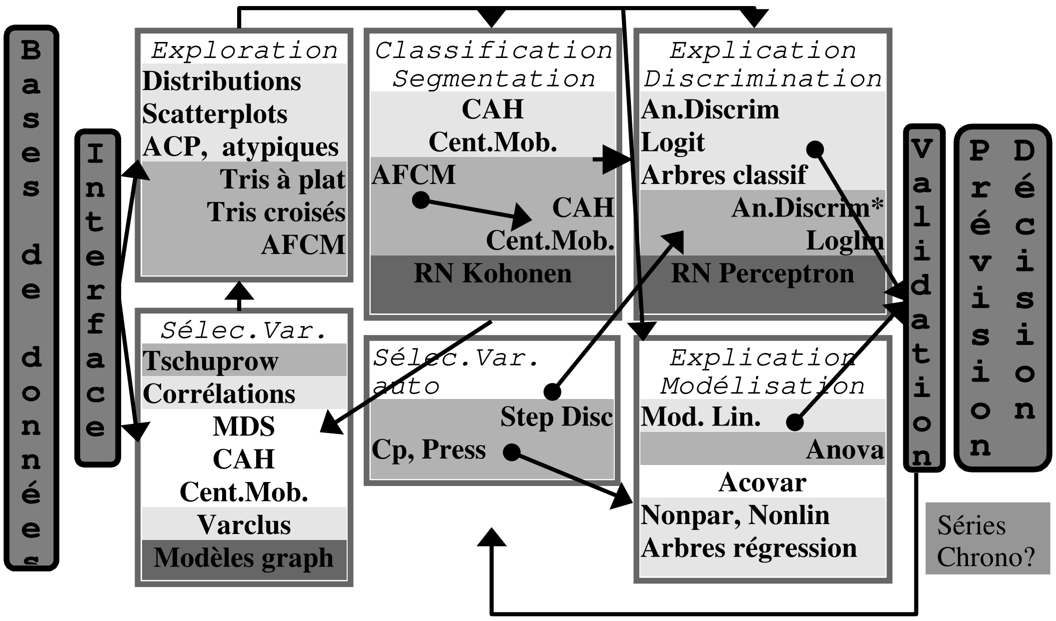
UNIVERSITE PAUL SABATIER
TOULOUSE III

PUBLICATIONS DU LABORATOIRE
DE
STATISTIQUE ET PROBABILITÉS

Statistique
&
Data Mining



PHILIPPE BESSE

Version septembre 2001 — mises à jour : www.lsp.ups-tlse.fr/Besse

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

Table des matières

Introduction **11**

Data mining et métier de statisticien 11

Méthodes 13

Notations **17**

I Exploration **19**

1 Description statistique élémentaire **21**

1 Exemple de données 21

2 Introduction 21

3 Decription d'une variable 23

3.1 Cas quantitatif 23

3.2 Cas qualitatif 25

4 Liaison entre variables 25

4.1 Deux variables quantitatives 26

4.2 Une variable quantitative et une qualitative 28

4.3 Deux variables qualitatives 29

5 Vers le cas multidimensionnel 32

5.1 Matrices des covariances et des corrélations 32

5.2 Tableaux de nuages 32

5.3 La matrice des coefficients de Tschuprow (ou de Cramer) 33

6 Problèmes 33

2 Analyse en Composantes Principales **35**

1 Introduction 35

1.1 Représentation vectorielle de données quantitatives 35

1.2 Interprétation statistique de la métrique des poids 36

1.3 La méthode 36

2 Modèle 37
2.1 Estimation 37
2.2 Définition équivalente 39
3 Représentations graphiques 39
3.1 Les individus 39
3.2 Les variables 42
3.3 Représentation simultanée ou “biplot” 43
4 Choix de dimension 45
4.1 Part d’inertie 45
4.2 Règle de Kaiser 45
4.3 Éboulis des valeurs propres 45
4.4 Boîtes-à-moustaches des variables principales 45
5 Interprétation 47

3 Analyse Factorielle Discriminante 49

1 Introduction 49
1.1 Données 49
1.2 Objectifs 49
1.3 Notations 50
2 Définition 51
2.1 Modèle 51
2.2 Estimation 51
3 Réalisation de l’AFD 52
3.1 Matrice à diagonaliser 52
3.2 Représentation des individus 52
3.3 Représentation des variables 52
3.4 Interprétations 53
4 Variantes de l’AFD 53
4.1 Individus de mêmes poids 53
4.2 Métrique de Mahalanobis 54
5 Exemples 54

4 Analyse Factorielle des Correspondances 59

1 Introduction 59
1.1 Données 59
1.2 Notations 59
1.3 Liaison entre deux variables qualitatives 60
1.4 Objectifs 61

2 Double ACP 61

2.1 Métriques du χ^2 61

2.2 ACP des profils–colonnes 62

2.3 ACP des profils–lignes 62

3 Modèles pour une table de contingence 63

3.1 Le modèle log–linéaire 63

3.2 Le modèle d'association 63

3.3 Le modèle de corrélation 64

3.4 Estimation Moindres Carrés dans le modèle de corrélation 64

4 Représentations graphiques 65

4.1 Biplot 65

4.2 Double ACP 66

4.3 Représentations barycentriques 66

4.4 Autre représentation 67

4.5 Aides à l'interprétation 67

5 Exemple 68

6 Compléments 68

6.1 Propriétés 68

6.2 Invariance 69

6.3 Choix de la dimension q 69

5 Analyse des Correspondances Multiples 71

1 Codages de variables qualitatives 71

1.1 Tableau disjonctif complet 71

1.2 Tableau de Burt 72

1.3 La démarche suivie dans ce chapitre 72

2 AFC du tableau disjonctif complet relatif à 2 variables 72

2.1 Données 72

2.2 ACP des profils–lignes 73

2.3 ACP des profils–colonnes 74

3 AFC du tableau de Burt relatif à 2 variables 75

4 Analyse Factorielle des Correspondances Multiples 76

4.1 Définition 76

4.2 AFC du tableau disjonctif complet $\mathbf{X}$ 76

4.3 AFC du tableau de Burt $\mathcal{B}$ 78

4.4 Variables illustratives 79

4.5 Interprétation 79

5 Exemple . . . 79
5.1 Les données . . . 79
5.2 Analyse brute . . . 80
5.3 Analyse des interactions . . . 80

6 Positionnement multidimensionnel 83
1 Introduction . . . 83
2 Distance, similarités . . . 83
2.1 Définitions . . . 83
3 Distances entre variables . . . 84
3.1 Variables quantitatives . . . 84
3.2 Variables qualitatives . . . 85
3.3 Variables quantitative et qualitative . . . 85
4 Recherche d'une configuration de points . . . 86
4.1 Propriétés . . . 86
5 Exemple . . . 87
6 Application au choix de variables . . . 87

7 Classification 91
1 Introduction . . . 91
2 Classification hiérarchique ascendante . . . 92
3 Centres mobiles . . . 93
4 Combinaison . . . 94

II Modélisation 95

8 Régression linéaire multiple 97
1 Introduction . . . 97
2 Modèle . . . 97
3 Estimation . . . 98
3.1 Estimation par M.C. . . . 98
3.2 Propriétés . . . 99
3.3 Sommes des carrés . . . 99
3.4 Coefficient de détermination . . . 100
4 Inférences dans le cas gaussien . . . 100
4.1 Inférence sur les coefficients . . . 100
4.2 Inférence sur le modèle . . . 101
4.3 Inférence sur un modèle réduit . . . 101

4.4 Prévision . . . 101
5 Sélection de variables, choix de modèle . . . 102
5.1 Critères . . . 102
5.2 Algorithmes de sélection . . . 104
6 Diagnostics et compléments . . . 105
6.1 Multicolinéarité . . . 105
6.2 Modèles curvilinéaires . . . 106
6.3 Influence, résidus, validation . . . 107
7 Analyse de covariance . . . 108
7.1 Modèle . . . 109
7.2 Tests . . . 109
7.3 Cas général . . . 110
8 Exemples . . . 111
8.1 Données comptables . . . 111
8.2 Données marketing . . . 114

9 Régression logistique 117
1 Introduction . . . 117
2 Odds et odds ratio . . . 117
3 Régression logistique . . . 118
3.1 Type de données . . . 118
3.2 Modèle binomial . . . 119
4 Choix de modèle . . . 120
4.1 Recherche pas à pas . . . 120
5 Exemples . . . 120
5.1 Débits$\times$Volumes . . . 121
5.2 Données bancaires . . . 122

10 Analyse Discriminante Décisionnelle 125
1 Introduction . . . 125
2 Règle de décision issue de l'AFD . . . 125
2.1 Cas général : m quelconque . . . 125
2.2 Cas particulier : $m = 2$ . . . 126
3 Règle de décision bayésienne . . . 126
3.1 Introduction . . . 126
3.2 Définition . . . 126
3.3 Coûts inconnus . . . 127
3.4 Détermination des *a priori* . . . 127

3.5 Cas particuliers . . . 127
4 Règle bayésienne avec modèle normal . . . 128
4.1 Hétéroscédasticité . . . 128
4.2 Homoscédasticité . . . 128
4.3 Commentaire . . . 129
5 Règle bayésienne avec estimation non paramétrique . . . 129
5.1 Introduction . . . 129
5.2 Méthode du noyau . . . 129
5.3 k plus proches voisins . . . 130
6 Exemple . . . 130

11 Arbres de classification 133
1 Introduction . . . 133
2 Construction d'un arbre binaire . . . 134
2.1 Principe . . . 134
2.2 Critère de division . . . 134
2.3 Règle d'arrêt . . . 136
2.4 Affectation . . . 136
3 Élagage . . . 136
3.1 Construction de la séquence d'arbres . . . 136
3.2 Recherche de l'arbre optimal . . . 138
4 Exemple des données bancaires . . . 138

12 Méthodes connexionistes 141
1 Historique . . . 141
2 Réseaux de neurones . . . 142
2.1 Neurone formel . . . 142
3 Perceptron multicouche . . . 143
3.1 Architecture . . . 143
3.2 Apprentissage . . . 144
3.3 Utilisation . . . 145

13 Comparaison des modèles 147
1 Estimation des erreurs . . . 147
1.1 Resubstitution . . . 147
1.2 Échantillon de validation . . . 147
2 Validation croisée . . . 148
3 Bootstrap . . . 148

4 Comparaison 149

Conclusion **153**

Stratégies du data mining 153

Pièges 155

Exploration versus décision 155

Complexité 156

Quelles compétences ? 156

III Annexes **161**

A Outils algébriques **163**

1 Matrices 163

1.1 Notations 163

1.2 Opérations sur les matrices 164

1.3 Propriétés des matrices carrées 164

2 Espaces euclidiens 166

2.1 Sous-espaces 166

2.2 Rang d'une matrice $\mathbf{A}_{(n \times p)}$ 166

2.3 Métrique euclidienne 167

2.4 Projection 167

3 Eléments propres 168

3.1 Définitions 168

3.2 Propriétés 169

3.3 Décomposition en Valeurs Singulières (DVS) 169

4 Optimisation 170

4.1 Norme d'une matrice 170

4.2 Approximation d'une matrice 170

B Enterprise miner **173**

1 Introduction 173

2 Organisation 173

2.1 Fenêtres 174

2.2 Menus 174

2.3 Démarrage 174

3 Descriptif sommaire des noeuds 174

3.1 Échantillonnage 175

3.2 Exploration 175

3.3 Transformations 175
3.4 Modélisation 176
3.5 Comparaison, certification 176

Introduction

Le développement des moyens informatiques de stockage (bases de données) et de calcul permet maintenant 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) 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 en données. L'accroche publicitaire souvent citée 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 (40%). Les entreprises sont maintenant très motivées pour tirer parti et amortir les coûts de stockage des teras octets que l'informatique sait maintenant enregistrer.

Data mining et métier de statisticien

Le contexte informationnel du data mining 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 (*data mart*) 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 ou au contraire les plus risquées (assurance) ;
- suivi en ligne des paramètres de production 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 ;
- ...

Il se caractérise par

- un environnement 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.
- 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 faible effectif mais à fort impact économique.

Les premières applications ont été, par exemple, l'étude des séquences d'achats des consommateurs en grande surface à partir des tickets de caisse. Il est ainsi possible de regrouper les articles correspondant fréquemment à des achats simultanés afin de limiter les appels promotionnels.

L'objectif est toujours d'extraire l'information pertinente en vue d'analyses puis de prises de décision. Le problème nouveau est l'articulation entre structure et traitement des données. On parle alors de *Knowledge Discovery in Databases (KDD)* que Fayyad (1997), de Microsoft Research, propose de le définir par

> *KDD is a non trivial process of identifying valid, novel, potentially usefull, and ultimately understandable pattern in data.*

Ce processus se décomposent en différentes étapes, dont certaines classiques du métier de statisticien, sans pour autant que ce "métier" soit clairement identifié, reconnu et budgétisé dans une entreprise :

i. Compréhension du domaine d'application : expliciter la connaissance *a priori* et les buts du commanditaire.
ii. Création d'un sous-ensemble cible des données (à partir de l'entrepôt) dans lequel appliquer la recherche. Les données se présentent alors sous la forme usuelle en Statistique d'un fichier, observations ou unités statistiques en lignes dont chaque champ ou colonne contient les valeurs prises par les variables considérées.
iii. Nettoyage : erreurs, données manquantes, valeurs atypiques (outliers).
iv. Transformation des données : "normalisation", linéarisation, compression.
v. Explicitation de l'objectif et de la stratégie d'analyse : exploration, classification, discrimination, segmentation, recherche de singularités, modélisation, prévision...
vi. Choix des méthodes, des algorithmes en privilégiant interprétabilité ou prédictibilité. Mise en œuvre des outils informatiques appropriés pour aboutir à une modélisation.
vii. Test : sur la base de critères à préciser (qualité d'ajustement, de prévision, simplicité, visualisations graphiques...).
viii. Prévision.
ix. Diffusion de l'information (par intranet) pour prise de décision.

De façon plus précise, le *Data Mining* concerne l'exécution des étapes 3 à 8 ci-dessus. Il nécessite la mise en œuvre, explicite ou non, de méthodes statistiques classiques (descriptives, sondage, réduction de dimension, classification, discrimination, segmentation, modélisation, prévision) ou moins classiques (arbres, modèles graphiques) ou algorithmiques (réseaux neuronaux).

Des éditeurs de logiciels proposent alors des "boîtes noires" munis d'interfaces conviviales et intuitives pouvant masquer certaines questions non résolues :

- Qu'est-ce qui peut être automatisé dans une démarche de K.D.D. sur un sujet donné ?
- Où doit nécessairement intervenir une expertise humaine, et de quel niveau (informaticien, statisticien confirmé, utilisateur) ?
- Quel contrôle, quelle validation doivent être effectués et par qui ?

Méthodes

Ce document propose un tour d'horizon des méthodes usuelles et d'autres moins classiques. Il contient nécessairement des redites concernant des contenus d'enseignements de statistique traditionnels mais il vise à les intégrer dans une démarche globale guidée par des objectifs classiques du data mining. Des précisions sont accessibles à partir de l'URL `www-sv.cict.fr/lsp/Besse` d'autres références comme Celeux et Nakache (1994), Chambers et Hastie (1992) apportent des compléments importants.

On peut schématiquement répartir les méthodes et techniques présentées en deux parties selon qu'elles visent à décrire, réduire, classifier des données ou encore à modéliser, expliquer, prévoir une variable cible :

Exploration statistique

- Des techniques simples résument les caractéristiques (tendance centrale, dispersion) d'une variable, estiment sa densité (boîte à moustaches, histogramme, estimation non paramétrique) et permettent de détecter des valeurs atypiques (outliers) ou encore de faire apparaître la nécessité d'une transformation (log, racine...).
- Elles évaluent également les relations entre couples de variables de même type quantitatif (coefficient de corrélation, nuage de points) ou qualitatif (χ^2, Cramer, Tchuprow) ou de types différents (rapport de corrélation, boîtes à moustaches parallèles).

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 à des valeurs atypiques ou des distributions dissymétriques. Ils constituent un préalable incontournable à une première approche des données.

- Différentes méthodes, selon le type des variables considérées permettent 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 n'est pas connue, les méthodes de classification (hiérarchiques ou par réallocation dynamique) déterminent une variable qualitative définissant une partition de l'ensemble des données. En marketing, on parle alors de *segmentation* lorsqu'il s'agit de la clientèle d'une banque, d'un VPCiste... Cet objectif est repris par les réseaux de neurones avec apprentissage non supervisé dit aussi auto-associatif ou de Kohonen.

– Enfin 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.

Modélisation

Plusieurs méthodes visent à expliquer une variable par combinaison des autres afin de construire un modèle de prévision. Elles diffèrent par le type des variables en présence ainsi que par leur domaine de validité. Pour faire synthétique : un grand volume de données (beaucoup d'individus et donc un très grand nombre de degrés de liberté) permet de s'affranchir d'hypothèses probabilistes sur la nature des distributions tout en conduisant à des méthodes plus algorithmiques et calculatoires.

– Le modèle gaussien pour expliquer une variable quantitative par d'autres variables quantitatives (régression linéaire multiple) ou un mélange de variables quantitatives et qualitatives (analyse de covariance).
– Le modèle binomial (régression logistique) pour expliquer une variable à deux modalités ou plus si elles sont ordonnées.
– Le modèle poissonnien (modèle log-linéaire) dans le cas d'une variable qualitative expliquée par plusieurs variables qualitatives.
– L'analyse discriminante décisionnelle (ADD) vise également à prévoir la classe d'un nouvel individu à partir de la connaissance de variables quantitatives.

Plusieurs méthodes, dont le développement fut directement lié à celui des moyens de calcul, sont récemment apparus en premier plan dans les opérations de médiatisation du data mining. Peu présentes dans les cours de Statistique ou d'Informatique en deuxième cycle, elles sont introduites brièvement et de façon intuitive, sans démonstration, dans la suite de ce document.

– Les arbres de classification et de régression (CART) poursuivent certains des objectifs des méthodes précédentes. Modélisation et prévision d'une variable qualitative (classification tree) ou quantitative (regression tree). Complémentaires des précédentes, elles apportent de plus une représentation graphique sous forme d'*arbre de décision* à l'interprétation très intuitive.
– Les méthodes connexionnistes sont basées sur le principe d'une simulation du fonctionnement du cerveau humain par des réseaux ou graphes de neurones formels. De conception ancienne, les prémices datent de 1943, le développement de la puissance de calcul a donné plus de potentialité a cette approche qui connaît des développements commerciaux importants depuis une dizaine d'année. Très souples et adaptatives, elles diffèrent entre elles par la structure du graphe, les propriétés des neurones formels et aussi par le mode d'apprentissage du réseau. Celui-ci peut être supervisé avec des applications principalement en discrimination ou non supervisé avec des applications en classification.

Tout au long de ce cours nous verrons que le principal enjeux est la recherche d'un *modèle parcimonieux* : moins de paramètres pour plus de précision des prévisions. Il est atteint par la recherche du meilleur compromis biais/variance par l'optimisation d'un critère : validation croisée, C_p de Mallows, Akaïke...

D'autres méthodes

Enfin, citons pour mémoire des méthodes et techniques qui connaissent des développements rapides :

- Les *algorithmes génétiques* sont souvent décrits dans les livres traitant de l'application du data mining en marketing. Cette méthode d'optimisation stochastique cherche un optimum global dans des cas où la fonction à optimiser n'est pas convexe. Même si elle peut apparaître dans des applications de transport ou d'algorithme d'apprentissage d'un réseau de neurones, ses objectifs sont très différents des autres méthode présentées et, curieusement, le *recuit simulé* qui est aussi une méthode d'optimisation stochastique tout aussi efficace reste dans l'ombre. Effet d'annonce ou de mode ?
- Les techniques de prévision de séries chronologiques par des modèles de type SARIMA, GARCH ou encore par des techniques non paramétriques (méthode du noyau) très efficaces lorsque les observations sont nombreuses. Elles font l'objet de cours spécifiques mais doivent prendre naturellement leur place dans le processus de modélisation/prévision.
- Les techniques de sondage (Ardilly, 1994) indispensables lorsque les données sont trop volumineuses pour envisager un traitement exhaustif avec certaines méthodes. Il est aussi préférable, en terme de qualité des prévisions, de travailler sur un échantillon représentatif de qualité, en ayant pris le temps d'un recueil soigneux, que de vouloir traiter beaucoup de données à la fiabilité douteuse. C'est encore un cours spécifique.
- Les modèles non linéaires paramétriques, semi-paramétriques, additifs (GAM) (Hastie et Tibshirani 1990, Chambers et Hastie 1992) peuvent s'adapter à des situations complexes par la recherche de transformations non linéaires des variables initiales. Il existe le même type d'extension non linéaire pour les méthodes factorielles (ACP curvilinéaire, Besse et Ferraty 1995).
- Les modèles d'indépendance conditionnelle ou modèles graphiques (Droesbeke 1992) qui, pouvant être estimés de façon adaptative, sont bien adaptés à des saisies automatiques de données volumineuses.
- Les méthodes adaptées aux données fonctionnelles (Ramsay et Silverman 1997, Besse et col. 1997) c'est-à-dire à des mesures observées sous forme de courbes ou de courbes discrétisées plutôt que sous forme de simples valeurs scalaires ou vectorielles.
- D'autres ont sûrement été oubliées...

Ce cours ne permet que d'entrevoir la complexité des problèmes soulevés. Il n'existe pas de stratégie gagnante à tous les coups, les choix en présence sont rarement sous-tendus par des critères théoriques bien justifiés dans les conditions pratiques d'utilisation. Une dernière partie tente donc d'expliciter les stratégies du data mining et les procédures de test et validation permettant d'optimiser les choix qui se présentent. Enfin, pour améliorer la cohérence de l'ensemble et la lisibilité de la démarche, un même jeu de données sert de fil conducteur pour illustrer les potentialités des outils présentés quand ils sont bien adaptés.

L'expérience et le savoir faire sont incontournables et, même si le prospecteur en données peut éviter d'aller au charbon, il n'échappera pas au cambouis !

Notations

Le recensement de techniques d'origines très diverses rend difficiles l'homogénéisation et la cohérence des notations et même de celles du vocabulaire. Chaque sous-discipline a ses habitudes ainsi, une *entrée*, un *poids*, une *sortie*, d'un réseau neuronal sont respectivement une variable statistique explicative, un paramètre de modèle, une variable à expliquer. Néanmoins, voici quelques conventions que nous allons essayer de respecter tout au long de ce cours.

X^j, j variant de 1 à p, désigne une variable statistique, généralement observée sur une population ou un échantillon de n individus ou unités statistiques de poids w_i. Les observations sont rangées dans une matrice[1] (n,p) notée $\mathbf{X}$ et $\overline{\mathbf{X}}$ la matrice centrée.

Y une variable quantitative à expliquée, T une variable qualitative à expliquée.

$\mathbf{M}$: matrice symétrique définie positive associée au produit scalaire dans l'espace des individus.

$\mathbf{D}$: matrice diagonale des pondérations w_i des observations.

$\mathbf{S}$: matrice de covariance empirique.

$\mathbf{\Lambda}_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$: matrice de valeurs propres restreinte aux q plus grandes et $\mathbf{V}_q$ la matrice des vecteurs propres associés.

[1]Les **matrices** et les **vecteurs** sont notés en gras pour les distinguer des fonctions.

Première partie

Exploration

Chapitre 1

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 **??**. 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, boîte-à-moustaches, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.

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

TAB. 1.1 – 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**

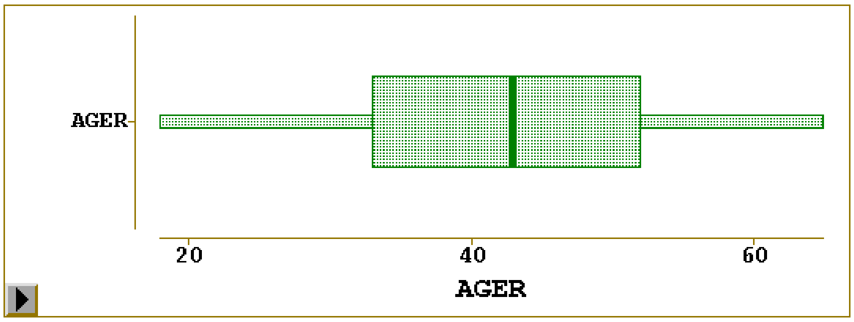


FIG. 1.1 – Boîte à moustaches illustrant la distribution des âges des clients.

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 boîtes à moustaches parallèles, les diagrammes de profils ou encore en mosaïque.

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).

Boîte-à-moustaches (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.

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

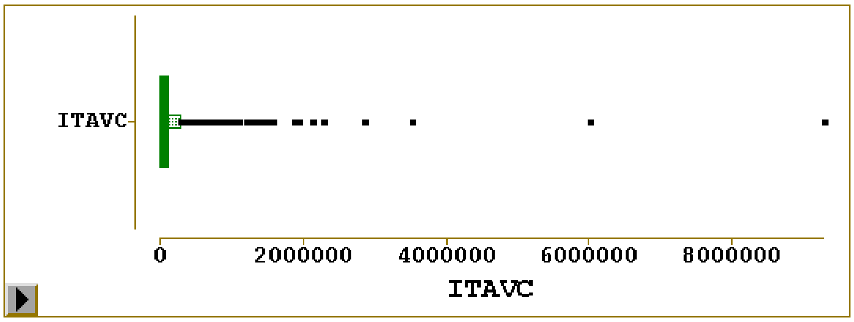


FIG. 1.2 – Boîte à moustaches illustrant la distribution de la variable cumulant les totaux des avoirs. Celle-ci apparaît comme très dissymétrique et avec de nombreuses valeurs atypiques. Une transformation s'impose.

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 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

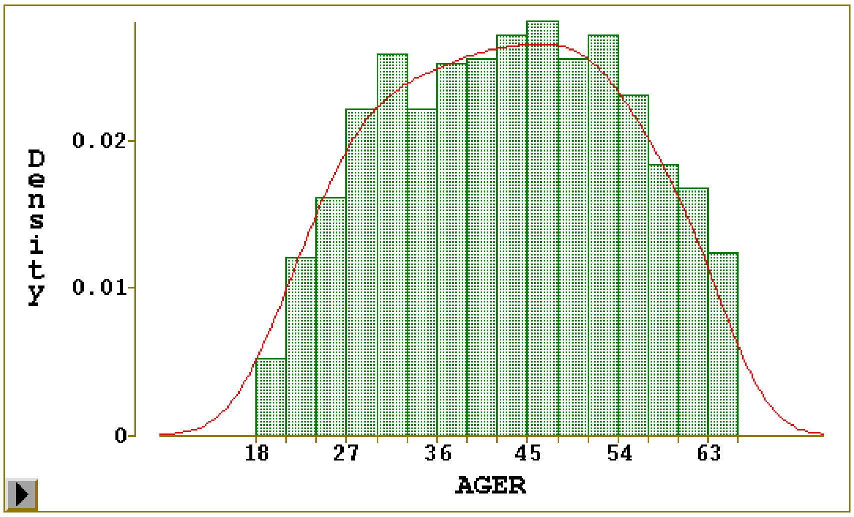


FIG. 1.3 – Histogramme et estimation fonctionnelle par la méthode du noyau de la distribution des âges.

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.

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

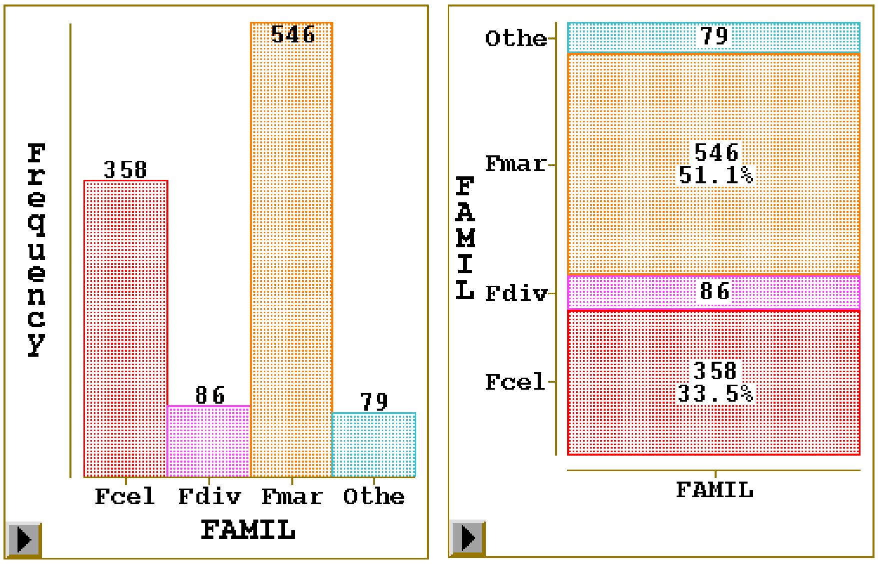


Fig. 1.4 – Diagramme en barres et diagramme en colonne de la répartition des situations familiales. Certaines modalités trop rares et regroupées automatiquement dans la classe *other* devront être recodées.

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*), boîtes-à-moustaches parallèles, diagramme de profils, tableau de nuages (*scatter-plot matrix*).

