Versailles 1977)* INRIA, p 487-500.

Govaert G. (2003) - *Analyse des données*, Hermès – Lavoisier, Paris.

Govaert, G. (1984) - Classification simultanée de tableaux binaires.- In: *Data Analysis and Informatics, 4*, E. Diday *et al.*, Eds, North Holland, p 223-236.

Gower J. C. (1966) - Some distance properties of latent and vector methods used in multivariate analysis. *Biometrika*, 53, p 325-328.

Gower J. C. (1968) - Adding a point to vector diagram in multivariate analysis. *Biometrika*, 55, 582-585.

Gower J. C. (1975) - Generalized Procrustes Analysis. *Psychometrika*, 40, (1), p 33-51.

Gower J. C. (1984) - Procrustes analysis. In : *Handbook of Applicable Mathematics*, 6, Lloyd E.H. (ed.), J. Wiley, Chichester, p 397-405.

Gower J. C., Banfield C. F. (1975) - Goodness-of-fit criteria in cluster analysis and their empirical distributions. In : *Proceeding of the 8th Intern. Biometric Conf.*, Corsten L. C. A., Postelnicu T., (eds), p 347-361.

Gower J. C., Dijksterhuis G. B. (2004) - *Procrustes Problems*, Oxford Univ. Press, Oxford.

Gower J. C., Harding A. (1988) - Nonlinear biplot. *Biometrika*, 75, p 445-455.

Gower J. C., Ross G. (1969) - Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics*, 18, p 54-64.

Graham R. L. and Hell P. (1985) – On the history of the minimum spanning tree problem. *Ann. Hist. Comput.* 7, 43-57.

Gras R. et Larher A., (1992), L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématique, Informatique et Sciences Humaines*, E.H.E.S.S. Paris, n°120, p 5-31

Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., (2004) : Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Editions*, p 3-32.

Gras R., Kuntz P. et Briand H. (2003), Hiérarchie orientée de règles généralisées en analyse implicative, *Extraction des Connaissances et apprentissage, Hermès*, p 145-157.

Gras R., Kuntz P. et Briand H., (2001), Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, p 9-29.

Green P. J. (1981) - Peeling bivariate data. In: *Interpreting multivariate data*, Barnett V. (ed.), J. Wiley, Chichester, p 3-20.

Greenacre M. (1984) - *Theory and Applications of Correspondence Analysis*. Academic press, London.

Greenacre M. (1988) - Clustering the rows and columns of a contingency table, *J. of Classification*, 5, p 39-51.

Greenacre M. (1993) - *Correspondence Analysis in Practice.* Academic Press, London.

Greenacre M., Blasius J. (Eds) (1994) - *Correspondence Analysis in the Social Sciences.* Academic Press, London.

Grelet Y. (1993) - Préparation des tableaux pour l'analyse des données : le codage des variables. In : *Traitement statistique des enquêtes*, Grangé D., Lebart L. (eds), Dunod, Paris.

Grizzle J. E., Starner C. F., Koch G. G. (1969) - Analysis of categorical data by linear models. *Biometrics*, 25, p 489-504.

Grosbras, J.-M. (1986) - *Méthodes statistiques des sondages.* Economica, Paris.

Guéguen A., Nakache J.-P. (1988) - Méthode de discrimination basée sur la construction d'un arbre de décision binaire. *Revue de Statist. Appl.*, 36, (1), p 19-38.

Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 251 -264, SSCR New York.

Guttman L. (1954) - Some necessary conditions for common factor analysis. *Psychometrika*, 19, p 149-161.

Haberman S. J. (1974) - *The Analysis of Frenquency Data.* University of Chicago University Press, Chicago.

Hand D. (1998) – Data Mining reaching statistics. *Research in Official Statistics*, vol. 2, 5-17.

Hand D. J. (1981) - *Discrimination and Classification.* J. Wiley, New York.

Hand D. J. (1982) - *Kernel Discriminant Analysis.* J. Wiley, New York.

Hand D. J. (1987) - A shrunken leaving-one-out estimator of error rate. *Comput. Math. Applic.*, 14, (3), p 161-167.

Hand D., J. (1986) - Recent advances in error-rate estimation. *Pattern Recogn. lett.*, 4, p 335-346.

Hand D. J. (1992) - Microdata, macrodata, metadata. In : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, 2, p 325-340.

Hardy A. (1994) - An examination of procedures for determining the number of clusters in a data set. In : *New Approaches in Classification and Data Analysis*, Diday E. et al. (eds), Springer Verlag, Berlin, p 178-185.

Harman H.H. (1967) - *Modern Factor Analysis* (2nd ed.). Chicago University Press, Chicago.

Harshman R. A. (1970) - Foundation of the PARAFAC procedure : Models and conditions for an explanatory multi-modal factor analysis. *UCLA working paper in Phonetics*, 16, UCLA, Los Angeles.

Harter H.L. (1974-1975) - The method of least squares and some alternatives. *Internat. Statist. Review*, Part 1 and 2: 42, p 147-174, p 235-264; Part 3 to 5: 43, p 1-44, p 125-190, p 269-278.

Hartigan J. A. (1972) - Direct clustering of a data matrix, *J. of Amer. Statist. Assoc.*, 67, p 123-129.

Hartigan J. A. (1975) *Clustering Algorithms.* J. Wiley, New York.

Hartigan J. A. (1985) - Statistical theory in clustering. *J. of Classification*, 2, 63-76.

Hastie T., Tibshirani R., Friedman J. (2001) - *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York.

Hayashi C.(1956) - Theory and examples of quantification. (II) *Proc. of the Institute of Statist. Math.* 4 (2), p 19-30.

Hayashi C., Hayashi F. (1982) - A new algorithm to solve PARAFAC model. *Behaviormetrika*, 14, p 27-48.

Heiser W. (2005) – Geometric representation of associations between categories. *Psychometrika*, vol. 69, 4, p 513-545.

Heiser W. J. (1986) - Undesired nonlinearities in nonlinear multivariate analysis. In : *Data Analysis and Informatics IV*, Diday E. et al. (eds), North Holland, Amsterdam, p 455-469.

Hérault J., Jutten C. (1994) - *Réseaux neuronaux et traitement du signal.* Hermès. Paris.

Hertz J., Krogh A., Palmer R.G. (1991) - *Introduction to the Theory of Neural Computation.* Addison- Wesley, Reading, (Mass.).

Highleyman W.H. (1962) - The design and analysis of pattern recognition experiments. *Bell Syst. Tech. Journal.*, 41, p 723-744.

Hilderman R.J., Hamilton H.J.(2001) *Knowledge Discovery and Measures of Interestingness.* Kluwer Academic Publisher.

