

E. Diday J. Lemaire
J. Pouget F. Testu

éléments d'analyse de données

Dunod

Table des matières

	Pages
AVANT-PROPOS	1
INTRODUCTION	3
CHAPITRE 1 - LES ÉTAPES PRÉLIMINAIRES À UNE ANALYSE DE DONNÉES	
1 - INTRODUCTION	13
2 - MODÉLISATION DE LA NOTION D'INDIVIDUS ET DE VARIABLES	13
2.1 - L'ensemble des individus	13
2.2 - Définition d'une variable	13
2.3 - Définition d'une variable vectorielle	13
3 - TABLEAUX DE DONNÉES	15
3.1 - Définition et notations	15
3.2 - Exemples de tableaux de données	15
3.2.1 - Tableau de données quantitatives	15
3.2.2 - Tableau de contingence	16
3.2.3 - Tableau de préférence	17
3.2.4 - Tableau binaire	18
3.2.5 - Tableau de modalités	18
3.2.6 - Tableau de proximité	19
3.2.7 - Tableau hétérogène	20
3.2.8 - Tableau à 3 dimensions	20
4 - ÉLÉMENTS DESCRIPTIFS D'UN TABLEAU DE DONNÉES	21
4.1 - Notations	21
4.2 - Éléments descriptifs dans l'espace des variables	22
4.3 - Éléments descriptifs des individus	24
4.4 - Représentations visuelles élémentaires	25
4.4.1 - Diagrammes, histogrammes, cumulées	25
4.4.2 - La méthode d'Andrews	28
4.4.3 - Polygones	28
4.4.4 - Méthode des visages de Chernof	29

4.5 - Représentation par plan et par hiérarchie	29
4.5.1 - La représentation plane	29
4.5.2 - La représentation hiérarchique	30
5 - LE CHANGEMENT DE VARIABLE	30
5.1 - But du changement de variable	30
5.2 - Formalisation de la notion de changement de variable	31
5.2.1 - Changement de variable par changement de structure	31
5.2.2 - Changement de variable par codage	32
5.2.3 - Codage optimal	32
5.3 - Codage d'une variable quantitative	33
5.3.1 - Par centrage et réduction	33
5.3.2 - Codage d'une variable quantitative par découpage	34
5.3.2.1 - Découpage par effectifs égaux : cumulée inverse	35
5.3.2.2 - Découpage à l'aide d'un histogramme ou par des intervalles égaux	36
5.3.2.3 - Découpage par minimisation de la variance	38
5.3.2.4 - Découpage par codage flou	39
5.4 - Codage d'une variable qualitative	40
5.4.1 - Codage à modalités avec ou sans ordre	40
5.4.2 - Codage disjonctif complet	42
5.5 - Codage d'une variable qualitative ordinale	43
5.5.1 - Codage additif	43
5.5.2 - Rang moyen	44
5.6 - Codage d'une variable vectorielle	44
5.6.1 - Codage par calcul de profil	44
5.6.2 - Codage par classification	45
5.6.3 - Codage optimal	45
5.6.3.1 - En terme de classification automatique optimale	45
5.6.3.2 - En terme de combinaison linéaire optimale	45
6 - CHOIX D'UNE MESURE DE RESSEMBLANCE	46
6.1 - Quelques définitions	46
6.2 - Transformation d'un indice de dissimilarité en une distance respectant le même ordre sur les couples	47
6.3 - Cas de tableaux où toutes les variables sont quantitatives	48

