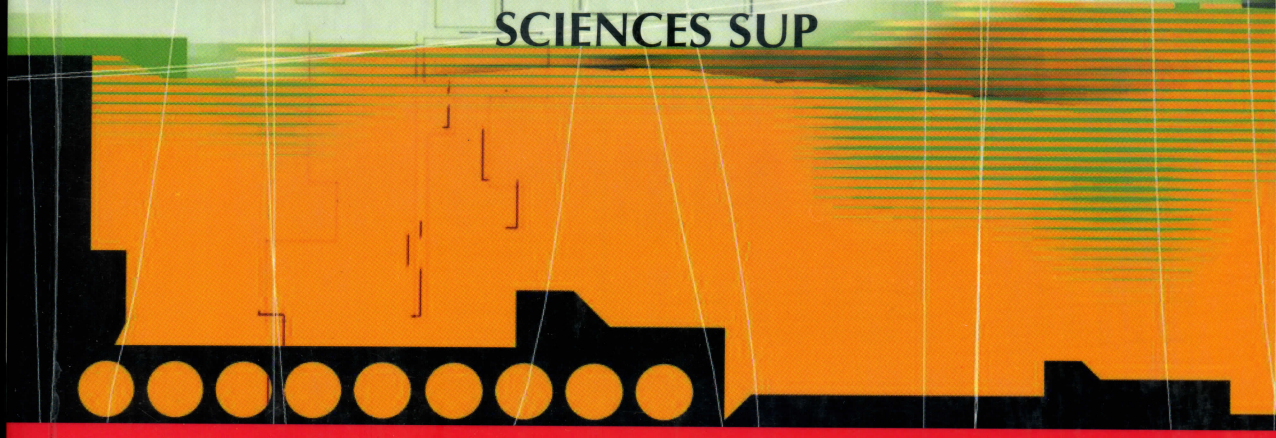
SCIENCES SUP



Cours et exercices corrigés

Masters • Écoles d'ingénieurs

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisation et inférence en fouilles de données

4e édition

Ludovic Lebart
Marie Piron
Alain Morineau

DUNOD

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisations et inférences en fouille de données

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

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

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

4e édition

DUNOD

Sommaire

Introduction 1

Chapitre 1
Analyses en axes principaux : principes de base 11

1.1 Le tableau de données 12
1.1.1 Représentation géométrique de base 12
1.1.2 Principaux types de tableaux et de méthodes 12

1.2 Analyse générale, décomposition aux valeurs singulières 16
1.2.1 Notions élémentaires et principe d'ajustement 16
1.2.2 Ajustement du nuage des individus dans l'espace des variables 18
a _ Droites d'ajustement 18
b _ Caractéristiques du sous-espace d'ajustement 20
1.2.3 Ajustement du nuage des variables dans l'espace des individus 20
1.2.4 Relation entre les ajustements dans les deux espaces 21
1.2.5 Reconstitution des données de départ 23
a _ Reconstitution exacte 23
b _ Reconstitution approchée 24
c _ Qualité numérique de l'approximation 24

1.3 Diversification de l'analyse générale 25
1.3.1 Analyse générale avec des métriques et des critères quelconques 25
1.3.2 Principe des éléments supplémentaires 27
1.3.3 Autres approches 28

1.4 Méthodes de validation empiriques : calculs de stabilité et de sensibilité 29
1.4.1 Aspects théoriques 29
1.4.2 Techniques de *bootstrap* 31

1.5 Annexe technique du chapitre 1 33
1.5.1 Démonstration sur les extrema de formes quadratiques sous contraintes quadratiques 33
1.5.2 Variations des valeurs et vecteurs propres 36