Hill M.O. (1974) - Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 3, p 340-354.

Hirschfeld H.D. (1935) - A Connection between correlation and contingency. *Proc. Camb. Phil. Soc.* 31, p 520-524.

Hochberg, Y. (1988) - A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75, p 800-803.

Holmes S. (1985) - *Outils Informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données.* Thèse USTL, Montpellier.

Holmes S. (1989) - Using the bootstrap and the RV coefficient in the multivariate context. in : *Data Analysis, Learning Symbolic and Numeric Knowledge*, E. Diday (ed.), Nova Science, New York, p 119-132.

Hornik K. (1994) - Neural networks : more than "statistics for amateurs". In : *COMPSTAT94, Dutter R., Grossmann W. (eds), Physica Verlag, Heidelberg, p 223-235.*

Horst P. (1961) - Relation among m sets of measures. *Psychometrika*, 26, p 129-149.

Horst P. (1965) - *Factor Analysis of Data Matrices.* Holt, Rinehart, Winston, New York.

Hosmer D. W., Lemeshow S. (1989) - *Applied Logistc Regression*, J. Wiley, New York.

Hotelling H. (1933) - Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, p 417-441, p 498-520.

Hotelling H. (1936) - Relation between two sets of variables. *Biometrika*, 28, p 129-149.

Householder A.S. (1953) - *Principles of Numerical Analysis.* Mc Graw-Hill, New York.

Hsu P. L. (1939) - On the distribution of the roots of certain determinantal equations. *Ann. Eugen.* 9, p 250-258.

Hsu, J. C. (1996) - *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Huber P.J. (1981) - *Robust Statistics.* J. Wiley, New York.

Huber P.J. (1987) - The place of the L_1-Norm in robust estimation, in: *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Dodge Y. ed., North Holland, Amsterdam, p 23-34.

Hudon G. (1990) - Une comparaison des résultats de modèles log-linéaires et de généralisations de l'analyse des correspondances. *Revue de Statist. Appl.*, 38, (2), p 43-53.

Hurley J. R. , Cattell R. B. (1962) - The Procrustes program : Producing direct rotation to test an hypothesized factor structure. *Behavioural Science*, 7, p 258-262.

Hyvärinen A. (1996) – Purely local neural principal components and independent component learning.. *Proc. Int. Conf. on Artificial Neural Networks*, Bochun, Germany, p 139-144.

Hyvärinen A. (1999) – Survey on Independent Component Analysis. *Neural Computing Surveys*, 2, p 94-128.

Jain A. K., Moreau J. V. (1987) - Bootstrap technique in cluster analysis. *Pattern Recognition*, 20, p 547-568.

Jambu M. (1991) - *Exploration statistique et informatique des données*. Dunod, Paris.

Jambu M., Lebeaux M.O. (1978) - *Classification automatique pour l'analyse des données*. Dunod, Paris.

Jeffreys H. (1946) - An Invariant form for the prior probability in estimation problems. *Proc. Roy. Soc.* (A), 186, p 453-461.

Johnson S. C. (1967) - Hierarchical clustering schemes. *Psychometrika*, 32, p 241-254.

Jolliffe I. (1986) - *Principal Component Analysis*. Springer-Verlag, New York.

Jones M.C., and Sibson R. (1987) - What is projection pursuit (with discussion). *J. of Royal Statist. Society*, A, 150, p 1-36.

Jordan C. (1874) - Mémoire sur les formes bilinéaires. *J. Math. Pures et Appliquées*. 19, p 35-54.

Joreskög K. (1963) - *Statistical Estimation in Factor Analysis : a New Technique and its Foundation*. Almqvist & Wiksell, Uppsala.

Joreskög K., Sörbom D. (1979) - *Advances in Factor Analysis and Structural Equation Models*. Abt, Cambridge (MA).

Jousselin B. (1972) - Les choix de consommation et les budgets des ménages. *Consommation*, Dunod, 1, p 41-72.

Jutten C. (1987) – *Calcul neuromimétique et traitement du signal ; Analyse en composantes indépendantes*. Thèse, INPG, Université de Grenoble.

Jutten C., Hérault J. (1991) – Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, vol. 24: p 1-10.

Kaiser H. F. (1961) - A note on Guttman's lower bound for the number of common factors. *Brit. J. Statist. Psychol.*, 14, p 1-2.

Kato T. (1966) - *Pertubation Theory for Linear Operators*. Springer, New York.

Kaufman L., Rousseeuw P. J. (1986) - Clustering large data sets (with discussion). *Pattern recognition in practice II* (E.S. Gelsema and L.N. Kanal, eds), North-Holland, Amsterdam, p 425-437.

Kaufman L., Rousseeuw P. J. (1990) - *Finding Groups in Data*. J. Wiley, New York.

Kayser H. F.(1958) – The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, vol. 23, p 187-200.

Kazmierczak J.-B. (1985) - Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.* , 33, (1), p 13-24.

Kendall M. G. (1962) - *Rank Correlation Methods*. Griffin, London.

Kendall M. G. (1966) - Discrimination and classification. In : *Proc. Symp. Mult. Analysis.* Dayton, Ohio, (Krishnaiah P. R. (ed.), Academic Press, New York, p 165-185.

Kendall M.G., Stuart A. (1961) - *The Advanced Theory of Statistics.* Charles Grifin, London.

Kettenring R. J. (1971) - Canonical analysis of several sets of variables. *Biometrika*, 58, (3), p 433-450.

Kharchaf I., Rousseau R. (1988, 1989) Reconnaissance de la structure de blocs d'un tableau de correspondance par la classification ascendante hiérarchique: parties 1 et 2, *Les Cahiers de l'Analyse des Données*, 13, p 439-443; et : 14, p 257-266.

Kissita G., Cazes P., Hanafi M., Lafosse R. (2004) – Deux méthodes d'analyse factorielle du lien entre deux tableaux de variables partitionnés. *Revue de Stat. Appl.*, p 73-92.

Kiers H. A. L. (1989) - *Three-way Methods for the Analysis of Quantitative and Qualitative Two-way Data.* DSWO Press, Leiden.

Kleiweg P. (1996) - *Een inleidende cursus met practica voor de studie,* Alfa-Informatica. Master's thesis, Rijksuniversiteit Groningen.

Kohonen T. (1989) - *Self-Organization and Associative Memory.* Springer Verlag, Berlin.

Koren Y., Carmel L., Harel D. (2002) - ACE: a Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs, *Proceedings of IEEE Information Visualization,* p 137-144.

Krishnaiah P.R., Chang T. C. (1971) - On the exact distribution of the extreme roots of the Wishart and MANOVA matrix. *J. of Multivariate Anal.*, 1, (1), p 108-116.

