

DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES



Data Analysis

Edited by Gérard Govaert

ISTE

 **WILEY**

Contents

Preface	xiii
Chapter 1. Principal Component Analysis: Application to Statistical Process Control	1
Gilbert SAPORTA, Ndèye NIANG	
1.1. Introduction	1
1.2. Data table and related subspaces	2
1.2.1. Data and their characteristics	2
1.2.2. The space of statistical units	5
1.2.3. Variables space	7
1.3. Principal component analysis	8
1.3.1. The method	8
1.3.2. Principal factors and principal components	8
1.3.3. Principal factors and principal components properties	10
1.4. Interpretation of PCA results	11
1.4.1. Quality of representations onto principal planes	11
1.4.2. Axis selection	12
1.4.3. Internal interpretation	13
1.4.4. External interpretation: supplementary variables and individuals	15
1.5. Application to statistical process control	18
1.5.1. Introduction	18
1.5.2. Control charts and PCA	20
1.6. Conclusion	22
1.7. Bibliography	23
Chapter 2. Correspondence Analysis: Extensions and Applications to the Statistical Analysis of Sensory Data	25
Jérôme PAGÈS	
2.1. Correspondence analysis	25

2.1.1. Data, example, notations	25
2.1.2. Questions: independence model	26
2.1.3. Intensity, significance and nature of a relationship between two qualitative variables	27
2.1.4. Transformation of the data	28
2.1.5. Two clouds	29
2.1.6. Factorial analysis of \mathbf{X}	30
2.1.7. Aid to interpretation	32
2.1.8. Some properties	33
2.1.9. Relationships to the traditional presentation	35
2.1.10. Example: recognition of three fundamental tastes	36
2.2. Multiple correspondence analysis	39
2.2.1. Data, notations and example	39
2.2.2. Aims	41
2.2.3. MCA and CA	41
2.2.4. Spaces, clouds and metrics	42
2.2.5. Properties of the clouds in CA of the CDT	43
2.2.6. Transition formulae	45
2.2.7. Aid for interpretation	45
2.2.8. Example: relationship between two taste thresholds	46
2.3. An example of application at the crossroads of CA and MCA	50
2.3.1. Data	50
2.3.2. Questions: construction of the analyzed table	51
2.3.3. Properties of the CA of the analyzed table	53
2.3.4. Results	54
2.4. Conclusion: two other extensions	63
2.4.1. Internal correspondence analysis	63
2.4.2. Multiple factor analysis (MFA)	63
2.5. Bibliography	64
Chapter 3. Exploratory Projection Pursuit	67
Henri CAUSSINUS, Anne RUIZ-GAZEN	
3.1. Introduction	67
3.2. General principles	68
3.2.1. Background	68
3.2.2. What is an interesting projection?	69
3.2.3. Looking for an interesting projection	70
3.2.4. Inference	70
3.2.5. Outliers	71
3.3. Some indexes of interest: presentation and use	71
3.3.1. Projection indexes based on entropy measures	71
3.3.2. Projection indexes based on L^2 distances	73
3.3.3. Chi-squared type indexes	75

3.3.4. Indexes based on the cumulative empirical function	75
3.4. Generalized principal component analysis	76
3.4.1. Theoretical background	76
3.4.2. Practice	78
3.4.3. Some precisions	79
3.5. Example	81
3.6. Further topics	86
3.6.1. Other indexes, other structures	86
3.6.2. Unsupervised classification	86
3.6.3. Discrete data	87
3.6.4. Related topics	88
3.6.5. Computation	89
3.7. Bibliography	89
Chapter 4. The Analysis of Proximity Data	93
Gerard D'AUBIGNY	
4.1. Introduction	93
4.2. Representation of proximity data in a metric space	97
4.2.1. Four illustrative examples	97
4.2.2. Definitions	100
4.3. Isometric embedding and projection	103
4.3.1. An example of computations	105
4.3.2. The additive constant problem	106
4.3.3. The case of observed dissimilarity measures blurred by noise	108
4.4. Multidimensional scaling and approximation	108
4.4.1. The parametric MDS model	109
4.4.2. The Shepard founding heuristics	111
4.4.3. The majorization approach	114
4.4.4. Extending MDS to a semi-parametric setting	119
4.5. A fielded application	122
4.5.1. Principal coordinates analysis	122
4.5.2. Dimensionality for the representation space	123
4.5.3. The scree test	125
4.5.4. Recourse to simulations	127
4.5.5. Validation of results	127
4.5.6. The use of exogenous information for interpreting the output configuration	131
4.5.7. Introduction to stochastic modeling in MDS	137
4.6. Bibliography	139
Chapter 5. Statistical Modeling of Functional Data	149
Philippe BESSE, Hervé CARDOT	
5.1. Introduction	149