Chapitre 2
Analyse canonique et régression linéaire 37

2.1 Analyse canonique 38
2.1.1 Formulation du problème et notations 38
2.1.2 Les variables canoniques 40
a _ Calcul des variables canoniques 40
b _ Interprétation géométrique 41

c _ Cas de matrices non inversibles 42

2.2 Régression multiple, modèle linéaire 43

2.2.1 Formulation du problème : le modèle linéaire 44

2.2.2 Ajustement par la méthode des moindres-carrés 46

a _ Calcul et propriétés de l'ajustement des moindres-carrés 47

b _ Approche géométrique dans $\mathcal{R}^n$ 47

c _ Le coefficient de corrélation multiple 48

2.2.3 Lien avec l'analyse canonique 49

2.2.4 Qualité de l'ajustement 50

a _ Spécification du modèle 50

b _ Moyenne et variance des coefficients 51

c _ Tests sous l'hypothèse de normalité des résidus 52

2.2.5 Régression sur variables nominales : l'analyse de la variance 53

2.2.6 Régression sur variables mixtes : analyse de la covariance 56

2.2.7 Choix des variables, généralisations du modèle 59

a _ Sélection et choix des variables explicatives 59

b_ Modèles linéaires généralisés 60

Chapitre 3

Analyse en composantes principales 61

3.1 Histoire, domaine, principes 61

3.1.1 Domaine d'application 62

3.1.2 Interprétations géométriques 63

3.2 Individus et variables 64

3.2.1 Analyse du nuage des individus 64

a _ Principe d'ajustement 64

b _ Distance entre individus 66

c _ Matrice à diagonaliser 66

d _ Axes factoriels 67

3.2.2 Analyse du nuage des points-variables 67

a _ distances entre points-variables 68

b _ Distance à l'origine 69

c _ Axes factoriels ou composantes principales 70

3.3 Compléments et variantes 71

3.3.1 Individus et variables supplémentaires 72

3.3.2 Représentation simultanée 75

a _ Représentation séparée des deux nuages 75

b _ Justification d'une autre représentation simultanée 75

3.3.3 Analyse en composantes principales non normée 77

3.3.4 Analyses non-paramétriques 79

a _ Analyse des rangs 80

b _ Analyse en composantes robustes 81

3.3.5 **L'analyse factorielle en facteurs communs et spécifiques** 81
a _ Le modèle 82
b_ Estimation des paramètres inconnus 84
3.3.6 **L'analyse en composantes indépendantes** 86
3.3.7 **Régression sur composantes principales et régression régularisée** 88
a _ Principe de la régression régularisée 89
b _ Variables supplémentaires et régression 90
c _ Expression des coefficients dans la nouvelle base 91
3.3.8 **Aperçu sur les autres méthodes dérivées** 91
3.4 **Interprétation et validation** 92
3.4.1 **Eléments pour l'interprétation** 93
3.4.2 **Choix du nombre d'axes : règles empiriques, validation externe** 96
3.4.3 **Critères statistiques pour les valeurs propres** 99
3.4.4 **Bootstrap pour l'analyse en composantes principales** 101
a _ Premiers travaux 101
b _ Diverses possibilités de bootstrap 102
3.5 **Deux exemples d'application** 106
3.5.1 **Exemple d'application 1** 106
3.5.2 **Exemple d'application 2** 116
3.6 **Annexe technique du chapitre 3**
3.6.1 **Travaux sur la loi des valeurs propres en analyse en composantes principales** 129

Chapitre 4
Analyse des correspondances 131
4.1 **Démarche et principe : introduction élémentaire** 132
4.1.1 **Tableau de contingence : hypothèse d'indépendance** 132
a _ Notations 132
b _ Transformations du tableau de contingence 133
c _ Hypothèse d'indépendance 134
4.1.2 **Représentation géométrique** 136
a _ Construction des nuages 136
b _ Critère d'ajustement 137
c _ Choix des distances 137
4.1.3 **Propriétés** 138
a _ Equivalence distributionnelle 138
b _ Relations de transition ou quasi-barycentriques 140
c _ Justification de la représentation simultanée 142
4.2 **Schéma général de l'analyse des correspondances** 143
4.2.1 **Eléments de base de l'analyse** 143
a _ Tableau de données, distance, géométrie des nuages 143
b _ Démonstration de l'équivalence distributionnelle 145