Krishnaiah P.R., Kanal L. (Eds) (1982) - *Handbook of Statistics* (2). North Holland, Amsterdam.

Kroonenberg P. (1983) - *Three-Mode Principal Component Analysis.* DSWO Press, Leiden.

Kroonenberg P. M., de Leeuw J. (1980) - Principal component analysis of three-mode data by means of alternating least-square algorithms. *Psychometrika*, 45, p 69-97.

Kruskal J. B. (1956) - On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* ,7, p 48-50.

Kruskal J. B., Wish M. (1978) - *Multidimensional Scaling.* Sage University Paper, 11, Sage, Beverly Hills.

Krzanowski W. J. (1984) - Sensitivity of principal components. *J. Royal Statist. Soc.* , 46,(3), p 558-563.

Krzanowski W. J. (1987) - Cross-validation choice in principal component analysis. *Biometrics*, 43, p 575-584.

Kshirsagar A.M. (1972) - *Multivariate Analysis.* Marcel Dekker, New York.

Kullback S. (1959) - *Information Theory and Statistics.* J. Wiley, New York.

Kullback S., Leibler R.A. (1951) - Information and sufficiency. *Ann. Math. Statist.*, 22, p 79-86.

Lachenbruch P.A., Goldstein M. (1979) - Discriminant Analysis. *Biometrics*, 35, p 68-85.

Lachenbruch P.A., Mickey M.R. (1968) - Estimation of error rate in discriminant analysis. *Technometrics*, 10, p 1-11.

Lafosse R. (1985) - *Analyses Procustéennes de deux Tableaux.* Thèse, Univ. Paul Sabatier, Toulouse.

Lallich S., Prudhomme E., Teytaud O. (2004) Contrôle du risque multiple pour la sélection de règles d'association significatives. *Revue des Nouvelles Technologies de l'Information. (EGC2004),* vol. 1, p 305-316.

Lancaster H. O. (1963) - Canonical correlation and partition of □. *Quart. J. Math.*, 14, p 220-224.

Lancaster H. O. (1969) - *The Chi-squared Distribution*. J. Wiley, New York.

Lance G. N., Williams W. T. (1967) - A general theory of classification sorting strategies. *Computer J.*, 9, p 373-380.

Laplace P.S. (1793) - Sur quelques points du système du monde. *Mémoires de l'Académie Royale des Sciences de Paris*, p 1-87; Réédition: *Oeuvres*, (1895), 11, Gauthier-Villars, Paris, p 477-558.

Lauro N. C., Decarli A. (1982) - Correspondence analysis and log-linear models. In : multiway contingency tables study. *Metron*, 1-2, p 213-234.

Lauro N. C, D'Ambra L. (1984) - L'Analyse non-symétrique des Correspondances. In : *Data Analysis and Informatics*, III, Diday et al. Ed., North-Holland, p 433-446.

Lavit C. (1988) - *Analyse Conjointe de Tableaux Quantitatifs*. Masson, Paris.

Lawley D. N. (1956) - Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43, p 128-136.

Lawley D. N., Maxwell A. E. (1963) - *Factor Analysis as a Statistical Method*. Methuen, London.

Le Calvé G. (1987) - L_1 -embeddings of a data structure (I,D). in: *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Dodge Y. ed., North Holland, Amsterdam, p 195-202.

Le Foll Y. (1982) - Pondération des distances en analyse factorielle. *Statist. et Anal. des Données*, 7, p 13-31.

Le Foll Y., Burtschy B. (1983) - Représentations optimales des matrices imports-exports. *Revue de Statist. Appl.* , 31, (3), p 57-72.

Lebart L. (1969 a) - L'Analyse statistique de la contiguïté. *Publications de l'ISUP*, XVIII- p 81 - 112.

Lebart L. (1969 b) - Introduction à l'analyse des données : Analyse des correspondances et validité des résultats. *Consommation*, Dunod. 4, p 65-87.

Lebart L. (1974) - On the Benzécri'smethod for finding eigenvectors by stochastic approximation. *Proceedings in Comp. Statist., In: COMPSTAT*, Physica verlag, Vienna, p 202-211.

Lebart L. (1975 a) - L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Consommation*, 2, p 73-96. Dunod.

Lebart L. (1975 b) - *Validité des résultats en analyse des données*. Rapport Credoc-Cordes. Credoc, Paris.

Lebart L. (1976) - The significance of eigenvalues issued from correspondence analysis. *Proceedings in Comp. Statist., In: COMPSTAT*, Physica verlag, Vienna, p 38-45.

Lebart L. (1982) - Exploratory analysis of large sparse matrices, with application to textual data. *COMPSTAT*, Physica Verlag, Vienna, p 67-76.

Lebart L. (1986) - Qui pense quoi ? Evolution et structure des opinions en France de 1978 à 1984. *Consommation Revue de Socio-Economie*, Dunod, 4, p 3-22.

Lebart L. (1987 a) - Some recent advances in data analysis practice. In : *New Perspective in Theoretical and Applied Statistics*. M.L. Puri and al., Eds. J. Wiley, New York.

Lebart L. (1987 b) - Conditions de vie et aspirations des Français, Evolution et structure des opinions de 1978 à 1984. *Futuribles*, 1, p 25-56.

Lebart L. (1988) - Contribution of classification to the processing of longitudinal socio-economic surveys. In : *Classification and Related Methods of Data Analysis*, H. Bock Ed., North Holland, p 113-120.

Lebart L. (1992) - Discrimination through the regularized nearest cluster method. *COMPSTAT; Proceedings of the 10th Symposium on Computational Statistics*, Physica Verlag, Vienna, p 103-118.

Lebart L. (1996) - Correspondence analysis, discrimination and neural networks. In: *Data Science, Classification and Related Methods*. Hayashi C., Ohsumi N., Yajima K., Tanaka Y., Bock H.-H., Baba Y. (eds), Springer, Berlin, p 423-430.

Lebart L. (1997) - Réseaux de neurones et analyse des correspondances. *La Revue de MODULAD*, 18, INRIA, p 21-37.

Lebart L. (2000) - Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (eds), *Data Analysis*. Berlin, Springer, p. 233--244.

Lebart L. (2004) - Validation techniques in Text Mining. In: *Text Mining and its Application*, S. Sirmakensis (ed.), Berlin- Heidelberg, Springer Verlag, p. 169-178.

Lebart L., Fénelon J.P. (1971) - *Statistique et informatique appliquées*. Dunod, Paris.

Lebart L., Houzel Y. (1980) - Le système d'enquête sur les aspirations des Français. *Consommation Revue de Socio-Economie*, Dunod, 1, p 3-25.

