

ANALYSE DE DONNÉES ET INFORMATIQUE

Fontainebleau, du 19 au 30 mars 1979

COURS DE LA COMMISSION DES COMMUNAUTÉS EUROPÉENNES

Dirigé par
R. TOMASSONE

réalisé par
M. AMIRCHAHY et D. NÉEL
INRIA - Service Formation

Édité par



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

B. P. 105 - 78150 LE CHESNAY - Tél. 954 90 20

c1195

ANALYSE DE DONNÉES ET INFORMATIQUE

Fontainebleau, du 19 au 30 mars 1979



COURS DE LA COMMISSION DES COMMUNAUTÉS EUROPÉENNES

Dirigé par
R. TOMASSONE



réalisé par
M. AMIRCHAHY et D. NÉEL

INRIA - Service Formation

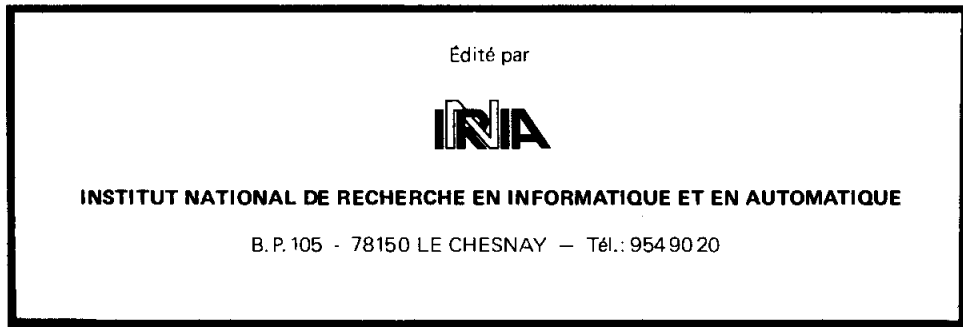
BIBLIOTHEQUE DU CERIST

Édité par



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

B.P. 105 - 78150 LE CHESNAY — Tél.: 954 90 20



C1485

BIBLIOTHEQUE DU CERIST

PREFACE

La décision d'organiser un cours d'analyse des données et d'informatique avait été prise immédiatement après les premières journées d'analyse des données et d'informatique qui s'étaient déroulées à Versailles du 7 au 9 Septembre 1977. Ceux qui avaient assisté à ces journées avaient, sans nul doute, été surpris par la variété des problèmes abordés et par la multiciplité des pistes de travail ; cette abondance permettait mal d'avoir une vue synthétique, cette multiplicité des sujets masquait souvent les grandes orientations. Le cours avait donc pour premier objectif de faire un bilan le plus exhaustif possible des grandes méthodes de l'analyse des données et de dégager quelques idées de recherche. Dans la mesure où le public auquel il s'adressait connaissait, au moins partiellement, les bases de l'analyse des données, peu de rappels ont été faits ; si bien que le lecteur devra se reporter aux ouvrages classiques, cités dans les références bibliographiques de chaque article.

C'est dans le cadre du château où se trouve l'Ecole des Mines de Fontainebleau qu'a eu lieu ce cours du 19 au 30 Mars 1979. Il a réuni 63 personnes venant des pays suivants : Belgique, Pays-Bas, Grande-Bretagne, Irlande, République Fédérale Allemande, Italie, Espagne, Portugal, Etats-Unis, Canada, Norvège, Suède, Pologne, Yougoslavie et France.

L'enseignement avait été scindé en deux grandes parties : analyse factorielle et classification, complétées par deux exposés sur la régression.

Les cinq premiers articles traitent de problèmes liés plus ou moins directement à l'analyse factorielle ; dans le premier, d'ESCOUFIER et al., une présentation formelle à partir du schéma de dualité dégage aussi bien quelques visions nouvelles de problèmes anciens que des orientations plus originales, en particulier dans le traitement simultané de plusieurs matrices de données. LEBART développe l'application de l'analyse des correspondances multiples au problème des dépouillements d'enquête ; ce faisant il propose une véritable stratégie d'analyse des données d'enquête.