c _ Critère à maximiser et matrice à diagonaliser 146
d _ Axes factoriels et facteurs 148
4.2.2. Représentation simultanée 148
a _ Relation entre les deux espaces 148
b _ Relations de transition (ou quasi-barycentriques) 149
c _ Représentation simultanée des lignes et colonnes 150
d _ Formule de reconstitution des données 150
4.2.3 Autre présentation de l'analyse des correspondances 151
4.3. Eléments pour l'interprétation des résultats 153
4.3.1 Inertie et formes de nuages 153
a _ Inertie et test d'indépendance 154
b _ Quelques formes caractéristiques de nuages de points 156
4.3.2 Contributions absolues et relatives 158
4.3.3 Eléments supplémentaires 162
4.4 Méthodes et critères de validation 163
4.4.1 Signification des valeurs propres et taux d'inertie 163
a _ Approximation de la distribution des valeurs propres 164
b – Indépendance des taux d'inertie et de la trace 165
c – Exemples d'abaques et tables statistiques 166
d – Autres critères de choix statistiques, résultats asymptotiques 167
e – Régions de confiance analytiques 169
4.4.2 Bootstrap pour l'analyse des correspondances 170
a _ Le principe des réplications 170
b _ Les zones de confiance 171
4.5 Exemple d'application 173
4.5.1 Données et premiers résultats 173
4.5.2 Visualisation et interprétation 175
4.5.3 Validation par bootstrap 177
4.6 Annexe technique du chapitre 4 180
4.6.1 Mise en œuvre pratique des calculs 180
4.6.2 Précisions sur l'approximation de la distribution des valeurs propres 183
4.6.3 Indépendance des taux d'inertie et de la trace 185

Chapitre 5
Analyse des correspondances multiples 186
5.1 Notations et définitions 188
5.1.1 Tableau disjonctif complet 188
a _ Hypercube de contingence 189
b _ Le codage disjonctif 189
5.1.2 Tableau de contingence de Burt 190

5.2 Principes de base de l'analyse des correspondances multiples 192
5.2.1 Schéma général 193
a _ Critère d'ajustement et distance du χ^2 193
b _ Axes factoriels et facteurs 194
c _ Facteurs et relations quasi-barycentriques 195
d _ Sous-nuage des modalités d'une même variable 196
e _ Support du nuage des modalités 197
f _ Meilleure représentation simultanée 197
5.2.2 Autres propriétés 198

5.3 Analyse du tableau de contingence de Burt 202
5.3.1 Equivalence avec l'analyse du tableau disjonctif complet 202
5.3.2 Equivalences dans le cas de deux questions 203
5.3.3 Autres équivalences 207
5.3.4 Liens avec l'analyse canonique 210
a _ Le cas de l'analyse des correspondances simples 211
b _ L'analyse des correspondances multiples 212

5.4 Méthodes de validation 214
5.4.1 Validation externe : éléments supplémentaires 214
a _ Valeurs-test pour les modalités supplémentaires 214
b _ Variables continues supplémentaires 217
5.4.2 Validation interne : inertie et méthode de bootstrap 217
a _ Taux d'inertie et information 217
b _ Bootstrap pour l'analyse des correspondances multiples 218

5.5 Interprétation et validation à propos d'un exemple 220
5.5.1 Description des données 220
5.5.2 Eléments d'interprétation 220
5.5.3 Eléments de validation 228
a _ Bootstrap partiel pour les variables actives 228
b _ Bootstrap partiel pour les variables supplémentaires 228
c _ Bootstrap total pour les variables actives 229

5.6 Modèles log-linéaires et analyse des correspondances multiples 231
5.6.1 Formulation du problème et principes de base 232
5.6.2 Ajustement d'un modèle log-linéaire 232
a _ Tableau de contingence à deux entrées 233
b _ Tableau de contingence à p entrées 233
c _ modèles hiérarchiques 235
5.6.3 Estimation et tests d'ajustement du modèle 235
a _ Estimation des paramètres 235
b _ Tests d'ajustement 236
c _ Choix du modèle 237
5.6.4 Lien avec l'analyse des correspondances 238
a _ Des champs d'application différents 239

b _ Liens théoriques entre l'analyse des correspondances et les modèles log-linéaires 241
c _ Difficultés de l'articulation exploration-inférence 243