4.1 Deux variables quantitatives

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

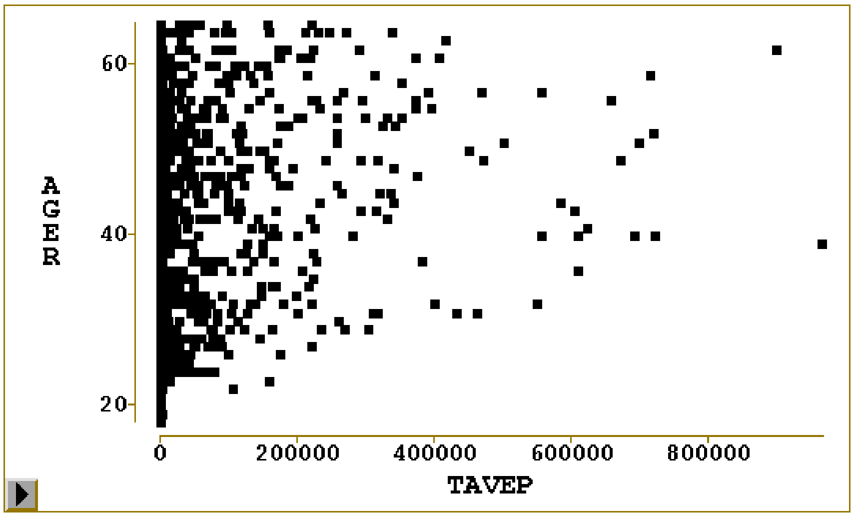


FIG. 1.5 – Nuage de points illustrant l'absence de liaison entre la variable âge et celle cumulant le total des épargnes monétaires (corrélation de 0,17).

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).

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 :

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

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 :

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

Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $\operatorname{corr}(X,Y) = \operatorname{cov}(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$. Par conséquent, $\operatorname{corr}(X,Y)$ 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.

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 $n = card(\Omega)$).

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 boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y une boîte-à-moustaches 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).

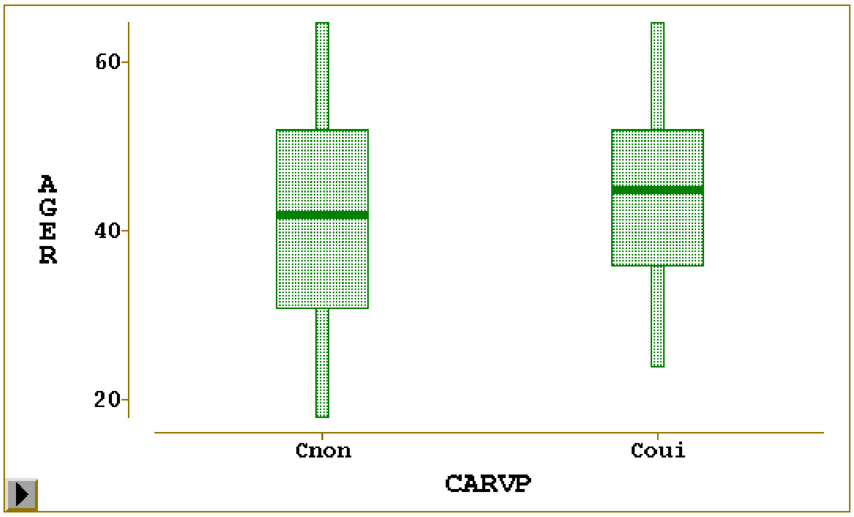


FIG. 1.6 – Boites à moustaches illustrant les différences de distribution des âges en fonction de la possession d'une carte Visa Premier.

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*.

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.

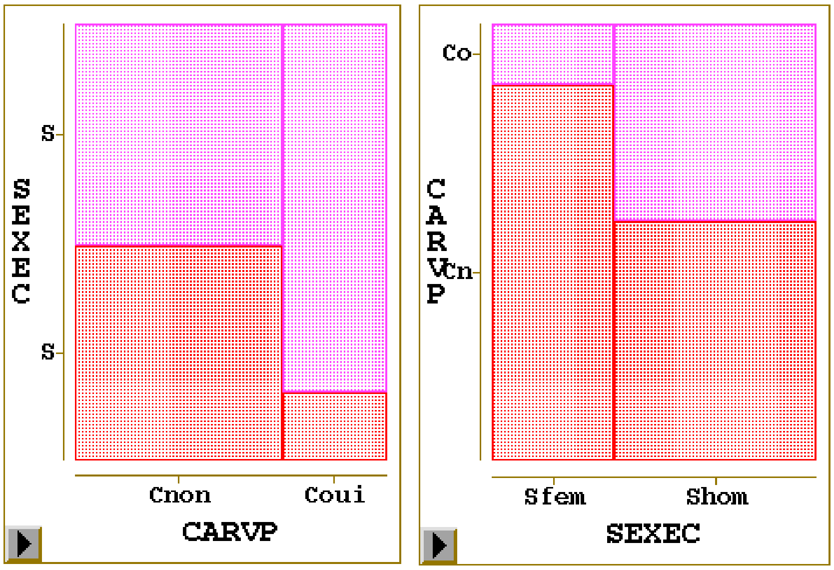


FIG. 1.7 – Diagrammes en barres des profils lignes et colonnes de la table de contingence croisant le sexe et la possession de la carte Visa Premier. La superficie de chaque case est en plus proportionnelle à l'effectif de la cellule associée.

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 : $0 \leq T \leq C \leq 1$.

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 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.

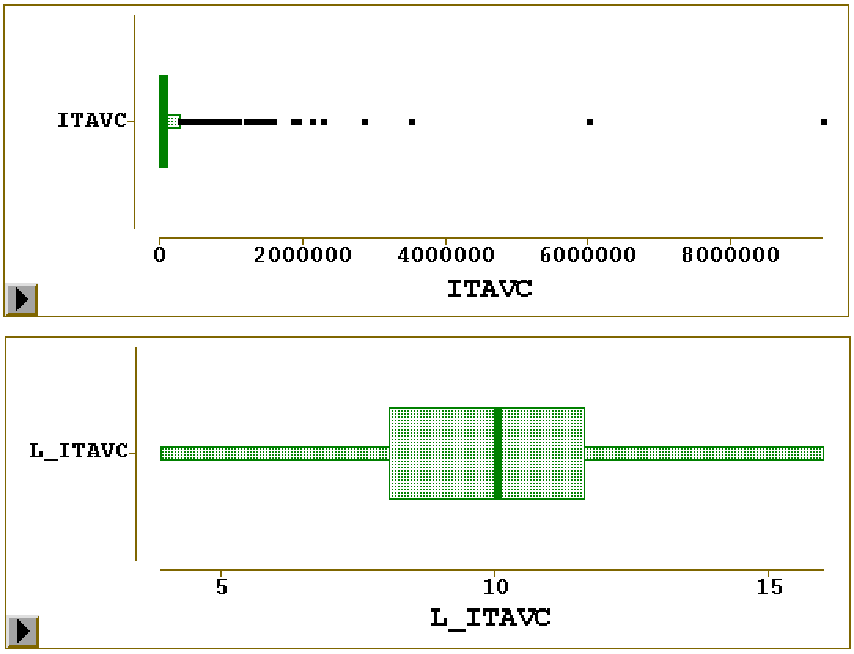


FIG. 1.8 – La simple transformation ($\log(50+x)$), de la variable cumulants les avoirs, résout bien les problèmes posés par l'allure "log-normale" de sa distribution avec son cortège de valeurs atypiques.

Chapitre 2

Analyse en Composantes Principales

L'analyse en Composantes Principales (ACP) est introduite ici comme l'estimation au sein d'un modèle, afin de préciser la signification statistique des résultats obtenus. Cette technique 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.

1 Introduction

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 :

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

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

1.1 Représentation vectorielle de données quantitatives

- 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 struc-

ture d'espace euclidien : F est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \mathrm{diag}(w_1, \ldots, w_n)$; F est alors appelé *espace des variables*.

1.2 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 :	$\overline{x^j}$	$=$	$\langle \mathbf{X}e^j, \mathbf{1}_n \rangle_{\mathbf{D}} = e^{j\prime}\mathbf{X}'\mathbf{D}\mathbf{1}_n$.
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	$=$	$(\mathbf{x}^{j\prime}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$.
Covariance de X^j et X^k :	$\mathbf{x}^{j\prime}\mathbf{D}\mathbf{x}^k$	$=$	$\langle \mathbf{x}^j, \mathbf{x}^k \rangle_{\mathbf{D}}$.
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*.

1.3 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.

2 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 $\mathbf{D} = \operatorname{diag}(w_1, \ldots, w_n)$.

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 :

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

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*).

2.1 Estimation

PROPOSITION 2.1. — *L'estimation des paramètres de (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 :

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

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

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \;; \mathbf{Z} \in \mathcal{M}_{n,p},\ \operatorname{rang}(\mathbf{Z}) = q \right\}. \tag{3}$$

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 :

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

Remarques

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

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

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 :

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

$\widetilde{\mathbf{S}}$ 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.

2.2 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 2.2. — *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* ;

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

- 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*.

3 Représentations graphiques

3.1 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 2.3. — *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.*

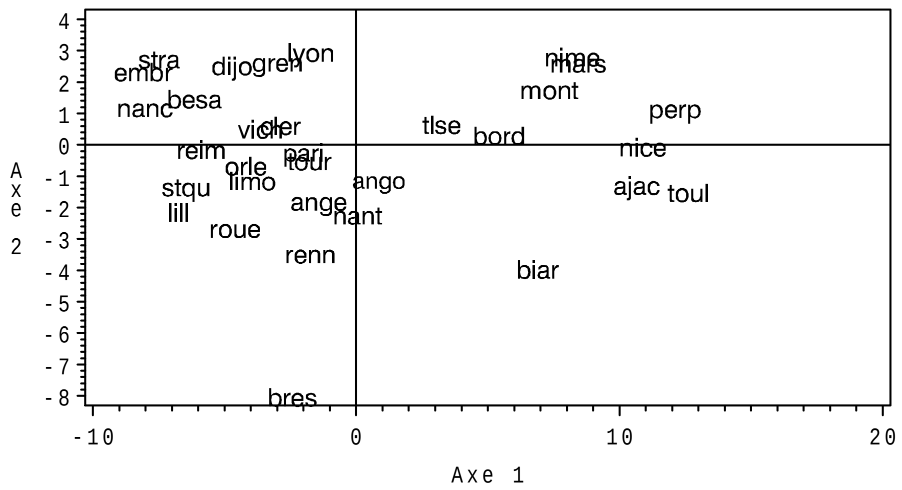


FIG. 2.1 – *Températures : premier plan des individus.*

Mesures de "qualité"

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

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

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 :

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

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

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

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

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

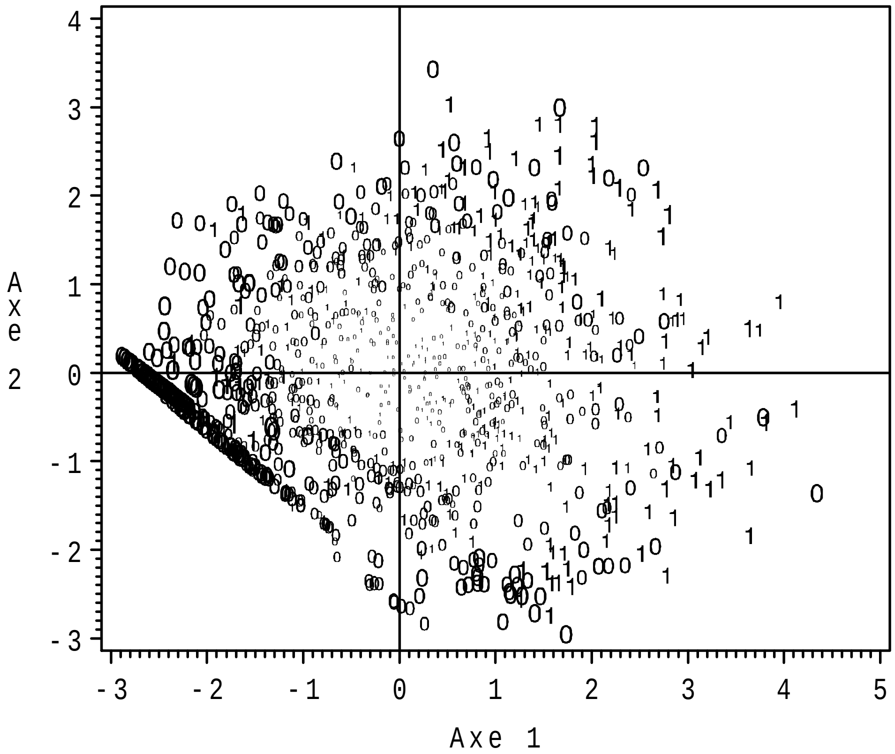


FIG. 2.2 – *Carte Visa : premier plan de l'ACP d'une sélection de variables. La discrimination des individus possèdant (1) ou non (0) la carte Visa premier n'est pas très claire sur cette représentation.*

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 boîtes-à-moustaches 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, 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 :

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

3.2 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 2.4. — *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 :

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

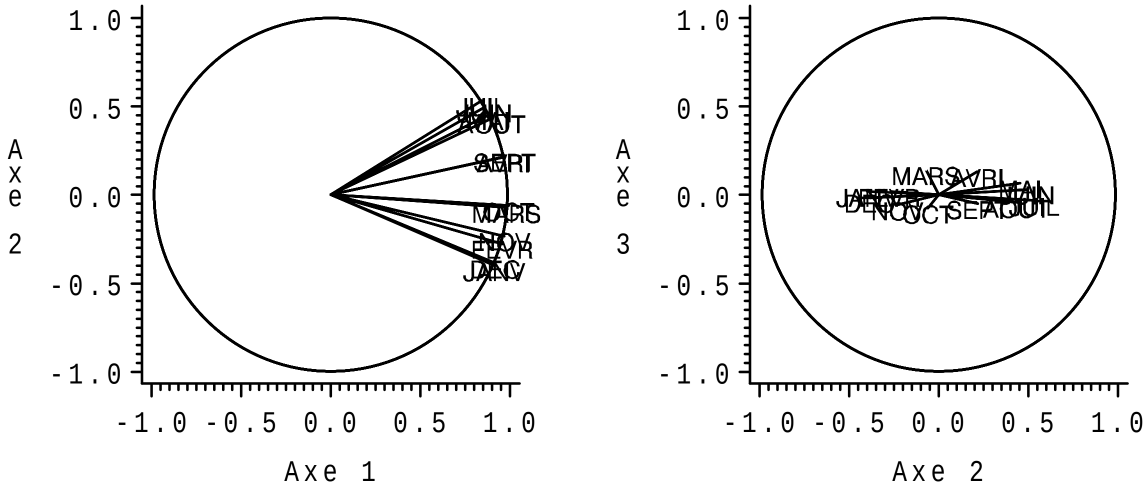


FIG. 2.3 – *Températures : Premier et deuxième plan des variables.*

Corrélations variables×facteurs

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

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

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.

3.3 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

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

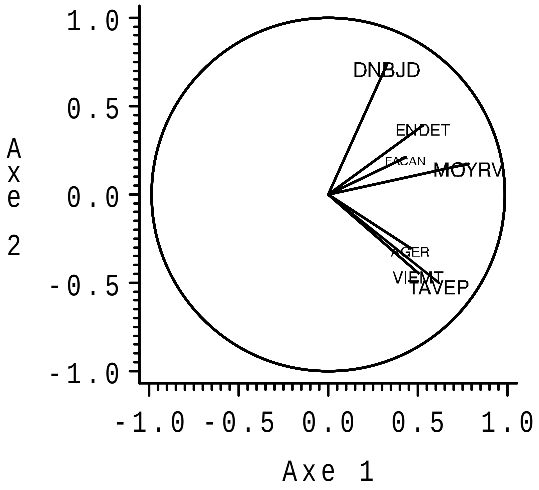


FIG. 2.4 – *Carte Visa : la représentation des variables dans le premier plan de l'ACP fournit une interprétation classique (stocks versus flux) de ce type de données.*

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.

4 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 les plus courants.

4.1 Part d'inertie

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

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

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.

4.2 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.

4.3 Éboulis des valeurs propres

C'est le graphique (figures 4.3 et 4.3) 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 $\widehat{E_q}$.

4.4 Boîtes-à-moustaches des variables principales

Un graphique (figure 4.4 et 4.4) 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 influents, dans une direction principale.

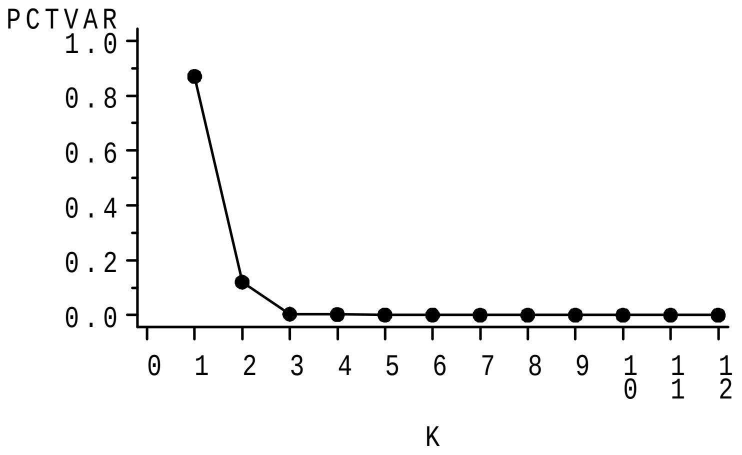


FIG. 2.5 – *Températures : éboulis des valeurs propres.*

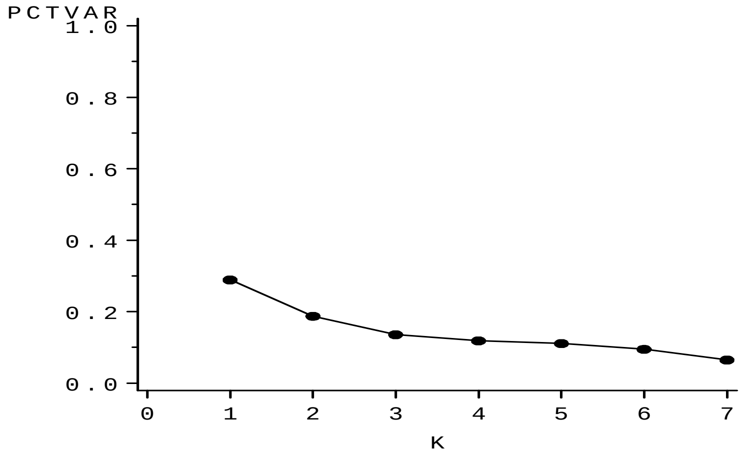


FIG. 2.6 – *Carte Visa : éboulis des valeurs propres.*

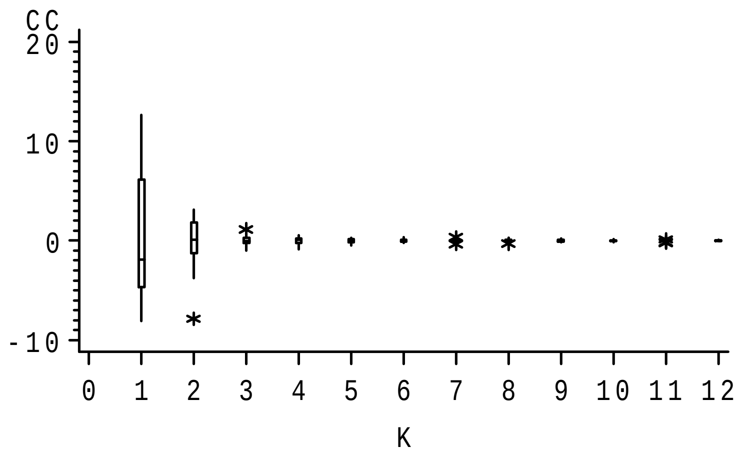


FIG. 2.7 – *Températures : composantes en boîtes.*

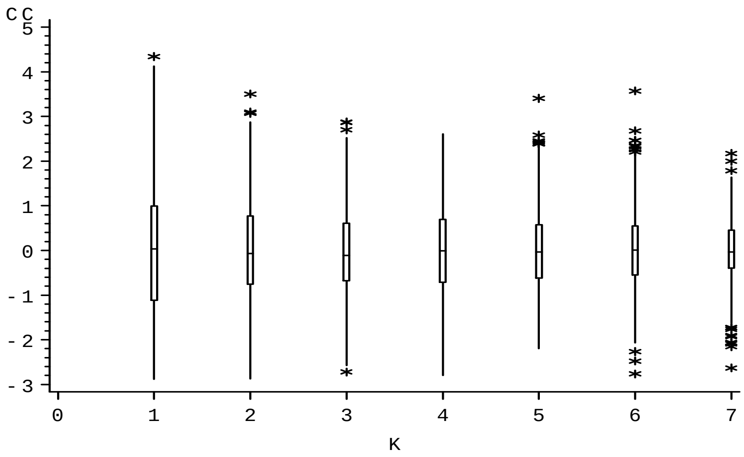


FIG. 2.8 – *Carte Visa Premier : composantes en boîtes.*

5 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.

- 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 3

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

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

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

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

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 :

$$\mathbf{X}_e = \mathbf{TG} = \mathbf{PG}\ ;$$

$\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$ 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 :

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

PROPOSITION 3.1. — *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 :

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

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 (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 3.2. — *L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle 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

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

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$\times$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,

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

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 3.3. — *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.

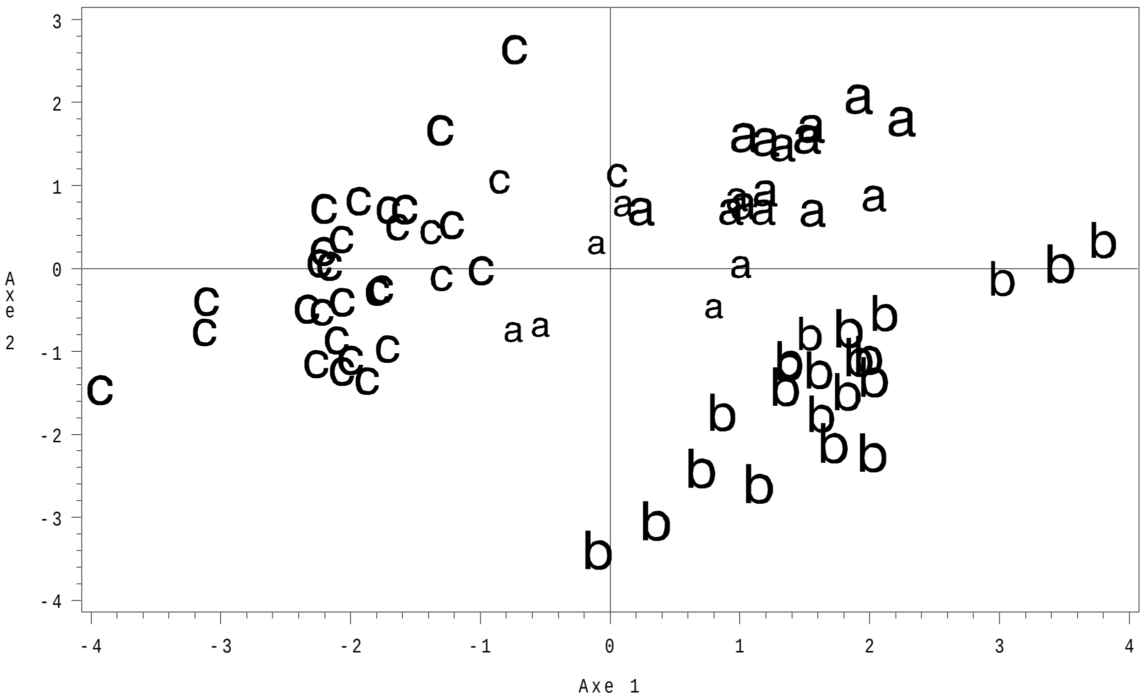


FIG. 3.1 – *Insectes : premier plan factoriel de l'ACP.*

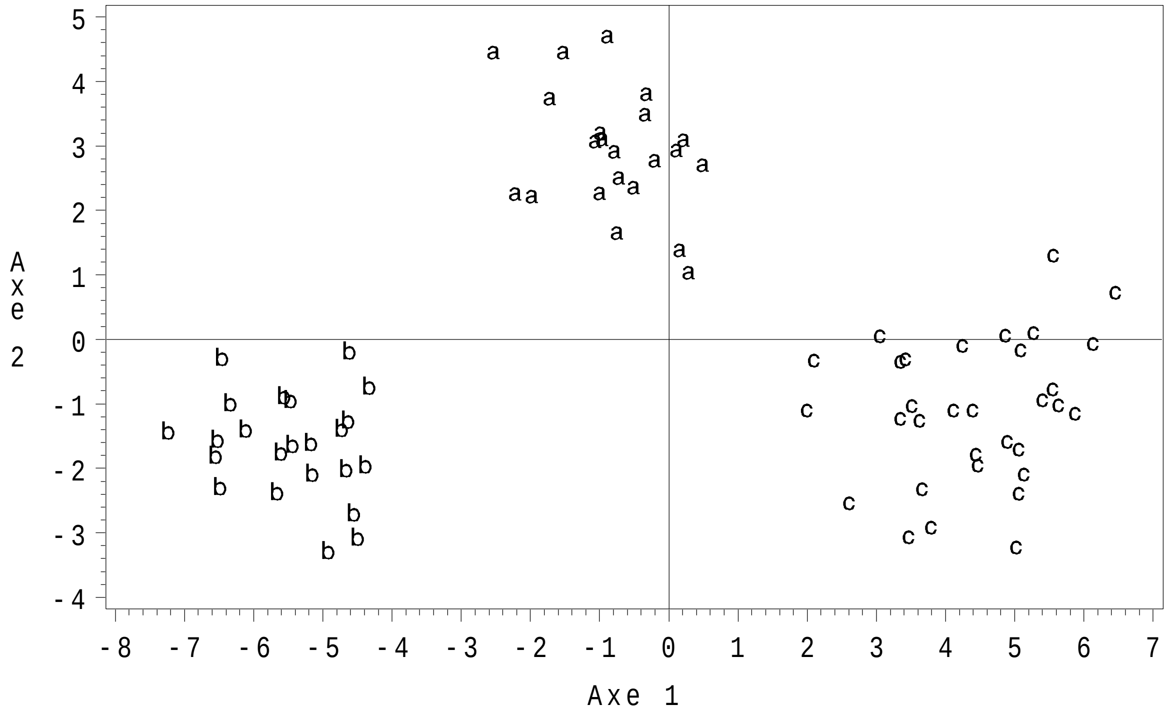


FIG. 3.2 – *Insectes : premier plan factoriel de l'AFD.*

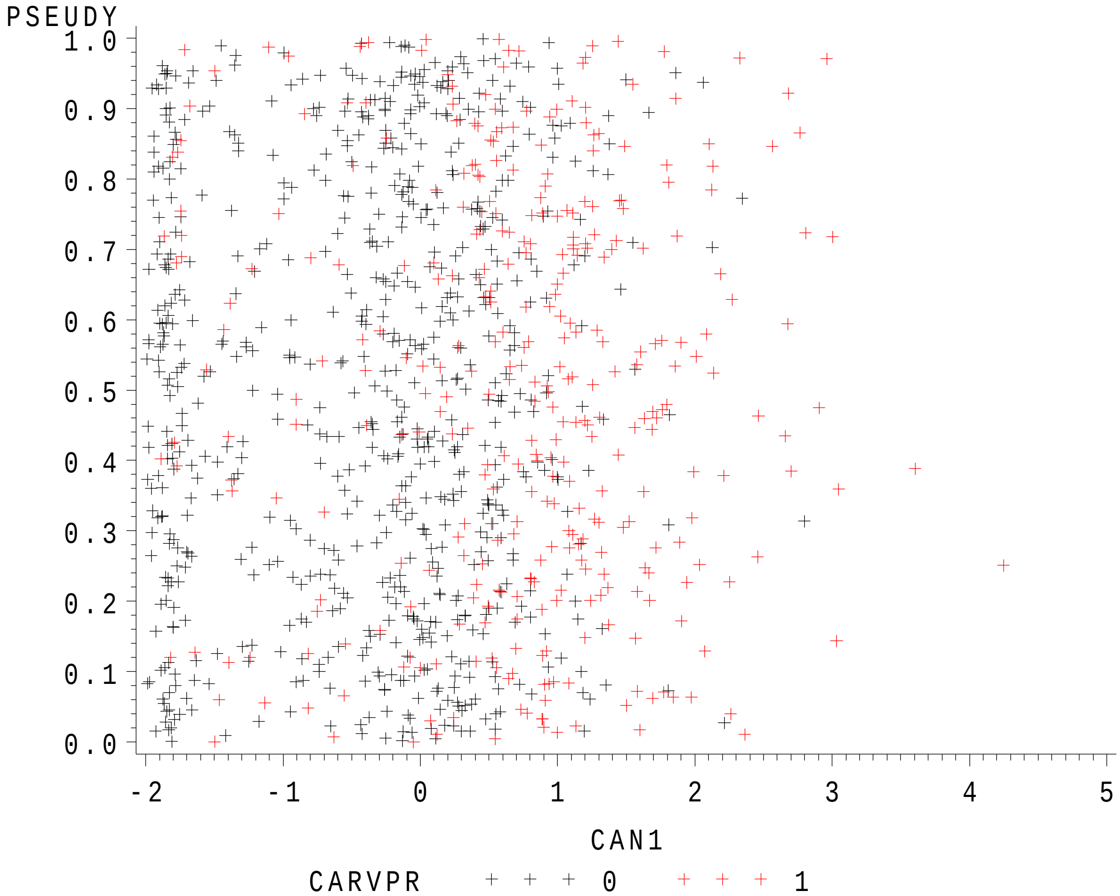


FIG. 3.3 – *Carte Visa : premier plan factoriel de l'AFD. L'axe 2 est issu d'un tirage aléatoire, l'axe 1 ne fournit ps une discrimination très marquée. Celle-ci pourrait être améliorée par un choix plus judicieux des variables.*

Chapitre 4

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 :

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

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 4.1. — *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 1).