Lebart L., Mirkin B. (1993) - Correspondence analysis and classification. In : *Multivariate Analysis: Future Directions 2*, C. M. Cuadras and C.R.Rao, Eds., North Holland, Amsterdam, p 341-357.

Lebart L., Morineau A., Fénelon J.P. (1981) - *Traitement des Données Statistiques*. Dunod, Paris.

Lebart L., Morineau A., Lambert T., Pleuvret P. (1991) - *SPAD.N version 2 Système Portable pour l'Analyse des Données*. CISIA, Saint-Mandé.

Lebart L., Morineau A., Tabard N. (1977) - *Techniques de la description statistique*. Dunod, Paris.

Lebart L., Morineau A., Warwick K. (1984) - *Multivariate Descriptive Statistical Analysis*. J. Wiley, New York.

Lebart L., Piron M., Steiner J.-F. (2003) – *La sémiométrie*. Dunod, Paris.

Lebart L., Salem A. (1994) - *Statistique textuelle*. Dunod, Paris.

Lebart L., Tabard N. (1973) - *Recherches sur la description automatique des données socio-economiques*. Rapport CORDES-CREDOC, C.R n°13/1971.

Lebreton J.-D., Chessel D., Prodon R., Yoccoz N. (1988) - L'analyse des relations espèces-milieu par analyse canonique des correspondances. *Acta Œcoligica, Œcol. Gener.*, 9, (1), p 53-67.

Leclerc A. (1975) - L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue Statist. Appl.*, 23, p 5-16.

Leclerc A. (1976) - Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives. *Int. Statist. Review*, 44, p 221-248.

Leclerc A., Chevalier A., Luce D., Blanc M. (1985) - Analyse des correspondances et modèle logistique : possibilités et intérêt d'approches complémentaires. *Revue Statist. Appl.*, 33, p 25-38.

Lejeune M. (ed) (2001) – *Traitement des fichiers d'enquêtes*. Presse Univ. de Grenoble.

Lelu A. (1991) - From data analysis to neural networks : new prospects for efficient browsing through databases. *Journal of Information Science*, 17, p 1-12.

Lelu A. (1994) - Clusters and factors : neural algorithm for a novel representation of highly multidimensional data sets. In : *New Approaches in Classification and Data Analysis*, Diday et al. (eds), Springer Verlag, Berlin, p 241-248.

Lerman I. C. (1970) - *Les Bases de la Classification Automatique*. Gauthier-Villars, Paris.

Lerman I. C. (1981) - *Classification et analyse ordinale des données*. Dunod, Paris.

Le Roux B., Rouanet H. (2004) – *Geometric Data Analysis*. Kluwer Ac. Publ., Dordrecht.

L'Hermier des Plantes H. (1976) - *STATIS : Structuration de tableaux à trois indices de statistique*. Thèse (3C), USTL, Montpellier.

Li G., and Chen Z. (1985) - Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *J. of Amer. Statist. Assoc.* 80, p 759-766.

Ling R. F. (1973) - A probability theory of cluster analysis. *J. Amer. Statist. Assoc.*, 68, p 159-164.

MacQueen J. B. (1967) - Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th)*, Berkeley, 1, p 281-297, Univ. of Calif. Press, Berkeley.

Mahalanobis P.C. (1936) - On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, 12, p 49-55.

Malinvaud E. (1964) - *Méthodes statistiques de l'économétrie*. Dunod, Paris.

Malinvaud E. (1987) - Data analysis in applied socio-economic statistics with special consideration of correspondence analysis. *Marketing Science Conference Proceedings*, HEC-ISA, Jouy en Josas.

Mallows C.L., Tukey J.W. (1982) - An overviews of technique of data analysis emphasizing its exploratory aspects. In : *Some Recent Advances in Statistics* (J. Tiago de Oliveira Ed.) Academic Press, p 11-172.

Mallows C.L. (1973) - Some comments on C_p. *Technometrics*, 15, p 661-675.

Marcotorchino F. (1987) - Block seriation problems: a unified approach, *Applied Stochastic Models and Data Analysis*, 3, p 73-93.

Markus M.Th. (1994) - Bootstrap Confidence Regions for Homogeneity Analysis.; the Influence of Rotation on Coverage Percentages. *COMPSTAT 1994*, (Dutter R. and Grossmann W. (eds)) , Heidelberg: Physica Verlag, p 337-342.

Marsaglia G., Bray T.A. (1964) - A convenient method for generating normal variables. *SIAM* Rev. 6, p 260-264.

Martin R.S., Reinsch C., Wilkinson J.H. (1968) - Householder's tridiagonalisation of a symmetric matrix. *Num. Math.* 11, p 181-195.

Martin R.S., Wilkinson J.H. (1968) - Implicit Ql algorithm. *Num. Math.* 12, p 377-383.

Masson M. (1974) - Analyse non linéaire de données. *C.R. Acad. Sc.*, 278 (11 mars).

Matheron G. (1963) - Principles of geostatistics. *Economic Geology*, 58, p 1246-1266.

Matheron G. (1965) - *Les variables régionalisées et leur estimation*. Masson, Paris.

Matusita K. (1955) - Decision rules based on the distance, for problems of fit, two samples, and estimation. *Ann. of Math. Statist.* 26, 4, p 631-640.

Matusita K., Ohsumi N. (1980) - A criterion for choosing the number of clusters in cluster analysis. In : *Recent Developement in Statistical Inference and Data Analysis*, Matusita K. (ed.) North-Holland, Amsterdam, p 203-213.

McCullagh P., Nelder J.A. (1989) - *Generalized Linear Models.* Chapman and Hall, London.

McLachlan G. J., Peel (2000) - Density estimation using Normal Mixture Models, J. Wiley, New York.

McLachlan G.J. (1992) - *Discriminant Analysis and Statistical Pattern Recognition.* J. Wiley, New York.

McQuitty L.L. (1966) - Single and multiple classification by reciprocal pairs and rank order type. *Educational Psychology Measurements.* 26, p 253-265.

Mehta C. R., Patel N., R. (1991) - *Statistique non-paramétrique exacte, Introduction à StatXact.* CISIA, Saint Mandé.

Mehta M.L. (1960) - On the statistical properties of the level spacing in nuclear spectra. *Nucl. Phys.* 18, p 395-419.

Mehta M.L. (1967) *Random Matrices and the Statistical Theory of Energy Levels.* Academic Press, New York.

Meot A., Chessel D., Sabatier R. (1993) - Opérateur de voisinage et analyse des données spatio-temporelles. In *Biométrie et environnement,* Lebreton J.-D., Asselain B., (eds), Masson, Paris, p 45-71.

Meulman J. (1982) - *Homogeneity Analysis of Incomplete Data.* DSWO Press, Leiden.

