UNIVERSITE
PAUL
SABATIER
TOULOUSE III

PUBLICATIONS DU LABORATOIRE
DE
STATISTIQUE ET PROBABILITÉS

Data mining I
Exploration Statistique

ALAIN BACCINI & PHILIPPE BESSE

Version septembre 2005 — mises à jour : `www.lsp.ups-tlse.fr/Besse`

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

Avant-propos

Motivations du *data mining*

Le développement des moyens informatiques et de calcul permet le stockage (bases de données), le traitement et l'analyse d'ensembles de données très volumineux. Plus récemment, le perfectionnement des interfaces offrent aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples des outils logiciels. Cette évolution, ainsi que la popularisation de nouvelles méthodes algorithmiques (réseaux de neurones, support vector machine...) et outils graphiques, conduit au développement et à la commercialisation de logiciels intégrant un sous-ensemble de méthodes statistiques et algorithmiques sous la terminologie de *Data Mining* : la prospection ou *fouille de données*. Cette approche, issue du marketing spécialisé dans la gestion de la relation client (GRC) (*client relation management* ou CRM) trouve également des développements et applications industrielles en contrôle de qualité ou même dans certaines disciplines scientifiques dès lors que les ingénieurs et chercheurs sont confrontés à un volume de données important. Besse et col. (2001) présente une introduction détaillée de cette démarche et des relations qu'elle entretien avec les disciplines traditionnelles Statistique et Informatique. L'accroche publicitaire souvent citée par les éditeurs de logiciels (SAS) est :

Comment trouver un diamant dans un tas de charbon sans se salir les mains.

Nous proposons d'évaluer et d'expérimenter la réalité de cette annonce qui s'adresse à un marché en pleine expansion. Les entreprises sont en effet très motivées pour tirer parti et amortir, par une aide à la décision quantifiée, les coûts de stockage des teras octets que leur service informatique s'emploie à administrer.

Le contexte informationnel de la fouille de données est celui des *data wharehouses*. Un entrepôt de données, dont la mise en place est assuré par un gestionnaire de données (data manager) est un ensemble de bases relationnelles extraites des données brutes de l'entreprise et relatives à une problématique :

- gestion des stocks (flux tendu), des ventes d'un groupe afin de prévoir et anticiper au mieux les tendances du marché,
- suivi des fichiers clients d'une banque, d'une assurance, associés à des données socio-économiques (INSEE), à l'annuaire, en vue de la constitution d'une segmentation (typologie) pour cibler des opérations de marketing ou des attributions de crédit. La *gestion de la relation client* vise à une individualisation ou personnalisation de la production et de la communication afin d'évacuer la notion de *client moyen*.
- recherche, spécification puis ciblage de *niches* de marché les plus profitables (banque) ou au contraire les plus risquées (assurance) ;
- suivi en ligne des paramètres de production (traçabilité) en contrôle de qualité pour détecter au plus vite l'origine d'une défaillance ;
- prospection textuelle (*text mining*) et veille technologique ;

- *web mining* et comportement des internautes ;
- ...

Cet environnement se caractérise par

- une informatique hétérogène faisant intervenir des sites distants (Unix, Dos, NT, VM...) à travers le réseau de l'entreprise (intranet) ou même des accès extérieurs (internet). Des contraintes d'efficacité, de fiabilité ou de sécurité conduisent à répartir, stocker l'information à la source plutôt qu'à la dupliquer systématiquement ou à la centraliser.
- L'incompatibilité logique des informations observées sur des échantillons différents ne présentant pas les mêmes strates, les mêmes codifications.
- Des volumes et flux considérables de données issues de saisies automatisées et chiffrés en téra-octets.
- Contrairement à une démarche statistique traditionnelle (planification de l'expérience), les données analysées sont stochées à d'autres fins (comptabilité, contrôle de qualité...) et sont donc *préalables* à l'anyle.
- La nécessité de ne pas exclure *a priori* un traitement *exhaustif* des données afin de ne pas laisser échapper, à travers le crible d'un *sondage*, des groupes de faibles effectifs mais à fort impact économique.

Stratégie du *data mining*

Dans tout ce qui suit, nous disposons d'un ensemble d'observations. Les caractéristiques ou variables $X = (X^1, \ldots, X^p)$ dites explicatives ont été observées sur un ensemble de n objets, individus ou unités statistiques. Un premier travail, souvent fastidieux mais incontournable, consiste à mener une exploration statistique de ces données : allure des distributions, présence de données atypiques, corrélations et cohérence, transformations éventuelles des données, description multidimensionnelle, classification. C'est l'objet de la première partie de ce cours. La deuxième partie décrit les outils de modélisation statistique ou encore d'apprentissage utilisables pour la modélisation à fin de prédiction d'une variable *cible* Y par les variables explicatives X^j.

L'enchaînement de ces étapes (exploration puis apprentissage) constitue le fondement de la fouille de données.

Pour comprendre la structure et bien appréhender le contenu de ce cours, il est important d'intégrer rapidement ce qu'est la stratégie à mettre en œuvre pour aboutir au bon *apprentissage* ou encore au bon *modèle prédictif* recherché à partir des données observées.

Attention, contrairement à une démarche statistique traditionnelle dans laquelle l'observation des données est intégrée à la méthodologie (plannification de l'expérience), les données sont ici *préalables* à l'analyse. Néanmoins il est clair que les préoccupations liées à leur analyse et à son objectif doivent intervenir le plus en amont possible pour s'assurer quelques chances de succès.

Les étapes de la fouille de données :

i. Extraction des données avec ou sans échantillonnage faisant référence à des techniques de sondage appliquées ou applicables à des bases de données.

ii. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour l'étude des distributions des structures de corrélation, recherche de typologies, pour des transformations des données...

iii. Partition aléatoire de l'échantillon (apprentissage, validation, test) en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prédiction en vue des choix de modèle, choix et certification de méthode.

iv. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial ou poissonien), discrimination paramétrique (linéaire ou quadratique) ou non paramétrique, k plus proches voisins, arbre, réseau de neurones (perceptron), support vecteur machine, combinaison de modèles (bagging, boosting).
   - estimer le modèle pour une valeur donnée d'un paramètre de *complexité* : nombre de variables, de voisins, de feuilles, de neurones, durée de l'apprentissage, largeur de fenêtre... ;
   - optimiser ce paramètre (sauf pour les combinaisons de modèles affranchies des problèmes de sur-apprentissage) en fonction de la technique d'estimation de l'erreur retenue : échantillon de validation, validation croisée, approximation par pénalisation de l'erreur d'ajustement.

v. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prévision sur l'échantillon test ou, si la présence d'un échantillon test est impossible, sur le critère de pénalisation de l'erreur (Akaîke par exemple) s'il en existe une version pour chacune des méthodes considérées.

vi. Itération éventuelle de la démarche précédente (valisation croisée), si l'échantillon test est trop réduit, depuis (iii). Partitions aléatoires successives de l'échantillon pour moyenner sur plusieurs cas l'estimation finale de l'erreur de prédiction et s'assurer de la robustesse du modèle obtenu.

vii. Choix de la méthode retenue en fonction de ses capacités de prédiction, de sa robustesse mais aussi, éventuellement, de l'interprétabillité du modèle obtenu.

Objectif

L'objet de ce cours est d'introduire, sous une forme homogène et synthétique, les principales techniques d'exploration, de modélisation ou encore d'apprentissage utilisées le plus couramment en fouille de données et citées dans la section précédente. Il a fallu faire des choix dans l'ensemble des techniques proposées et leurs nombreux avatars. La forme et le contenu sont guidés par les besoins exprimés lors des stages réalisées par les étudiants du Master professionnel de Statistique & Econométrie ou encore par les thèmes des collaborations industrielles du laboratoire de Statistique et Probabilités[1]. Le lecteur peut se faire une idée du nombre très important de méthodes et variantes concernées par l'apprentissage supervisée ou non supervisé en consultant une boîte à outil Mathlab de classification[2]. Remarquons que les principaux logiciels commerciaux (SAS, Splus, SPSS, Matlab...) ou gratuits (R), performants et s'imposant par des interfaces très conviviales (Enterprise Miner, Insightfull Miner, Clementine), contribuent largement à la diffusion, voire la pénétration, de méthodes très sophistiquées dans des milieux imperméables à une conceptualisation mathématique trop abstraite.

Le choix a été fait de conserver et expliciter, dans la mesure du possible, les concepts originaux de chaque méthode dans son cadre disciplinaire tout en tâchant d'homogénéiser notations et terminologies. L'objectif principal est de faciliter la compréhension et l'interprétation des techniques des principaux logiciels pour en faciliter une *utilisation pertinente et réfléchie*. Un exemple élémentaire de recherche d'un score d'appétance issu du marketing bancaire illustre les différents points abordés. Traité avec les logiciels SAS, Splus ou R, il sert de "fil rouge" tout au long du cours.

[1] http ://www.lsp.ups-tlse.fr
[2] http ://tiger.technion.ac.il/ eladyt/classification/

Chapitre 1

Introduction

1 Objectif

Toute étude sophistiquée d'un corpus de données doit être précédée d'une étude *exploratoire* à l'aide d'outils, certes rudimentaires mais robustes, en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec des données et surtout de dépister les sources de problèmes :

- valeurs manquantes, erronées ou atypiques,
- modalités trop rares,
- distributions "anormales" (dissymétrie, multimodalité, épaisseur des queues),
- incohérences, liaisons non linéaires.
- ...

C'est ensuite la recherche de prétraitements des données afin de les rendre conformes aux techniques de modélisation ou d'apprentissage qu'il sera nécessaire de mettre en œuvre afin d'atteindre les objectifs fixés :

- transformation : logarithme, puissance, réduction, rangs... des variables,
- codage en classe ou recodage de classes,
- imputations ou non des données manquantes,
- lissage, décompositions (ondelettes, fourier) de courbes,
- réduction de dimension, classification et premier choix de variables,
- classification ou typologie des observations.

Attention, le côté rudimentaire voire trivial de ces outils ne doit pas conduire à les négliger au profit d'une mise en œuvre immédiate de méthodes beaucoup plus sophistiquées, donc beaucoup plus sensibles aux problèmes cités ci-dessus. S'ils ne sont pas pris en compte, ils réapparaîtront alors comme autant d'*artefacts* susceptibles de dénaturer voire de fausser toute tentative de modélisation.

2 Contenu

Cette partie se propose tout d'abord d'introduire brièvement les techniques permettant de résumer les caractéristiques (tendance centrale, dispersion, boîte à moustaches, histogramme, estimation non paramétrique) d'une variable statistique ou les relations entre variables de même type quantitatif (coefficient de corrélation, nuage de points, ou qualitatif (χ^2, Cramer, Tchuprow) ou de types différents (rapport de corrélation, diagrammes en boîtes parallèles). Les notions présentées sont illustrées sur un jeu de données typique d'un score d'appétance en marketing bancaire.

Après cette approche uni et bidimensionnelle, les techniques multidimensionnelles[1] sont décrites et illustrées. Elles diffèrent selon le type des variables considérées mais permettent toutes de réduire la dimension afin de résumer un tableau $(n \times p)$ de grande dimension et révéler ses caractéristiques. L'analyse en composantes principales (ACP) pour les variables quantitatives, l'analyse des correspondances simples ou multiples (AFCM) pour les variables qualitatives. L'analyse factorielle discriminante (AFD) permet de juger de la qualité de discrimination d'un ensemble de variables quantitatives afin d'expliquer une typologie décrite par une variable qualitative. Lorsqu'une typologie est recherchée, les méthodes de classification (hiérarchiques ou par réallocation dynamique) déterminent une variable qualitative définissant une partition de l'ensemble des données. D'autres techniques sont plus spécifiques, le positionnement multidimensionnel ou ACP sur tableau de distances est adapté à des données particulières mais permet également de structurer un ensemble de variables trop important. Enfin, ce document se termine par une introduction à l'étude exploratoire de données fonctionnelles illustrées par des exemples de séries climatiques.

[1] Elles constituent un ensemble communément appelé en France "Analyse de Données".

Chapitre 2

Description statistique élémentaire

1 Exemple de données

Un même ensemble de données bancaires[1] va servir à illustrer la plupart des outils et méthodes décrits dans ce document. En voici le descriptif sommaire.

Le service marketing d'une banque dispose de fichiers décrivant ses clients et leurs comportements (mouvements, soldes des différents comptes). Deux types d'études sont habituellement réalisées sur des données bancaires ou même plus généralement dans le tertiaire afin de personnaliser les relations avec les clients.

i. une classification ou *segmentation* de la clientèle permettant de déterminer quelques classes ou segments de comportements types.

ii. l'estimation d'un score en vue d'un objectif particulier. Il s'agit ici de prévoir l'intérêt ou l'*appétence* d'un client pour le produit bancaire *carte Visa Premier*. C'est une carte de paiement haut de gamme qui cherche à renforcer le lien de proximité avec la banque en vue de fidéliser une clientèle aisée.

La liste des variables est issue d'une base de données retraçant l'historique mensuel bancaire et les caractéristiques de tous les clients. Un sondage a été réalisé afin d'alléger les traitements ainsi qu'une première sélection de variables. Les variables contenues dans le fichier sont explicitées dans le tableau 2.1. Elles sont observées sur un échantillon de 1425 clients.

2 Introduction

l'objectif des outils de Statistique descriptive élémentaire est de fournir des résumés synthétique de séries de valeurs, adaptés à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon.

Dans le cas d'une seule variable, Les notions les plus classiques sont celles de médiane, quantile, moyenne, fréquence, variance, écart-type définies parallèlement à des représentations graphiques : diagramme en bâton, histogramme, diagramme-boîte, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.

Dans le cas de deux variables, on s'intéresse à la corrélation, au rapport de corrélation ou encore à la statistique d'un test du χ^2 associé à une table de contingence. Ces notions sont associées à différents graphiques comme le nuage de points (scatterplot), les diagrammes-boîtes parallèles, les diagrammes de profils ou encore en mosaïque.

[1] Merci à Sophie Sarpy de Informatique Banque Populaire à Balma pour la mise à disposition de ces données.

TAB. 2.1 – Libellés des variables des données bancaires.

Identif.	**Libellé**
matric	Matricule (identifiant client)
depts	Département de résidence
pvs	Point de vente
sexec	Sexe (qualitatif)
ager	Age en années
famil	Situation familiale
	(Fmar : marié, Fcel : célibataire, Fdiv :divorcé,
	Fuli :union libre, Fsep : séparé de corps, Fveu :veuf)
relat	Ancienneté de relation en mois
prcsp	Catégorie socio-professionnelle (code num)
quals	Code “qualité” client évalué par la banque
GxxGxxS	plusieurs variables caractérisant les interdits
	bancaires
impnbs	Nombre d’impayés en cours
rejets	Montant total des rejets en francs
opgnb	Nombre d’opérations par guichet dans le mois
moyrv	Moyenne des mouvements nets créditeurs
	des 3 mois en Kf
tavep	Total des avoirs épargne monétaire en francs
endet	Taux d’endettement
gaget	Total des engagements en francs
gagec	Total des engagements court terme en francs
gagem	Total des engagements moyen terme en francs
kvunb	Nombre de comptes à vue
qsmoy	Moyenne des soldes moyens sur 3 mois
qcred	Moyenne des mouvements créditeurs en Kf
dmvtp	Age du dernier mouvement (en jours)
boppn	Nombre d’opérations à M-1
facan	Montant facturé dans l’année en francs
lgagt	Engagement long terme
vienb	Nombre de produits contrats vie
viemt	Montant des produits contrats vie en francs
uemnb	Nombre de produits épargne monétaire
uemmts	Montant des produits d’épargne monétaire en francs
xlgnb	Nombre de produits d’épargne logement
xlgmt	Montant des produits d’épargne logement en francs
ylvnb	Nombre de comptes sur livret
ylvmt	Montant des comptes sur livret en francs
nbelts	Nombre de produits d’épargne long terme
mtelts	Montant des produits d’épargne long terme en francs
nbcats	Nombre de produits épargne à terme
mtcats	Montant des produits épargne à terme
nbbecs	Nombre de produits bons et certificats
mtbecs	Montant des produits bons et certificats en francs
rocnb	Nombre de paiements par carte bancaire à M-1
jntca	Nombre total de cartes
nptag	Nombre de cartes point argent
segv2s	Segmentation version 2
itavc	Total des avoirs sur tous les comptes
havef	Total des avoirs épargne financière en francs
dnbjd1s	Nombre de jours à débit à M
dnbjd2s	Nombre de jours à débit à M-1
dnbjd3s	Nombre de jours à débit à M-2
carvp	**Possession de la carte VISA Premier**

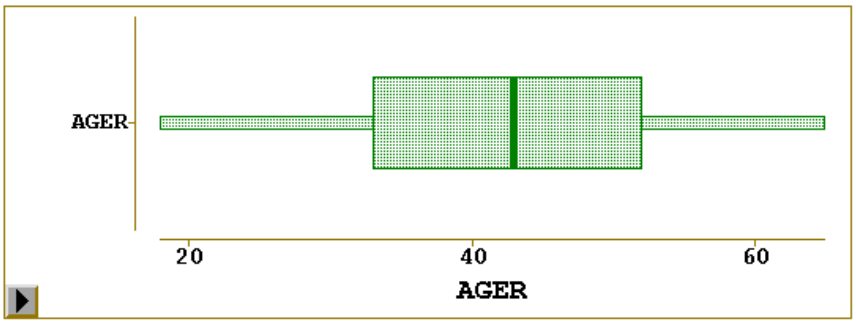

FIG. 2.1 – Diagramme-boîte illustrant la distribution des âges des clients.

Les définitions de ces différentes notions se trouvent dans n'importe quel ouvrage élémentaire de Statistique[2], nous nous proposons simplement de rappeler dans ce chapitre certains outils moins classiques mais efficaces et présents dans la plupart des logiciels statistiques. Cela nous permettra également d'illustrer les premières étapes exploratoires à réaliser sur un jeu de données.

3 Decription d'une variable

3.1 Cas quantitatif

Une variable quantitative prend des valeurs entières ou réelles, elle est dite alors discrète ou continue. Cette propriété ayant des incidences sur la nature de sa distribution et donc sur les graphiques associés. Nous nous intéresserons surtout aux variables continues.

La distribution d'un variable statistique quantitative est résumée par différents indicateurs empiriques de *tendance centrale* (moyenne $\overline{x} = \sum_{i=1}^{n} w_i x_i$, médiane) ou de *dispersion* (écart-type σ, intervalle inter-quartiles). D'autres indicateurs s'intéressent à la dissymétrie (skeeness, associée au moment d'ordre 3) ou encore à l'applatissement (kurtosis à partir du moment d'ordre 4)

Deux graphiques permettent de rendre compte précisément de la nature de la distribution. La statistique de Kolmogorov est la plus couramment utilisée pour tester l'adéquation à une loi (normale).

Diagramme-boîte (box-and-whiskers plot)

Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane.

Histogramme

Dans le cas d'un échantillon, on cherche à approcher par une estimation empirique le graphe de la densité de la loi théorique associée à la population. L'*histogramme* en est un exemple. Une fois déterminée un découpage en classes de l'ensemble des valeurs et les fréquences f_ℓ d'occurences de ces classes, un histogramme est la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_\ell = b_\ell - b_{\ell-1}$) et dont les hauteurs sont les quantités $\frac{f_\ell}{b_\ell - b_{\ell-1}}$, appelées *densités de fréquence*. L'aire du ℓ-ème rectangle vaut donc f_ℓ, fréquence de la

[2] Un support de cours accessible à la page `www-sv.cict.fr/lsp/Besse`.

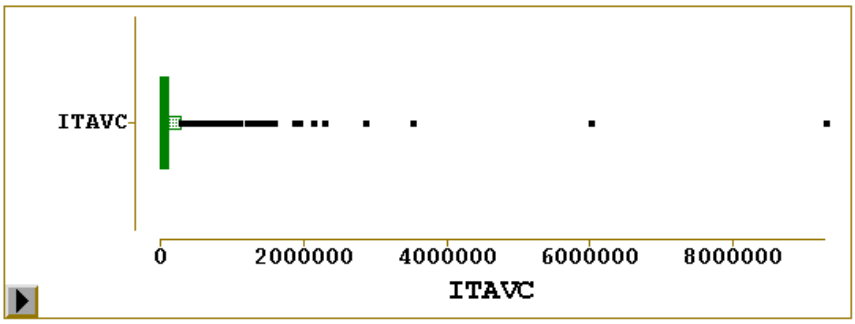

FIG. 2.2 – Diagramme-boîte illustrant la distribution de la variable cumulant les totaux des avoirs. Celle-ci apparaît comme très dissymétrique et avec de nombreuses valeurs atypiques. Une transformation s'impose.

classe correspondante.

Estimation fonctionnelle

La qualité de l'estimation d'une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d'effectifs égaux et donc de mieux répartir l'imprécision, les logiciels utilisent des classes d'amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d'estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d'estimer la distribution d'une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l'ordinateur) mais nécessitent le choix d'un paramètre dit de *lissage*. Les démonstrations du caractère optimal de ces estimations fonctionnelles, liée à l'optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank, 1988, Silverman, 1986).

L'estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

où λ est le paramètre de lissage optimisée par une procédure automatique qui minimise une approximation de l'erreur quadratique moyenne intégrée (norme de l'espace L^2) ; K est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C'est souvent la fonction densité de la loi gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un "élément de densité" de la forme du noyau K et à sommer tous ces éléments. Un histogramme est une version particulière d'estimation dans laquelle l"'élément de densité" est un "petit rectangle" dans la classe de l'observation.

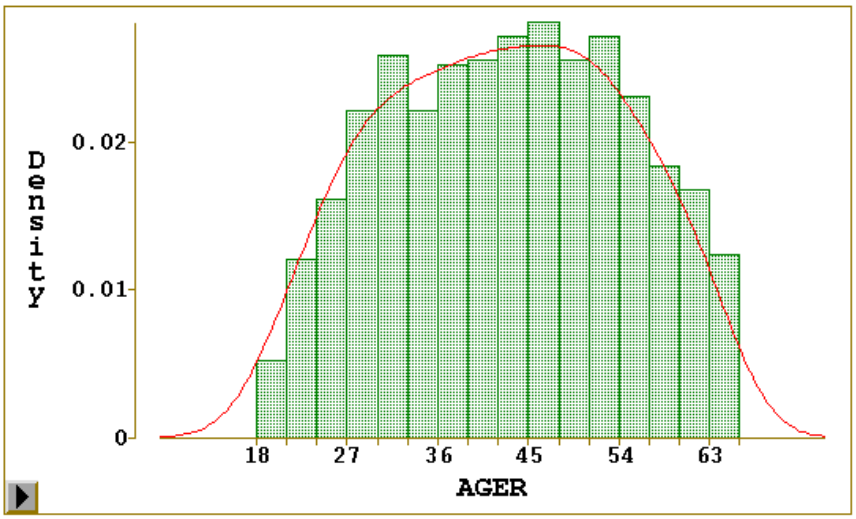

FIG. 2.3 – Histogramme et estimation fonctionnelle par la méthode du noyau de la distribution des âges.

3.2 Cas qualitatif

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées *modalités*. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac ou une classe d'âge), la variable est dite *ordinale*. Dans le cas contraire (par exemple, la profession dans une population de personnes actives ou la situation familiale) la variable est dite *nominale*.

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont les diagrammes en colonnes, en barre, en secteurs. Tous visent à représenter la répartition en effectif ou fréquences des individus dans les différentes classes ou modalités.

4 Liaison entre variables

Dans cette section, on s'intéresse à l'étude simultanée de deux variables X et Y. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors *liaison*. Dans certains cas, cette liaison peut être considérée *a priori* comme *causale*, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité. Sont ainsi introduites les notions de covariance, coefficient de corrélation linéaire, régression linéaire, rapport de corrélation, indice de concentration, khi-deux et autres indicateurs qui lui sont liés. De même, nous présentons les graphiques illustrant les liaisons entre variables : nuage de points (*scatter-plot*), diagrammes-oîtes parallèles, diagramme de profils, tableau de nuages (*scatter-plot matrix*).

4.1 Deux variables quantitatives

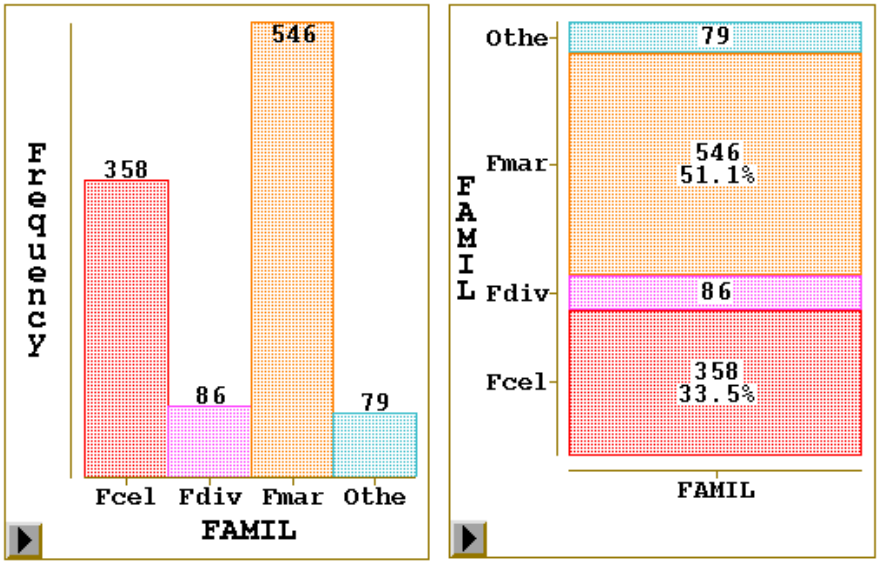

FIG. 2.4 – Diagramme en barres et diagramme en colonne de la répartition des situations familiales. Certaines modalités trop rares et regroupées automatiquement dans la classe *other* devront être recodées.

Nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé par les coordonnées des valeurs observées. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction plus fidèle de l'anglais *scatter-plot*.

Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables *homogènes* (représentant la même grandeur et exprimées dans la même unité) de celui des variables *hétérogènes*. Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).

Indice de liaison

le coefficient de corrélation linéaire est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Il est défini à partir de la covariance qui généralise à deux variables la notion de variance :

$$\begin{aligned} \mathrm{cov}(X,Y) &= \sum_{i=1}^{n} w_i [x_i - \overline{x}][y_i - \overline{y}] \\ &= [\sum_{i=1}^{n} w_i x_i y_i] - \overline{x}\,\overline{y}. \end{aligned}$$

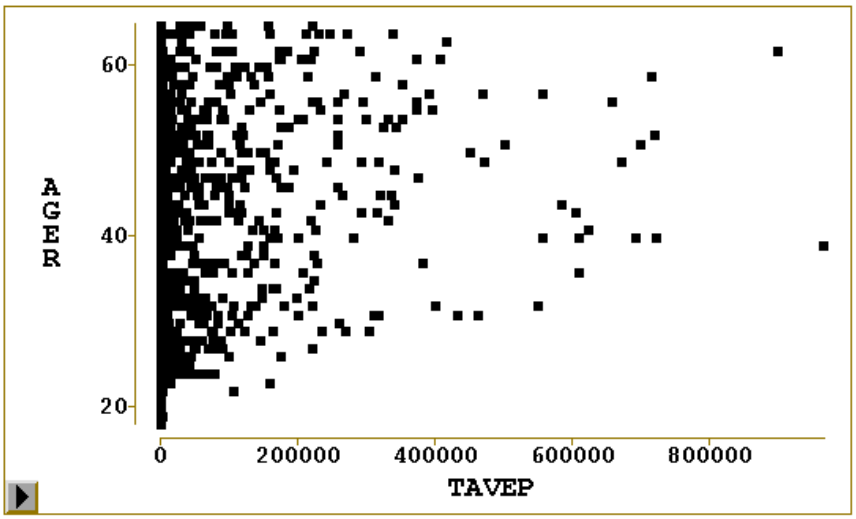

FIG. 2.5 – Nuage de points illustrant l'absence de liaison entre la variable âge et celle cumulant le total des épargnes monétaires (corrélation de 0,17).

La covariance est une forme bilinéaire symétrique qui peut prendre toute valeur réelle et dont la variance est la forme quadratique associée. Elle dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (parfois appelé coefficient de Pearson ou de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types :

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $\text{corr}(X,Y) = \text{cov}(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$. Par conséquent, $\text{corr}(X,Y)$ est indépendant des unités de mesure de X et de Y. Le coefficient de corrélation est *symétrique* et prend ses valeurs entre -1 et +1.

Notons pour mémoire la possibilité d'utiliser d'autres indicateurs de liaison entre variables quantitatives. Construits sur les rangs (corrélation de Spearman) ils sont plus robustes faces à des situations de non linéarité ou des valeurs atypiques mais restent très réducteurs.

4.2 Une variable quantitative et une qualitative

Notations

Soit X la variable qualitative considérée, supposée à r modalités notées

$$x_1, \ldots, x_\ell, \ldots, x_r$$

et soit Y la variable quantitative de moyenne $\overline{y}$ et de variance σ_Y^2. Désignant par Ω l'échantillon considéré, chaque modalité x_ℓ de X définit une sous-population (un sous-ensemble) Ω_ℓ de Ω : c'est l'ensemble des individus, supposés pour simplifier de poids $w_i = 1/n$ et sur lesquels on a observé x_ℓ ; on obtient ainsi une *partition* de Ω en m classes dont nous noterons $n_1, \ldots, n_m$ les cardinaux (avec toujours $\sum_{\ell=1}^{m} n_\ell = n$, où $n = card(\Omega)$).

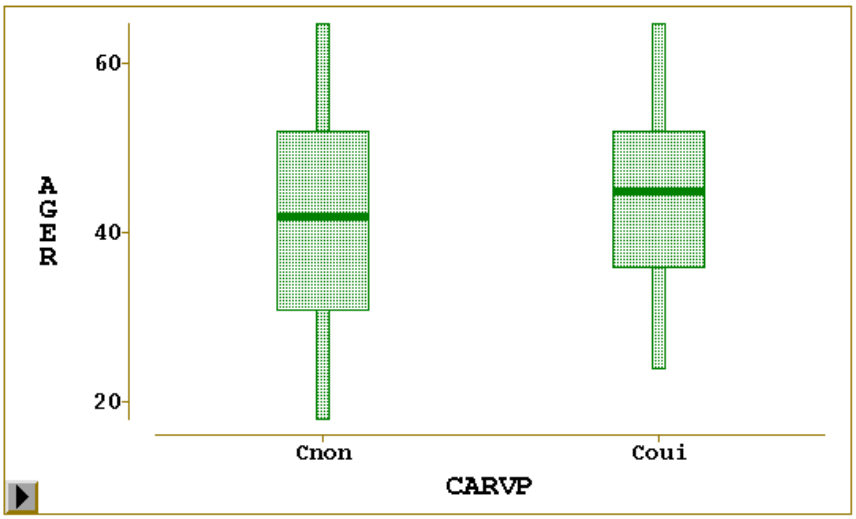

FIG. 2.6 – Diagrammes-boites illustrant les différences de distribution des âges en fonction de la possession d'une carte Visa Premier.

Considérant alors la restriction de Y à Ω_ℓ ($l = 1, \ldots, m$), on peut définir la moyenne et la variance partielles de Y sur cette sous-population ; nous les noterons respectivement $\overline{y_\ell}$ et σ_ℓ^2 :

$$\overline{y_\ell} = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,;$$

$$\sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

Boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des diagrammes-boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y un diagramme-boîte pour chacune des sous-populations définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y, c'est-à-dire de la liaison entre les deux variables.

Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-à-dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables.

$$\begin{aligned} \overline{y} &= \frac{1}{n} \sum_{\ell=1}^{r} n_\ell \overline{y_\ell}\,; \\ \sigma_Y^2 &= \frac{1}{n} \sum_{\ell=1}^{r} n_\ell (\overline{y_\ell} - \overline{y})^2 \;+\; \frac{1}{n} \sum_{\ell=1}^{r} n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,. \end{aligned}$$

Le premier terme de la décomposition de σ_Y^2, noté σ_E^2, est appelé *variance expliquée* (par la partition, c'est-à-dire par X) ou *variance inter* (between) ; le second terme, noté σ_R^2, est appelé *variance résiduelle* ou *variance intra* (within).

Rapport de corrélation

Il s'agit d'un indice de liaison entre les deux variables X et Y qui est défini par :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}} ;$$

X et Y n'étant pas de même nature, $s_{Y/X}$ n'est pas symétrique et vérifie $0 \leq s_{Y/X} \leq 1$. Cet encadrement découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont une signification particulière intéressante.

4.3 Deux variables qualitatives

Notations

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

Ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Les quantités $n_{\ell +} (\ell = 1, \ldots, r)$ et $n_{+h} (h = 1, \ldots, c)$ sont appelées les *effectifs marginaux* ; ils sont définis par $n_{\ell +} = \sum_{h=1}^{c} n_{\ell h}$ et $n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}$, et ils vérifient $\sum_{\ell=1}^{r} n_{\ell +} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

Représentations graphiques

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.

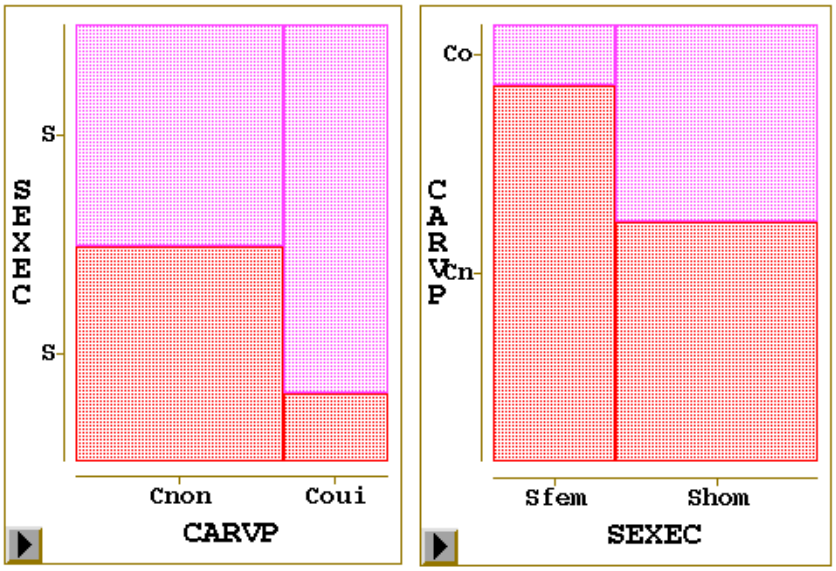

FIG. 2.7 – Diagrammes en barres des profils lignes et colonnes de la table de contingence croisant le sexe et la possession de la carte Visa Premier. La superficie de chaque case est en plus proportionnelle à l'effectif de la cellule associée.

Profils

On appelle ℓ-ème profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité x_ℓ de X (c'est-à-dire définies au sein de la sous-population Ω_ℓ de Ω associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\}.$$

On définit de façon analogue le h-ème profil-colonne :

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

La représentation graphique des profils-lignes ou des profils-colonnes, au moyen, par exemple, de diagrammes en barre parallèles, donne alors une idée assez précise de la variation conjointe des deux variables.

Indices de liaison

Lorsque tous les profils-lignes sont égaux, ce qui est équivalent à ce que tous les profils-colonnes soient égaux et que

$$\forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\} \ : n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n} \ ,$$

on dit qu'il n'existe aucune forme de liaison entre les deux variables considérées X et Y. Par suite, la mesure de la liaison va se faire en évaluant l'écart entre la situation observée et l'état de non liaison défini ci-dessus.

Khi-deux

Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique $n_{\ell h}$, à une table de contingence donnée a priori (et appelée *standard*), d'effectif conjoint générique $s_{\ell h}$, en calculant la quantité

$$\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{(n_{\ell h}-s_{\ell h})^2}{s_{\ell h}}\,.$$

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise donc l'indice appelé khi-deux (chi-square) et défini comme suit :

$$\chi^2=\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{(n_{\ell h}-\frac{n_{\ell+}n_{+h}}{n})^2}{\frac{n_{\ell+}n_{+h}}{n}}=n\left[\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{n_{\ell h}^2}{n_{\ell+}n_{+h}}-1\right].$$

Le coefficient χ^2 est toujours positif ou nul et il est d'autant plus grand que la liaison entre les deux variables considérées est forte. Malheureusement, il dépend aussi des dimensions r et c de la table étudiée, ainsi que de la taille n de l'échantillon observé ; en particulier, il n'est pas majoré. C'est la raison pour laquelle on a défini d'autres indices, liés au khi-deux, et dont l'objectif est de palier ces défauts.

Autres indicateurs

Nous en citerons trois.

- Le *phi-deux* : $\Phi^2=\frac{\chi^2}{n}$. Il ne dépend plus de n, mais dépend encore de r et de c.
- Le coefficient T de Tschuprow :

$$T=\sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

  On peut vérifier : $0\leq T\leq 1$.
- Le coefficient C de Cramer :

$$C=\sqrt{\frac{\Phi^2}{d-1}},$$

  avec : $d=\inf(r,c)$. On vérifie maintenant : $0\leq T\leq C\leq 1$.

Enin, la p-value d'un test d'indépendance (test du χ^2) est aussi utilisée pour compararr des liaisons entre variables.

5 Vers le cas multidimensionnel

L'objectif des prochains chapitres de ce cours est d'exposer les techniques de la statistique descriptive multidimensionnelle. Or, sans connaître ces techniques, il se trouve qu'il est possible de débuter une exploration de données multidimensionnelles en adaptant simplement les méthodes déjà étudiées.

5.1 Matrices des covariances et des corrélations

Lorsqu'on a observé simultanément plusieurs variables quantitatives (p variables, $p\geq 3$) sur le même échantillon, il est possible de calculer d'une part les variances de toutes ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités

peut alors être disposé dans une matrice carrée ($p \times p$) et symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale ; cette matrice, appelée matrice des variances-covariances (ou encore matrice des covariances) sera notée $\mathbf{S}$. Elle sera utilisée par la suite, mais n'a pas d'interprétation concrète. Notons qu'il est possible de vérifier que $\mathbf{S}$ est semi définie positive.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée matrice des corrélations, elle est également semi définie positive, et nous la noterons $\mathbf{R}$. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

5.2 Tableaux de nuages

Notons $X^1, \ldots, X^p$ les p variables quantitatives considérées ; on appelle tableau de nuages le graphique obtenu en juxtaposant, dans une sorte de matrice carrée $p \times p$, p^2 sous-graphiques ; chacun des sous-graphiques diagonaux est relatif à l'une des p variables, et il peut s'agir, par exemple, d'un histogramme ; le sous-graphique figurant dans le bloc d'indice $(j, j'), j \neq j'$, est le nuage de points réalisé avec la variable X^j en abscisses et la variable $X^{j'}$ en ordonnées. Dans certains logiciels anglo-saxons, ces graphiques sont appelés *splom* (Scatter PLOt Matrix). Le tableau de nuages, avec la matrice des corrélations, fournit ainsi une vision globale des liaisons entre les variables étudiées.

5.3 La matrice des coefficients de Tschuprow (ou de Cramer)

Considérons maintenant le cas où l'on étudie simultanément plusieurs variables qualitatives (p variables, $p \geq 3$). La matrice des coefficients de Tschuprow est la matrice carrée d'ordre p, symétrique, comportant des 1 sur la diagonale et, en dehors de la diagonale, les coefficients de Tschuprow entre les variables prises deux à deux. Il s'agit donc d'une matrice du même type que la matrice des corrélations (elle est d'ailleurs, elle aussi, semi définie positive), et son utilisation pratique est analogue. Notons que l'on peut, de la même façon, utiliser les coefficients de Cramer au lieu des coefficients de Tschuprow.

6 Problèmes

Les quelques outils de ce chapitre permettent déjà de se faire une première idée d'un jeu de données mais surtout, en préalable à toute analyse, ils permettent de s'assurer de la fiabilité des données, de repérer des valeurs extrêmes atypiques, éventuellement des erreurs de mesures ou de saisie, des incohérences de codage ou d'unité.

Les erreurs, lorsqu'elle sont décelées, conduisent naturellement et nécessairement à leur correction ou à l'élimination des données douteuses mais d'autres problèmes pouvant apparaître n'ont pas toujours de solutions évidentes.

- Le mitage de l'ensemble des données ou absence de certaines valeurs en fait partie. Faut-il supprimer les individus incriminés ou les variables ? Faut-il compléter, par une modélisation et prévision partielles, les valeurs manquantes ? Les solutions dépendent du taux de valeurs manquantes, de leur répartition (sont-elles aléatoires) et du niveau de tolérance des méthodes qui vont être utilisées.
- La présence de valeurs atypiques peut influencer sévèrement des estimations de méthodes peu robustes car basées sur le carré d'une distance. Ces valeurs sont-elles des erreurs ? Sinon faut-il les conserver en transformant les variables ou en adoptant des méthodes robustes

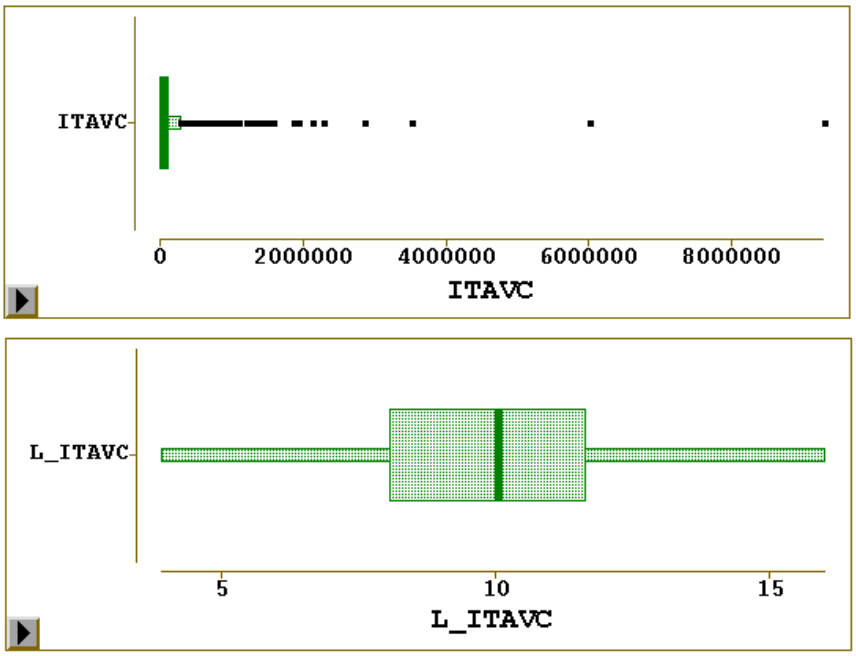

FIG. 2.8 – La simple transformation ($\log(50 + x)$), de la variable cumulants les avoirs, résout bien les problèmes posés par l'allure "log-normale" de sa distribution avec son cortège de valeurs atypiques.

basées sur des écarts absolus ?

- Même sans hypothèse explicite de normalité des distributions, il est préférable d'avoir à faire à des distributions relativement symétriques. Une transformation des variables par une fonction monotone (log, puissance) est hautement recommandée afin d'améliorer la symétrie de leur distribution ou encore pour linéariser (nuage de points) la nature d'une liaison.

Chapitre 3

Analyse en Composantes Principales

1 introduction

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

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

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

L'analyse en Composantes Principales (ACP) est un grand classique de l"'analyse des données" en France pour l'étude exploratoire ou la compression d'un grand tableau $n \times p$ de données quantitatives. Le livre de Jolliffe (2002) en détaille tous les aspects et utilisations de façon exhaustive. Elle est introduite ici comme l'estimation des paramètres d'un modèle, afin de préciser la signification statistique des résultats obtenus. Une approche plus sophistiquée adaptée à l'étude de courbes ou données fonctionnelles est proposée au chapitre 9. L'ACP est illustrée dans ce chapitre à travers l'étude de données élémentaires. Elles sont constituées des moyennes sur dix ans des températures moyennes mensuelles de 32 villes françaises. La matrice initiale $\mathbf{X}$ est donc (32×12). Les colonnes sont l'observation à différents instants d'une même variable ; elles sont homogènes et il est inutile de les réduire.