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 1). 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 4.2. — *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 $\mathbf{BD}_c\mathbf{B}' - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r'$ 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**)

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

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 4.3. — *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 :

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

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

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

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 :

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

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

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

Remarque. — :

i. Le modèle (1) ci-dessus est équivalent au modèle considéré par Goodman (1991) :

$$p_{\ell h} = p_{\ell +} p_{+h} \left(1 + \sum_{k=1}^{q} \sqrt{\lambda_k} \xi_\ell^k \eta_h^k \right), \tag{2}$$

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 (1). Le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n} \mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1} D_c^{-1}}}. \tag{3}$$

[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.

Estimation

PROPOSITION 4.4. — *L'estimation des paramètres de (1) en résolvant (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 :

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

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

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

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

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

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

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 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

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

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 ($\frac{1}{n}\mathbf{T}$) :

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

- Les valeurs propres vérifient :

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

En effet, on vérifie facilement :

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

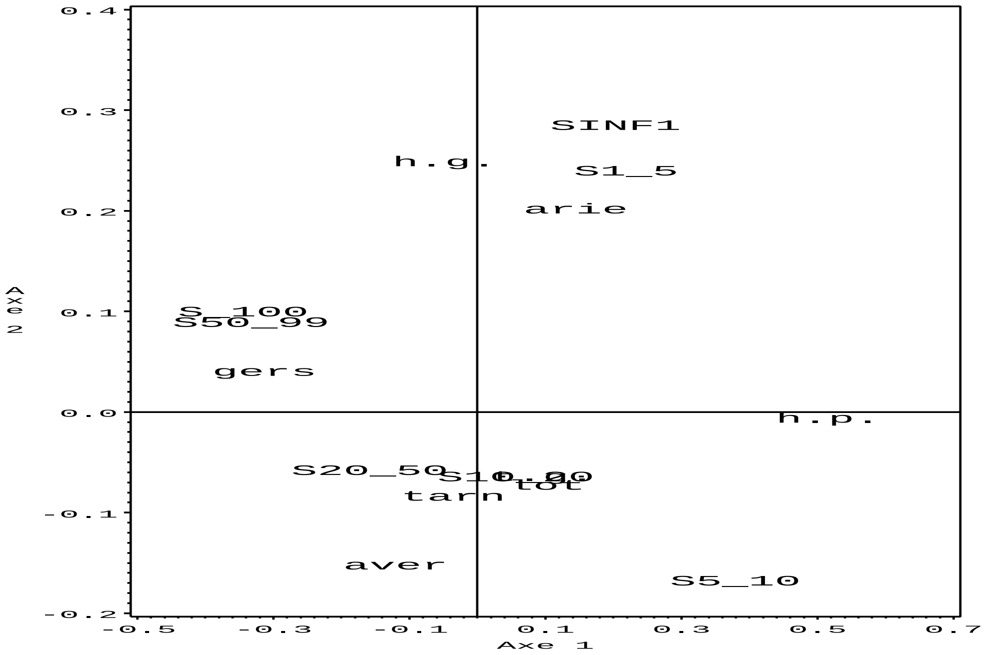


FIG. 4.1 – *Répartition des exploitations agricoles par taille et par département. Premier plan de l'AFC.*

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 5

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 ($k = 1, \dots, c$), 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 $X^1, \dots, X^p$. 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 :

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

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 :

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

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 5.1. — *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 5.2. — *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

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

(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 :

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

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

Proposition 5.3. — *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 :*

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

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 :

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

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 5.4. — *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 $\mathbf{\Delta} =$ diag $(\mathbf{D}_1 \ldots \mathbf{D}_p)$ ($\mathbf{\Delta}$ 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 5.5. — *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 5.6. — *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 :*

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

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 5.7. — *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 :

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

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

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

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`,

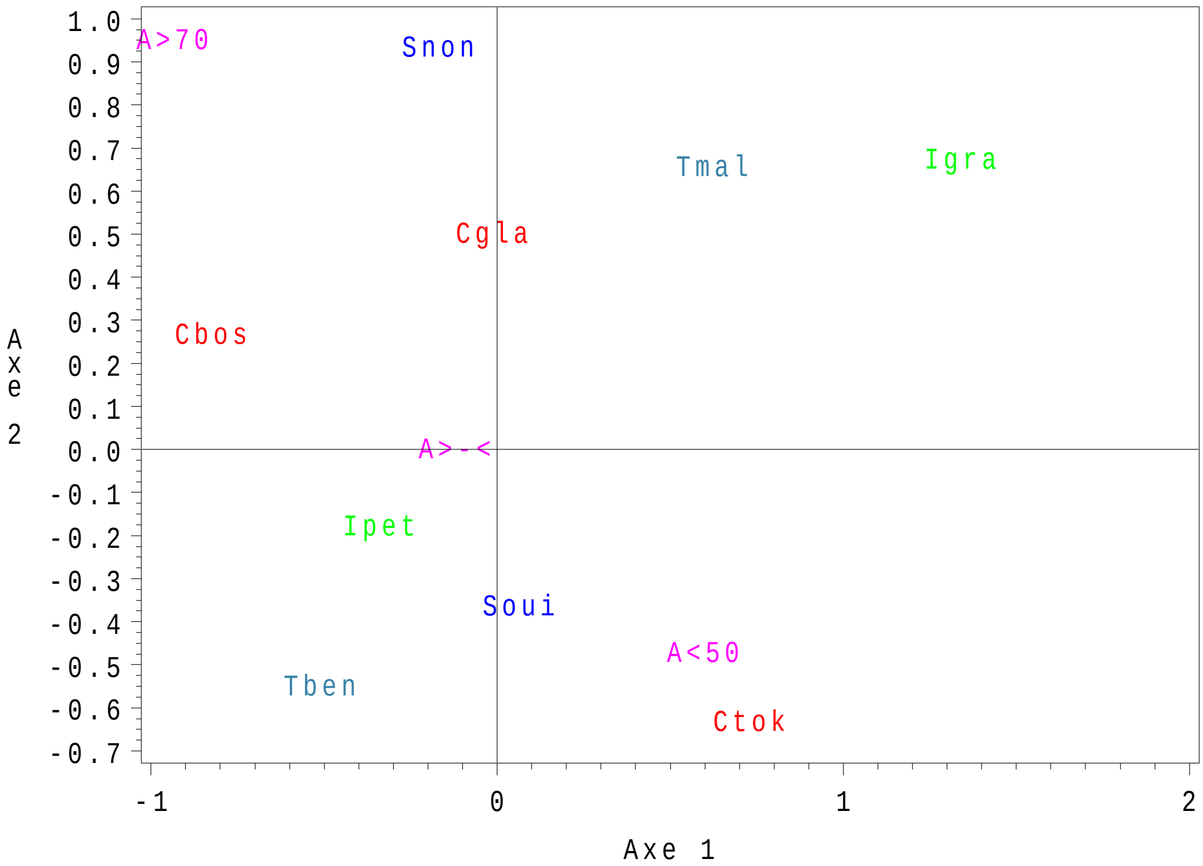


FIG. 5.1 – *Cancer du sein : analyse des données brutes.*

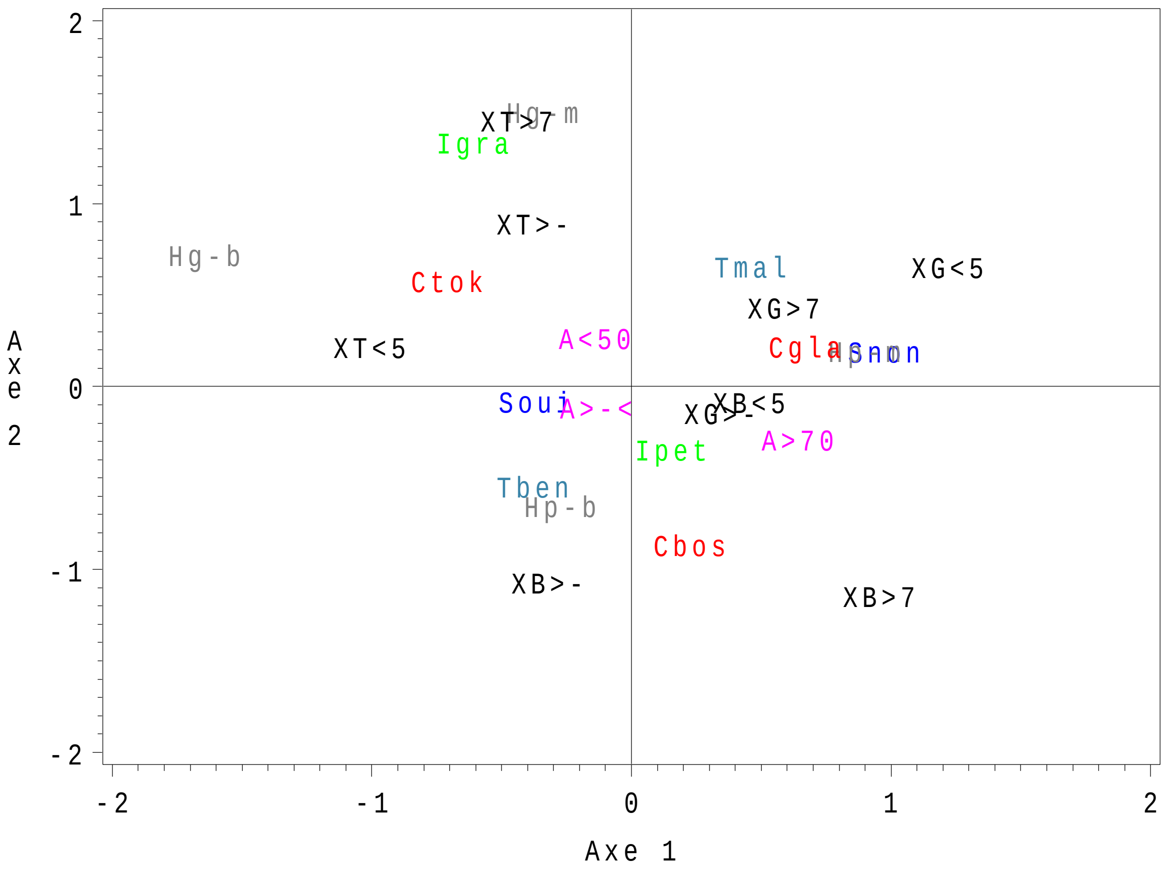


FIG. 5.2 – *Cancer du sein : analyse des interactions.*

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.

Chapitre 6

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 6.1. —

- *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 6.2. — *Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs* $\{\mathbf{x_1}, \dots, \mathbf{x_n}\}$ *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 $\mathcal{D}$ 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 6.3. —

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

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

  alors $\mathcal{D}$ *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* $\mathcal{C}$ *est positive alors la matrice de distance* $\mathcal{D}$ *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 :

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

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 :

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

PROPOSITION 6.4. — *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* **D***-orthogonaux sur les directions engendrées par les vecteurs* **x** *et* **y**.

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

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

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

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

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

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 **X** et **Y** les matrices contenant les variables indicatrices des modalités des variables et $\mathbf{P_X}$, $\mathbf{P_Y}$ les projecteurs **D**-orthogonaux sur les sous-espaces engendrés par ces indicatrices. On montre (cf. Saporta 1976) alors la

PROPOSITION 6.5. — *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 6.6. — *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 6.7. — *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 $\mathbf{B} = \mathbf{U\Delta U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U\Delta}^{1/2}$.*

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 $\mathbf{B} = \mathbf{U\Delta U}'$;
iv. les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $\mathbf{X} = \mathbf{U\Delta}^{1/2}$.

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

Proposition 6.8. — *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

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

où $\mathbf{V}$ est la matrice des vecteurs propres de la matrice $1/n\mathbf{X}'\mathbf{X}\mathbf{M}$ et $\mathbf{U}$ ceux des vecteurs propres de la matrice $1/n\mathbf{X}\mathbf{M}\mathbf{X}'$ associés aux mêmes valeurs propres $\mathbf{\Lambda}$. De son côté, le MDS conduit 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 6.9. — *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.

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

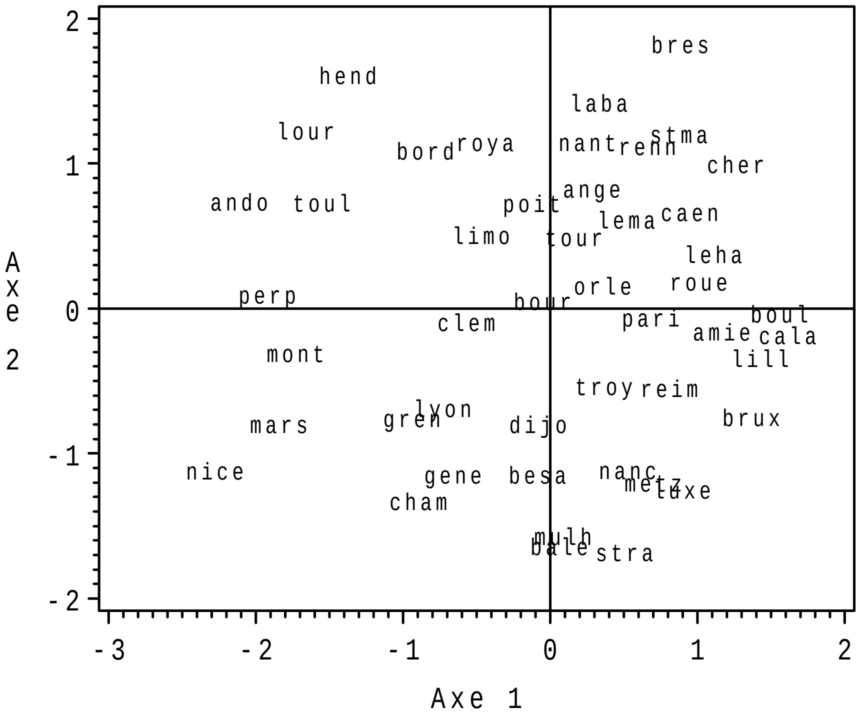


FIG. 6.1 – *Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.*

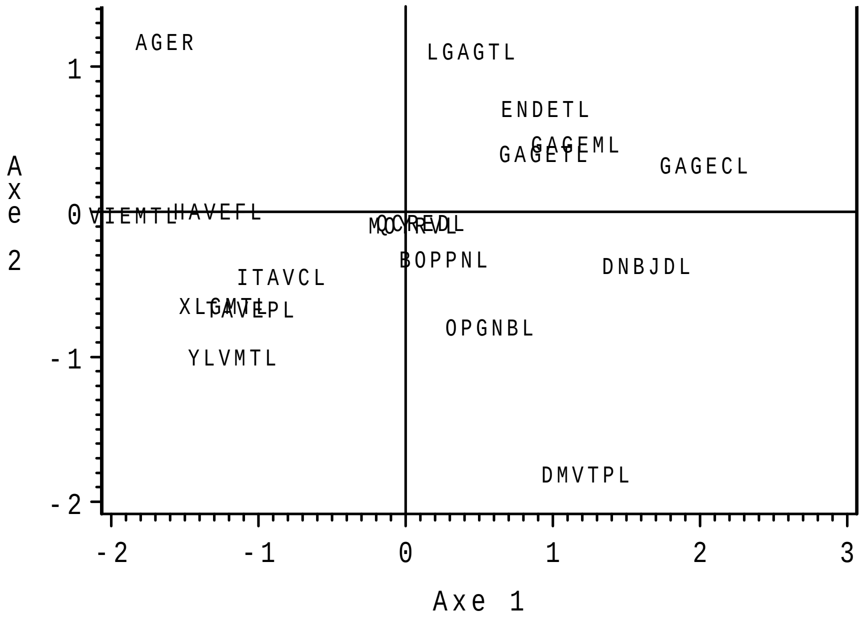


FIG. 6.2 – *Positionnement, conformément aux carrés de leurs corrélations, des variables quantitatives observées sur les données bancaires.*

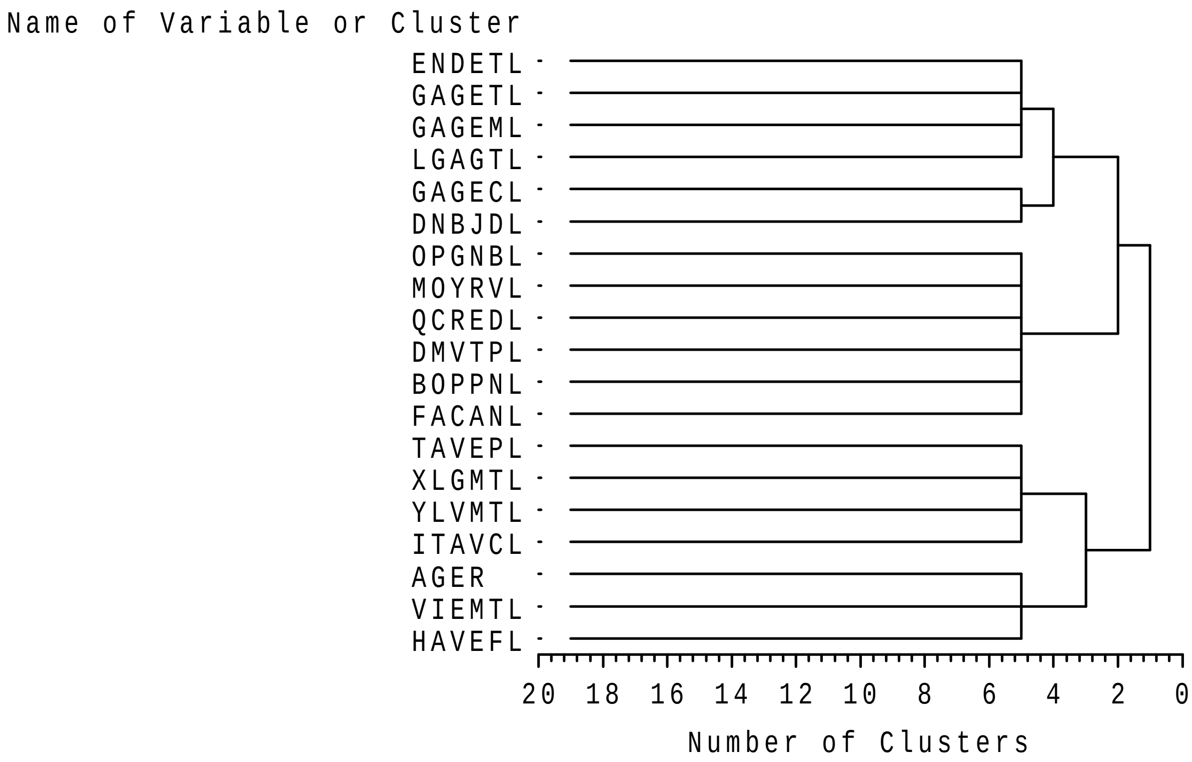


FIG. 6.3 – *Classification (*`varclus`*) des variables quantitatives observées sur les données bancaires.*

(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 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 7

Classification

1 Introduction

Les données, se présentent sous la forme d'une matrice (n, p) contenant les observations de p variables quantitatives sur un échantillon de n individus ou encore sous celle d'un tableau de distances (ou mesure de dissemblance) des individus 2 à 2. 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 homogènes et le plus distinctes entre elles. Cet objectif est à distinguer des procédures de discrimination ou encore de classement pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage.

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}. Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une"bonne" partition qui peut correspondre 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 :

- le critère à optimiser ; il est généralement défini à partir des traces des matrices de variances inter ou intra (ou inertie) de la partition ;
- 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.

La variance totale se décomposant en la somme des variances inter et intra, il est équivalent de maximiser la trace de la variance inter (classes les plus distinctes) que de minimiser la trace de la variance intra (classes les plus homogènes) pour un nombre de classes fixé *a priori* afin d'éviter la solution triviale d'une partition en n singletons. Dans ce cas extrême, la variance inter est maximale car égale à la variance totale.

2 Classification hiérarchique ascendante

L'initialisation consiste à calculer un tableau de distances entre les individus à classer. L'algorithme démarre alors de la partition triviale des n singletons et cherche, à chaque étape, à constituer des classes par agglomération de 2 éléments de la partition de l'étape précédente. Une mise à jour du tableau de distance calcule la "distance" du nouveau groupe constitué avec tous les autres groupes selon un indice fixé par l'utilisateurs.

Notons A et B deux groupes ou éléments d'une partition, g_A et g_B leurs barycentres, w_A et w_B leurs pondérations, $d_{i,j}$ la distance entre deux individus (distance euclidienne dans $\mathbb{R}^p$ ou indice de dissemblance), les choix possibles sont :

$$\begin{aligned}
d(A,B) &= \min_{i\in A, j\in B}(d_{ij}) \quad \text{(saut minimum)},\\
d(A,B) &= \sup_{i\in A, j\in B}(d_{ij}) \quad \text{(saut maximum ou diamètre)},\\
d(A,B) &= \frac{1}{\mathrm{card}(A)\mathrm{card}(B)}\sum_{i\in A, j\in B} d_{ij} \quad \text{(saut moyen)},\\
d(A,B) &= d(g_A, g_B) \quad \text{(distance des barycentres)},\\
d(A,B) &= \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}.
\end{aligned}$$

Le saut de Ward mesure, dans le cas d'une distance euclidienne entre individus, la perte de variance inter par fusion de deux classes. Il s'agit donc, à chaque étape, de minimiser cette quantité.

Il est d'usage de tracer un graphique représentant, à rebours, la décroissance en fonction du nombre de classes, du rapport de la variance intra 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. L'indice de Ward est le plus généralement utilisé.

Algorithme de classification hiérarchique ascendante

Initialisation Les classes initiales sont les singletons. Calculer la matrice de leurs distances 2 à 2.

Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe.
- regrouper les deux classes les plus proches au sens de la "distance" entre groupes choisie,
- mise à jour du tableau de distance en remplaçant les deux classes par la nouvelle et en calculant sa "distance" avec les autres classes.

À l'issue de l'exécution, cette méthode fournit donc deux graphiques, le premier aide au choix du nombre de classes tandis que le deuxième : le *dendogramme* est une représentation graphique, sous forme d'arbre, des agrégations successives jusqu'à la réunion en une seule classe. La hauteur d'une branche est proportionnel à la perte de variance inter au niveau d'agrégation considéré. Une fois un nombre de classes sélectionné, une coupure de l'arbre fournit, dans chaque sous-arbre, la répartition des individus en classes.

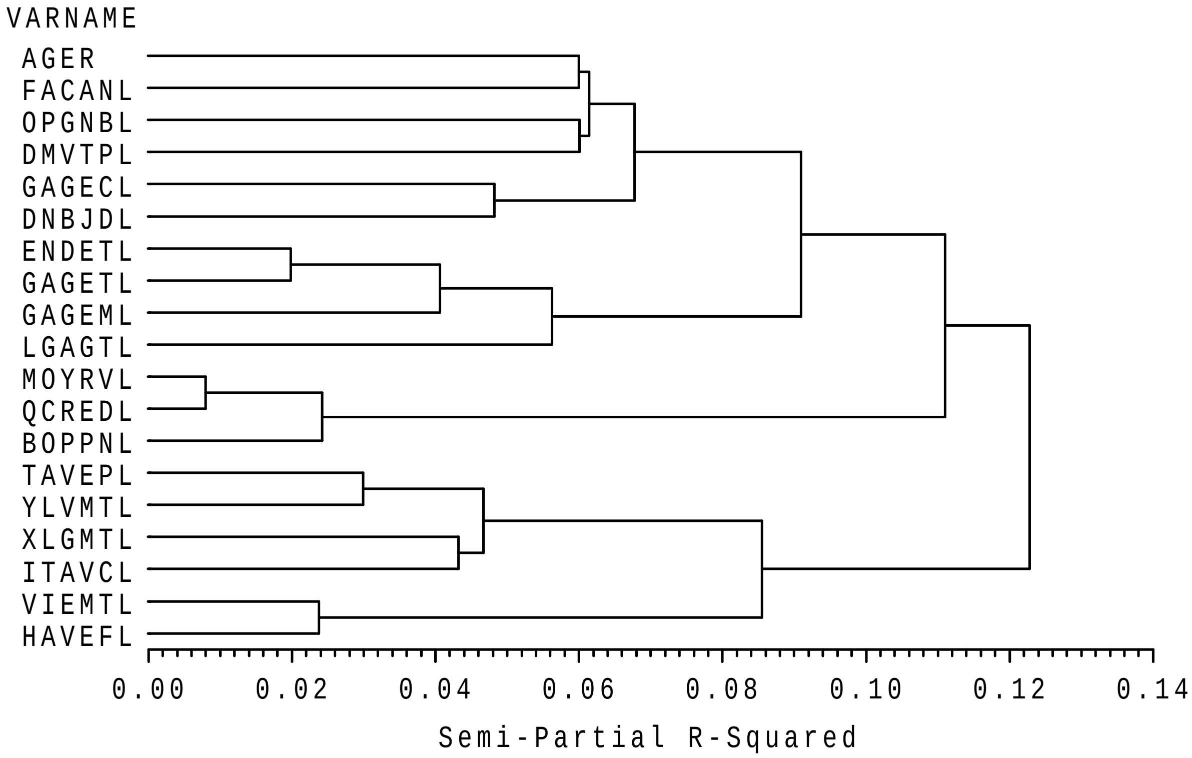


FIG. 7.1 – *Classification, conformément aux carrés de leurs corrélations, des variables quantitatives observées sur les données bancaires.*

3 Centres mobiles

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. En France, ce sont les nuées dynamiques. Contrairement à la méthode hiérarchique précédente, le nombre de classes k doit être déterminé *a priori*.

Algorithme de Forgy

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 inter classe ne croisse plus de manière significative, lorsque les classes se sont stabilisées.
- Allouer chaque individu à l'un des noyaux, c'est-à-dire à une classe. On obtient ainsi, à chaque étape, une classification en k classes ou moins si finalement une des classes devient vide.
- Calculer le centre de gravité de chaque classe, il devient le nouveau noyau. Si une classe s'est vidée, on peut éventuellement retirer un noyau complémentaire.

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.

Cette méthode se généralise en considérant pour "noyaux" d'autres types d'objets comme, par exemple, une famille de densités ou encore des axes factoriels.

Algorithme kmeans

Il s'agit d'une modification de l'algorithme précédent. Les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; l'algorithme est ainsi plus efficace.

4 Combinaison

Chaque méthode précédente peut être plus ou moins adaptée à la situation rencontrée. La classification hiérarchique qui construit la matrice des distances n'accepte qu'un nombre limité d'individus, il est indispensable de fixer le nombre de classes dans le cas de la réallocation dynamique. 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 et en choisissant pour noyaux les barycentres (calculés en pondérant par les effectifs de classes) des classes de l'étape précédente.

Dans tous les cas, le résultat fourni est une variable qualitative T dont les modalités précisent la classe retenue pour chaque individu. D'autres méthodes sont ensuite traditionnellement enchaînées : ACP 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.

Si partie ou totalité des variables observées sont qualitatives, une AFC préalable du tableau disjonctif complet fournit un ensemble de coordonnées, relatives aux premiers axes factoriels, qui permet de se ramener au cas quantitatif.

Deuxième partie

Modélisation

Chapitre 8

Régression linéaire multiple

1 Introduction

Ce chapitre ne propose qu'une introduction au modèle gaussien, à sa définition et à son estimation. Il s'attarde sur le problème délicat du choix de modèle afin, principalement, d'en introduire les grands principes pour les adapter au cas de la régression logistique largement utilisée en prospection de données. Une dernière section introduit le modèle d'analyse de covariance mais de nombreux aspects : colinéarité, points influents, tests, analyse de variance, modèle multinomial ou poissonien (modèle log-linéaire)...sont négligés et à rechercher dans la bibliographie de même qu'une présentation globale du *modèle linéaire généralisé* incluant toutes ces approches.

Le but premier de ce chapitre est donc l'explication ou la modélisation d'une variable *quantitative* par plusieurs variables quantitatives (régression linéaire multiple) ou par un mélange de variables quantitatives et qualitatives (analyse de covariance).