Meyer R. (1994) - An eigenvector algorithm to fit L_p-distance matrices. In : *New Approches in Classification and Data Analysis,* Diday E. et al. (eds), Springer Verlag, Berlin, p 502-509.

Michelat G., Simon M. (1985) - Les sans-réponses aux questions politiques, *Revue Pouvoirs,* 33, PUF, Paris.

Milan L., Whittaker J. (1995) - Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics,* 44, 1, 31-49.

Milgram M. (1993) - *Reconnaissance des formes, méthodes numériques et connexionnistes.* Armand Colin, Paris.

Miller R. G. (1966) - *Simultaneous Statistical Inference.* Mac Graw Hill, New York.

Miller R. G. (1974) - The Jakknife-a review. *Biometrika,* 61, p 1-15.

Milligan G. W., Cooper M. C. (1985) - An examination of procedures for determining the number of cluster in a data set. *Psychometrika,* 50, p 159-179.

Mirkin, B.G. (1990) - A sequential fitting procedure for linear data analysis models, *J. of Classification,* 7, p 167-195.

Mohar B. (1991) - The Laplacian Spectrum of Graphs, *Graph Theory Combinatorics and Application,* 2, , p 871-898.

Mohar B.(1997) - Some Applications of Laplace Eigenvalues of Graphs, *Graph Symmetry, Algebraic Methods and Application,* Hahn G., Sabidussi G., NATO Ser. C., 497, Kluwer, p 225-275.

Mollière J.L. (1986) - What's the real number of clusters? In : *Classification as Tool of Research,* Gaul W., Schader M. (eds), North-Holland, Amsterdam, p 311-320.

Mollière J.L. (1989) - Stratégie de classification pour de grands ensembles de données. *La Revue de Modulad (INRIA),*3, p 31-69.

Mom A. (1988) - *Méthodologie statistique de la classification des réseaux de transport.* Thèse, U.S.T.L., Montpellier.

Mood A.M. (1951) - On the distribution of the charasteristic roots of normal second moment matrices. *Ann. Math. Statist.* 22, p 266-273.

Moran P.A.P. (1948) - The interpretation of statistical maps. *J. Royal Statist. Soc.*, B, 10, p 243-251.

Moran P.A.P. (1954) - Notes on continuous stochastic phenomena, *Biometrika*, 37, p 17-23.

Moreau J. (1992) - *Analyse de données structurées par des graphes. Cas de l'analyse des correspondances*. Thèse, E.P.F.L., Lausanne.

Moreau J., Doudin P.A., Cazes P. (2000) - *L'analyse des correspondances et les techniques connexes*. Springer, Berlin.

Morgan J.M., Messenger R.C. (1973) - *THAID : a sequential search program for the analysis of nominal scale dependent variables*. Institute for Social Research, University of Michigan, Ann Arbor.

Morgenthaler S., Tukey J.W. (1989) - The next future of data analysis. In *Data Analysis, Learning Numeric and Symbolic Knowledge*, 1989, Diday ed., Novascience, p 1-12.

Morin A., Bosc P., Hébrail G., Lebart L. (2001) - *Bases de données et statistique*. Dunod.

Morineau A. (1983) - Etude de stabilité en analyse en composantes principales. *Bull. Techn. du Centre de Statist. et d'Infor. Appl.*, 1, p 9-12.

Morineau A. (1984) - Note sur la caractérisation statistique d'une classe et les valeurs-tests, *Bull. Techn. du Centre de Statist. et d'Infor. Appl.*, 2, p 20-27.

Morineau A. (1992) - L'Analyse de données et les tests de cohérence dans les données d'enquête. In : *La Qualité de l'Information dans les Enquêtes*, ASU (ed.), Dunod, Paris.

Morineau A., Lebart L. (1986) - Specific clustering algorithms for large data sets and implementation in SPAD Software. In : *Classification as a tool of research*, Gaul W., Schader M., Eds, North Holland, Amsterdam, p 321-330

Morineau A., Rakotomalala R. (2006) – Critère VT100 de sélection de règles d'association. *Revue des Nouvelles Technologies de l'Information, EGC-2006*, p 581 – 591.

Mosteller F., Tukey J. W. (1977) - *Data Analysis and Regression*. Addison Wesley, Reading, (Mass).

Muirhead R. J. (1982) - *Aspects of Multivariate Statistical Theory*. J. Wiley, New York.

Mulaik S. A. (1972) - *The Foundation of Factor Analysis*. McGraw Hill, New York.

Murtagh F. (1985) - *Multidimensional Clustering Algorithms*.COMPSTAT Lectures 4, Physica Verlag, Vienna.

Nakache J. P., Confais J. (2003) - *Statistique explicative appliquée*. Editions Technip, Paris.

Nakache J. P., Confais J. (2005) - *Approche pragmatique de la classification*. Editions Technip, Paris.

Nakache J.P. (1973) - Influence du codage des données en analyse factorielle des correspondances. Etude d'un exemple pratique médical. *Revue Statist. Appl.*, 21, (2).

Nakache J.-P., Lorente P., Benzécri J.-P., Chastang J.-F. (1977) - Aspects Pronostics et thérapeutiques de l'infarctus myocardique aigu. *Les Cahiers de l'Analyse des Données*, 2, p 415-434.

Nakhlé F. (1976) - Sur l'analyse d'un tableau de notes dédoublées. *Les Cahiers de l'Analyse des Données*, 1, p 243-257.

Neave H.R. (1973) - On using Box-Muller transformation with multiplicative congruential pseudo-random number generators. *Appl. Statist.*, 22, p 92-97.

Nelder J.A., Wedderburn R.W.M. (1972) - Generalized linear models. *J. R. Statist. Soc.*, A, 135, p 370-384.

Newman T.G., Odell P.L. (1971) - *The Generation of Random Variates.* GRIFFIN's Statistical Methods and Courses, n°29, Griffin.

Nijenhuis A., Wilf H.S. (1975) - *Combinatorial Algorithms.* Academic Press, New York.

Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application.* Univ. of Toronto Press.

Ohsumi N. (1988) - Role of computer graphics in interpretation of clustering results. In : *Recent Developments in Clustering and Data Analysis,* Diday E. et al. (eds), Academic Press, Boston.

Oja E. (1982) - A simplified neuron model as a principal components analyzer. *J. of Math. Biology,* 15, p 267-273.

Oja E. (1992) - Principal components, minor components, and linear neural networks. *Neural Networks,* 5, p 927-935.

Oja E., Karhunen J. (1981) On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. Report of the Helsinki University of Technology (Dept of Technical Physics). Otaniemi, Finland.