L'ACP joue dans ce cours un rôle central ; cette méthode sert de fondement théorique aux autres méthodes de statistique multidimensionnelle dites *factorielles* qui en apparaissent comme des cas particuliers. Cette méthode est donc étudiée en détail et abordée avec différents niveaux de lecture. La première section présente les grands principes de façon très élémentaire, voire intuitive, tandis que les suivantes explicitent les expressions matricielles des résultats.

2 Présentation élémentaire de l'ACP

2.1 Les données

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Nous savons comment analyser séparément chacune de ces 4 variables, soit en faisant un *graphique*, soit en calculant des *résumés numériques*. Nous savons également qu'on peut regarder les *liaisons entre 2 variables* (par exemple mathématiques et français), soit en faisant un graphique du type nuage de points, soit en calculant leur *coefficient de corrélation linéaire*, voire en réalisant la *régression* de l'une sur l'autre.

Mais comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique ? La difficulté vient de ce que les individus (les élèves) ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension 4 (chacun étant caractérisé par les 4 notes qu'il a obtenues). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple, ici, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir *le résumé le plus pertinent* des données initiales.

2.2 Résultats préliminaires

Tout logiciel fournit la moyenne, l'écart-type, le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'*études univariées*.

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Notons au passage la grande homogénéité des 4 variables considérées : même ordre de grandeur pour les moyennes, les écarts-types, les minima et les maxima.

Le tableau suivant est la *matrice des corrélations*. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'*analyses bivariées*, constituant un premier pas vers l'*analyse multivariée*.

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Remarquons que toutes les corrélations linéaires sont positives (ce qui signifie que toutes les variables varient, en moyenne, dans le même sens), certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres enfin plutôt faibles (0.40 et 0.23).

2.3 Résultats généraux

Continuons l'analyse par celui de la **matrice des variances-covariances**, matrice de même nature que celle des corrélations, bien que moins "parlante" (nous verrons néanmoins plus loin comment elle est utilisée concrètement). La diagonale de cette matrice fournit les variances des 4 variables considérées (on notera qu'au niveau des calculs, il est plus commode de manipuler la variance que l'écart-type ; pour cette raison, dans de nombreuses méthodes statistiques, comme en A.C.P., on utilise la variance pour prendre en compte la dispersion d'une variable quantitative).

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Les *valeurs propres* données ci-dessous sont celles de la matrice des variances-covariances.

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	-----	----	
	40.30	1.00	

Interprétation

Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (voilà les *facteurs*) dont la colonne VAL. PR. (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne PCT. VAR, ou pourcentage de variance, correspond

au pourcentage de variance de chaque ligne par rapport au total. La colonne PCT. CUM. représente le cumul de ces pourcentages.

Additionnons maintenant les variances des 4 variables initiales (diagonale de la matrice des variances-covariances) : $11.39 + 8.94 + 12.06 + 7.91 = 40.30$. La dispersion totale des individus considérés, en dimension 4, est ainsi égale à 40.30.

Additionnons par ailleurs les 4 valeurs propres obtenues : $28.23+12.03+0.03+0.01 = 40.30$. Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changé. Il s'agit d'un simple changement de base dans un espace vectoriel. C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouve modifiée : les 2 premiers facteurs restituent à eux seuls la quasi-totalité de la dispersion du nuage, ce qui permet de négliger les 2 autres.

Par conséquent, les graphiques en dimension 2 présentés ci-dessous résument presque parfaitement la configuration réelle des données qui se trouvent en dimension 4 : l'objectif (résumé pertinent des donnée en petite dimension) est donc atteint.

2.4 Résultats sur les variables

Le résultat fondamental concernant les variables est le tableau des **corrélations variables-facteurs**. Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui vont permettre de donner un sens aux facteurs (de les interpréter).

Corrélations variables-facteurs

FACTEURS -->	F1	F2	F3	F4
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le *graphique des variables* (version SAS) donné ci-dessous.

Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

On notera que les deux dernières colonnes ne seront pas utilisées puisqu'on ne retient que deux dimensions pour interpréter l'analyse.

Interprétation

Ainsi, on voit que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif. En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques. Cette interprétation peut être précisée avec les graphiques et tableaux relatifs aux individus que nous présentons maintenant.

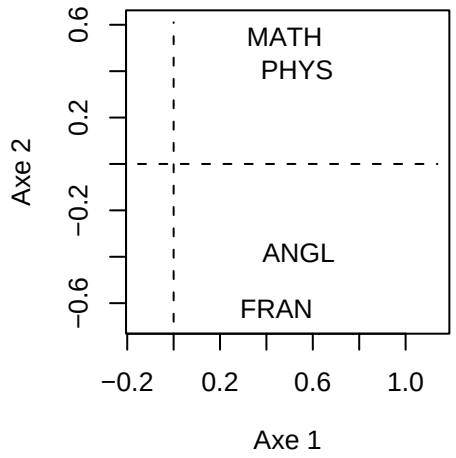

FIG. 3.1 – *Représentation des variables*

2.5 Résultats sur les individus

Le tableau ci-dessous contient tous les résultats importants sur les individus.

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de $1/9 = 0.11$, ce qui est fourni par la première colonne du tableau ci-dessus.

Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le **graphique des individus**. Ce dernier permet de préciser la signification des axes, donc des facteurs.

Interprétation

On peut ainsi voir que l'axe 1 représente le résultat d'ensemble des élèves (si on prend leur score – ou coordonnée – sur l'axe 1, on obtient le même classement que si on prend leur moyenne générale). Par ailleurs, l'élève "le plus haut" sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7 et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disci-

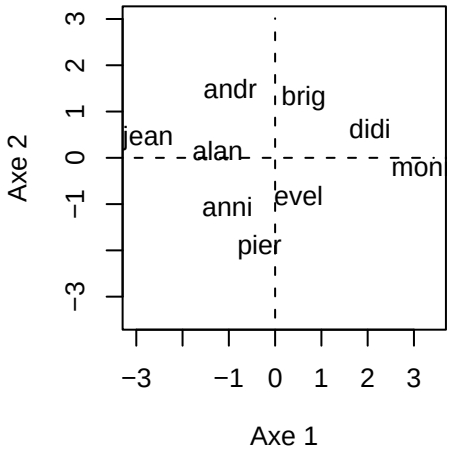

FIG. 3.2 – *Données fictives : Représentation des individus*

plines littéraires (7 et 5.5). On notera que Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1).

Les 3 colonnes suivantes du tableau fournissent des **contributions** des individus à diverses dispersions : CONT1 et CONT2 donnent les contributions des individus à la variance selon les axes 1 et 2 (rappelons que c'est la variance qui caractérise la dispersion) ; CONTG les contributions à la dispersion en dimension 4 (il s'agit de ce que l'on appelle l'**inertie** du nuage des élèves ; la notion d'inertie généralise celle de variance en dimension quelconque, la variance étant toujours relative à une seule variable). Ces contributions sont fournies en pourcentages (chaque colonne somme à 100) et permettent de repérer les individus les plus importants au niveau de chaque axe (ou du nuage en dimension 4). Elles servent en général à affiner l'interprétation des résultats de l'analyse.

Ainsi, par exemple, la variance de l'axe 1 vaut 28.23 (première valeur propre). On peut la retrouver en utilisant la formule de définition de la variance :

$$Var(C^1) = \frac{1}{9}\sum_{i=1}^{9}(c_i^1)^2$$

(il faut noter que, dans une A.C.P., les variables étant centrées, il en va de même pour les facteurs ; ainsi, la moyenne de C^1 est nulle et n'apparaît pas dans la formule de la variance). La coordonnée de Jean (le premier individu du fichier) sur l'axe 1 vaut $c_1^1 = -8.61$; sa contribution est donc :

$$\frac{\frac{1}{9}(-8.61)^2}{28.23} \times 100 = 29.19\ \%.$$

À lui seul, cet individu représente près de 30 % de la variance : il est prépondérant (au même titre que Monique) dans la définition de l'axe 1 ; cela provient du fait qu'il a le résultat le plus faible, Monique ayant, à l'opposé, le résultat le meilleur.

Enfin, les 2 dernières colonnes du tableau sont des cosinus carrés qui fournissent la (* qualité de la représentation *) de chaque individu sur chaque axe. Ces quantités s'additionnent axe par

axe, de sorte que, en dimension 2, Évelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.

Lorsqu'on considère les données initiales, chaque individu (chaque élève) est représenté par un vecteur dans un espace de dimension 4 (les éléments – ou coordonnées – de ce vecteur sont les notes obtenues dans les 4 disciplines). Lorsqu'on résume les données en dimension 2, et donc qu'on les représente dans un plan, chaque individu est alors représenté par la projection du vecteur initial sur le plan en question. Le cosinus carré relativement aux deux premières dimensions (par exemple, pour Évelyne, 0.98 ou 98 %) est celui de l'angle formé par le vecteur initial et sa projection dans le plan. Plus le vecteur initial est proche du plan, plus l'angle en question est petit et plus le cosinus, et son carré, sont proches de 1 (ou de 100 %) : la représentation est alors très bonne. Au contraire, plus le vecteur initial est loin du plan, plus l'angle en question est grand (proche de 90 degrés) et plus le cosinus, et son carré, sont proches de 0 (ou de 0 %) : la représentation est alors très mauvaise. On utilise les carrés des cosinus, parce qu'ils s'additionnent suivant les différentes dimensions.

3 Représentation vectorielle de données quantitatives

3.1 Notations

Soit p *variables statistiques* réelles X^j ($j = 1, \ldots, p$) observées sur n individus i ($i = 1, \ldots, n$) affectés des poids w_i :

$$\begin{aligned} \forall i = 1, \ldots, n \ : w_i &> 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ; \\ \forall i = 1, \ldots, n \ : x_i^j &= X^j(i), \text{ mesure de } X^j \text{ sur le } i^{ème} \text{ individu.} \end{aligned}$$

Ces mesures sont regroupées dans une matrice $\mathbf{X}$ d'ordre $(n \times p)$.

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

- À chaque individu i est associé le vecteur $\mathbf{x}_i$ contenant la i-ème ligne de $\mathbf{X}$ mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice $\mathbf{M}$ lui conférant une structure d'espace euclidien : E est isomorphe à $(\mathbb{R}^p, \mathcal{E}, \mathbf{M})$; E est alors appelé *espace des individus*.
- À chaque variable X^j est associé le vecteur $\mathbf{x}^j$ contenant la j-ème colonne *centrée* (la moyenne de la colonne est retranchée à toute la colonne) de $\mathbf{X}$. C'est un élément d'un espace vectoriel noté F de dimension n ; nous choisissons $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice $\mathbf{D}$ diagonale des *poids* lui conférant une structure d'espace euclidien : F est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$; F est alors appelé *espace des variables*.

3.2 Interprétation statistique de la métrique des poids

L'utilisation de la métrique des poids dans l'espace des variables F donne un sens très particulier aux notions usuelles définies sur les espaces euclidiens. Ce paragraphe est la clé permettant de fournir les interprétations en termes statistiques des propriétés et résultats mathématiques.

Moyenne empirique de X^j :	$\overline{x^j}$	$=$	$\left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n.$
Barycentre des individus :	$\overline{\mathbf{x}}$	$=$	$\mathbf{X}'\mathbf{D}\mathbf{1}_n.$
Matrice des données centrées :	$\overline{\mathbf{X}}$	$=$	$\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'.$
Ecart-type de X^j :	σ_j	$=$	$(\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}.$
Covariance de X^j et X^k :	$\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k$	$=$	$\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}.$
Matrice des covariances :	$\mathbf{S}$	$=$	$\sum_{i=1}^n w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}.$
Corrélation de X^j et X^k :	$\frac{\langle \mathbf{x}^j, \mathbf{x}^k \rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}} \|\mathbf{x}^k\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k).$

Attention : Par souci de simplicité des notations, on désigne toujours par $\mathbf{x}^j$ les colonnes de la matrice **centrée** $\overline{\mathbf{X}}$. On considère donc que des vecteurs "variables" sont toujours centrés.

Ainsi, lorsque les variables sont centrées et représentées par des vecteurs de F :
- la *longueur* d'un vecteur représente un *écart-type*,
- le *cosinus* d'un angle entre deux vecteurs représente une *corrélation*.

3.3 La méthode

Les objectifs poursuivis par une ACP sont :
- la représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace E_q de dimension q $(q < p)$,
- la représentation graphique des variables dans un sous-espace F_q en explicitant au "mieux" les liaisons initiales entre ces variables,
- la réduction de la dimension (compression), ou approximation de X par un tableau de rang q $(q < p)$.

Les derniers objectifs permettent d'utiliser l'ACP comme préalable à une autre technique préférant des variables orthogonales (régression linéaire) ou un nombre réduit d'entrées (réseaux neuronaux).

Des arguments de type géométrique dans la littérature francophone, ou bien de type statistique avec hypothèses de normalité dans la littérature anglo-saxonne, justifient la définition de l'ACP. Nous adoptons ici une optique intermédiaire en se référant à un modèle "allégé" car ne nécessitant pas d'hypothèse "forte" sur la distribution des observations (normalité). Plus précisément, l'ACP admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

4 Modèle

Les notations sont celles du paragraphe précédent :
- $\mathbf{X}$ désigne le tableau des données issues de l'observation de p variables *quantitatives* X^j sur n individus i de *poids* w_i,
- E est l'espace des individus muni de la base canonique et de la métrique de matrice $\mathbf{M}$,
- F est l'espace des variables muni de la base canonique et de la métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

De façon générale, un modèle s'écrit :

$$\textbf{Observation} = \textbf{Modèle} + \textbf{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de n vecteurs aléatoires indépendants $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, de même matrice de covariance $\sigma^2\mathbf{\Gamma}$, mais d'espérances différentes $\mathbf{z}_i$, toutes contenues dans un sous-espace affine de dimension q $(q < p)$ de E. Dans ce modèle, $E(\mathbf{x}_i) = \mathbf{z}_i$ est un paramètre spécifique attaché à chaque individu i et appelé *effet fixe*, le modèle étant dit *fonctionnel*. Ceci s'écrit en résumé :

$$\begin{array}{l} \{\mathbf{x}_i\,; i = 1, \ldots, n\},\ n \text{ vecteurs aléatoires indépendants de } E, \\ \mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i\ , i = 1, \ldots, n \text{ avec } \left\{ \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = 0,\ \text{var}(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{\Gamma}, \\ \sigma > 0 \text{ inconnu, } \mathbf{\Gamma} \text{ régulière et connue,} \end{array} \right. \\ \exists A_q,\ \text{sous-espace affine de dimension } q \text{ de } E \text{ tel que } \forall i, \mathbf{z}_i \in A_q\ (q < p). \end{array} \tag{3.1}$$

Soit $\overline{\mathbf{z}} = \sum_{i=1}^n w_i\mathbf{z}_i$. Les hypothèses du modèle entraînent que $\overline{\mathbf{z}}$ appartient à A_q. Soit donc E_q le sous-espace vectoriel de E de dimension q tel que :

$$A_q = \overline{\mathbf{z}} + E_q.$$

Les paramètres à estimer sont alors E_q et $\mathbf{z}_i, i = 1, \ldots, n$, éventuellement σ ; $\mathbf{z}_i$ est la part systématique, ou *effet*, supposée de rang q ; éliminer le bruit revient donc à réduire la dimension.

Si les $\mathbf{z}_i$ sont considérés comme *aléatoires*, le modèle est alors dit *structurel* ; on suppose que $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ est un échantillon statistique i.i.d. Les unités statistiques jouent des rôles symétriques, elles ne nous intéressent que pour l'étude des relations entre les variables. On retrouve alors le principe de l'analyse en facteurs (ou en facteurs communs et spécifiques, ou *factor analysis*).

4.1 Estimation

PROPOSITION 3.1. — *L'estimation des paramètres de (3.1) est fournie par l'ACP de* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *c'est-à-dire par la décomposition en valeurs singulières de* $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ *:*

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2}\mathbf{u}^k\mathbf{v}^{k\prime} = \mathbf{U}_q\mathbf{\Lambda}^{1/2}\mathbf{V}_q'.$$

Preuve

Sans hypothèse sur la distribution de l'erreur, une estimation par les moindres carrés conduit à résoudre le problème :

$$\min_{E_q, \mathbf{z}_i} \left\{ \sum_{i=1}^n w_i \left\|\mathbf{x}_i - \mathbf{z}_i\right\|_{\mathbf{M}}^2\ ;\ \dim(E_q) = q, \mathbf{z}_i - \overline{\mathbf{z}} \in E_q \right\}. \tag{3.2}$$

Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$.

$$\sum_{i=1}^n w_i \left\|\mathbf{x}_i - \mathbf{z}_i\right\|_{\mathbf{M}}^2 = \sum_{i=1}^n w_i \left\|\mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i\right\|_{\mathbf{M}}^2 + \left\|\overline{\mathbf{x}} - \overline{\mathbf{z}}\right\|_{\mathbf{M}}^2\ ;$$

le problème (3.2) conduit alors à prendre $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et devient équivalent à résoudre :

$$\min_{\mathbf{Z}} \left\{ \left\|\overline{\mathbf{X}} - \mathbf{Z}\right\|_{\mathbf{M},\mathbf{D}}\ ; \mathbf{Z} \in \mathcal{M}_{n,p},\ \text{rang}(\mathbf{Z}) = q \right\}. \tag{3.3}$$

La fin de la preuve est une conséquence immédiate du théorème (A.5).

□

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant.
- Les $\mathbf{v}_k$, appelés *vecteurs principaux*, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres ; ils engendrent des s.e.v. de dimension 1 appelés axes principaux.

Les estimations sont donc données par :

$$\begin{aligned}
\widehat{\overline{\mathbf{z}}} &= \overline{\mathbf{x}}, \\
\widehat{\mathbf{Z}_q} &= \sum_{k=1}^{q} \lambda^{1/2}\mathbf{u}^k\mathbf{v}^{k'} = \mathbf{U}_q\mathbf{\Lambda}^{1/2}\mathbf{V}_q' = \overline{\mathbf{X}}\widehat{\mathbf{P}_q}', \\
\text{où } \widehat{\mathbf{P}_q} &= \mathbf{V}_q\mathbf{V}_q'\mathbf{M} \text{ est la matrice de projection} \\
&\quad \mathbf{M}\text{-orthogonale sur } \widehat{E_q}, \\
\widehat{E_q} &= \text{vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}, \\
\widehat{E_2} &\text{ est appelé plan principal}, \\
\widehat{\mathbf{z}_i} &= \widehat{\mathbf{P}_q}\mathbf{x}_i + \overline{\overline{x}}.
\end{aligned}$$

Remarques

i. Les solutions sont emboîtées pour $q = 1, \ldots, p$:

$$E_1 = \text{ vect}\{\mathbf{v}^1\} \subset E_2 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2\} \subset E_3 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3\} \subset \ldots$$

ii. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.

iii. Si les variables ne sont pas homogènes (unités de mesure différentes, variances disparates), elles sont préalablement réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\mathbf{\Sigma}^{-1/2} \text{ où } \mathbf{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2), \text{ avec } \sigma_j^2 = \text{Var } (X^j) \text{ ;}$$

$\widetilde{\mathbf{S}}$ est alors la matrice $\mathbf{R} = \mathbf{\Sigma}^{-1/2}\mathbf{S}\mathbf{\Sigma}^{-1/2}$ des *corrélations*.

Sous l'hypothèse que la distribution de l'erreur est gaussienne, une estimation par maximum de vraisemblance conduit à la même solution.

4.2 Définition équivalente

On considère p variable statistiques *centrées* $X^1, \ldots, X^p$. Une *combinaison linéaire* de coefficients $\mathbf{f}_j$ de ces variables,

$$\mathbf{c} = \sum_{j=1}^{p} f_j\mathbf{x}^j = \overline{\mathbf{X}}\mathbf{f},$$

définit une nouvelle variable centrée C qui, à tout individu i, associe la "mesure"

$$C(i) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{f}.$$

PROPOSITION 3.2. — *Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est aussi la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.*

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les *facteurs principaux*. Ils permettent de définir les combinaisons linéaires des X^j optimales au sens ci-dessus.
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les *composantes principales*.
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les *variables principales* ;

$$\begin{aligned}\text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell.\end{aligned}$$

- Les $\mathbf{f}^k$ sont les vecteurs propres $\mathbf{M}^{-1}$-orthonormés de la matrice $\mathbf{MS}$.
- La matrice

$$\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$$

  est la matrice des composantes principales.
- Les axes définis par les vecteurs $\mathbf{D}$-orthonormés u^k sont appelés *axes factoriels*.

5 Représentations graphiques

5.1 Les individus

Les graphiques obtenus permettent de représenter "au mieux" les distances euclidiennes inter-individus mesurées par la métrique $\mathbf{M}$.

Projection

Chaque individu i représenté par $\mathbf{x}_i$ est approché par sa projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}}_i^q$ sur le sous-espace $\widehat{E_q}$ engendré par les q premiers vecteurs principaux $\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$. En notant $\mathbf{e}_i$ un vecteur de la base canonique de E, la coordonnée de l'individu i sur $\mathbf{v}^k$ est donnée par :

$$\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k.$$

PROPOSITION 3.3. — *Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.*

Mesures de "qualité"

La "qualité globale" des représentations est mesurée par la *part de dispersion expliquée* :

$$r_q = \frac{\text{tr}\mathbf{S}\mathbf{M}\widehat{\mathbf{P}_q}}{\text{tr}\mathbf{S}\mathbf{M}} = \frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k}.$$

Remarque. — La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage $\mathcal{N}$ par rapport à son barycentre $\overline{\mathbf{x}}$ se mesure par l'*inertie*, généralisation de la variance :

$$I_g(\mathcal{N}) = \sum_{i=1}^n w_i \left\|\mathbf{x}_i - \overline{\mathbf{x}}\right\|_{\mathbf{M}}^2 = \left\|\overline{\mathbf{X}}\right\|_{\mathbf{M},\mathbf{D}}^2 = \text{ tr }(\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M}) = \text{ tr }(\mathbf{S}\mathbf{M}).$$

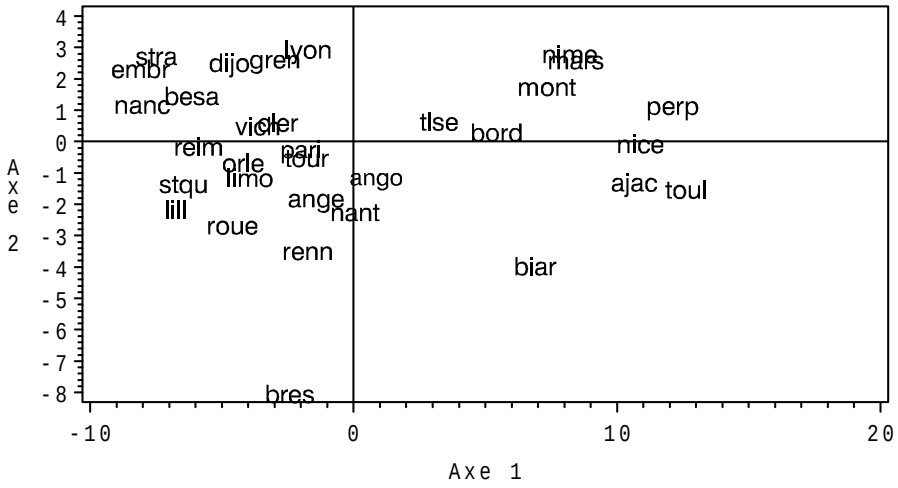

FIG. 3.3 – *Températures : premier plan des individus.*

La qualité de la représentation de chaque x_i est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$[\cos\theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}_i}^q)]^2 = \frac{\left\|\widehat{\mathbf{P}_q}(\mathbf{x}_i - \overline{\mathbf{x}})\right\|^2_{\mathbf{M}}}{\|\mathbf{x}_i - \overline{\mathbf{x}}\|^2_{\mathbf{M}}} = \frac{\sum_{k=1}^{q}(c_i^k)^2}{\sum_{k=1}^{p}(c_i^k)^2}.$$

Pour éviter de consulter un tableau qui risque d'être volumineux (n lignes), les étiquettes de chaque individu sont affichées sur les graphiques avec des caractères dont la *taille est fonction de la qualité*. Un individu très mal représenté est à la limite de la lisibilité.

Contributions

Les contributions de chaque individu à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \|\mathbf{x}_i - \overline{\mathbf{x}}\|^2_{\mathbf{M}}}{\mathrm{tr}\mathbf{SM}} = \frac{w_i \sum_{k=1}^{p}(c_i^k)^2}{\sum_{k=1}^{p}\lambda_k},$$

ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k},$$

permettent de déceler les observations les plus *influentes* et, éventuellement, aberrantes. Ces points apparaissent visiblement lors du tracé des diagrammes-boîtes parallèles des composantes principales qui évitent ainsi une lecture fastidieuse de ce tableau des contributions. En effet, ils se singularisent aussi comme "outliers" hors de la boîte (au delà des moustaches) correspondant à une direction principale. Les individus correspondants, considérés comme *individus supplémentaires*, peuvent être éliminés lors d'une nouvelle analyse.

Individus supplémentaires

Il s'agit de représenter, par rapport aux axes principaux d'une analyse, des individus qui n'ont pas participé aux calculs de ces axes. Soit $\mathbf{s}$ un tel vecteur, il doit être centré, éventuellement réduit,

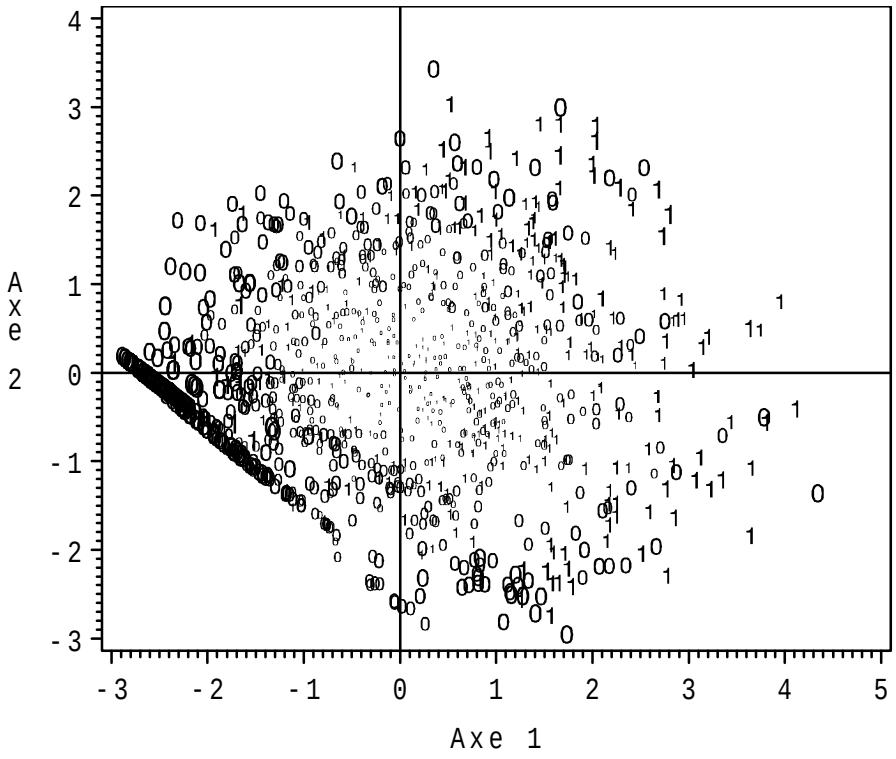

FIG. 3.4 – *Carte Visa : premier plan de l'ACP d'une sélection de variables. La discrimination des individus possèdant (1) ou non (0) la carte Visa premier n'est pas très claire sur cette représentation.*

puis projeté sur le sous-espace de représentation. Les coordonnées sont fournies par :

$$\left\langle \mathbf{v}^k, \mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}})\right\rangle_{\mathbf{M}} = \mathbf{v}^{k\prime}\mathbf{M}\mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}) = \mathbf{e}^{k\prime}\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

Les coordonnées d'un individu supplémentaire dans la base des vecteurs principaux sont donc :

$$\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

5.2 Les variables

Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables (cosinus des angles) et, si celles-ci ne sont pas réduites, leurs variances (longueurs).

Projection

Une variable X^j est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q}\mathbf{x}^j$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de $\mathbf{x}^j$ sur $\mathbf{u}^k$ est :

$$\left\langle \mathbf{x}^j, \mathbf{u}^k\right\rangle_{\mathbf{D}} = \mathbf{x}^{j\prime}\mathbf{D}\mathbf{u}^k = \frac{1}{\sqrt{\lambda_k}}\mathbf{x}^{j\prime}\mathbf{D}\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = \frac{1}{\sqrt{\lambda_k}}\mathbf{e}^{j\prime}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = \sqrt{\lambda_k}v_j^k.$$

PROPOSITION 3.4. — *Les coordonnées de la projection $\mathbf{D}$-orthogonale de $\mathbf{x}^j$ sur le sous-espace F_q sont les q premiers éléments de la j-ème ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.*

Mesure de "qualité"

La qualité de la représentation de chaque $\mathbf{x}^j$ est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q}\mathbf{x}^j)\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q}\mathbf{x}^j\right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^q \lambda_k (v_k^j)^2}{\sum_{k=1}^p \lambda_k (v_k^j)^2}.$$

Corrélations variables×facteurs

Ces indicateurs aident à l'interprétation des axes factoriels en exprimant les corrélations entre variables principales et initiales.

$$\mathrm{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\left\langle \mathbf{x}^j, \mathbf{u}^k\right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j}v_j^k\ ;$$

ce sont les éléments de la matrice $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Cercle des corrélations

Dans le cas de variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, les $\widetilde{\mathbf{x}}^j$ sont sur la sphère unité $\mathcal{S}_n$ de F. L'intersection $\mathcal{S}_n \cap F_2$ est un cercle centré sur l'origine et de rayon 1 appelé *cercle des corrélations*. Les projections de $\widetilde{\mathbf{x}}^j$ et $\mathbf{x}^j$ sont colinéaires, celle de $\widetilde{\mathbf{x}}^j$ étant à l'intérieur du cercle :

$$\left\|\widehat{\mathbf{Q}_2}\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2}x_j) \leq 1.$$

Ainsi, plus $\widehat{\mathbf{Q}_2}\widetilde{\mathbf{x}}^j$ est proche de ce cercle, meilleure est la qualité de sa représentation. Ce graphique est commode à interpréter à condition de se méfier des échelles, le cercle devenant une ellipse si elles ne sont pas égales. Comme pour les individus, la taille des caractères est aussi fonction de la qualité des représentations.

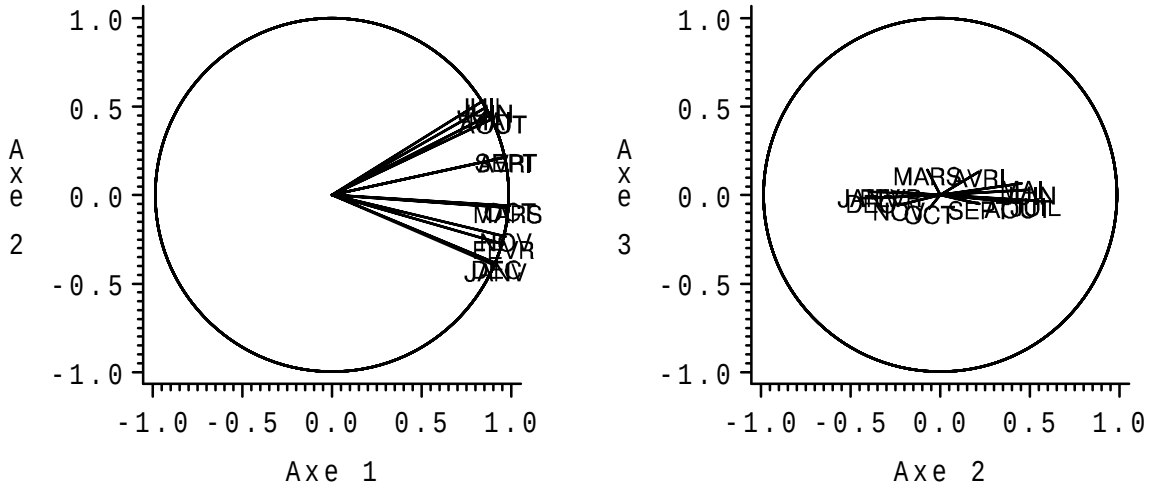

FIG. 3.5 – *Températures : Premier et deuxième plan des variables.*

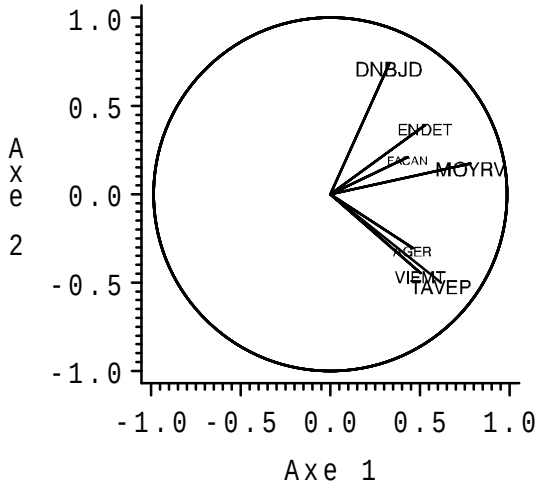

FIG. 3.6 – *Carte Visa : la représentation des variables dans le premier plan de l'ACP fournit une interprétation classique (stocks versus flux) de ce type de données.*

5.3 Représentation simultanée ou "biplot"

À partir de la décomposition en valeurs singulières de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$, on remarque que chaque valeur

$$x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$$

s'exprime comme produit scalaire usuel des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j .$$

Pour $q = 2$, la quantité $\widehat{z_i}^j$ en est une approximation limitée aux deux premiers termes.

Cette remarque permet d'interpréter deux autres représentations graphiques en ACP projetant *simultanément* individus et variables.

i. la représentation *isométrique ligne* utilise les matrices $\mathbf{C}$ et $\mathbf{V}$; elle permet d'interpréter les distances entre individus ainsi que les produits scalaires entre un individu et une variable qui sont, dans le premier plan principal, des approximations des valeurs observées $X^j(\omega_i)$;

ii. la représentation *isométrique colonne* utilise les matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.

Remarques

i. Dans le cas fréquent où $\mathbf{M} = \mathbf{I}_p$ et où les variables sont réduites, le point représentant X^j, en superposition dans l'espace des individus se confond avec un pseudo individu supplémentaire qui prendrait la valeur 1 (écart-type) pour la variable j et 0 pour les autres.

ii. En pratique, ces différents types de représentations (simultanées ou non) ne diffèrent que par un changement d'échelle sur les axes ; elles sont très voisines et suscitent souvent les mêmes interprétations.

6 Choix de dimension

La qualité des estimations auxquelles conduit l'ACP dépend, de façon évidente, du choix de q, c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentation.

De nombreux critères de choix pour q ont été proposés dans la littérature. Nous présentons ici ceux, les plus courants, basés sur une heuristique et un reposant sur une quantification de la stabilité du sous-espace de représentation. D'autres critères, non explicités, s'inspirent des pratiques statistiques décisionnelles ; sous l'hypothèse que l'erreur admet une distribution *gaussienne*, on peut exhiber les lois asymptotiques des valeurs propres et donc construire des tests de nullité ou d'égalité de ces dernières. Malheureusement, outre la nécessaire hypothèse de normalité, ceci conduit à une procédure de tests emboîtés dont le niveau global est incontrôlable. Leur utilisation reste donc heuristique.

6.1 Part d'inertie

La "qualité globale" des représentations est mesurée par la *part d'inertie expliquée* :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

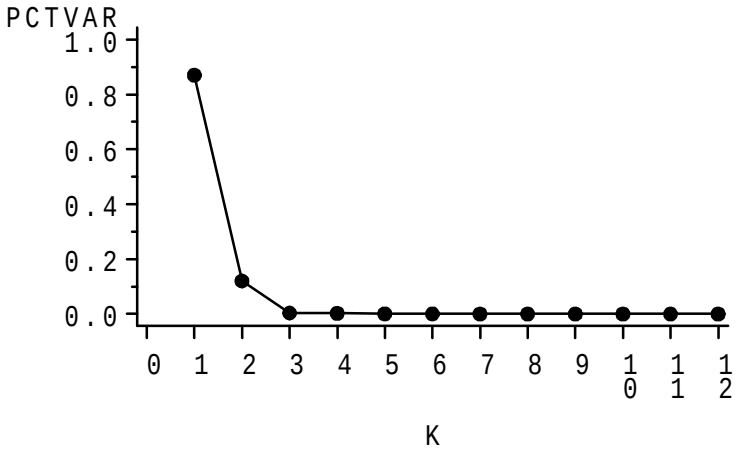

FIG. 3.7 – *Températures : éboulis des valeurs propres.*

La valeur de q est choisie de sorte que cette part d'inertie expliquée r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur. C'est souvent le seul critère employé.

6.2 Règle de Kaiser

On considère que, si tous les éléments de Y sont indépendants, les composantes principales sont toutes de variances égales (égales à 1 dans le cas de l'ACP réduite). On ne conserve alors que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales ; dans le cas d'une ACP réduite, ne sont donc retenues que celles plus grandes que 1. Ce critère, utilisé implicitement par SAS/ASSIST, a tendance à surestimer le nombre de composantes pertinentes.

6.3 Éboulis des valeurs propres

C'est le graphique (figures 6.3 et 6.3) présentant la décroissance des valeurs propres. Le principe consiste à rechercher, s'il existe, un "coude" (changement de signe dans la suite des différences d'ordre 2) dans le graphe et de ne conserver que les valeurs propres jusqu'à ce coude. Intuitivement, plus l'écart $(\lambda_q - \lambda_{q+1})$ est significativement grand, par exemple supérieur à $(\lambda_{q-1} - \lambda_q)$, et plus on peut être assuré de la stabilité de $\widehat{E_q}$.

6.4 Boîtes-à-moustaches des variables principales

Un graphique (figure 6.4 et 6.4) présentant, en parallèle, les boîtes-à-moustaches des variables principales illustre bien leurs qualités : stabilité lorsqu'une grande boîte est associée à de petites moustaches, instabilité en présence d'une petite boîte, de grandes moustaches et de points isolés. Intuitivement, on conserve les premières "grandes boîtes". Les points isolés ou "outliers" désignent les points à forte contribution, ou potentiellement influents, dans une direction principale. Ils nécessitent une étude clinique : une autre analyse dans laquelle ils sont déclarés supplémentaires (poids nuls) afin d'évaluer leur impact sur l'orientation des axes.

6.5 Stabilité du sous-espace

La présentation de l'ACP, comme résultat de l'estimation d'un modèle, offre une autre approche au problème du choix de dimension. La qualité des estimations est évaluée de façon habituelle en statistique par un risque moyen quadratique définissant un critère de stabilité du sous-

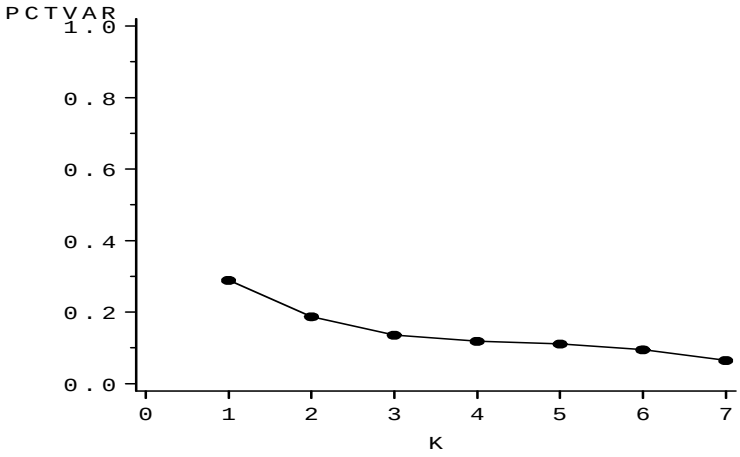

FIG. 3.8 – *Carte Visa : éboulis des valeurs propres.*

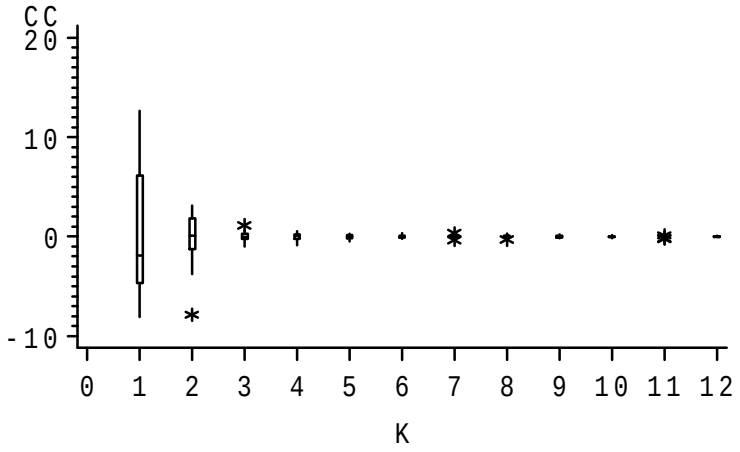

FIG. 3.9 – *Températures : composantes en boîtes.*

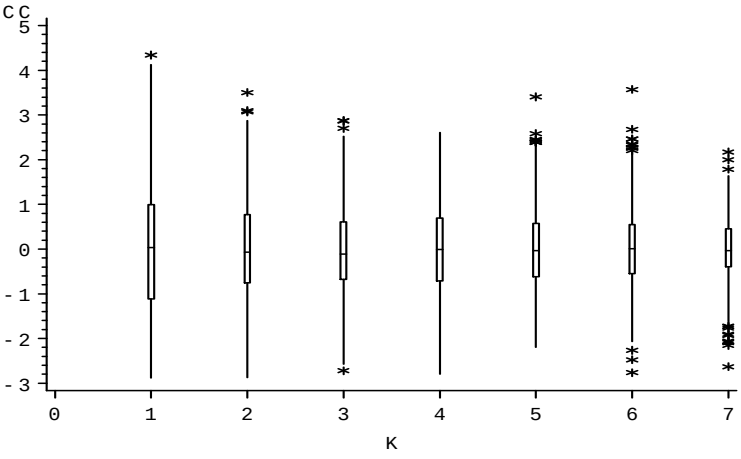

FIG. 3.10 – *Carte Visa Premier : composantes en boîtes.*

espace de représentation. Il est défini comme l'espérance d'une distance entre le modèle "vrai" et l'estimation qui en est faite. Besse (1992) propose d'étudier la qualité de l'estimation du sous-espace de représentation $\widehat{E_q}$ en considèrant la fonction perte :

$$L_q = Q(E_q, \widehat{E_q}) = \frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|^2_{\mathbf{M},\mathbf{D}} = q - \text{tr}\mathbf{P}_q\widehat{\mathbf{P}_q},$$

où Q mesure la distance entre deux sous-espaces par la distance usuelle entre les matrices de projection qui leur sont associées. C'est aussi la somme des carrés des coefficients de corrélation canonique entre les ensembles de composantes ou de variables principales qui engendrent respectivement E_q et son estimation $\widehat{E_q}$.