2 Modèle

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

Une variable quantitative Y dite *à expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, \ldots, X^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i) \quad i = 1, \ldots, n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, X^1, \ldots, X^p\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de "1" . C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + u_i \quad i = 1, 2, \ldots, n$$

avec les hypothèses suivantes :

i. Les u_i sont des termes d'erreur, d'une variable U, non observés, indépendants et identiquement distribués ; $E(u_i) = 0, Var(U) = \sigma_u^2\mathbf{I}$.

ii. Les termes x^j sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur U est indépendante de la distribution conjointe de $X^1, \ldots, X^p$. On écrit dans ce dernier cas que :

$$E(Y|X^1, \ldots, X^p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p \text{ et } Var(Y|X^1, \ldots, X^p) = \sigma_u^2.$$

iii. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.

iv. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur U ($\mathcal{N}(0, \sigma_u^2\mathbf{I})$). Les u_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma_u^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général y_i. En notant les vecteurs $\mathbf{u} = [u_1 \cdots u_p]'$ et $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

3 Estimation

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ_u^2 (paramètre de nuisance), sont estimés par minimisation du critère des moindres carrés (M.C.) ou encore, en supposant (iv), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

3.1 Estimation par M.C.

L'expression à minimiser sur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \cdots - \beta_p x_i^p)^2 &= \|\mathbf{y} - \mathbf{X}\beta\|^2 \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient les *"équations normales"* :

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Des diagnostics de colinéarité et des critères aident au choix des variables.

Alors, l'estimation des paramètres β_j est donnée par :

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

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{y}$ ont pour expression :

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est appelée "*hat matrix*" ; elle met un chapeau à $\mathbf{y}$. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace $\mathrm{Vect}(\mathbf{X})$ engendré par les vecteurs colonnes de $\mathbf{X}$.

On note

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

le vecteur des résidus ; c'est la projection de $\mathbf{y}$ sur le sous-espace orthogonal de $\mathrm{Vect}(\mathbf{X})$ dans $\mathbb{R}^n$.

3.2 Propriétés

Les estimateurs des M.C. $b_0, b_1, \ldots, b_p$ sont des estimateurs sans biais : $E(\mathbf{b}) = \boldsymbol{\beta}$, et, parmi les estimateurs sans biais fonctions linéaires des y_i, ils sont de variance minimum (propriété de Gauss-Markov) ; ils sont donc "BLUE" : *best linear unbiaised estimators*. Sous hypothèse de normalité, les estimateurs du M.V. sont uniformément meilleurs ; ils sont efficaces c'est-à-dire que leur matrice de covariance atteint la borne inférieure de Cramer-Rao.

On montre que la matrice de covariance des estimateurs se met sous la forme

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1},$$

celle des prédicteurs est

$$E[(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'] = \sigma_u^2\mathbf{H}$$

et celle des estimateurs des résidus est

$$E[(\mathbf{e} - \mathbf{u})((\mathbf{e} - \mathbf{u}))'] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

tandis qu'un estimateur sans biais de σ_u^2 est fourni par :

$$s^2 = \frac{\|\mathbf{e}\|^2}{n-p-1} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-p-1} = \frac{\mathrm{SSE}}{n-p-1}.$$

Ainsi, les termes $s^2 h_i^i$ sont des estimations des variances des prédicteurs $\widehat{y}_i$.

3.3 Sommes des carrés

SSE est la somme des carrés des résidus (*sum of squared errors*),

$$\mathrm{SSE} = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \|\mathbf{e}\|^2 .$$

On définit également la somme totale des carrés (*total sum of squares*) par

$$\mathrm{SST} = \|\mathbf{y} - \bar{y}\mathbf{1}\|^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2$$

et la somme des carrés de la régression (*regression sum of squares*) par

$$\mathrm{SSR} = \|\widehat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2 = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\bar{y}^2 = \mathbf{y}'\mathbf{H}\mathbf{y} - n\bar{y}^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2.$$

On vérifie alors : $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$.

3.4 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

qui est donc la part de variation de Y expliquée par le modèle de régression. Géométriquement, c'est un rapport de carrés de longueur de deux vecteurs. C'est donc le cosinus carré de l'angle entre ces vecteurs : $\mathbf{y}$ et sa projection $\widehat{\mathbf{y}}$ sur $\text{Vect}(\mathbf{X})$.

Attention, dans le cas extrême où $n = (p+1)$, c'est-à-dire si le nombre de variables explicatives est grand comparativement au nombre d'observations, $R^2 = 1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination.

La quantité R est appelée *coefficient de corrélation multiple* entre Y et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{y}$ et sa prédiction (ou projection) $\widehat{\mathbf{y}}$.

4 Inférences dans le cas gaussien

En principe, l'hypothèse optionnelle (iv) de normalité des erreurs est nécessaire pour cette section. En pratique, des résultats asymptotiques, donc valides pour de grands échantillons, ainsi que des études de simulation, montrent que cette hypothèse n'est pas celle dont la violation est la plus pénalisante pour la fiabilité des modèles.

4.1 Inférence sur les coefficients

Pour chaque coefficient β_j on montre que la statistique

$$\frac{b_j - \beta_j}{\sigma_{b_j}}$$

où $\sigma^2_{b_j}$, variance de b_j est le terme diagonal de la matrice $s^2(\mathbf{X}'\mathbf{X})^{-1}$, suit une loi de Student à $(n-p-1)$ degrés de liberté. Cette statistique est donc utilisée pour tester une hypothèse $H_0 : \beta_j = a$ ou pour construire un intervalle de confiance de niveau $100(1-\alpha)\%$:

$$b_j \pm t_{\alpha/2;(n-p-1)}\sigma_{b_j}.$$

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement (cf. §3.4) sur d'autres coefficients car ils sont corrélés entre eux ; de plus elle dépend des absences ou présences des autres variables X^k dans le modèle. Par exemple, dans le cas particulier de deux variables X^1 et X^2 très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, elles peuvent chacune apparaître avec des coefficients insignifiants.

De façon plus générale, si $\mathbf{c}$ désigne un vecteur de $(p+1)$ constantes réelles, il est possible de tester la valeur d'une combinaison linéaire $\mathbf{c}'\mathbf{b}$ des paramètres en considérant l'hypothèse nulle H_0 : $\mathbf{c}'\mathbf{b} = a$. Sous H_0, la statistique

$$\frac{\mathbf{c}'\mathbf{b} - a}{(s^2\mathbf{c}'\mathbf{X}'\mathbf{X}^{-1}\mathbf{c})^{1/2}}$$

suit une loi de Student à $(n-p-1)$ degrés de liberté.

4.2 Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle H_0 : $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, la statistique

$$\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)} = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté. Les résultats sont habituellement présentés dans un tableau *"d'analyse de la variance"* sous la forme suivante :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	F
Régression	p	SSR	MSR=SSR/p	MSR/MSE
Erreur	$n-p-1$	SSE	MSE=SSE/$(n-p-1)$	
Total	$n-1$	SST		

4.3 Inférence sur un modèle réduit

Le test précédent amène à rejeter H_0 dès que l'une des variables X^j est liée à Y. Il est donc d'un intérêt limité. Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients sont nuls contre le modèle complet avec toute les variables. En ayant éventuellement réordonné les variables, on considère l'hypothèse nulle H_0 : $\beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$.

Notons respectivement SSR_q, SSE_q, R_q^2 les sommes de carrés et le coefficient de détermination du modèle réduit à $(p-q)$ variables. Sous H_0, la statistique

$$\frac{(\mathrm{SSR} - \mathrm{SSR}_q)/q}{\mathrm{SSE}/(n-p-1)} = \frac{(R^2 - R_q^2)/q}{(1-R^2)/(n-p-1)}$$

suit une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

Dans le cas particulier où $q = 1$ $(\beta_j = 0)$, la F-statistique est alors le carré de la t-statistique de l'inférence sur un paramètre et conduit donc au même test.

4.4 Prévision

Connaissant les valeurs des variables X^j pour une nouvelle observation : $\mathbf{x}_0' = [x_0^1, x_0^2, \ldots, x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y}_0$ de Y ou $E(Y)$ est donnée par :

$$\widehat{y}_0 = b_0 + b_1 x_0^1 + \cdots + b_p x_0^p.$$

Les intervalles de confiance des prévisions de Y et $E(Y)$, pour *une* valeur x_0, sont respectivement

$$\begin{aligned}\widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}s(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2},\\ \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}s(\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}.\end{aligned}$$

5 Sélection de variables, choix de modèle

De façon un peu schématique, on peut associer la pratique de la modélisation statistique à trois objectifs qui peuvent éventuellement être poursuivis en complémentarité.

Descriptif : Il vise à rechercher de façon exploratoire les liaisons entre Y et d'autres variables, potentiellement explicatives, X^j qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyses en Composantes Principales, correspond des algorithmes de recherche (pas à pas) moins performants mais économiques en temps de calcul si p est grand.

Attention, si n est petit, et la recherche suffisamment longue avec beaucoup de variables explicatives, il sera toujours possible de trouver un “bon” modèle expliquant y ; c'est l'effet *data mining* dans les modèles économétriques.

Explicatif : Le deuxième objectif est sous-tendu par une connaissance *a priori* du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels précédents permettent de construire *le* bon test conduisant à la prise de décision recherchée. Utilisées hors de ce contexte, les statistiques de test n'ont plus alors qu'une valeur indicative au même titre que d'autres critères plus empiriques.

Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des estimateurs et des prédicteurs qui doivent, par exemple, minimiser une erreur quadratique moyenne. Ceci conduit à rechercher des modèles *parcimonieux* c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Le “meilleur” modèle ainsi obtenu peut donner des estimateurs légèrement biaisés au profit d'un compromis pour une variance plus faible. Un bon modèle n'est donc plus celui qui explique le mieux les données au sens d'une déviance (SSE) minimale (ou d'un R^2 max) au prix d'un nombre important de variables pouvant introduire des colinéarités. Le bon modèle est celui qui conduit aux prédictions les plus fiables.

5.1 Critères

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple. Citons le critère d'information d'Akaïke (`AIC`), celui bayésien de Sawa (`BIC`), l'erreur quadratique moyenne de prédiction (cas gaussien).... Ils sont équivalents lorsque le nombre de variables à sélectionner, ou niveau du modèle, est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de niveaux différents. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux. En pratique, les plus utilisés ou ceux généralement fournis par les logiciels sont les suivants.

Statistique du F de Fisher

Ce critère, justifié dans le cas explicatif est aussi utilisé à titre indicatif pour comparer des séquences de modèles emboîtés. La statistique partielle de Fisher est

$$\frac{(\mathrm{SSR}-\mathrm{SSR}_q)/s}{\mathrm{SSE}/(n-p-1)}=\frac{(R^2-R_q^2)}{1-R^2)}\frac{n-p-1}{q}$$

dans laquelle l'indice q désigne les expressions concernant le modèle réduit avec $(p-q)$ variables explicatives. On considère alors que si l'accroissement $(R^2-R_q^2)$ est suffisamment grand :

$$R^2-R_R^2>\frac{q}{(n-p-1)}F_{\alpha;q,(n-p-1)},$$

l'ajout des q variables au modèle est justifié.

R^2 et R^2 ajusté

Le coefficient de détermination $R^2=1-\mathrm{SSE}/\mathrm{SST}$, directement lié à la déviance (SSE) est aussi un indice de qualité mais qui a la propriété d'être monotone croissant en fonction du nombre de variables. Il ne peut donc servir qu'à comparer deux modèles de même niveau c'est-à-dire avec le même nombre de variables.

En revanche, le R^2 *ajusté* :

$$R'^2=1-\frac{n-1}{n-p-1}(1-R^2)=1-\frac{\mathrm{SSE}/(n-p-1)}{\mathrm{SST}/(n-1)}.$$

dans lequel le rapport SSE/SST est remplacé par un rapport des estimations sans biais des quantités σ_u^2 et σ_y^2 introduit une pénalisation liée au nombre de paramètres à estimer.

Ce coefficient s'exprime encore par

$$1-\frac{(n-1)\mathrm{MSE}}{\mathrm{SST}}$$

ainsi dans la comparaison de deux modèles partageant la même SST, on observe que $R'^2>R'^2_j$ si et seulement si $\mathrm{MSE}<\mathrm{MSE}_j$; MSE et MSE_j désignant respectivement l'erreur quadratique moyenne du modèle complet et celle d'un modèle à j variables explicatives. Maximiser le R^2 ajusté revient donc à minimiser l'erreur quadratique moyenne.

C_p de Mallow

Une erreur quadratique moyenne s'écrit comme la somme d'une variance et du carré d'un biais :

$$\mathrm{MSE}(\widehat{y}_i)=\mathrm{Var}(\widehat{y}_i)+[\mathrm{Biais}(\widehat{y}_i)]^2$$

puis après sommation et réduction :

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\mathrm{MSE}(\widehat{y}_i)=\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\mathrm{Var}(\widehat{y}_i)+\frac{1}{\sigma_u^2}\sum_{i=1}^{n}[\mathrm{Biais}(\widehat{y}_i)]^2.$$

En supposant que les estimations du modèle complet sont sans biais et en utilisant des estimateurs de $Var(\widehat{y}_i)$ et σ_u^2, l'expression de l'erreur quadratique moyenne totale standardisée (ou réduite) pour un modèle à j variables explicatives s'écrit :

$$C_j = (n-j-1)\frac{\text{MSE}_j}{\text{MSE}} - [n - 2(j+1)]$$

et définit la valeur du C_p de Mallow pour les j variables considérées. Il est alors d'usage de rechercher un modèle qui minimise le C_p tout en fournissant une valeur inférieure et proche de $(j+1)$. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise.

PRESS de Allen

On désigne par $\widehat{y}_{(i)}$ la prédiction de y_i calculée sans tenir compte de la $i$ème observation $(y_i, x_i^1, \ldots, x_i^p)$, la somme des erreurs quadratiques de prédiction (PRESS) est définie par

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2$$

et permet de comparer les capacités prédictives de deux modèles.

5.2 Algorithmes de sélection

Lorsque p est grand, il n'est pas raisonnable de penser explorer les 2^p modèles possibles afin de sélectionner le "meilleur" au sens de l'un des critères ci-dessus. Différentes stratégies sont donc proposées qui doivent être choisies en fonction de l'objectif recherché et des moyens de calcul disponibles ! Trois types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire c'est-à-dire par nombre croissant de modèles considérés parmi les 2^p et donc par capacité croissante d'optimalité. On donne pour chaque algorithme l'option `selection` à utiliser dans la procédure `REG` de SAS.

Pas à pas

Sélection (`forward`) À chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur p ("prob value")associée à la statistique partielle du test de Fisher qui compare les deux modèles est minimum. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque p reste plus grande qu'une valeur seuil fixée par défaut à $0,50$.

Élimination (`backward`) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable associée à la plus grande valeur p est éliminée du modèle. La procédure s'arrête lorsque les variables restant dans le modèle ont des valeurs p plus petites qu'un seuil fixé par défaut à $0,10$.

Mixte (`stepwise`) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Global

L'algorithme de Furnival et Wilson est utilisé pour comparer tous les modèles possibles en cherchant à optimiser l'un des critères : R^2, R^2 ajusté, ou C_p de Mallow (`rsquare, adjrsq, cp`) choisi par l'utilisateur. Par souci d'économie, cet algorithme évite de considérer des modèles de certaines sous-branches de l'arborescence dont on peut savoir a priori qu'ils ne sont pas compétitifs. En général les logiciels exécutant cet algorithme affichent le (`best=1`) ou les meilleurs modèles de chaque niveau.

6 Diagnostics et compléments

6.1 Multicolinéarité

Problème

L'estimation des paramètres ainsi que celle de leur écart-type (standard error) nécessite le calcul explicite de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Dans le cas dit *mal conditionné* où le déterminant de la matrice $\mathbf{X}'\mathbf{X}$ n'est que légèrement différent de 0, les résultats conduiront à des estimateurs de variances importantes et même, éventuellement, à des problèmes de précision numérique. Il s'agit donc de diagnostiquer ces situations critiques puis d'y remédier. Dans les cas descriptif ou prédictif on supprime des variables à l'aide des procédures de choix de modèle mais, pour un objectif explicatif nécessitant toutes les variables, d'autres solutions doivent être envisagées : algorithme de résolution des équations normales par transformations orthogonales (procédure `orthoreg` de SAS) sans calcul explicite de l'inverse pour limiter les problèmes numériques, régression biaisée (ridge), régression sur composantes principales.

VIF

La plupart des logiciels proposent des diagnostics de colinéarité. Le plus classique est le *facteur d'inflation de la variance* (VIF)

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j^2 désigne le coefficient de détermination de la régression de la variable X^j sur les autres variables explicatives ; R_j est alors un coefficient de corrélation multiple, c'est le cosinus de l'angle dans $\mathbb{R}^n$ entre X^j et le sous-espace vectoriel engendré par les variables $\{X^1, \ldots, X^{j-1}, X^{j-1}, \ldots, X^p\}$. Plus X^j est "linéairement" proche de ces variables et plus R_j est proche de 1 ; on montre alors que la variance de l'estimateur de β_j est d'autant plus élevée. Évidemment, cette variance est minimum lorsque X^j est orthogonale au sous-espace engendré par les autres variables.

Conditionnement

De façon classique, les qualités numériques de l'inversion d'une matrice sont quantifiées par son *indice de conditionnement*. On note $\lambda_1, \ldots, \lambda_p$ les valeurs propres de la matrice des corrélations $\mathbf{R}$ rangées par ordre décroissant. Le déterminant de $\mathbf{R}$ est égal au produit

des valeurs propres. Ainsi, des problèmes numériques, ou de variances excessives apparaissent dès que les dernières valeurs propres sont relativement trop petites. L'*indice de conditionnement* est le rapport
$$\kappa = \lambda_1/\lambda_p$$
de la plus grande sur la plus petite valeur propre.

En pratique, si $\kappa < 100$ on considère qu'il n'y a pas de problème. Celui-ci devient sévère pour $\kappa > 1000$. Cet indice de conditionnement donne un aperçu global des problèmes de colinéarité tandis que les VIF, les tolérances ou encore l'étude des vecteurs propres associés au plus petites valeurs propres permettent d'identifier les variables les plus problématiques.

Régression ridge

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur légèrement biaisé des paramètres. L'estimateur "ridge" est donné par
$$\mathbf{b}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$
qui a pour effet de décaler de la valeur k toutes les valeurs propres de la matrice à inverser et, plus particulièrement, les plus petites qui reflètent la colinéarité. On montre que cela revient encore à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur $\boldsymbol{\beta}$ des paramètres ne soit pas trop grande. C'est alors un problème de moindres carrés pénalisés.

Régression sur composantes principales

L'Analyse en Composantes Principales est, entre autre, la recherche de p variables dites principales qui sont des combinaisons linéaires des variables initiales de variance maximale sous une contrainte d'orthogonalité. En désignant par $\mathbf{V}$ la matrice des vecteurs propres de la matrice des corrélations $\mathbf{R}$ rangés dans l'ordre décroissant des valeurs propres, les valeurs prises par ces variables principales sont obtenues dans la matrice des composantes principales
$$\mathbf{C} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{V}.$$
Elles ont chacune pour variance la valeur propre λ_j associée. Le sous-espace engendré par ces variables principales est le même que celui engendré par les variables initiales. Il est donc géométriquement équivalent de régresser Y sur les colonnes de $\mathbf{C}$ que sur celles de $\mathbf{X}$. Les problèmes de colinéarité sont alors résolu en supprimant les variables principales de plus faibles variances c'est-à-dire associées aux plus petites valeurs propres ou encore en exécutant un algorithme de choix de modèle sur les composantes.

La solution obtenue présente ainsi de meilleures qualités prédictives mais les coefficients de la régression s'appliquant aux composantes principales un calcul complémentaire est nécessaire afin d'évaluer et d'interpréter les effets de chacune des variables initiales.

6.2 Modèles curvilinéaires

En cas d'invalidation de l'hypothèse de linéarité, il peut être intéressant de considérer des modèles polynômiaux, très classiques pour décrire des phénomènes physiques, de la forme

$$Y = \beta_0 + \cdots + \beta_j X^j + \cdots + \gamma_{kl} X^k X^l + \cdots + \delta_j X^{j2}$$

qui sont encore appelés *surfaces de réponse*. Ces modèles sont faciles à étudier dans le cadre linéaire, il suffit d'ajouter des nouvelles variables constituées des produits ou des carrés des variables explicatives initiales. Les choix : présence ou non d'une interaction entre deux variables, présence ou non d'un terme quadratique se traitent alors avec les mêmes outils que ceux des choix de variable mais en intégrant une contrainte lors de la lecture des résultats : ne pas considérer des modèles incluant des termes quadratiques dont les composants linéaires auraient été exclus ou encore, ne pas supprimer d'un modèle une variable d'un effet linéaire si elle intervient dans un terme quadratique.

La procédure `rsreg` de SAS est plus particulièrement adaptée aux modèles quadratiques. Elle ne comporte pas de procédure de choix de modèle mais fournit des aides et diagnostics sur l'ajustement de la surface ainsi que sur la recherche des points optimaux.

Attention : Ce type de modèle accroît considérablement les risques de colinéarité, il est peu recommandé de considérer des termes cubiques.

6.3 Influence, résidus, validation

Avant toute tentative de modélisation complexe, il est impératif d'avoir conduit des analyses uni et bivariées afin d'identifier des problèmes sur les distributions de chacune des variables : dissymétrie, valeurs atypiques (outliers) ou sur les liaisons des variables prises deux par deux : non-linéarité. Ces préliminaires acquis, des aides ou diagnostics associés à la régression linéaire multiple permettent de détecter des violations d'hypothèses (homoscédasticité, linéarité) ou des points influents dans ce contexte multidimensionnel.

Points influents

Comme toute méthode quadratique, l'estimation des paramètres est très sensible à la présence de points extrêmes susceptibles de perturber gravement les résultats. Une observation est *influente* sur les paramètres d'une régression si, à la fois,

- elle est éloignée du barycentre, et ce dans la direction d'un vecteur propre associé à une petite valeur propre (effet levier),
- elle provoque un grand résidu.

L'observation de la diagonale de la matrice $\mathbf{H}$ (hat matrix) révèle un effet levier potentiel tandis que l'analyse des résidus *studentisés* pointe ceux susceptibles de poser des problèmes (valeur absolue plus grande que 2).

Les deux diagnostics précédents sont combinés dans des mesures synthétiques proposées par différents auteurs. La plus utilisée est la distance de *Cook*

$$D_i = \frac{1}{s^2(p+1)}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}) = \left[\frac{h_i^i}{1-h_i^i}\right]\frac{r_i^2}{(p+1)}$$

qui quantifie l'influence de la i-ème observation sur l'écart entre le prédicteur $\widehat{\mathbf{y}}$ et le prédicteur $\widehat{\mathbf{y}}_{(i)}$ calculé sans cette $i$ème observation. On conclut à une influence de l'observation i lorsque la valeur de D_i dépasse 1.

Régression partielle

Un modèle de régression multiple est une technique *linéaire*. Il est raisonnable de s'interroger sur la pertinence du caractère linéaire de la contribution d'une variable explicative à l'ajustement du modèle. Ceci peut être réalisé en considérant une *régression partielle*.

On calcule alors deux régressions :

- la régression de Y sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$, dans laquelle la $j$ème variable est omise, soit $\mathbf{r}_{y(j)}$ le vecteur des résidus obtenus.
- La régression de X^j sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$. Soit $\mathbf{r}_{x(j)}$ le vecteur des résidus obtenus.

La comparaison des résidus par un graphe (nuage de points $\mathbf{r}_{y(j)} \times \mathbf{r}_{x(j)}$) permet alors de représenter la nature de la liaison entre X^j et Y *conditionnellement* aux autres variables explicatives du modèle.

Graphes

Différents graphiques permettent finalement de contrôler le bien fondé des hypothèses de linéarité, d'homoscédasticité, éventuellement de normalité des résidus.

- Le premier considère le nuage de points des résidus studentisés croisés avec les valeurs prédites. Les points doivent être uniformément répartis entre les bornes -2 et $+2$ et ne pas présenter de formes suspectes.
- Le deuxième croise les valeurs observées de Y avec les valeurs prédites. Il illustre le coefficient de détermination R qui est aussi la corrélation linéaire simple entre $\widehat{\mathbf{y}}$ et $\mathbf{y}$. Les points doivent s'aligner autour de la première bissectrice. Il peut être complété par l'intervalle de confiance des y_i ou celui de leurs moyennes.
- La qualité, en terme de linéarité, de l'apport de chaque variable est étudiée par des régressions partielles. Chaque graphe de résidus peut être complété par une estimation fonctionnelle ou régression non-paramétrique (loess, noyau, spline) afin d'en facilité la lecture.
- Le dernier trace la droite de Henri (Normal QQplot) des résidus dont le caractère linéaire de la représentation donne une idée de la normalité de la distribution.

7 Analyse de covariance

L'*analyse de covariance* se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, ont peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynômiale. Le principe général est toujours d'estimer des modèles "*intra-groupes*" et de faire apparaître (tester) des effets différentiels "*inter-groupes*" des paramètres des régressions. Ainsi, dans le cas plus simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

7.1 Modèle

Le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à J niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y ; $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon.

En pratique, avant de lancer une procédure de modélisation et tests, une démarche exploratoire s'appuyant sur une représentation en couleur (une par modalité j de T) du nuage de points croisant Y et X et associant les droites de régression permet de se faire une idée sur les effets respectifs des variables : parallélisme des droites, étirement, imbrication des sous-nuages.

On suppose que les moyennes conditionnelles $E[Y|T]$, c'est-à-dire calculées à l'intérieur de chaque cellule, sont dans le sous-espace vectoriel engendré par les variables explicatives quantitatives, ici X. Ceci s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \ ; \quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij}|i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{x}$ le vecteur $[x_{ij}|i = 1, n_j; j = 1, J]'$, $\mathbf{u} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. On note encore $\mathbf{x}.\mathbf{1}_j$ le produit terme à terme des deux vecteurs, c'est-à-dire le vecteur contenant les observations de $\mathbf{X}$ sur les individus prenant le niveau j de T et des zéros ailleurs.

La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le modèle :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

dans lequel $\mathbf{X}$ est la matrice $n \times 2J$ constituée des blocs $[\mathbf{1}_j|\mathbf{x}.\mathbf{1}_j]$; $j = 1, \ldots, J$. L'estimation de ce modèle global conduit, par bloc, à estimer les modèles de régression dans chacune des cellules.

Comme pour l'analyse de variance, les logiciels opèrent une reparamétrisation faisant apparaître des effets différentiels par rapport au dernier niveau (SAS/GLM, SAS/INSIGHT) ou par rapport à un effet moyen (Systat), afin d'obtenir directement les bonnes hypothèses dans les tests. Ainsi, dans le premier cas, on considère la matrice de même rang (sans la Jème indicatrice)

$$\mathbf{X} = [\mathbf{1}|\mathbf{x}|\mathbf{1}_1|\cdots|\mathbf{1}_{J-1}|\mathbf{x}.\mathbf{1}_1|\cdots|\mathbf{x}.\mathbf{1}_{J-1}]$$

associée aux modèles :

$$y_{ij} = \beta_{0J} + (\beta_{0j} - \beta_{0J}) + \beta_{1J} x_{ij} + (\beta_{1j} - \beta_{1J}) x_{ij} + \varepsilon_{ij} \ ; \quad j = 1, \ldots, J-1; i = 1, \ldots, n_j.$$

7.2 Tests

Différentes hypothèses sont alors testées en comparant le modèle complet

$$\begin{aligned} \mathbf{y} = \beta_{0J}\mathbf{1} \quad &+ \quad (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \\ &+ \quad (\beta_{11} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u} \end{aligned}$$

à chacun des modèles réduits :

$$
\begin{array}{ll}
(i) & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \mathbf{u} \\
(ii) & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \\
 & +(\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u} \\
(iii) & \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u}
\end{array}
$$

par un test de Fisher. Ceci revient à considérer les hypothèses suivantes :

- H_0^i : pas d'interaction, $\beta_{11} = \cdots = \beta_{1J}$, les droites partagent la même pente β_{1J},
- H_0^{ii} : β_{1J}=0,
- H_0^{iii} :$\beta_{01} = \cdots = \beta_{0J}$, les droites partagent la même constante à l'origine β_{0J}.

On commence donc par évaluer i), si le test n'est pas significatif, on regarde ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si i) n'est pas significatif, on s'intéresse à iii) pour juger de l'effet du facteur T.

7.3 Cas général

Ce cadre théorique et les outils informatiques (SAS/GLM) permettent de considérer des modèles beaucoup plus complexes incluant plusieurs facteurs, plusieurs variables quantitatives, voire des polynômes de celles-ci, ainsi que les diverses interactions entre qualitatives et quantitatives. Le choix du "bon" modèle devient vite complexe d'autant que la stratégie dépend, comme pour la régression linéaire multiple, de l'objectif visé :

descriptif : des outils multidimensionnels descriptifs (ACP, AFD, AFCM…) s'avèrent souvent plus efficaces pour sélectionner, en première approche, un sous-ensemble de variables explicatives avant d'opérer une modélisation,

explicatif : de la prudence est requise d'autant que les hypothèses ne peuvent être évaluées de façon indépendante surtout si, en plus, des cellules sont déséquilibrées ou vides,

prédictif : la recherche d'un modèle efficace, donc parcimonieux, peut conduire à négliger des interactions ou effets principaux lorsqu'une faible amélioration du R^2 le justifie et même si le test correspondant apparaît comme significatif. L'utilisation du C_p est possible mais en général ce critère n'est pas calculé et d'utilisation délicate pour définir ce qu'est le "vrai" modèle de référence. En revanche, le PRESS donne des indications pertinentes. L'algorithme de recherche descendant est le plus couramment utilisé avec la contrainte suivante : *un effet principal est supprimé qu'à la qu'il n'apparaisse plus dans une interaction.*

8 Exemples

8.1 Données comptables

Elles sont extraites de Jobson (1991) et décrivent les résultats comptables de 40 entreprises du Royaume Uni.