CONTENTS

(Y. ESCOUFLIER, J.P. PAGES and R. TOMASSONE)	AN OVERVIEW ON FACTOR ANALYSIS	1
1.	Notation and background	2
1.1.	Some historical aspects	2
1.2.	Duality diagram	4
1.3.	How to represent a qualitative variable	6
1.4.	Operators related to a data matrix	7
2.	Classical problems : a new glance	10
3.	Some new problems	12
3.1.	Joint analysis of several data matrices	12
3.2.	Choice of a metric	14
3.3.	Selection of variables	17
3.4.	Qualitative variables	19
	References	21
(L. LEBART)	MULTIPLE CORRESPONDENCE ANALYSIS AS A METHOD FOR PROCESSING SURVEY DATA	23
1.	Operational survey procedures	25
2.	Multicorrespondence analysis	27
2.1.	Introduction	27
2.2.	The study of two modalities : (cardQ = 2)	29
2.3.	Generalization to more than two questions	31
2.4.	The properties of multianalysis	33
3.	Application to survey processing	40
3.1.	An example of a socio-administrative chart	40
3.2.	Visualized regression and supplementary variables	44
3.3.	An example of chart illustration	45
4.	Practical application of the methods	47
4.1.	Data preparation and control	47
	References	51
(J.B. KRUSKAL)	MULTILINEAR AND GEOMETRICAL MODELS : OVERVIEW AND RELATIONSHIPS	55
1.	Some bilinear models (R dimensions)	54
2.	Least squares fitting	56
3.	An example	56
4.	Computation of least squares fitted values for the bilinear models	56
5.	Some bilinear methods	60
6.	Classification of multilinear models	61

7.	Multidimensional scaling models (R dimensions)	61
8.	Some trilinear models (R dimensions)	63
9.	A three-way multidimensional scaling model (R dimensions)	64
	References	66
(J.B. KRUSKAL) PATH MODELS		67
1.	Some simple examples of path models	68
2.	Path models as explanations of correlation matrices	73
3.	Alternative path models for the same data	73
4.	A relationship between principal components analysis and canonical correlation	73
5.	The MIMIC model	77
6.	The uses of path models, and standardized vs. nonstandardized coefficients	77
	References	78
(J.D.F. HABBEMA and J. HERMANS) DISCRIMINANT ANALYSIS		79
1.	Introduction to discriminant analysis methods	80
2.	The Bayes Discriminant analysis model	85
2.1.	Introduction	85
2.2.	Bayes theorem and Bayes allocation rules	85
2.3.	Estimation of probability density functions	89
2.4.	Some further problems	90
3.	Measurement of performance of Bayes discriminant analysis systems	94
3.1.	The forced classification matrix	94
3.2.	Performance measures derived from the forced classification matrix	94
3.3.	Classification matrix with doubt	96
3.4.	Exclusion matrix	96
3.5.	Reliability	97
3.6.	The approach to reliability checking	97
3.7.	Continuous scoring rules derived from the posterior probabilities	99
4.	A Kernel method for density estimation and its implementation in the ALLOC computer programs	102
4.1.	Introduction	102
4.2.	Implementation in the ALLOC computer programs	104
4.3.	Variable Kernel density estimation	107
4.4.	Kernel density estimation in mixed data	110

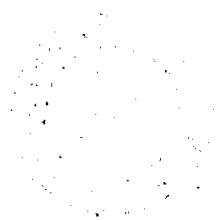
5.	Selection of variables	115
5.1.	Introduction	115
5.2.	Criterion and algorithm	115
5.3.	Comparison of F-statistic and error rate	117
(R.M. CORMACK)	CLASSIFICATION : AN OVERVIEW	125
1.	Introduction	126
2.	Formal listing of possibilities	128
2.1.	Objects	128
2.2.	Variables	128
2.3.	Scaling	128
2.4.	Similarity measures	129
2.5.	Output structure	130
2.6.	Use of clusters	131
3.	Single-link clustering	131
4.	General agglomerative algorithm	132
4.1.	Algorithms and methods	133
5.	Partitioning algorithms	134
5.1.	Graph theory	135
6.	Density estimation and mixtures	136
7.	Similarities of similarities	137
8.	Clusterability	139
9.	Validation	140
9.1.	Measures of distortion	141
10.	Cluster analysis and spatial point processes	143
	References	145
(A.D. GORDON)	ON THE ASSESSMENT AND COMPARISON OF CLASSIFICATIONS	149
1.	Introduction	150
1.1.	Adaptative clustering schemes	151
1.2.	Significance testing approach	151
1.3.	Properties of clustering criteria	152
1.4.	Comparative studies	152
2.	Comparing classifications	152
2.1.	Comparing partitions and coverings	153
2.2.	The intersection of trees	155
2.3.	A pruning algorithm	156
2.4.	Regrafting pruned branches	157
3.	Discussion	158
	References	159