6.4 - Cas de tableaux binaires	50
6.5 - Cas de tableaux où toutes les variables sont qualitatives nominales	51
6.5.1 - Mesure de ressemblance entre individus et comment gagner de la place en mémoire centrale	51
6.5.2 - Mesure de ressemblance entre variables nominales	52
6.6 - Cas de tableaux où les variables sont qualitatives ordinales	54
6.6.1 - Corrélation des rangs de Kendall entre deux variables v_1 et v_2	54
6.6.2 - Corrélation des rangs de Spearman	55
6.6.3 - Le coefficient de corrélation des rangs (dit de monotonie) de Guttman	55
6.7 - Mesures de ressemblance entre groupes	56
6.7.1 - Mesures de ressemblance entre groupes d'individus	56
6.7.1.1 - Distance du lien minimum	56
6.7.1.2 - Distance du lien maximum	57
6.7.1.3 - Distance des noyaux	57
6.7.1.4 - Utilisation de l'inertie intra-classe et inter-classe	58
6.7.2 - Mesure de ressemblance entre groupes de variables	62
6.8 - Cas de tableaux hétérogènes	63
6.8.1 - Utilisation de changements de variables par un codage qui rend toutes les variables compatibles (i.e. de même type)	64
6.8.2 - Utilisation de plusieurs distances appropriées	64
6.8.3 - Utilisation du codage optimal d'une variable vectorielle	65
6.9 - Le problème des données manquantes	66
6.10 - Mesures de ressemblance entre indices de dissimilarité	66
6.11 - Comment choisir une mesure de ressemblance ?	68
EXERCICES	70
CHAPITRE 2 - CLASSIFICATION AUTOMATIQUE	
1 - PRINCIPES GÉNÉRAUX	73
2 - LA CLASSIFICATION PAR HIÉRARCHIE ET PAR ARBRE	74
2.1 - Définition d'une hiérarchie	74

2.2 - Hiérarchie indicée	75
2.2.1 - Définition d'une hiérarchie indicée	75
2.2.2 - Définition d'un indice de dissimilarité à partir d'une hiérarchie indicée	76
2.2.3 - Indice d'agrégation entre groupes d'individus	76
2.3 - Les principaux indices d'agrégation utilisés	77
2.4 - Construction de hiérarchies indicées	80
2.4.1 - L'algorithme général de la classification ascendante hiérarchique	81
2.4.2 - Conditions à satisfaire par un indice d'agrégation δ pour qu'une hiérarchie indicée par $f : f(h_1 \cup h_2) = \delta(h_1, h_2)$ ne présente pas d'inversions	83
2.4.3 - La formule de récurrence générale	86
2.4.4 - Valeur des coefficients de récurrence pour les principaux choix de l'indice d'agrégation	88
2.4.5 - Algorithmes accélérés pour les grands tableaux	91
2.5 - Hiérarchies indicées et ultramétriques	96
2.5.1 - Définition et propriétés d'une ultramétrique	96
2.5.2 - Existence d'une bijection entre les hiérarchies et les ultramétriques	98
2.6 - Optimisation d'une classification hiérarchique	102
2.7 - La sous-dominante	103
2.7.1 - Définition	103
2.7.2 - Quelques propriétés caractéristiques de la sous-dominante	103
2.8 - Construction de la sous-dominante	107
2.8.1 - Construction de la sous-dominante par un arbre de longueur minimum	107
2.8.2 - Algorithmes de construction d'un arbre de longueur minimum	109
2.8.3 - Construction de la sous-dominante par la hiérarchie du saut minimum	113
3 - LA CLASSIFICATION PAR PARTITIONS	116
3.1 - Un algorithme de type "nuées dynamiques"	117
3.1.1 - Choix d'un mode de représentation	117
3.1.2 - Comment se déroule l'algorithme ?	118
3.1.3 - Un exemple de critère à optimiser	119
3.1.4 - L'algorithme dans le cas où les noyaux sont des groupes d'individus	119