Un risque moyen quadratique est alors défini en prenant l'espérance de la fonction perte :

$$R_q = EQ(E_q, \widehat{E_q}). \tag{3.4}$$

Sans hypothèse sur la distribution de l'erreur, seules des techniques de ré-échantillonnage (bootstrap, jackknife) permettent de fournir une estimation de ce risque moyen quadratique. Leur emploi est justifié, car le risque est invariant par permutation des observations, mais coûteux en temps de calcul. On se pose donc la question de savoir pour quelles valeurs de q les représentations graphiques sont fiables, c'est-à-dire stables pour des fluctuations de l'échantillon. Besse (1992) propose d'utiliser une approximation de l'estimateur par jackknife ; elle fournit, directement à partir des résultats de l'A.C.P. (valeurs propres et composantes principales), une estimation satisfaisante du risque :

$$\widehat{R_{JKq}} = \widehat{R_{\mathbf{P}q}} + O((n-1)^{-2}).$$

$\widehat{R_{\mathbf{P}q}}$ est une approximation analytique de l'estimateur jackknife qui a pour expression :

$$\widehat{R_{\mathbf{P}q}} = \frac{1}{n-1}\sum_{k=1}^{q}\sum_{j=q+1}^{p}\frac{\frac{1}{n}\sum_{i=1}^{n}(c_i^k)^2(c_i^j)^2}{(\lambda_j - \lambda_k)^2} \tag{3.5}$$

où c_i^j désigne le terme général de la matrice des composantes principales $\mathbf{C}$.

Ce résultat souligne l'importance du rôle que joue l'écart $(\lambda_q - \lambda_{q+1})$ dans la stabilité du sous-espace de représentation. Le développement est inchangé dans le cas d'une ACP réduite ; de plus, il est valide tant que

$$n > \frac{\|\mathbf{S}\|_2^2}{\inf\{(\lambda_k - \lambda_{k+1}); k = 1, \ldots, q\}}.$$

La figure 3.11 montrent la stabilité du sous-espace de représentation en fonction de la dimension q pour l'A.C.P. des données de températures. Comme souvent, le premier axe est très stable tandis que le premier plan reste fiable. Au delà, les axes étant très sensibles à toute perturbation des données, ils peuvent être associés à du bruit. Ces résultats sont cohérents avec les deux critères graphiques précédents mais souvent, en pratique, le critère de stabilité conduit à un choix de dimension plus explicite.

7 Interprétation

Les macros SAS décrites en exemple, de même que la plupart des logiciels, proposent, ou autorisent, l'édition des différents indicateurs (contributions, qualités, corrélations) et graphiques définis dans les paragraphes précédents.

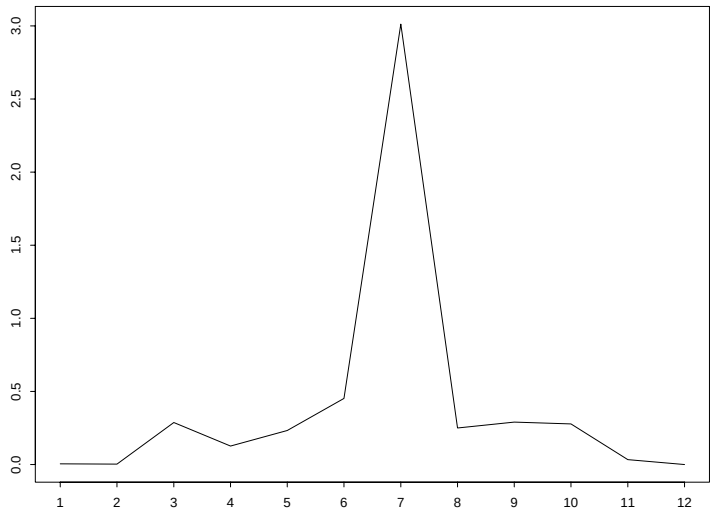

FIG. 3.11 – *Températures : stabilité des sous-espaces.*

- Les *contributions* permettent d'identifier les individus très influents pouvant déterminer à eux seuls l'orientation de certains axes ; ces points sont vérifiés, caractérisés, puis éventuellement considérés comme *supplémentaires* dans une autre analyse.
- Il faut choisir le nombre de composantes à retenir, c'est-à-dire la dimension des espaces de représentation.
- Les axes factoriels sont interprétés par rapport aux variables initiales bien représentées.
- Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales.

Les quelques graphiques présentés suffisent, dans la plupart des cas, à l'interprétation d'une ACP classique et évitent la sortie volumineuse, lorsque n est grand, des tableaux usuels d'aide à l'interprétation. On échappe ainsi à une critique fréquente, et souvent justifiée, des anglo-saxons vis-à-vis de la pratique française de "l'analyse des données" qui, paradoxalement, cherche à "résumer au mieux l'information" mais produit plus de chiffres en sortie qu'il n'y en a en entrée !

Remarque. — L'ACP est une technique *linéaire* optimisant un critère *quadratique* ; elle ne tient donc pas compte d'éventuelles liaisons non linéaires et présente une forte sensibilité aux valeurs extrêmes.

Chapitre 4

Analyse Factorielle Discriminante

1 Introduction

1.1 Données

Les données sont constituées de

- p variables *quantitatives* $X^1, \ldots, X^p$ jouant le rôle de variables explicatives comme dans le modèle linéaire,
- une variable *qualitative* T, à m modalités $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$, jouant le rôle de variable à expliquer.

La situation est analogue à celle de la régression linéaire multiple mais, comme la variable à expliquer est qualitative, on aboutit à une méthode très différente. Les variables sont observées sur l'ensemble Ω des n individus affectés des poids $w_i > 0,\ (\sum_{i=1}^n w_i = 1)$, et l'on pose

$$\mathbf{D} = \text{ diag}(w_i\,; i = 1, \ldots, n).$$

La variable T engendre une partition $\{\Omega_\ell\,; \ell = 1, \ldots, m\}$ de l'ensemble Ω des individus dont chaque élément est d'effectif n_ℓ.

On note $\mathbf{T}$ $(n \times m)$ la matrice des indicatrices des modalités de la variable T ; son terme général est

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases}.$$

En posant

$$\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i,$$

il vient

$$\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{ diag}(\overline{w_1}, \ldots, \overline{w_m}).$$

1.2 Objectifs

Deux techniques cohabitent sous la même appellation d'analyse discriminante :

descriptive : cette méthode recherche, parmi toutes les ACP possibles sur les variables X^j, celle dont les représentations graphiques des individus *discriminent* "au mieux" les m classes engendrées par la variable T (e.g. recherche de facteurs de risque en statistique médicale) ;

décisionnelle : connaissant, pour un individu donné, les valeurs des Y^j mais pas la modalité de T, cette méthode consiste à affecter cet individu à une modalité (e.g. reconnaissance de formes). Cette méthode est décrite dans la partie *modélisation* de ce cours.

Remarque. — Lorsque le nombre et les caractéristiques des classes sont connues, il s'agit d'une *discrimination* ; sinon, on parle de *classification* ou encore, avec des hypothèses sur les distributions, de *reconnaissance de mélanges*.

1.3 Notations

On note $\mathbf{X}$ la matrice $(n \times p)$ des données quantitatives, $\mathbf{G}$ la matrice $(m \times p)$ des barycentres des classes :

$$\mathbf{G} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{X} = \begin{bmatrix} \mathbf{g}_1{}' \\ \vdots \\ \mathbf{g}_m{}' \end{bmatrix} \text{ où } \mathbf{g}_\ell = \frac{1}{\overline{w_\ell}} \sum_{i \in \Omega_\ell} w_i \mathbf{x}_i,$$

et $\mathbf{X}_e$ la matrice $(n \times p)$ dont la ligne i est le barycentre $\mathbf{g}_\ell$ de la classe Ω_ℓ à laquelle appartient l'individu i :

$$\mathbf{X}_e = \mathbf{T}\mathbf{G} = \mathbf{P}\mathbf{G} \; ;$$

$\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur le sous-espace engendré par les indicatrices de T ; c'est encore l'espérance conditionnelle sachant T.

Deux matrices "centrées" sont définies de sorte que $\overline{\mathbf{X}}$ se décompose en

$$\overline{\mathbf{X}} = \overline{\mathbf{X}}_r + \overline{\mathbf{X}}_e$$

avec

$$\overline{\mathbf{X}}_r = \mathbf{X} - \mathbf{X}_e \text{ et } \overline{\mathbf{X}}_e = \mathbf{X}_e - \mathbf{1}_n \overline{\mathbf{x}}'.$$

On note également $\overline{\mathbf{G}}$ la matrice centrée des barycentres :

$$\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \overline{\mathbf{x}}'.$$

On appelle alors variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}'\mathbf{D}\overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)',$$

et variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}'\overline{\mathbf{D}\mathbf{G}} = \overline{X}_e'\mathbf{D}\overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \overline{\mathbf{x}})(\mathbf{g}_\ell - \overline{\mathbf{x}})'.$$

PROPOSITION 4.1. — *La matrice des covariances se décompose en*

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

2 Définition

2.1 Modèle

Dans l'espace des individus, le principe consiste à projeter les individus dans une direction permettant de mettre en évidence les groupes. À cette fin, Il faut privilégier la variance interclasse au détriment de la variance intraclasse considérée comme due au bruit.

En ACP, pour chaque effet $\mathbf{z}_i$ à estimer, on ne dispose que d'une observation $\mathbf{x}_i$; dans le cas de l'AFD on considère que les éléments d'une même classe Ω_ℓ sont les observations répétées n_ℓ fois du même effet $\mathbf{z}_\ell$ pondéré par $\overline{w_\ell} = \sum_{i\in\Omega_\ell} w_i$. Le modèle devient donc :

$$\begin{array}{l} \{\mathbf{x}_i\,; i=1,\ldots,n\},\ n \text{ vecteurs indépendants de } E,\\ \forall \ell, \forall i \in \Omega_\ell, \mathbf{x}_i = \mathbf{z}_\ell + \boldsymbol{\varepsilon}_i \text{ avec } \left\{ \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = 0,\ \text{var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Gamma},\\ \boldsymbol{\Gamma} \text{ régulière et inconnue,} \end{array} \right.\\ \exists A_q,\ \text{sous-espace affine de de dimension } q \text{ de } E \text{ tel que}\\ \forall \ell, \mathbf{z}_\ell \in A_q,\ (q < \min(p, m-1)). \end{array} \tag{4.1}$$

Remarque. — Soit $\overline{\mathbf{z}} = \sum_{\ell=1}^{m} \overline{w_\ell}\mathbf{z}_\ell$. Le modèle entraîne que $\overline{\mathbf{z}} \in A_q$. Soit E_q le sous-espace de dimension q de E tel que $A_q = \overline{\mathbf{z}} + E_q$. Les paramètres à estimer sont E_q et $\{\mathbf{z}_\ell\,; \ell = 1,\ldots,m\}$; $\overline{w_\ell}$ est un paramètre de nuisance qui ne sera pas considéré.

2.2 Estimation

L'estimation par les moindres carrés s'écrit ainsi :

$$\min_{E_q, \mathbf{z}_\ell} \left\{ \sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|^2_{\mathbf{M}}\ ;\ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}.$$

Comme on a

$$\sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|^2_{\mathbf{M}} = \sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{g}_\ell\right\|^2_{\mathbf{M}} + \sum_{\ell=1}^{m} \overline{w_\ell} \left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|^2_{\mathbf{M}},$$

on est conduit à résoudre :

$$\min_{E_q, z_\ell} \left\{ \sum_{\ell=1}^{m} \overline{w_\ell} \left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|^2_{\mathbf{M}}\ ;\ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}.$$

La covariance $\sigma^2\boldsymbol{\Gamma}$ du modèle (4.1) étant inconnue, il faut l'estimée. Ce modèle stipule que l'ensemble des observations d'une même classe Ω_l suit une loi (inconnue) de moyenne $\mathbf{z}_e ll$ et de variance $\boldsymbol{\Gamma}$. Dans ce cas particulier, la matrice de covariances intraclasse ou matrice des covariances résiduelles empiriques $\mathbf{S}_r$ fournit donc une estimation "optimale" de la métrique de référence :

$$\mathbf{M} = \widehat{\boldsymbol{\Gamma}}^{-1} = \mathbf{S}_r^{-1}$$

.

PROPOSITION 4.2. — *L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle 4.1 est obtenue par l'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (AFD) de $(\mathbf{X}|\mathbf{T}, \mathbf{D})$.*

3 Réalisation de l'AFD

Les expressions matricielles définissant les représentations graphiques et les aides à l'interprétation découlent de celles de l'ACP.

3.1 Matrice à diagonaliser

L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à l'analyse spectrale de la matrice positive $\mathbf{S}_r^{-1}$-symétrique :

$$\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}.$$

Comme $\mathbf{S}_r^{-1}$ est régulière, cette matrice est de même rang que $\mathbf{S}_e$ et donc de même rang que $\mathbf{G}$ qui est de dimension $(m \times p)$. Les données étant centrées lors de l'analyse, le rang de la matrice à diagonaliser est

$$h = \text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p),$$

qui vaut en général $m-1$ c'est-à-dire le nombre de classes moins un.

On note $\lambda_1 \geq \cdots \geq \lambda_h > 0$ les valeurs propres de $\mathbf{S}_e\mathbf{S}_r^{-1}$ et $\mathbf{v}^1, \ldots, \mathbf{v}^h$ les vecteurs propres$\mathbf{S}_r^{-1}$-orthonormés associés. On pose

$$\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h) \text{ et } \mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h].$$

Les vecteurs $\mathbf{v}^k$ sont appelés *vecteurs discriminants* et les sous-espaces vectoriels de dimension 1 qu'ils engendrent dans $\mathbb{R}^p$ les *axes discriminants*.

3.2 Représentation des individus

L'espace des individus est ($\mathbb{R}^p$, b. c., $\mathbf{S}_r^{-1}$). Une représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ des classes par rapport aux mêmes axes discriminants est obtenue dans cet espace au moyen des coordonnées :

$$\begin{aligned} \mathbf{C} &= \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V} \text{ pour les individus et} \\ \overline{\mathbf{C}} &= \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C} \text{ pour les barycentres.} \end{aligned}$$

Les individus initiaux sont projetés comme des individus supplémentaires dans le système des axes discriminants. Comme en ACP, on peut calculer des cosinus carrés pour préciser la qualité de représentation de chaque individu.

Il est utile de différencier graphiquement la classe de chaque individu afin de pouvoir apprécier visuellement la qualité de la discrimination.

3.3 Représentation des variables

L'espace des variables est ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$). Chaque variable X^j est représenté par un vecteur dont les coordonnées dans le système des axes factoriels est une ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.

3.4 Interprétations

Les interprétations usuelles : la norme est un écart-type, un cosinus d'angle est un coefficient de corrélation, doivent être faites en termes d'écarts-types et de corrélations *expliquées* par la partition.

La représentation des variables est utilisée pour interprétée les axes en fonction des variables initiales conjointement avec la matrice des corrélations expliquées variables×facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$. La matrice $\mathbf{\Sigma}_e^{-1}$ étant la matrice diagonale des écarts-types expliqués σ_e^j c'est-à-dire des racines carrées des éléments diagonaux de la matrice $\mathbf{S}_e$.

Le point pratique essentiel est de savoir si la représentation des individus-barycentres et des individus initiaux permet de faire une bonne discrimination entre les classes définies par la variable

T. Si ce n'est pas le cas, l'AFD ne sert à rien, les X^j n'expliquent pas T. Dans le cas favorable, le graphique des individus permet d'interpréter la discrimination en fonction des axes et, celui des variables, les axes en fonction des variables initiales. La synthèse des deux permet l'interprétation de T selon les X^j.

4 Variantes de l'AFD

4.1 Individus de mêmes poids

L'AFD peut être définie de différentes façon. Dans la littérature anglo-saxonne, et donc dans la version standard d'AFD du logiciel SAS (procédure `candisc`), ce sont les estimations sans biais des matrices de variances "intra" (within) et "inter" (between) qui sont considérées dans le cas d'individus de mêmes poids $1/n$.

Dans ce cas particulier,

$$\mathbf{D} = \frac{1}{n}\mathbf{I}_n \text{ et } \overline{\mathbf{D}} = \frac{1}{n}\text{diag}(n_1, \ldots, n_m) \text{ où } n_\ell = \text{card}(\Omega_\ell)$$

et les matrices de covariances empiriques ont alors pour termes généraux :

$$\begin{aligned}
(\mathbf{S})_j^k &= \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}^j)(x_i^k - \overline{x}^k),\\
(\mathbf{S}_e)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m} n_\ell(g_\ell^j - \overline{x}^j)(g_\ell^k - \overline{x}^k),\\
(\mathbf{S}_r)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(x_i^j - g_\ell^j)(x_i^k - g_\ell^k).
\end{aligned}$$

Du point de vue de le Statistique inférentielle, on sait que les quantités calculées ci-dessus ont respectivement $(n-1)$, $(m-1)$ et $(n-m)$ degrés de liberté. En conséquence, ce point de vue est obtenu en remplaçant dans les calculs

$$\begin{aligned}
\mathbf{S} \quad &\text{par} \quad \mathbf{S}^* = \frac{n}{n-1}\mathbf{S},\\
\mathbf{S}_e \quad &\text{par} \quad \mathbf{S}_e^* = \mathbf{B} = \frac{n}{m-1}\mathbf{S}_e,\\
\mathbf{S}_r \quad &\text{par} \quad \mathbf{S}_r^* = \mathbf{W} = \frac{n}{n-m}\mathbf{S}_r.
\end{aligned}$$

Les résultats numériques de l'AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e^*\mathbf{S}_r^{*-1}$	$=$	$\frac{n-m}{m-1}\mathbf{S}_e\mathbf{S}_r^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}^*$	$=$	$\frac{n-m}{m-1}\mathbf{\Lambda}$,
– vecteurs propres :	$\mathbf{V}^*$	$=$	$\sqrt{\frac{n}{n-m}}\mathbf{V}$,
– représentation des barycentres :	$\overline{\mathbf{C}}^*$	$=$	$\sqrt{\frac{n-m}{n}}\overline{\mathbf{C}}$,
– représentation des variables :	$\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\sqrt{\frac{n}{m-1}}\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{*-1}\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Ainsi, les représentations graphiques sont identiques à un facteur d'échelle près tandis que les parts de variance expliquée et les corrélations variables-facteurs sont inchangées.

4.2 Métrique de Mahalanobis

L'AFD est souvent introduite dans la littérature francophone comme un cas particulier d'Analyse Canonique entre un ensemble de p variables quantitatives et un ensemble de m variables indicatrices des modalités de T. La proposition suivante établit les relations entre les deux approches :

PROPOSITION 4.3. — *l'ACP de* $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ *conduit aux mêmes vecteurs principaux que l'ACP de* $(\mathbf{G}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$. *Cette dernière est l'ACP des barycentres des classes lorsque l'espace des individus est muni de la métrique dite de Mahalanobis* $\mathbf{M} = \mathbf{S}^{-1}$ *et l'espace des variables de la métrique des poids des classes* $\overline{\mathbf{D}}$.

Les résultats numériques de l'AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e\mathbf{S}^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
– vecteurs propres :	$\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
– représentation des barycentres :	$\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
– représentation des variables :	$\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Les représentations graphiques des individus (voir ci-dessus) ne diffèrent alors que d'une homothétie et conduisent à des interprétations identiques, les corrélations variables-facteurs ainsi que les représentations des variables sont inchangées.

5 Exemples

Ce chapitre est illustrée par une comparaison des sorties graphiques issues d'une ACP et d'une AFD. Les données décrivent trois classes d'insectes sur lesquels ont été réalisées 6 mesures anatomiques. On cherche à savoir si ces mesures permettent de retrouver la typologie de ces insectes. Ce jeu de données "scolaire" conduit à une bien meilleure discrimination que ce que l'on peut obtenir dans une situation concrète.

C'est ce qui se passe avec les données bancaires. La discrimination obtenue n'est pas très nette, une meilleure le sera en considérant une sélection de variables plus adaptée. D'autre part, la situation est ici très particulière car la variable à expliquer n'ayant que deux modalités, la dimension du sous-espace est réduite à un. Une deuxième dimension est générée de façon aléatoire afin de rendre plus lisible la représentation des individus.

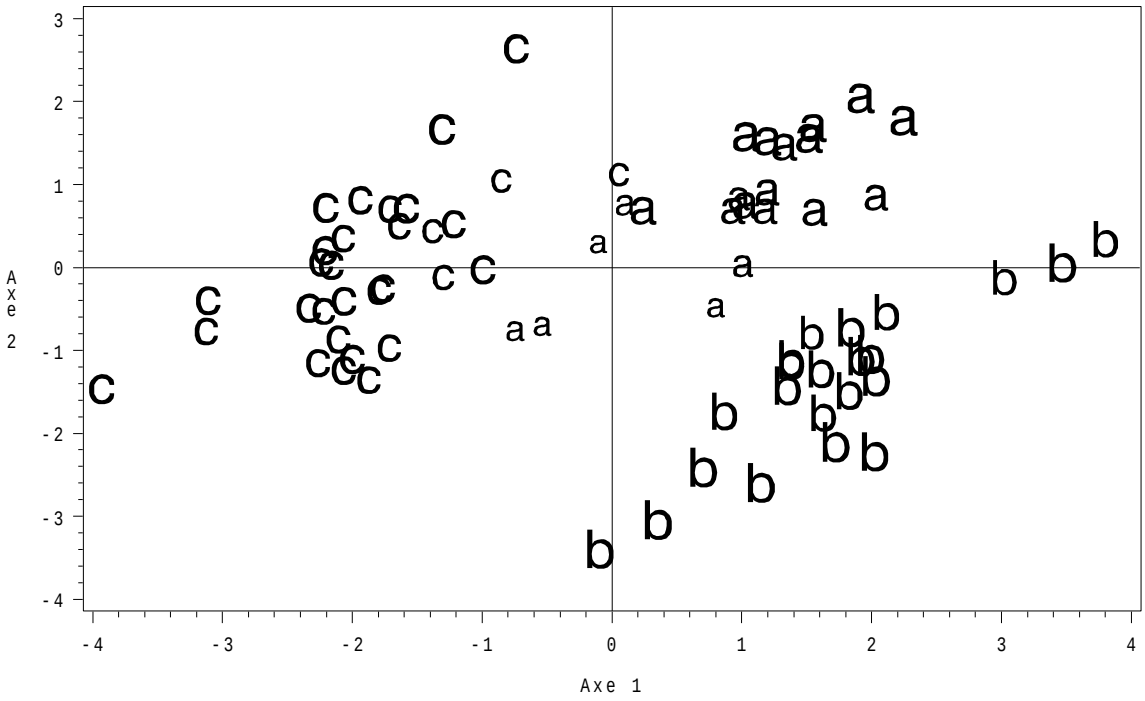

FIG. 4.1 – *Insectes : premier plan factoriel de l'ACP.*

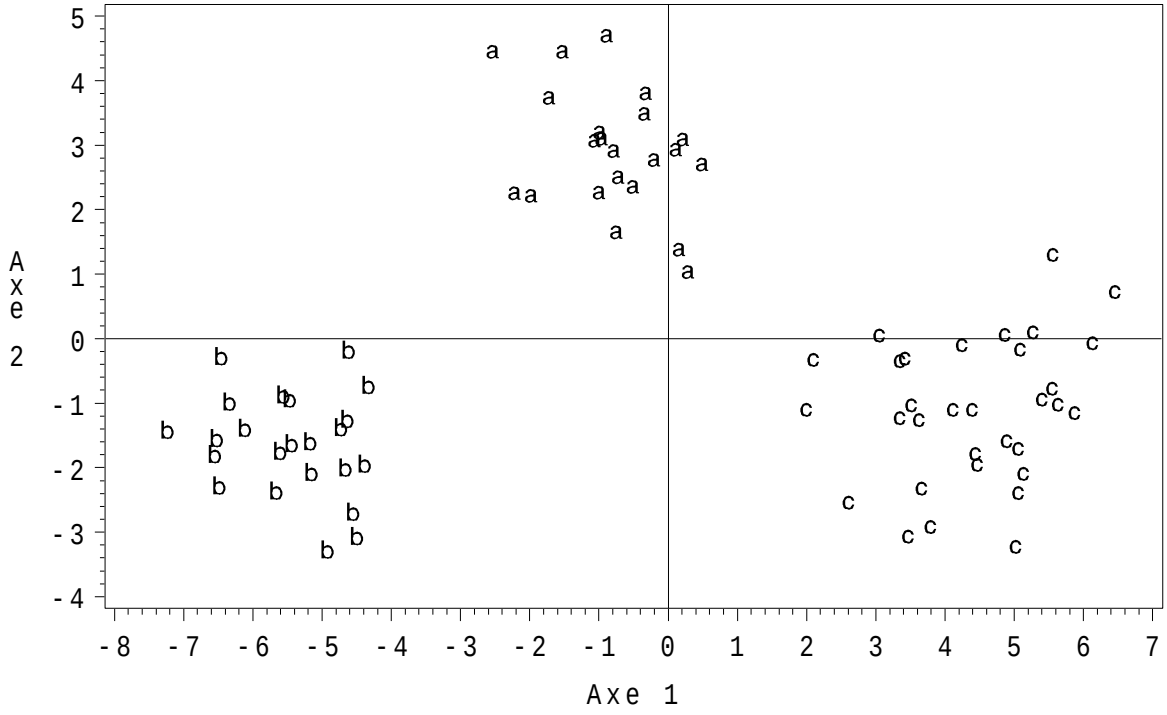

FIG. 4.2 – *Insectes : premier plan factoriel de l'AFD.*

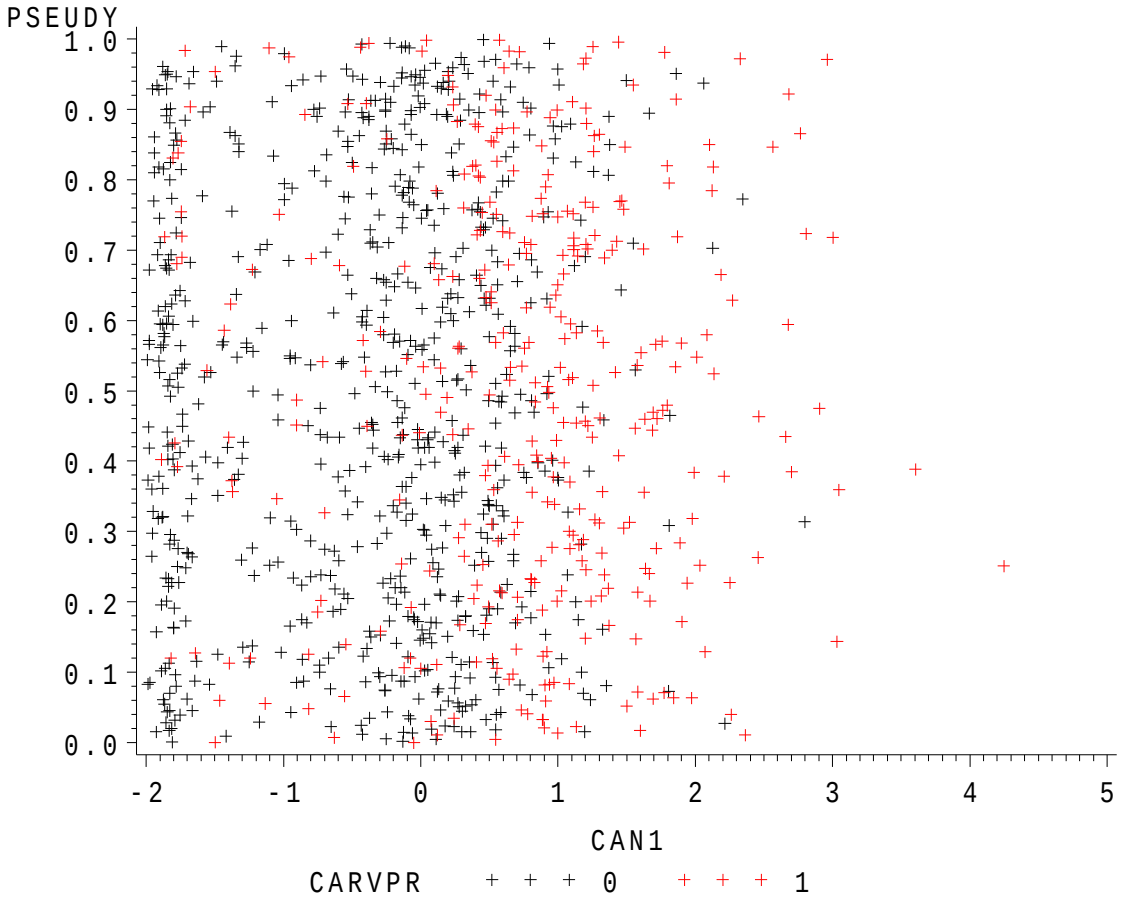

FIG. 4.3 – *Carte Visa : premier plan factoriel de l'AFD. L'axe 2 est issu d'un tirage aléatoire, l'axe 1 ne fournit pas une discrimination très marquée. Cela remet en cause la popssibilité de discrimination linéaire des deuc classes.*

Chapitre 5

Analyse Factorielle des Correspondances

1 Introduction

1.1 Données

On considère dans ce chapitre deux variables qualitatives observées simultanément sur n individus affectés de poids identiques $1/n$. On suppose que la première variable, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

La table de contingence associée à ces observations, de dimension $r \times c$, est notée $\mathbf{T}$; son élément générique est $n_{\ell h}$, effectif conjoint. Elle se présente sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

1.2 Notations

Les quantités $\{n_{\ell +} = \sum_{h=1}^{c} n_{\ell h}; \ell = 1, \ldots, r\}$ et $\{n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}; h = 1, \ldots, c\}$ sont les *effectifs marginaux* vérifiant $\sum_{\ell=1}^{r} n_{\ell +} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on définit les notions de *fréquences conjointes* ($f_{\ell h} = n_{\ell h}/n$) et de *fréquences marginales*. Ces dernières sont rangées dans les vecteurs :

$$\begin{aligned} \mathbf{g}_r &= [f_{1+}, \ldots, f_{r+}]', \\ \text{et } \mathbf{g}_c &= [f_{+1}, \ldots, f_{+c}]'. \end{aligned}$$

Elles permettent de définir les matrices :

$$\begin{aligned} \mathbf{D}_r &= \text{diag}(f_{1+}, \ldots, f_{r+}), \\ \text{et } \mathbf{D}_c &= \text{diag}(f_{+1}, \ldots, f_{+c}). \end{aligned}$$

On sera également amené à considérer les profils–lignes et les profils–colonnes déduits de $\mathbf{T}$. Le ℓème profil-ligne est

$$\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\}.$$

Il est considéré comme un vecteur de $\mathbb{R}^c$ et les r vecteurs ainsi définis sont disposés en colonnes dans la matrice $c \times r$

$$\mathbf{A} = \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}.$$

De même, le $h$ème profil-colonne est

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\},$$

vecteur de $\mathbb{R}^r$, et la matrice $r \times c$ des profils-colonnes est

$$\mathbf{B} = \frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}.$$

1.3 Liaison entre deux variables qualitatives

DÉFINITION 5.1. — *On dit que deux variables X et Y sont* non liées *relativement à T si et seulement si :*

$$\forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\} : n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n}.$$

Il est équivalent de dire que tous les profils-lignes sont égaux, ou encore que tous les profils-colonnes sont égaux (voir chapitre 2).

Cette notion est cohérente avec celle d'indépendance en probabilités. En effet, soit $\Omega = \{1, \ldots, n\}$ l'ensemble des individus observés et $(\Omega, \mathcal{P}(\Omega), P)$ l'espace probabilisé associé où P est l'équiprobabilité ; $\mathcal{M}_X = \{x_1, \ldots, x_r\}$ et $\mathcal{M}_Y = \{y_1, \ldots, y_c\}$ désignent les ensembles de modalités, ou valeurs prises par les variables X et Y. On note $\widetilde{X}$ et $\widetilde{Y}$ les variables aléatoires associées aux 2 variables statistiques X et Y :

$$\begin{aligned} \widetilde{X} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), \\ \widetilde{Y} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y)) ; \end{aligned}$$

P_X, P_Y et P_{XY} désignent respectivement les probabilités images définies par $\widetilde{X}, \widetilde{Y}$ et le couple $(\widetilde{X}, \widetilde{Y})$ sur $(\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y))$ et $(\mathcal{M}_X \times \mathcal{M}_Y, \mathcal{P}(\mathcal{M}_X) \times \mathcal{P}(\mathcal{M}_Y))$; ce sont les probabilités empiriques. Alors, X et Y sont *non liées* si et seulement si $\widetilde{X}$ et $\widetilde{Y}$ sont *indépendantes en probabilité* (la vérification est immédiate).

On suppose maintenant qu'il existe une liaison entre X et Y que l'on souhaite étudier. La représentation graphique des profils-lignes ou des profils-colonnes, au moyen de diagrammes en barres parallèles, ainsi que le calcul de coefficients de liaison (Cramer ou Tschuprow) donnent une première idée de la variation conjointe des deux variables (voir chapitre 2). Le test du χ^2 permet de plus de s'assurer du caractère significatif de cette liaison. Il est construit de la manière suivante :

l'hypothèse nulle est $H0$: $\widetilde{X}$ et $\widetilde{Y}$ sont indépendantes en probabilités ;

l'hypothèse alternative est $H1$: les variables $\widetilde{X}$ et $\widetilde{Y}$ ne sont pas indépendantes.

La statistique de test est alors

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{\left(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n}\right)^2}{\frac{n_{\ell +} n_{+h}}{n}};$$

elle suit asymptotiquement (pour les grandes valeurs de n), et si l'hypothèse $H0$ est vraie, une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté. On rejette donc $H0$ (et l'on conclut au caractère significatif de la liaison) si χ^2 dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori – en général 0,05 – d' être dépassée par une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté).

1.4 Objectifs

Pour préciser la liaison existant entre les variables X et Y, on souhaite définir un modèle statistique susceptible de fournir des paramètres dont la représentation graphique (de type biplot) illustrera les "*correspondances*" entre les modalités de ces 2 variables. Cette approche sera développée au paragraphe 3.

Une autre approche, très courante dans la littérature francophone, consiste à définir l'Analyse Factorielle des Correspondances (AFC) comme étant le résultat d'une double Analyse en Composantes Principales

- l'ACP des profils–lignes,
- l'ACP des profils–colonnes,

relativement à la métrique dite du χ^2. Cette approche est présentée au paragraphe 2.

Remarque. — :

i. Toute structure d'ordre existant éventuellement sur les modalités de X ou de Y est ignorée par l'AFC

ii. Tout individu présente une modalité et une seule de chaque variable.

iii. Chaque modalité doit avoir été observée au moins une fois ; sinon, elle est supprimée.

2 Double ACP

2.1 Métriques du χ^2

Les correspondances entre modalités évoquées au paragraphe précédant se trouvent exprimées en termes de distances au sens d'une certaine métrique. Ainsi, chaque modalité x_ℓ de X est caractérisée par son profil–ligne représenté par le vecteur $\mathbf{a}^\ell$ de l'espace $\mathbb{R}^c$ muni de la base canonique (les coordonnées de $\mathbf{a}^\ell$ sont les éléments de la ℓème colonne de $\mathbf{A}$). De même, chaque modalité y_h de Y est caractérisée par son profil–colonne représenté par le vecteur $\mathbf{b}^h$ de l'espace $\mathbb{R}^r$ muni de la base canonique.

Ces espaces sont respectivement munis des métriques, dites du χ^2, de matrices $\mathbf{D}_c^{-1}$ et $\mathbf{D}_r^{-1}$. Ainsi, la distance entre deux modalités x_ℓ et x_i de X s'écrit

$$\|\mathbf{a}^\ell - \mathbf{a}^i\|^2_{\mathbf{D}_c^{-1}} = \sum_{h=1}^{c} \frac{1}{f_{+h}}(a_h^\ell - a_h^i)^2,$$

et de même pour les modalités de Y. La métrique du χ^2 introduit les inverses des fréquences marginales des modalités de Y comme *pondérations* des écarts entre éléments de deux profils relatifs à X (et réciproquement) ; elle attribue donc plus de poids aux écarts correspondants à des modalités de *faible effectif* (rares) pour Y.

2.2 ACP des profils–colonnes

On s'intéresse ici à l'ACP du triplet $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$. Dans cette ACP, les "individus" sont les modalités de Y, caractérisées par les profils–colonnes de $\mathbf{T}$, pondérées par les fréquences marginales correspondantes et rangées en lignes dans la matrice $\mathbf{B}'$.

PROPOSITION 5.2. — *Les éléments de l'ACP de* $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_r^{-1}$*–symétrique et semi–définie positive* $\mathbf{BA}$.

Preuve Elle se construit en remarquant successivement que :

i. le barycentre du nuage des profils–colonnes est le vecteur $\mathbf{g}_r$ des fréquence marginales de X,

ii. la matrice $\mathbf{B}\mathbf{D}_c\mathbf{B}' - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r'$ joue le rôle de la matrice des variances–covariances,

iii. la solution de l'ACP est fournie par la D.V.S. de $(\mathbf{B}' - \mathbf{1}\mathbf{g}_r', \mathbf{D}_r^{-1}, \mathbf{D}_c)$, qui conduit à rechercher les valeurs et vecteurs propres de la matrice (**SM**)

$$\mathbf{B}\mathbf{D}_c\mathbf{B}'\mathbf{D}_r^{-1} - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r' = \mathbf{B}\mathbf{A} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1} \quad (\text{ car } \mathbf{B}'\mathbf{D}_r^{-1} = \mathbf{D}_c^{-1}\mathbf{A})$$

iv. les matrices $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et $\mathbf{BA}$ ont les mêmes vecteurs propres associées aux mêmes valeurs propres, à l'exception du vecteur $\mathbf{g}_r$ associé à la valeur propre $\lambda_0 = 0$ de $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et à la valeur propre $\lambda_0 = 1$ de $\mathbf{BA}$.

□

On note $\mathbf{U}$ la matrice contenant les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés de $\mathbf{BA}$. La représentation des "individus" de l'ACP réalisée fournit une représentation des modalités de la variable Y. Elle se fait au moyen des lignes de la matrice des "composantes principales" (**XMV**) :

$$\mathbf{C}_c = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U}.$$

2.3 ACP des profils–lignes

De façon symétrique (ou duale), on s'intéresse à l'ACP des "individus" modalités de X ou profils–lignes (la matrice des données est $\mathbf{A}'$), pondérés par les fréquences marginales des lignes de $\mathbf{T}$ (la matrice diagonale des poids est $\mathbf{D}_r$) et utilisant la métrique du χ^2. Il s'agit donc de l'ACP de $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$.

PROPOSITION 5.3. — *Les éléments de l'ACP de* $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_c^{-1}$*–symétrique et semi–définie positive* $\mathbf{AB}$.

On obtient directement les résultats en permutant les matrices $\mathbf{A}$ et $\mathbf{B}$, ainsi que les indices c et r. Notons $\mathbf{V}$ la matrice des vecteurs propres de la matrice $\mathbf{AB}$; les coordonnées permettant la représentation les modalités de la variable X sont fournies par la matrice :

$$\mathbf{C}_r = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}.$$

Sachant que $\mathbf{V}$ contient les vecteurs propres de $\mathbf{AB}$ et $\mathbf{U}$ ceux de $\mathbf{BA}$, le théorème (A.1) montre qu'il suffit de réaliser une seule analyse, car les résultats de l'autre s'en déduisent simplement :

$$\begin{aligned} \mathbf{V} &= \mathbf{A}\mathbf{U}\mathbf{\Lambda}^{-1/2}, \\ \mathbf{U} &= \mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2}; \end{aligned}$$

$\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) communes aux deux ACP

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U} = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{A}\mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V} = \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2}.\end{aligned}$$

On en déduit les formules dites de *transition* :

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{C}_r\mathbf{\Lambda}^{-1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2}.\end{aligned}$$

La représentation simultanée habituellement construite à partir de ces matrices (option par défaut de SAS) n'est pas a priori justifiée. On lui donnera un sens dans les paragraphes suivants.

3 Modèles pour une table de contingence

On écrit d'abord que chaque fréquence $f_{\ell h}$ de $\mathbf{T}$ correspond à l'observation d'une probabilité théorique $p_{\ell h}$; on modélise donc la table de contingence par cette distribution de probabilités. On précise ensuite le modèle en explicitant l'écriture de $p_{\ell h}$. Différents modèles classiques peuvent être considérés.

3.1 Le modèle log–linéaire

Il consiste à écrire :

$$\ln(p_{\ell h}) = \mu + \alpha_\ell + \beta_h + \gamma_{\ell h}$$

avec des contraintes le rendant identifiable. Ce modèle, très classique, ne sera pas développé ici. On pourra se reporter, par exemple, à Bishop *et al.* (1975).

3.2 Le modèle d'association

Il est encore appelé RC-modèle, ou modèle de Goodman (1991) :

$$p_{\ell h} = \gamma\alpha_\ell\beta_h \exp\left(\sum_{k=1}^{q}\phi_k\mu_{\ell k}\nu_{hk}\right).$$

Ce modèle, muni des contraintes nécessaires, permet de structurer les interactions et de faire des représentations graphiques des lignes et des colonnes de $\mathbf{T}$ au moyen des paramètres $\mu_{\ll k}$ et ν_{hk}. Ces paramètres peuvent être estimés par maximum de vraisemblance ou par moindres carrés.

3.3 Le modèle de corrélation

On écrit ici :

$$p_{\ell h} = p_{\ell +}p_{+h} + \sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k, \tag{5.1}$$

avec $q \leq \inf(r-1, c-1)$, $\lambda_1 \geq \cdots \geq \lambda_q > 0$ et sous les contraintes d'identifiabilité suivantes :

$$\begin{aligned}\sum_{\ell=1}^{r}u_\ell^k = \sum_{h=1}^{c}v_h^k &= 0,\\ \mathbf{u}^{k'}\mathbf{D}_r^{-1}\mathbf{u}^j = \mathbf{v}^{k'}\mathbf{D}_c^{-1}\mathbf{v}^j &= \delta_{kj}.\end{aligned}$$

Remarque. — :

i. Le modèle (5.1) ci-dessus est équivalent au modèle considéré par Goodman (1991) :

$$p_{\ell h} = p_{\ell +} p_{+h} \left(1 + \sum_{k=1}^{q} \sqrt{\lambda_k} \xi_\ell^k \eta_h^k \right), \tag{5.2}$$

moyennant une homothétie sur les paramètres.

ii. La quantité $\sum_{k=1}^{q} \sqrt{\lambda_k} u_\ell^k v_h^k$ exprime l'écart à l'indépendance pour la cellule considérée.

iii. Le modèle suppose que cet écart se décompose dans un sous–espace de dimension $q < \min(c-1, r-1)$.

iv. Les estimations des paramètres $p_{\ell +}, p_{+h}, \lambda_k, \mathbf{u}^k, \mathbf{v}^k$ peuvent être réalisées par maximum de vraisemblance [1] ou par moindres carrés. Dans le contexte de la statistique descriptive, qui est celui de ce cours, il est naturel de retenir cette dernière solution.

3.4 Estimation Moindres Carrés dans le modèle de corrélation

Critère

Considérons les espaces $\mathbb{R}^c$ et $\mathbb{R}^r$ munis de leur base canonique et de leur métrique du χ^2 respectives et notons $\mathbf{P}$ le tableau des probabilités théoriques définies selon le modèle (5.1). Le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n}\mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1}D_c^{-1}}}. \tag{5.3}$$

Estimation

PROPOSITION 5.4. — *L'estimation des paramètres de (5.1) en résolvant (5.3) est fournie par la D.V.S. de* $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$ *à l'ordre* q*. Les probabilités marginales* $p_{\ell +}$ *et* p_{+h} *sont estimées par* $f_{\ell +}$ *et* f_{+h} *tandis que les vecteurs* $\mathbf{u}^k$ *(resp.* $\mathbf{v}^k$*) sont vecteurs propres de la matrice* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*) associés aux valeurs propres* λ_k*.*

On obtient ainsi, d'une autre façon, l'AFC de la table de contingence $\mathbf{T}$.