5.7 Annexe technique du chapitre 5 245

Chapitre 6
Méthodes de classification 247

6.1 Méthodes de partitionnement 250
6.1.1 Agrégation autour des centres mobiles 250
a _ Bases théoriques de l'algorithme 250
b _ Justification élémentaire de l'algorithme 252
c _ Techniques connexes 253
d _ Formes fortes et groupements stables 254
6.1.2 Cartes auto-organisées 256
a _ Principe 256
b _ L'algorithme de Kohonen 258
c _ Application au jeu de données sémiométriques 259

6.2 Classification hiérarchique 261
6.2.1 Principe 262
a _ Distances entre éléments et entre groupes 262
b _ Algorithme de classification 263
c _ Eléments de vocabulaire 264
6.2.2 Classification ascendante selon le saut minimal et arbre de longueur minimale 266
a _ Définition d'une ultramétrique 266
b _ Équivalence entre ultramétrique et hiérarchie indicée 266
c _ L'ultramétrique sous dominante 268
d _ Arbre de longueur minimale : définition et généralités 270
f _ Arbre de longueur minimale : algorithme de Florek 271
g _ Arbre de longueur minimale : algorithme de Florek 272
h _ Exemple d'application 273

277

régation de deux éléments :
ralisé 278
erche en chaîne des voisins réciproques 280
281
e 282
on 1 282
on 2 285
cription statistique des classes 287
cation mixte 288

b _ Cas général
c _ Les critères extern

6.6 Recherche non superv
6.6.1 Algorithme Apri
a _ Les étapes de l'alg
b _ Support, confianc
c _ Règles et visualisa
6.6.2 Méthodes d'anal
a _ Extraction de règl
b _ Mesure et évaluat

a _ Les étapes de l'algorithme 288
b _ Choix du nombre de classes par coupure de l'arbre 289
c _ Procédure de consolidation 290

6.3.2 Description statistique des classes 291
a _ Valeurs-test pour les variables continues 291
b _ Valeurs-test pour les variables nominales 292
c _ Variables caractéristiques d'une classe 294

6.4 Complémentarité entre analyse factorielle et classification 295

6.4.1 Utilisation conjointe des axes principaux et de la classification 295
a _ Nécessité... et insuffisance des méthodes factorielles 295
b _ Mise en œuvre pratique dans le cas de la classification mixte 297
c _ Autres travaux sur la complémentarité 298

6.4.2 Aspects techniques et théoriques de la complémentarité 299
a _ Classification des lignes ou colonnes d'un tableau de contingence 299
b _ Un exemple de coïncidence entre les deux approches 299

6.4.3 Valeurs propres et indices de niveau 302
a _ Quelques inégalités 302
b _ Le cas des tables de contingence structurées par blocs 303
c _ Lien entre valeurs propres et indices 303

6.4.4 La complémentarité en pratique : un exemple 304
a _ Les étapes 304
b _ L'espace des variables actives 305
c _ Exemples de description automatique de trois classes 307
d _ Projection de variables signalétiques en supplémentaires 309

6.5 Validation des classifications 311

6.5.1 Cadre général 312
a _ Cadre inférentiel général 312
b _ Validation empirique, calculs de stabilité 312
c _ Importance des critères externes 312

6.5.2 L'hypothèse d'absence de structure, les modèles 313
a _ Modèles de mélanges 313
b _ Modèles de partitions fixes 315
c _ Autre modèles 315

6.5.3 Nombre de classes à retenir 316
a _ Cas de la classification mixte 316
b _ Cas général 317
c _ Les critères externes 317

6.6 Recherche non supervisée de règles d'associations 318

6.6.1 Algorithme Apriori pour la recherche de règles 319
a _ Les étapes de l'algorithme 319
b _ Support, confiance, confiance attendue, *Lift* 320
c _ Règles et visualisation 321