3.1.4.1 - Etude de la convergence de l'algorithme	120
3.1.4.2 - Stationnarité de la suite $(L^{(n)}, p^{(n)})$	121
3.1.5 - La variante du centre de gravité	122
3.2 - Autres algorithmes de partitionnement et de typologie	123
3.3 - Stabilité et formes fortes	125
3.4 - Un algorithme de partitionnement optimal dans le cas d'une variable unique	129
3.5 - La segmentation	132
EXERCICES	136
CHAPITRE 3 - LES MÉTHODES LINÉAIRES	
1 - RÉGRESSION LINÉAIRE	139
1.1 - Introduction	139
1.2 - Forme des données, but de la méthode	139
1.3 - Solution du problème 1	141
1.4 - Changement de variable	142
1.5 - Qualité de la régression	143
1.6 - Introduction de variables explicatives supplémentaires. Amélioration de la qualité de la régression	147
1.7 - Elimination de variables explicatives non significatives	149
1.8 - Tests statistiques appliqués aux modèles de régression linéaire	150
1.9 - Méthodes de régression pas à pas	155
1.10 - Annexe mathématique : le théorème des projections	156
1.10.1 - Projection d'un vecteur sur U	156
1.10.2 - Détermination pratique de la projection	156
1.10.3 - Propriétés du projection P_U	157
1.11 - Exemples	157
2 - ANALYSE FACTORIELLE	167
2.1 - Introduction	167
2.2 - Analyse en composantes principales	167

2.2.1 - Problème à traiter et critère à optimiser	167
2.2.1.1 - But de la méthode	167
2.2.1.2 - Le problème d'optimisation	169
2.2.1.3 - Formulation mathématique du problème d'optimisation	170
2.2.2 - Solution du problème 1	170
2.2.2.1 - Choix de $\underline{a}$	170
2.2.2.2 - Choix de $(\underline{u}_1, \dots, \underline{u}_q)$	173
2.2.3 - Principaux résultats	180
2.2.3.1 - Choix de différentes métriques	180
2.2.3.2 - Axes principaux d'inertie	181
2.2.3.3 - Composantes principales	182
i) Ecriture matricielle des composantes principales	183
ii) Propriétés des composantes principales	183
2.2.3.4 - Valeur du critère I_{W_0}	184
2.2.3.5 - Choix du nombre d'axes	185
2.2.4 - Interprétations des résultats de l'analyse en composantes principales	187
2.2.4.1 - Description des individus	187
i) Qualité globale de la représentation	187
ii) Qualité ponctuelle de la représentation	187
2.2.4.2 - Interprétation des composantes principales	189
2.2.4.3 - Représentation des axes associés à chaque variable	189
2.2.4.4 - Description des variables	190
2.2.5 - Diagramme structuré et exemples	193
2.3 - Analyse factorielle sur un tableau de distance	207
2.3.1 - Le problème	207
2.3.2 - Décomposition d'ECKART-YOUNG	207
2.3.2.1 - Le problème et sa solution	207
2.3.2.2 - Exemple	211
2.3.3 - Représentation euclidienne d'un tableau de distances	211
2.3.3.1 - Notations utilisées	211
2.3.3.2 - Résolution du problème	213
2.3.4 - Exemples	216
2.4 - Analyse des correspondances	222
2.4.1 - Formalisation du problème	222

2.4.1.1 - Tableau des données	222
2.4.1.2 - But de la méthode	222
2.4.1.3 - Transformation des données initiales : profils	223
2.4.2 - Analyse en composantes principales des tableaux des profils	225
2.4.2.1 - Nuages $N(I)$ et $N(J)$	225
2.4.2.2 - Choix de la métrique	226
2.4.2.3 - Centres de gravité	227
2.4.2.4 - Matrices d'inertie	228
2.4.2.5 - Recherche des axes principaux d'inertie	235
2.4.2.6 - Calcul des composantes principales	237
2.4.2.7 - Calcul des contributions absolues et relatives	237
2.4.2.8 - Graphiques et interprétations	237
i) Divers types de représentation	237
ii) Représentation de points supplémentaires	238
2.4.3 - Diagramme structuré et exemples	239
2.5 - Analyse factorielle des correspondances multiples	239
2.5.1 - Données analysées	239
2.5.2 - But de l'analyse	251
2.5.3 - Critère de représentation	251
2.5.4 - Résolution du problème 2	252
2.5.5 - Conduite des calculs lorsque n est grand devant $s = m(1) + \dots + m(p)$	257
2.5.6 - Exemple	260
2.5.7 - Lien avec l'analyse factorielle des correspondances du tableau de BURT	261
2.5.8 - Exemple d'application	263
2.6 - Analyse canonique	272
2.6.1 - But de l'analyse canonique - Exemples	272
2.6.2 - Le modèle - Forme des liaisons	274
2.6.2.1 - Introduction	274
2.6.2.2 - Définition de l'analyse canonique de deux sous-espaces de $\mathbb{R}^n$	275
2.6.2.3 - Recherche des variables canoniques et propriétés	276
2.6.3 - Méthode pratique de détermination d'une analyse canonique	281
2.6.4 - Représentation et interprétation	283
2.6.4.1 - Représentation des variables	283
2.6.4.2 - Représentation des individus	284
2.6.5 - Diagramme structuré et exemples	285
EXERCICES	294