RETCAP	Return on capital employed
WCFTDT	Ratio of working capital flow to total debt
LOGSALE	Log to base 10 of total sales
LOGASST	Log to base 10 of total assets
CURRAT	Current ratio
QUIKRAT	Quick ratio
NFATAST	Ratio of net fixed assets to total assets
FATTOT	Gross sixed assets to total assets
PAYOUT	Payout ratio
WCFTCL	Ratio of working capital flow to total current liabilities
GEARRAT	Gearing ratio (debt-equity ratio)
CAPINT	Capital intensity (ratio of total sales to total assets)
INVTAST	Ratio of total inventories to total assets

Modèle complet

La procédure SAS/REG est utilisée dans le programme suivant. La plupart des options sont actives afin de fournir la plupart des résultats même si certains sont redondants ou peu utiles.

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp1 all;
 model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       /dw covb  Influence cli clm tol vif collin R P;
output out=resout h=lev p=pred r=res student=resstu ;
run;
```

```
Analysis of Variance
                           Sum of          Mean
Source          DF        Squares         Square         F Value       Prob>F
               (1)
Model           12       0.55868 (2)     0.04656 (5)    8.408 (7)     0.0001 (8)
Error           27       0.14951 (3)     0.00554 (6)
C Total         39       0.70820 (4)
  Root MSE       0.07441 (9)    R-square       0.7889 (12)
  Dep Mean       0.14275 (10)    Adj R-sq      0.6951 (13)
  C.V.          52.12940 (11)
```

(1)	degrés de liberté de la loi de Fisher du test global
(2)	SSR
(3)	SSE ou déviance
(4)	SST=SSE+SSR
(5)	SSR/DF
(6)	s^2 =MSE=SSE/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10)
(12)	Coefficient de détermination R^2
(13)	Coefficient de détermination ajusté ${R'}^2$

Parameter Estimates

Variable	DF	Parameter Estimate (1)	Standard Error (2)	T for H0: Parameter=0 (3)	Prob>|T| (4)	Tolerance (5)	Variance Inflation (6)
INTERCEP	1	0.188072	0.13391661	1.404	0.1716	.	0.00000000
WCFTCL	1	0.215130	0.19788455	1.087	0.2866	0.03734409	26.77799793
WCFTDT	1	0.305557	0.29736579	1.028	0.3133	0.02187972	45.70441500
GEARRAT	1	-0.040436	0.07677092	-0.527	0.6027	0.45778579	2.18442778
LOGSALE	1	0.118440	0.03611612	3.279	0.0029	0.10629382	9.40788501
LOGASST	1	-0.076960	0.04517414	-1.704	0.0999	0.21200778	4.71680805

...

(1)	estimations des paramètres (b_j)
(2)	écarts-types de ces estimations (s_{b_j})
(3)	statistique T du test de Student de $H_0\ : b_j = 0$
(4)	$P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$
(5)	$1 - R^2_{(j)}$
(6)	VIF=$1/(1 - R^2_{(j)})$

Ces résultats soulignent les problèmes de colinéarités. De grands "VIF" sont associés à de grands écart-types des estimations des paramètres. D'autre part les nombreux tests de Student non significatifs renforcent l'idée que trop de variables sont présentes dans le modèle. Cette idée est renforcée par le calcul de l'indice de conditionnement (8.76623/0.00125).

Obs	Dep Var RETCAP (1)	Predict Value (2)	Std Err Predict (3)	Lower95 Mean (4)	Upper95 Mean (5)	Lower95 Predict (6)	Upper95 Predict (7)	Residual (8)	Std Err Residual (9)	Student Residual (10)
1	0.2600	0.2716	0.053	0.1625	0.3808	0.0839	0.4593	-0.0116	0.052	-0.223
2	0.5700	0.3690	0.039	0.2882	0.4497	0.1962	0.5417	0.2010	0.063	3.183
3	0.0900	0.00897	0.063	-0.1205	0.1385	-0.1912	0.2092	0.0810	0.039	2.055
4	0.3200	0.2335	0.021	0.1903	0.2768	0.0748	0.3922	0.0865	0.071	1.212
5	0.1700	0.1164	0.046	0.0215	0.2113	-0.0634	0.2961	0.0536	0.058	0.920

...

```
                          Cook's           Hat Diag    Cov          INTERCEP   WCFTCL    WCFTDT
Obs   -2-1-0 1 2            D     Rstudent     H      Ratio  Dffits  Dfbetas   Dfbetas   Dfbetas
         (11)             (12)      (13)      (14)     (15)   (15)     (15)      (15)      (15)
 1   |       |       | 0.004   -0.2194    0.5109   3.2603  -0.2242   0.0299    0.0632 -0.0911
 2   |       |******| 0.302     3.9515    0.2795   0.0050   2.4611   0.9316   -0.3621  0.3705
 3   |       |****  | 0.832     2.1955    0.7192   0.6375   3.5134   0.5543    2.1916 -2.0241
 4   |       |**    | 0.010     1.2228    0.0803   0.8585   0.3613  -0.0132   -0.0835  0.1207
 5   |       |*     | 0.041     0.9175    0.3864   1.7591   0.7280  -0.0386    0.0906  0.0060
        ...
```

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y}_i$
(3)	écart-type de cette estimation$s_{\widehat{y}_i}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés e_i
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau **H**
(15)	autres indicateurs d'influence

```
Sum of Residuals                      0
Sum of Squared Residuals         0.1495  (SSE)
Predicted Resid SS (Press)       1.0190  (PRESS)
```

Sélection du modèle

Parmi les trois types d'algorithmes et les différents critères de choix, une des façons les plus efficaces consistent à choisir les options du programme ci-dessous. Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur p du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p ou BIC ou ...).

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp2 ;
model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       / selection=rsquare cp rsquare bic best=1;
run;
```

```
   N = 40     Regression Models for Dependent Variable: RETCAP
  R-square Adjusted  C(p)     BIC    Variables in Model
In         R-square
 1  0.1055 0.0819 78.3930 -163.26 WCFTCL
 2  0.3406 0.3050 50.3232 -173.72 WCFTDT QUIKRAT
 3  0.6154 0.5833 17.1815 -191.14 WCFTCL NFATAST CURRAT
 4  0.7207 0.6888  5.7146 -199.20 WCFTDT LOGSALE NFATAST CURRAT
 5  0.7317 0.6923  6.3047 -198.05 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
 6  0.7483 0.7025  6.1878 -197.25 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
 7  0.7600 0.7075  6.6916 -195.77 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
 8  0.7692 0.7097  7.5072 -193.87 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
 9  0.7760 0.7088  8.6415 -191.59 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                                  CURRAT
10  0.7830 0.7082  9.7448 -189.15 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                                  QUIKRAT CURRAT
11  0.7867 0.7029 11.2774 -186.40 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                                  PAYOUT QUIKRAT CURRAT
12  0.7888 0.6950 13.0000 -183.51 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                                  INVTAST PAYOUT QUIKRAT CURRAT
```

Dans cet example, C_p et BIC se comportent de la même façon. Avec peu de variables, le modèle est trop biaisé. Ils atteignent un minimum pour un modèle à 4 variables explicatives puis croissent de nouveau selon la première bissectrice. La maximisation du R^2 ajusté conduirait à une solution beaucoup moins parcimonieuse. On note par ailleurs que l'algorithme remplace WCFTCL par WCFTDT. Un algorithme par sélection ne peut pas aboutir à la solution optimale retenue.

8.2 Données marketing

Les données, extraites de Jobson (1991), sont issues d'une étude marketing visant à étudier l'impact de différentes campagnes publicitaires sur les ventes de différents aliments. Un échantillon ou "panel" de familles a été constitué en tenant compte du lieu d'habitation ainsi que de la constitution de la famille. Chaque semaine, chacune de ces familles ont rempli un questionnaire décrivant les achats réalisés. Nous nous limitons ici à l'étude de l'impact sur la *consommation de lait* de quatre campagnes diffusées sur des chaînes locales de télévision. Quatre villes, une par campagne publicitaire, ont été choisies dans cinq différentes régions géographiques. Les consommations en lait par chacune des six familles par ville alors été mesurées (en dollars) après deux mois de campagne.

Les données se présentent sous la forme d'un tableau à 6 variables : la région géographique, les 4 consommations pour chacune des villes ou campagnes publicitaires diffusées, la taille de la famille. Cette situation est celle classique d'un modèle danalyse de variance. Nous choisissons ici de conserver quantitative la variable taille de la famille et donc de modéliser la consommation de lait par un modèle d'analyse de covariance plus *économique* en dgrés de liberté moins de paramètres sont à estimer.

On s'intéresse à différents modèles de régression visant à expliquer la consommation en fonction de la taille de la famille conditionnellement au type de campagne publicitaire.

```
proc glm data=sasuser.milk;
class pub;
model consom=pub taille pub*taille;
run;
```

Les résultats ci-dessous conduiraient à conclure à une forte influence de la taille mais à l'absence d'influence du type de campagne. Les droites de régression ne semblent pas significativement différentes.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PUB	3	227.1807	75.7269	0.57	0.6377 (1)
TAILLE	1	40926.0157	40926.0157	306.57	0.0001 (2)
TAILLE*PUB	3	309.8451	103.2817	0.77	0.5111 (3)

(1)	Test de la significativité des différences des termes constants.
(2)	Test de l'influence du facteur quantitatif.
(3)	Test de la significativité des différences des pentes (interaction).

Néanmoins, le même calcul est effectué séparément pour chaque région :

```
proc glm data=sasuser.milk;
by region;
class pub;
model consom=pub taille pub*taille;
run;
```

Région	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	PUB	3	72.02974	24.00991	4.62	0.0164
1	TAILLE	1	7178.32142	7178.32142	1380.25	0.0001

	TAILLE*PUB	3	217.37048	72.45683	13.93	0.0001
	PUB	3	231.73422	77.24474	30.36	0.0001
2	TAILLE	1	8655.25201	8655.25201	3402.34	0.0001
	TAILLE*PUB	3	50.15069	16.71690	6.57	0.0042
	PUB	3	79.54688	26.51563	6.01	0.0061
3	TAILLE	1	6993.30160	6993.30160	1585.35	0.0001
	TAILLE*PUB	3	173.19305	57.73102	13.09	0.0001
	PUB	3	415.66664	138.55555	15.23	0.0001
4	TAILLE	1	9743.37830	9743.37830	1071.32	0.0001
	TAILLE*PUB	3	361.39556	120.46519	13.25	0.0001
	PUB	3	15.35494	5.11831	0.79	0.5168
5	TAILLE	1	8513.28516	8513.28516	1314.71	0.0001
	TAILLE*PUB	3	52.75119	17.58373	2.72	0.0793

Il apparaît alors qu'à l'intérieur de chaque région (sauf région 5), les campagnes de publicité ont un effet tant sur la constante que sur la pente.

Ceci incite donc à se méfier des *interactions* et encourage à toujours conserver le facteur bloc dans une analyse de variance. Une approche complète, considérant *a priori* toutes les variables (3 facteurs), est ici nécessaire (cf. TP).

Chapitre 9

Régression logistique

1 Introduction

Dans ce chapitre, nous définissons le contexte pratique de la *régression logistique* qui s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs comme, par exemple, le nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais. Nous laissons de coté le *modèle log-linéaire* qui vise à expliquer un nombre d'individus prenant une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs. Contrairement aux modèles du chapitre précédent basés sur l'hypothèse de normalité des observations, les lois concernées sont discrètes et associées à des dénombrements : loi de Poisson, binomiale, multinomiale. Néanmoins, tous ces modèles appartiennent à la famille des *modèles linéaires généralisés* et partagent à ce titre beaucoup d'aspects (estimation, tests, diagnostic) et dont la stratégie de mise en œuvre, similaire au cas gaussien, n'est pas reprise.

2 Odds et odds ratio

Une variable

Soit Y une variable qualitative à J modalités. On désigne la chance ou *l'odds*[1] de voir se réaliser la $j$ème modalité plutôt que la $k$ème par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k}$$

où π_j est la probabilité d'apparition de la $j$ème modalité. Cette quantité est estimée par le rapport n_j/n_k des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernouilli de paramètre π, l'odds est le rapport $\pi/(1-\pi)$ qui exprime une cote ou chance de gain.

[1] Il n'existe pas, même en Québécois, de traduction consensuelle de "odds".

Table de contingence

On considère maintenant une table de contingence 2×2 croisant deux variables qualitatives binaires X^1 et X^2. les paramètres de la loi conjointe se mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X^1 = i\} \text{ et } \{X^2 = j\}]$ est la probabilité d'occurence de chaque combinaison.

- Dans la ligne 1, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$

- Dans la ligne 2, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}}.$$

On appelle odds ratio le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Ce rapport prend la valeur 1 si les variables sont indépendantes, il est supérieur à 1 si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon.

L'odds ratio est également défini pour deux lignes (a, b) et deux colonnes (c, d) quelconques d'une table de contingence croisant deux variables à J et K modalités. L'odds ratio est le rapport

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \quad \text{estimé par l'odds ratio empirique} \quad \widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

3 Régression logistique

3.1 Type de données

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, panne d'un équipement, faillite d'une entreprise, bon ou mauvais client.... Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\boldsymbol{\beta}$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités

$$\pi = P(Z = 1) \quad \text{ou} \quad 1 - \pi = P(Z = 0),$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme :

$$g(\pi_i) = \mathbf{x}_i'\beta.$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.

log-log avec g définie par

$$g(\pi) = \ln[-\ln(1-\pi)]$$

mais cette fonction est dissymétrique.

logit est définie par

$$g(\pi) = \mathit{logit}(\pi) = \ln\frac{\pi}{1-\pi} \quad \text{avec} \quad g^{-1}(x) = \frac{e^x}{1+e^x}.$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1-\pi)$, qui exprime une "cote", est l'*odds* et la *régression logistique* s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

Cette section se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'explication d'une variable qualitative ordinale (plusieurs modalités), l'intervention de variables explicatives avec effet aléatoire, l'utilisation de mesures répétées donc dépendantes, sont à rechercher dans la bibliographie.

3.2 Modèle binomial

On considère, pour $i = 1,\ldots,I$, différentes valeurs *fixées* $x_i^1,\ldots,x_i^q$ des variables explicatives $X^1,\ldots,X^q$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives, c'est-à-dire des facteurs issus d'une planification expérimentale.

Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^I n_i$) de la variable Z qui se mettent sous la forme $y_1/n_1,\ldots,y_I/n_I$ où y_i désigne le nombre de "succès" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, la probabilité π_i de succès est constante. Alors, la variable Y_i sachant n_i et d'espérance $E(y_i) = n_i\pi_i$ suit une loi *binomiale* $\mathcal{B}(n_i,\pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{(n_i-y_i)}.$$

On suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace $\mathrm{vect}\{X^1,\ldots,X^q\}$ engendré par les variables explicatives :

$$\mathit{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1,\ldots,I$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad i = 1,\ldots,I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation $\mathbf{b}$ de $\boldsymbol{\beta}$, il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1 + e^{\mathbf{x}_i'\mathbf{b}}}$$

et ainsi celles des effectifs

$$\widehat{y}_i = n_i \widehat{p}_i.$$

Remarques

i. La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de l'analyse de covariance mixant variables explicatives quantitatives et qualitatives. Ainsi, les logiciels gèrent avec plus ou moins de clarté le choix des variables indicatrices et donc des paramètres estimables ou contrastes associés.

ii. La situation décrite précédemment correspond à l'observation de données *groupées*. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1; i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernouilli remplace la loi binomiale. Certaines méthodes ne sont alors plus applicables et les comportements asymptotiques des distributions des statistiques de test ne sont plus valides, le nombre de paramètres tendant vers l'infini.

4 Choix de modèle

4.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald, cf.bibliographie), sont utilisés de façon analogue au test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (logit), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction.

5 Exemples

Il y a au moins 6 façons différentes d'estimer une régression logistique avec SAS :

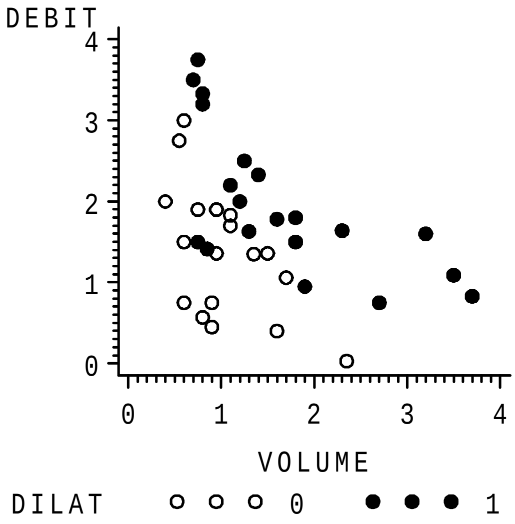


FIG. 9.1 – Nuage des modalités de Y dans les coordonnées des variables explicatives.

`sas/logistic` cette procédure n'est utilisable que lorsque toutes les variables explicatives sont quantitatives, elle inclut en option un algorithme de recherche de modèles par sélection ascendante, descendante ou pas à pas.

`sas/catmod` est adaptée à toute modélisation impliquant des variables qualitatives (modèle log-linéaire, logit, probit...) mais la paramétrisation mise en œuvre rend les résultats difficiles à interpréter.

`sas/genmod` Cette procédure plus récente est arrivée avec la version 6.09. Elles est directement issue des travaux sur le modèle linéaire généralisé (Mc Cullagh et Nelder 1983) et s'inscrit donc dans la logique du logiciel GLIM pour la définition des modèles.

`sas/insight` est, pour le modèle linéaire généralisé, une version interactive et graphique de `sas/genmod`.

Enfin, pour mémoire, il existe en plus une option `logit` dans la procédure `sas/probit` et la possibilité d'utiliser SAS Enterprise Miner qui propose un algorithme de choix de modèle.

5.1 Débits×Volumes

On étudie l'influence du débit et du volume d'air inspiré sur l'occurence (codée 1) de la dilatation des vaisseaux sanguins superficiels des membres inférieurs.

Référence : Pregibon, D. (1981) Logistic regression diagnostics, Annals of Stat., 9, 705-724. Un graphique élémentaire représentant les modalités de Y dans les coordonnées de $X^1 \times X^2$ est toujours instructif. Il montre une séparation raisonnable et de bon augure des deux nuages de points. Dans le cas de nombreuses variables explicatives quantitatives, une analyse en composantes principales s'impose. Les formes des nuages représentés, ainsi que l'allure des distributions (étudiées préalablement), incitent dans ce cas à considérer par la suite les logarithmes des variables. Une variable (un) ne contenant que des "1" dénombrant le nombre d'essais est nécessaire dans la syntaxe de `genmod`. Les données sont en effet non groupées.

```
proc logistic data=sasuser.debvol;
```

```
model dilat=l_debit l_volume;
run;
proc genmod data=sasuser.debvol;
model dilat/un=l_debit l_volume/d=bin;
run;
```

```
The LOGISTIC Procedure
                          Intercept
              Intercept      and
 Criterion      Only      Covariates    Chi-Square for Covariates
 AIC           56.040       35.216            .
 SC            57.703       40.206            .
 -2 LOG L      54.040       29.216(1)     24.824 with 2 DF (p=0.0001)
 Score           .             .          16.635 with 2 DF (p=0.0002)

            Parameter(2) Standard    Wald(3)     Pr >    Standardized  Odds
Variable DF  Estimate    Error   Chi-Square  Chi-Square  Estimate     Ratio
INTERCPT 1    2.8782    1.3214     4.7443     0.0294          .         .
L_DEBIT  1   -4.5649    1.8384     6.1653     0.0130    -2.085068   0.010
L_VOLUME 1   -5.1796    1.8653     7.7105     0.0055    -1.535372   0.006

       Association of Predicted Probabilities and Observed Responses(4)
                Concordant = 93.7          Somers' D = 0.874
                 Discordant =  6.3          Gamma     = 0.874
                 Tied       =  0.0          Tau-a     = 0.448
                 (380 pairs)                c         = 0.937
```

Cette procédure fournit des critères de choix de modèle dont la déviance (1), le vecteur **b** des paramètres (2) et les statistiques des tests (3) comparant le modèle excluant un terme par rapport au modèle complet tel qu'il est décrit dans la commande.

```
The GENMOD Procedure
                Criteria For Assessing Goodness Of Fit
            Criterion            DF         Value      Value/DF
            Deviance             36        29.2156       0.8115 (1)
            Scaled Deviance      36        29.2156       0.8115 (2)
            Pearson Chi-Square   36        34.2516       0.9514 (3)
            Scaled Pearson X2    36        34.2516       0.9514
            Log Likelihood        .       -14.6078          .

                     Analysis Of Parameter Estimates
        Parameter     DF    Estimate (4)  Std Err   ChiSquare (5)  Pr>Chi
        INTERCEPT      1     -2.8782      1.3214       4.7443      0.0294
        L_DEBIT        1      4.5649      1.8384       6.1653      0.0130
        L_VOLUME       1      5.1796      1.8653       7.7105      0.0055
        SCALE (6)      0      1.0000      0.0000          .          .
```

(1)	Déviance du modèle par rapport au modèle saturé.
(2)	Déviance pondérée si le paramètre d'échelle est différent de 1 en cas de sur-dispersion.
(3)	Statistique de Pearson, voisine de la déviance, comparant le modèle au modèle saturé .
(4)	Paramètres du modèle.
(5)	Statistique des tests comparant le modèle excluant un terme par rapport au modèle complet.
(6)	Estimation du paramètre d'échelle si la quasi-vraisemblance est utilisée.

5.2 Données bancaires

Plusieurs stratégies peuvent être mises en œuvre sur les données bancaires. Les premières consistent à ne s'intéresser qu'aux variables quantitatives et à rechercher un "meilleur" modèle à l'aide de la procédure `logistic` en association avec l'un des trois algorithmes de sélection.

```
proc logistic data=sasuser.vispremt;
model carvpr =ager relat opgnbl--dnbjdl/selection=stepwise;
run;
```

Ainsi, l'algorithme qui procède par élimination retient finalement 14 des 20 variables. pour un taux de mal classés de 15,4%. Par sélection ou pas à pas, la même solution est proposée avec 12 variables pour un taux de 15,6%. *Attention*, ces taux, calculés sur l'échantillon ayant servi à estimer les paramètres, sont nécessairement trop optimistes.

À moins d'utiliser SAS Enterprise Miner, la prise en compte des variables qualitatives nécessitent une procédure de choix de modèle manuelle. Le module SAS/INSIGHT est alors le plus efficace pour réaliser une approche descendante avant de réestimer le modèle obtenu à l'aide de `genmod`.

```
proc genmod data=sasuser.vispremt ;
    class sexec  PRCSP;
    make 'OBSTATS' out=outglm;
    model carvpr/poids = SEXEC PRCSP OPGNBL MOYRVL BOPPNL
    DNBJDL UEMNB XLGNB YLVNB JNTCA NPTAG / obstats d=bin;
    run;
/* Estimation du taux de mal classés */
data prev ;
set outglm (keep=yvar1 pred);
if pred ge 0.5 then predy=1; else predy=0;
proc freq data=prev;
tables yvar1*predy/ nocol norow;
run;
```

Les résultats semblent alors bien meilleurs mais il faut rester prudent quant à la précision de cette estimation du taux d'erreur. On pourrait enfin s'interroger sur les qualités d'un modèle considérant toutes les variables qualitatives.

```
 YVAR1(CARVPR)     PREDY
Frequency|
Percent  |       0|       1|  Total
---------+--------+--------+
       0 |    659 |     53 |    712
         |  61.65 |   4.96 |  66.60
---------+--------+--------+
       1 |     70 |    287 |    357
         |   6.55 |  26.85 |  33.40
---------+--------+--------+
Total         729      340     1069
            68.19    31.81   100.00
```

Chapitre 10

Analyse Discriminante Décisionnelle

1 Introduction

Ce chapitre reprend le contexte déjà décrit pour l'analyse factorielle discriminante ; p variables quantitatives explicatives $X^j, j = 1, \ldots, p$ et une variable qualitative T (m modalités) à expliquer sont observées sur un échantillon Ω de taille n. L'objectif de l'analyse discriminante décisionnelle déborde le simple cadre descriprif. Disposant d'un nouvel individu (ou de plusieurs, c'est la même chose) sur lequel on a observé les X^j mais pas T, il s'agit maintenant de *décider* de la modalité $\mathcal{T}_\ell$ de T (ou de la classe correspondante) de ce nouvel individu. On parle aussi de problème d'*affectation*. L'ADD s'applique donc également à la situation précédente de la régression logistique ($m = 2$) mais aussi lorsque le nombre de classes est plus grand que 2.

Pour cela, on va définir et étudier dans ce chapitre des *règles de décision* (ou d'affectation) et donner ensuite les moyens de les évaluer sur un seul individu ; $\mathbf{x} = (x^1, \ldots, x^p)$ désigne les observations des variables explicatives sur cet individu, $\{\mathbf{g}_\ell; \ell = 1, \ldots, m\}$ les barycentres des classes calculés sur l'échantillon et $\overline{\mathbf{x}}$ le barycentre global.

2 Règle de décision issue de l'AFD

2.1 Cas général : m quelconque

Définition 10.1. — *On affectera l'individu x à la modalité de T minimisant :*

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell), \ell = 1, \ldots, m.$$

Cette distance se décompose en

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell) = \|\mathbf{x} - \mathbf{g}_\ell\|^2_{\mathbf{S}_r^{-1}} = (\mathbf{x} - \mathbf{g}_\ell)'\mathbf{S}_r^{-1}(\mathbf{x} - \mathbf{g}_\ell)$$

et le problème revient donc à maximiser

$$\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{g}_\ell.$$

Il s'agit bien d'une règle linéaire en $\mathbf{x}$ car elle peut s'écrire : $\mathbf{A}_\ell\mathbf{x} + \mathbf{b}_\ell$.

2.2 Cas particulier : $m = 2$

Dans ce cas, la dimension r de l'AFD vaut 1. Il n'y a qu'une seule valeur propre non nulle λ_1, un seul vecteur discriminant v^1 et un seul axe discriminant Δ_1. Les 2 barycentres $\mathbf{g}_1$ et $\mathbf{g}_2$ sont sur Δ_1, de sorte que v^1 est colinéaire à $\mathbf{g}_1 - \mathbf{g}_2$.

L'application de la règle de décision permet d'affecter $\mathbf{x}$ à T_1 si :

$$\mathbf{g}_1'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_1'\mathbf{S}_r^{-1}\mathbf{g}_1 > \mathbf{g}_2'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_2'\mathbf{S}_r^{-1}\mathbf{g}_2$$

c'est-à-dire encore si

$$(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{S}_r^{-1}\mathbf{x} > (\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{S}_r^{-1}\frac{\mathbf{g}_1 + \mathbf{g}_2}{2}.$$

Remarque

La règle de décision liée à l'AFD est simple mais elle est limitée et insuffisante notamment si les variances des classes ne sont pas identiques. De plus, elle ne tient pas compte de l'échantillonnage pour $\mathbf{x}$: tous les groupes n'ont pas nécessairement la même probabilité d'occurence.

3 Règle de décision bayésienne

3.1 Introduction

Dans cette optique, on considère que la variable T, qui indique le groupe d'appartenance d'un individu, prend ses valeurs dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ et est munie d'une loi de probabilité $\pi_1, \ldots, \pi_m$. Les probabilités $\pi_\ell = P[\mathcal{T}_\ell]$ représentent les probabilités *a priori* des classes ou groupes ω_ℓ. On suppose que les vecteurs $\mathbf{x}$ des observations des variables explicatives suivent, connaissant leur classe, une loi de densité

$$f_\ell(\mathbf{x}) = P[\mathbf{x} \mid \mathcal{T}_\ell]$$

par rapport à une mesure de référence[1].

3.2 Définition

Une règle de décision est une application δ de Ω dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ qui, à tout individu, lui affecte une classe connaissant $\mathbf{x}$. Sa définition dépend du contexte de l'étude et prend en compte la

- connaissance ou non de coûts de mauvais classement,
- connaissance ou non des lois *a priori* sur les classes,
- nature aléatoire ou non de l'échantillon.

On désigne par $c_{\ell \mid k}$ le coût du classement dans $\mathcal{T}_\ell$ d'un individu de $\mathcal{T}_k$. Le *risque de Bayes* d'une règle de décision δ exprime alors le coût moyen :

$$R_\delta = \sum_{k=1}^{m} \pi_k \sum_{\ell=1}^{m} c_{\ell \mid k} \int_{\{\mathbf{x} \mid \delta(\mathbf{x}) = \mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$$

[1] La mesure de Lebesgues pour des variables réelles, celle de comptage pour des variables qualitatives

où $\int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} f_k(\mathbf{x})d\mathbf{x}$ représente la probabilité d'affecté $\mathbf{x}$ à $\mathcal{T}_\ell$ alors qu'il est dans $\mathcal{T}_k$.