6.6.2 Méthodes d'analyse statistique implicative 322
a _ Extraction de règles, indices d'implication et graphe orienté 322
b _ Mesure et évaluation de règles 323

c _ Graphes de règles 324

6.7 Annexe technique du chapitre 6 325

6.7.1 Les correspondances hiérarchiques 325

6.7.2 L'algorithme EM 327

Chapitre 7
Analyse discriminante, classification supervisée 329

7.1 Analyse linéaire discriminante 330

7.1.1 Formulation du problème et notations 330

7.1.2 Fonctions linéaires discriminantes 332

a _ Décomposition de la matrice de covariance 332

b _ Calcul des fonctions linéaires discriminantes 334

c _ Diagonalisation d'une matrice symétrique 334

7.2 Lien avec d'autres méthodes 335

7.2.1 Cas de deux classes : équivalence avec la régression multiple 335

7.2.2 Lien avec l'analyse canonique 337

7.2.3 Lien avec l'analyse des correspondances 338

7.2.4 Une analyse avec une métrique particulière 340

7.3 Règles de classement 341

7.3.1 Le modèle bayésien d'affectation 341

7.3.2 Le modèle bayésien dans le cas normal 342

7.3.3 Autres règles d'affectation 343

a _ Estimation de la densité par noyaux 343

b _ Règle des m plus proches voisins 345

7.3.4 Qualité des règles de classement 345

7.4 Régularisation en analyse discriminante 346

7.4.1 Analyse régularisée 347

7.4.2 Analyse régularisée par axes principaux 347

a _ Axes principaux de l'échantillon total 348

b _ Axes principaux de l'échantillon projeté 349

c _ Axes principaux dans les groupes 349

d _ Exemple numérique d'application 350

e _ Analyse discriminante sur variables qualitatives 352

f _ Analyse discriminante barycentrique 353

g _ Note sur le "scoring" 353

7.5 Régression logistique 354

7.5.1 Le modèle logistique 355

7.5.2 Estimation et tests des coefficients 356

a _ Procédure d'estimation 356

b _ Comparaison de deux modèles 358

c _ Modèle avec interaction 358

7.6 Segmentation 358
7.6.1 Formulation du problème, principe et vocabulaire 359
7.6.2 Construction d'un arbre de décision binaire 361
a _ Algorithme général de segmentation 361
b _ Cas de la régression 363
c _ Cas de la discrimination 366
7.6.3 Sélection du "meilleur sous-arbre" 369
a _ Procédures de sélection 369
b _ Estimation de l'Erreur Théorique de Prévision 370
c _ Estimation du Taux d'Erreur Théorique de classement 370
7.6.4 Divisions équi-réductrices et équi-divisantes 372
a _ Divisions équi-réductrices 372
b _ Divisions équi-divisantes 373
7.6.5 Lien avec les méthodes de classement 373

7.7 Discrimination et réseaux de neurones 374
7.7.1 Schéma et modèle du perceptron multi-couches 375
7.7.2 Modèles supervisés 376
7.7.3 Modèles non-supervisés ou auto-organisés 378
7.7.4 SVM : « Séparateurs à vastes marges » ou « Support Vector Machines » 379
a _ Hyperplan séparateur 380
b _ Cas de deux groupes séparables 380
c _ Cas de deux groupes non séparables 381
d _ Extension des descripteurs 382
7.7.5 Les modèles statistiques et les réseaux de neurones 383

7.8 Annexe technique du chapitre 7 384
7.8.1 Distances entre distributions 384
7.8.2 Distance de Mahalanobis et information 385

Chapitre 8
Analyse de données structurées 387

8.1 Analyses partielles et projetées 389
8.1.1 Définition du coefficient de corrélation partielle 389
8.1.2 Calcul des covariances et corrélations partielles 390
a _ Cas de deux variables 390
b _ Cas de p variables (X) et de q variables (Z) 391
8.1.3 Analyse du nuage résiduel ou analyse partielle 392
8.1.4 Autres analyses partielles ou projetées 393
a _ Analyse canonique des correspondances 394
b _ Analyse non-symétrique des correspondances 395
c _ Régression PLS (Partial Least Squares) 396