CHAPITRE 4 - ANALYSE DISCRIMINANTE

1 - INTRODUCTION	307
2 - FORME DES DONNÉES, NOTATION, ÉLÉMENTS DESCRIPTIFS	309
2.1 - Notations	309
2.2 - Centre de gravité	311
2.3 - Matrices de covariances	311
2.3.1 - Matrice de covariance totale	311
2.3.2 - Matrice de covariance intra-classe	312
2.3.3 - Matrice de covariance inter-classe	312
3 - BUTS DES MÉTHODES	313
3.1 - Introduction	313
3.2 - Exemple de la transmission de données en présence de bruit	314
4 - CRITÈRE DE DÉCISION, RÈGLE DE DÉCISION OU D'AFFECTATION ASSOCIÉE À UN CRITÈRE DE DÉCISION	316
4.1 - Introduction	316
4.2 - Le modèle mathématique	316
4.3 - Méthode Bayésienne	317
4.3.1 - Critère du risque minimum ou critère de Bayes	317
4.3.2 - Critère de la probabilité d'erreur (critère II)	320
4.3.3 - Critère du maximum de vraisemblance a posteriori (critère III)	320
4.3.4 - Comparaison des critères II et III	320
4.4 - Applications pratiques de la méthode Bayésienne	321
4.4.1 - Fonctions de potentiel, fonctions discriminantes associées à une fonction de potentiel	322
4.4.2 - Méthode des noyaux ou méthode de Parzen	323
4.4.3 - Méthode du régressogramme	325
4.4.4 - Cas particulier des distributions gaussiennes	327
4.5 - Méthodes utilisant les notions de distance ou de similarité	329
Méthodes géométriques	
4.5.1 - Introduction	329
4.5.2 - Distance assurant la cohésion maximum d'un groupe d'individus	330
4.5.3 - Choix d'une similitude ou d'une distance	332
4.6 - Méthodes basées sur la recherche de fonctions discriminantes de formes données	335

Table des matières

4.6.1 - Cas des fonctions discriminantes affines	335
5 - PROBLÈMES DE RÉDUCTIBILITÉ, ÉLIMINATION DE VARIABLES	336
5.1 - Méthodes de pas à pas	338
5.1.1 - Introduction	338
5.1.2 - Méthode de Sebestyen	339
5.1.3 - Echantillon test - interprétation des résultats obtenus	341
5.2 - Analyse factorielle discriminante	343
5.2.1 - Introduction	343
5.2.2 - Variances et covariances de nouvelles variables Combinaisons linéaires des anciennes variables	343 344
5.2.3 - Choix des variables	347
5.2.4 - Relation avec l'analyse en composantes principales du nuage des centres de gravité	347
5.2.5 - Approche géométrique	348
5.2.6 - Règles d'affectation après réduction des variables	349
5.3 - Approche de la discrimination par une méthode de régression linéaire	349
5.3.1 - Introduction	350
5.3.2 - Mise en oeuvre de la méthode	352
5.3.3 - Application au cas de deux classes	354
6 - TESTS STATISTIQUES UTILISÉS EN ANALYSE DISCRIMINANTE	355
6.1 - Le modèle probabiliste utilisé	355
6.2 - Résultats théoriques de base	355
6.2.1 - Loi de Wishart	357
6.2.2 - T^2 de Hotelling	358
6.2.3 - Critère de Wilks	358
6.3 - Les tests en analyse discriminante	358
6.3.1 - Test d'égalité des matrices de covariance Σ_k ou test de Box	359 360
6.3.2 - Tests de l'égalité des moyennes des groupes	360
6.3.3 - Test de l'appartenance des vecteurs "moyennes" μ_k à une variété de dimension donnée	360 362
6.3.4 - Test pour l'analyse discriminante pas à pas	363
6.3.5 - Comparaison entre les groupes, contrastes linéaires	366
7 - GÉNÉRALITÉS SUR LES MÉTHODES D'ANALYSE DISCRIMINANTE SUR VARIABLES QUALITATIVES	366
7.1 - Introduction	366
7.2 - Les méthodes de réduction des variables	367