3.3 Coûts inconnus

L'estimation des coûts n'est pas du ressort de la Statistique et, s'ils ne sont pas connus, on suppose simplement qu'ils sont tous égaux. La minimisation du risque ou règle de Bayes revient alors à affecter tout $\mathbf{x}$ à la classe la plus probable c'est-à-dire à celle qui maximise la probabilité conditionnelle *a posteriori* : $P[\mathcal{T}_\ell \mid \mathbf{x}]$. Par le théorème de Bayes, on a :

$$P[\mathcal{T}_\ell \mid \mathbf{x}] = \frac{P[\mathcal{T}_\ell \text{ et } \mathbf{x}]}{P[\mathbf{x}]} = \frac{P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]}{P[\mathbf{x}]}$$

avec le principe des probabilités totales : $P[\mathbf{x}] = \sum_{\ell=1}^{m} P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]$.

Comme $P[\mathbf{x}]$ ne dépend pas de ℓ, la règle consistera à choisir $\mathcal{T}_\ell$ maximisant

$$P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell] = \pi_\ell.P[\mathbf{x} \mid \mathcal{T}_\ell];$$

$P[\mathbf{x} \mid \mathcal{T}_\ell]$ est la probabilité d'observer $\mathbf{x}$ au sein de la classe $\mathcal{T}_\ell$. Pour une loi discrète, il s'agit d'une probabilité du type $P[\mathbf{x} = \mathbf{x}_k^l \mid \mathcal{T}_\ell]$ et d'une densité $f(\mathbf{x} \mid \mathcal{T}_\ell)$ pour une loi continue. Dans tous les cas nous utiliserons la notation $f_\ell(\mathbf{x})$.

La règle de décision s'écrit finalement sous la forme :

$$\delta(\mathbf{x}) = \arg \max_{\ell=1,\ldots,m} \pi_\ell f_\ell(\mathbf{x}).$$

3.4 Détermination des *a priori*

Les probabilités *a priori* π_ℓ peuvent effectivement être connues *a priori* : proportions de divers groupes dans une population, de diverses maladies… ; sinon elles sont estimées sur l'échantillon d'apprentissage :

$$\widehat{\pi}_\ell = w_\ell = \frac{n_\ell}{n} \quad \text{(si tous les individus ont le même poids)}$$

à condition qu'il soit bien un échantillon aléatoire susceptible de fournir des estimations correctes des fréquences. Dans le cas contraire il reste à considérer tous les π_ℓ égaux.

3.5 Cas particuliers

- Dans le cas où les probabilités *a priori* sont égales, c'est par exemple le cas du choix de probabilités non informatives, la règle de décision bayésienne revient alors à maximiser $f_\ell(\mathbf{x})$ qui est la vraisemblance, au sein de T_ℓ, de l'observation $\mathbf{x}$. La règle consiste alors à choisir la classe pour laquelle cette vraisemblance est maximum.
- Dans le cas où $m = 2$, on affecte $\mathbf{x}$ à T_1 si :

  $$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

  faisant ainsi apparaître un rapport de vraisemblance. D'autre part, l'introduction de coûts de mauvais classement différents selon les classes amène à modifier la valeur limite π_2/π_1.

Finalement, il reste à estimer les densités conditionnelles $f_\ell(\mathbf{x})$. Les différentes méthodes d'estimation considérées conduisent aux méthodes classiques de discrimination bayésienne objets des sections suivantes.

4 Règle bayésienne avec modèle normal

On suppose dans cette section que, conditionnellement à T_ℓ, $\mathbf{x} = (x_1, \dots, x_p)$ est l'observation d'un vecteur aléatoire gaussien $\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$; $\boldsymbol{\mu}_\ell$ est un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\Sigma}_\ell$ une matrice $(p \times p)$ symétrique et définie-positive. La densité de la loi, au sein de la classe $\mathcal{T}_\ell$, s'écrit donc :

$$f_\ell(\mathbf{x}) = \frac{1}{\sqrt{2\pi}(\det(\boldsymbol{\Sigma}_\ell))^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)\right].$$

L'affectation de $\mathbf{x}$ à une classe se fait en maximisant $\pi_\ell.f_\ell(\mathbf{x})$ par rapport à l soit encore la quantité :

$$\ln(\pi_\ell) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_\ell)) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell).$$

4.1 Hétéroscédasticité

Dans le cas général, il n'y a pas d'hypothèse supplémentaire sur la loi de $\mathbf{x}$ et donc les matrices $\boldsymbol{\Sigma}_\ell$ sont fonction de ℓ. Le critère d'affectation est alors *quadratique* en $\mathbf{x}$. Les probabilités π_ℓ sont supposées connues mais il est nécessaire d'estimer les moyennes $\boldsymbol{\mu}_\ell$ ainsi que les covariances $\boldsymbol{\Sigma}_\ell$ en maximisant, compte tenu de l'hypothèse de normalité, la vraisemblance. Ceci conduit à estimer la moyenne

$$\widehat{\boldsymbol{\mu}_\ell} = \mathbf{g}_\ell$$

par la moyenne empirique de $\mathbf{x}$ dans la classe l pour l'échantillon d'apprentissage et $\boldsymbol{\Sigma}_\ell$ par la matrice de covariance empirique $\mathbf{S}^*_{Rl}$:

$$\mathbf{S}^*_{Rl} = \frac{1}{n_\ell - 1}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

pour ce même échantillon.

4.2 Homoscédasticité

On suppose dans ce cas que les lois de chaque classe partagent la même structure de covariance $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$. Supprimant les termes indépendants de l, le critère à maximiser devient

$$\ln(\pi_\ell) - \frac{1}{2}\boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell + \boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}_\ell^{-1}\mathbf{x}$$

qui est cette fois *linéaire* en $\mathbf{x}$. Les moyennes $\boldsymbol{\mu}_\ell$ sont estimées comme précédemment tandis que $\boldsymbol{\Sigma}$ est estimée par la matrice de covariance intra empirique :

$$\mathbf{S}^*_R = \frac{1}{n-m}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'.$$

Si, de plus, les probabilités π_ℓ sont égales, après estimation le critère s'écrit :

$$\overline{\mathbf{x}_\ell}'\mathbf{S}_R^{*-1}\mathbf{x} - \frac{1}{2}\overline{\mathbf{x}_\ell}'\mathbf{S}_R^{*-1}\overline{\mathbf{x}_\ell}.$$

On retrouve alors le critère de la section 2 issu de l'AFD.

4.3 Commentaire

Les hypothèses : normalité, éventuellement l'homoscédasticité, doivent être vérifiées par la connaissance *a priori* du phénomène ou par une étude préalable de l'échantillon d'apprentissage. L'hypothèse d'homoscédasticité, lorqu'elle est vérifiée, permet de réduire très sensiblement le nombre de paramètres à estimer et d'aboutir à des estimateurs plus fiables car de variance moins élevée. Dans le cas contraire, l'échantillon d'apprentissage doit être de taille importante.

5 Règle bayésienne avec estimation non paramétrique

5.1 Introduction

En Statistique, on parle d'estimation non paramétrique ou fonctionnelle lorsque le nombre de paramètres à estimer est infini. L'objet statistique à estimer est alors une fonction par exemple de régression $y = f(x)$ ou encore une densité de probabilité. Dans ce cas, au lieu de supposer qu'on a affaire à une densité de type connu (normale) dont on estime les paramètres, on cherche une estimation $\widehat{f}$ de la fonction de densité f. Pour tout x de $\mathbb{R}$, $f(x)$ est donc estimée par $\widehat{f}(x)$.

Cette approche très souple a l'avantage de ne pas nécessiter d'hypothèse particulière sur la loi (seulement la régularité de f pour de bonnes propriétés de convergence), en revanche elle n'est applicable qu'avec des échantillons de grande taille d'autant plus que le nombre de dimensions p est grand (curse of dimensionality).

Dans le cadre de l'analyse discriminante, ces méthodes permettent d'estimer directement les densités $f_\ell(\mathbf{x})$. On considère ici deux approches : la méthode du noyau et celle des k plus proches voisins.

5.2 Méthode du noyau

Estimation de densité

Soit $y_1, \ldots, y_n$ n observations équipondérées d'une v.a.r. continue Y de densité f inconnue. Soit $K(y)$ (le *noyau*) une densité de probabilité unidimensionnelle (sans rapport avec f) et h un réel strictement positif. On appelle estimation de f par la méthode du noyau la fonction

$$\widehat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

Il est immédiat de vérifier que

$$\forall y \in \mathbb{R}, \widehat{f}(y) \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} \widehat{f}(y)dy = 1;$$

h est appelé *largeur de fenêtre* ou paramètre de *lissage*; plus h est grand, plus l'estimation $\widehat{f}$ de f est régulière. Le noyau K est choisi centré en 0, unimodal et symétrique. Les cas les plus usuels sont la densité gaussienne, celle uniforme sur $[-1, 1]$ ou triangulaire : $K(x) = [1 - |x|]\mathbf{1}_{[-1,1]}(x)$. La forme du noyau n'est pas très déterminante sur la qualité de l''estimation contrairement à la valeur de h.

Application à l'analyse discriminante

La méthode du noyau est utilisée pour calculer une estimation non paramétrique de chaque densité $f_\ell(\mathbf{x})$ qui sont alors des fonctions définies dans $\mathbb{R}^p$. Le noyau K^* dont donc être choisi multidimensionnel et

$$\widehat{f}_\ell(\mathbf{x}) = \frac{1}{n_\ell h^p} \sum_{i\in\Omega_\ell} K^* \left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right).$$

Un noyau multidimensionnel peut être défini à partir de la densité usuelle de lois : multinormale $\mathcal{N}_p(0,\Sigma_p)$ ou uniforme sur la sphère unité ou encore par produit de noyaux unidimensionnels :

$$K^*(\mathbf{x}) = \prod_{j=1}^{p} K(x^j).$$

5.3 k plus proches voisins

Cette méthode d'affectation d'un vecteur $\mathbf{x}$ consiste à enchaîner les étapes suivantes :

i. choix d'un entier $k : 1 \geq k \geq n$;

ii. calculer les distances $d_{\mathbf{M}}(\mathbf{x},\mathbf{x}_i)$, $i = 1,\dots,n$ où $\mathbf{M}$ est la métrique de Mahalanobis c'est-à-dire la matrice inverse de la matrice de variance (ou de variance intra) ;

iii. retenir les k observations $\mathbf{x}_{(1)},\dots,\mathbf{x}_{(k)}$ pour lesquelles ces distances sont les plus petites

iv. compter les nombres de fois $k_1,\dots,k_m$ que ces k observations apparaissent dans chacune des classes,

v. estimer les densités par

$$\widehat{f}_\ell(\mathbf{x}) = \frac{k_\ell}{kV_k(\mathbf{x})};$$

où $V_k(\mathbf{x})$ est le volume de l'ellipsoïde $\{\mathbf{z}|(\mathbf{z}-\mathbf{x})'\mathbf{M}(\mathbf{z}-\mathbf{x}) = d_{\mathbf{M}}(\mathbf{x},\mathbf{x}_{(k)})\}$.

Pour $k = 1$, $\mathbf{x}$ est affecté à la classe du plus proche élément.

Comme toute technique, celles présentées ci-dessus nécessitent le réglage d'un paramètre (largeur de fenêtre, nombre de voisins considérés). Ce choix s'apparente à un choix de modèle et nécessite le même type d'approche à savoir l'optiomisation d'un critère (erreur de classement, validation croisée,...). Ce points sera abordé dans un prochain chapitre.

6 Exemple

Une première étape de traitement des données bancaires permet tout d'abord de sélectionner par élimination un sous-ensemble des variables à l'aide de la procédure `stepdisc`. La variable qualitative `sexe` est considérée comme une variable quantitative (0, 1). Ceci pourrait,abusivement mais fréquent en pratique, se généraliser à d'autres variables qualitatives codées numériquement. Les variables discriminantes n'ont plus guère de sens mais, si la discrimination fonctionne...

```
proc stepdisc data=sasuser.vispremt;
class carvp;
var sexer ager relat opgnbl--dnbjdl;
run;
```

Les variables ainsi sélectionnées sont utilisées dans deux algorithmes de discrimination. Le premier, non-paramétrique, utilise les k plus proches voisins tandis que le deuxième fait implicitement l'hypothèse de normalité des distributions. dans les deux cas, une pseudo procédure de validation croisée permet d'estimer les taux de mauvais classement. Il ne s'agit en effet pas d'une procédure de validation croisée explicite car les matrices de variances sont calculées une fois pour toute et donc dépendent des individus à prévoir.

```
proc discrim data= sasuser.vispremt
  method=npar k=11 crossvalidate;
  class CARVP;
  var MOYRVL SEXER BOPPNL GAGECL OPGNBL QCREDL
FACANL XLGMTL RELAT HAVEFL GAGEML ENDETL LGAGTL
VIEMTL TAVEPL ITAVCL AGER;
  run;
```

```
                     Error Count Estimates for CARVP:
                          Cnon          Coui         Total
            Rate        0.2191        0.2801        0.2496
            Priors      0.5000        0.5000
```

```
proc discrim data= sasuser.vispremt
  method=NORMAL crossvalidate;
  class CARVP;
  var MOYRVL SEXER BOPPNL GAGECL OPGNBL QCREDL
FACANL XLGMTL RELAT HAVEFL GAGEML ENDETL LGAGTL
VIEMTL TAVEPL ITAVCL AGER;
  run;
```

```
                      Error Count Estimates for CARVP:
                          Cnon          Coui         Total
            Rate        0.1784        0.2689        0.2236
            Priors      0.5000        0.5000
```

La valeur de k pourrait être sans doute améliorée mais il semble que dans ce cas, l'approche paramétrique fasse un peu mieux. La comparaison entre régression logistique et analyse discriminante demande un peu d'attention et surtout la constitution préalable d'un échantillon test.

Chapitre 11

Arbres de classification

1 Introduction

Ce chapitre s'intéresse aux méthodes ayant pour objectif la construction d'*arbres binaires*, ou dendogrammes, modélisant une discrimination ou une régression. Complémentaires des méthodes usuelles : analyse discriminante, régression linéaire, les solutions obtenues sont présentées sous une forme graphique simple à interpréter, même pour des néophytes, et constituent une aide efficace pour l'aide à la décision. Nommées initialement partitionnement récursif ou segmentation, les développements importants de Breiman et col. (1984) les ont fait connaître sous l'acronyme de CART : Classification and Regression Tree. Cela correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (classification) ou quantitative (régression).

Ces méthodes ne sont efficaces que pour des tailles d'échantillons importantes et elles sont très calculatoires. Les deux raisons : modèle graphique de décision simple à interpréter, puissance de calcul nécessaire, suffisent à expliquer leur popularité toute récente. De plus, elles requièrent plutôt moins d'hypothèses que des méthodes classiques et semblent particulièrement adaptées au cas où les variables explicatives sont nombreuses. En effet, la procédure de sélection des variables est intégrée à l'algorithme construisant l'arbre, d'autre part, les interactions sont prises en compte. Néanmoins, cet algorithme suivant une stratégie pas à pas hiérarchisée, il peut, comme dans le cas du choix de modèle en régression, passer à coté d'un optimum global. Ceci souligne encore l'importance de confronter plusieurs approches sur les mêmes données.

Nous nous limiterons dans cet exposé introductif au cas des arbres de classification (variable à expliquer qualitative) dans la situation la plus standard. Les compléments sont à rechercher dans la littérature (Breiman et col. 1984, Celeux et Nakache (1994)). D'autres approches (CHAID) considèrent des arbres où les nœuds peuvent avoir plus de deux nœuds descendants.

2 Construction d'un arbre binaire

2.1 Principe

Les données sont constituées de l'observation de p variables quantitatives ou qualitatives explicatives X^j et d'une variable à expliquer T qualitative, à m modalités $\{\mathcal{T}_\ell ; \ell = 1 \ldots, m\}$, observées sur un échantillon de n individus.

La construction d'un arbre de discrimination binaire (cf. figure 2.2) consiste à déterminer une séquence de *nœuds*.

- Un nœud est défini par le choix conjoint d'une variable parmi les explicatives et d'une *division* qui induit une partition en deux classes.
- Implicitement, à chaque nœud correspond donc un sous-ensemble de l'échantillon auquel est appliquée une dichotomie.
- Une division est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.
- À la racine ou nœud initial correspond l'ensemble de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles.

L'algorithme considéré nécessite :

i. la définition d'un critère permettant de sélectionner la "meilleure" division parmi toutes celles *admissibles* pour les différentes variables ;
ii. une règle permettant de décider qu'un nœud est terminal : il devient ainsi une *feuille* ;
iii. l'affectation de chaque feuille à l'un des groupes de la variable à expliquer.

Le point 2 est le plus délicat, il correspond encore à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une surparamétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations. La contribution majeure de Breiman et col. (1984) est justement une stratégie de recherche d'arbre optimal. Elle consiste à

i. construire l'arbre maximal $A_{\max}$,
ii. ordonner les sous-arbres selon une séquence emboîtée suivant la décroissance d'un critère pénalisé de déviance ou de taux de mal-classés,
iii. puis à sélectionner le sous-arbre optimal ; c'est la procédure d'*élagage*.

Tous ces points sont détaillés ci-dessous.

2.2 Critère de division

Une division est dite *admissible* si aucun des deux nœuds descendants qui en découlent n'est vide. Si la variable explicative est qualitative ordinale avec m modalités, elle fournit $(m-1)$ divisions binaires admissibles. Si elle est seulement nominale le nombre de divisions passe à $2^{(m-1)} - 1$. Une variable quantitative se ramène au cas ordinal.

Le critère de division repose sur la définition d'une fonction d'*hétérogénéité* ou de désordre. L'objectif étant de partager les individus en deux groupes les plus homogènes au sens de la variable à expliquer. L'hétérogénéité d'un nœud se mesure par une fonction non négative qui doit être

i. nulle si, et seulement si, le nœud est homogène : tous les individus appartiennent à la même modalité de T.

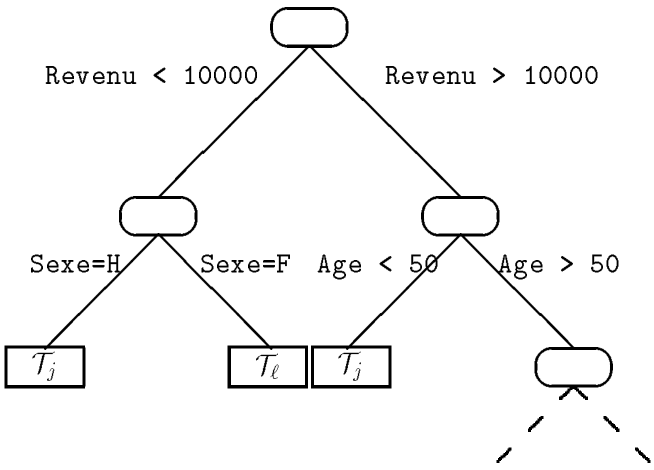


FIG. 11.1 – Exemple élémentaire d'arbre de classification.

ii. Maximale lorsque les modalités de T sont équiréparties ou équiprobables.

Un nœud k de l'arbre représente un sous-ensemble de l'échantillon d'effectif $n_{+k} = \sum_{\ell=1}^{m} n_{\ell k}$ où $n_{\ell k}$ est l'effectif de la ℓème classe dans le $k$ème nœud. On note

$$p_{\ell k} = P[\mathcal{T}_\ell \mid k] \quad \text{avec} \quad \sum_{\ell=1}^{m} p_{\ell k} = 1$$

la probabilité qu'un élément du $k$ème nœud appartienne à la ℓème classe. Comme pour l'ADD, ces probabilités conditionnelles sont définies par la règle de Bayes lorsque les probabilités *a priori* π_ℓ d'appartenance à la ℓème classe sont connues. Dans le cas contraire, les probabilités de chaque classe sont estimées sur l'échantillon et les probabilités conditionnelles s'estiment simplement par des rapports d'effectifs : $p_{\ell k}$ est estimée par $n_{\ell k}/n_{+k}$.

La fonction d'hétérogénéité ou de désordre d'un nœud est définie à partir de la notion d'entropie, du critère de concentration de Gini ou encore d'une statistique de test du χ^2. Le choix du critère importe peu, le premier est souvent préféré car il s'interprète comme un terme de déviance d'un modèle gaussien (arbre de régression) ou multinomial (arbre de classification.

Dans le cas d'une variable à expliquer qualitative, le *désordre* du $k$ème nœud défini à partir de l'entropie s'écrit[1] :

$$D_k = -2\sum_{\ell=1}^{m} n_{+k} p_{\ell k} \log(p_{\ell k}) \quad \text{et s'estime par} \quad \widehat{D}_k = -2\sum_{\ell=1}^{m} n_{\ell k} \log(\frac{n_{\ell k}}{n_{+k}}).$$

La division du nœud k crée deux fils gauche et droit. Pour simplifier, ils sont notés $(k+1)$ et $(k+2)$ mais une Re-numérotation est nécessaire pour respecter la séquence de sous-arbres qui sera décrite dans la section suivante.

Parmi toutes les divisions admissibles du nœud k, l'algorithme retient celle qui rend la somme $D_{(k+1)} + D_{(k+2)}$ des désordres des nœuds fils minimales. Ceci revient encore à résoudre à chaque étape k de construction de l'arbre :

$$\min_{\{divisions\ de X^j; j=1,p\}} D_k - (D_{(k+1)} + D_{(k+2)})$$

[1] Avec la convention $0\log(0) = 0$.

Graphiquement, la longueur de chaque branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division.

2.3 Règle d'arrêt

La croissance de l'arbre s'arrête à un nœud donné, qui devient donc terminal ou feuille, lorsqu'il est homogène c'est-à-dire lorsqu'il n'existe plus de partition admissible ou, pour éviter un découpage inutilement fin, si le nombre d'observations qu'il contient est inférieur à une valeur seuil à choisir en général entre 1 et 5.

2.4 Affectation

Chaque feuille ou nœud terminal est affecté à une classe $\mathcal{T}_\ell$ de T :
- celle la mieux représentée dans le nœud et il est ensuite facile de compter le nombre d'objets mal classés;
- la classe *a posteriori* la plus probable au sens bayesien si des probabilités *a priori* sont connues;
- la classe la moins coûteuse si des coûts de mauvais classement sont donnés.

3 Élagage

Dans des situations complexes, la règle précédente peut conduire à des arbres extrêmement raffinés et donc à des modèles de prévision très instables car fortement dépendants des échantillons qui ont permis leur estimation. On se trouve donc dans une situation de sur-ajustement à éviter au profit de modèles plus parcimonieux donc plus robuste au moment de la prévision. Cet objectif est obtenu par une procédure d'*élagage* (pruning) de l'arbre.

Le principe de la démarche, introduite par Breiman et col. (1984), consiste à construire une suite emboîtée de sous-arbres de l'arbre maximum par élagage successif puis à choisir, parmi cette suit, l'arbre optimal au sens d'un critère. La solution ainsi obtenue par un algorithme pas à pas n'est pas nécessairement celle optimale mais l'efficacité et la fiabilité sont préférées à l'optimalité.

3.1 Construction de la séquence d'arbres

Pour un arbre A donné, on note H le nombre de feuilles ou nœuds terminaux de A; la valeur de H exprime la complexité de A. La mesure de qualité de discrimination d'un arbre A s'exprime par un critère

$$D(A) = \sum_{h=1}^{H} D_h(A)$$

où $D_h(A)$ est le nombre de mal classés ou la déviance ou le coût de mauvais classement de la $h$ème feuille de l'arbre A.

La construction de la séquence d'arbres emboîtés repose sur une pénalisation (analogue au C_p de Mallow) de la complexité de l'arbre :

$$C(A) = D(A) + \gamma H.$$

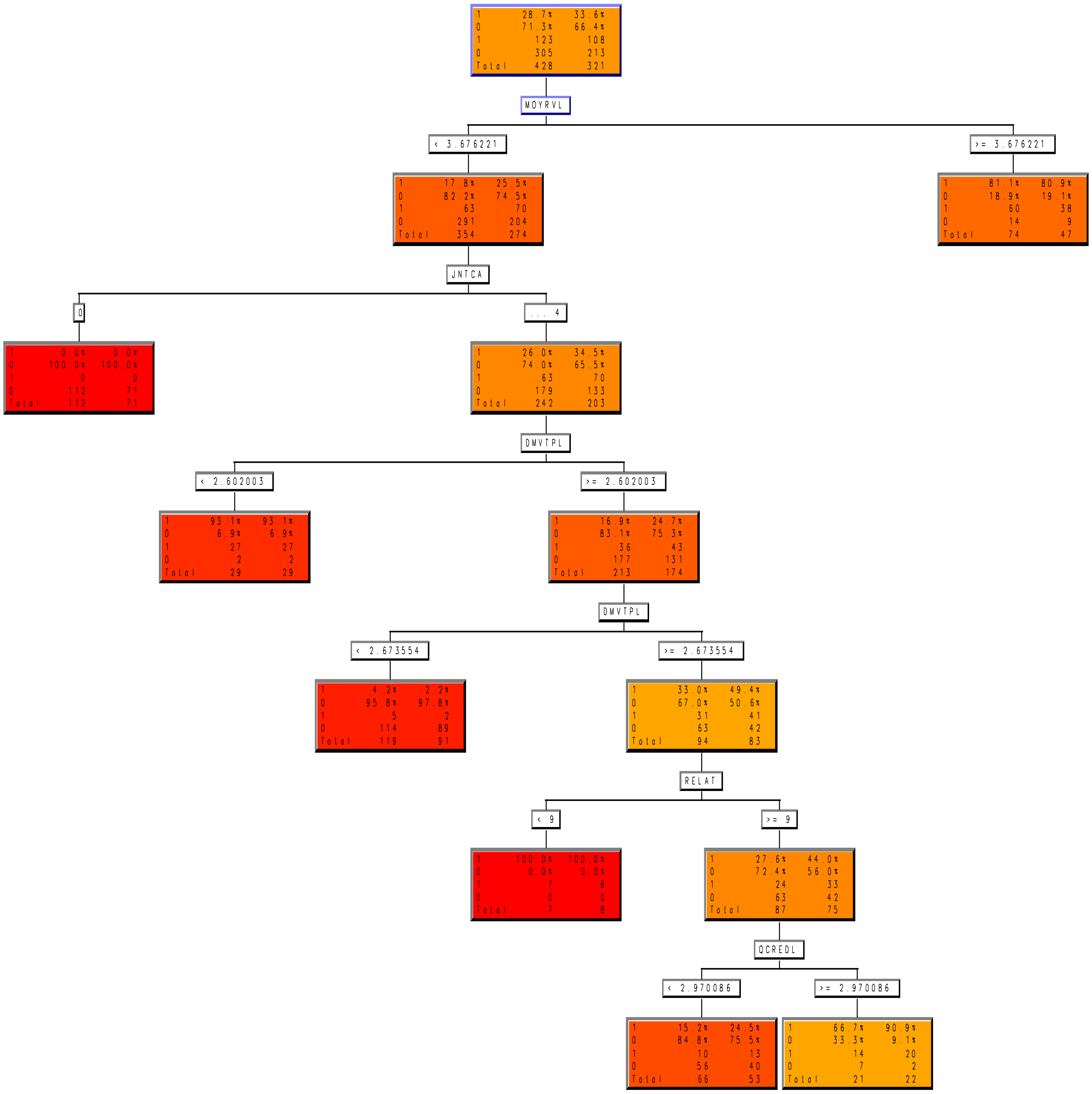


Fig. 11.2 – *Carte Visa : arbre de décision.*

Pour $\gamma = 0$ $A_{\max} = A_H$ minimise $C(A)$. En faisant croître γ, l'une des divisions de A_H, celle pour laquelle l'amélioration de D est la plus faible (inférieure à γ), apparaît comme superflue et les deux feuilles obtenues sont regroupées (élaguées) dans le nœud père qui devient terminal ; A_H devient A_{H-1}.

Le procédé est itéré pour la construction de la séquence emboîté :

$$A_{\max} = A_H \supset A_{H-1} \supset \cdots A_1$$

où A_1, le nœud racine, regroupe l'ensemble de l'échantillon.

Un graphe représente la décroissance ou éboulis de la déviance (ou du taux de mal classés) en fonction du nombre de feuilles dans l'arbre.

3.2 Recherche de l'arbre optimal

Le graphe précédemment obtenu peut se lire comme un éboulis de valeur propre, quand l'amélioration du critère est jugé trop petite ou négligeable, on élague l'arbre au nombre de feuilles obtenu. L'évaluation de la déviance ou du taux de mauvais classement estimé alors sur est biaisée (trop optimiste). Une estimation sans biais est obtenu par l'utilisation d'un autre échantillon ou encore par validation croisée.

Le principe de sélection d'un arbre optimal est donc le suivant :

i. construction de $A_{\max}$
ii. construction de la séquence $A_H \ldots A_1$
iii. estimation sans biais (échantillon de validation ou validation croisée) de $D(A_H), \ldots, D(A_1)$,
iv. représentation de $D(A_h)$ en fonction de h et choix du h rendant $D(A_h)$ minimum.

4 Exemple des données bancaires

On reprend le même ensemble de variables explicatives que pour la régression logistique et on construit l'arbre complet. L'élagage est ensuite opéré à l'aide d'un échantillon de validation (SAS) ou encore par validation croisée (Splus).