5.2. Functional framework	152
5.2.1. Functional random variable	152
5.2.2. Smoothness assumption	153
5.2.3. Smoothing splines	154
5.3. Principal components analysis	156
5.3.1. Model and estimation	156
5.3.2. Dimension and smoothing parameter selection	158
5.3.3. Some comments on discretization effects	159
5.3.4. PCA of climatic time series	160
5.4. Linear regression models and extensions	161
5.4.1. Functional linear models	162
5.4.2. Principal components regression	163
5.4.3. Roughness penalty approach	163
5.4.4. Smoothing parameters selection	164
5.4.5. Some notes on asymptotics	165
5.4.6. Generalized linear models and extensions	165
5.4.7. Land use estimation with the temporal evolution of remote sensing data	166
5.5. Forecasting	169
5.5.1. Functional autoregressive process	169
5.5.2. Smooth ARH(1)	171
5.5.3. Locally ARH(1) processes	172
5.5.4. Selecting smoothing parameters	173
5.5.5. Some asymptotic results	173
5.5.6. Prediction of climatic time series	173
5.6. Concluding remarks	176
5.7. Bibliography	177
Chapter 6. Discriminant Analysis	181
Gilles CELEUX	
6.1. Introduction	181
6.2. Main steps in supervised classification	182
6.2.1. The probabilistic framework	182
6.2.2. Sampling schemes	183
6.2.3. Decision function estimation strategies	184
6.2.4. Variables selection	185
6.2.5. Assessing the misclassification error rate	187
6.2.6. Model selection and resampling techniques	189
6.3. Standard methods in supervised classification	190
6.3.1. Linear discriminant analysis	191
6.3.2. Logistic regression	192
6.3.3. The K nearest neighbors method	195
6.3.4. Classification trees	197

6.3.5. Single hidden layer back-propagation network	199
6.4. Recent advances	204
6.4.1. Parametric methods	204
6.4.2. Radial basis functions	207
6.4.3. Boosting	208
6.4.4. Support vector machines	209
6.5. Conclusion	211
6.6. Bibliography	212
Chapter 7. Cluster Analysis	215
Mohamed NADIF, Gérard GOVAERT	
7.1. Introduction	215
7.2. General principles	217
7.2.1. The data	217
7.2.2. Visualizing clusters	218
7.2.3. Types of classification	218
7.2.4. Objectives of clustering	222
7.3. Hierarchical clustering	224
7.3.1. Agglomerative hierarchical clustering (AHC)	225
7.3.2. Agglomerative criteria	226
7.3.3. Example	227
7.3.4. Ward's method or minimum variance approach	227
7.3.5. Optimality properties	228
7.3.6. Using hierarchical clustering	231
7.4. Partitional clustering: the k -means algorithm	233
7.4.1. The algorithm	233
7.4.2. k -means: a family of methods	234
7.4.3. Using the k -means algorithm	236
7.5. Miscellaneous clustering methods	239
7.5.1. Dynamic cluster method	239
7.5.2. Fuzzy clustering	240
7.5.3. Constrained clustering	241
7.5.4. Self-organizing map	242
7.5.5. Clustering variables	243
7.5.6. Clustering high-dimensional datasets	244
7.6. Block clustering	245
7.6.1. Binary data	247
7.6.2. Contingency table	248
7.6.3. Continuous data	249
7.6.4. Some remarks	250
7.7. Conclusion	251
7.8. Bibliography	251

Chapter 8. Clustering and the Mixture Model	257
Gérard GOVAERT	
8.1. Probabilistic approaches in cluster analysis	257
8.1.1. Introduction	257
8.1.2. Parametric approaches	258
8.1.3. Non-parametric methods	259
8.1.4. Validation	260
8.1.5. Notation	260
8.2. The mixture model	261
8.2.1. Introduction	261
8.2.2. The model	261
8.2.3. Estimation of parameters	262
8.2.4. Number of components	263
8.2.5. Identifiability	263
8.3. EM algorithm	263
8.3.1. Introduction	263
8.3.2. Complete data and complete-data likelihood	264
8.3.3. Principle	264
8.3.4. Application to mixture models	265
8.3.5. Properties	266
8.3.6. EM: an alternating optimization algorithm	266
8.4. Clustering and the mixture model	267
8.4.1. The two approaches	267
8.4.2. Classification likelihood	267
8.4.3. The CEM algorithm	268
8.4.4. Comparison of the two approaches	269
8.4.5. Fuzzy clustering	270
8.5. Gaussian mixture model	271
8.5.1. The model	271
8.5.2. CEM algorithm	272
8.5.3. Spherical form, identical proportions and volumes	273
8.5.4. Spherical form, identical proportions but differing volumes	274
8.5.5. Identical covariance matrices and proportions	275
8.6. Binary variables	275
8.6.1. Data	275
8.6.2. Binary mixture model	276
8.6.3. Parsimonious model	277
8.6.4. Example of application	279
8.7. Qualitative variables	279
8.7.1. Data	279
8.7.2. The model	279
8.7.3. Parsimonious model	281
8.8. Implementation	282