O'Neill M. E. (1978) - Distributional expansion from canonical correlation from contingency tables. *J. Roy. Statist. Soc.,* B, 40, p 301-312.

O'Neill M. E. (1981) - A note on the canonical correlation from contingency tables. *Austr. J. Statist.,* 23, p 58-66.

Pagès J.-P., Escoufier Y., Cazes P. (1976) - Opérateurs et analyse de tableaux à plus de deux dimensions. *Cahiers du BURO,* ISUP, Paris, p 61-89

Palm R., Iemma A. F. (1995) - Quelques alternatives à la regression classique dans le cas de la colinéarité. *Revue Statist. Appl.,* 43, (2), p 5-33.

Parzen E. (1962) - On the estimation of a probability density function and mode. *Ann. of Math. Statist.,* 33, p 1065-1076.

Pearson K. (1901) - On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°ll, p 559-572.

Perneger T.,V. (1998). What is wrong with Bonferroni adjustments, *British Medical Journal,* 136, 1236-1238.

Perruchet C. (1983) - Une analyse bibliographique des épreuves de classifiabilité en analyse des données. *Statist. et Anal. des Données,* 8, p 18-41.

Pillai K.C.S. (1965) - On the distribution of the largest root of a matrix in multivariate analysis. *Biometrika,* 52, p 405-414.

Pillai K.C.S. (1967) - Upper percentage point of the largest root of a matrix in multivariate analysis. *Biometrika,* 54, p 189-194.

Pillai K.C.S., Chang T.C. (1970)) An approximation to the c.d.f. of the largest root of a covariance matrix. *Ann. of the Inst. of Statist. Math.,* p 115-124.

Piron M. (1990) - *Structuration de l'information à plusieurs niveaux et analyse des données.* Thèse, Université Pierre et Marie Curie.

Piron M. (1992) - *Analyse statistique d'un système d'échelles.* Réseau ADOC, doc. 4, ORSTOM, Bondy.

Pousse A. (1992) - Résultats asymptotiques. In : *Modèles pous l'analyse des données multidimensionnelles,* Droesbeke *et al.,* eds, Economica, Paris.

Prim R.C. (1957) - Shortest connection matrix network and some generalizations. *Bell System Techn. J.,* 36, p 1389-1401.

Proriol J. (1991) MLP - Programme de réseau de neurone multicouche. *La Revue de MODULAD*, 8, INRIA, p 23-29.

Ramsay J.O. (1978) - Confidence region for multidimensional scaling analysis. *Psychometrika*, 43, p 145-160.

Rao C.R. (1964) - The use and interpretation of principal component analysis in applied research. *Sankhya* serie A, 26, p 329-357.

Rao C.R. (1973) - *Linear Statistical Inference and its Application*. (1st ed. : 1965) J. Wiley, New York.

Rasson J.-P., Kubushishi T. (1994) - The gap test : an optimal method for determining the number of natural classes in cluster analysis. In : *New Approaches in Classification and Data Analysis*, Diday E. et al. (eds), Springer Verlag, Berlin, p 186-193.

Reinert M. (1986) - Classification descendante hiérarchique : un algorithme pour le traitement des tableaux logiques de grandes dimensions. In *Data Analysis and Informatics*, 4, Diday *et al.* Ed., North-Holland, p 23-28.

Richardson M., Kuder G. F. (1933) - Making a rating scale that measures. *Personnel Journal.*, 12, p 71-75.

Ripley B. D. (1981) - *Spatial Statistics*. J. Wiley, New York.

Ripley B. D. (1983) - Computer generation of random variables : a tutorial. *Inter. Statist. Review*, 51, p 301-319.

Ripley B. D. (1993) - Statistical aspects of neural networks. In : *Networks and Chaos-Statistical and Probabilistic Aspects*, Barndorff-Nielsen O.E., Jensen J. L., Kendall W. S., (eds), Chapman and Hall, London, p 40-123..

Ripley B. D. (1994) - Neural nerworks and related methods of classification. *J. R. Statist. Soc. B*, 56, n°3, p 409-456.

Ritter H., Martinez T., Schulten K. (1992) - *Neural Computation and Self-Organizing Maps : An Introduction*. Addison Wesley, Reading.

Robert Ch. (2006) – *Le choix bayésien*. Springer, Paris.

Robert P., Escoufier Y. (1976) - A unifying tool for linear multivariate methods : the Rv coefficient. *Applied Statistics*, 25, (3), p 257-265.

Romeder J.M. (1973) - *Méthodes et Programmes d'Analyse Discriminante*. Dunod, Paris.

Rosenblatt M. (1956) - Remarks on some nonparametric estimates of the density function. *Ann. of Math. Statist.*, 27, p 823-835.

Rouanet H., Le Roux B. (1993) - *Analyse des données Multidimensionnelles*. Dunod, Paris.

Rousset P. and Guinot C. (2002) - Visualisation des distances entre les classes de la carte de Kohonen pour le développement d'un outil d'analyse et de représentation des données. *Revue de Statistique Appliquée*, p. 35-47.

Roux M. (1985) - *Algorithmes de Classification*. Masson, Paris.

Roux M. (1991) - Basic procedures in hierarchical cluster analysis. *Applied Multivariate Analysis in SAR and Environmental Studies* (J. Devillers and W. Karcher, eds), p 115-135, ECSC, EEC, EAEC, Brussels and Luxembourg.

Roy S.N. (1939) - *p*-Statistics or some generalisations of analysis of variance appropriate to multivariate problems. *Sankhya*, 4, p 381-396.

Rumelhart D. E., Hinton G. E., Williams R. J. (1986) - Learning internal representation by back-propagating errors. *Nature*, 323, p 533-536.

Sabatier R. (1984) - Quelques généralisations de l'analyse en composantes principales de variables intrumentales. *Statist. et Anal. des Données*, 9, (3), p 75-103.

Sabatier R. (1987) - Analyse factorielle de données structurées et métriques. *Statist. et Anal. des Données*, 12, (3), p 75-96.

Sabatier R., Lebreton J.-D., Chessel D. (1989) - Principal component analysis with instrumental variables as a tool for modeling composition data. In : *Mutiway Data Analysis*, Coppi R., Bolasco S. (eds), Elsevier, Amsterdam.

Saporta G. (1990) - *Probabilités, analyse des données et statistiques*. Technip, Paris.

Saporta G. (1975 a) - *Liaisons entre plusieurs ensembles de variables et codages de données qualitatives*. Thèse 3°C., Université Paris VI.

Saporta G. (1975 b) - Dépendance et codage de deux variables aléatoires. *Revue Statist. Appl.* 23, p 43-63.

Saporta G. (1977) - Une méthode et un programme d'analyse discriminante sur variables qualitatives. In : *Premières Journées Int. Analyse des Données et informatiques*, INRIA, Rocquencourt.