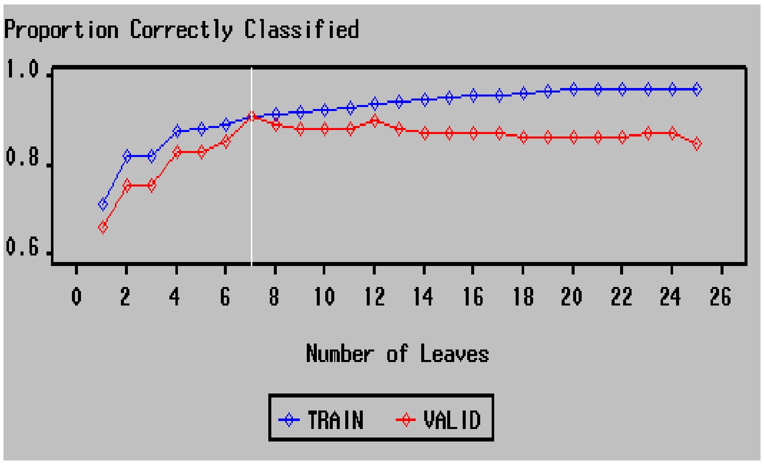


Fig. 11.3 – *Carte Visa : choix du nombre de feuilles par échantillon de validation.*

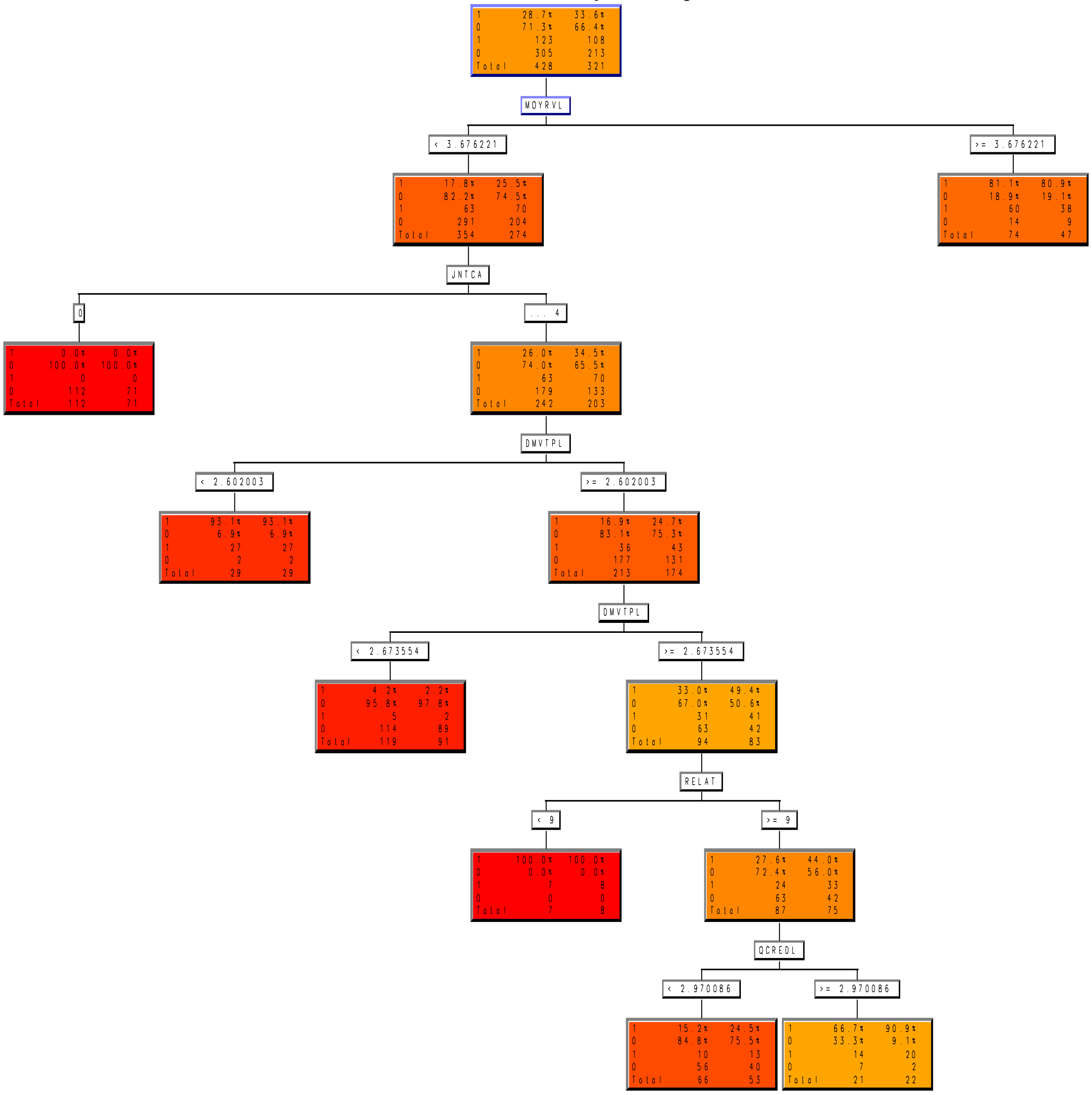


Fig. 11.4 – *Carte Visa : arbre de décision élagué.*

Chapitre 12

Méthodes connexionistes

1 Historique

Nous nous intéressons ici à une branche de l'Informatique fondamentale qui, sous l'appellation d'*Intelligence Artificielle*[1], a pour objectif de simuler des comportements du cerveau humain. Les premières tentatives de modélisation du cerveau sont anciennes et précèdent même l'ère informatique. C'est en 1943 que Mc Culloch (neurophysiologiste) et Pitts (logicien) ont proposé les premières notions de *neurone formel*. Ce concept fut ensuite mis en réseau avec une couche d'entrée et une sortie par Rosenblatt en 1959 pour simuler le fonctionnement rétinien et tacher de reconnaître des formes. C'est l'origine du *perceptron*. Cette approche dite *connexioniste* a atteint ses limites technologiques, compte tenu de la puissance de calcul de l'époque, mais aussi théoriques au début des années 70 (1970).

L'approche connexioniste à *connaissance répartie* a alors été supplantée par l'approche *symbolique* ou séquentielle qui promouvait les *systèmes experts* à *connaissance localisée*. L'objectif était alors d'automatiser le principe de l'expertise humaine en associant trois concepts :

- une *base de connaissance* dans laquelle étaient regroupées "toutes" les connaissances d'experts humains sous forme de propositions logiques élémentaires (Si ... Alors... Sinon...) ou plus élaborées en utilisant des quantificateurs (logique du premier ordre).
- une *base de faits* contenant les observations du cas à traiter comme, par exemple, des résultats d'examens, d'analyses de sang, de salive pour des applications biomédicales de choix d'un antibiotique,
- un *moteur d'inférence* chargé d'appliquer les règles expertes sur la base de faits afin d'en déduire de nouveaux faits jusqu'à la réalisation d'un objectif comme l'élaboration du traitement d'un infection bactérienne.

Face aux difficultés rencontrées lors de la modélisation des connaissances d'un expert humain, au volume considérable des bases de connaissance qui en découlait et au caractère exponentiel de la complexité des algorithmes d'inférence mis en jeu, cette approche s'est éteinte avec les années 80. En effet, pour les systèmes les plus compliqués à base de calcul des prédicats du premier ordre, on a pu montrer qu'ils conduisaient à des problèmes NP complets et donc dont la solution pouvait être atteinte mais pas nécessairement en un

[1] *La tristesse de l'Intelligence Artificielle c'est qu'elle est sans artifice donc sans intelligence.* Beaudrillard in *Cool memories*, Gallimard, 1987.

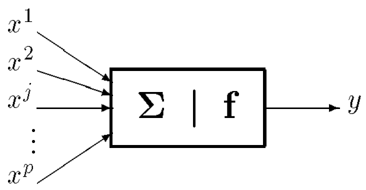


FIG. 12.1 – Représentation d'un neurone formel.

temps fini !

L'essor technologique et surtout quelques avancées théoriques :
- algorithme d'estimation par rétropropagation de l'erreur (Hopkins 1982, Le Cun 1985),
- analogie de la phase d'apprentissage avec les modèles markoviens de systèmes de particules de la mécanique statistique (verres de spin, Hopfield 1982),

au début des années 80 ont permis de relancer l'approche connexioniste. Celle-ci connaît depuis le début des années 90 un développement considérable si l'on considère le nombre de publications et de congrès qui lui sont consacrés mais aussi les domaines d'applications très divers où elle apparaît. Alors que les systèmes experts n'ont jamais atteint un niveau de développement industriel, ou du moins un niveau rentable, de nombreuses sociétés proposent actuellement des logiciels de reconnaissance de forme, de discrimination pour le marketing, de prévision de séries temporelles... à base de réseaux neuronaux. Sur de nombreux objectifs, justement ceux propres au data mining, les réseaux neuronaux ne rentrent pas nécessairement en concurrence avec des méthodes statistiques bientôt centenaires mais apportent un point de vue complémentaire qu'il est important de considérer (Thiria et col. 1997).

2 Réseaux de neurones

Un *réseau neuronal* est l'association, en un graphe plus ou moins complexe, d'objets élémentaires, les *neurones formels*. Les principaux réseaux se distinguent par l'organisation du graphe (en couches, complets...), c'est-à-dire leur architecture, son niveau de complexité (le nombre de neurones) et par le type des neurones (leurs fonctions de transition).

2.1 Neurone formel

De façon très réductrice, un neurone biologique est une cellule qui se caractérise par
- des synapses, les points de connexion avec les autres neurones, fibres nerveuses ou musculaires ;
- des dentrites, les "entrées" du neurones ;
- l'axone, la "sortie" du neurone vers d'autres neurones ou fibres musculaires ;
- le noyau qui active la sortie en fonction des stimuli en entrée.

Par analogie, le neurone formel est un modèle qui se caractérise par un état interne $s \in \mathcal{S}$, des signaux d'entrée $x_1, \dots, x_p$ et une fonction de transition d'état

$$s = h(x_1, \dots, x_p) = f\left(\sum_{j=1}^{p} \beta_j x_j\right).$$

La fonction de transition opère une transformation d'une combinaison linéaire des signaux d'entrée. Cette combinaison linéaire est déterminée par un *vecteur de poids* $[\beta_1, \dots, \beta_p]$ associé à chaque neurone et dont les valeurs sont estimées dans la phase d'apprentissage. Ils constituent "la mémoire" ou "connaissance répartie" du réseau.

Les différents types de neurones se distinguent par la nature f de leur fonction de transition. Les principaux types sont :

linéaire f est la fonction identité,

sigmoïde $f(x) = 1/(1 + e^x)$,

seuil $f(x) = \mathbf{1}_{[0,+\infty[}(x)$,

stochastiques $f(x) = 1$ avec la probabilité $1/(1 + e^{-x/H})$, 0 sinon (H intervient comme une température dans un algorithme de recuit simulé).

Les modèles linéaires et sigmoïdaux sont bien adaptés aux algorithmes d'apprentissage comme celui de rétropropagation du gradient car leur fonction de transition est différentiable. Ce sont les plus utilisés. Le modèle à seuil est sans doute plus conforme à la "réalité" biologique mais pose des problèmes d'apprentissage. Enfin le modèle stochastique est utilisé pour des problèmes d'optimisation globale de fonctions perturbées ou encore pour les analogies avec les systèmes de particules. On ne le rencontre pas en data mining.

3 Perceptron multicouche

3.1 Architecture

Le perceptron multicouche (PMC) est un réseau composé de couches successives. Une *couche* est un ensemble de neurones n'ayant pas de connexion entre eux. Une couche d'entrée lit les signaux entrant, un neurone par entrée x_j, une couche en sortie fournit la réponse du système. Une ou plusieurs couches cachées participent au transfert. Un neurone d'une couche cachée est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche suivante.

Un perceptron multicouche réalise donc une transformation

$$y = F(x_1, \dots, x_p; \boldsymbol{\beta})$$

où $\boldsymbol{\beta}$ est le vecteur contenant chacun des paramètres $\beta_{jk\ell}$ de la $j$ème entrée du $k$ème neurone de la ℓème couche ; la couche d'entrée ($\ell = 0$) n'est pas paramétrée, elle ne fait que distribuer les entrées sur tous les neurones de la couche suivante.

Par souci de cohérence, nous avons tâché de conserver les mêmes notations à travers les différents chapitres. Ainsi, les *entrées* d'un réseau sont encore notées $x^1, \dots, x^p$ comme les variables explicatives d'un modèle tandis que les *poids* des entrées sont des paramètres $\boldsymbol{\beta}$ à estimer lors de la procédure d'*apprentissage* et que la *sortie* est la variable à expliquer ou cible du modèle.

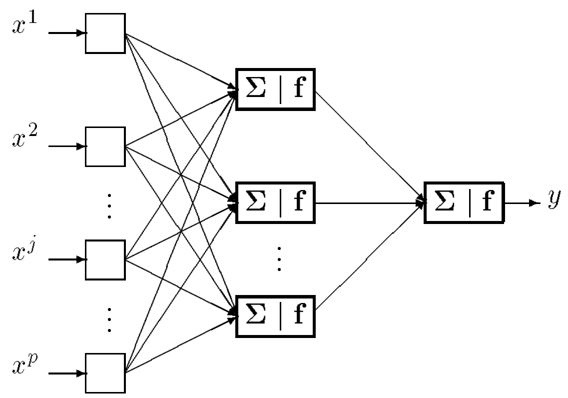


FIG. 12.2 – Exemple de perceptron multicouche élémentaire avec une couche d'entrée, une couche cachée et une couche de sortie.

3.2 Apprentissage

Supposons que l'on dispose d'une base d'apprentissage de taille n d'observations $(x_i^1, \ldots, x_i^p; y_i)$ des variables explicatives $X^1, \ldots, X^p$ et de la variable à prévoir Y. L'apprentissage est l'estimation $\widehat{\boldsymbol{\beta}}$ des paramètres du modèle solutions du problème des moindres carrés :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} Q(\mathbf{b}) \quad \text{avec} \quad Q(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}[y_i - F(x_i^1, \ldots, x_i^p; (b))]^2.$$

L'algorithme d'optimisation le plus utilisé est celui de rétropropagation du gradient basé sur l'idée suivante : en tout point $\mathbf{b}$, le vecteur gradient de Q pointe dans la direction de l'erreur croissante. Pour faire décroître Q il suffit donc de se déplacer en sens contraire. Il s'agit d'un algorithme itératif modifiant les poids de chaque neurone selon :

$$b_{jk\ell}(i) = b_{jk\ell}(i-1) + \Delta b_{jk\ell}(i)$$

où la correction $\Delta b_{jk\ell}(i)$ est proportionnelle au gradient et à l'erreur attribuée à l'entrée concernée $\varepsilon_{jk\ell}(i)$ et incorpore un terme d'"inertie" $\alpha b_{jk\ell}(i-1)$ permettant d'amortir les oscillations du système :

$$\Delta b_{jk\ell}(i) = -\tau\varepsilon_{jk\ell}(i)\frac{\partial Q}{\partial b_{jk\ell}} + \alpha b_{jk\ell}(i-1).$$

Le coefficient de proportionnalité τ est appelé le *taux d'apprentissage*. Il peut être fixe à déterminer par l'utilisateur ou encore varier en cours d'exécution selon certaines règles paramétrées par l'utilisateur. Il paraît en effet intuitivement raisonnable que, grand au début pour aller plus vite, ce taux décroisse pour aboutir à un réglage plus fin au fur et à mesure que le système s'approche d'une solution. La formule de *rétropropagation de l'erreur* (voir par exemple Bourret et al. pour des détails) fournit, à partir des erreurs observées sur les sorties, l'expression de l'erreur attribuée à chaque entrée de la couche de sortie à la couche d'entrée.

La littérature sur le sujet propose quantités de recettes destinées à améliorer la vitesse de convergence de l'algorithme ou bien lui éviter de rester collé à une solution locale défavorable. Des propriétés (dynamique markovienne ergodique et convergence vers la mesure stationnaire) de cet algorithme impliquent une convergence presque sûre ; la probabilité d'atteindre une précision fixée *a priori* tend vers 1 lorsque la taille de l'échantillon d'apprentissage tend vers l'infini.

Algorithme de rétropropagation

Initialisation

i. Les poids $b_{jk\ell}$ sont initialisés par tirage aléatoire selon une loi uniforme sur $[0,1]$.

ii. Normaliser dans $[0,1]$ les données d'apprentissage.

Tant que $Q >$ ***errmax ou niter<itermax*** La base d'apprentissage est rangée dans un nouvel ordre aléatoire.

Répéter pour chaque élément i ***de la base***

i. Calculer $\varepsilon(i) = y_i - F(x_i^1, \ldots, x_i^p; (b)(i-1))$ en propageant les entrées vers l'avant.

ii. L'erreur est "rétropropagée" dans les différentes couches afin d'affecter à chaque entrée une responsabilité dans l'erreur globale.

iii. Mise à jour de chaque poids $b_{jk\ell}(i) = b_{jk\ell}(i-i) + \Delta b_{jk\ell}(i)$

Frep

FTQ

3.3 Utilisation

On pourra se reporter à l'abondante littérature sur le sujet pour obtenir des précisions sur l'algorithme d'apprentissage, ses nombreuses variantes et ses concurrents. Il est important de rappeler la liste des choix qui sont laissés à l'utilisateur. En effet, même si les logiciels proposent des valeurs par défaut, il est fréquent que cet algorithme connaisse quelques soucis de convergence.

L'utilisateur doit donc déterminer

i. les variables d'entrée et la variable de sortie ; leur faire subir comme pour toutes méthodes statistiques, d'éventuelles transformations.

ii. L'architecture du réseau : le nombre de couches cachées (en général une ou deux) qui correspond à une aptitude à traiter des problèmes de non-linéarité, le nombre de neurones par couche cachée. Ces deux choix conditionnent directement le nombre de paramètres (de poids) à estimer. Ils participent à la recherche d'un bon compromis biais/variance c'est-à-dire à l'équilibre entre qualité d'apprentissage et qualité de prévision. À la louche, on considère en pratique qu'il faut un échantillon d'apprentissage au moins dix fois plus grand que le nombre de paramètres à estimer.

iii. Deux autres paramètres interviennent également sur ce compromis : le nombre maximum d'itérations et l'erreur maximum tolérée. En renforçant ces critères on améliore la qualité de l'apprentissage ce qui peut se faire au détriment de celle de la prévision.

iv. Le taux d'apprentissage ainsi qu'une éventuelle stratégie d'évolution de celui-ci.

Les champs d'application des PMC sont très nombreux : discrimination, prévision d'une série temporelle, reconnaissance de forme... Ils sont en général bien explicités dans les documentations des logiciels spécialisés.

Les critiques principales énoncées à l'encontre du PMC concernent les difficultés liés à l'apprentissage (temps de calcul, taille de l'échantillon, expérience de l'utilisateur) ainsi que son statut de boîte noir. En effet, contrairement à un modèle de discrimination ou un arbre de régression il est *a priori* impossible de connaître l'influence effective d'une entrée (une variable) sur le système dès qu'une couche cachée intervient. Néanmoins, des techniques de recherche de sensibilité du système à chacune des entrées permettent de préciser les idées et, éventuellement de simplifier le système en supprimant certaines des entrées.

Enfin il est clair que toute cette méthodologie doit être mise en œuvre avec beaucoup de prudence. L'expérience de l'utilisateur doit suppléer au manque de recul de ces innovations très récentes. Il ne peut se retrancher derrière tout le savoir faire des outils statistiques maintenant bien rôdés. Les procédures de validation par échantillons tests sont à utiliser sans retenue.

Chapitre 13

Comparaison des modèles

En dehors d'une situation expérimentale planifiée, c'est-à-dire sans le secours de modèles probabilistes, l'usage intensif du calcul (computational statistics), permet, essentiellement par des simulations, d'estimer des risques dans le cas d'une variable à expliquer Y quantitative ou des taux (ou des coûts) d'erreur de classement dans le cas qualitatif (T). Ces estimations, dont le principe est développé par Efron (1982), ont pour objectif de comparer, choisir, certifier des modèles estimés par une même méthode ou encore des méthodes entre elles. Ce chapitre décrit les outils élémentaires d'estimation des erreurs et propose un protocole de comparaison et test de modèles dans le cadre du data mining.

1 Estimation des erreurs

Trois types d'estimation de l'erreur (ou des coûts) sont considérés.

1.1 Resubstitution

Cette méthode consiste à appliquer la règle de décision ou le modèle estimé sur l'échantillon d'apprentissage pour calculer l'erreur quadratique de prévision ou le taux de mal classées. Il s'appelle le *taux apparent d'erreur* et est fourni en standard par les logiciels. Il n'est pas coûteux en temps de calcul mais sous-estime souvent ce taux d'erreur. Il est en effet biaisé car est estimé sur les données qui ont également servi à estimer les paramètres du modèle ou la règle de décision. Il est d'autant plus faible que le modèle est complexe et que la taille de l'échantillon est faible. Il est peu recommandé car conduit nécessairement à une sur paramétrisation.

1.2 Échantillon de validation

Si l'échantillon disponible est de taille importante, ce biais peut être évité simplement en éclatant l'échantillon en une partie *apprentissage* sur laquelle les paramètres sont estimés et une partie *validation* pour l'estimation de l'erreur. La taille de l'échantillon influe sur la variance de l'estimation de l'erreur et cette variance ne peut être estimée. D'autre part, la réduction de la taille de l'échantillon pénalise la précision de l'estimation des paramètres. Ce sont les raisons qui conduisent à préférer les approches ci-dessous. Elles consomment beaucoup plus de ressources informatiques mais les moyens actuels le permettent.

2 Validation croisée

Le principe décrit ci-dessus peut-être systématisé même (et surtout) pour des échantillons de taille plus modeste. Historiquement, la première idée développée consistait à réduire successivement l'échantillon de validation à chacune des observations. La *validation croisée* consiste donc à exécuter l'algorithme suivant où n désigne la taille de l'échantillon : ,

Répéter pour chaque observation i de l'échantillon

i. estimer le modèle sur l'échantillon privé de la $i$ème observation,

ii. prévoir la $i$ème observation avec le modèle ainsi estimé,

iii. calculer l'erreur $\varepsilon_{(i)}$ de prévision (resp. de classement) de la $i$ème observation.

Fin répéter

Calculer PRESS (resp. le taux de bon classement) qui est la moyenne des carrés des erreurs de prédiction $\varepsilon_{(i)}$.

L'objectif est évidemment de minimiser cette quantité PRESS. L'effort de calcul n'est pas toujours aussi important car pour certains modèles il existe des formules de mise à jour de l'estimation par suppression d'une observation. Le biais de l'estimation de l'erreur ainsi obtenue par PRESS est négligeable mais il peut être affecté d'une variance importante. Des extensions ont donc été suggérées afin de réduire cette variance et donc d'améliorer la précision.

- En partitionnant l'échantillon :
  i. partager l'échantillon en N parties de même taille,
  ii. exécuter le principe de la validation croisée en supprimant, prévoyant successivement chacune des N parties conduisant au calcul des erreurs ε_j.
- En itérant les étapes suivantes N fois :
  i. tirage aléatoire d'un échantillon d'apprentissage,
  ii. estimation du modèle sur cet échantillon,
  iii. prévision du reste de l'échantillon et calcul de l'erreur.

Calculer, dans un cas comme dans l'autre, la moyenne des N erreurs tandis que l'étude de leur distribution fournit une estimation de la variance de l'erreur ainsi estimée.

3 Bootstrap

Cette idée, d'approcher par simulation (*Monte Carlo*) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent lorsque l'on ne peut pas supposer qu'elle est gaussienne, est l'objectif même du *bootstrap* (Efron, 1982).

Le principe fondamental de cette technique de rééchantillonnage est de substituer, à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit *échantillon bootstrap* selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales.

Il est ainsi facile de construire un grand nombre d'échantillons bootstrap sur lesquels calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables[1] de la loi de l'estimateur. Cette approximation fournit ainsi des estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance de l'estimateur sans hypothèse (normalité) sur la vraie loi.

4 Comparaison

Les outils précédents sont mis à profit dans le cadre d'une démarche de prospection pour l'optimisation des modèles, leur comparaison ou leur certification. Il s'agit de bien distinguer les deux étapes suivantes :

i. optimisation d'un modèle attaché à une méthode donnée (régression logistique, analyse discriminante, arbre, réseau de neurones),
ii. comparaison des méthodes et certification du meilleur modèle.

Ces deux étapes doivent être traitées indépendamment et nécessitent donc la partition de l'échantillon en deux parties selon la démarche suivante :

i. procéder à un tirage aléatoire d'un *échantillon test* qui ne sera utilisé que lors de la dernière étape.
ii. Sur la partie restante qui sera itérativement[2] découpé en échantillon d'apprentissage et échantillon de validation, optimiser par validation croisée les choix afférents à chacune des méthodes de modélisation (resp. discrimination) :
   - choix des variables, interactions et méthode de modélisation (resp. discrimination),
   - nombre de nœuds dans l'arbre de régression (resp. classification),
   - architecture (nombre de couches, de neurones par couche, fonctions de transferts, nombre de cycle. . .) du perceptron.

   Attention, si cette étape est trop raffinée (trop d'essais, trop de paramètres à régler) relativement à la taille de l'échantillon, elle peut conduire à un sur apprentissage qui nuira à la qualité de prévision.
iii. Comparaison finale, entre chacun des meilleurs modèles de chaque méthode, des qualités de prévision calculées sur la base de l'erreur de prévision (resp. du taux de mal classés) estimées sur le seul échantillon test qui est resté à l'écart des efforts d'optimisations des modèles.

   > *Attention, ne pas tricher en modifiant le modèle obtenu lors de l'étape précédente afin d'améliorer le résultat sur l'échantillon de test !*

[1] Échantillon indépendant de même loi et estimateur indépendant de l'ordre des observations.

[2] Éventuellement, à condition que l'échantillon soit de grande taille, le nombre d'itérations est réduit à un. C'est le seul choix offert dans SAS Enterprise miner.

Conclusion

Stratégies du data mining

Les chapitres précédents décrivent les outils de base du prospecteur de données tandis que les logiciels commerciaux en proposent une intégration plus ou moins complète, plus ou moins conviviale de mise en œuvre. En pratique, l'enchaînement de ces techniques permet la mise en place de *stratégies de prospection* bien définies. Celles-ci dépendent essentiellement des *types* de variables considérés et des *objectifs* poursuivis.

Types de variables

Explicatives L'ensemble des p variables explicatives est noté X, il est constitué de variables
- $X_{\mathbb{R}}$ toutes quantitatives[3],
- X_E toutes qualitatives,
- $X_{\mathbb{R}\cup E}$ un mélange de qualitatives et quantitatives.

À expliquer La variable à expliquer ou *cible* (target) peut être
- Y quantitative,
- Z qualitative à 2 modalités,
- T qualitative.

Objectifs

Trois objectifs principaux sont poursuivis dans les applications classiques de dData mining :

i. **Exploration multidimensionnelle** ou réduction de dimension : production de graphes, d'un sous-ensemble de variables représentatives X_r, d'un ensemble de composantes C_q préalables à une autre technique.

ii. **Classification** (clustering) ou segmentation : production d'une variable qualitative T_r.

iii. **Modélisation (Y ou Z)/Discrimination (Z ou T)** production d'un modèle de prévision de Y (resp. Z, T).

D'autres méthodes plus spécifiques à certaines problématiques peuvent apparaître (analyse sensorielle, analyse conjointe, SARIMA... mais leur usage reste limité à des problématiques ou des contextes bien particuliers.

[3]Une variables explicative qualitative à 2 modalités (0,1) peut être considérée comme quantitative ; c'est l'indicatrice des modalités.

Outils

Les méthodes utilisables se classent en fonction de leur objectif et des types de variables explicatives ou à expliquer.

Exploration

`ACP` $X_{\mathbb{R}}$ et $\emptyset$

`AFCM` X_E et $\emptyset$

`AFD` $X_{\mathbb{R}}$ et T

Classification

`CAH` $X_{\mathbb{R}}$ et $\emptyset$

`NuéeDyn` $X_{\mathbb{R}}$ et $\emptyset$

`RNKoho` $X_{\mathbb{R}}$ et $\emptyset$

Modélisation

i. Modèle linéaire généralisé
   - `RLM` $X_{\mathbb{R}}$ et Y
   - `ANOVA` X_E et Y
   - `ACOVA` $X_{\mathbb{R}\cup E}$ et Y
   - `Rlogi` $X_{\mathbb{R}\cup E}$ et Z
   - `Lglin` X_T et T
ii. Analyse discriminante
   - `ADpar/nopar` $X_{\mathbb{R}}$ et T
iii. Classification and regression Tree
   - `ArbReg` $X_{\mathbb{R}\cup E}$ et Y
   - `ArbCla` $X_{\mathbb{R}\cup E}$ et T
iv. Réseaux neuronaux
   - `percep` $X_{\mathbb{R}\cup E}$ et Y ou T

Stratégies

Les stratégies classiques de la prospection de données consistent à enchaîner les étapes suivantes :

i. **Extraction** de l'entrepôt des données éventuellement par sondage pour renforcer l'effort sur la qualité des données plutôt que sur la quantité.
ii. **Exploration**
   - Tri à plat, étape élémentaire mais essentielle de vérification des données, de leur cohérence. Étude des distributions, transformation, recodage éventuel des variables quantitatives, regroupement de modalités des variables qualitatives, élimination de certaines variables (trop de données manquantes, quasi constantes, redondantes...). Gérer rigoureusement les codes des variables et de leurs modalités.
   - Étude bivariée Recherche d'éventuelles relations non linéaires. Si les variables sont trop nombreuses, sélectionner les plus liées à la variable cible. Complétion des données manquantes.
iii. **Analyse**

   Classification :
   Pas de variable à expliquer
   - En cas de variables $X_{\mathbb{R}\cup E}$ ou X_T, la classification est exécutée sur les C_q issues d'une AFCM des variables codées en classes.
   - Caractérisation des classes par les variables initiales à l'aide des outils de discrimination.