8.2 Structures de graphe, analyses locales 397

8.2.1 Variance locale et covariance locale d'une variable 397
a _ Matrice de contiguïté 398
b _ Coefficient de contiguïté de Geary (1954) 399
c _ Nouvelle définition de la variance locale 399
d _ Bornes pour c(x) 400
e _ Analyse des correspondances des matrices associées M 401

8.2.2 Analyse locale 402

8.2.3 Analyse de contiguïté et projections révélatrices 403
a _ Analyse de contiguïté 403
b _ Représentation de groupes par projection 404
c _ Liens avec les analyses partielles 405

8.2.4 Extensions, généralisations, applications 405

8.2.5 Cas particuliers : Structure de partition 406
a _ Analyse inter-classes 406
b _ Analyse intra-classes 407

8.3 Tableaux multiples, groupes de variables 408

8.3.1 Quelques travaux de référence 408

8.3.2 Analyses procrustéennes 410
a _ Analyse procrustéenne orthogonale 410
b _ Analyse procrustéenne sans contrainte 412
c _ Formulaire de quelques méthodes d'analyse impliquant deux groupes de variables 413

8.3.3 Méthode STATIS 413
a _ Notations 413
b _ Comparaison globale entre les tableaux : l'interstructure 414
c _ Le nuage moyen ou compromis : l' intrastructure 414
d _ Représentation simultanée des nuages partiels : les trajectoires 415

8.3.4 Analyse factorielle multiple 415
a _ Une analyse en composantes principales pondérée 416
b _ Recherche de facteurs communs (intrastructures) 416
c _ Représentation des groupes de variables (interstructure) 417
d _ Représentations superposées des nuages partiels des groupes actifs (trajectoires) 417

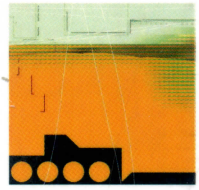
8.3.5 Analyse canonique généralisée 418
a _ Formulation générale 419
b _ Propriétés de l'Analyse Canonique Généralisée 420
c _ Utilisation en pratique de l'analyse canonique généralisée 423

Bibliographie 425

Index des auteurs 454

Index des matières 460

SCIENCES SUP



4e édition

Ludovic Lebart
Marie Piron
Alain Morineau

STATISTIQUE EXPLORATOIRE MULTIDIMENSIONNELLE

Visualisation et inférence en fouilles de données

Cette quatrième édition entièrement refondue et complétée s'adresse aux étudiants, chercheurs, ingénieurs, professeurs de toutes disciplines confrontés dans leurs travaux aux recueils de données multidimensionnelles. Les enquêtes socio-économiques, épidémiologiques et de marketing en sont des exemples courants.

Appuyé sur de nombreux exemples, l'ouvrage présente les concepts de base et les fondements des méthodes exploratoires et rend compte des développements récents. Il insiste sur la place centrale, dans la démarche « Fouille de données » (ou *Data Mining*), des visualisations fondées sur des principes géométriques et algébriques simples, sous le contrôle de méthodes inférentielles robustes.

Le livre peut être lu à plusieurs niveaux : celui de l'étudiant (Master, écoles d'ingénieur), celui du praticien, celui de l'utilisateur exigeant, enfin celui du chercheur en méthodologie statistique.

LUDOVIC LEBART est directeur de recherche CNRS à l'École nationale supérieure des télécommunications (ENST).

MARIE PIRON est chargée de recherche à l'Institut de recherche pour le développement (IRD).

ALAIN MORINEAU ancien directeur du Centre international de statistiques et d'informatique appliquées (CISIA), dirige la revue électronique MODULAD.

MATHÉMATIQUES
PHYSIQUE
CHIMIE
SCIENCES DE L'INGÉNIEUR
INFORMATIQUE
SCIENCES DE LA VIE
SCIENCES DE LA TERRE

9 782100 496167

ISBN 2 10 049616 6

1er cycle | 2e cycle | 3e cycle
1 2 3 4 5 6 7 8
LICENCE | MASTER | DOCTORAT

www.dunod.com

DUNOD