Saporta G. (2001) – *Data Mining* et statistique. *Journal de la Société Française de Statistique*, vol. 142, p 81-84.

Saporta G., Hatabian G. (1986) - Régions de confiance en analyse factorielle. In : *Data Analysis and Informatics*, 4, Diday E. *et al.* (eds), North-Holland, Amsterdam, p 499-508.

Sarle W. S. (1983) - *Cubic clustering criterion*. SAS Technical Report. A-108. SAS Institute Limited. Cary, NC.

Saville D. J. (1990) - Multiple comparison procedures : The practical solution. *American Statistician*, 44, p 174-180.

Schiffman S. S., Lance G. N., Reynolds M., Young F. W. (1981) - *Introduction to Multidimensional Scaling*. Academic Press, New York.

Schönemann P. H. (1968) - On two-sided orthogonal procrustes problems. *Psychometrika*, 33, p 19-33.

Schönemann P. H., Carroll R. M. (1970) - Fitting one matrix to another under choice of a central dilation an a rigid motion. *Psychometrika*, 35, p 245-255.

Schriever B.F. (1983) - Scaling of order dependent categorical variables with correspondence analysis. *Inter. Statist. Review*, 51, p 225-238.

Searle S.E. (1971) - *Linear Models*. J. Wiley, New York.

Seber G.A.F. (1977) - *Linear Regression Analysis*, J. Wiley, New York.

Shepard R. N. (1974) - Representation of structure in similarity data : problems and prospects. *Psychometrika*, 39, (4), p 373-421.

Silverman B. W. (1986) - *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Sirat J. A. (1991) - A fast neural algorithm for principal component analysis and singular value decomposition. *Internat. J. of Neural Systems*, 2, p 147-155.

Sneath P. H. A. (1957) - The Application of computers to taxonomy. *J. General Microbiology*, 17, p 201-226.

Snee R.D. (1974) - Graphical displays of two-ways contingency tables. *Amer. Statistician* 28, p 9-12.

Sokal R. R., Sneath P. H. A. (1963) - *Principles of Numerical Taxonomy*, Freeman and co., San-Francisco.

Sonquist J. A. and Morgan J. N. (1964) - *The Detection of Interaction Effects*. Institute for Social Research, University of Michigan, Ann Arbor.

Spearman C. (1904) - General intelligence, objectively determined and measured. *Amer. Journal of Psychology*, 15, p 201-293.

Stauffer D. F., Garton E. O., Steinhorst R. K. (1985) - A comparison of principal component from real and random data. *Ecology*, 66, p 1693-1698.

Stone M. (1974) - Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B 36, p 111-147.

Stone M. (1977) - An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Royal Statist. Soc. B*, 39, p 44-47.

Sugiyama T. (1966) - On the distribution of the largest latent root and the corresponding latent vector for principal component analysis. *Ann. Math. Statist.* 37, p 995-1001.

Sylvester J.J. (1889) - *Messenger of Mathematics* (cité par Eckart, Young, 1939). 19, n°42.

Tabard N. (1972) - Consommation et statut social, analyse multidimensionnelle des budgets familiaux. *Consommation*, 2, p 41-63.

Tanaka Y. (1984) - Sensitivity analysis in Hayashi's third method of quantification. *Behaviormetrika*, 16, p 31-44.

Tenenhaus M. (1994) - *Méthodes statistiques en gestion*. Dunod, Paris.

Tenenhaus M. (1998) - *La régression PLS, Théorie et Pratique*. Technip, Paris.

Tenenhaus M., Leroux Y., Guimart C., Gonzales P. L. (1993) - Modèle linéaire généralisé et analyse des correspondances. *Revue de Statist. Appl.* , 41, (2) p 59-86.

Tenenhaus M., Young F. W. (1985) - An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, p 91-119.

Ter Braak C. J. F. (1986) - Canonical Correspondence Analysis. : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67, (5), p 1167-1179.

Ter Braak C. J. F. (1987) - The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69, p 69-77.

Ter Braak C. J. F. (1988) - Partial Canonical Correspondence Analysis. In : *Classification and Related Methods of Data Analysis*, Bock H.H. (ed.) North Holland, p 551-558.

Theil H. (1971) - *Principles of Econometrics*. J. Wiley, New York.

Thionet P. (1976) - Construction et reconstruction de tableaux statistiques. *Annales de l'INSEE*, 22-23, p 5-28.

Thiria S., Gascuel O., Lechevallier Y., Canu S. (1997) - *Statistique et méthodes neuronales*. Dunod, Paris.

Thom R. (1974) - *Modèles mathématiques de la morphogenèse*. 10/18, Bourgois, Paris.

Thorndike R.L. (1953) - Who belongs in the family. *Psychometrika*, 18, p 267-276.

Thurstone L. L. (1947) - *Multiple Factor Analysis*. The Univ. of Chicago Press, Chicago.

Tomassone R., Danzart M., Daudin J.J., Masson J.P. (1988) - *Discrimination et classement*. Masson, Paris.

Tomassone R., Dervin C., Masson J.-P. (1993) - *Biométrie, Modélisation de phénomènes biologiques*. Masson, Paris.

Tomassone R., Lesquoy E., Millier C.(1983) - *La régression : nouveaux regards sur une ancienne méthode statistique*. Masson, Paris.

Toussaint G.T. (1974) - Bibliography on estimation of misclassification. *IEEE Trans. Inform. Theory*, IT-20, p 472-479.

Tucker L. R. (1958) - An inter-battery method of factor analysis. *Psychometrika*, 23, (2).

Tucker L. R. (1964) - The extension of factor analysis to three-dimensional matrices. In : *Contribution to Mathematical Psychology*, Harris C. W. (ed.), Univ. of Wisconsin Press, Madison, p 109-127.

Tucker L. R. (1966) - Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, p 279-311.

Tuffery S.(2005) - *Data mining et statistique décisionnelle*, Editions Technip, Paris.

Tukey J. W. (1958) - Bias and confidence in not quite large samples. *Ann. Math. Statist.*, (Abstract), 29, p 614.

Tukey J. W. (1977) - *Exploratory Data Analysis*. Addison Wesley, Reading, Mass.

Vaillant B., Lenca P., Lallich S. (2004a) - Etude expérimentale de mesure de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information. (EGC2004)*, Vol. 1, p 341-352.

Vaillant B., Lenca P., Lallich S. (2004b) – A clustering of interestingness measures. *Proceedings of the Seventh International Conference on Discovery Science)*, DS2004, Padova, 290-297..

van Buuren S., and Heiser W.J. (1989) - Clustering N objects into k groups under optimal scaling of variables. *Psychometrika*, 54, 4, p 699-706.