   Modélisation/Discrimination :
   Une variable à expliquer Y, Z ou T
   - Extraction d'un échantillon *test*,
   - Estimation, optimisation (validation croisée) des modèles pour chacune des méthodes utilisables.
   - Comparaison des performances des modèles optimaux de chaque méthode sur l'échantillon test.

iv. *Exploitation* du modèle et diffusion des résultats. Finalement, une fois que la bonne méthode associé au bon modèle ont été choisie, tout l'échantillon est regroupé pour faire une dernière estimation du modèle qui sera utilisé en exploitation.

Pièges

Les principaux pièges qui peuvent être rencontrés au cours d'une prospection sont souvent le résultat d'un *acharnement* en quête de sens (*data snooping*). Cela signifie qu'à force de creuser, contrairement à un prospecteur minier à la recherche de diamants bien réels, le prospecteur en données disposant d'un grand nombre de variables finit bien, en mode exploratoire, par trouver des relations semblant hautement significatives. Un exemple simple permet d'illustrer cette réserve.

> *Réalisons un tirage (pseudo) aléatoire de 15 variables quantitatives indépendantes et de même loi. Une étude systématique des corrélations linéaires entre ces variables conduit au calcul de* $15 \times 7 = 105$ *coefficients. La théorie nous dit que le test de nullité d'un coefficient de corrélation (même remarque pour le test d'indépendance du* χ^2*) appliqué très naïvement 105 fois avec un risque de première espèce de 5% sera rejeté en moyenne 5 fois. Sans précaution évidente, le statisticien débutant sera amené à considérer comme significatives les liaisons entre 5 couples de variables. Ce piège est facile à éviter en rectifiant le seuil de confiance (Bonferroni), d'autres sont nettement moins évidents et dans beaucoup de situations mal contrôlées, les méthodes statistiques doivent être employées avec prudence, plus comme des outils exploratoires que comme des outils décisionnels.*

En phase de modélisation, une sur paramétrisation ou un sur ajustement du modèle expliquera parfaitement les données. C'est particulièrement sensible pour les réseaux de neurones qui sont définis par un nombre important de paramètres et sont donc capables d'apprendre “par cœur” l'échantillon d'apprentissage. Dans tous les cas, les résultats ne seront pas extrapolables à d'autres données que celles étudiées. Les résultats de prédiction seront donc entachés d'une forte erreur relative liée à la variance des estimations des paramètres. C'est toujours le problème de trouver un compromis entre le biais d'un modèle qui est toujours plus ou moins faux et la variance des estimateurs. Nous insistons donc sur les indispensables phases de choix et comparaison des modèles.

Exploration *versus* décision

Nous touchons l'un des sujets sensibles, point d'achoppement entre représentants de différentes disciplines. Le statisticien traditionnel, fidèle aux procédés de la planification expérimental, s'attachera à organiser l'observation de façon à se mettre dans les conditions optimales le conduisant à des modèles fiables, à des tests d'hypothèses et donc à des décisions sous le contrôle des modèles probabilistes inférés. À l'autre extrémité, la tentation est grande, compte tenu des possibilités des moyens de stockage et de calcul et des coûts très lourds engendrés par une démarche expérimentale, de vouloir prétendre à une exhaustivité de l'exploration, de vouloir tester toutes les relations ou modèles possibles.

> *Par exemple, l'impact ou l'effet d'une campagne de marketing directe pourrait être préalablement testé, comme c'est l'usage et même l'obligation légale pour*

un médicament. Considérant deux échantillons représentatifs (cf. théorie des sondages) l'un reçoit une information ciblée, l'autre celle usuelle des clients d'une entreprise. Il est alors facile de tester avec rigueur les comportements de ces deux échantillons et de juger, à un niveau de risque fixé, s'ils sont significativement différents.

Entre les deux attitudes, tous les intermédiaires sont envisageables mais il faut bien comprendre que la pertinence des méthodes statistiques, la fiabilité d'une décision issue d'un test d'hypothèse sont soumises à leurs conditions strictes d'application : échantillons identiquement distribués, homogénéité des variances, de celle des résidus suivant des lois généralement gaussiennes, observation sans erreur des variables explicatives, dans les cas les plus contraignants.

Complexité

Une bonne pratique du Data Mining nécessite de savoir articuler toutes les méthodes entrevues dans ce document, listées schématiquement en couverture et reprises ci-dessus. Rude tâche qui ne peut être entreprise qu'à la condition d'avoir très bien spécifié les objectifs de l'étude. On peut noter que certaines méthodes (i.e. ADD, régression logistique, perceptron, arbres) poursuivent les mêmes objectifs. Dans les bons cas, données bien structurées, elles fourniront des résultats très similaires, dans d'autres une méthode peut se révéler plus efficace compte tenu de la taille de l'échantillon ou géométriquement mieux adaptée à la topologie des groupes à discriminer. D'autre part, les méthode ne présentent pas toutes les mêmes facilités d'interprétation. Il n'y a pas de choix *a priori*, seul l'expérience et un protocole de *test* soignée permettent de se déterminer. C'est la raison pour laquelle des logiciels généralistes comme SAS (module Enterprise Miner) ne font pas de choix et offrent ces méthodes en parallèle pour mieux s'adapter aux données, aux habitudes de chaque utilisateur/client potentiel et à la mode.

Le travail demandé déborde rapidement du rôle d'un statisticien car la masse et la complexité des données nécessitent le développement d'interfaces et d'outils graphiques sophistiqués ainsi que de nouveaux algorithmes adaptatifs assurant une mise à jour et une diffusion en temps réel (ouèbe) des résultats. Néanmoins l'objectif principal reste une "quête de sens" en vue de faciliter les prises de décision tout en en préservant la fiabilité. Ainsi, une expertise statistique reste importante car la méconnaissance des limites et pièges des méthodes employées peut en effet conduire à des aberrations discréditant la démarche et rendant caducs les investissements consentis.

Quelles compétences ?

En ajoutant les problèmes de définition, gestion des bases de données, les problèmes de réseau..., le champ des compétences requises pour prétendre à une gestion efficace de l'information est d'une étendue redoutable. Néanmoins l'utilisateur peut maintenant disposer d'outils très conviviaux dont il est facile et rapide d'obtenir des résultats. Un petit quart d'heure suffit pour se familiariser avec une interface graphique qui exécute des méthodes dont une compréhension fine nécessite plusieurs heures de cours ou réflexion à Bac+5.

Cette situation pour le moins paradoxale invite à s'interroger sur l'intérêt des logiciels spécifiques et très coûteux de data mining tel SAS Enterprise Miner face à une utilisation plus conventionnelle d'un logiciel de statistique : SPSS, SAS ou encore Splus. Certes, ils sont simples à utiliser et on peut arguer qu'il n'est pas nécessaire d'être un mécanicien accompli pour conduire une voiture. Néanmoins, la conduite d'une modélisation, d'une segmentation, d'une discrimination imposent à son auteur des choix plus ou moins implicites qui sont loin d'être neutres et qui dépassent largement en complexité celui du choix d'un carburant par le conducteur à la pompe.

Les investissements matériels et humains consentis par une entreprise pour mettre en place des entrepôts de données, même s'ils sont sous-traités, sont considérables. Il serait périlleux de chercher à les rentabiliser, par exemple, par une optimisation de stratégies commerciales basée sur l'utilisation naïve de boîtes noires imposant des méthodes et options par défaut et sans le contrôle ou l'expertise d'un, disons le mot, *statisticien*. Statisticien qui, justement, sait mettre en œuvre à moindre coût (i.e. Splus) les méthodes statistiques en question sans faire appel à un logiciel spécifique de prospection des données. En résumé, n'est-il pas plus efficace de payer un statisticien plutôt qu'un logiciel ? Ceci rappelle le débat apparu au cours des années 80 (1980) qui ont vu l'expansion puis le déclin et la disparition des logiciels dits *Systèmes Experts* chargés de simulé le travail d'un expert humain.

De toute évidence, le champ des compétences concernées ne peut être couvert que par une équipe pluridisciplinaire ayant dépassé ses querelles de chapelles. Mais, comme l'association entre *Information* et *Pouvoir* se fait de plus en plus présente, on peut prévoir, avec une bonne certitude, que les blocages et retards de mise place des systèmes d'information publics ou privés sont ou seront d'ordre hiérarchique, institutionnel, politique, financier plus que technique.

Bibliography

Agresti, A. (1990). *Categorical data analysis.* Wiley.

Antoniadis, A., Berruyer, J., and Carmona, R. (1992). *Régression non linéaire et applications.* Economica.

Ardilly, P. (1994). *Les techniques de sondage.* Technip.

Berry, M. and Gordon, L. (1997). *Data Mining, techniques appliquées au marketing, à la vente et aux services clients.* Masson.

Besse, P., Cardot, H., and Ferraty, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis*, 24 :255–270.

Besse, P. and Ferraty, F. (1995). A fixed effect curvilinear model. *Computational Statistics*, 10 :339–351.

Bourret, P., Reggia, J., and Samuelides, M. (1991). *Réseaux neuronaux.* Teknea.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees.* Wadsworth & Brooks.

Celeux, G. (1990). *Analyse discriminante sur variables continues.* INRIA.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., and Ralambondrainy, H. (1989). *Classification automatique des données.* Dunod.

Celeux, G. and Nakache, J.-P. (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., Fichet, B., and Tassi, P. (1992). *Modèles pour l'Analyse des Données Multidimensionnelles.* Economica.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Methods.* SIAM.

Everitt, B. and Dunn, G. (1991). *Applied Multivariate Data Analysis.* Edward Arnold.

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.

Lefèbure, R. and Venturi, G. (1998). *Le data Mining.* Eyrolles.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models.* Chapman & Hall.

Monfort, A. (1982). *Cours de Statistique Mathématique.* Economica.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis.* Springer-Verlag.

S-plus (1993). *Guide to statistical and mathematical analysis.* MathSoft.

SAS (1989). *SAS/STAT User's Guide*, fourth edition. version 6.

SAS (1995). *SAS/INSIGHT User's Guide*, third edition. version 6.

Thiria, S., Lechevallier, Y., Gascuel, O., and Canu, S. (1997). *Statistique et méthodes neuronales.* Dunod.

Tomassonne, R., Audrain, S., Lesquoy-de Turckheim, E., and Millier, C. (1992). *La régression, nouveaux regards sur une ancienne méthode statistique.* Masson.

Troisième partie

Annexes

Annexe 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 $(p \times n)$.

$$(\mathbf{A}')' = \mathbf{A} \; ; (\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}' \; ; (\mathbf{A}\mathbf{B})' = \mathbf{B}'\mathbf{A}' \; ; \begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix}' = \begin{bmatrix} \mathbf{A}_1^{1'} & \mathbf{A}_2^{1'} \\ \mathbf{A}_1^{2'} & \mathbf{A}_2^{2'} \end{bmatrix}.$$

Produit scalaire élémentaire : $a'b = \sum_{i=1}^n a_i b_i$ où a et b sont des vecteurs-colonnes.

Produit : $[\mathbf{A}\mathbf{B}]_i^j = a_i' b^j$ avec $\mathbf{A}_{(n \times p)}$, $\mathbf{B}_{(p \times q)}$ et $\mathbf{A}\mathbf{B}_{(n \times q)}$, 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 :

$$\begin{aligned}
\mathrm{tr}\alpha &= \alpha, \\
\mathrm{tr}\alpha\mathbf{A} &= \alpha\mathrm{tr}\mathbf{A}, \\
\mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}\mathbf{A}+\mathrm{tr}\mathbf{B}, \\
\mathrm{tr}\mathbf{A}B &= \mathrm{tr}\mathbf{B}A, \\
&\quad \text{reste vrai si } \mathbf{A} \text{ est } (n\times p) \text{ et si } \mathbf{B} \text{ est } (p\times n) \\
\mathrm{tr}\mathbf{C}C' &= \mathrm{tr}\mathbf{C}'\mathbf{C} = \sum_{i=1}^{n}\sum_{j=1}^{p}(c_i^j)^2 \\
&\quad \text{dans ce cas, } \mathbf{C} \text{ est } (n\times p).
\end{aligned}$$

Déterminant

On note $|\mathbf{A}|$ le *déterminant* de la matrice carrée $\mathbf{A}$ $(p \times p)$. Il vérifie :

$$\begin{aligned}
|\mathbf{A}| &= \prod_{j=1}^{p} a_j^j, \text{ si } \mathbf{A} \text{ est triangulaire ou diagonale,} \\
|\alpha\mathbf{A}| &= \alpha^p|\mathbf{A}|, \\
|\mathbf{A}\mathbf{B}| &= |\mathbf{A}||\mathbf{B}|, \\
\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{vmatrix} &= |\mathbf{A}||\mathbf{C}|, \\
\begin{vmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{vmatrix} &= |\mathbf{A}_1^1||\mathbf{A}_2^2 - \mathbf{A}_2^1(\mathbf{A}_1^1)^{-1}\mathbf{A}_1^2| && (1) \\
&= |\mathbf{A}_2^2||\mathbf{A}_1^1 - \mathbf{A}_1^2(\mathbf{A}_2^2)^{-1}\mathbf{A}_2^1|, && (2) \\
&\quad \text{sous réserve de la régularité de } \mathbf{A}_1^1 \text{ et } \mathbf{A}_2^2.
\end{aligned}$$

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 :

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}, \quad (\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

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 $\{\mathbf{x}_1, \dots, \mathbf{x}_q\}$ 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 $\mathcal{E}_q = \{\mathbf{e}_1, \dots, \mathbf{e}_q\}$ qui engendre dans E un s.e.v. $E_q = \text{vec}\{\mathbf{e}_1, \dots, \mathbf{e}_q\}$ 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 $F = \mathbb{R}^n$.

$$\begin{aligned}
\text{Im}(\mathbf{A}) &= \text{vect}\{\mathbf{a}^1, \dots, \mathbf{a}^p\} \text{ est le s.e.v. de } F \textit{ image } \text{de } \mathbf{A}\,; \\
\text{Ker}(\mathbf{A}) &= \{x \in E\ ;\ \mathbf{A}x = 0\} \text{ est le s.e.v. de } E \textit{ noyau } \text{de } \mathbf{A}\,; \\
E &= \text{Im}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A}) \text{ si } \mathbf{A} \text{ est carrée associée à un endomorphisme de } E \\
\text{et } p &= \dim(\text{Im}(\mathbf{A})) + \dim(\text{Ker}(\mathbf{A})).
\end{aligned}$$

$$\begin{array}{rcl} \text{rang}(\mathbf{A}) & = & \dim(\text{Im}(\mathbf{A})), \\ 0 \leq & \text{rang}(\mathbf{A}) & \leq \min(n,p), \\ \text{rang}(\mathbf{A}) & = & \text{rang}(\mathbf{A}'), \\ \text{rang}(\mathbf{A}+\mathbf{B}) & \leq & \text{rang}(\mathbf{A}) + \text{rang}(\mathbf{B}), \\ \text{rang}(\mathbf{AB}) & \leq & \min(\text{rang}(\mathbf{A}), \text{rang}(\mathbf{B})), \\ \text{rang}(\mathbf{BAC}) & = & \text{rang}(\mathbf{A}),\ \text{si } \mathbf{B} \text{ et } \mathbf{C} \text{ sont régulières}, \\ \text{rang}(\mathbf{A}) & = & \text{rang}(\mathbf{AA}') = \text{rang}(\mathbf{A}'\mathbf{A}). \end{array}$$

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* : $\mathrm{d}_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}}$,
- 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) :

$$\begin{array}{lll} \mathbf{u} \in W, & \mathbf{Pu} = \mathbf{u}, & \lambda = 1, \text{ de multiplicité } \dim(W), \\ \mathbf{v} \perp W, \text{(on note } \mathbf{v} \in W^{\perp}) & \mathbf{Pv} = 0, & \lambda = 0, \text{ de multiplicité } \dim(W^{\perp}). \end{array}$$

- $\mathrm{tr}\mathbf{P} = \dim(W)$.

- $\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1}\mathbf{B}'\mathbf{M}$, 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 :

$$\mathbf{P} = \mathbf{B}\mathbf{B}'\mathbf{M} = \sum_{i=1}^{q} \mathbf{b}^j \mathbf{b}^{j\prime}\mathbf{M}.$$

- Dans le cas particulier où $q = 1$ alors :

$$\mathbf{P} = \frac{\mathbf{b}\mathbf{b}'}{\mathbf{b}'\mathbf{M}\mathbf{b}}\mathbf{M} = \frac{1}{\|\mathbf{b}\|_{\mathbf{M}}}\mathbf{b}\mathbf{b}'\mathbf{M}.$$

- 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 :

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

- 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 $\mathbf{v} = \mathbf{A}\mathbf{u}$ 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 :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k\prime}$$

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 :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{M} = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M}$$

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 $\lambda_k \neq \lambda_j$,	$\mathbf{v}^k \perp_{\mathbf{M}} \mathbf{v}^j$;
$\mathrm{tr}\mathbf{A} = \sum_{k=1}^{p} \lambda_k$;	$\lvert\mathbf{A}\rvert = \prod_{k=1}^{p} \lambda_k$;
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 :*

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' = \sum_{k=1}^{r} \sqrt{\lambda_k} \mathbf{u}^k \mathbf{v}^{k'}; \tag{3}$$

$\mathbf{U}$ $(n \times r)$ *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* :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M},\mathbf{D}} = \mathrm{tr}\mathbf{X}MY'\mathbf{D}. \tag{4}$$

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 :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{I_p},\mathbf{I_n}} = \mathrm{tr}\mathbf{X}Y' = \sum_{i=1}^{n}\sum_{j=1}^{p} x_i^j y_i^j = \mathrm{vec}(\mathbf{X})'\mathrm{vec}(\mathbf{Y}).$$

La *norme* associée à ce produit scalaire (4) est appelée norme trace :

$$\begin{aligned} \|\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} &= \mathrm{tr}\mathbf{X}MX'\mathbf{D}, \\ \|\mathbf{X}\|^2_{\mathbf{I_p},\mathbf{I_n}} &= \mathrm{tr}\mathbf{X}X' = \mathrm{SSQ}(\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_i^j)^2 \\ & \quad (\text{SSQ signifie "sum of squares"}). \end{aligned}$$

La *distance* associée à cette norme devient, dans le cas où $\mathbf{D}$ est une matrice diagonale ($\mathbf{D} =$ diag$(w_1, \ldots, w_n)$), le critère usuel des *moindres carrés* :

$$\mathrm{d}^2(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{M},\mathbf{D}} = \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{y}_i\|^2_{\mathbf{M}}.$$

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 :*

$$\min_{\mathbf{Z}} \left\{ \|\mathbf{X} - \mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}} ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ rang(\mathbf{Z}) = q < r \right\} \tag{5}$$

est donnée par la somme des q premiers termes de la décomposition en valeurs singulières (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 $E_q = \operatorname{Im}(\mathbf{V}_q)$ (resp. $\mathbf{D}$-orthogonale sur $F_q = \operatorname{Im}(\mathbf{U}_q)$) :

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \sum_{k=1}^{q} \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M} \\
\widehat{\mathbf{Q}_q} &= \sum_{k=1}^{q} \mathbf{u}^k \mathbf{u}^{k'} \mathbf{D} = \mathbf{U}_q \mathbf{U}'_q \mathbf{D}, \\
\mathbf{Z}_q &= \widehat{\mathbf{Q}_q} \mathbf{X} = \mathbf{X} \widehat{\mathbf{P}_q}'.
\end{aligned}$$

Proposition A.6. — *Avec les notations précédentes :*

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \quad \mathit{arg} \max_{\mathbf{P}_q} \left\{ \left\| \mathbf{X}\mathbf{P}'_q \right\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\qquad \left. \mathbf{P}_q \textit{ projection } \mathbf{M}\textit{-orthogonale de rang } q < r \right\}, \\
\widehat{\mathbf{Q}_q} &= \quad \mathit{arg} \max_{\mathbf{Q}_q} \left\{ \left\| \mathbf{Q}_q \mathbf{X} \right\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\qquad \left. \mathbf{Q}_q \textit{ projection } \mathbf{D}\textit{-orthogonale de rang } q < r \right\}.
\end{aligned}$$

Annexe B

Enterprise miner

1 Introduction

SAS propose de nombreux modules plus spécifiques à certains environnement : statistique élémentaire guidée (SAS/LAB), contrôle de qualité, planification expérimentale, prévision.... Parmi ceux-ci et depuis la version 8.e, SAS Enterprise Miner (SEM) est la réponse de SAS Institute à la demande spécifique autour du *data mining* et de ses applications à la *gestion de la relation client*. Bien sur, cet outil a d'autres champs d'application que le marketing comme par exemple la détection d'une défaillance dans un procédé industriel.

SEM est avant tout une *interface graphique* à des procédures ou macros SAS permettant une gestion, en première lecture facile, de l'ensemble de la chaîne des traitements nécessaires à une prospection de données (Besse, 2000)[1]. En deuxième lecture, les choses se compliquent rapidement. Un utilisateur ne peut pas échapper à une bonne compréhension des méthodes proposées, de leurs conditions d'application, de leur interprétation et de leurs limites. D'autre part une personnalisation de l'environnement pour ajouter ses propres traitements nécessitent également des compétences informatiques et une bonne connaissance de la programmation SAS classique.

Ce court chapitre ne propose pas un cours détaillé ou un manuel de référence de SEM, tout ceci est disponible dans la documentation en ligne. En effet, l'interface graphique, très intuitive, permet de se retrouver rapidement dans les différents outils disponibles. Comme pour SAS/Insight, nous nous limiterons donc à un descriptif sommaire des possibilités offertes de façon à préciser l'organisation et les capacités de traitement de ce module. Enfin il est bon de préciser que ce module est d'un coût relativement élevé par rapport aux autres modules de SAS, cet investissement lourd ne pourrait se justifier que par un gain très significatif de productivité dans le traitement et l'analyse des données.

2 Organisation

Le principe de base de l'utilisation de l'interface est la réalisation d'un *graphe* schématisant l'ensemble des traitements d'une prospection. Chaque *nœud* symbolise un traitement tan-

[1] *Statistique et Data mining*, `http ://www.lsp.ups-tlse/Besse`

dis qu'une arrête orientée matérialise la circulation des données entre les différentes étapes. L'édition du graphe engendre ainsi l'écriture puis l'exécution de commandes SAS.

SEM propose sous cette forme la mise en œuvre d'une méthodologie en 5 étapes (SEMMA) qui reprend le schéma logique d'une étude statistique : échantillonnage (Sample), Exploration, transformations (Modify), Modélisation, comparaisons et certifications (Assess).

2.1 Fenêtres

Outre des barres d'outils et menus de la version 8, SEM se présente sous la forme de deux fenêtres :

Outils Avec 3 onglets :

- **Diagrams** : gestion hiérarchisée des projets et études.
- **Tools** : liste des nœuds disponibles.
- **Reports** : gestion des résultats en format `html`.

Graphe en cours de définition, exécution. La construction du graphe se fait avec la souris en sélectionnant et en faisant glisser les icônes de la barre des outils ou ceux de la fenêtre `tools`. Il suffit ensuite de les relier.

2.2 Menus

À chaque nœud est associé un menu identique proposant différents choix dont :

open ouvre une fenêtre dans lesquels sont définis les paramètres et options du nœud considéré,

run provoque l'exécution du programme associé à ce nœud,

resul la fenêtre de présentation des résultats.

2.3 Démarrage

i. Le module SEM est lancé par la commande *miner* ou dans le menu `global>solutions>enterprise miner`.

ii. Il suffit ensuite s'ouvrir un projet existant ou un nouveau projet (menu `file>new>project`). Un projet correspond à une librairie SAS c'est-à-dire encore à un répertoire de l'environnement de l'utilisateur.

iii. Nommer le projet créé choisir le type de localisation : *client* sur le poste de l'utilisateur ou sur un *serveur* distant à travers le réseau et le module SAS/Connect.

3 Descriptif sommaire des noeuds

Pour une utilisation efficace de tout logiciel, il est nécessaire d'avoir une vision complète de ses possibilités afin d'utiliser les outils les mieux adaptés à l'objectif à atteindre. Les outils de SEM ou nœuds sont répartis en plusieurs classes. Certains reprennent des fonctionnalités usuelles des modules SAS/Stat ou SAS/Graph, d'autres (notés par une '*') apportent des compléments.

3.1 Échantillonnage

Input data source Sélectionne une table SAS contenant les données et spécifie son rôle : données brutes (raw), train (apprentissage), validate (validation), test, score (paramètres d'un modèle). Spécification du rôle des variables et en particulier de celle à expliquer (cible) ; les autres sont par défaut explicatives (input).

Sampling Tirage aléatoire simple ou stratifié d'un échantillon.

Data partition Partition de la table (données brutes) en échantillons d'apprentissage, de validation et de test en spécifiant la part respective de chacun.

3.2 Exploration

Distribution explorer graphe dynamique de la distribution de la loi conjointe de 1 à 3 variables.

Multiplot distribution et distribution croisée avec la variable cible de chacune des variables explicatives.

Insight est intégré pour ses aspects exploratoires et ses tests mais sans certaines fonctionnalités (ACP, modélisation).

Association* Recherche de co-occurences, ou de séquences d'évènements. Exemple : achats simultanés d'articles.

Variable selection Recherche, dans un grand ensemble, des variables explicatives les plus liées (coef. de corrélation, χ^2 ou les plus aptes à la prédiction de la cible. Des discrétisations sont proposées pour les variables qualitatives si elles améliorent les qualités prédictives.

3.3 Transformations

Data set attributes permet de redéfinir les rôles des variables (input, target, rejected, id, freq, trial, predict, residual, group, assess), leur type (binary, interval, nominal, ordinal, unary), leur libellé.

Transform variables par des fonctions classiques ($\log, \exp, \sqrt{}\ldots$) définies par l'utilisateur ou automatiquement pour optimiser la normalité de la distribution ou encore la corrélation avec la cible.

Filter outliers élimine les modalités rares, les valeurs extrêmes par rapport à la moyenne ou la médiane ou à l'extérieur d'un intervalle.

Replacement Plutôt que de supprimer des lignes ou des colonnes, il est possible de remplacer les valeurs manquantes par différentes méthodes : valeur moyenne ou médiane, tirage aléatoire, valeur prédite par un arbre de classification ou de régression.

Clustering Création d'une typologie (nuées dynamiques ou CAH) et des graphiques liant la typologie avec les variables explicatives.

SOM/Kohonen crée une typologie par la construction d'une "carte auto-organisatrice" de Kohonen.

3.4 Modélisation

Tree Estimation d'un arbre de classification ou de régression avec différents critères optionnels : test de Fisher et réduction de la variance pour une cible quantitative, entropie, concentration de Gini, test du χ^2 pour une cible qualitative. La procédure d'élagage cherche à minimiser l'erreur sur l'échantillon de validation.

Regression Régression logistique ou régression linéaire avec choix de modèle par par différents algorithmes (forward, backward, stepwise) et selon différents critères (AIC, BIC, validation error, validation, cross-validation error, profit/loss).

Neural network Perceptron multi couches avec différents options pour l'apprentissage : Levenberg Marquardt, Quasi Newton, gradient conjugué, rétropropagation dont certains sont issus du module SAS/OR. Le choix dépend de la capacité mémoire, de la taille du réseau et du volume de l'échantillon d'apprentissage. Le premier est plus rapide mais très pénalisé par le nombre de paramètres à estimer et la mémoire requise. Le dernier est jugé moins fiable et sensible à la présence de minimas locaux. Les options permettent de fixer le nombre de couches cachées, le nombre de neurones par couches et les fonctions de transfert de chaque neurone. Le nombre d'itérations et donc la durée de l'apprentissage est contrôlée par l'erreur calculée sur l'échantillon de validation.

User defined model reçoit une table de type score ou encore le résultat d'un programme SAS inclut dans un nœud `SAS Code` et permet d'intégrer dans SEM un modèle définit par l'utilisateur à l'aide des procédures classiques de SAS. Par exemple : une analyse discriminante décisionnelle, une régression sur composantes principales. . . .

Ensemble permet de calculer de nouvelles prévisions par combinaisons de plusieurs modèles ou par moyennage de prédictions issues de rééchantillonnages (bagging ou bootstrap aggregating). Enfin une autre procédure de rééchantillonnage (arcing ou adaptative resampling & combining) renforce dans son tirage les probabilités proportionnellement aux résidus. Ceci à pour effet d'améliorer l''ajustement des valeurs extrêmes.

3.5 Comparaison, certification

Assessment compare les capacités prédictives des modèles sur l'échantillon test.

Score applique un modèle estimée sur une première table à une autre table possédant les mêmes variables et permettant ainsi d'établir les prévisions sur d'autres données.

Reporter placé en fin de graphe, il produit un rapport en format `html` regroupant les tableaux et graphiques résultants des différentes étapes.