Preuve Elle se construit à partir de la D.V.S. de $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$:

$$\frac{1}{n} t_\ell^h = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

où les vecteurs $\mathbf{u}^k$ (resp. $\mathbf{v}^k$) sont vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés (resp. $\mathbf{D}_c^{-1}$–orthonormés) de la matrice

$$\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}\frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1} = \mathbf{BA} \quad (\text{resp. } \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1} = \mathbf{AB}),$$

associés aux valeurs propres λ_k.

De plus, le vecteur $\mathbf{g}_r = \mathbf{u}^0$ (resp. $\mathbf{g}_c = \mathbf{v}^0$) est vecteur propre $\mathbf{D}_r^{-1}$–normé (resp. $\mathbf{D}_c^{-1}$–normé) de la matrice $\mathbf{BA}$ (resp. $\mathbf{AB}$) associé à la valeur propre $\lambda_0 = 1$. Enfin, les matrices $\mathbf{AB}$ et $\mathbf{BA}$ sont stochastiques[2] et donc les valeurs propres vérifient :

$$1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_q > 0.$$

[1] On suppose alors que les n $p_{\ell h}$ sont les paramètres de lois de Poisson indépendantes conditionnellement à leur somme qui est fixée et égale à n.

[2] Matrice réelle, carrée, à termes positifs, dont la somme des termes de chaque ligne (ou chaque colonne) vaut 1.

En identifiant les termes, l'approximation de rang $(q+1)$ de la matrice $\mathbf{P}$ s'écrit donc :

$$\widehat{\mathbf{P}}_q = \mathbf{g}_r \mathbf{g}_c' + \sum_{k=1}^{q} \sqrt{\lambda_k} \mathbf{u}^k \mathbf{v}^{k'}$$

et les propriétés d'orthonormalité des vecteurs propres assurent que les contraintes du modèle sont vérifiées.

□

4 Représentations graphiques

4.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{T}$ se transforme encore en :

$$\frac{f_{\ell h} - f_{\ell +} f_{+h}}{f_{\ell +} f_{+h}} = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} \frac{u_\ell^k}{f_{\ell +}} \frac{v_h^k}{f_{+h}}.$$

En se limitant au rang q, on obtient donc, pour chaque cellule (ℓ, h) de la table $\mathbf{T}$, une approximation de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs

$$\frac{u_\ell^k}{f_{\ell +}} \lambda_k^{1/4} \text{ et } \frac{v_h^k}{f_{+h}} \lambda_k^{1/4},$$

termes génériques respectifs des matrices

$$\mathbf{D}_r^{-1} \mathbf{U} \mathbf{\Lambda}^{1/4} \text{ et } \mathbf{D}_c^{-1} \mathbf{V} \mathbf{\Lambda}^{1/4},$$

qui sont encore les estimations des vecteurs ξ_ℓ et η_h du modèle 5.2. Leur représentation (par exemple avec $q = 2$) illustre alors la *correspondance* entre les deux modalités x_ℓ et y_h : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'AFC apparaît ainsi comme la meilleure reconstitution des fréquences $f_{\ell h}$, ou encore la meilleure représentation des écarts relatifs à l'indépendance. La représentation simultanée des modalités de X et de Y se trouve ainsi pleinement justifiée.

4.2 Double ACP

Chacune des deux ACP réalisée permet une représentation des "individus" (modalités) approchant, au mieux, les distances du χ^2 entre les profils–lignes d'une part, les profils–colonnes d'autre part. Les coordonnées sont fournies cette fois par les matrices (de composantes principales)

$$\mathbf{C}_r = \mathbf{D}_r^{-1} \mathbf{U} \mathbf{\Lambda}^{1/2} \text{ et } \mathbf{C}_c = \mathbf{D}_c^{-1} \mathbf{V} \mathbf{\Lambda}^{1/2}.$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de "caution" puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double ACP est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

4.3 Représentations barycentriques

D'autres représentations simultanées, appelées barycentriques, sont proposées en utilisant les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda},$$

ou encore les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

Si l'on considère alors, par exemple, la formule de transition

$$\mathbf{C}_r = \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2} \Longleftrightarrow \mathbf{C}_r\mathbf{\Lambda}^{1/2} = \mathbf{A}'\mathbf{C}_c \Longleftrightarrow \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},$$

on voit que dans la seconde des représentations ci–dessus, chaque modalité x_ℓ de X est représentée par un vecteur qui est barycentre de l'ensemble des vecteurs associés aux modalités de Y, chacun d'eux ayant pour poids l'élément correspondant du l–ième profil–ligne. Là encore, la représentation simultanée s'en trouve parfaitement justifiée. Malheureusement, dans la pratique, les représentations barycentriques sont souvent illisibles ; elles sont, de ce fait, très peu utilisées.

4.4 Autre représentation

La pratique de l'AFC montre que l'interprétation des graphiques est toujours la même, quelle que soit la représentation simultanée choisie parmi les 3 ci–dessus.

On peut ainsi envisager d'utiliser, pour une représentation simultanée des modalités de X et de Y, les coordonnées fournies respectivement par les lignes des matrices

$$\mathbf{D}_r^{-1}\mathbf{U} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}.$$

L'interprétation du graphique sera toujours la même et les matrices ci–dessus, outre leur simplicité, présentent l'avantage de conduire a une représentation graphique qui reste invariante lorsque l'on utilise la technique d'Analyse Factorielle des Correspondances Multiples (voir chapitre suivant) sur les données considérées ici.

4.5 Aides à l'interprétation

Les qualités de représentation dans la dimension choisie et les contributions des modalités de X ou de Y se déduisent aisément de celles de l'ACP Ces quantités sont utilisées à la fois pour choisir la dimension de l'AFC et pour interpréter ses résultats dans la dimension choisie.

Mesure de la qualité globale

Pour une dimension donnée q ($1 \leq q \leq d = \inf(r-1, c-1)$), la qualité globale des représentations graphiques en dimension q se mesure par le rapport entre la somme des q premières valeurs propres de l'AFC et leur somme complète de 1 à d.

Compte–tenue de la propriété $\sum_{k=1}^{d} \lambda_k = \Phi^2$ (voir en 6.1), la qualité de la représentation dans la k–ième dimension s'écrit

$$\frac{n\lambda_k}{\chi^2}.$$

On parle encore de part du khi–deux expliquée par la k–ième dimension (voir les sorties du logiciel SAS).

Mesure de la qualité de chaque modalité

Pour chaque modalité de X (resp. de Y), la qualité de sa représentation en dimension q se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^c$ (resp. dans $\mathbb{R}^r$) et sa projection $\mathbf{D}_c^{-1}$–orthogonale (resp. $\mathbf{D}_r^{-1}$–orthogonale) dans le sous–espace principal de dimension q.

Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de $\mathbf{C}_r$ (resp. de $\mathbf{C}_c$).

Contributions à l'inertie totale

L'inertie totale (en dimension d) du nuage des profils–lignes (resp. des profils–colonnes) est égale à la somme des d valeurs propres. La part due au i–ième profil–ligne (resp. au j–ième profil–colonne) valant $f_{\ell+} \sum_{k=1}^{d} (c_{r\ell}^k)^2$ (resp. $f_{+h} \sum_{k=1}^{d} (c_{ch}^k)^2$), les contributions à l'inertie totale s'en déduisent immédiatement.

Contributions à l'inertie selon chaque axe

Il s'agit de quantités analogues à celles ci–dessus, dans lesquelles il n'y a pas de sommation sur l'indice k. Ces quantités sont utilisées dans la pratique pour sélectionner les modalités les plus importantes, c'est–à–dire celles qui contribuent le plus à la définition de la liaison entre les 2 variables X et Y.

Remarque

En général, on n'interprète pas les axes d'une AFC (en particulier parce qu'il n'y a pas de variable quantitative intervenant dans l'analyse). L'interprétation s'appuie surtout sur la position relative des différentes modalités repérées comme les plus importantes.

5 Exemple

L'exemple des données bancaires se prête mal à une analyse des correspondances, aucun couple de variable qualitative ne conduit à des représentations intéressantes. La table de contingence étudiée décrit la répartition des exploitations agricoles de la région Midi–Pyrénées dans les différents départements en fonction de leur taille. Elle croise la variable qualitative *département*, à 8 modalités, avec la variable *taille de l'exploitation*, quantitative découpée en 6 classes. Les données, ainsi que les résultats numériques obtenus avec la procédure `corresp` de SAS/STAT, sont fournis en annexe.

La figure 5 présente le premier plan factoriel utilisant les coordonnées obtenues par défaut, c'est–à–dire celles de la double ACP.

6 Compléments

6.1 Propriétés

- *Formule de reconstitution des données.* On appelle ainsi l'approximation d'ordre q (c'est–à–dire fournie par l'AFC en dimension q) de la table des fréquences initiales ($\frac{1}{n}\mathbf{T}$) :

$$f_{\ell h} \simeq f_{\ell+} f_{+h} \sum_{k=1}^{q} \sqrt{\lambda_k} u_\ell^k v_h^k.$$

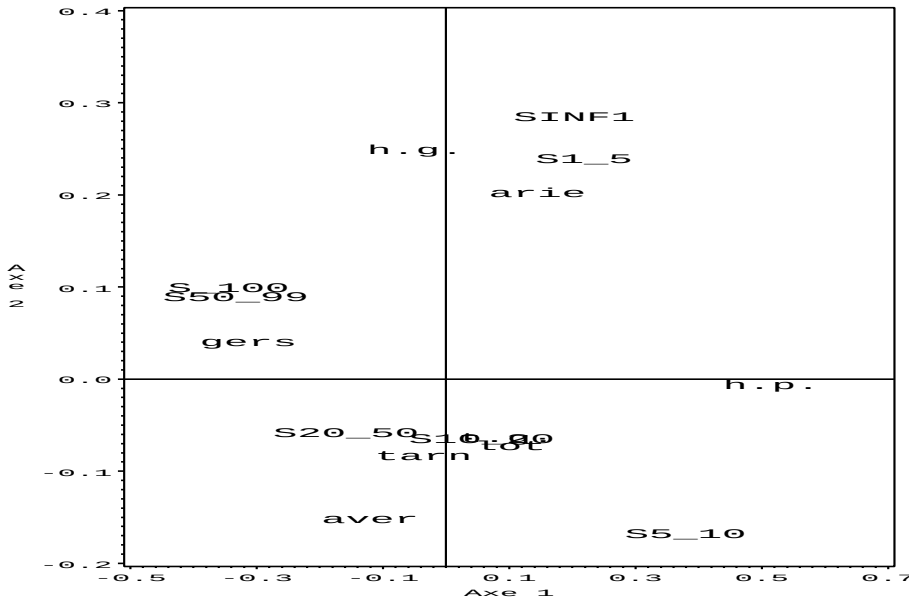

FIG. 5.1 – *Répartition des exploitations agricoles par taille et par département. Premier plan de l'AFC.*

- Les valeurs propres vérifient :

$$\sum_{k=1}^{d} \lambda_k = \Phi^2.$$

En effet, on vérifie facilement :

$$\mathrm{tr}\mathbf{AB} = \sum_{k=0}^{d} \lambda_k = 1 + \frac{\chi^2}{n} = 1 + \Phi^2;$$

d'où le résultat.

6.2 Invariance

- Les tables de contingence $\mathbf{T}$ et $\alpha\mathbf{T}, \alpha \in \mathbb{R}_+^*$, admettent la même AFC (évident).
- *Propriété d'équivalence distributionnelle* : si deux lignes de $\mathbf{T}$, ℓ et i, ont des effectifs proportionnels, alors les représentations de x_ℓ et x_i sont confondues (leurs profils sont identiques) et le regroupement de x_ℓ et x_i en une seule modalité (en additionnant les effectifs) laisse inchangées les représentations graphiques (même chose pour les colonnes de $\mathbf{T}$). Cette propriété est une conséquence de la métrique du χ^2.

6.3 Choix de la dimension q

Le choix de la dimension pose les mêmes problèmes qu'en ACP De nombreuses techniques empiriques ont été proposées (essentiellement : part d'inertie expliquée, éboulis des valeurs propres). Il existe également une approche probabiliste qui peut donner des indications intéressantes. Nous la détaillons ci–dessous.

Posons

$$\widehat{n_{\ell h}^q} = n f_{\ell +} f_{+h} + n \sum_{k=1}^{q} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

estimation d'ordre q de l'effectif conjoint de la cellule (ℓ, h). Alors, sous certaines conditions (échantillonnage, n grand, modèle multinomial ...), on peut montrer que

$$K_q = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \widehat{n_{\ell h}^q})^2}{\widehat{n_{\ell h}^q}} \simeq n \sum_{k=q+1}^{d} \lambda_k$$

suit approximativement une loi de χ^2 à $(r-q-1)(c-q-1)$ degrés de liberté. On peut donc retenir pour valeur de q la plus petite dimension pour laquelle K_q est inférieure à la valeur limite de cette loi. Le choix $q = 0$ correspond à la situation où les variables sont proche de l'indépendance en probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.

Chapitre 6

Analyse des Correspondances Multiples

Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p ($p > 2$) variables qualitatives simultanément observées sur n individus. Elle est aussi souvent utilisée pour la construction de *scores* comme préalable à une méthode de classification (nuées dynamiques) nécessitant des données quantitatives.

1 Codages de variables qualitatives

1.1 Tableau disjonctif complet

Soit X une variable qualitative à c modalités. On appelle *variable indicatrice* de la k–ième modalité de x ($k = 1, \ldots, c$), la variable $X_{(k)}$ définie par

$$X_{(k)}(i) = \begin{cases} 1 & \text{si } X(i) = \mathcal{X}_k, \\ 0 & \text{sinon,} \end{cases}$$

où i est un individu quelconque et $\mathcal{X}_k$ est la k–ième modalité de X. On notera n_k l'effectif de $\mathcal{X}_k$.

On appelle *matrice des indicatrices* des modalités de X, et l'on notera $\mathbf{X}$, la matrice $n \times c$ de terme général :

$$x_i^k = X_{(k)}(i).$$

On vérifie :

$$\sum_{k=1}^{c} x_i^k = 1, \forall i \text{ et } \sum_{i=1}^{n} x_i^k = n_k.$$

Considérons maintenant p variables qualitatives $X^1, \ldots, X^p$. On note c_j le nombre de modalités de X^j, $c = \sum_{j=1}^{p} c_j$ et $\mathbf{X}_j$ la matrice des indicatrices de X^j.

On appelle alors *tableau disjonctif complet* la matrice $\mathbf{X}, n \times c$, obtenue par concaténation des matrices $\mathbf{X}_j$:

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p].$$

$\mathbf{X}$ vérifie :

$$\sum_{k=1}^{c} x_i^k = p, \forall i \text{ et } \sum_{i=1}^{n} \sum_{k=1}^{c} x_i^k = np.$$

D'autre part, la somme des éléments d'une colonne de $\mathbf{X}$ est égale à l'effectif marginal de la modalité de la variable X^j correspondant à cette colonne.

1.2 Tableau de Burt

On observe toujours p variables qualitatives sur un ensemble de n individus. On appelle *tableau de Burt* la matrice $\boldsymbol{\mathcal{B}}$, $c \times c$, définie par :

$$\boldsymbol{\mathcal{B}} = \mathbf{X}'\mathbf{X}.$$

On peut écrire $\boldsymbol{\mathcal{B}} = [\boldsymbol{\mathcal{B}}_{jl}]$ $(j = 1, \ldots, p\,;\, l = 1, \ldots, p)$; chaque bloc $\boldsymbol{\mathcal{B}}_{jl}$, de dimension $c_j \times c_l$, est défini par :

$$\boldsymbol{\mathcal{B}}_{jl} = \mathbf{X}'_j\mathbf{X}_l.$$

Si $j \neq l$, $\boldsymbol{\mathcal{B}}_{jl}$ est la table de contingence obtenue par croisement des variables X^j en lignes et X^l en colonnes. Si $j = l$, le bloc diagonal $\boldsymbol{\mathcal{B}}_{jj}$ est lui–même une matrice diagonale vérifiant :

$$\boldsymbol{\mathcal{B}}_{jj} = \text{ diag } (n_1^j, \ldots, n_{c_j}^j).$$

La matrice $\boldsymbol{\mathcal{B}}$ est symétrique, d'effectifs marginaux $n_l^j p$ et d'effectif total np^2.

1.3 La démarche suivie dans ce chapitre

La généralisation de l'AFC à plusieurs variables qualitatives repose sur certaines propriétés observées dans le cas élémentaire où $p = 2$. On s'intéresse tout d'abord aux résultats fournis par l'AFC usuelle réalisée sur le tableau disjonctif complet $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ relatif à 2 variables qualitatives X^1 et X^2 ; $\mathbf{X}$ est alors considéré comme une table de contingence (paragraphe 2). Ensuite, on suit la même démarche avec l'AFC réalisée sur le tableau de Burt $\boldsymbol{\mathcal{B}}$ relatif à X^1 et X^2 (paragraphe 3). Enfin, en utilisant les propriétés obtenues dans les deux premiers cas, on généralise cette double approche à un nombre quelconque p de variables qualitatives ; on définit ainsi l'Analyse Factorielle des Correspondances Multiples (paragraphe 4).

2 AFC du tableau disjonctif complet relatif à 2 variables

2.1 Données

On note toujours X^1 et X^2 les 2 variables qualitatives considérées et r et c leurs nombres respectifs de modalités.

Les matrices intervenant dans l'AFC usuelle sont reprises ici avec les mêmes notations, mais surlignées. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]\,; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n\,; \\
\overline{\mathbf{D}}_c &= \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\boldsymbol{\Delta}\,; \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}'\,; \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}}\,\overline{\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

On considère ici l'AFC comme une double ACP : celle des profils–lignes $\overline{\mathbf{A}}$, puis celle des profils–colonnes $\overline{\mathbf{B}}$.

2.2 ACP des profils–lignes

Les profils–lignes, provenant de $\overline{\mathbf{T}}$, sont associés aux n individus observés. Leur ACP conduit ainsi à une représentation graphique des individus, inconnue en AFC classique.

PROPOSITION 6.1. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}}_c^{-1}$*–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}.$$

Les $r + c$ *valeurs propres de* $\overline{\mathbf{A}}\,\overline{\mathbf{B}}$ *s'écrivent*

$$\mu_k = \frac{1 \pm \sqrt{\lambda_k}}{2},$$

où les λ_k *sont les valeurs propres de la matrice* $\mathbf{AB}$ *(donc celles de l'AFC classique de* X^1 *et* X^2*).*

Les vecteurs propres $\overline{\mathbf{D}}_c^{-1}$*–orthonormés associés se mettent sous la forme*

$$\overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix};$$

la matrice $\mathbf{U}$ *(resp.* $\mathbf{V}$*) contient les vecteurs propres* $\mathbf{D}_r^{-1}$*–orthonormés (resp.* $\mathbf{D}_c^{-1}$*-orthonormés) de la matrice* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*) ; autrement dit, les matrices* $\mathbf{U}$ *et* $\mathbf{V}$ *sont les matrices de vecteurs propres obtenues en faisant l'AFC classique de la table de contingence croisant* X^1 *et* X^2*.*

La matrice des composantes principales s'écrit

$$\overline{\mathbf{C}}_r = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r + \mathbf{X}_2\mathbf{C}_c\right]\mathbf{\Lambda}^{-1/2},$$

où $\mathbf{C}_r$ *et* $\mathbf{C}_c$ *sont encore les matrices de composantes principales de l'AFC classique.*

Dans la pratique, on ne considère que les $d = \inf(r-1, c-1)$ plus grandes valeurs propres différentes de 1, ainsi que les vecteurs propres associés. Les valeurs propres sont rangées dans la matrice

$$\mathbf{M} = \text{ diag }(\mu_1, \ldots, \mu_d) = \frac{1}{2}\left[\mathbf{I}_d + \mathbf{\Lambda}^{1/2}\right].$$

Les autres valeurs propres non nulles sont dues à l'artifice de construction de la matrice à diagonaliser ; elles n'ont donc pas de signification statistique.

On notera que la matrice $\overline{\mathbf{C}}_r$, $n \times d$, fournit les coordonnées permettant la représentation graphique des individus sur les axes factoriels.

2.3 ACP des profils–colonnes

Les profils–colonnes sont associés aux $r + c$ modalités des variables. Leur ACP conduit donc à une représentation graphique de ces modalités dont on verra qu'elle est très voisine de celle fournie par une AFC classique.

PROPOSITION 6.2. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}}_r^{-1}$*–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{2n}\left[\mathbf{X}_1\mathbf{D}_r^{-1}\mathbf{X}_1' + \mathbf{X}_2\mathbf{D}_c^{-1}\mathbf{X}_2'\right].$$

Les $r+c$ *valeurs propres non nulles de* $\overline{\mathbf{B}}\,\overline{\mathbf{A}}$ *sont les* μ_k*.*

Les vecteurs propres $\overline{\mathbf{D}}_r^{-1}$*–orthonormés associés se mettent sous la forme :*

$$\overline{\mathbf{U}} = \frac{1}{n}\overline{\mathbf{C}}_r\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{M}^{1/2}.$$

Ainsi, l'AFC du tableau disjonctif complet permet, grâce aux coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_c$, une représentation simultanée des modalités des 2 variables. Cette représentation est très voisine de celle obtenue par l'AFC classique, définie au chapitre précédent. Une simple homothétie sur chaque axe factoriel, de rapport $\sqrt{\frac{1+\sqrt{\lambda_k}}{2\lambda_k}}$, permet de passer de l'une à l'autre.

De plus, cette approche permet aussi de réaliser une représentation graphique des individus avec les coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_r$. À un facteur près, chaque individu apparaît comme le barycentre des 2 modalités qu'il a présentées. Dans le cas où n est grand, le graphique des individus a néanmoins peu d'intérêt ; seule sa forme générale peut en avoir un.

Remarque. — Si, dans l'AFC classique, on choisit d'utiliser, pour la représentation simultanée des modalités de X^1 et de X^2, les lignes des matrices

$$\mathbf{C}_r^* = \mathbf{D}_r^{-1}\mathbf{U} = \mathbf{C}_r\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{C}_c^* = \mathbf{D}_c^{-1}\mathbf{V} = \mathbf{C}_c\mathbf{\Lambda}^{-1/2}$$

(voir chapitre précédent, sous–section 4.4), alors on obtient par AFC du tableau disjonctif complet la matrice

$$\overline{\mathbf{C}}_c^* = \overline{\mathbf{C}}_c\mathbf{M}^{-1/2} = \begin{bmatrix} \mathbf{C}_r^* \\ \mathbf{C}_c^* \end{bmatrix};$$

il y a invariance de la représentation des modalités lorsqu'on passe d'une méthode à l'autre. Pour les individus, on obtient

$$\overline{\mathbf{C}}_r^* = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r^* + \mathbf{X}_2\mathbf{C}_c^*\right]\mathbf{M}^{-1/2}$$

(le commentaire est alors le même qu'avec $\overline{\mathbf{C}}_r$).

3 AFC du tableau de Burt relatif à 2 variables

Dans cette section, on s'intéresse aux résultats fournis par l'AFC réalisée sur le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, $(r+c)\times(r+c)$, relatif aux 2 variables X^1 et X^2 ; $\mathcal{B}$ est encore considéré comme une table de contingence. La matrice $\mathcal{B}$ étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; il suffit donc de considérer une seule ACP

Les notations des matrices usuelles de l'AFC sont maintenant réutilisées surmontées d'un tilde. On obtient ainsi :

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}; \\
\widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\mathbf{\Delta} = \overline{\mathbf{D}}_c\,; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.
\end{aligned}$$

On considère encore l'AFC comme l'ACP des profils–lignes $\widetilde{\mathbf{A}}$ (ou des profils–colonnes $\widetilde{\mathbf{B}}$).

PROPOSITION 6.3. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$*–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$*–orthonormés*

$$\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}.$$

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix}\mathbf{\Lambda}^{-1/2}\mathbf{M}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant une représentation simultanée des modalités des deux variables. À une homothétie près, cette représentation est identique à celle de l'AFC classique, réalisée sur la table de contingence $\mathbf{T}$ (mais le rapport d'homothétie, sur chaque axe, n'est plus le même qu'avec $\overline{\mathbf{C}}_c$).

Remarque. —

- En reprenant les notations de la remarque 2.3, on obtient ici :

$$\widetilde{\mathbf{C}_r^*}\,(=\widetilde{\mathbf{C}_c^*}) = \widetilde{\mathbf{C}_r}\mathbf{M}^{-1} = \overline{\mathbf{C}}_c^* = \begin{bmatrix} \mathbf{C}_r^* \\ \mathbf{C}_c^* \end{bmatrix}.$$

  Ainsi, si l'on utilise ce mode de représentation graphique, les trois approches de l'AFC que nous avons présentées conduisent à la même représentation simultanée des modalités des 2 variables : il y a donc invariance de cette représentation.
- Dans les deux cas d'AFC considérés dans ce chapitre (sur tableau disjonctif complet et sur tableau de Burt) on trouve, par construction, des valeurs propres non nulles sans signification statistique. En conséquence, les critères de qualité s'exprimant comme une "part d'inertie expliquée" n'ont plus de signification.
- L'AFC sur tableau de Burt ne prend en compte que l'information contenue dans $\mathcal{B}$ qui ne considère que les croisements de variables prises deux à deux. En conséquence, les interactions de niveau plus élevé sont ignorées par cette approche, à moins de procéder à des recodages de variables comme l'explique l'exemple présenté dans la section 5.

4 Analyse Factorielle des Correspondances Multiples

4.1 Définition

On considère maintenant p variables qualitatives $(p \geq 3)$ notées $\{X^j\,; j = 1, \ldots, p\}$, possédant respectivement c_j modalités, avec $c = \sum_{j=1}^{p} c_j$. On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids $1/n$.

Soit $\mathbf{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_p]$ le tableau disjonctif complet des observations ($\mathbf{X}$ est $n \times c$) et $\boldsymbol{\mathcal{B}} = \mathbf{X}'\mathbf{X}$ le tableau de Burt correspondant ($\boldsymbol{\mathcal{B}}$ est carré d'ordre c, symétrique).

DÉFINITION 6.4. — *On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables $(X^1, \ldots, X^p)$ relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice $\mathbf{X}$ soit sur la matrice $\boldsymbol{\mathcal{B}}$.*

On note n_k^j $(1 \leq j \leq p, 1 \leq k \leq c_j)$ l'effectif de la k–ième modalité de X^j, $\mathbf{D}_j = \dfrac{1}{n}$ diag $(n_1^j, \ldots, n_{c_j}^j)$ et $\boldsymbol{\Delta} =$ diag $(\mathbf{D}_1 \ldots \mathbf{D}_p)$ ($\boldsymbol{\Delta}$ est carrée d'ordre c et diagonale).

4.2 AFC du tableau disjonctif complet X

Comme dans le cas $p = 2$, on reprend les notations de l'AFC classique en les surlignant. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X}\,; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n\,; \\
\overline{\mathbf{D}}_c &= \frac{1}{p}\boldsymbol{\Delta}\,; \\
\overline{\mathbf{A}} &= \frac{1}{p}\mathbf{X}'\,; \\
\overline{\mathbf{B}} &= \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

ACP des profils–lignes

PROPOSITION 6.5. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_c^{-1}$–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{np}\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}^{-1}.$$

Il y a m $(m \leq c - p)$ valeurs propres notées μ_k, $(0 < \mu_k < 1)$ rangées dans la matrice diagonale $\mathbf{M}$.

La matrice des vecteurs propres $\overline{\mathbf{D}}_c^{-1}$–orthonormés associés se décompose en blocs de la façon suivante :

$$\overline{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_p \end{bmatrix};$$

chaque bloc $\mathbf{V}_j$ est de dimension $c_j \times m$.

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_r = \sum_{j=1}^{p} \mathbf{X}_j \mathbf{D}_j^{-1} \mathbf{V}_j.$$

Comme dans le cas $p = 2$, la matrice des composantes principales permet de réaliser une représentation graphique des individus dans laquelle chacun apparaît, à un facteur près, comme le barycentre des p modalités qu'il a présentées.

Remarque. — La généralisation au cas $p > 2$ restreint les propriétés. Ainsi, les vecteurs des blocs $\mathbf{V}_j$ ne sont pas les vecteurs propres $\mathbf{D}_j^{-1}$–orthonormés d'une matrice connue.

ACP des profils–colonnes

PROPOSITION 6.6. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_r^{-1}$–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{np}\mathbf{X}\boldsymbol{\Delta}^{-1}\mathbf{X}' = \frac{1}{np}\sum_{j=1}^{p} \mathbf{X}_j \mathbf{D}_j^{-1} \mathbf{X}_j'.$$

La matrice des vecteurs propres $\overline{\mathbf{D}}_r^{-1}$–orthonormés vérifie :

$$\overline{\mathbf{U}} = \overline{\mathbf{B}\mathbf{V}}\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = p\boldsymbol{\Delta}^{-1}\overline{\mathbf{V}}\mathbf{M}^{1/2}\ ;$$

elle se décompose en blocs sous la forme :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}.$$

Chaque bloc $\mathbf{C}_j$, de dimension $c_j \times m$, fournit en lignes les coordonnées des modalités de la variable X^j permettant la représentation graphique simultanée.

4.3 AFC du tableau de Burt $\mathcal{B}$

Le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, carré d'ordre c, étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; on ne considère donc ici qu'une seule ACP

En utilisant encore le tilde dans ce cas, les matrices usuelles de l'AFC deviennent :

$$\begin{aligned} \widetilde{\mathbf{T}} &= \mathcal{B}\ ; \\ \widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{p}\boldsymbol{\Delta} = \overline{\mathbf{D}}_c\ ; \\ \widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{np}\mathcal{B}\boldsymbol{\Delta}^{-1} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}. \end{aligned}$$

PROPOSITION 6.7. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à p variables qualitatives conduit à l'analyse spectrale de la matrice $\widetilde{\mathbf{D}_c}^{-1}$–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$–orthonormés $\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}}$.

Les valeurs propres associées vérifient $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \overline{\mathbf{C}}_c\mathbf{M}^{1/2} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix} \mathbf{M}^{1/2}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant la représentation simultanée des modalités de toutes les variables (on ne peut pas faire de représentation des individus si l'on fait l'AFC du tableau de Burt).

4.4 Variables illustratives

Soit X^0 une variable qualitative, à c_0 modalités, observée sur les mêmes n individus que les X^j et n'étant pas intervenue dans l'AFCM Soit $\mathbf{T}_{0j}$ la table de contingence $c_0 \times c_j$ croisant les variables X^0 en lignes et X^j en colonnes. L'objectif est maintenant de représenter les modalités de cette variable supplémentaire X^0 dans le graphique de l'AFCM réalisée sur $X^1, \ldots, X^p$. Pour cela, on considère les matrices :

$$\begin{aligned} \mathbf{B}_0 &= [\mathbf{T}_{01}|\ldots|\mathbf{T}_{0p}] \ ; \\ \mathbf{D}_0 &= \frac{1}{n}\,\text{diag}\,(n_1^0, \ldots, n_{c_0}^0) \ ; \\ \mathbf{A}_0 &= \frac{1}{np}\mathbf{D}_0^{-1}\mathbf{B}_0. \end{aligned}$$

Les coordonnées des modalités de la variable supplémentaires X^0 sur les axes factoriels sont alors fournies dans les lignes de la matrice

$$\mathbf{C}_0 = \mathbf{A}_0\widetilde{\mathbf{D}_c}^{-1}\widetilde{\mathbf{V}} = p\mathbf{A}_0\mathbf{\Delta}^{-1}\overline{\mathbf{V}}.$$

4.5 Interprétation

Les représentations graphiques sont interprétées de manière analogue à ce qui est fait dans l'AFC de deux variables, bien que la représentation simultanée des modalités de toutes les variables ne soit pas, en toute rigueur, réellement justifiée.

Les "principes" suivants sont donc appliqués :

- on interprète globalement les proximités et les oppositions entre les modalités des différentes variables, comme en AFC, en privilégiant les modalités suffisamment éloignées du centre du graphique (attention aux modalités à faible effectif !) ;

- les rapports de valeurs propres ne sont pas interprétables comme indicateurs de qualité globale ; on peut néanmoins regarder la décroissance des premières valeurs propres pour choisir la dimension ;
- les coefficients de qualité de chaque modalité ne peuvent pas être interprétés ; seules les contributions des modalités à l'inertie selon les axes sont interprétées, selon le même principe qu'en AFC

5 Exemple

L'AFCM ne donne pas non plus de résultats intéressants sur les données bancaires.

5.1 Les données

La littérature anglo–américaine présente souvent des données relatives à plusieurs variables qualitatives sous la forme d'une table de contingence *complète* (5.1). C'est le cas de l'exemple ci–dessous qui décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokyo) sur des patientes atteintes d'un cancer du sein. On se propose d'étudier la survie de ces patientes, trois ans après le diagnostic. En plus de cette information, quatre autres variables sont connues pour chacune des patientes :

- le centre de diagnostic,
- la tranche d'âge,
- le degré d'inflammation chronique,
- l'apparence relative (bénigne ou maligne).

L'objectif de cette étude est une analyse descriptive de cette table en recherchant à mettre en évidence les facteurs de décès.

5.2 Analyse brute

On se reportera à la figure 5.1. La variable survie, qui joue en quelques sortes le rôle de variable à expliquer, est très proche de l'axe 2 et semble liée à chacune des autres variables.

5.3 Analyse des interactions

Pour essayer de mettre en évidence d'éventuelles interactions entre variables, les données sont reconsidérées de la façon suivante :

- les variables `centre` et `âge` sont croisées, pour construire une variable `c_x_âge`, à 9 modalités ;
- les variables `inflam` et `appar` sont également croisées pour définir la variable `histol`, à 4 modalités.

Une nouvelle analyse est alors réalisée en considérant comme actives les deux variables nouvellement créées, ainsi que la variable `survie`, et comme illustratives les variables initiales : `centre, âge, inflam, appar`. Les résultats sont donnés dans la figure 5.3.

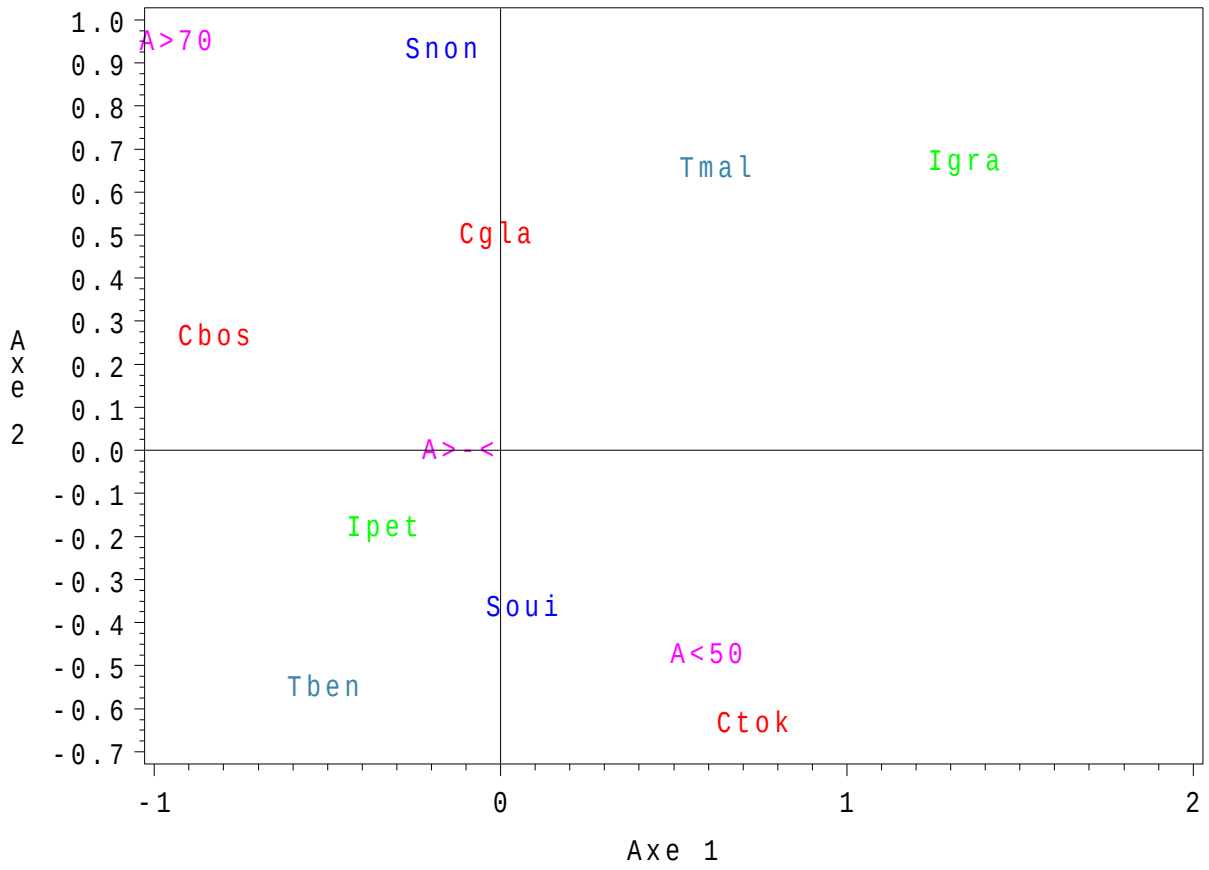

FIG. 6.1 – *Cancer du sein : analyse des données brutes.*

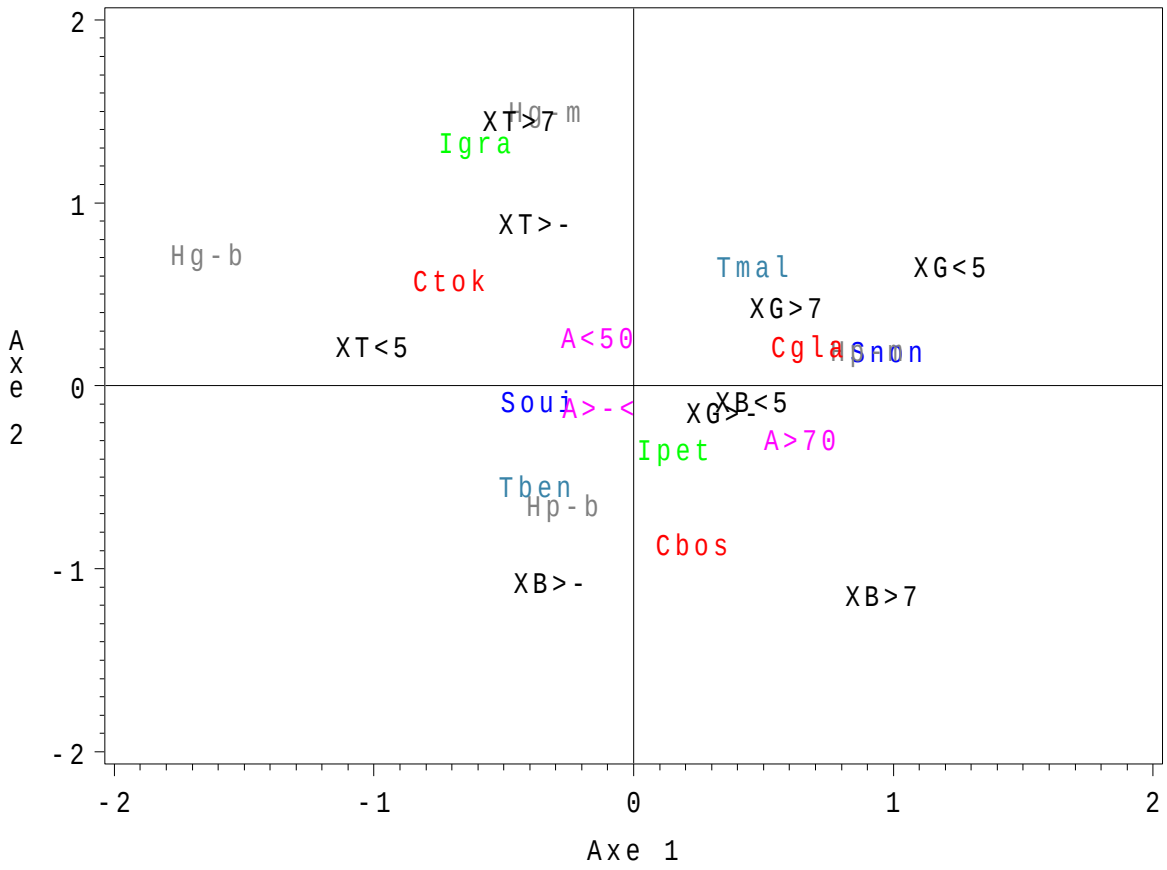

FIG. 6.2 – *Cancer du sein : analyse des interactions.*

			Histologie			
			Inflammation minime		Grande inflammation	
Centre	*Âge*	*Survie*	Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

TAB. 6.1 – *Données sous la forme d'une table de contingence complète*

Chapitre 7

Positionnement multidimensionnel

1 Introduction

Considérons n individus. Contrairement aux chapitres précédents, on ne connaît pas les observations de p variables sur ces n individus mais les $1/2n(n-1)$ valeurs d'un indice (de distance, similarité ou dissimilarité) observées ou construites pour chacun des couples d'individus. Ces informations sont contenues dans une matrice $(n \times n)$ $\boldsymbol{\mathcal{D}}$. L'objectif du *positionnement multidimensionnel* (multidimensional scaling ou MDS ou ACP d'un tableau de distances) est de construire, à partir de cette matrice, une représentation euclidienne des individus dans un espace de dimension réduite q qui approche au "mieux" les indices observés.

Exemple : Considérons un tableau avec, en ligne, les individus d'un groupe et en colonne les pays de la C.E. La valeur 1 est mise dans une case lorsque l'individu de la ligne a passé au moins une nuit dans le pays concerné. Il est alors facile de construire une matrice de similarité avec un indice qui compte le nombre de 1 apparaissant dans les mêmes colonnes de tous les couples d'individus. L'objectif est ensuite d'obtenir une représentation graphique rapprochant les individus ayant visité les mêmes pays.

Les preuves des propositions sont omises dans cet exposé succinct, elles sont à chercher dans la bibliographie. Voir par exemple Mardia et col. (1979).

2 Distance, similarités

2.1 Définitions

DÉFINITION 7.1. —

- *Une matrice* $(n \times n)$ $\boldsymbol{\mathcal{D}}$ *est appelée matrice de distance si elle est symétrique et si :*

$$d_j^j = 0 \text{ et } \forall (j,k), j \neq k, d_j^k \geq 0.$$

- *Une matrice* $(n \times n)$ $\boldsymbol{\mathcal{C}}$ *est appelée matrice de similarité si elle est symétrique et si*

$$\forall (j,k), c_j^k \leq c_j^j.$$

Une matrice de similarité se transforme en matrice de distance par :

$$d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}.$$

DÉFINITION 7.2. — *Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs* $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ *dans un espace euclidien* E *de sorte que*

$$d_j^{k\,2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle \ .$$

On note $\mathbf{A}$ la matrice issue de $\boldsymbol{\mathcal{D}}$ de terme général $d_j^k = -1/2 d_j^{k\,2}$ et $\mathbf{H}$ la matrice de centrage :

$$\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D},$$

qui est la matrice de projection sur le sous-espace $\mathbf{D}$-orthogonal au vecteur $\mathbf{1}$ dans l'espace euclidien F des variables muni de la métrique des poids.

PROPOSITION 7.3. —

- *Soit* $\boldsymbol{\mathcal{D}}$ *une matrice de distance et* $\mathbf{B}$ *la matrice obtenue par double centrage de la matrice* $\mathbf{A}$ *issue de* $\boldsymbol{\mathcal{D}}$ *:*

  $$\mathbf{B} = \mathbf{HAH},$$

  alors $\boldsymbol{\mathcal{D}}$ *est une matrice euclidienne si et seulement si* $\mathbf{B}$ *est positive (toutes ses valeurs propres sont positives ou nulles).*
- *Si la matrice de similarité* $\boldsymbol{\mathcal{C}}$ *est positive alors la matrice de distance* $\boldsymbol{\mathcal{D}}$ *déduite est euclidienne.*

3 Distances entre variables

L'un des intérêts pratique du positionnement multidimensionnel est d'aider à comprendre, visualiser, les structures de liaison dans un grand ensemble de variables. On obtient ainsi des indications pour guider le choix d'un sous-ensemble de variables, par exemple les plus liées à une variable à expliquer. Cette approche nécessite la définition d'indices de similarité entre variables. Beaucoup sont proposés dans la littérature. Nous en retenons trois pour différents types de variables.