8 - EXEMPLES

8.1 - Comparaison des zébus et des charolais 371

8.2 - Les nématodes 373

EXERCICES

379

CHAPITRE 5 - METHODES ORDINALES

1 - INTRODUCTION

383

2 - EXEMPLES DE DONNEES ORDINALES

385

2.1 - Données de préférence

385

2.1.1 - Comparaisons par paires

385

2.1.2 - Classement des objets à comparer

386

2.1.3 - Comparaison par blocs

387

3 - RELATIONS SUR UN ENSEMBLE FINI

388

3.1 - Définitions et notations

388

3.2 Exemples de relations

390

3.2.1 - Relations d'équivalence

390

3.2.2 - Relations d'ordre et de préordre. Applications

391

3.2.3 - Relations de quasi-ordre

395

3.3 - Extension de la notion de relation de préférence

396

3.3.1 - Signature d'une relation de préférence

396

3.3.2 - Relations floues de préférence

397

4 - CODAGE D'UNE RELATION (FLOUE) DE PRÉFÉRENCE

399

4.1 - Fonction score

399

4.1.1 - Cas d'une relation floue de préférence

399

4.1.2 - Cas d'une relation de préférence

401

4.1.3 - Cas d'une relation d'ordre

402

4.1.4 - Cas d'une relation de préordre

402

4.2 - Codage optimal

403

4.2.1 - Principe

403

4.2.2 - Codage d'une relation de pré-ordre

404

4.2.3 - Algorithme de J.B. Kruskal

405

5 - REPRÉSENTATION EUCLIDIENNE DE DONNÉES DE PRÉFÉRENCE

409

5.1 - Formulation du problème

409

Table des matières

XV

5.2 - Le modèle MDPREF	410
5.3 - Analyse factorielle ordinale	417
6 - AGRÉGATION DE DONNÉES DE PRÉFÉRENCE	425
6.1 - Méthodes de type ELECTRE	425
6.2 - Méthodes métriques	429
6.2.1 - Relations d'ordre médiane	429
6.2.2 - Relations de préordre moyennes	436
6.3 - Analyse typologique de données de préférence	439
EXERCICES	442
ANNEXE : La programmation structurée	447
BIBLIOGRAPHIE GÉNÉRALE	455

Les progrès de l'informatique et le développement des ordinateurs conduisent de plus en plus à l'accumulation d'informations de différentes sortes. La forme la plus répandue et la plus classique de stockage est celle correspondant à des tableaux de données. Dans de nombreux domaines on est ensuite conduit à tirer parti de cette information pour la synthétiser, pour servir de base à un processus de décision, de reconnaissance ou, plus généralement, pour appréhender d'une certaine manière la nature des phénomènes sous-jacents aux données. L'analyse de données répond à un bon nombre de ces questions.

Cet ouvrage développe l'ensemble des principales méthodes de cette discipline : classification automatique, analyse factorielle, analyse discriminante, méthodes ordinales. Les principaux résultats théoriques utiles y sont démontrés et des exemples d'applications illustrent chaque méthode. En outre, des exercices complètent l'ouvrage, qui s'adresse aux étudiants, universitaires, chercheurs et ingénieurs ainsi qu'aux utilisateurs, relevant de toutes les disciplines, des méthodes d'analyse de données.



ISBN 2-04-015430-2

F134/85-05