(A.D. GORDON)	METHODS OF CONSTRAINED CLASSIFICATION	161
1.	Introduction	162
2.	Simpler constrained classification problems	164
2.1.	Hierarchical constrained classification	164
2.2.	Dynamic programming algorithm	165
2.3.	Variable barriers algorithm	166
2.4.	Generalisations	168
3.	More complicated constrained classification problems	168
	References	169
(H.H. BOCK)	CLUSTERING BY DENSITY ESTIMATION	173
1.	Introduction	174
2.	Unimodal and multimodal densities; density estimation	175
3.	The gradient algorithm of SCHNELL and IHM	178
4.	The gradient method of KOPP	180
5.	The smoothing parameter h	181
6.	Special cases	182
7.	Conclusion and remarks	185
	Bibliography	184
(H.H. BOCK)	SIMULTANEOUS CLUSTERING OF OBJECTS AND VARIABLES	187
1.	Introduction and examples	188
2.	Mathematical clustering problem	189
3.	No-interaction model	191
4.	The general model	193
5.	The interaction model	198
6.	Generalizations and final remarks	201
	Bibliography	202
(H.H. BOCK)	FUZZY CLUSTERING PROCEDURES	205
1.	Introduction	205
2.	Fuzzy sets and partitions	207
3.	Optimal fuzzy partitions	208
4.	A fuzzy clustering algorithm	211
5.	Evaluation of the method	213
	Bibliography	216
(E. DIDAY)	CANONICAL ANALYSIS FROM THE AUTOMATIC CLASSIFICATION POINT OF VIEW	219
1.	Introduction	226

2.	Canonical typological analysis	221
2.1.	The data	221
2.2.	The criterion to optimize	222
2.3.	The centred and non-centred cases	224
2.4.	Comparison of both cases	229
2.5.	The search for several canonical factors in each cluster	231
2.6.	Some numerical aspects	233
3.	Factor analysis of the local correspondances	233
3.1.	Statement of the problem	233
3.2.	Expression of the criterion to maximize	233
3.3.	Relation between the criterion to minimize and the contingency X^2	235
3.4.	The algorithm	236
3.5.	Maximization of the criterion	237
3.6.	Another possible algorithm	237
4.	Local discriminant analysis	237
4.1.	Statement of the problem	237
4.2.	Expression of the criterion to minimize	238
4.3.	The algorithm	240
4.4.	Choice of the initial partition	241
4.5.	Assistance in the decision making processes	241
5.	Canonical typological analysis in the case of several parameters groups	242
5.1.	Notations	242
5.2.	The problem	243
5.3.	The algorithm	243
6.	Conclusion	243
	Bibliography	244

(C. CHARLES and Y. LECHEVALLIER) PATTERN RECOGNITION BY A PIECEWISE

	POLYNOMIAL APPROXIMATION WITH VARIABLE JOINTS	245
1.	Introduction	246
2.	The problem	246
3.	Algorithms	247
3.1.	First method - Transfer method	247
3.2.	Second method - Dynamic programming algorithm	248
4.	Polynomial approximation of curves	250
5.	The modified model	251
6.	Generalization of this algorithm to the case of a spline function	252



6.1.	Spline function	252
6.2.	Interpolation SPLINE function	255
7.	The second algorithm with a spline function	253
8.	Generalization	254
9.	Conclusion	254
	Bibliography	259

(J. BRENOT)	BIASED ESTIMATION IN THE LINEAR MODEL	261
1.	Introduction	262
2.	Notation and classical results	262
3.	Some biased estimators	264
3.1.	Shrinkage estimator	264
3.2.	Ridge regression	266
3.3.	Equality constrained least squares estimator	269
4.	Comparisons	273
5.	Some synthetical aspects	274
5.1.	Class of generalised ridge estimators	274
5.2.	Class of James estimators	274
5.3.	Class of constrained estimators	274
6.	Conclusion	274
	References	275

(A. MORINEAU)	LEAST SQUARES FITTING AND ROBUST REGRESSION	279
1.	Introduction	280
2.	Least squares fitting. A local description of the adjustment	282
3.	Least absolute values criteria	296
4.	Robust regression : M-estimators	298
5.	Regression and data analysis	304
	Bibliography	307



AN OVERVIEW ON FACTOR ANALYSIS

ESCOUFIER Y. (*), PAGES J.P. (**), TOMASSONE R. (***)

Résumé :

Cette contribution fournit un bilan général et formel de l'analyse factorielle. Elle permet ainsi de voir des problèmes classiques sous un angle nouveau, en plus, elle permet de dégager quelques voies nouvelles de recherche liées à des problèmes pratiques.

Abstract :

This paper gives a general and formal overview on factor analysis. So, it is possible to have a new glance at old problems, but it also gives the opportunity to point out new research orientations linked with practical problems.

(*) Centre de Recherche en Informatique et Gestion. Avenue d'Occitanie, 34075 Montpellier Cedex.

(**) Centre d'Etudes Nucléaires, 92260 Fontenay aux Roses.

(***) Institut National de la Recherche Agronomique, 78350 Jouy en Josas.