3.1 Variables quantitatives

On note X et Y deux variables statistiques dont les observations sur les mêmes n individus sont rangées dans les vecteurs *centrés* $\mathbf{x}$ et $\mathbf{y}$ de l'espace euclidien F muni de la métrique des poids $\mathbf{D}$. On vérifie facilement :

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \text{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}} \|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

La valeur absolue ou le carré du coefficient de corrélation définissent des indices de similarité entre deux variables quantitatives. Il est facile d'en déduire des distances. On préfère par la suite utiliser le carré du coefficient de corrélation qui induit une distance euclidienne :

$$d^2(X,Y) = 2(1 - \text{cor}^2(X,Y)).$$

PROPOSITION 7.4. — *La distance entre variables quantitatives $d^2(X,Y)$ est encore le carré de la distance $\|\mathbf{P_x} - \mathbf{P_y}\|_{\mathbf{D}}$ entre les projecteurs $\mathbf{D}$-orthogonaux sur les directions engendrées par les vecteurs $\mathbf{x}$ et $\mathbf{y}$.*

Démonstration. — Un projecteur de rang 1 s'écrit :$\mathbf{P_x} = \mathbf{x}\mathbf{x}'\,1/(\|\mathbf{x}\|^2_{\mathbf{D}})\mathbf{D}$,

$$\|\mathbf{P_x} - \mathbf{P_y}\|^2_{\mathbf{D}} = \text{tr}(\mathbf{P_x} - \mathbf{P_y})'\mathbf{D}(\mathbf{P_x} - \mathbf{P_y}) = \|\mathbf{P_x}\|^2_{\mathbf{D}} + \|\mathbf{P_y}\|^2_{\mathbf{D}} - 2\text{tr}\mathbf{P_x}'\mathbf{D}\mathbf{P_y}.$$

Comme un projecteur est de norme son rang c'est-à-dire ici 1 et que :

$$\text{tr}\mathbf{P_x}'\mathbf{D}\mathbf{P_y} = \text{tr}\frac{\mathbf{x}\mathbf{x}'}{\|\mathbf{x}\|^2_{\mathbf{D}}}\mathbf{D}\frac{\mathbf{y}\mathbf{y}'}{\|\mathbf{y}\|^2_{\mathbf{D}}}\mathbf{D} = \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\|\mathbf{y}\|_{\mathbf{D}}}\frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\|\mathbf{y}\|_{\mathbf{D}}} = \text{cor}^2(X,Y)$$

alors, $\|\mathbf{P_x} - \mathbf{P_y}\|^2_{\mathbf{D}} = 2(1 - \text{cor}^2(X,Y))$. ∎

3.2 Variables qualitatives

Considérons maintenant deux variables qualitatives, X à r modalités et Y à c modalités. De nombreux indices de similarité ont été proposés : la "prob value" du test du χ^2 d'indépendance, le V de Cramer, le Φ^2 de Pearson, le T de Tschuprow (cf. T1)…Ce dernier a une signification particulière. Soit $\mathbf{X}$ et $\mathbf{Y}$ les matrices contenant les variables indicatrices des modalités des variables et $\mathbf{P_X}$, $\mathbf{P_Y}$ les projecteurs $\mathbf{D}$-orthogonaux sur les sous-espaces engendrés par ces indicatrices. On montre (cf. Saporta 1976) alors la

PROPOSITION 7.5. — *Dans le cas de 2 variables qualitatives,*

$$\|\mathbf{P_X} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - T^2(X,Y)).$$

Ainsi, en utilisant comme indice de similarité le carré du T de Tschuprow entre deux variables qualitatives, on définit une distance euclidienne entre ces variables.

3.3 Variables quantitative et qualitative

La même démarche s'adapte à l'étude d'une liaison entre une variable quantitative X, son projecteur associé $\mathbf{P_x}$ et une variable qualitative Y représentée par le projecteur $\mathbf{P_Y}$. On montre alors (cf. Saporta 1976)

PROPOSITION 7.6. — *Dans le cas d'une variable quantitative X et d'une variable qualitative Y,*

$$\|\mathbf{P_x} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - R^2_c(X,Y))$$

où R_c désigne le rapport de corrélation.

Le rapport de corrélation (Cf. T1) est, dans ce cas, l'indice de similarité qui conduit à la construction d'une distance euclidienne entre variables de types différents.

On aboutit ainsi à une certaine généralisation de la notion de similarité entre variables conduisant, quelque soit le type des variables, à des distances euclidiennes. Néanmoins, en pratique, il n'apparaît pas simple de comparer, sur la même échelle entre 0 et 1, des liaisons entre variables de types différents. Les coefficients de corrélations se répartissent plus communément sur toute l'échelle alors que les indices de Tschuprow sont souvent confinés sur des petites valeurs.

4 Recherche d'une configuration de points

Le positionnement multidimensionnel est la recherche d'une configuration de points dans un espace euclidien qui admette $\mathcal{D}$ comme matrice de distances si celle-ci est euclidienne ou, dans le cas contraire, qui en soit la meilleure approximation à un rang q fixé (en général 2) au sens d'une norme sur les matrices. Nous ne nous intéressons dans ce chapitre qu'à la version "métrique" du MDS, une autre approche construite sur les rangs est développée dans la bibliographie.

Ainsi posé, le problème admet une infinité de solutions. En effet, la distance entre deux vecteurs $\mathbf{x}_i$ et $\mathbf{x}_k$ d'une configuration est invariante par toute transformation affine $\mathbf{z}_i = \mathbf{F}\mathbf{x}_i + \mathbf{b}$ dans laquelle $\mathbf{F}$ est une matrice orthogonale quelconque et $\mathbf{b}$ un vecteur de $\mathbb{R}^p$. Une solution n'est donc connue qu'à une rotation et une translation près.

4.1 Propriétés

La solution est décrite dans les théorèmes (Mardia 1979) ci-dessous :

THÉORÈME 7.7. — *Soit $\mathcal{D}$ une matrice de distance et $\mathbf{B} = \mathbf{HAH}$ la matrice centrée en lignes et colonnes associée.*

- *Si $\mathcal{D}$ est la matrice de distance euclidienne d'une configuration $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ alors $\mathbf{B}$ est la matrice de terme général*
$$b_j^k = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}})$$
qui se met sous la forme
$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'.$$
Elle est donc positive et appelée matrice des produits scalaires de la configuration centrée.
- *Réciproquement, si $\mathbf{B}$ est positive de rang p, une configuration de vecteurs admettant $\mathbf{B}$ pour matrice des produits scalaires est obtenue en considérant sa décomposition spectrale $\mathbf{B} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}^{1/2}$.*

Ainsi, dans le cas d'une matrice $\mathcal{D}$ euclidienne supposée de rang q, la solution est obtenue en exécutant les étapes suivantes :

i. construction de la matrice $\mathbf{A}$ de terme général $-1/2{d_j^k}^2$,
ii. calcul de la matrice des produits scalaires par double centrage $\mathbf{B} = \mathbf{HAH}$,
iii. diagonalisation de $\mathbf{B} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}'$;
iv. les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}^{1/2}$.

Dans le cas euclidien, ACP et MDS sont directement connectés.

PROPOSITION 7.8. — *Soit $\mathbf{Y}$ la matrice des données habituelles en ACP. L'ACP de $(\mathbf{Y}, \mathbf{M}, 1/n\mathbf{I})$ fournit les mêmes représentations graphiques que le positionnement calculé à partir de la matrice de distances de terme général $\|\mathbf{y}_i - \mathbf{y}_j\|_{\mathbf{M}}$. Si $\mathbf{C}$ désigne la matrice des composantes principales, alors les coordonnées principales sont $\sqrt{n}\mathbf{C}$.*

Démonstration. — Posons $\mathbf{X} = \mathbf{HY}$. Les composantes principales de l'ACP sont données par

$$\mathbf{C} = \mathbf{XMV} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$$

où $\mathbf{V}$ est la matrice des vecteurs propres de la matrice $1/n\mathbf{X}'\mathbf{XM}$ et $\mathbf{U}$ ceux des vecteurs propres de la matrice $1/n\mathbf{XMX}'$ associés aux mêmes valeurs propres $\boldsymbol{\Lambda}$. De son côté, le MDS conduit

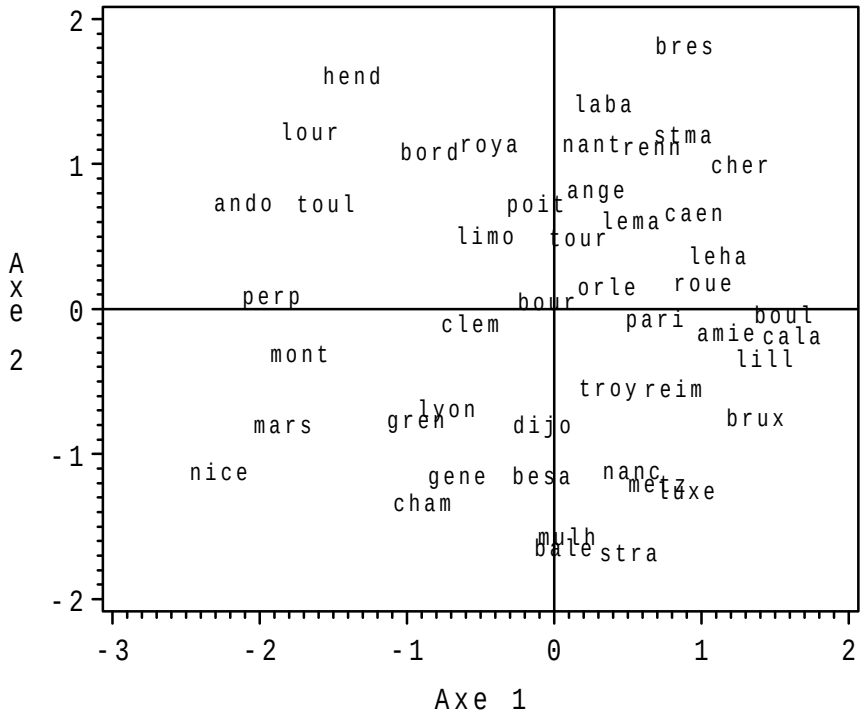

FIG. 7.1 – *Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.*

à considérer la matrice des produits scalaires $\mathbf{HYM}(\mathbf{HY})' = \mathbf{XMX}'$ qui amène aux mêmes vecteurs propres et aux valeurs propres $\mathbf{\Delta} = \sqrt{n}\mathbf{\Lambda}$. ∎

L'intérêt du MDS apparaît évidemment lorsque les observations $\mathbf{Y}$ sont inconnues ou encore si l'on cherche la meilleure représentation euclidienne de distances non-euclidiennes entre les individus ; c'est l'objet du théorème suivant. En ce sens, le MDS "généralise" l'ACP et permet, par exemple, de considérer une distance de type robuste à base de valeurs absolues mais la représentation des variables pose alors quelques soucis car le "biplot" n'est plus linéaire (Gower 19xx).

THÉORÈME 7.9. — *Si $\boldsymbol{\mathcal{D}}$ est une matrice de distance, pas nécessairement euclidienne, $\mathbf{B}$ la matrice de produit scalaire associée, alors, pour une dimension q fixée, la configuration issue du MDS a une matrice de distance $\widehat{\boldsymbol{\mathcal{D}}}$ qui rend $\sum_{j,k=1}^{n}(\{d_j^k\}^2 - \widehat{d_j^k}^2)$ minimum et, c'est équivalent, une matrice de produit scalaire $\widehat{\mathbf{B}}$ qui minimise $\left\|\mathbf{B} - \widehat{\mathbf{B}}\right\|^2$.*

5 Exemple

Cet exemple s'intéresse aux distances kilométriques par route (Source : IGN) entre 47 grandes villes en France et dans les pays limitrophes. Toutes ces valeurs sont rangées dans le triangle inférieur d'une matrice carrée avec des 0 sur la diagonale. La structure du réseau routier fait que cette matrice de distance n'est pas euclidienne, mais, comme le montre le graphique issu d'un positionnement multidimensionnel, l'approximation euclidienne en est très proche.

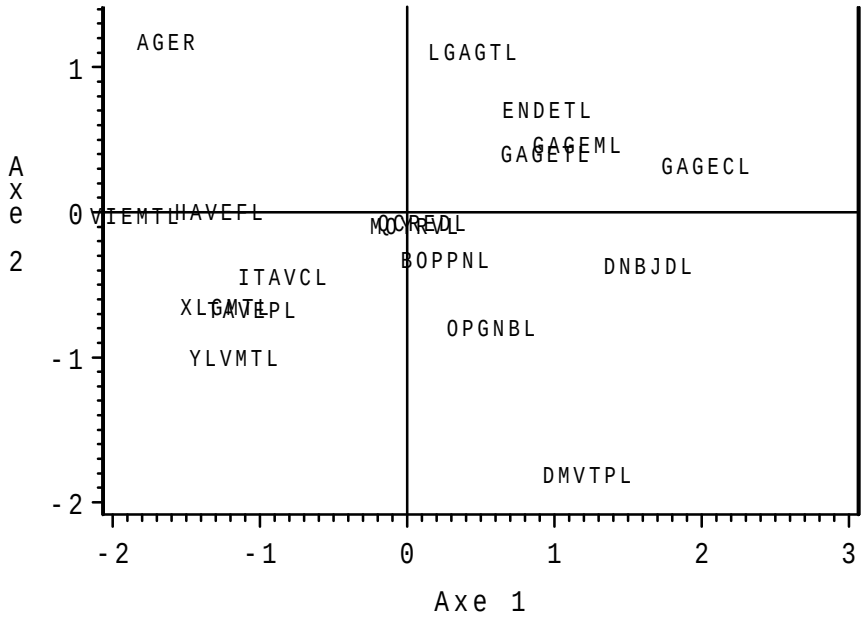

FIG. 7.2 – *Positionnement, conformément aux carrés de leurs corrélations, des variables quantitatives observées sur les données bancaires.*

6 Application au choix de variables

La sélection d'un sous-ensemble de variables pour la mise en œuvre de techniques factorielles (Jolliffe 19xx) n'est pas aussi claire que dans le cadre de la recherche d'un modèle linéaire parcimonieux. Le problème vient souvent de la confusion de deux objectifs :

- supprimer des variables très liées, donc redondantes, et dont la multiplicité vient renforcer artificiellement l'influence de certains phénomènes,
- supprimer des variables afin de simplifier l'interprétation des axes tout en conservant au mieux les représentations graphiques.

Le premier objectif modifie donc les représentations en visant à être plus proche de la "réalité" ou au moins d'une réalité moins triviale tandis que, par principe, le deuxième objectif recherche le sous-ensemble restreint de variables susceptibles d'engendrer le même sous-espace de représentation.

Il n'existe pas de solution miracle néanmoins, les outils présentés dans ce chapitre : indices de similarité entre variable et positionnement multidimensionnel, peuvent aider à ces choix surtout lorsque l'analyse d'un grand nombre de variables nécessite de segmenter l'analyse en sous-groupes. Les algorithmes de classification (hiérarchique ou centres mobiles) appliqués sur les mêmes tableaux de distance apportent un éclairage complémentaire.

D'autres techniques sont également disponibles pour aider à l'interprétation des axes. Elles ont été développées dans le cadre de l'analyse en facteurs communs et spécifiques (factor analysis) mais sont transposables en ACP. L'objectif est la recherche de rotations orthogonales (varimax) ou obliques des axes dans le sous-espace retenu pour la représentation de sorte que ceux-ci soient le plus corrélés avec les variables initiales. Ils n'ont plus les mêmes propriétés optimales d'axes de plus grande dispersion mais, dans le sous-espace qui globalement est de plus grande dispersion, ils peuvent être plus simples à interpréter à partir des variables initiales.

Un algorithme (`varclus` dans SAS) de classification des variables dans le cas quantitatif suit ce même type d'objectifs et fournit des résultats sous une forme identique à la recherche d'une

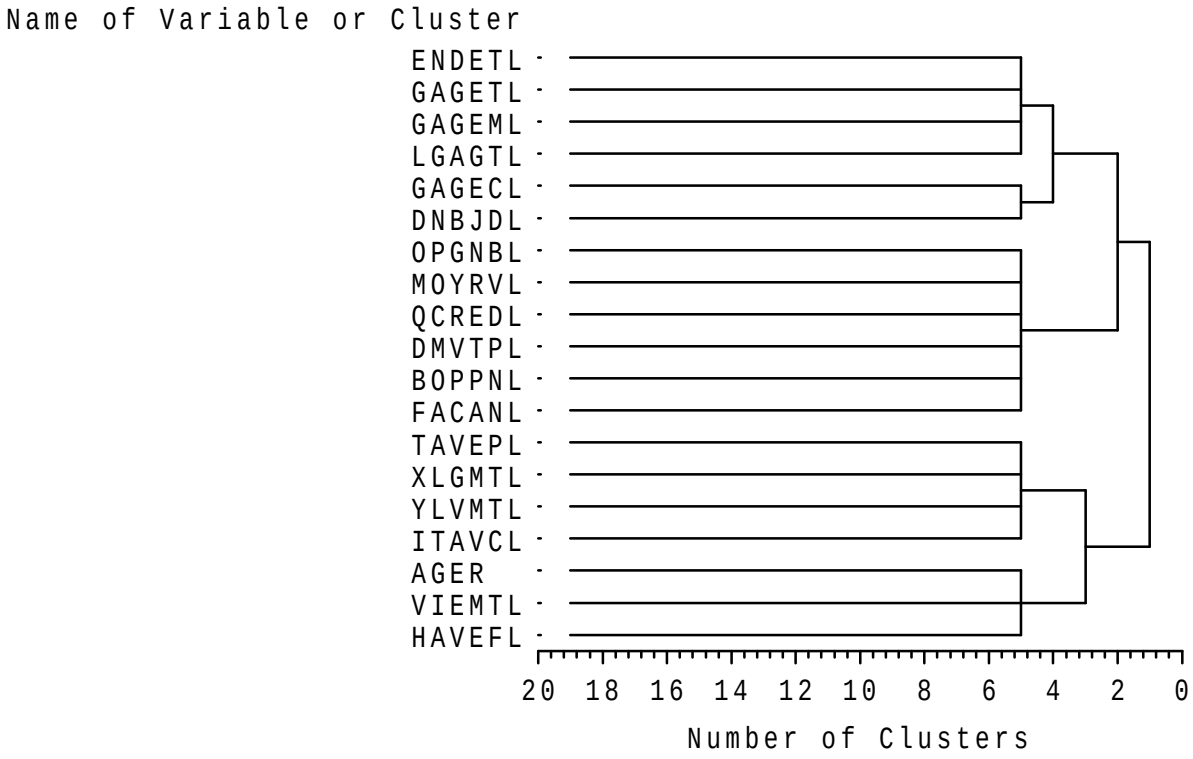

FIG. 7.3 – *Classification (*`varclus`*) des variables quantitatives observées sur les données bancaires.*

rotation oblique. Il procède par classification hiérarchique descendante de l'ensemble des variables et réalise à chaque étape les traitements suivants :

- sélection du sous-groupe de variable dont l'ACP conduit à la plus faible part de variance expliquée par le premier axe ou (en option) la plus forte du 2ème axe,
- rotation des deux premiers axes de l'ACP pour les rapprocher des variables et segmentation des variables en deux groupes par affectation à l'axe avec lequel elles sont le plus corrélées.

L'algorithme s'arrête lorsque la dimension dans chaque groupe est jugée être égale à 1. Par défaut, lorsque dans chaque groupe, une seule valeur propre est plus grande que 1.

Chapitre 8

Classification

1 Introduction

1.1 Les données

Comme dans le cas du chapitre précédent (MDS), les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

i. un tableau de distances (ou dissimilarités, ou mesure de dissemblance), $(n \times n)$ entre les individus pris deux à deux ;

ii. les observations de p variables quantitatives sur ces n individus ;

iii. les observations, toujours sur ces n individus, de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se ramener au tableau des distances deux à deux entre les individus (c'est-à-dire au premier cas). Le choix d'une matrice de produit scalaire permet de prendre en compte simplement un ensemble de variables quantitatives tandis que le troisième cas nécessite plus de développements, objets de la section suivante.

1.2 Objectif

L'objectif d'une méthode de classification déborde le cadre strictement exploratoire. C'est la recherche d'une *typologie* ou *segmentation* c'est-à-dire d'une partition ou répartition des individus en *classes*, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune la plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination ou encore de classement (en anglais *classification*) pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage. Nous sommes dans une situation d'apprentissage *non-supervisé*, ou en anglais de *clustering*[1].

1.3 Les méthodes

Un calcul élémentaire de combinatoire montre que le nombre de partitions possibles d'un ensemble de n éléments croît plus qu'exponentiellement avec n. Ainsi, pour $n = 20$, il est de l'ordre de 10^{13}. Plus précisément, le nombre de partitions en K groupes de n éléments est donné

[1] Faire attention aux faux amis français / anglais : discrimination / classification (supervisée) et classification / clustering (non-supervisée)

par la formule :

$$\frac{1}{K!}\sum k = 0K_k(-1)^k(K-k)^n\binom{k}{K}$$

Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions possibles. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une "bonne" partition qui correspond en général à un optimum local. Même si le besoin de classer des objets est très ancien, seule la généralisation des outils informatiques en a permis l'automatisation dans les années 70. Celeux et col. (1989) décrivent en détail ces algorithmes.

Différents choix sont laissés à l'initiative de l'utilisateur :

- une mesure d'éloignement (dissemblance, dissimilarité ou distance) entre individus ;
- le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la traces d'une matrice de variances-covariances ; soit les variances et covariances interclasses (la trace correspond alors à l'inertie de la partition), soit les variances et covariances intraclasse ;
- la méthode : la classification ascendante hiérarchique ou celle par réallocation dynamique sont les plus utilisées, seules ou combinées,
- le nombre de classes ; c'est un point délicat.

Enfin, différents outils recherchent une interprétation ou des caractérisations des classes obtenues.

Les principes algorithmiques de ces méthodes sont relativement élémentaires.

Classification ascendante hiérarchique ou CAH

Il s'agit de regrouper itérativement les individus, en commençant par le bas (les deux plus proches) et en construisant progressivement un arbre ou *dendrogramme*, regroupant finalement tous les individus en une seule classe, à la racine (cf. figure 3.4 qui reprend les données élémentaires du chapitre précédent). Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ou la distance entre deux groupes. Ceci nécessite donc, pour l'utilisateur de cette méthode, de faire un choix supplémentaire : comment définir la distance entre deux groupes connaissant celles de tous les couples d'individus entre ces deux groupes. Différents choix, appelés *saut* en français et *linkage* en anglais, sont détaillés plus loin. Le nombre de classes est déterminé *a posteriori*, à la vue du dendrogramme ou d'un graphique représentant la décroissance de la hauteur de chaque saut, ou écart de distance, opéré à chaque regroupement.

Réallocation dynamique

Dans ce cas, le nombre de classes, k, est fixé *a priori*. Ayant initialisé k centres de classes par tirage aléatoire, tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie (en principe, euclidienne pour cette méthode). Dans une deuxième étape, l'algorithme calcule des barycentres de ces classes qui deviennent les nouveaux centres. Le procédé (affectation de chaque individu à un centre, détermination des centres) est itéré jusqu'à convergence vers un minimum (local) ou un nombre d'itérations maximum fixé.

Classification mixte

La CAH nécessite impérativement la construction d'un tableau de distances $n \times n$ et son stockage en mémoire ; le nombre maximum d'individus traités peut s'en trouver limité. Ce n'est pas le cas dans l'algorithme de réallocation, d'où l'intérêt possible d'une approche mixte pour, à la fois, classer de grands volumes de données et sélectionner le nombre de classes par CAH.

2 Mesures d'éloignement

Notons $\Omega = \{i = 1, \ldots, n\}$ l'ensemble des individus. Cette section se propose de définir sur $\Omega \times \Omega$ différentes mesures d'éloignement entre deux individus. Les hypothèses et propriétés étant de plus en plus fortes.

2.1 Indice de ressemblance, ou similarité

C'est une mesure de proximité définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ et vérifiant :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie ;} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance d'un individu avec lui-même ;} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \text{la ressemblance est majorée par } S. \end{aligned}$$

Un indice de ressemblance normé s^* est facilement défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega \, ;$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$.

2.2 Indice de dissemblance, ou dissimilarité

Une dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie ;} \\ d(i,i) &= 0, \forall i \in \Omega \ : \ \text{nullité de la dissemblance d'un individu avec lui-même.} \end{aligned}$$

Les notions de similarité et dissimilarité se correspondent de façon élémentaire. Si s est un indice de ressemblance, alors

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

est un indice de dissemblance. De façon réciproque, si d est un indice de dissemblance avec $D = \sup_{(i,j) \in \Omega \times \Omega} d(i,j)$, alors $s(i,j) = D - d(i,j)$ est un indice de ressemblance. Comme s^*, un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$. Du fait de cette correspondance immédiate, seule la notion de dissemblance, ou dissimilarité, normée est considérée par la suite.

2.3 Indice de distance

Un indice de distance est, par définition, un indice de dissemblance qui vérifie de plus la propriété :

$$d(i,j) = 0 \Longrightarrow i = j.$$

Cette propriété évite des incohérences pouvant apparaître entre dissemblances, par exemple :

$$\exists \, k \in \Omega : d(i,k) \neq d(j,k), \quad \text{avec pourtant } i \neq j \text{ et } d(i,j) = 0.$$

2.4 Distance

Une distance sur Ω est, par définition, un indice de distance vérifiant en plus la propriété d'*inégalité triangulaire*. Autrement dit, une distance d est une application de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} d(i,j) &= d(j,i), \ \forall (i,j) \in \Omega \times \Omega \, ; \\ d(i,i) &= 0 \Longleftrightarrow i = j \, ; \\ d(i,j) &\leq d(i,k) + d(j,k), \ \forall (i,j,k) \in \Omega^3. \end{aligned}$$

Si Ω est fini, la distance peut être normée.

2.5 Distance euclidienne

Dans le cas où Ω est un espace vectoriel muni d'un produit scalaire, donc d'une norme, la distance définie à partir de cette norme est appelée distance euclidienne :

$$d(i,j) = < i-j, i-j >^{1/2} = \|i-j\|.$$

La condition pour qu'une matrice donnée de distances entre éléments d'un espace vectoriel soit issue d'une distance euclidienne est explicitée dans le chapitre précédent. Toute distance n'est pas nécessairement euclidienne ; voir, par exemple, celle construite sur la valeur absolue.

2.6 Utilisation pratique

Concrètement, il peut arriver que les données à traiter soient directement sous la forme d'une matrice d'un indice de ressemblance ou de dissemblance. Il est alors facile de la transformer en une matrice de dissemblances normées avant d'aborder une classification.

Nous précisons ci-dessous les autres cas.

Données quantitatives

Lorsque les p variables sont toutes quantitatives, il est nécessaire de définir une matrice $\mathbf{M}$ de produit scalaire sur l'espace $\mathbb{R}^P$. Le choix $\mathbf{M} = \mathbf{I}_p$, matrice identité, est un choix élémentaire et courant ; mais il est vivement conseillé de *réduire* les variables de variances hétérogènes, comme en ACP, ce qui revient à considérer, comme matrice de produit scalaire, la matrice diagonale composée des inverses des écarts-types :

$$\mathbf{M} = \Sigma^{-1} = \text{diag} \, (\frac{1}{\sigma_1} \cdots \frac{1}{\sigma_p}).$$

La métrique dite de Mahalanobis (inverse de la matrice des variances-covariances) peut aussi être utilisée pour atténuer la structure de corrélation.

Données qualitatives

Dans le cas très particulier où toutes les variables sont binaires (présence, absence de caractéristiques), de nombreux indices de ressemblances ont été proposés dans la littérature. Ils sont basés sur les quantités suivantes définis pour deux individus i et j distincts :

- a_{ij} = nombre de caractères communs à i et j sur les p considérés,
- b_{ij} = nombre de caractères possédés par i mais pas par j,
- c_{ij} = nombre de caractères possédés par j mais pas par i,

- $c_{ij} =$ nombre de caractères que ne possèdent ni i ni j.
- bien sûr, $a_{ij} + b_{ij} + b_{ij} + d_{ij} = p$.

Les indices de ressemblance les plus courants sont :

$$\frac{a_{ij} + d_{ij}}{p}(concordance), \frac{a_{ij}}{a_{ij} + b_{ij} + b_{ij}}(Jaccard), \frac{2a_{ij}}{2a_{ij} + b_{ij} + b_{ij}}(Dice).$$

Puis, il est facile de construire un indice de dissemblance.

Dans le cas plus général de p variables qualitatives, la distance la plus utilisée est celle, euclidienne, dite du χ^2 entre profils-lignes du tableau disjonctif complet (cf. chapitre 6 AFCM). La distance entre deux individus i et k est alors définie par :

$$d^2_{\chi^2} = \frac{n}{p} \sum_{j=1} p \sum_{\ell=1}^{m_j} \delta_{ik}^{j\ell} \frac{1}{n_\ell^j}.$$

où m_j est le nombre de modalités de la variable qualitative Y^j, n_ℓ^j est l'effectif de la ℓième modalité de Y^j et $\delta_{ik}^{j\ell}$ vaut 1 si les individus i et k présentent une discordance pour la ℓième modalité de la variables Y^j et 0 sinon. L'importance donnée à une discordance est d'autant plus importante que les modalités considérées sont rares. Le coefficient n/p peut être omis.

Mélange quantitatif, qualitatif

Différentes stratégies sont envisageables dépendant de l'importance relative des nombres de variables qualitatives et quantitatives.

Rendre tout qualitatif . Les variables quantitatives sont rendues qualitatives par découpage en classes. Les classes d'une même variable sont généralement recherchées d'effectifs sensiblement égaux : bornes des classes égales à des quantiles. La métrique à utiliser est alors celle du χ^2 décrite ci-dessus.

Rendre tout quantitatif à l'aide d'une AFCM. Une AFCM est calculée sur les seules variables qualitatives ou sur l'ensemble des variables après découpage en classes des variables quantitatives. L'AFCM calculée par AFC du tableau disjonctif complet produit des *scores* (cf. chapitre 6) qui sont les composantes principales de l'ACP des profils-lignes. Dans le cas d'une AFCM partielle des seules variables qualitatives, les variables quantitatives restantes doivent être nécessairement réduites. Ces scores sont ensuite utilisés commes coordonnées quantitatives des individus en vue d'une classification.

2.7 En résumé

Une fois ces préliminaires accomplis, nous nous retrouvons donc avec

- soit un tableau de mesures quantitatives $n \times p$ associé à une matrice de produit scalaire $p \times p$ (en général $\mathbf{I}_p$) définissant une métrique euclidienne,
- soit directement un tableau $n \times n$ de dissemblances ou distances entre individus.

Attention, si n est grand, la deuxième solution peut se heurter rapidement à des problèmes de stockage en mémoire pour l'exécution des algorithmes.

3 Classification ascendante hiérarchique

3.1 Principe

L'initialisation de cet algorithme consiste, s'il n'est déjà donné, à calculer un tableau de distances (ou de dissemblances) entre les individus à classer. L'algorithme démarre alors de la partition triviale des n singletons (chaque individu constitue une classe) et cherche, à chaque étape, à constituer des classes par agrégation des deux éléments les plus proches de la partition de l'étape précédente. L'algorithme s'arrête avec l'obtention d'une seule classe. Les regroupements successifs sont représentés sous la forme d'un arbre binaire ou *dendrogramme*.

3.2 Dissemblance ou distance entre deux classes

À chaque étape de l'algorithme, il est nécessaire de mettre à jour le tableau des distances (ou des dissemblances). Après chaque regroupement, de deux individus ou de deux classes ou d'un individu à une classe, les distances entre ce nouvel objet et les autres sont calculées et viennent remplacer, dans la matrice, les dissemblances des objets qui viennent d'être agrégés. Différentes approches sont possibles à ce niveau correspondant à différentes CAH.

Notons A et B deux groupes ou éléments d'une partition donnée, w_A et w_B leurs pondérations, et $d_{i,j}$ la distance entre deux individus quelconques i et j.

Le problème est de définir $d(A, B)$ la distance entre deux éléments d'une partition de Ω.

Cas d'une dissemblance

Les stratégies ci-dessous s'accomodent d'un simple indice de dissemblance défini entre les individus. Elles s'appliquent également à des indices plus structurés (distance) mais n'en utilisent pas toutes les propriétés.

$$
\begin{aligned}
d(A,B) &= \min_{i\in A, j\in B}(d_{ij}) \quad \text{(saut minimum, } \textit{single linkage}\text{)},\\
d(A,B) &= \sup_{i\in A, j\in B}(d_{ij}) \quad \text{(saut maximum ou diamètre, } \textit{complete linkage}\text{)},\\
d(A,B) &= \frac{1}{\text{card}(A)\text{card}(B)}\sum_{i\in A, j\in B} d_{ij} \quad \text{(saut moyen, } \textit{group average linkage}\text{)}.
\end{aligned}
$$

Cas d'une distance euclidienne

Les stratégies suivantes nécessitent la connaissance de représentations euclidiennes des individus : matrice $n \times p$ des individus afin, au minimum, de pouvoir définir les barycentres notés g_A et g_B des classes.

$$
\begin{aligned}
d(A,B) &= d(g_A, g_B) \quad \text{(distance des barycentres, centroïd)},\\
d(A,B) &= \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}.
\end{aligned}
$$

Important

Le saut de Ward joue un rôle particulier et est la stratégie la plus courante, c'est même l'option par défaut (SAS), dans le cas d'une distance euclidienne entre individus. En effet, ce critère induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse.

3.3 Algorithme

ALGORITHME 8.1 *:* **classification ascendante hiérarchique**

- `Initialisation` *Les classes initiales sont les singletons. Calculer la matrice de leurs distances deux à deux.*
- `Itérer` *les deux étapes suivantes jusqu'à l'agrégation en une seule classe.*
  - *i. regrouper les deux classes les plus proches au sens de la "distance" entre groupes choisie,*
  - *ii. mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa "distance" avec chacune des autres classes.*

3.4 Graphes

Ce chapitre est illustré par l'étude de données décrivant le traffic sur 50 lignes de chemin de fer pour les mois de mars, juillet août et octobre. On s'intéresse plus partivulièrement aux profils de ces traffics. En effet, les données étant des effectifs de voyageurs, pour éviter une classification triviale basée sur le traffic absolu de chaque ligne, on utilise la métrique du $\chi 2$ entre profils lignes des données considérées comme une table de contingence. La classification est donc construite à partir de la matrice de ces distances entre lignes. À l'issue de l'exécution, la classification ascendante hiérarchique fournit deux graphiques :

- un graphique aide au choix du nombre de classes (cf. figure3.4). Il représente, à rebours, la décroissance en fonction du nombre de classes de la distance entre les agrégations de classes. Dans le cas du saut de Ward, il s'agit des écarts observés par le rapport de la variance inter sur la variance totale (R^2 partiel). La présence d'une rupture importante dans cette décroissante aide au choix du nombre de classes comme dans le cas du choix de dimension en ACP avec l'éboulis des valeurs propres. Dans ce cas, il faut lire le graphe de droite à gauche et s'arrêter avanr le premier saut juger significatif. L'indice de Ward est le plus généralement utilisé, cela revient à couper l'arbre avant une perte, juger trop importante, de la variance inter classe.
- le *dendrogramme* (cf. figure3.4) est une représentation graphique, sous forme d'arbre binaire, des agrégations successives jusqu'à la réunion en une seule classe de tous les individus. La hauteur d'une branche est proportionnel à l'indice de dissemblance ou distance entre les deux objets regroupés. Dans le cas du saut de Ward, c'est la perte de variance inter-classe.

Une fois un nombre de classes sélectionné (ici 4) à l'aide du premier graphique, une coupure de l'arbre (deuxième graphique) fournit, dans chaque sous-arbre, la répartition des individus en classes. Ces classes sont ensuite représentées dans les axes d'une analyse factorielle, en général une ACP, mais qui peut être un MDS lorsque les données initiales sont un tableau de distances ou encore, dans le cas présent, une AFCM. Cette représentation (fig. 3.4) est indispensable pour se faire une bonne idée intuitive de la qualilté de séparation des classes. La même démarche appliquée aux données constituées de la matrice des distances kilométriques entre villes conduit à une classification en 5 classes représentées dans les coordonnées du MDS (figure 3.4).

Il est à noter que ces exemples sont relativement simples et bien structurés. Dans ce cas, modifier le critère de saut ne change pas grand chose. Mais, *attention,* il est facile de vérifier expérimentalement qu'une classification ascendante est un objet très sensible. En effet, il suffit de modifier une distance dans le tableau, par exemple de réduire sensiblement la distance de Grenoble à Brest, pour que la classification (nombre de classes, organisation) devienne très sensible au choix du critère de saut. En revanche, la structure des données fait que la représentation factorielle de l'ACP du tableau de distance (MDS) soit plus robuste à ce type d"'erreur de mesure".

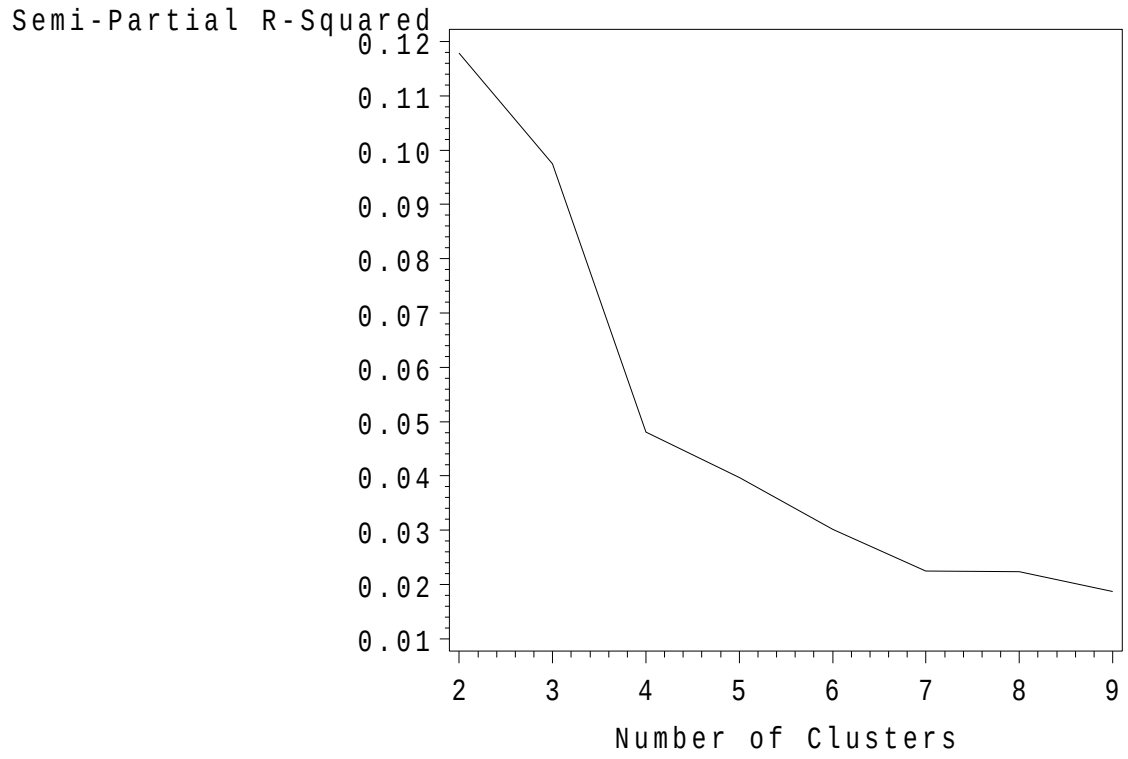

FIG. 8.1 – *Traffic : Décroissance de la variance inter classes à chaque regroupemet dans le cas du saut de Ward.*

4 Agrégation autour de centres mobiles

4.1 Principes

Différents types d'algorithmes ont été définis autour du même principe de *réallocation dynamique* des individus à des centres de classes, eux-mêmes recalculés à chaque itération. Ces algorithmes requièrent une représentation vectorielle des individus dans $\mathbb{R}^p$ muni d'une métrique généralement euclidienne. Une adaptation de cet algorithme, PAM (pour *Partitioning — clustering — of the data into k clusters Around Medoïds* ; Kaufman & Rousseeuw, 1990), en est une version robuste, également adaptée à une matrice de dissimilarités. Ce dernier algorithme est en revanche limité au niveau du nombre d'observations (200).

Il est important de noter que, contrairement à la méthode hiérarchique précédente, le nombre de classes k doit être déterminé *a priori*.

Ces méthodes sont itératives : après une initialisation des centres consistant, le plus souvent, à tirer aléatoirement k individus, l'algorithme répète deux opérations juqu'à la convergence d'un critère :

i. Chaque individu est affecté à la *classe* dont le centre est le plus proche.

ii. Calcul des k *centres* des classes ainsi constituées.