8.8.1. Choice of model and of the number of classes	282
8.8.2. Strategies for use	283
8.8.3. Extension to particular situations	283
8.9. Conclusion	284
8.10. Bibliography	284
Chapter 9. Spatial Data Clustering	289
Christophe AMBROISE, Mo DANG	
9.1. Introduction	289
9.1.1. The spatial data clustering problem	289
9.1.2. Examples of applications	290
9.2. Non-probabilistic approaches	293
9.2.1. Using spatial variables	293
9.2.2. Transformation of variables	293
9.2.3. Using a matrix of spatial distances	293
9.2.4. Clustering with contiguity constraints	294
9.3. Markov random fields as models	295
9.3.1. Global methods and Bayesian approaches	295
9.3.2. Markov random fields	297
9.3.3. Markov fields for observations and classes	300
9.3.4. Supervised segmentation	301
9.4. Estimating the parameters for a Markov field	305
9.4.1. Supervised estimation	305
9.4.2. Unsupervised estimation with EM	307
9.4.3. Classification likelihood and inertia with spatial smoothing	310
9.4.4. Other methods of unsupervised estimation	312
9.5. Application to numerical ecology	313
9.5.1. The problem	313
9.5.2. The model: Potts field and Bernoulli distributions	314
9.5.3. Estimating the parameters	315
9.5.4. Resulting clustering	315
9.6. Bibliography	316
List of Authors	319
Index	323

Preface

Statistical analysis has traditionally been separated into two phases: an exploratory phase, drawing on a set of descriptive and graphical techniques, and a decisional phase, based on probabilistic models. Some of the tools employed as part of the exploratory phase belong to *descriptive statistics*, whose elementary exploratory methods consider only a very limited number of variables. Other tools belong to *data analysis*, the subject matter of this book. This topic comprises more elaborate exploratory methods to handle multidimensional data, and is often seen as stepping beyond a purely exploratory context.

The first part of this book is concerned with methods for obtaining the pertinent dimensions from a collection of data. The variables so obtained provide a synthetic description, often leading to a graphical representation of the data. A considerable number of methods have been developed, adapted to different data types and different analytical goals. Chapters 1 and 2 discuss two reference methods, namely Principal Components Analysis (PCA) and Correspondence Analysis (CA), which we illustrate with examples from statistical process control and sensory analysis. Chapter 3 looks at a family of methods known as Projection Pursuit (less well known, but with a promising future), that can be seen as an extension of PCA and CA, which makes it possible to specify the structures that are being sought. Multidimensional positioning methods, discussed in Chapter 4, seek to represent proximity matrix data in low-dimensional Euclidean space. Chapter 5 is devoted to functional data analysis where a function such as a temperature or rainfall graph, rather than a simple numerical vector, is used to characterize individuals.

The second part is concerned with methods of clustering, which seek to organize data into homogenous classes. These methods provide an alternative means, often complementary to those discussed in the first part, of synthesizing and analyzing data. In view of the clear link between clustering and discriminant analysis – in pattern recognition the former is termed unsupervised and the latter supervised learning – Chapter 6 gives a general introduction to discriminant analysis. Chapter 7 then

provides an overall picture of clustering. The statistical interpretation of clustering in terms of mixtures of probability distributions is discussed in Chapter 8 and Chapter 9 looks at how this approach can be applied to spatial data.

I would like to express my heartfelt thanks to all the authors who were involved in this publication. Without their expertise, their professionalism, their invaluable contributions and the wealth of their experience, it would not have been possible.

Gérard GOVAERT

Traditionally, statistical analysis is divided into two stages: an exploratory stage, based on a comprehensive process of descriptive and graphic techniques; and a decision stage, based on probabilistic models. The exploratory phase, also called data analysis - and which is the object of this book - is the process of transforming data and extracting useful information. This process is normally composed of detailed and often complex exploratory methods that apply to multidimensional data and often exceed the capacity of the exploratory framework itself.

The first part of this book is devoted to methods aimed at identifying relevant dimensions of the data. The variables thus obtained provide a synthetic description which often results in a graphical representation of the data. After a general presentation of discriminating analysis, the second part of the book is devoted to a presentation of clustering methods which constitute another method (which is often complementary to the methods described in the first part of the book) to synthesize and analyze the data. The book concludes by examining the interrelations between data mining and data analysis.

Gérard Govaert is Professor at the University of Technology of Compiègne, France. He is also a member of the CNRS Laboratory Heudiasyc (Heuristics and diagnostics of complex systems). His research interests include latent structure modeling, model selection, model-based cluster analysis, block clustering and statistical pattern recognition. He is one of the authors of the MIXMOD (MIXture MODeling) software.

ISTE

www.iste.co.uk



WILEY

wiley.com



9 781848 210981