# Contents

**Preface** xiii

**Acknowledgement** xv

**1 An Introduction to classification and clustering** **1**

1.1 Introduction 1

1.2 Reasons for classifying 3

1.3 Numerical methods of classification – cluster analysis 4

1.4 What is a cluster? 7

1.5 Examples of the use of clustering 9

1.5.1 Market research 9

1.5.2 Astronomy 9

1.5.3 Psychiatry 10

1.5.4 Weather classification 11

1.5.5 Archaeology 12

1.5.6 Bioinformatics and genetics 12

1.6 Summary 13

**2 Detecting clusters graphically** **15**

2.1 Introduction 15

2.2 Detecting clusters with univariate and bivariate plots of data 16

2.2.1 Histograms 16

2.2.2 Scatterplots 16

2.2.3 Density estimation 19

2.2.4 Scatterplot matrices 24

2.3 Using lower-dimensional projections of multivariate data for graphical representations 29

2.3.1 Principal components analysis of multivariate data 29

2.3.2 Exploratory projection pursuit 32

2.3.3 Multidimensional scaling 36

2.4 Three-dimensional plots and trellis graphics 38

2.5 Summary 41

**3 Measurement of proximity** **43**

3.1 Introduction 43

3.2 Similarity measures for categorical data 46

3.2.1 Similarity measures for binary data 46

3.2.2 Similarity measures for categorical data with more than two levels 47

3.3 Dissimilarity and distance measures for continuous data 49

3.4 Similarity measures for data containing both continuous and categorical variables 54

3.5 Proximity measures for structured data 56

3.6 Inter-group proximity measures 61

3.6.1 Inter-group proximity derived from the proximity matrix 61

3.6.2 Inter-group proximity based on group summaries for continuous data 61

3.6.3 Inter-group proximity based on group summaries for categorical data 62

3.7 Weighting variables 63

3.8 Standardization 67

3.9 Choice of proximity measure 68

3.10 Summary 69

**4 Hierarchical clustering** **71**

4.1 Introduction 71

4.2 Agglomerative methods 73

4.2.1 Illustrative examples of agglomerative methods 73

4.2.2 The standard agglomerative methods 76

4.2.3 Recurrence formula for agglomerative methods 78

4.2.4 Problems of agglomerative hierarchical methods 80

4.2.5 Empirical studies of hierarchical agglomerative methods 83

4.3 Divisive methods 84

4.3.1 Monothetic divisive methods 84

4.3.2 Polythetic divisive methods 86

4.4 Applying the hierarchical clustering process 88

4.4.1 Dendrograms and other tree representations 88

4.4.2 Comparing dendrograms and measuring their distortion 91

4.4.3 Mathematical properties of hierarchical methods 92

4.4.4 Choice of partition – the problem of the number of groups 95

4.4.5 Hierarchical algorithms 96

4.4.6 Methods for large data sets 97

4.5 Applications of hierarchical methods 98

4.5.1 Dolphin whistles – agglomerative clustering 98

4.5.2 Needs of psychiatric patients – monothetic divisive clustering 101

4.5.3 Globalization of cities – polythetic divisive method 101

4.5.4 Women's life histories – divisive clustering of sequence data 105
4.5.5 Composition of mammals' milk – exemplars, dendrogram seriation and choice of partition 107
4.6 Summary 110

**5 Optimization clustering techniques 111**
5.1 Introduction 111
5.2 Clustering criteria derived from the dissimilarity matrix 112
5.3 Clustering criteria derived from continuous data 113
5.3.1 Minimization of trace(W) 114
5.3.2 Minimization of det(W) 115
5.3.3 Maximization of trace ($BW^{-1}$) 115
5.3.4 Properties of the clustering criteria 115
5.3.5 Alternative criteria for clusters of different shapes and sizes 116
5.4 Optimization algorithms 121
5.4.1 Numerical example 124
5.4.2 More on *k*-means 125
5.4.3 Software implementations of optimization clustering 126
5.5 Choosing the number of clusters 126
5.6 Applications of optimization methods 130
5.6.1 Survey of student attitudes towards video games 130
5.6.2 Air pollution indicators for US cities 133
5.6.3 Aesthetic judgement of painters 136
5.6.4 Classification of 'nonspecific' back pain 141
5.7 Summary 142