4.2 Principale méthode

ALGORITHME 8.2 *: de* **Forgy** *(1965)*

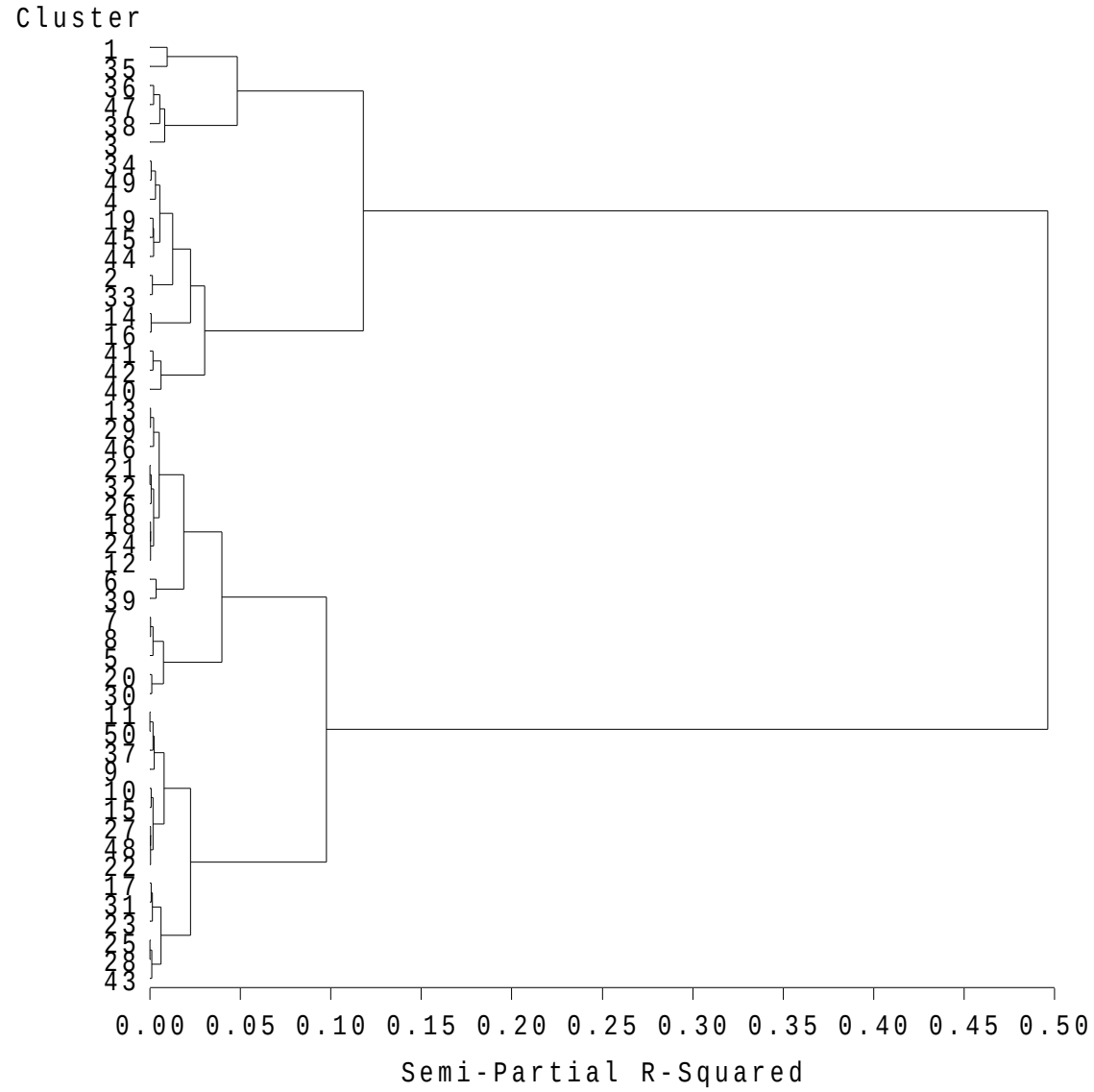

FIG. 8.2 – *Traffic : Exemple d'un dendrogramme issu de la classification de données fictives par CAH et saut de Ward.*

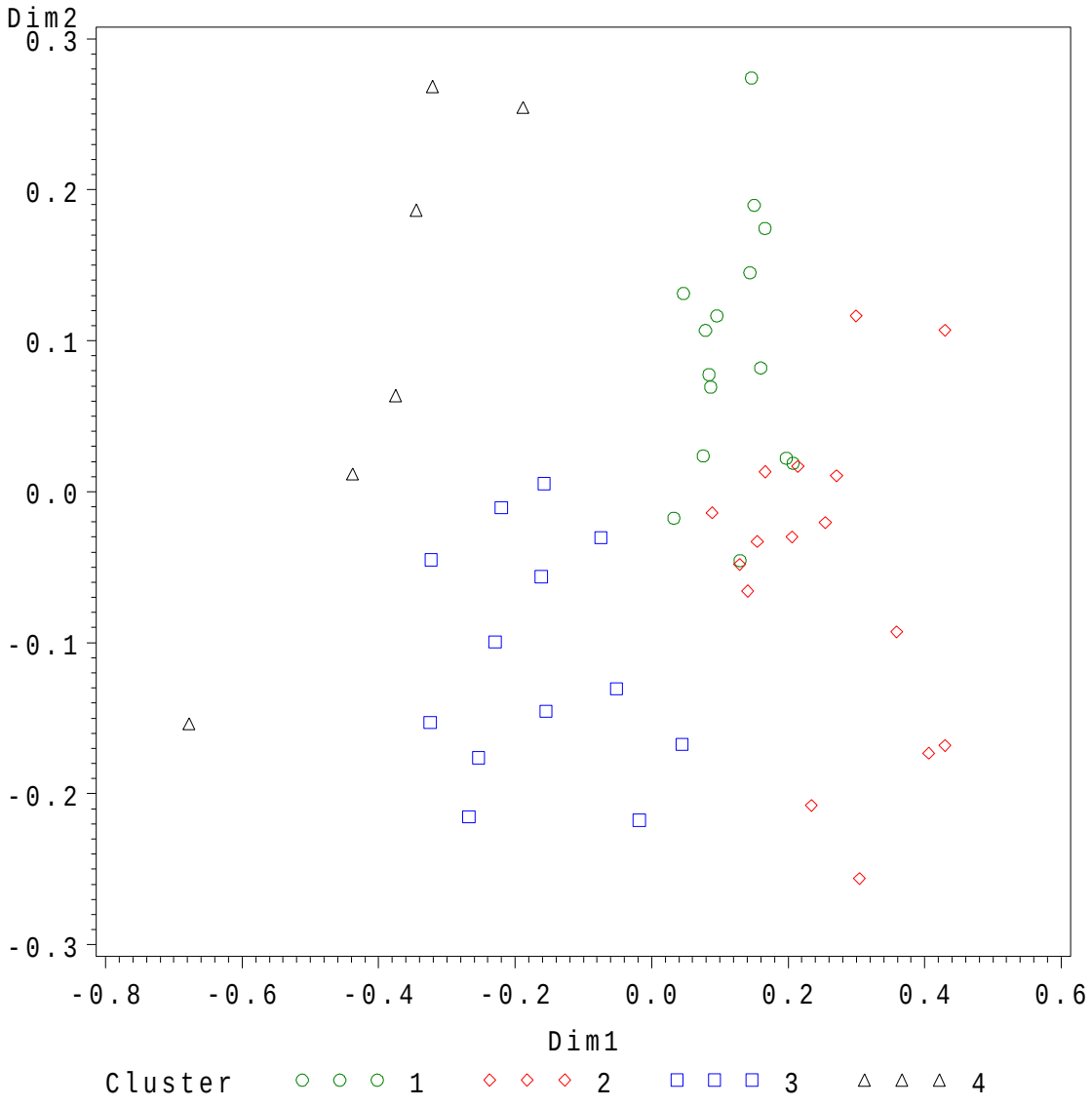

FIG. 8.3 – *Traffic : Représentation des classes dans les coordonnées de l'AFCM.*

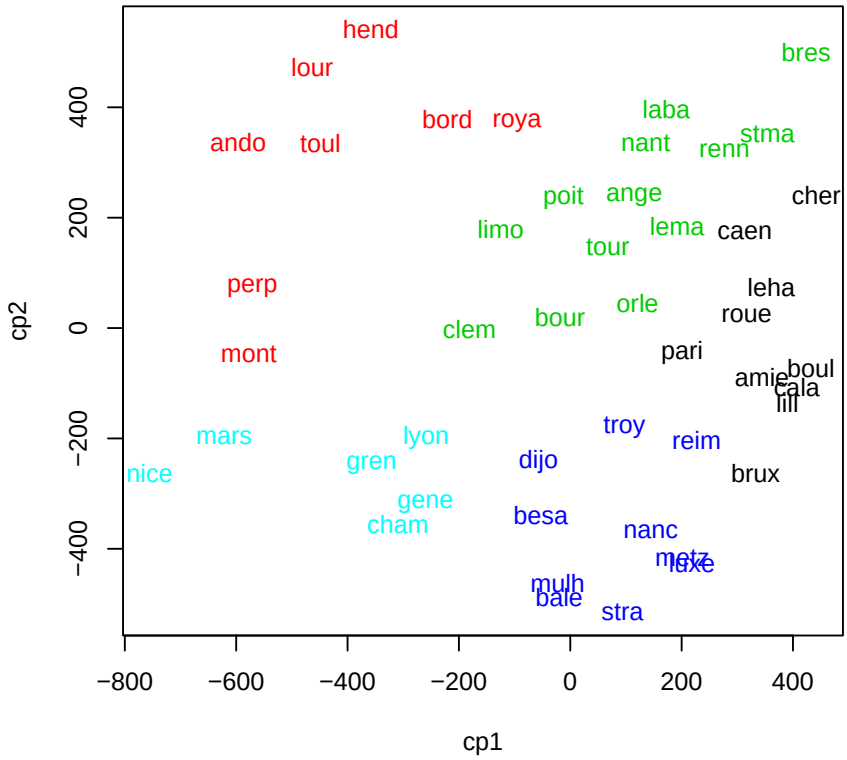

FIG. 8.4 – *Villes : Représentation des classes (couleurs) obtenues par CAH dans les coordonnées du MDS.*

- `Initialisation` *Tirer au hasard ou sélectionner, pour des raisons extérieures à la méthode, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux.*
- `Itérer` *les deux étapes suivantes jusqu'à ce que le critère de variance interclasse ne croisse plus de manière significative, c'est-à-dire jusqu'à la stabilisation des classes.*
  i. *Allouer chaque individu au centre, c'est-à-dire à une classe, le plus proche au sens de la métrique euclidienne choisie ; on obtient ainsi, à chaque étape, une classification en k classes ou moins si finalement une des classes devient vide.*
  ii. *Calculer le centre de gravité de chaque classe : il devient le nouveau noyau ; si une classe s'est vidée, on peut éventuellement retirer aléatoirement un noyau complémentaire.*

4.3 Propriétés

Convergence Le critère (la variance interclasse) est majoré par la variance totale. Il est simple de montrer qu'il ne peut que croître à chaque étape de l'algorithme, ce qui en assure la convergence. Il est équivalent de maximiser la variance interclasse ou de minimiser la variance intraclasse. Cette dernière est alors décroissante et minorée par 0. Concrètement, une dizaine d'itérations suffit généralement pour atteindre la convergence.

Optimum local La solution obtenue est un optimum local, c'est-à-dire que la répartition en classes dépend du choix initial des noyaux. Plusieurs exécution de l'algorithme permettent de s'assurer de la présence de *formes fortes* c'est-à-dire de classes ou portions de classes présentes de manière stable dans la majorité des partitions obtenues.

4.4 Variantes

$k$means

Il s'agit d'une modification de l'algorithme précédent proposée par Mac Queen (1967). Les noyaux des classes, ici les barycentres des classes concernées, sont recalculés à chaque allocation d'un individu à une classe. L'algorithme est ainsi plus efficace, mais il dépend de l'odre des individus dans le fichier.

Nuées dynamiques

La variante proposée par Diday (1971) consiste à remplacé chaque centre de classe par un noyau constitué d'éléments représentatifs de cette classe. Cele permet de corriger l'influence d'éventuelles valeurs extrèmes sur le calcul du barycentre.

Partitionning around medoïds

Cet algorithme, proposé par Kaufman & Rousseeuw (1990), permet de classifier des données de façon plus robuste c'est-à-dire moins sensible à des valeurs atypiques. Il permet également de traiter des matrices de dissimilarités. Les résultats sont fournis dans la figure 4.4, pour lequel le nombre de classe est fixé *a priori* à 5 comme le suggère la CAH, mais pour lesquels les classes obtenues sont sensiblement différentes.

5 Combinaison

Chaque méthode précédente peut être plus ou moins adaptée à la situation rencontrée. La classification hiérarchique, qui construit nécessairement la matrice des distances, n'accepte qu'un

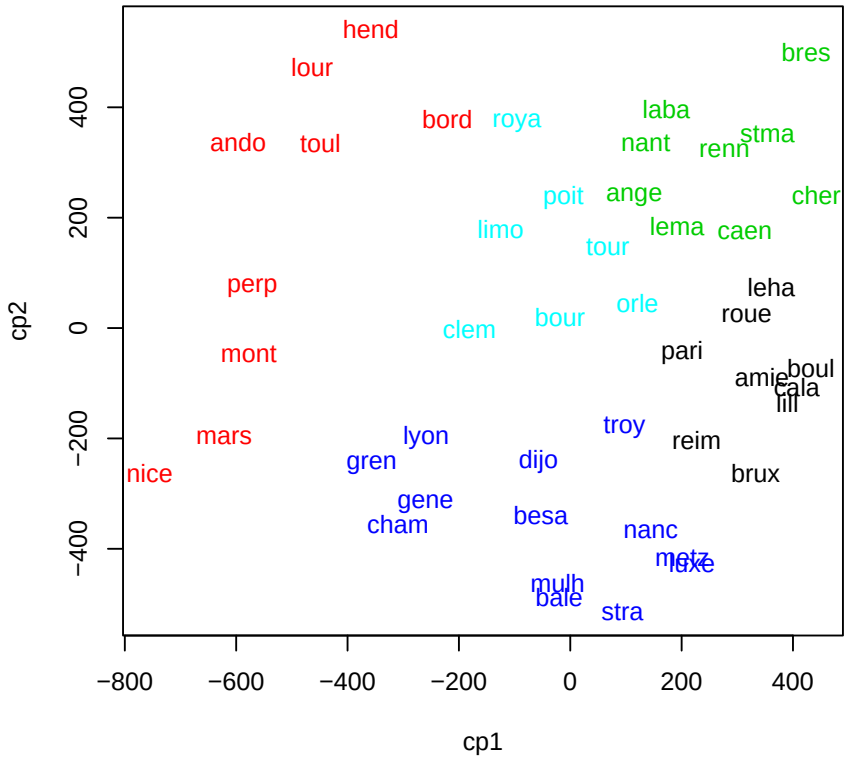

FIG. 8.5 – *Villes : Représentation des classes (couleurs) obtenues par PAM dans les coordonnées du MDS.*

nombre limité d'individus ; de son côté, la réallocation dynamique nécessite de fixer *a priori* le nombre de classes. La stratégie suivante, adaptée aux grands ensembles de données, permet de contourner ces difficultés.

i. Exécuter une méthode de réallocation dynamique en demandant un grand nombre de classes, de l'ordre de 10% de n.

ii. Sur les barycentres des classes précédentes, exécuter une classification hiérarchique puis déterminer un nombre "optimal" k de classes.

iii. Exécuter une méthode de réallocation dynamique sur tout l'ensemble en fixant à k le nombre de classes. Pour initialiser l'algorithme, il est habituel de choisir pour noyaux les barycentres (calculés en pondérant par les effectifs de classes) des classes de l'étape précédente.

6 Interprétation

Dans tous les cas, le résultat fourni est une variable qualitative T dont les modalités précisent la classe retenue pour chaque individu. Il est alors important de caractériser chaque classe à partir des variables initiales afin d'en synthétiser les propriétés.

Les outils élémentaires de statistiques descriptive bidimensionnelle sont, dans un premier temps adaptés à cet objectif. Statistiques (moyenne, écart-type...) par classe, diagrammes boîtes, rapports de corrélations, pour les variables quantitatives, profils, tests d'indépendance, pour les variables qualitatives, permettent de déterminer les variables les plus liées à la classification obtenue.

D'autres méthodes sont ensuite traditionnellement enchaînées : ACP, MDS avec représentation des classes et de leur enveloppe convexe, pour apprécier la qualité de la classification, AFD et/ou arbre de classification afin d'aider à l'interprétation de chacune des classes de la typologie par les variables initiales, AFCM dans le cas de variables qualitatives.

Chapitre 9

Exploration de données fonctionnelles

1 Introduction

Ce chapitre est une introduction à l'étude exploratoire d'ensembles de données dans lesquels les n individus ou observations ne sont plus considérées comme de simples vecteurs de $\mathbb{R}^p$ mais sont des courbes ou plus généralement des fonctions. Ces fonctions dépendent d'un indice, traditionnellement le temps t, évoluant dans un intervalle que l'on supposera être, sans perte de généralité, un intervalle $T = [a, b]$ de $\mathbb{R}$. En pratique, ces fonctions sont observées en des instants de discrétisation qui peuvent être équirépartis ou non, identiques ou non, pour chaque courbe. La figure 9.1 donne un exemple type représentant des cumuls mensuels de précipitations.

Depuis une vingtaine d'années, ce type de données se rencontre de plus en plus fréquemment avec l'automatisation et l'informatisation des procédures de mesure : télémétrie, spectrographie... En conséquence, la littérature consacrée à l'étude de données *fonctionnelles* s'est considérablement développée. Ce chapitre ne s'intéresse qu'à un objectif d'exploration ou de réduction de la dimension. L'aspect modélisation ou apprentissage est développé dans le deuxième volet[1].

Historiquement, les premiers travaux peuvent être attribués à des météorologues ou encore des chimistes qui furent les premiers à être confrontés à ce type de données ou encore à des techniques de traitement du signal associant Analyse en Composantes Principales (ACP) et décomposition de Karhunen-Loeve. En France, Deville (1974) introduisit une ACP de courbe ou *analyse harmonique* et Dauxois et Pousse (1976) proposèrent un cadre synthétique généralisant l'analyse des données multidimensionnelles aux variables aléatoires hilbertiennes qui constituent le cadre théorique à l'exploration statistique de courbes. Différents développements impliquant des outils d'interpolation ou de lissage (splines) ont permis d'adapter finement l'ACP à ce contexte (Besse et Ramsay, 1986 ; Besse et col. 1997) tandis que Ramsay et Silverman (1997) fournissent une bibliographie détaillée.

L'adaptation de méthodes statistiques à des données fonctionnelles requiert un arsenal mathématique pouvant paraître sophistiqué voire rebutant. Certains de ces outils théoriques ne sont indispensables que pour aborder les aspects asymptotiques[2]. Une introduction élémentaire est proposée en annexe B. Mais, en pratique, les données sont de toute façon discrétisées et les calculs réalisés matriciellement dans des espaces de dimension finie. D'autres outils, essentiellement issus de l'analyse numérique, sont alors indispensables pour rendre leur caractère fonctionnel aux observations

[1] Data mining 2. Modélisation statistique et apprentissage.

[2] Il est en effet important de pouvoir exhiber les analyses limites dans des espaces fonctionnels de dimension infinie lorsque le pas de discrétisation décroît et que la taille de l'échantillon croît indéfiniment. C'est le moyen de s'assurer de la stabilité des solutions proposées.

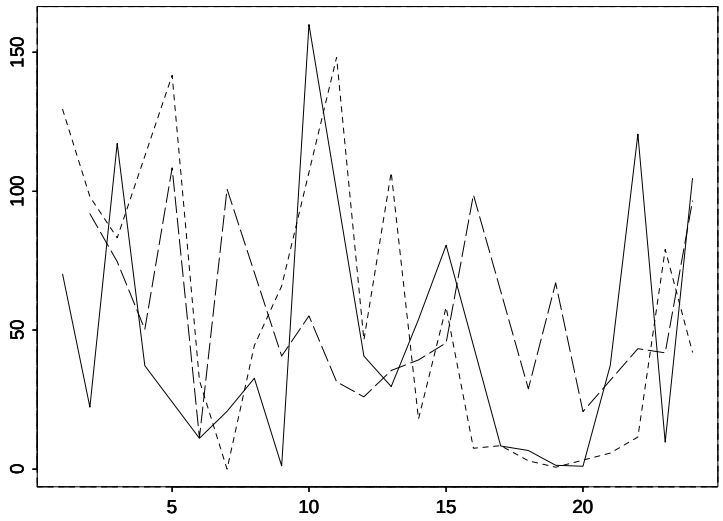

FIG. 9.1 – Trois exemples de courbes décrivant la pluviométrie mensuelle durant 2 ans.

discrétisées. Il s'agit principalement de techniques d'interpolation et de lissage (splines, noyaux) ou de décomposition (Fourier, ondelettes) sur différents types de bases. La question principale qui se pose alors est la suivante.

> Dans quelles situations une approche fonctionnelle peut s'avérer plus efficace que celle vectorielle classique dans le cadre euclidien de $\mathbb{R}^p$?

Deux situations relativement fréquentes ont été identifiées comme répondant à cette question :

- lorsque les *courbes* ou fonctions observées sont très *régulières*, les variables issues de la discrétisation sont très corrélées deux à deux. Ceci a pour effet de masquer, par un effet "taille" trivial, l'essentiel des phénomènes d'intérêt. Dans le cadre de l'ACP, Besse et Ramsay (1986) proposent des pistes développées ensuite par Ramsay (1996, 2000). L'objectif est de décomposer l'espace des courbes ou individus en deux sous-espaces orthogonaux. Le premier contient l'effet trivial exprimé comme la solution d'une équation différentielle et correspond donc au noyau de l'opérateur différentiel correspondant. Le deuxième, l'espace orthogonal à ce noyau, s'attache à représenter la partie restante du phénomène.
- Lorsque les données sont l'observation bruitée d'un phénomène que l'on peut supposer relativement régulier, il est important de faire intervenir une action de lissage ou débruitage. Le problème soulevé est alors celui d'une coordination optimale entre une technique de débruitage et celle multidimensionnelle considérée.

Évitant les développements trop théoriques, nous insistons dans ce chapitre sur la mise en œuvre matricielle d'une ACP de courbes supposées régulières mais observées bruitées. Dans ce cas, l'ACP rejoint l'objectif de la régression non paramétrique en proposant une estimation *simultanée* de plusieurs courbes. Elle doit incorporer des outils d'analyse numérique adaptés pour définir des approximations de ces courbes. Les fonctions splines d'interpolation et de lissage remplissent bien ce rôle mais d'autres techniques, comme la décomposition en ondelettes, auraient pu être utilisées notamment si les fonctions que l'on cherche à estimer présentent des singularités.

Pour simplifier la présentation de la méthodologie proposée, nous supposons dans ce chapitre que toutes les courbes sont observées selon le même plan de discrétisation c'est-à-dire aux

mêmes instants. Dans le cas contraire, une adaptation a été proposée par Besse et coll. (1997) afin, également, de pouvoir prendre en compte des données manquantes. Celle-ci repose sur l'utilisation d'une approximation par splines hybrides associant B-splines et splines de lissage. Tous les programmes[3] utilisés dans les exemples sont écrits en Splus (1997).

2 ACP de courbes bruitées

Nous nous intéressons dans cette section à la description et à l'estimation des réalisations de trajectoires z_i d'un processus Z ou, c'est équivalent, d'une variable aléatoire prenant ses valeurs dans un espace hilbertien. Nous considérons que la variable aléatoire X constitue l'observation bruitée des trajectoires, supposées régulières, de la variable aléatoire Z. La figure 9.1 donne un exemple illustratif de telles données.

Chacune des n réalisations ou trajectoires z_i, est donc supposée observée pour un nombre p d'instants de discrétisation $t_1, \ldots, t_p$ de l'intervalle T, les mêmes pour chaque trajectoire. Cette mesure introduit des erreurs aléatoires indépendantes et identiquement distribuées de variance σ^2. La situation correspond donc à n répétitions supposées indépendantes d'un modèle de régression non-paramétrique (B.4) :

$$\begin{array}{l} x_j = z(t_j) + \varepsilon_j \; ; \quad E(\varepsilon_j) = 0, \; E(\varepsilon_j \varepsilon_k) = \sigma^2 \delta k, \; j,k = 1, ..., p \\ \qquad\qquad\qquad\qquad a \leq t_1 < t_2 < ... < t_p \leq b \end{array}$$

auquel il faut ajouter l'hypothèse d'indépendance entre les différentes réalisations de Z et le bruit : $\mathbb{E}(\boldsymbol{\varepsilon}_i \mathbf{z}'_{i'}) = 0$.

À ce niveau, il serait possible de considérer l'estimation des n trajectoires de Z comme n problèmes classiques d'estimation non paramétrique de fonctions de régression. Néanmoins, intuitivement et c'est vérifié par des simulations (Besse et coll. 1997), il est important de tenir compte du fait qu'il s'agit de l'estimation simultanée de n réalisations d'un même processus et donc de tenir compte de la structure de covariance qu'il est possible d'estimer. C'est réalisé en introduisant une contrainte supplémentaire issue de l'hypothèse que la variable Z évolue dans un sous-ensemble de dimension finie de l'espace de Sobolev $W^2(T)$ (fonctions continues admettant une dérivée dans L^2). Ceci revient encore à écrire que ses trajectoires s'expriment comme combinaisons linéaires d'un nombre réduit q de composantes. Ces composantes étant par ailleurs régulières du fait de la première hypothèse.

2.1 Modèle et estimation

Les observations de chacune des trajectoires sont rangées dans des vecteurs $\mathbf{x}_i$ de $\mathbb{R}^p$ et A_q désigne un sous-espace affine de $\mathbb{R}^p$ de dimension $q < p$. La situation impliquée par l'estimation simultanée de n régressions non paramétriques sous une double contrainte de régularité et de dimension se résume par le modèle suivant :

$$\begin{array}{l} \mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i; \\ i = 1, \ldots, n \end{array} \quad \text{avec} \quad \left\{ \begin{array}{l} \mathbb{E}(\boldsymbol{\varepsilon}_i) = 0 \text{ et } \mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i) = \sigma^2 \mathbf{I}, \\ \sigma \text{ inconnue}, \; (\sigma > 0) \\ \mathbf{x}_i \text{ indépendant de } \boldsymbol{\varepsilon}_{i'}, \; i' = 1, \ldots, n, \\ \mathbf{x}_i \in A_q \text{ p.s. et } \|\mathbf{x}_i\|_m^2 \leq c \text{ p.s.}. \end{array} \right. \tag{9.1}$$

Ce modèle présente donc la particularité d'associer deux types de contraintes, la première, de dimension, conduisant à une définition de l'analyse en composantes principales (cf. chapitre 3),

[3] Ils sont accessibles à partir de l'URL `www.inra.fr/bia/T/cardot/`

la deuxième de régularité, habituelle en statistique fonctionnelle ; $\ell \|x\|_m$ désigne une semi-norme définie par la norme dans $L^2[0,1]$ de la dérivée $m$ième de x. L'estimation par les moindres carrés pondérés amène à résoudre un problème d'optimisation dans lequel la contrainte de régularité a été remplacée par un multiplicateur de Lagrange ℓ dépendant de c.

Avec les notations matricielles où $\mathbf{M}$ désigne la matrice associée à la semi-norme $\ell \|.\|_m$ (cf. annexe B) et en supposant que les observations sont pondérées par les éléments diagonaux w_i de la matrice $\mathbf{D}$, il s'agit de résoudre :

$$\min_{\mathbf{z}_i, A_q} \left\{ \sum_{i=1}^{n} w_i \left(\|\mathbf{z}_i - \mathbf{x}_i\|^2_{\mathbf{I}} + \ell \|\mathbf{z}_i\|^2_{\mathbf{M}} \right) \; ; \mathbf{z} \in A_q \; , \; \dim A_q = q \right\} \tag{9.2}$$

Notons par $\overline{\mathbf{x}} = \sum_{i=1}^{n} w_i \mathbf{x}_i$ la moyenne des coordonnées et par $\overline{\mathbf{X}}$ la matrice des observations centrées c'est-à-dire dans un contexte d'études climatiques, la matrice des anomalies $(\mathbf{x}_i - \overline{\mathbf{x}})$ des observations par rapport à la moyenne annuelle ; $\mathbf{S}$ désigne la matrice de covariance empirique : $\mathbf{S} = \overline{\mathbf{X}}' \mathbf{D} \overline{\mathbf{X}}$.

PROPOSITION 9.1. — *La solution du problème(9.2) est donnée par :*

$$\widehat{\mathbf{z}}_i = \mathbf{A}_\ell^{1/2} \widehat{\mathbf{P}}_q \mathbf{A}_\ell^{1/2} \mathbf{x}_i + \mathbf{A}_\ell \overline{\mathbf{x}} \; , \; i = 1, \ldots, n.$$

La matrice $\widehat{\mathbf{P}}_q = \mathbf{V}_q \mathbf{V}_q'$ est la projection orthogonale sur le sous-espace $\widehat{E}_q$ engendré par les q vecteurs propres de la matrice

$$\mathbf{A}_\ell^{1/2} \mathbf{S} \mathbf{A}_\ell^{1/2}.$$

associés aux q plus grandes valeurs propres.

Les estimations lisses des trajectoires s'obtiennent alors par interpolation spline des valeurs contenues dans le vecteur $\widehat{z}_i$.

Démonstration. — Notons $\mathbf{z}_i$ le vecteur de $\mathbb{R}^p$ contenant les valeurs de z_i et $\overline{\mathbf{z}} = \sum_{i=1}^{n} w_i \mathbf{z}_i$. On définit la matrice centrée $\mathbf{Z}$ $(n \times p)$ dont les vecteurs lignes $(\mathbf{z}_i - \overline{\mathbf{z}}')$ sont contraints à appartenir au sous-espace vectoriel $E_q = A_q - \overline{\mathbf{z}}$. Cette contrainte est équivalente à imposer à la matrice $\mathbf{Z}$ d'être au plus de rang q.

Le critère à minimiser se décompose de la façon suivante :

$$\begin{aligned} \sum_{i=1}^{n} w_i \left(\|\mathbf{z}_i - \mathbf{x}_i\|^2_{\mathbf{I}} + \ell \|\mathbf{z}_i\|^2_{\mathbf{M}} \right) &= \sum_{i=1}^{n} w_i \|\mathbf{x}_i - (\mathbf{z}_i - \overline{\mathbf{z}})\|^2_{\mathbf{I}} + \\ &+ \ell \sum_{i=1}^{n} w_i \|\mathbf{z}_i - \overline{\mathbf{z}}\|^2_{\mathbf{M}} + \|\overline{\mathbf{x}} - \overline{\mathbf{z}}\|^2_{\mathbf{I}} + \ell \|\overline{\mathbf{z}}\|^2_{\mathbf{M}} . \end{aligned}$$

Les deux derniers termes de cette expression conduisent à estimer $\overline{\mathbf{z}}$ par lissage spline de la moyenne empirique :

$$\widehat{\overline{\mathbf{z}}} = \mathbf{A}_\ell \overline{\mathbf{x}} \quad \text{donc} \quad \widehat{A}_q = \widehat{\overline{\mathbf{z}}} + \widehat{E}_q.$$

Les deux premiers termes nous amènent ensuite à résoudre :

$$\min_{\mathbf{Z}(n \times p)} \left\{ \left\| \mathbf{Z} - \overline{\mathbf{X}} \right\|^2_{\mathbf{I},\mathbf{D}} + \ell \|\mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}} \, ; \, \text{rang}(\mathbf{Z}) = q, \; q < p \right\} \tag{9.3}$$

où $\|\mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}} = \text{tr}\mathbf{Z}'\mathbf{D}\mathbf{Z}\mathbf{M}$ désigne la norme euclidienne des matrices $(n \times p)$.

Notons $\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\mathbf{A}_\ell$ la matrice des lignes lissées de $\overline{\mathbf{X}}$, de sorte que

$$\begin{aligned}\left\|\mathbf{Z}-\overline{\mathbf{X}}\right\|^2_{\mathbf{I},\mathbf{D}} + \ell\left\|\mathbf{Z}\right\|^2_{\mathbf{M},\mathbf{D}} &= \mathrm{tr}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}} - 2\mathrm{tr}\overline{\mathbf{X}}'\mathbf{D}\mathbf{Z} + \mathrm{tr}\mathbf{Z}'\mathbf{D}\mathbf{Z}(\mathbf{I}+\ell\mathbf{M}) \\ &= \mathrm{tr}\widetilde{\overline{\mathbf{X}}}'\mathbf{D}\widetilde{\overline{\mathbf{X}}}(\mathbf{I}+\ell\mathbf{M})^2 - 2\mathrm{tr}\widetilde{\overline{\mathbf{X}}}'\mathbf{D}\mathbf{Z}(\mathbf{I}+\ell\mathbf{M}) + \\ &+ \mathrm{tr}\mathbf{Z}'\mathbf{D}\mathbf{Z}(\mathbf{I}+\ell\mathbf{M}) \\ &= \left\|\widetilde{\overline{\mathbf{X}}}-\mathbf{Z}\right\|^2_{(\mathbf{I}+\ell\mathbf{M}),\mathbf{D}} + \ell\left\|\widetilde{\overline{\mathbf{X}}}\right\|^2_{\mathbf{M},\mathbf{D}} + \ell^2\left\|\widetilde{\overline{\mathbf{X}}}\right\|^2_{\mathbf{M^2},\mathbf{D}}.\end{aligned}$$

Seul le premier terme de cette équation dépend de $\mathbf{Z}$. Par conséquent, la solution est la meilleure approximation de rang q de la matrice $\widetilde{\overline{\mathbf{X}}}$. Elle est obtenue par la décomposition en valeurs singulières (DVS) de $\overline{\mathbf{X}}\mathbf{A}_\ell$ relativement aux métriques $\mathbf{A}_\ell^{-1}$ et $\mathbf{D}$:

$$\widehat{\mathbf{Z}_q} = \widetilde{\mathbf{U}}_q\widetilde{\mathbf{L}}_q^{1/2}\widetilde{\mathbf{V}}_q',$$

$$\text{où } \begin{cases} \overline{\mathbf{X}}\mathbf{A}_\ell\overline{\mathbf{X}}'\mathbf{D}\widetilde{\mathbf{U}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{L}} & \text{et} \quad \widetilde{\mathbf{U}}'\mathbf{D}\widetilde{\mathbf{U}} = \mathbf{I}, \\ \mathbf{A}_\ell\overline{\mathbf{X}}'\mathbf{D}\mathbf{X}\widetilde{\mathbf{V}} = \widetilde{\mathbf{V}}\widetilde{\mathbf{L}} & \text{et} \quad \widetilde{\mathbf{V}}'\mathbf{A}_\ell^{-1}\widetilde{\mathbf{V}} = \mathbf{I}. \end{cases}$$

Cette décomposition en valeurs singulières généralisée est aussi déduite de celle de $\overline{\mathbf{X}}\mathbf{A}_\ell^{1/2}$ relativement à $\mathbf{I}$ et $\mathbf{D}$:

$$\overline{\mathbf{X}}\mathbf{A}_\ell^{1/2} = \mathbf{U}\mathbf{L}^{1/2}\mathbf{V}'$$

$$\text{où } \begin{cases} \overline{\mathbf{X}}\mathbf{A}_\ell^{1/2}\mathbf{A}_\ell^{1/2}\overline{\mathbf{X}}'\mathbf{D}\mathbf{U} = \mathbf{U}\mathbf{L} & \text{et} \quad \mathbf{U}'\mathbf{D}\mathbf{U} = \mathbf{I}, \\ \mathbf{A}_\ell^{1/2}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{A}_\ell^{1/2}\mathbf{V} = \mathbf{V}\mathbf{L} & \text{et} \quad \mathbf{V}'\mathbf{V} = \mathbf{I}. \end{cases}$$

On retrouve ensuite $\widetilde{\mathbf{L}} = \mathbf{L}, \widetilde{\mathbf{U}} = \mathbf{U}$ et $\widetilde{\mathbf{V}} = \mathbf{A}_\ell^{1/2}\mathbf{V}$. ∎

La décomposition en valeurs singulières de $(\mathbf{X}\mathbf{A}_\ell^{1/2}, \mathbf{I}, \mathbf{D})$ conduit à l'analyse spectrale de la matrice $\mathbf{A}_\ell^{1/2}\mathbf{S}\mathbf{A}_\ell^{1/2}$. Les trajectoires discrètes du processus sont projetées sur le sous-espace engendré par les vecteurs

$$\widetilde{\mathbf{v}}_i = \mathbf{A}_\ell^{1/2}\mathbf{v}_j, \quad j = 1, \ldots, q.$$

Les trajectoires discrètes estimées $\hat{\mathbf{z}}_i$ se décomposent de manière équivalente sur la base $\mathbf{A}_\ell^{-1}$-orthonormée des $\{\widetilde{\mathbf{v}}_j\}$ par projection des données transformées $\mathbf{A}_\ell\mathbf{x}_i$:

$$\hat{\mathbf{z}}_i = \mathbf{A}_\ell\bar{\mathbf{x}} + \sum_{j=1}^{q}\left\langle\widetilde{\mathbf{v}}_j, \mathbf{A}_\ell\mathbf{x}_i\right\rangle_{\mathbf{A}_\ell^{-1}}\widetilde{\mathbf{v}}_j. \tag{9.4}$$

2.2 Dimension et paramètre de lissage

Cette méthode nécessite de régler les valeurs de deux paramètres : la dimension q du sous-espace ainsi que celle du paramètre de lissage ℓ. Ce choix doit être réalisé conjointement car, en pratique, la réduction de dimension opère également une sorte de lissage ou filtre passe-bas. En effet, il est courant d'observer sur les derniers vecteurs propres les composantes les plus perturbées de la fonction aléatoire. Cela s'explique simplement car dans le cas d'un processus stationnaire ou "peu" éloigné de la stationnarité, son opérateur de covariance commute avec l'opérateur retard et possède donc les mêmes fonctions propres périodiques. L'ACP ressemble alors fortement à une décomposition en séries de Fourier et c'est pourquoi, dans les premiers travaux sur ce type de données, Deville (1974) associait déjà ACP et analyse harmonique.

Les deux paramètres : dimension et lissage, interfèrent donc l'un sur l'autre. Plus précisément, la réduction de dimension permet de moins lisser à l'aide des splines et donc de trouver une valeur

optimale de ℓ plus petite que celle qui serait obtenue avec le lissage seul. C'est une des raisons qui fait que cette ACP fonctionnelle conduit à de meilleures estimations des courbes qu'une succession de régression non paramétrique pour laquelle chaque paramètre de lissage serait optimisé indépendamment par validation croisée.

Le même critère, aidant au choix de dimension (cf. chapitre 3 équation 3.5) peut être utilisé. Il est basé sur une approximation du risque moyen quadratique mesurant la qualité d'estimation du sous-espace de représentation E_q :

$$R_q = \mathbb{E}\frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|^2 = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q}$$

L'approximation par la théorie des perturbations de l'estimation jackknife est donnée par :

$$\widehat{R_{Pq}} = \frac{1}{n-1}\sum_{k=1}^{q}\sum_{j=k+1}^{p}\frac{\frac{1}{n}\sum_{i=1}^{n} c_{ik}^2 c_{ij}^2}{(\lambda_j - \lambda_k)^2} \tag{9.5}$$

où c_{ij} désigne le terme général de la matrice $\mathbf{X}\mathbf{A}_\ell^{1/2}\mathbf{V}_q$.

Besse et coll. (1997) ont montré, sur des données simulées, l'efficacité de cette approche associant dans le même problème d'optimisation des contraintes de réduction de rang et de régularité. Un lissage de chaque trajectoire prise séparément à base de validation croisée conduit à des résultats moins performants. La prise en compte de la structure de covariance à travers l'ACP permet une meilleure extraction du signal pour différents rapports signal sur bruit. On note encore que, lorsque la variance du bruit devient relativement importante, c'est-à-dire plus grande que la ou les dernières valeurs propres de la partie signal, il est préférable de réduire la dimension en conséquence. Cette étude montre également que le critère $\widehat{R_{Pq}}$ de choix de dimension fournit des résultats suffisamment précis pour être opérationnels.

3 Exemples : ACP de séries climatiques

3.1 ACP des précipitations

Nous préférons illustrer cette section par un exemple de données réelles particulièrement bruitées. Il s'agit des racines carrées des cumuls mensuels des précipitations de 26 villes en France observées pendant 10 ans (ECOSTAT 1991). La transformation (racine) s'avère nécessaire afin de stabiliser la variance comme dans le cadre d'un processus de Poisson. Pour traiter un problème suffisamment complexe on s'intéresse à 26×5 courbes observées durant 2 années consécutives. L'étude des données annuelles fournit le même type de résultats mais avec une composante en moins.

Une ACP classique calculée sur ces données fournit les résultats de la figure 9.2. Il s'agit donc des trois premiers vecteurs ou plutôt fonctions propres qui, très bruitées, sont difficiles à interpréter. Une ACP fonctionnelle incluant une contrainte de régularité contrôlé par le paramètre de lissage ℓ a ensuite été calculée. Le choix simultané de la dimension et de ce paramètre de lissage est guidé par les résultats de la figure 9.3. Celle-ci représente l'évolution de la stabilité du sous-espace de représentation en fonction de la valeur du paramètre de lissage et pour différentes dimensions. Cet indice $\widehat{R_{Pq}}$ lié au comportement d'écarts entre valeurs propres est très instable donc délicat à interpréter. Néanmoins, il apparaît que pour de petites valeurs de ℓ $(\log(\ell) < -5)$, seule la première composante associée à une simple tendance est stable. Pour de plus grandes valeurs $(\log(\ell) > 6)$, les données sont sur-lissées et beaucoup de composantes disparaissent. Le comportement de R_{P5} présentant un minimum conduit finalement à retenir $q = 5$ et $\rho \approx 1$.

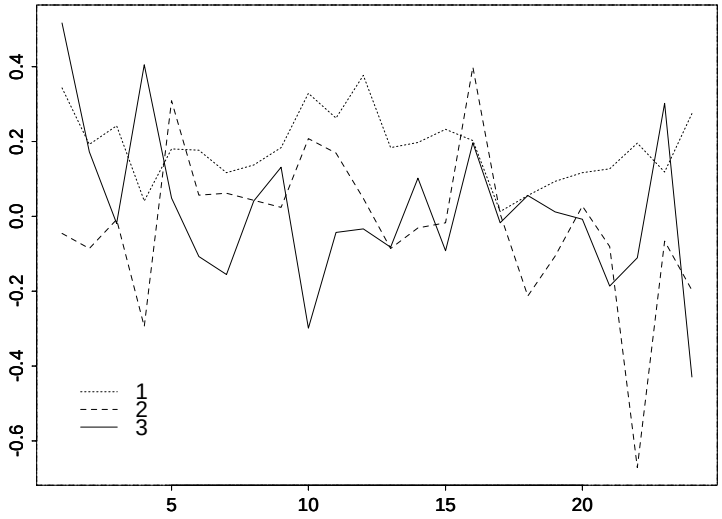

FIG. 9.2 – Les trois premières fonctions propres de l'ACP classique (sans contrainte de régularité) des données pluviométriques. Très irrégulières, elles sont difficiles à interpréter.

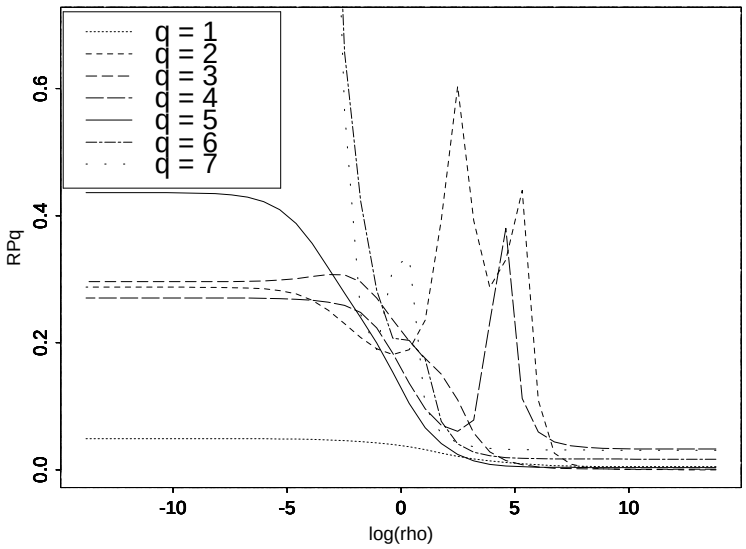

FIG. 9.3 – Estimation de la stabilité $\widehat{R_{Pq}}$ du sous-espace de projection en fonction de $\log(\rho)$ et pour différents choix de dimension.

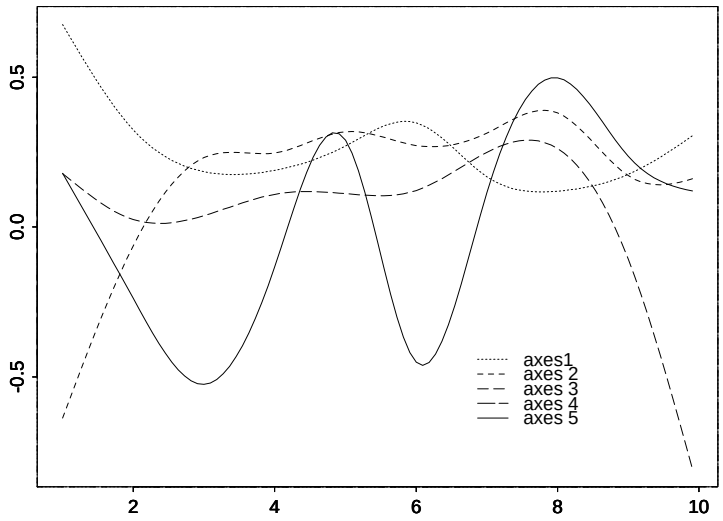

FIG. 9.4 – Les cinq premières composantes principales engendrant le sous-espace de projection $\widehat{\mathbf{P}}_q$.

Les composantes principales d'une telle ACP avec contraintes de régularité devient alors nettement plus facile à interpréter à partir du graphique des fonctions propres plus régulières (figure 9.4) qui révèlent différentes composantes périodiques.