Van Cutsem B. (ed.) (1994) - *Classification and Dissimilarity Analysis*. Springer-Verlag, New York.

van der Heijden P. G. M., de Falguerolles A., de Leeuw J. (1989) - A combined approach to contingency table analysis with correspondence analysis and log-linear analysis (with discussion). *Applied Statistics*, 38, p 249-292.

van der Heijden P. G. M., de Leeuw J. (1985) - Correspondence analysis used complementary to log-linear analysis. *Psychometrika*, 50, p 429-447.

van der Heijden, P. G. M. (1987) - *Correspondence Analysis of Longitudinal Categorical Data*. DSWO Press, Leiden.

van Rijckevorsel J. (1987) - *The application of fuzzy coding and horseshoes in multiple correspondances analysis*. DSWO Press, Leiden.

Vapnik W. (1995) - *The Nature of Statistical Learning*. Springer, New York.

Vapnik W. (1998) - *Statistical Learning Theory*. Wiley, New York.

Volle M. (1981) – *Analyse des données*, Economica, Paris.

Wakimoto K., Taguri M. (1978) - Constellation graphical methods for representing multidimensional data. *Ann. of the Inst. of Statist. Math.*, 30, (1), p 97-104.

Ward J.H. (1963) - Hierarchical grouping to optimize an objective function. *J. of Amer. Statist. Assoc.*, 58, p 236-244.

Waternaux C. M. (1976) - Asymptotic distribution of the sample roots for a non-normal population. *Biometrika*, 63, p 639-645.

Werbos P. J. (1974) - *Beyond Regression : New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph.D. Thesis, Harvard University.

Werbos P. J. (1990) - Backpropagation through time : what it does and how to do it. *Proceedings of the IEEE*, 78, (10), p 1550-1560.

Wermuth N. (1976) - Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32, p 95-108.

Wermuth N., Cox D. R. (1992) - Graphical models for dependencies and associations. In : *Computational Statistics* (Dodge Y., Whittaker J., eds), 1, p 235-250, Physica Verlag, Heidelberg.

Westfall P. H., Young S. S. (1993). *Resampling Based Multiple Testing: Examples and Methods for p-values Adjustment.* Wiley, New York.

Whittaker J. (1990) - *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.

Wilkinson J. H. (1965) - *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford.

Wilkinson J. H., Reinsch C. (1971) - *Handbook for Automatic Computation.* 2, Linear Algebra, Springer-Verlag.

Williams W. T. and Lambert J. M. (1959) - Multivariate methods in plant ecology. (I) Association analysis in plant communities. *J. Ecology*, 47, p 83-101.

Williams W. T., Lance G. N. (1965) - Logic of computer based intrinsic classifications. *Nature*, 207, p 159-161.

Wishart D. (1969) - Mode analysis : a generalization of nearest neighbour which reduces chaining effects. *Numerical Taxonomy* (A.J. Cole ed.) p 282-311, Academic Press, London, .

Wishart J. (1928) - The generalized product-moment distribution in samples from a normal multivariate population. *Biometrika*, 20A, p 32-43.

Wold H. (1966) - Estimation of principal components and related models by iterative least squares, in *Multivariate analysis*, Krishnaiah P.R. (Ed.), Academic Press, New York, pp391-420.

Wold H. (1976) - Pattern recognition by means of disjoint principal component models. *Pattern Recognition*, 8, p 127-139.

Wold H. (1978) - Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, 20, p 397-405.

Wong M. A. (1985) - A bootstrap testing procedure for investigating the number of subpopulations. *J. Statist. Comput. and Simul.*, 22, p 99-112.

Wong M.A. (1982) - A hybrid clustering method for identifying high density clusters. *J. of Amer. Statist. Assoc.*, 77, p 841-847.

Worsley K. J. (1987) - Un exemple d'identification d'un modèle log-linéaire grâce à une analyse des correspondances. *Revue de Statist. Appl.* 35, p 13-20.

Yenyukov I. S. (1988) - Detecting structure by mean of projection pursuit. *COMPSTAT Proceedings*, Physica Verlag, Heidelberg, p 47-58.

Young G. A. (1994) - Bootstrap : More than a stab in the dark. *Statistical Science.* 9, p 382-418.

Zarraga A., Goitisolo B. (2002) – Méthode factorielle pour l'analyse simultanée de tableaux de contingence. *Revue de Statist. Appl.* vol L, p 47-70.

Zighed D. A., Rakotomalala R. (2000) *Graphes d'induction, apprentissage et Data Mining*, Hermès, Paris.

Index des auteurs

D

E

F

Index des matières

49616 – (VI) – OSB 80° – AUT – SPU
Dépôt légal . juin 2006 – Suite du tirage . octobre 2017

Achevé d'imprimer par Dupli-Print à Domont (95)
N° d'impression : 2017101521
www.dupli-print.fr

Imprimé en France

SCIENCES SUP

4e édition

Ludovic Lebart
Marie Piron
Alain Morineau

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisation et inférence en fouilles de données

Cette quatrième édition entièrement refondue et complétée s'adresse aux étudiants, chercheurs, ingénieurs, professeurs de toutes disciplines confrontés dans leurs travaux aux recueils de données multidimensionnelles. Les enquêtes socio-économiques, épidémiologiques et de marketing en sont des exemples courants.

Appuyé sur de nombreux exemples, l'ouvrage présente les concepts de base et les fondements des méthodes exploratoires et rend compte des développements récents. Il insiste sur la place centrale, dans la démarche « Fouille de données » (ou *Data Mining*), des visualisations fondées sur des principes géométriques et algébriques simples, sous le contrôle de méthodes inférentielles robustes.

Le livre peut être lu à plusieurs niveaux : celui de l'étudiant (Master, écoles d'ingénieur), celui du praticien, celui de l'utilisateur exigeant, enfin celui du chercheur en méthodologie statistique.

LUDOVIC LEBART est directeur de recherche CNRS à l'École nationale supérieure des télécommunications (ENST).

MARIE PIRON est chargée de recherche à l'Institut de recherche pour le développement (IRD).

ALAIN MORINEAU ancien directeur du Centre international de statistiques et d'informatique appliquées (CISIA), dirige la revue électronique MODULAD.

MATHÉMATIQUES
PHYSIQUE
CHIMIE
SCIENCES DE L'INGÉNIEUR
INFORMATIQUE
SCIENCES DE LA VIE
SCIENCES DE LA TERRE

LICENCE | MASTER | DOCTORAT
1 2 3 4 5 6 7 8

9 782100 496167

6465108
ISBN 978-2-10-049616-7

www.dunod.com

DUNOD