**6 Finite mixture densities as models for cluster analysis 143**
6.1 Introduction 143
6.2 Finite mixture densities 144
6.2.1 Maximum likelihood estimation 145
6.2.2 Maximum likelihood estimation of mixtures of multivariate normal densities 146
6.2.3 Problems with maximum likelihood estimation of finite mixture models using the EM algorithm 150
6.3 Other finite mixture densities 151
6.3.1 Mixtures of multivariate *t*-distributions 151
6.3.2 Mixtures for categorical data – latent class analysis 152
6.3.3 Mixture models for mixed-mode data 153
6.4 Bayesian analysis of mixtures 154
6.4.1 Choosing a prior distribution 155
6.4.2 Label switching 156
6.4.3 Markov chain Monte Carlo samplers 157

6.5 Inference for mixture models with unknown number of components and model structure 157
6.5.1 Log-likelihood ratio test statistics 157
6.5.2 Information criteria 160
6.5.3 Bayes factors 161
6.5.4 Markov chain Monte Carlo methods 162
6.6 Dimension reduction – variable selection in finite mixture modelling 163
6.7 Finite regression mixtures 165
6.8 Software for finite mixture modelling 165
6.9 Some examples of the application of finite mixture densities 166
6.9.1 Finite mixture densities with univariate Gaussian components 166
6.9.2 Finite mixture densities with multivariate Gaussian components 173
6.9.3 Applications of latent class analysis 177
6.9.4 Application of a mixture model with different component densities 178
6.10 Summary 185

**7 Model-based cluster analysis for structured data 187**
7.1 Introduction 187
7.2 Finite mixture models for structured data 190
7.3 Finite mixtures of factor models 192
7.4 Finite mixtures of longitudinal models 197
7.5 Applications of finite mixture models for structured data 202
7.5.1 Application of finite mixture factor analysis to the ‘categorical versus dimensional representation’ debate 202
7.5.2 Application of finite mixture confirmatory factor analysis to cluster genes using replicated microarray experiments 205
7.5.3 Application of finite mixture exploratory factor analysis to cluster Italian wines 207
7.5.4 Application of growth mixture modelling to identify distinct developmental trajectories 208
7.5.5 Application of growth mixture modelling to identify trajectories of perinatal depressive symptomatology 211
7.6 Summary 212

**8 Miscellaneous clustering methods 215**
8.1 Introduction 215
8.2 Density search clustering techniques 216
8.2.1 Mode analysis 216
8.2.2 Nearest-neighbour clustering procedures 217
8.3 Density-based spatial clustering of applications with noise 220

8.4 Techniques which allow overlapping clusters 222
8.4.1 Clumping and related techniques 222
8.4.2 Additive clustering 223
8.4.3 Application of MAPCLUS to data on social relations in a monastery 225
8.4.4 Pyramids 226
8.4.5 Application of pyramid clustering to gene sequences of yeasts 230
8.5 Simultaneous clustering of objects and variables 231
8.5.1 Hierarchical classes 232
8.5.2 Application of hierarchical classes to psychiatric symptoms 234
8.5.3 The error variance technique 234
8.5.4 Application of the error variance technique to appropriateness of behaviour data 237
8.6 Clustering with constraints 237
8.6.1 Contiguity constraints 240
8.6.2 Application of contiguity-constrained clustering 242
8.7 Fuzzy clustering 242
8.7.1 Methods for fuzzy cluster analysis 245
8.7.2 The assessment of fuzzy clustering 246
8.7.3 Application of fuzzy cluster analysis to Roman glass composition 246
8.8 Clustering and artificial neural networks 249
8.8.1 Components of a neural network 250
8.8.2 The Kohonen self-organizing map 252
8.8.3 Application of neural nets to brainstorming sessions 254
8.9 Summary 255

**9 Some final comments and guidelines 257**
9.1 Introduction 257
9.2 Using clustering techniques in practice 260
9.3 Testing for absence of structure 262
9.4 Methods for comparing cluster solutions 264
9.4.1 Comparing partitions 264
9.4.2 Comparing dendrograms 265
9.4.3 Comparing proximity matrices 267
9.5 Internal cluster quality, influence and robustness 267
9.5.1 Internal cluster quality 268
9.5.2 Robustness – split-sample validation and consensus trees 269
9.5.3 Influence of individual points 271
9.6 Displaying cluster solutions graphically 273
9.7 Illustrative examples 278
9.7.1 Indo-European languages – a consensus tree in linguistics 279

9.7.2 Scotch whisky tasting – cophenetic matrices for comparing clusterings 279
9.7.3 Chemical compounds in the pharmaceutical industry 281
9.7.4 Evaluating clustering algorithms for gene expression data 285
9.8 Summary 287

**Bibliography** **289**

**Index** **321**