3.2 ACP de températures

Cette section est le résultat d'une collaboration d'Antoniadou et coll. 2000 développée au sein d'un projet européen. Les données étudiées sont celles de la série CET des moyennes mensuelles des températures centrales en Angleterre qui débute en 1659. C'est la plus longue des séries de températures enregistrées disponibles pour des études climatiques. Elle représente une moyenne calculée sur plusieurs stations du centre de l'Angleterre ce qui permet, entre autres, de suppléer à des valeurs manquantes. Une étude préliminaire montre que cette série fait apparaître une tendance linéaire montrant un réchauffement de l'ordre de 0,5 ˚ C par siècle pour les moyennes des mois d'hiver mais seulement de 0,2 ˚ C pour les mois d'été.

Les moyennes mensuelles de la température en Angleterre peuvent être considérées comme l'observation d'un processus aléatoire réel et représentées par une série chronologique. Ces données peuvent également être considérées comme des observations discrétisées d'un processus aléatoire $(\mathbf{X}_i)_{i\in\mathbb{Z}}$ à valeurs dans un espace fonctionnel. Supposons que n trajectoires $\mathbf{x}_i$, $i = 1, \ldots, n$ du processus ont été mesurées en p instants de discrétisation $\{t_1, t_2, \ldots, t_p\}$. Ainsi, les données peuvent être rangées dans une matrice $\mathbf{X}$ d'éléments : $x_{ij} = \mathbf{x}_i(t_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, p$.

L'objectif de l'étude était l'étude conjointe du processus de température conjointement avec celui relatantle phénomène de balancier atmosphérique (north atlantic oscillation) présent dans l'Atlantique nord et dont l'influence est marquante sur le climat européen. Un traitement préalable a conduit à centrer les séries autour des moyennes climatiques afin d'éliminer la forte composante saisonnière puis à les lisser par la méthode du noyau. Les paramètres de lissage ont été optimisés afin de maximiser la corrélation linéaire des deux séries lissées centrées. Seule l'étude des températures est reprise ici.

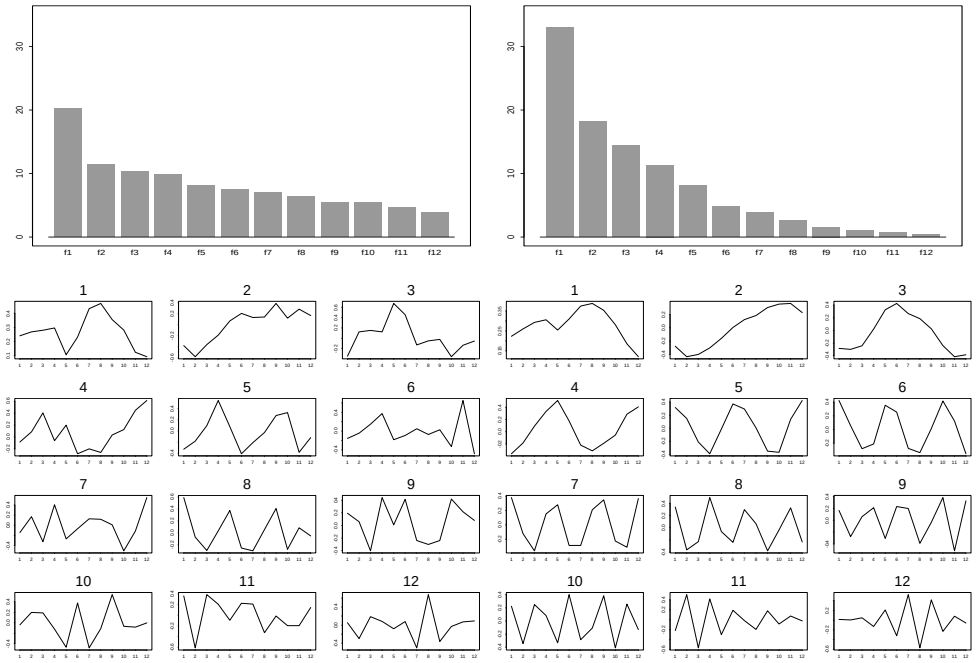

FIG. 9.5 – Éboulis des valeurs et fonctions propres de l'ACP des tableaux des données brutes (à gauche) et lissées (à droite) de température.

L'ACP des données brutes, qui apparaissent très bruitées, ne présente que peu d'intérêt. La décroissance des valeurs propres (cf. figure 9.5) est très lente, seul le premier vecteur propre, un peu trivial, semble fiable. L'axe associé (effet taille) distingue entre années chaudes et années froides. Lorsque l'ACP est combinée à un lissage, d'autres axes apparaissent comme pertinents dans la décomposition (figure 9.5). Compte tenu de la forme particulière des vecteurs propres, celle-ci ressemble beaucoup à une décomposition en série de Fourier. Cela signifie, qu'une fois lissée, la série centrée se comporte approximativement comme un processus stationnaire à accroissements indépendants avec décalage à l'origine.

La représentation des individus dans l'ACP des courbes de température mensuelle (figures 9.6 et 9.7) révèle la tendance déjà signalée et amplement médiatisée : la majorité des 25 dernières années apparaissent parmi celles qui sont en moyenne plus chaudes (Axe 1). Le plan (2,3) de cette même ACP apporte des résultats plus originaux. Il attribue principalement ce réchauffement moyen aux hivers. En effet, les 25 dernières années se projettent dans le demi-plan associé à des hivers plus doux que la moyenne générale. Ce réchauffement général expliqué principalement par des hivers moins rigoureux se confirme par létude d'Antoniadou et coll. (2000) du comportement des valeurs extrêmes.

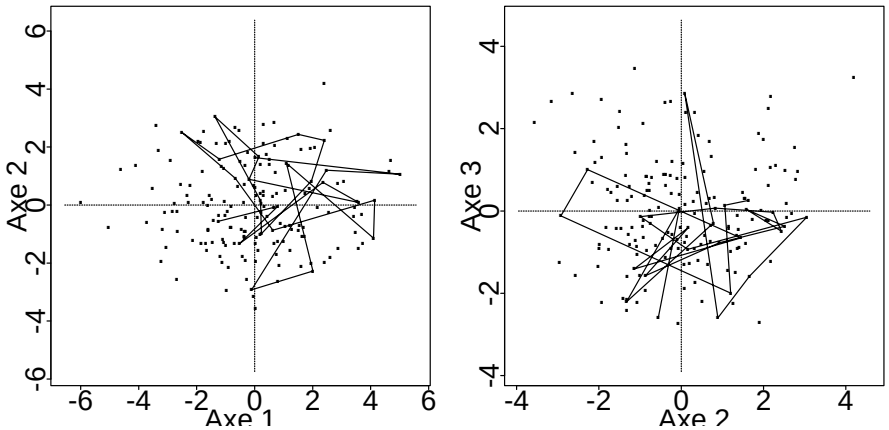

FIG. 9.6 – Représentation des individus sur les deux premiers plans de l'ACP des courbes annuelles lissées de température. La ligne brisée relie les 25 dernières années.

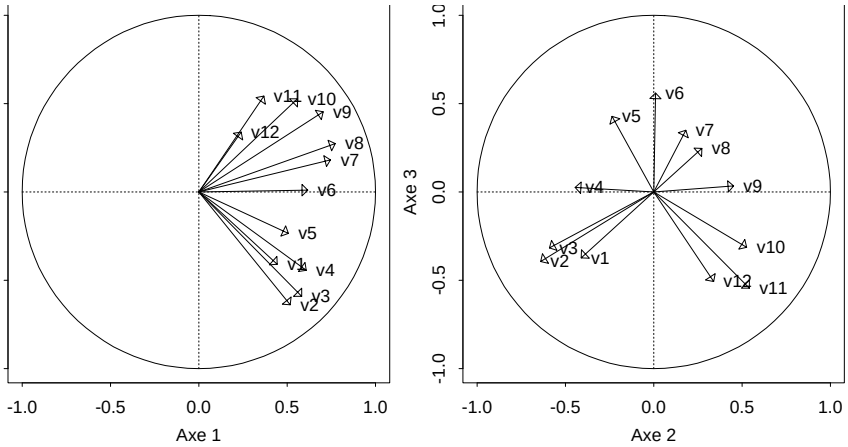

FIG. 9.7 – Représentation des variables (les mois) de l'ACP des courbes annuelles lissées de température.

Chapitre 10

Analyse Canonique

1 Introduction

L'analyse canonique (AC ou en anglais *canonical correlation analysys*) est une méthode de statistique descriptive multidimensionnelle qui présente des analogies à la fois avec l'analyse en composantes principales (ACP), pour la construction et l'interprétation de graphiques, et avec la régression linéaire, pour la nature des données. L'objectif général de l'analyse canonique est d'explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d'individus. L'étude des relations entre deux groupes de variables constitue la principale particularité de l'AC par rapport à l'ACP. De ce point de vue, l'AC est d'avantage proche de la régression linéaire multiple (explication d'une variable quantitative par un ensemble d'autres variables quantitatives), méthode dont elle constitue, d'ailleurs, une généralisation (on retrouve la régression lorsqu'un des deux groupes de l'AC ne comporte qu'une seule variable).

En fait, l'analyse canonique est, sur le plan théorique, une méthode centrale de la statistique descriptive multidimensionnelle, dans la mesure où elle généralise diverses autres méthodes et peut aussi être considérée comme un cas particulier d'ACP de deux paquets de variables dans un espace muni d'une métrique particuli-re (inverse par blocs des matrices de variance covariance).

Outre la régression linéaire, l'A.C. redonne en effet l'analyse factorielle discriminante lorsqu'un des deux groupes de variables est remplacé par les indicatrices d'une variable qualitative. Elle redonne également l'analyse factorielle des correspondances lorsque chacun des deux groupes est remplacé par les indicatrices d'une variable qualitative. Signalons également qu'il existe certaines généralisations de l'AC à plus de deux groupes de variables quantitatives et qu'elles permettent de retrouver l'analyse des correspondances multiples (en remplaçant chaque groupe par les indicatrices d'une variable qualitative), ainsi que l'ACP (en ne mettant qu'une seule variable quantitative dans chaque groupe). Nous ne nous intéresserons ici qu'à l'AC classique, entre deux groupes de variables.

En dépit de sa place centrale au sein des méthodes de statistique multidimensionnelle, pendant longtemps, l'A.C. n'était pas (ou très peu) enseignée dans ces cursus, compte tenu du petit nombre d'applications auxquelles elle donnait lieu. Les choses ont changé, d'abord vers le milieu des années 1990, avec le développement de la régression P.L.S. (*partial least squares*), méthode assez voisine de l'A.C., ensuite, plus récemment, avec l'apparition des données d'expression génomique (biopuces) combinées à des variables biologiques, dans une situtation qui relève typiquementde l'analyse canonique.

2 La méthode

2.1 Notations

Dans toute la suite de ce chapitre, on notera n le nombre d'individus considérés (autrement dit, la taille de l'échantillon observé), p le nombre de variables (quantitatives) du premier groupe et q le nombre de variables (également quantitatives) du second groupe. On désignera par $\mathbf{X}$ la matrice, de dimension $n \times p$, contenant les observations relatives au premier groupe de variables et par $\mathbf{Y}$ la matrice, de dimension $n \times q$, contenant celles relatives au second groupe. La j-ième colonne de $\mathbf{X}$ ($j = 1, \ldots, p$) contient les observations x_i^j de la j-ième variable du premier groupe (notée X^j) sur les n individus considérés ($i = 1, \ldots, n$). De même, la k-ième colonne de $\mathbf{Y}$ ($k = 1, \ldots, q$) contient les observations y_i^k de la k-ième variable du second groupe (notée Y^k).

Généralement, en A.C., on suppose $n \geq p$, $n \geq q$, $\mathbf{X}$ de rang p et $\mathbf{Y}$ de rang q. De plus, sans perte de généralité, on suppose également $p \leq q$ (on désigne donc par premier groupe celui qui comporte le moins de variables). Compte tenu des particularités des données de biopuces, les quatre premières hypothèses ci-dessus pourront ne pas être vérifiées dans certains exemples.

2.2 Représentations vectorielles des données

Comme en A.C.P., on peut considérer plusieurs espaces vectoriels réels associés aux observations.

Tout d'abord, l'espace des variables ; c'est $F = \mathbb{R}^n$, muni de la base canonique et d'une certaine métrique, en général l'identité. À chaque variable X^j est associé un vecteur unique x^j de F dont les coordonnées sur la base canonique sont les x_i^j ($i = 1, \ldots, n$). De même, à chaque variable Y^k est associé un vecteur unique y^k de F, de coordonnées les y_i^k. On peut ainsi définir dans F deux sous-espaces vectoriels : F_X, engendré par les vecteurs x^j ($j = 1, \ldots, p$), en général de dimension p, et F_Y, engendré par les vecteurs y^k ($k = 1, \ldots, q$), en général de dimension q.

Remarque. — Il est courant de munir l'espace vectoriel F de la métrique dite "des poids", définie, relativement à la base canonique, par la matrice diag $(p_1, \ldots, p_n)$, où les p_i ($i = 1, \ldots, n$) sont des poids (positifs et de somme égale à 1) associés aux individus observés. Lorsque tous ces poids sont égaux, ils valent nécessairement $\frac{1}{n}$ et la matrice définissant la métrique des poids vaut $\frac{1}{n}\mathbf{I}_n$, où $\mathbf{I}_n$ est la matrice identité d'ordre n. Dans ce cas, il est équivalent d'utiliser la métrique identité, ce que nous ferons par la suite, dans la mesure où les individus seront systématiquement équipondérés.

On peut ensuite considérer deux espaces vectoriels pour les individus, $E_1 = \mathbb{R}^p$ et $E_2 = \mathbb{R}^q$, eux aussi munis de leur base canonique et d'une certaine métrique. Dans E_1, chaque individu i est représenté par le vecteur x_i, de coordonnées x_i^j ($j = 1, \ldots, p$) sur la base canonique. De même, dans E_2, l'individu i est représenté par le vecteur y_i, de coordonnées les y_i^k.

En fait, c'est surtout l'espace F que nous considèrerons par la suite, la définition de l'A.C. y étant plus naturelle.

2.3 Principe de la méthode

Le principe général de l'A.C. est décrit ci-dessous, dans l'espace des variables F.

Dans un premier temps, on cherche un couple de variables (V^1, W^1), V^1 étant une combinaison linéaire des variables X^j (donc un élément de F_X), normée, et W^1 une combinaison linéaire des variables Y^k (donc un élément de F_Y), normée, telles que V^1 et W^1 soient le plus corrélées possible.

Ensuite, on cherche le couple normé (V^2, W^2), V^2 combinaison linéaire des X^j non corrélée

à V^1 et W^2 combinaison linéaire des Y^k non corrélée à W^1, telles que V^2 et W^2 soient le plus corrélées possible. Et ainsi de suite...

Remarque. — Dans la mesure où l'A.C. consiste à maximiser des corrélations, quantités invariantes par translation et par homothétie de rapport positif sur les variables, on peut centrer et réduire les variables initiales X^j et Y^k sans modifier les résultats de l'analyse. Pour des raisons de commmodité, on le fera systématiquement. Par conséquent, les matrices $\mathbf{X}$ et $\mathbf{Y}$ seront désormais supposées centrées et réduites (en colonnes).

L'A.C. produit ainsi une suite de p couples de variables (V^s, W^s), $s = 1, \ldots, p$. Les variables V^s constituent une base orthonormée de F_X (les V^s, combinaisons linéaires de variables centrées, sont centrées ; comme elles sont non corrélées, elles sont donc orthogonales pour la métrique identité). Les variables W^s constituent, de même, un système orthonormé de F_Y (ils n'en constituent une base que si $q = p$). Les couples (V^s, W^s), et plus particulièrement les premiers d'entre eux, rendent compte des liaisons linéaires entre les deux groupes de variables initiales. Les variables V^s et W^s sont appelées les *variables canoniques*. Leurs corrélations successives (décroissantes) sont appelées les *coefficients de corrélation canonique* (ou *corrélations canoniques*) et notées ρ_s $(1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_p \geq 0)$.

Remarque. — Toute variable canonique V^{s_0} est, par construction, non corrélée (donc orthogonale) avec les autres variables canoniques V^s, $s \neq s_0$. On peut également montrer que V^{s_0} est non corrélée avec W^s, si $s \neq s_0$ (la même propriété est bien sûr vraie pour toute variable W^{s_0} avec les variables V^s, $s \neq s_0$).

Remarque. — Si nécessaire, on peut compléter le système des variables W^s $(s = 1, \ldots, p)$ pour obtenir une base orthonormée de F_Y dans laquelle les dernières variables W^s $(s = p+1, \ldots, q)$ sont associées à des coefficients de corrélation canonique nuls ($\rho_s = 0$, pour $s = p+1, \ldots, q$).

2.4 Aspects mathématiques

Dans l'espace vectoriel F muni de la métrique identité, notons $\mathbf{P_X}$ et $\mathbf{P_Y}$ les matrices des projecteurs orthogonaux sur les sous-espaces F_X et F_Y. Les formules usuelles de définition des projecteurs permettent d'écrire ($\mathbf{X}'$ désignant la matrice transposée de $\mathbf{X}$) :

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; ; \; \mathbf{P_Y} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'.$$

On peut alors montrer la propriété ci-dessous.

PROPOSITION 10.1. — *Les vecteurs V^s sont les vecteurs propres normés de la matrice $\mathbf{P_X P_Y}$ respectivement associés aux valeurs propres λ_s rangées par ordre décroissant (on peut vérifier que ces valeurs propres sont comprises entre 1 et 0). De même, les vecteurs W^s sont les vecteurs propres normés de la matrice $\mathbf{P_Y P_X}$ respectivement associés aux mêmes valeurs propres λ_s. De plus, les coefficients de corrélation canonique ρ_s sont les racines carrées positives de ces valeurs propres : $\rho_s = \sqrt{\lambda_s}$, $s = 1, \ldots, p$ (le logiciel SAS fournit les corrélations canoniques ρ_s ainsi que leurs carrés λ_s).*

2.5 Représentations graphiques

Comme en A.C.P., les représentations graphiques des résultats d'une A.C. se font en dimension réduite (souvent 2 ou 3). Nous noterons d cette dimension, avec : $1 \leq d \leq p$. Plusieurs représentations sont envisageables, à la fois pour les variables et pour les individus.

Représentation des variables dans le sous-espace F_X

Désignons par v^s et w^s les vecteurs de F_X et F_Y respectivement associés aux variables canoniques V^s et W^s.

Dans F_X, on considère la base orthonormée $(v^1, \ldots, v^p)$ que l'on restreint à $(v^1, \ldots, v^d)$ pour les représentations graphiques.

On peut tout d'abord représenter chacune des variables initiales X^j au moyen de ses coordonnées sur les v^s. Ces coordonnées s'obtiennent en calculant les produits scalaires $< x^j, v^s >$, $j = 1, \ldots, p$, $s = 1, \ldots, d$. Les variables X^j étant centrées et réduites, les vecteurs x^j sont centrés et normés (et il en va de même pour les vecteurs v^s), de sorte que ces produits scalaires sont égaux aux corrélations entre variables initiales X^j et variables canonique V^s (au coefficient n près, puisqu'on a considéré la métrique identité).

Dans le même espace, on peut également représenter les variables de l'autre groupe, les Y^k, en projetant tout d'abord les vecteurs y^k dans F_X, au moyen de $\mathbf{P_X}$, puis en prenant le produit scalaire de ces projections avec les vecteurs v^s. On doit donc calculer pour cela les produits scalaires

$$< \mathbf{P_X}(y^k), v^s > = < y^k, \mathbf{P_X}(v^s) > = < y^k, v^s >,$$

encore égaux aux corrélations entre les variables initiales Y^k et les variables canoniques V^s.

Dans la mesure où le graphique ainsi obtenu est "bon" (sur ce point, voir plus loin), on peut l'utiliser pour interpréter les relations (proximités, oppositions, éloignements) entre les deux ensembles de variables. Par construction, ce graphique représente les corrélations entre les variables canoniques V^s et les variables initiales X^j et Y^k, corrélations à la base de son interprétation. On peut aussi conforter cette interprétation en utilisant les coefficients de corrélation linéaire entre variables X^j, entre variables Y^k, et entre variables X^j et Y^k. Tous ces coefficients sont en général fournis par les logiciels.

Représentation des variables dans le sous-espace F_Y

De façon symétrique, on restreint le système $(w^1, \ldots, w^p)$ de F_Y aux premières variables $(w^1, \ldots, w^d)$, par rapport auxquelles on représente aussi bien les variables initiales X^j que les Y^k, selon le même principe que celui décrit ci-dessus (les coordonnées sont les corrélations).

Là encore, dans la mesure où ce graphique est "bon", il permet d'interpréter les relations entre les deux ensembles de variables.

Les deux graphiques (dans F_X et dans F_Y) ayant la même qualité et conduisant aux mêmes interprétations, un seul suffit pour interpréter les résultats d'une analyse.

Représentation des individus

Dans chacun des espaces relatifs aux individus (E_1 et E_2), il est encore possible de faire une représentation graphique de ces individus en dimension d, ces deux représentations graphiques étant comparables (d'autant plus comparables que les corrélations canoniques sont élevées).

En fait, on peut vérifier que les coordonnées des individus sur les axes canoniques pour ces deux représentations sont respectivement données par les lignes des matrices $\mathbf{V_d}$ (dans E_1) et $\mathbf{W_d}$ (dans E_2), $\mathbf{V_d}$ et $\mathbf{W_d}$ désignant les matrices $n \times d$ dont les colonnes contiennent les coordonnées des d premières variables canoniques sur la base canonique de F.

Choix de la dimension

Comme dans toute méthode factorielle, différents éléments doivent être pris en compte pour le choix de la dimension d dans laquelle on réalise les graphiques (et dans laquelle on interprète les résultats).

- Tout d'abord, il est clair que d doit être choisi petit, l'objectif général de la méthode étant d'obtenir des résultats pertinents dans une dimension réduite ; ainsi, le plus souvent, on choisi d égal à 2 ou à 3.
- Plus l'indice de dimension s augmente, plus la corrélation canonique ρ_s diminue ; or, on ne s'intéresse pas aux corrélations canoniques faibles, puisqu'on cherche à expliciter les relations entre les deux groupes de variables ; par conséquent, les dimensions correspondant à des ρ_s faibles peuvent être négligées.
- Le pourcentage que chaque valeur propre représente par rapport à la somme, c'est-à-dire par rapport à la trace de la matrice diagonalisée, facilitent également le choix de d (voir la remarque 5).

2.6 Compléments : analyse canonique et régression multivariée

Introduction

Ouvrages et logiciels anglo-saxons de statistique présentent souvent l'analyse canonique parallèlement à la régression linéaire multivariée (régression d'un ensemble de variables Y^k sur un autre ensemble de variables X^j). Cette approche est, en fait, assez naturelle, dans la mesure où les données sont de même nature dans les deux méthodes et où l'on cherche, dans l'une comme dans l'autre, des relations linéaires entre variables.

Il convient toutefois de noter les deux différences fondamentales entre les deux métho-des : contrairement à ce qu'il se passe en A.C., les deux ensembles de variables X^j et Y^k ne sont pas symétriques en régression, puisqu'il s'agit d'expliquer les variables Y^k au moyen des variables X^j ; d'autre part, toujours en régression, on suppose la normalité des variables réponses Y^k, alors qu'aucune hypothèse de cette nature n'est nécessaire en A.C. L'avantage de cette hypothèse (lorsqu'elle est "raisonnable") est de permettre de réaliser des tests dans le modèle de régression.

Le modèle de régression multivariée

Le modèle de régression multivariée des variables Y^k sur les variables X^j s'écrit :

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \ ;$$

les matrices $\mathbf{Y}$, $n \times q$ et $\mathbf{X}$, $n \times p$, sont celles introduites en A.C. ; $\mathbf{B}$ est la matrice $p \times q$ des paramètres inconnus, à estimer ; $\mathbf{U}$ est la matrice $n \times q$ des erreurs du modèle. Chaque ligne U_i de $\mathbf{U}$ est un vecteur aléatoire de $\mathbb{R}^q$ supposé $\mathcal{N}_q(0, \Sigma)$, les U_i étant indépendants (Σ est une matrice inconnue, à estimer, supposée constante en i).

L'estimation maximum de vraisemblance de $\mathbf{B}$ conduit à la solution :

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

On appelle alors *valeurs prédites* (de $\mathbf{Y}$ par le modèle) les quantités :

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} \ ;$$

d'autre part, on appelle *résidus* les quantités :

$$\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$$

(dans l'écriture ci-dessus, $\mathbf{P}_{\mathbf{X}}^{\perp}$ désigne, dans $\mathbb{R}^n$, le projecteur orthogonal sur le sous-espace supplémentaire orthogonal à F_X dans $\mathbb{R}^n$; on sait que ce projecteur s'écrit : $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$).

Matrices intervenant dans les tests

Dans le cadre du modèle gaussien, on peut tester la significativité du modèle en généralisant le test de Fisher, bien connu dans le cas unidimensionnel. Au numérateur de la statistique de Fisher figure la norme carrée du vecteur $\hat{y} - \overline{y}$, ici remplacée par $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ (cette matrice est centrée). Au dénominateur figure la norme carrée des résidus, ici remplacée par $\hat{\mathbf{U}}'\hat{\mathbf{U}}$ (on néglige, pour l'instant, les degrés de liberté de ces quantités). La statistique de Fisher est donc remplacée par le produit matriciel $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}$. Comme on a $\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$, il vient : $\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{H}$ (la notation $\mathbf{H}$ est standard, car il s'agit d'une matrice proche de 0 sous l'hypothèse nulle de non significativité du modèle). D'autre part, $\hat{\mathbf{U}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$ entraîne : $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y} = \mathbf{E}$ (il s'agit encore d'une notation standard, cette matrice représentant les erreurs du modèle). Les tets multidimensionnels de significativité du modèle sont ainsi basés sur l'étude des valeurs propres soit du produit matriciel

$$\mathbf{H}\mathbf{E}^{-1} = (\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y})(\mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y})^{-1},$$

soit encore du produit $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$, les valeurs propres de ces deux matrices se déduisant les unes des autres. Développons le second produit matriciel :

$$\mathbf{H} + \mathbf{E} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} + \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y} = \mathbf{Y}'\mathbf{Y};$$

d'où :

$$\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1},$$

matrice ayant les mêmes valeurs propres que

$$\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}' = \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}},$$

c'est-à-dire les λ_s ($s = 1, \ldots, p$), carrés des corrélations canoniques.

Remarque. — On peut vérifier (le résultat est classique) que les valeurs propres de la matrice $\mathbf{H}\mathbf{E}^{-1}$ valent $\dfrac{\lambda_s}{1-\lambda_s}$. Ces valeurs propres sont fournies par le logiciel SAS, ainsi que les pourcentages (et les pourcentages cumulés) qu'elles représentent par rapport à leur somme, trace de la matrice $\mathbf{H}\mathbf{E}^{-1}$.

En interprétant ces pourcentages comme la part d'inertie globale du nuage des individus restituée par les différents axes canoniques (ce qu'elles sont, par exemple, en analyse factorielle discriminante), ces quantités facilitent le choix de la dimension d retenue pour les graphiques et les interprétations.

Tests

Il existe plusieurs tests de significativité du modèle de régression multivariée, en général équivalents (au moins au niveau des décisions qu'ils entraînent). Ces tests sont les généraliSations classiques du test de Fisher au cas multivarié (on les retrouve, par exemple, en analyse de variance multivariée). Le logiciel SAS fournit les trois premiers ci-dessous, mais pas le quatrième. Il fournit également le test de Roy, basé sur la plus grande valeurs propre de la matrice $\mathbf{H}\mathbf{E}^{-1}$, soit $\dfrac{\lambda_1}{1-\lambda_1}$, mais ce test est à déconseiller.

- Le test de Wilks, adaptation du test du rapport des vraisemblances, est basé sur la statistique

$$\Lambda = \prod_{s=1}^{p}(1-\lambda_s) = \prod_{s=1}^{p}(1-\rho_s^2).$$

- Le test de la trace de Pillai est basé sur la statistique

$$Z = \text{trace } \mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \sum_{s=1}^{p}\lambda_s.$$

- Le test de la trace de Lawley-Hotelling est basé sur la statistique

$$T^2 = \text{trace } \mathbf{H}\mathbf{E}^{-1} = \sum_{s=1}^{p}\frac{\lambda_s}{1-\lambda_s}.$$

- Le test du khi-deux, basé sur la statistique

$$K = -[(n-1)-\frac{1}{2}(p+q+1)]\ln\prod_{s=1}^{p}(1-\lambda_s).$$

Le test du khi-deux présente l'avantage d'être directement utilisable, puisqu'on compare la statistique K à une loi de khi-deux à pq degrés de libertés (il s'agit d'un test approché).

Dans les trois autres tests ci-dessus, on doit transformer la statistique (Λ, Z ou T^2) pour obtenir un test de Fisher approché, les transformations étant assez compliquées à expliciter (toutefois, SAS les réalise automatiquement).

Remarque. — Dans un article de 1951, Rao a montré que, dans la plupart des cas, l'approximation de Fisher du test de Wilks est la meilleure. C'est donc le test que nous conseillerons dans ce cas là.

Si le modèle de régression est signifcatif (il en va alors de même pour l'analyse canonique), on peut tester la significativité d'une dimension et de l'ensemble des suivantes, en particulier pour guider le choix de la dimension en A.C. Ainsi, supposons que les corrélations canoniques soient significatives depuis la première jusqu'à la k-ième ($1 \leq k \leq p$). On peut alors tester l'hypothèse nulle

$$\{H_0 : \rho_{k+1} = \cdots = \rho_p = 0\} \ (\Longleftrightarrow \{H_0 : d = k\})$$

contre l'alternative

$$\{H_1 : \rho_{k+1} > 0\} \ (\Longleftrightarrow \{H_1 : d > k\}).$$

Pour cela, il faut adapter soit le test de Wilks, soit le test du khi-deux.

Pour le test de Wilks, il suffit de faire le produit des quantités $(1-\lambda_s)$ de l'indice $k+1$ à l'indice p et d'adapter la transformation en fonction des nouvelles dimensions. SAS le fait automatiquement. Pour le test du khi-deux, il faut considérer la statistique

$$K_k = -[(n-1-k)-\frac{1}{2}(p+q+1)+\sum_{s=1}^{k}\frac{1}{\lambda_s}]\ln\prod_{s=k+1}^{p}(1-\lambda_s)$$

et la comparer à une loi de khi-deux à $(p-k)(q-k)$ degrés de liberté.

Remarque. — Dans l'utilisation de ces tests, il convient de ne pas perdre de vue d'une part qu'il s'agit de tests approchés (d'autant meilleurs que la taille de l'échantillon, n, est grande), d'autre part qu'ils ne sont valables que sous l'hypothèse de normalité des variables Y^k.

3 Un exemple : la nutrition des souris

3.1 Les données

Ces données nous ont été proposées par l'Unité Pharmacologie-Toxicologie de l'INRA de Saint Martin du Touch, près de Toulouse. Elles ont été produites par Pascal Martin et Thierry Pineau.

Il s'agit d'une population de 40 souris sur lesquelles, entre autres choses, on a observé deux groupes de variables. Un premier groupe est constitué par 10 gènes spécifiques de la nutrition chez la souris. Chaque variable est en fait la mesure (quantitative) de l'expression du gène correspondant, réalisée par *macroarrays* sur membranes de nylon avec marquage radioactif. En fait, dans l'expérience, on disposait de 120 gènes parmi lesquels 10 ont été selectionnés (a priori, parmi les plus pertinents) pour réduire le volume des données. Pour mémoire, les codes de ces gènes sont les suivants :

```
CAR1 BIEN CYP3A11 CYP4A10 CYP4A14 AOX THIOL CYP2c29 S14 GSTpi2 .
```

Un deuxième groupe de variables est constitué par les pourcentages de 21 acides gras hépatiques ; il s'agit de variables quantitatives, avec la particularité que, tous les acides gras hépatiques ayant été pris en compte, la somme de ces variables vaut 100 pour tout individu. Pour mémoire, les codes de ces acides gras sont les suivants :

```
   C14_0    C16_0    C18_0 C16_1n_9 C16_1n_7 C18_1n_9 C18_1n_7
C20_1n_9 C20_3n_9 C18_2n_6 C18_3n_6 C20_2n_6 C20_3n_6 C20_4n_6
C22_4n_6 C22_5n_6 C18_3n_3 C20_3n_3 C20_5n_3 C22_5n_3 C22_6n_3 .
```

Le but de l'analyse canonique de ces données est donc d'étudier les relations pouvant exister entre gènes et acides gras.

Remarque. — On notera que les hypothèses usuelles relatives aux données d'une A.C., que nous avons mentionnées à la fin du 2.1, sont ici toutes vérifiées.

3.2 Traitements préliminaires

Nous donnons ci-dessous les statistiques élémentaires relatives aux deux groupes de variables. Pour les corrélations entre les variables de chaque groupe, on se reportera aux annexes A et B.

Variable	N	Mean	Std Dev	Minimum	Maximum
CAR1	40	220.85000	60.76881	135	376
BIEN	40	214.67500	58.14191	105	385
CYP3A11	40	518.15000	294.13415	170	1327
CYP4A10	40	179.17500	83.91873	89	399
CYP4A14	40	171.37500	112.53733	99	658
AOX	40	830.55000	237.60385	452	1529
THIOL	40	644.05000	277.55461	206	1260
CYP2c29	40	1062.0	336.10239	371	1934
S14	40	328.65000	216.91881	132	1350
GSTpi2	40	2266.0	717.60913	965	3903

Variable	N	Mean	Std Dev	Minimum	Maximum

C14_0	40	0.76300	0.80057	0.22000	3.24000
C16_0	40	23.02600	3.57303	14.65000	29.72000
C18_0	40	6.74700	2.64016	1.68000	10.97000
C16_1n_9	40	0.68700	0.28498	0.29000	1.50000
C16_1n_7	40	4.41875	2.98497	1.59000	13.90000
C18_1n_9	40	25.27325	7.33966	14.69000	41.23000
C18_1n_7	40	4.42600	3.37585	1.53000	15.03000
C20_1n_9	40	0.28400	0.13965	0	0.65000
C20_3n_9	40	0.30675	0.72116	0	2.89000
C18_2n_6	40	15.27750	8.76020	2.31000	40.02000
C18_3n_6	40	0.37450	0.87840	0	5.07000
C20_2n_6	40	0.18525	0.20236	0	0.83000
C20_3n_6	40	0.77600	0.46167	0.11000	1.64000
C20_4n_6	40	5.27925	4.45999	0.75000	15.76000
C22_4n_6	40	0.18400	0.25213	0	0.73000
C22_5n_6	40	0.43700	0.66392	0	2.52000
C18_3n_3	40	2.88800	5.82863	0	21.62000
C20_3n_3	40	0.09100	0.17930	0	0.64000
C20_5n_3	40	1.78950	2.59001	0	9.48000
C22_5n_3	40	0.87175	0.85598	0	2.58000
C22_6n_3	40	5.91400	5.33487	0.28000	17.35000

Remarque. — Comme indiqué dans la remarque 2, ces variables ont été centrées et rédui-tes avant la réalisation de l'A.C.

3.3 Analyse canonique

Généralités

Les premiers résultats fournis par une A.C. sont les corrélations croisées entre les deux groupes de variables. Nous donnons ces corrélations dans l'annexe C.

Ensuite sont données les corrélations canoniques reproduites ci-dessous.

	Canonical Correlation
1	0.990983
2	0.978581
3	0.957249
4	0.891429
5	0.799633
6	0.794380
7	0.770976
8	0.635902
9	0.626384
10	0.325094

On notera que "le plus petit" groupe ne comportant que 10 variables, on ne peut déterminer que 10 corrélations canoniques. L'objectif principal de l'A.C. étant d'étudier les relations entre variables des deux groupes, on peut noter ici qu'il existe effectivement des relations fortes entre ces deux groupes, puisque les premiers coefficiens canoniques sont très élevés. Compte tenu des valeurs importantes des premiers coefficients, on peut raisonnablement se contenter de deux ou trois dimensions pour étudier les résultats fournis par la méthode et nous avons choisi ici seulement deux dimensions, compte tenu qu'il s'agit essentiellement d'une illustration.

Remarque. — Les valeurs propres de la matrice $\mathbf{HE}^{-1}$ et les pourcentages d'inertie restitués par les différentes dimensions sont les suivants :

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	54.7032	32.1069	0.5553	0.5553
2	22.5963	11.6451	0.2294	0.7847
3	10.9512	7.0816	0.1112	0.8958
4	3.8696	2.0964	0.0393	0.9351
5	1.7732	0.0629	0.0180	0.9531
6	1.7103	0.2448	0.0174	0.9705
7	1.4655	0.7866	0.0149	0.9854
8	0.6789	0.0332	0.0069	0.9922
9	0.6457	0.5275	0.0066	0.9988
10	0.1182		0.0012	1.0000

Par ailleurs, les tests de Wilks, de significativité de chaque dimension, sont les suivants :

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.00000023	2.19	210	104.61	<.0001
2	0.00001272	1.63	180	100.74	0.0035
3	0.00030012	1.25	152	95.57	0.1202
4	0.00358677	0.96	126	89.05	0.5890
5	0.01746624	0.82	102	81.12	0.8259
6	0.04843824	0.78	80	71.72	0.8542
7	0.13128287	0.69	60	60.78	0.9228
8	0.32367928	0.53	42	48.23	0.9807
9	0.54342420	0.47	26	34	0.9762
10	0.89431401	0.18	12	18	0.9980

On voit que le choix de la dimension 2 est recommandé.

Graphique des individus

Dans un premier temps, nous avons réalisé le graphique des individus (les 40 souris) relativement aux deux premiers axes canoniques de l'espace des gènes E_1 (voir la Figure 1). Ce graphique a pour seul but de regarder l'homogéneité de l'ensemble des individus. S'il ne présente aucune particularité notable, il y a néanmoins des individus occupant des positions assez différenciées et il pourrait être intéressant d'étudier en détail ce qui les caractérise.

On notera qu'on a également réalisé le graphique des individus relativement aux deux premiers axes de l'autre espace (espace des acides gras, E_2) et qu'il est très semblable à celui-ci.

Graphique des variables

Pour la représentation des variables, nous avons considéré le sous-espace F_X, engendré par les 10 gènes, et nous avons représenté à la fois les gènes et les acides gras relativement aux deux

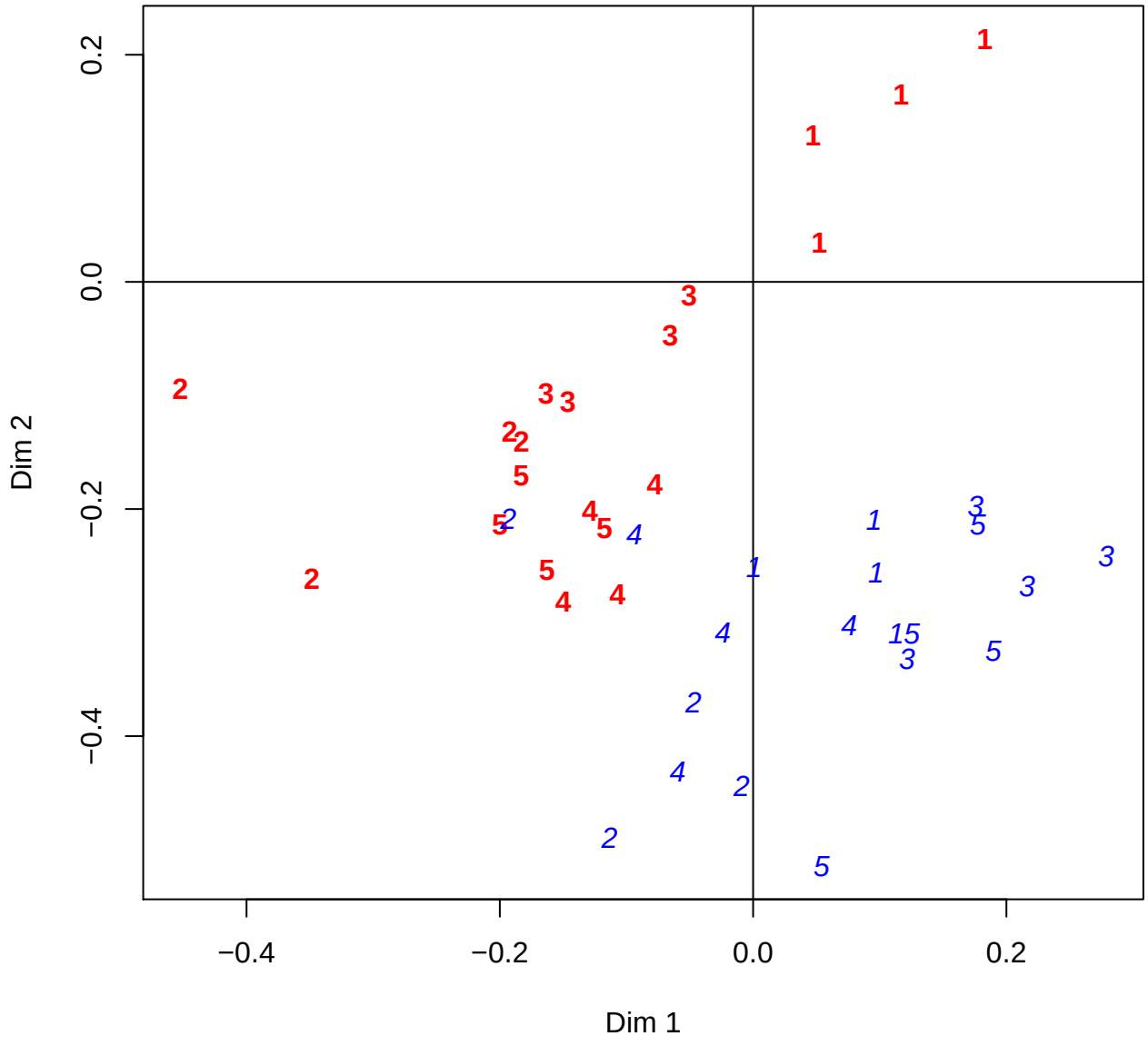

FIG. 10.1 – *Nutrition : Représentation des individus (souris) dans l'espace des gènes.*

premières variables canoniques, V^1 et V^2 (voir la Figure 2). Comme indiqué en 2.5, les coordonnées des variables initiales sont fournies par leur corrélations avec les variables canoniques.

Certaines associations entre gènes et acides gras, en particulier celles correspondant à des points éloignés de l'origine, sont intéressantes à noter.

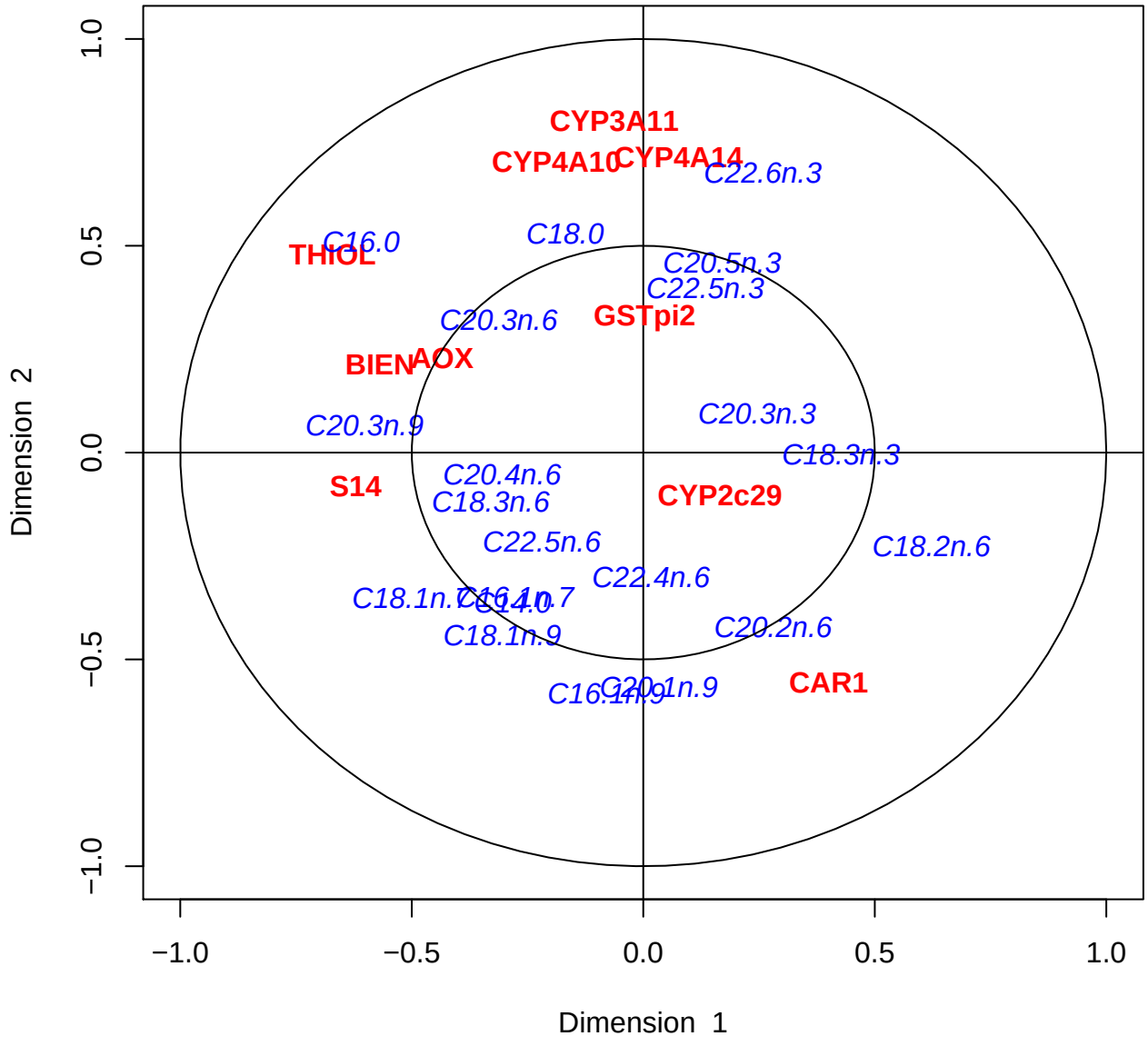

FIG. 10.2 – *Nutrition : Représentation des gènes et des acides dans le sous-espace des gènes.*

Bibliography

Agresti, A. (1990). *Categorical data analysis*. Wiley.

Antoniadis, A., J. Berruyer, and R. Carmona (1992). *Régression non linéaire et applications*. Economica.

Ardilly, P. (1994). *Les techniques de sondage*. Technip.

Berry, M. and L. Gordon (1997). *Data Mining, techniques appliquées au marketing, à la vente et aux services clients*. Masson.

Besse, P. (1992). Pca stability and choice of dimensionality. *Statistics & Probability Letters 13*, 405–410.

Besse, P., H. Cardot, and F. Ferraty (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis 24*, 255–270.

Besse, P. and F. Ferraty (1995). A fixed effect curvilinear model. *Computational Statistics 10*, 339–351.

Besse, P. and J. Ramsay (1986). Principal component analysis of sampled curves. *Psychometrika 51*, 285–311.

Bourret, P., J. Reggia, and M. Samuelides (1991). *Réseaux neuronaux*. Teknea.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and regression trees*. Wadsworth & Brooks.

Celeux, G. (1990). *Analyse discriminante sur variables continues*. INRIA.

Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy (1989). *Classification automatique des données*. Dunod.

Celeux, G. and J.-P. Nakache (1994). *Analyse discriminante sur variables qualitatives*. Polytechnica.

Collett, D. (1991). *Modelling binary data*. Chapman & Hall.

Dobson, A. (1990). *An introduction to generalized linear models*. Chapman and Hall.

Droesbeke, J., B. Fichet, and P. Tassi (1992). *Modèles pour l'Analyse des Données Multidimensionnelles*. Economica.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Methods*. SIAM.

Everitt, B. and G. Dunn (1991). *Applied Multivariate Data Analysis*. Edward Arnold.

Green, P. and B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall.

Jobson, J. (1991). *Applied Multivariate Data Analysis*, Volume I : Regression and experimental design. Springer-Verlag.

Jobson, J. (1992). *Applied Multivariate Data Analysis*, Volume II : Categorical and multivariate methods. Springer-Verlag.

Jolliffe, I. (2002). *Principal Component Analysis* (2nd edition ed.). Springer-Verlag.

Kaufman, L. and J. Rousseeuw, P. (1990). *Finding groups in data*. Wiley.

Lefèbure, R. and G. Venturi (1998). *Le data Mining*. Eyrolles.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press.

McCullagh, P. and J. Nelder (1983). *Generalized Linear Models*. Chapman & Hall.

Monfort, A. (1982). *Cours de Statistique Mathématique*. Economica.

Ramsay, J. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, B 53*, 539–572. with discussion.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis*. Springer-Verlag.

Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*. Technip.

SAS (1989). *SAS/STAT User's Guide* (fourth ed.), Volume 2. Sas Institute Inc. version 6.

SAS (1995). *SAS/INSIGHT User's Guide* (Third ed.). Sas Institute Inc. version 6.

Thiria, S., Y. Lechevallier, O. Gascuel, and S. Canu (1997). *Statistique et méthodes neuronales*. Dunod.

Tomassonne, R., S. Audrain, E. Lesquoy-de Turckheim, and C. Millier (1992). *La régression, nouveaux regards sur une ancienne méthode statistique*. Masson.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Chapitre A

Outils algébriques

Ce chapitre se propose de rassembler des notations et rappels d'algèbre linéaire ainsi que quelques compléments mathématiques du niveau du premier cycle des Universités.

Dans tout ce qui suit, E et F sont deux espaces vectoriels réels munis respectivement des bases canoniques $\mathcal{E} = \{\mathbf{e}_j \ ; \ j = 1, \ldots, p\}$ et $\mathcal{F} = \{\mathbf{f}_i \ ; \ i = 1, \ldots, n\}$. On note indifféremment soit un vecteur de E ou de F, un endomorphisme de E, ou une application linéaire de E dans F, soit leurs représentations matricielles dans les bases définies ci-dessus.

1 Matrices

1.1 Notations

La matrice d'ordre $(n \times p)$ associée à une application linéaire de E dans F est décrite par un tableau :

$$\mathbf{A} = \begin{bmatrix} a_1^1 & \ldots & a_1^j & \ldots & a_1^p \\ \vdots & & \vdots & & \vdots \\ a_i^1 & \ldots & a_i^j & \ldots & a_i^p \\ \vdots & & \vdots & & \vdots \\ a_n^1 & \ldots & a_n^j & \ldots & a_n^p \end{bmatrix}.$$

On note par la suite :

$$\begin{aligned} a_i^j &= [\mathbf{A}]_i^j \text{ le terme général de la matrice,} \\ \mathbf{a}_i &= [a_i^1, \ldots, a_i^p]' \text{ un vecteur-ligne mis en colonne,} \\ \mathbf{a}^j &= [a_1^j, \ldots, a_n^j]' \text{ un vecteur-colonne.} \end{aligned}$$

Types de matrices

Une matrice est dite :

- *vecteur-ligne (colonne)* si $n = 1$ $(p = 1)$,
- *vecteur-unité* d'ordre p si elle vaut $\mathbf{1}_p = [1, \ldots, 1]'$,
- *scalaire* si $n = 1$ et $p = 1$,
- *carrée* si $n = p$.

Une matrice carrée est dite :

- *identité* $(\mathbf{I}_p)$ si $a_i^j = \delta_i^j = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$,

- *diagonale* si $a_i^j = 0$ lorsque $i \neq j$,
- *symétrique* si $a_i^j = a_j^i, \forall (i,j)$,
- *triangulaire* supérieure (inférieure) si $a_i^j = 0$ lorsque $i > j$ $(i < j)$.

Matrice partitionnée en blocs

Matrices dont les éléments sont eux-mêmes des matrices. Exemple :

$$\mathbf{A}(n \times p) = \begin{bmatrix} \mathbf{A}_1^1(r \times s) & \mathbf{A}_1^2(r \times (p-s)) \\ \mathbf{A}_2^1((n-r) \times s) & \mathbf{A}_2^2((n-r) \times (p-s)) \end{bmatrix}.$$

1.2 Opérations sur les matrices

Somme : $[\mathbf{A} + \mathbf{B}]_i^j = a_i^j + b_i^j$ pour $\mathbf{A}$ et $\mathbf{B}$ de même ordre $(n \times p)$.

Multiplication par un scalaire : $[\alpha \mathbf{A}]_i^j = \alpha a_i^j$ pour $\alpha \in \mathbf{R}$.

Transposition : $[\mathbf{A}']_i^j = a_j^i$, $\mathbf{A}'$ est d'ordre $(p \times n)$.

$$(\mathbf{A}')' = \mathbf{A} \; ; (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \; ; (\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \; ; \begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix}' = \begin{bmatrix} \mathbf{A}_1^{1'} & \mathbf{A}_2^{1'} \\ \mathbf{A}_1^{2'} & \mathbf{A}_2^{2'} \end{bmatrix}.$$

Produit scalaire élémentaire : $a'b = \sum_{i=1}^n a_i b_i$ où a et b sont des vecteurs-colonnes.

Produit : $[\mathbf{AB}]_i^j = a_i' b^j$ avec $\mathbf{A}_{(n \times p)}, \mathbf{B}_{(p \times q)}$ et $\mathbf{AB}_{(n \times q)}$, et pour des matrices par blocs :

$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1^1 & \mathbf{B}_1^2 \\ \mathbf{B}_2^1 & \mathbf{B}_2^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^1\mathbf{B}_1^1 + \mathbf{A}_1^2\mathbf{B}_2^1 & \mathbf{A}_1^1\mathbf{B}_1^2 + \mathbf{A}_1^2\mathbf{B}_2^2 \\ \mathbf{A}_2^1\mathbf{B}_1^1 + \mathbf{A}_2^2\mathbf{B}_2^1 & \mathbf{A}_2^1\mathbf{B}_1^2 + \mathbf{A}_2^2\mathbf{B}_2^2 \end{bmatrix}$$

sous réserve de compatibilité des dimensions.

1.3 Propriétés des matrices carrées

La *trace* et le *déterminant* sont des notions intrinsèques, qui ne dépendent pas des bases de représentation choisies, mais uniquement de l'application linéaire sous-jacente.

Trace

Par définition, si $\mathbf{A}$ est une matrice $(p \times p)$,

$$\mathrm{tr}\mathbf{A} = \sum_{j=1}^p a_j^j,$$

et il est facile de montrer :

$$\begin{aligned} \mathrm{tr}\alpha &= \alpha, \\ \mathrm{tr}\alpha\mathbf{A} &= \alpha \mathrm{tr}\mathbf{A}, \\ \mathrm{tr}(\mathbf{A} + \mathbf{B}) &= \mathrm{tr}\mathbf{A} + \mathrm{tr}\mathbf{B}, \\ \mathrm{tr}\mathbf{AB} &= \mathrm{tr}\mathbf{BA}, \\ &\quad \text{reste vrai si } \mathbf{A} \text{ est } (n \times p) \text{ et si } \mathbf{B} \text{ est } (p \times n) \\ \mathrm{tr}\mathbf{CC}' &= \mathrm{tr}\mathbf{C}'\mathbf{C} = \sum_{i=1}^n \sum_{j=1}^p (c_i^j)^2 \\ &\quad \text{dans ce cas, } \mathbf{C} \text{ est } (n \times p). \end{aligned}$$

Déterminant

On note $|\mathbf{A}|$ le *déterminant* de la matrice carrée $\mathbf{A}$ $(p \times p)$. Il vérifie :

$$\begin{aligned}
|\mathbf{A}| &= \prod_{j=1}^{p} a_j^j, \text{ si } \mathbf{A} \text{ est triangulaire ou diagonale,} \\
|\alpha\mathbf{A}| &= \alpha^p |\mathbf{A}|, \\
|\mathbf{AB}| &= |\mathbf{A}||\mathbf{B}|, \\
\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{vmatrix} &= |\mathbf{A}||\mathbf{C}|, \\
\begin{vmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{vmatrix} &= |\mathbf{A}_1^1||\mathbf{A}_2^2 - \mathbf{A}_2^1(\mathbf{A}_1^1)^{-1}\mathbf{A}_1^2| \qquad \text{(A.1)} \\
&= |\mathbf{A}_2^2||\mathbf{A}_1^1 - \mathbf{A}_1^2(\mathbf{A}_2^2)^{-1}\mathbf{A}_2^1|, \qquad \text{(A.2)} \\
&\quad \text{sous réserve de la régularité de } \mathbf{A}_1^1 \text{ et } \mathbf{A}_2^2.
\end{aligned}$$

Cette dernière propriété se montre en considérant les matrices :

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & -\mathbf{A}_1^2(\mathbf{A}_2^2)^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \text{ et } \mathbf{BAB}',$$

puis en comparant les déterminants $|\mathbf{BAB}'|$ et $|\mathbf{A}|$.

Inverse

L'*inverse* de $\mathbf{A}$, lorsqu'elle existe, est la matrice unique notée $\mathbf{A}^{-1}$ telle que :

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}\,;$$

elle existe si et seulement si $|\mathbf{A}| \neq 0$. Quelques propriétés :

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}, \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

Définitions

Une matrice carrée $\mathbf{A}$ est dite :

symétrique si $\mathbf{A}' = \mathbf{A}$,

singulière si $|\mathbf{A}| = 0$,

régulière si $|\mathbf{A}| \neq 0$,

idempotente si $\mathbf{AA} = \mathbf{A}$,

définie-positive si, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{x}'\mathbf{Ax} \geq 0$, et si $x'\mathbf{A}x = 0 \Rightarrow x = 0$,

positive, ou semi-définie-positive, si, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{x}'\mathbf{Ax} \geq 0$,

orthogonale si $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$ $(\mathbf{A}' = \mathbf{A}^{-1})$.

2 Espaces euclidiens

E est un espace vectoriel réel de dimension p isomorphe à $\mathbb{R}^p$.

2.1 Sous-espaces

- Un sous-ensemble E_q de E est un *sous-espace vectoriel* (s.e.v.) de E s'il est non vide et stable :
$$\forall(\mathbf{x},\mathbf{y}) \in E_q^2, \forall\alpha \in \mathbb{R}, \alpha(\mathbf{x}+\mathbf{y}) \in E_q.$$
- Le q-uple $\{\mathbf{x}_1,\ldots,\mathbf{x}_q\}$ de E constitue un système *linéairement indépendant* si et seulement si :
$$\sum_{i=1}^{q} \alpha_i \mathbf{x}_i = 0 \Rightarrow \alpha_1 = \cdots = \alpha_q = 0.$$
- Un système linéairement indépendant $\mathcal{E}_q = \{\mathbf{e}_1,\ldots,\mathbf{e}_q\}$ qui engendre dans E un s.e.v. $E_q = \text{vec}\{\mathbf{e}_1,\ldots,\mathbf{e}_q\}$ en constitue une *base* et $\dim(E_q) = \text{card}(\mathcal{E}_q) = q$.

2.2 Rang d'une matrice $\mathbf{A}_{(n\times p)}$

Dans ce sous-paragraphe, $\mathbf{A}$ est la matrice d'une application linéaire de $E = \mathbb{R}^p$ dans $F = \mathbb{R}^n$.

$$\begin{aligned}
\text{Im}(\mathbf{A}) &= \text{vect}\{\mathbf{a}^1,\ldots,\mathbf{a}^p\} \text{ est le s.e.v. de } F \textit{ image} \text{ de } \mathbf{A}\,;\\
\text{Ker}(\mathbf{A}) &= \{x \in E\,;\ \mathbf{A}x = 0\} \text{ est le s.e.v. de } E \textit{ noyau} \text{ de } \mathbf{A}\,;\\
E &= \text{Im}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A}) \text{ si } \mathbf{A} \text{ est carrée associée à un endomorphisme de } E\\
\text{et } p &= \dim(\text{Im}(\mathbf{A})) + \dim(\text{Ker}(\mathbf{A})).
\end{aligned}$$

$$\begin{aligned}
\text{rang}(\mathbf{A}) &= \dim(\text{Im}(\mathbf{A})),\\
0 \le \text{rang}(\mathbf{A}) &\le \min(n,p),\\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{A}'),\\
\text{rang}(\mathbf{A}+\mathbf{B}) &\le \text{rang}(\mathbf{A}) + \text{rang}(\mathbf{B}),\\
\text{rang}(\mathbf{AB}) &\le \min(\text{rang}(\mathbf{A}), \text{rang}(\mathbf{B})),\\
\text{rang}(\mathbf{BAC}) &= \text{rang}(\mathbf{A}), \text{ si } \mathbf{B} \text{ et } \mathbf{C} \text{ sont régulières},\\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{AA}') = \text{rang}(\mathbf{A}'\mathbf{A}).
\end{aligned}$$

Enfin, si $\mathbf{B}$ $(p \times q)$ est de rang $q (q < p)$ et $\mathbf{A}$ est carrée $(p \times p)$ de rang p, alors la matrice $\mathbf{B}'\mathbf{AB}$ est de rang q.

2.3 Métrique euclidienne

Soit $\mathbf{M}$ une matrice carrée $(p \times p)$, symétrique, définie-positive ; $\mathbf{M}$ définit sur l'espace E :

- un *produit scalaire* : $\langle \mathbf{x},\mathbf{y}\rangle_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{y}$,
- une *norme* : $\|\mathbf{x}\|_{\mathbf{M}} = \langle \mathbf{x},\mathbf{x}\rangle_{\mathbf{M}}^{1/2}$,
- une *distance* : $\text{d}_{\mathbf{M}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_{\mathbf{M}}$,
- des *angles* : $\cos\theta_{\mathbf{M}}(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{y}\rangle_{\mathbf{M}}}{\|\mathbf{x}\|_{\mathbf{M}}\|\mathbf{y}\|_{\mathbf{M}}}$.

La matrice $\mathbf{M}$ étant donnée, on dit que :

- une matrice $\mathbf{A}$ est $\mathbf{M}$-*symétrique* si $(\mathbf{MA})' = \mathbf{MA}$,
- deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ sont $\mathbf{M}$-*orthogonaux* si $\langle \mathbf{x},\mathbf{y}\rangle_{\mathbf{M}} = 0$,
- un vecteur $\mathbf{x}$ est $\mathbf{M}$-*normé* si $\|\mathbf{x}\|_{\mathbf{M}} = 1$,
- une base $\mathcal{E}_q = \{\mathbf{e}_1,\ldots,\mathbf{e}_q\}$ est $\mathbf{M}$-*orthonormée* si
$$\forall(i,j), \langle \mathbf{e}_i,\mathbf{e}_j\rangle_{\mathbf{M}} = \delta_i^j.$$

2.4 Projection

Soit W un sous-espace de E et $\mathcal{B} = \{\mathbf{b}^1, \ldots, \mathbf{b}^q\}$ une base de W ; $\mathbf{P}(p \times p)$ est une matrice de projection $\mathbf{M}$-orthogonale sur W si et seulement si :

$$\forall \mathbf{y} \in E, \mathbf{Py} \in W \text{ et } \langle \mathbf{Py}, \mathbf{y} - \mathbf{Py} \rangle_{\mathbf{M}} = 0.$$

Toute matrice idempotente ($\mathbf{P}^2 = \mathbf{P}$) et $\mathbf{M}$-symétrique ($\mathbf{P}'\mathbf{M} = \mathbf{MP}$) est une matrice de projection $\mathbf{M}$-orthogonale et réciproquement.

Propriétés

- Les valeurs propres de $\mathbf{P}$ sont 0 ou 1 (voir § 3) :

$$\begin{array}{lll} \mathbf{u} \in W, & \mathbf{Pu} = \mathbf{u}, & \lambda = 1, \text{ de multiplicité } \dim(W), \\ \mathbf{v} \perp W, (\text{on note } \mathbf{v} \in W^{\perp}) & \mathbf{Pv} = 0, & \lambda = 0, \text{ de multiplicité } \dim(W^{\perp}). \end{array}$$

- $\mathrm{tr}\mathbf{P} = \dim(W)$.
- $\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{MB})^{-1}\mathbf{B}'\mathbf{M}$, où $\mathbf{B} = \left[\mathbf{b}^1, \ldots, \mathbf{b}^q\right]$.
- Dans le cas particulier où les $\mathbf{b}^j$ sont $\mathbf{M}$-orthonormés :

$$\mathbf{P} = \mathbf{BB}'\mathbf{M} = \sum_{i=1}^{q} \mathbf{b}^j \mathbf{b}^{j'} \mathbf{M}.$$

- Dans le cas particulier où $q = 1$ alors :

$$\mathbf{P} = \frac{\mathbf{bb}'}{\mathbf{b}'\mathbf{Mb}}\mathbf{M} = \frac{1}{\|\mathbf{b}\|_{\mathbf{M}}}\mathbf{bb}'\mathbf{M}.$$

- Si $\mathbf{P}_1, \ldots, \mathbf{P}_q$ sont des matrices de projection $\mathbf{M}$-orthogonales alors la somme $\mathbf{P}_1 + \cdots + \mathbf{P}_q$ est une matrice de projection $\mathbf{M}$-orthogonale si et seulement si : $\mathbf{P}_k\mathbf{P}_j = \delta_k^j \mathbf{P}_j$.
- La matrice $\mathbf{I} - \mathbf{P}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $W^{\perp}$.

3 Eléments propres

Soit $\mathbf{A}$ une matrice carrée $(p \times p)$.

3.1 Définitions

- Par définition, un vecteur $\mathbf{v}$ définit une *direction propre* associée à une *valeur propre* λ si l'on a :

$$\mathbf{Av} = \lambda \mathbf{v}.$$

- Si λ est une valeur propre de $\mathbf{A}$, le noyau $\mathrm{Ker}(\mathbf{A} - \lambda\mathbf{I})$ est un s.e.v. de E, appelé sous-espace propre, dont la dimension est majoré par l'ordre de multiplicité de λ. Comme cas particulier, $\mathrm{Ker}(\mathbf{A})$ est le sous-espace propre associé, si elle existe, à la valeur propre nulle.
- Les valeurs propres d'une matrice $\mathbf{A}$ sont les racines, avec leur multiplicité, du *polynôme caractéristique* :

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

THÉORÈME A.1. — *Soit deux matrices $\mathbf{A}(n \times p)$ et $\mathbf{B}(p \times n)$; les valeurs propres non nulles de $\mathbf{AB}$ et $\mathbf{BA}$ sont identiques avec le même degré de multiplicité. Si $\mathbf{u}$ est vecteur propre de $\mathbf{BA}$ associé à la valeur propre λ différente de zéro, alors $\mathbf{v} = \mathbf{Au}$ est vecteur propre de la matrice $\mathbf{AB}$ associé à la même valeur propre.*

Les applications statistiques envisagées dans ce cours ne s'intéressent qu'à des types particuliers de matrices.

THÉORÈME A.2. — *Une matrice* $\mathbf{A}$ *réelle symétrique admet* p *valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base orthonormée de* E *;* $\mathbf{A}$ *se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k'}$$

où $\mathbf{V}$ *est une matrice orthogonale* $[\mathbf{v}^1, \ldots, \mathbf{v}^p]$ *des vecteurs propres orthonormés associés aux valeurs propres* λ_k*, rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}$*.*

THÉORÈME A.3. — *Une matrice* $\mathbf{A}$ réelle $\mathbf{M}$-symétrique *admet* p *valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base* $\mathbf{M}$*-orthonormée de* E *;* $\mathbf{A}$ *se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}V'\mathbf{M} = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M}$$

où $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^p]$ *est une matrice* $\mathbf{M}$*-orthogonale* ($\mathbf{V}'\mathbf{M}\mathbf{V} = \mathbf{I}_p$ *et* $\mathbf{V}\mathbf{V}' = \mathbf{M}^{-1}$) *des vecteurs propres associés aux valeurs propres* λ_k*, rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}$*.*

Les décompositions ne sont pas uniques : pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que pour une valeur propre multiple, une infinité de bases $\mathbf{M}$-orthonormées peuvent être extraites du sous-espace propre unique associé.

Le rang de $\mathbf{A}$ est aussi le rang de la matrice $\mathbf{\Lambda}$ associée et donc le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.

Par définition, si $\mathbf{A}$ est positive, on note la racine carrée de $\mathbf{A}$:

$$\mathbf{A}^{1/2} = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}'\mathbf{M}.$$

3.2 Propriétés

Si $\lambda_k \neq \lambda_j$,	$\mathbf{v}^k \perp_{\mathbf{M}} \mathbf{v}^j$;
tr$\mathbf{A} = \sum_{k=1}^{p} \lambda_k$;	$\lvert\mathbf{A}\rvert = \prod_{k=1}^{p} \lambda_k$;
si $\mathbf{A}$ est régulière,	$\forall k, \lambda_k \neq 0$;
si $\mathbf{A}$ est positive,	$\lambda_p \geq 0$;
si $\mathbf{A}$ est définie-positive,	$\lambda_p > 0$;

3.3 Décomposition en Valeurs Singulières (DVS)

Il s'agit, cette fois, de construire la décomposition d'une matrice $\mathbf{X}(n \times p)$ rectangulaire relativement à deux matrices symétriques et positives $\mathbf{D}(n \times n)$ et $\mathbf{M}(p \times p)$.

THÉORÈME A.4. — *Une matrice* $\mathbf{X}$ $(n \times p)$ *de rang* r *peut s'écrire :*

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' = \sum_{k=1}^{r} \sqrt{\lambda_k} \mathbf{u}^k \mathbf{v}^{k'}; \tag{A.3}$$

$\mathbf{U}$ $(n\times r)$ *contient les vecteurs propres* $\mathbf{D}$*-orthonormés* $(\mathbf{U}'\mathbf{D}\mathbf{U}=\mathbf{I}_r)$ *de la matrice* $\mathbf{D}$*-symétrique positive* $\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}$ *associés aux* r *valeurs propres non nulles* λ_k *rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}(r\times r)$*;* $\mathbf{V}$ $(p\times r)$ *contient les vecteurs propres* $\mathbf{M}$*-orthonormés* $(\mathbf{V}'\mathbf{M}\mathbf{V}=\mathbf{I}_r)$ *de la matrice* $\mathbf{M}$*-symétrique positive* $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$ *associés aux mêmes valeurs propres. De plus,*

$$\mathbf{U}=\mathbf{X}\mathbf{M}\mathbf{V}\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{V}=\mathbf{X}'\mathbf{D}\mathbf{U}\mathbf{\Lambda}^{-1/2}.$$

4 Optimisation

4.1 Norme d'une matrice

L'espace vectoriel E de dimension p (resp. F de dimension n) est muni de sa base canonique et d'une métrique de matrice $\mathbf{M}$ (resp. $\mathbf{D}$). Soit $\mathbf{X}$ une matrice $(n\times p)$. L'ensemble $\mathcal{M}_{n,p}$ des matrices $(n\times p)$ est un espace vectoriel de dimension np ; on le munit du *produit scalaire* :

$$\langle\mathbf{X},\mathbf{Y}\rangle_{\mathbf{M},\mathbf{D}}=\mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{Y}'\mathbf{D}. \tag{A.4}$$

Dans le cas particulier où $\mathbf{M}=\mathbf{I}_p$ et $\mathbf{D}=\mathbf{I}_n$, et en notant $\mathrm{vec}(\mathbf{X})=\left[\mathbf{x}^{1'},\ldots,\mathbf{x}^{p'}\right]'$ la matrice "vectorisée", ce produit scalaire devient :

$$\langle\mathbf{X},\mathbf{Y}\rangle_{\mathbf{I_p},\mathbf{I_n}}=\mathrm{tr}\mathbf{X}\mathbf{Y}'=\sum_{i=1}^{n}\sum_{j=1}^{p}x_i^jy_i^j=\mathrm{vec}(\mathbf{X})'\mathrm{vec}(\mathbf{Y}).$$

La *norme* associée à ce produit scalaire (A.4) est appelée norme trace :

$$\begin{aligned}\|\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} &= \mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D},\\ \|\mathbf{X}\|^2_{\mathbf{I_p},\mathbf{I_n}} &= \mathrm{tr}\mathbf{X}\mathbf{X}'=\mathrm{SSQ}(\mathbf{X})=\sum_{i=1}^{n}\sum_{j=1}^{p}(x_i^j)^2\\ &\quad\text{(SSQ signifie "sum of squares").}\end{aligned}$$

La *distance* associée à cette norme devient, dans le cas où $\mathbf{D}$ est une matrice diagonale ($\mathbf{D}=\mathrm{diag}(w_1,\ldots,w_n)$), le critère usuel des *moindres carrés* :

$$\mathrm{d}^2(\mathbf{X},\mathbf{Y})=\|\mathbf{X}-\mathbf{Y}\|^2_{\mathbf{M},\mathbf{D}}=\sum_{i=1}^{n}w_i\left\|\mathbf{x}_i-\mathbf{y}_i\right\|^2_{\mathbf{M}}.$$

4.2 Approximation d'une matrice

Les matrices $\mathbf{X}$, $\mathbf{M}$ et $\mathbf{D}$ sont définies comme ci-dessus ; $\mathbf{X}$ est supposée de rang r. On cherche la matrice $\mathbf{Z}_q$, de rang q inférieur à r, qui soit la plus proche possible de $\mathbf{X}$.

THÉORÈME A.5. — *La solution du problème :*

$$\min_{\mathbf{Z}}\left\{\|\mathbf{X}-\mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}}\,;\mathbf{Z}\in\mathcal{M}_{n,p},\ \mathit{rang}(\mathbf{Z})=q<r\right\} \tag{A.5}$$

est donnée par la somme des q *premiers termes de la décomposition en valeurs singulières (A.3) de* $\mathbf{X}$ *:*

$$\mathbf{Z}_q=\sum_{k=1}^{q}\sqrt{\lambda_k}\mathbf{u}^k\mathbf{v}^{k'}=\mathbf{U}_q\mathbf{\Lambda}_q^{1/2}\mathbf{V}_q'.$$

Le minimum atteint est :

$$\|\mathbf{X} - \mathbf{Z}_q\|^2_{\mathbf{M},\mathbf{D}} = \sum_{k=q+1}^{r} \lambda_k.$$

Les matrices $\mathbf{U}_q, \mathbf{\Lambda}_q$ et $\mathbf{V}_q$ contiennent les q premiers vecteurs et valeurs propres donnés par la DVS de $\mathbf{X}$; $\mathbf{Z}_q$ est appelée approximation de rang q de $\mathbf{X}$.

Ce théorème peut se reformuler d'une manière équivalente. On note $\widehat{\mathbf{P}_q}$ (resp. $\widehat{\mathbf{Q}_q}$) la projection $\mathbf{M}$-orthogonale sur $E_q = \text{ Im}(\mathbf{V}_q)$ (resp. $\mathbf{D}$-orthogonale sur $F_q = \text{ Im}(\mathbf{U}_q)$) :

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \sum_{k=1}^{q} \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M} = \mathbf{V}_q \mathbf{V}_q' \mathbf{M} \\
\widehat{\mathbf{Q}_q} &= \sum_{k=1}^{q} \mathbf{u}^k \mathbf{u}^{k'} \mathbf{D} = \mathbf{U}_q \mathbf{U}_q' \mathbf{D}, \\
\mathbf{Z}_q &= \widehat{\mathbf{Q}_q} \mathbf{X} = \mathbf{X} \widehat{\mathbf{P}_q}'.
\end{aligned}$$

PROPOSITION A.6. — *Avec les notations précédentes :*

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \textit{arg } \max_{\mathbf{P}_q} \left\{ \|\mathbf{X}\mathbf{P}_q'\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\quad \left. \mathbf{P}_q \textit{ projection } \mathbf{M}\textit{-orthogonale de rang } q < r \right\}, \\
\widehat{\mathbf{Q}_q} &= \textit{arg } \max_{\mathbf{Q}_q} \left\{ \|\mathbf{Q}_q\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\quad \left. \mathbf{Q}_q \textit{ projection } \mathbf{D}\textit{-orthogonale de rang } q < r \right\}.
\end{aligned}$$

Chapitre B

Cadre fonctionnel

Cette annexe fournit une introduction sommaire au cadre mathématique nécessaire à l'étude de courbes. Un premier objectif est de définir les notations nécessaires à la manipulation de variables ou processus aléatoires à valeurs dans un espace fonctionnel. Incontournables pour des études asymptotiques, ces notions peuvent être survolées en première lecture. Le deuxième objectif est de définir des critères de *régularité* d'une fonction qui interviendront comme contraintes dans les optimisations ou termes de pénalisation. Ils s'exprimeront pratiquement par l'explicitation matricielle de normes ou semi-normes dans un espace euclidien de dimension finie.

0.3 Variable aléatoire hilbertienne

On considère une variable aléatoire Z à valeurs dans l'espace de Hilbert supposé séparable H muni de la tribu des boréliens. On note $\|x\|_H$ la norme dans cet espace. On suppose que Z est du second ordre c'est-à-dire qu'elle vérifie $\mathbb{E}\,\|Z\|_H^2 < \infty$. Sous cette hypothèse, Z admet une espérance dans H notée $\mathbb{E}(Z)$ et un opérateur de covariance compact Γ admettant donc un spectre discret.

L'existence des moments de Z et leur définition sont fournies par le théorème de Riesz (H' désigne le dual topologique de H) :

$$\begin{aligned} \forall f \in H, \qquad & \langle \mathbb{E}(Z), f\rangle_{\mathbf{H}} = \mathbb{E}[\langle f, Z\rangle_{\mathbf{H}}], \\ \forall (u,v) \in H' \times H', \qquad & \langle \Gamma u, v\rangle_{\mathbf{H},\mathbf{H}'} = \mathbb{E}\left[\langle Z - \mathbb{E}(Z), u\rangle_{\mathbf{H},\mathbf{H}'} \langle Z - \mathbb{E}(Z), v\rangle_{\mathbf{H},\mathbf{H}'}\right]. \end{aligned}$$

L'opérateur de covariance s'écrit alors avec la notation de produit tensoriel :

$$\Gamma = \mathbb{E}\left[(Z - \mathbb{E}(Z)) \otimes (Z - \mathbb{E}(Z))\right].$$

Dans le cas particulier où $H = L^2(T)$, il est facile de vérifier que l'opérateur de covariance s'écrit sous la forme d'un opérateur intégral :

$$\forall f \in L^2(T),\ \forall t \in T, \quad \Gamma f(t) = \int_T \gamma(s,t) f(s)\, ds, \tag{B.1}$$

où $\gamma(s,t)$ est la fonction de covariance du processus à temps continu $Z(t)$ d'espérance $a(t)$:

$$\gamma(s,t) = \mathbb{E}\left[(Z(t) - a(t))(Z(s) - a(s))\right], \quad \forall (s,t) \in T \times T. \tag{B.2}$$

Nous serons également amenés à considérer un processus à temps discret $(Z_i)_{i\in\mathbb{Z}}$ supposé du second ordre, auto-régressif d'ordre 1 et prenant ses valeurs dans un espace hilbertien : $(Z_i)_{i\in\mathbb{Z}}$

est dit ARH(1). Notons Γ l'opérateur de covariance et Δ celui de covariance croisée du processus. Le processus étant supposé stationnaire, ces opérateurs ne dépendent pas de i et sont définis par :

$$\begin{aligned}\Gamma &= \mathbb{E}\left[(Z_0 - \mathbb{E}(Z_0)) \otimes (Z_0 - \mathbb{E}(Z_0))\right], \\ \Delta &= \mathbb{E}\left[(Z_0 - \mathbb{E}(Z_0)) \otimes (Z_1 - \mathbb{E}(Z_1))\right], \\ \Delta^* &= \mathbb{E}\left[(Z_1 - \mathbb{E}(Z_1)) \otimes (Z_0 - \mathbb{E}(Z_0))\right]\end{aligned}$$

où la fonction μ représente la moyenne du processus $\mathbb{E}(Z_i)$.. Ils vérifient :

$$\begin{aligned}\Delta^*(Z_i - \mu) &= \Gamma(E(Z_{i+1}|Z_i, Z_{i-1}, \ldots) - \mu), \\ \Delta &= \rho\Gamma\end{aligned}$$

et possèdent un spectre discret.

0.4 Condition de régularité

Les différentes techniques proposées reposent sur la recherche de solutions régulières ou lisses au sens d'un critère faisant intervenir les normes des dérivées successives. Ce critère est couramment utilisé pour la définition et la construction des fonctions splines, il se définit comme une semi-norme dans un espace de Sobolev

$$W^m = \{z : z, z', \ldots, z^{(m-1)} \text{ absolument continues }, z^{(m)} \in L^2\}.$$

Pour toute fonction z de W^m, sa régularité est contrôlée par la semi-norme :

$$\|z\|_m^2 = \|D^m z\|_{L^2(T)}^2 = \int_T (z^{(m)}(t))^2 dt. \tag{B.3}$$

Ce critère peut être généralisé à d'autres semi-normes équivalentes (Wahba 1990) en remplaçant l'opérateur D^m par tout opérateur différentiel linéaire faisant au moins intervenir le même ordre m de dérivation et conduisant ainsi à la définition de familles plus générales de splines dites de Tchebicheff.

0.5 Splines de lissage

L'estimation non-paramétrique par lissage spline a donné lieu à une importante littérature : Wahba (1990), Green et Silverman (1994) en fournissent par exemple des présentations détaillées. Plaçons-nous dans le cadre usuel du modèle de régression non paramétrique :

$$\begin{aligned}x_j = z(t_j) + \varepsilon_j \; ; \quad & E(\varepsilon_j) = 0, \; E(\varepsilon_j \varepsilon_k) = \sigma^2 \delta k, \; j, k = 1, ..., p \\ & a \leq t_1 < t_2 < ... < t_p \leq b.\end{aligned} \tag{B.4}$$

où x_j est l'observation bruitée en t_j de la fonction z supposée régulière et lisse : $z \in W^m$.

L'estimation spline $\widehat{z}$ de la fonction z est alors la solution du problème d'optimisation sous contrainte de régularité

$$\min_{z \in W^2} \left\{ \frac{1}{p} \sum_{j=1}^{p} (z(t_j) - x(t_j))^2; \; \|z\|_m^2 < \mathrm{c} \; (\mathrm{c} \; \in \mathbb{R}_+) \right\}. \tag{B.5}$$

En introduisant un multiplicateur de Lagrange, ce problème d'optimisation est équivalent à :

$$\min_{z\in W^2}\left\{\frac{1}{p}\sum_{j=1}^{p}(z(t_j)-x(t_j))^2+\ell\,\|D^m z\|^2_{L^2}\right\}. \tag{B.6}$$

Le paramètre de lissage ℓ, qui dépend directement de c, permet d'effectuer un arbitrage entre la fidélité aux données ($\ell \to 0$) et la régularité de la solution ($\ell \to +\infty$). En régression non paramétrique sa valeur est choisie en minimisant le critère de validation croisée généralisée (GCV, Wabha 1990).

La solution $\widehat{z}$ de ce problème est une fonction polynômiale par morceaux ; polynômes de degré (2m-1) entre deux nœuds t_j et t_{j+1}, de degré $(m-1)$ aux extrémités entre a et t_1 et entre t_p et b. À la limite ($\ell = 0$) la solution est la fonction d'interpolation spline passant par les données observées et minimisant le critère. À l'autre limite (ℓ infini), la solution est une régression polynômiale de degré $(m-1)$ rendant nulle la pénalisation.

La construction explicite dépend alors de la base choisie pour définir le sous-espace S_p des fonctions splines et différentes solutions sont proposées dans la littérature. La plus simple, sur le plan théorique, consiste à utiliser les propriétés d'auto-reproduction de l'espace de Sobolev et donc son noyau (Wahba 1990) comme base. Cette approche, adoptée aussi par Besse et Ramsay (1986), pose des problème numériques lorsque, en pratique, le nombre de nœuds p est grand car elle conduit à l'inversion d'une matrice mal conditionnée. Dans la version simplifiée de ce chapitre, nous nous limitons aux fonctions splines cubiques dites naturelles, c'est-à-dire polynômiales de degré 3 par morceaux ($m = 2$) et dont les dérivées secondes et troisièmes s'annulent aux bornes a et b. L'algorithme de Reisch (Green et Silverman, 1994) conduit alors à la résolution numérique d'un système d'équations tridiagonal par décomposition de Cholesky puis substitution en un nombre d'opérations qui croît linéairement avec p. Il ne pose alors plus de problèmes numériques.

Soit $\mathbf{Q}$ la matrice bande $(p \times (p-2))$ d'éléments

$$q_{j-1,j}=\frac{1}{t_j-t_{j-1}}, q_{j,j}=-\frac{1}{t_j-t_{j-1}}-\frac{1}{t_{j+1}-t_j}, q_{j+1,j}=-\frac{1}{t_{j+1}-t_j},$$

sur les diagonales (0 à l'extérieur) et $\mathbf{R}$ la matrice bande symétrique $(p-2)\times(p-2)$ d'éléments

$$r_{j,j}=\frac{1}{3}(t_{j+1}-t_{j-1}), r_{j+1,j}=r_{j,j+1}=\frac{1}{6}(t_{j+1}-tj).$$

Notons $\mathbf{M}$ la matrice associée au semi produit scalaire de $W^2(T)$ induit sur S_p : si z désigne une fonction de $W^2(T)$ et $\mathbf{z}$ le vecteur de ses valeurs aux nœuds,

$$\int_T z''(t)^2 dt = \|z\|_2^2 = \mathbf{z}'\mathbf{M}\mathbf{z} = \|\mathbf{z}\|^2_{\mathbf{M}}.$$

La matrice $\mathbf{M}$ est définie par $\mathbf{M} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}$.

Formellement, si le vecteur $\mathbf{x}$ contient les p observations aux nœuds t_j, le lissage spline revient à calculer le vecteur

$$\widehat{\mathbf{z}} = \mathbf{A}_\ell \mathbf{x} \quad \text{avec} \quad \mathbf{A}_\ell = (\mathbf{I}+\ell\mathbf{M})^{-1}$$

et où $\mathbf{A}_\ell$ est la matrice de lissage (ou hat matrix). Enfin, la fonction $\widehat{z}$ est obtenue par simple interpolation spline aux valeurs du vecteur $\widehat{\mathbf{z}}$. On remarque que les solutions obtenues ne dépendent pas des positions des bornes a et b de l'intervalle T.

Une autre matrice de produit scalaire $\mathbf{N}$ est nécessaire, il s'agit de celle associant aux vecteurs $\mathbf{y}_1$ et $\mathbf{y}_2$ issus d'un même schéma de discrétisation le produit scalaire dans l'espace $L^2(T)$ entre leur interpolant spline $\widehat{y}_1$ et $\widehat{y}_2$:

$$\mathbf{y}_1'\mathbf{N}\mathbf{y}_2 \;=\; \int_T \hat{y}_1(t)\hat{y}_2(t)\; dt. \tag{B.7}$$

D'une manière générale cette matrice s'obtient à l'aide des noyaux reproduisants associés aux fonctions splines (Besse et Ramsay 1986, Ramsay et Dalzell, 1991). Il est également possible de l'approcher en utilisant une méthode de quadrature. On peut considérer par exemple $\mathbf{N} = \text{diag}(w_1, \ldots, w_p)$ où $w_1 = (t_2 - t_1)/2$, $w_j = (t_{j+1} - t_{j-1})/2,\ \ j = 2, \ldots, p-1$ et $w_p = (t_p - t_{p-1})/2$. Le calcul est alors rapide, stable et généralement suffisamment précis.

Table des matières