

Model-based Clustering and Classification for Data Science

Cluster analysis consists of methods for finding groups in data automatically. Most methods have been heuristic and leave open such central questions as: How many clusters are there? Which clustering method should I use? How should I handle outliers? Classification involves assigning new observations to groups given previously classified observations, and also has open questions about parameter tuning, robustness and uncertainty assessment. This book frames cluster analysis and classification in terms of statistical models, thus yielding principled estimation, testing and prediction methods, and soundly-based answers to the central questions. It develops the basic ideas of model-based clustering and classification in an accessible but rigorous way, using extensive real-world data examples and providing R code for many methods, and describes modern developments for high-dimensional data and for networks. It explains recent methodological advances, such as Bayesian regularization methods, non-Gaussian model-based clustering, cluster merging, variable selection, semi-supervised classification, robust classification, clustering of functional data, text and images, and co-clustering. Written for advanced undergraduates and beginning graduate students in data science, as well as researchers and practitioners, it assumes basic knowledge of multivariate calculus, linear algebra, probability and statistics.

CHARLES BOUVEYRON is Professor of Statistics at Université Côte d'Azur and the Chair of Excellence in Data Science at Inria Sophia-Antipolis. He has published extensively on model-based clustering, particularly for networks and high-dimensional data.

GILLES CELEUX is Director of Research Emeritus at Inria Saclay Île-de-France. He is one of the founding researchers in model-based clustering, having published extensively in the area for 35 years.

T. BRENDAN MURPHY is Professor of Statistics at University College Dublin. His research interests include model-based clustering, classification, network modeling and latent variable modeling.

ADRIAN E. RAFTERY is Professor of Statistics and Sociology at University of Washington, Seattle. He was one of the founding researchers in model-based clustering, having published in the area since 1984.

CAMBRIDGE SERIES IN STATISTICAL AND
 PROBABILISTIC MATHEMATICS

Editorial Board

- Z. Ghahramani (Department of Engineering, University of Cambridge)
 R. Gill (Mathematical Institute, Leiden University)
 F. P. Kelly (Department of Pure Mathematics and Mathematical Statistics,
 University of Cambridge)
 B. D. Ripley (Department of Statistics, University of Oxford)
 S. Ross (Department of Industrial and Systems Engineering,
 University of Southern California)
 M. Stein (Department of Statistics, University of Chicago)

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
 Recent titles include the following:

30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition)*, by Rick Durrett
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability (Second Edition)*, by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
39. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, by Jianfeng Yao, Shurong Zheng and Zhidong Bai
40. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, by Evarist Giné and Richard Nickl
41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks (Volume 1)*, by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vlasos Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke
47. *High-Dimensional Probability*, by Roman Vershynin
48. *High-Dimensional Statistics*, by Martin J. Wainwright
49. *Probability: Theory and Examples (Fifth Edition)*, by Rick Durrett

Model-Based Clustering and Classification for Data Science

With Applications in R

Charles Bouveyron
Université Côte d'Azur

Gilles Celeux
Inria Saclay Île-de-France

T. Brendan Murphy
University College Dublin

Adrian E. Raftery
University of Washington





Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University’s mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108494205
DOI: 10.1017/9781108644181

© Charles Bouveyron, Gilles Celeux, T. Brendan Murphy and Adrian E. Raftery 2019

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

First published 2019

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data

Names: Bouveyron, Charles, 1979– author. | Celeux, Gilles, author. |
Murphy, T. Brendan, 1972– author. | Raftery, Adrian E., author.
Title: Model-based clustering and classification for data science : with applications in R /
Charles Bouveyron, Université Côte d’Azur, Gilles Celeux, Inria Saclay Île-de-France,
T. Brendan Murphy, University College Dublin, Adrian E. Raftery, University of Washington.
Description: Cambridge ; New York, NY : Cambridge University Press, 2019. |
Series: Cambridge series in statistical and probabilistic mathematics |
Includes bibliographical references and index.
Identifiers: LCCN 2019014257 | ISBN 9781108494205 (hardback)
Subjects: LCSH: Cluster analysis. | Mathematical statistics. |
Statistics–Classification. | R (Computer program language)
Classification: LCC QA278.55 .M63 2019 | DDC 519.5/3–dc23
LC record available at <https://lccn.loc.gov/2019014257>

ISBN 978-1-108-49420-5 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

To Nathalie, Alexis, Romain and Nathan
Charles

To Mailys and Maya
Gilles

To Trish, Áine and Emer
Brendan

To Hana, Isolde and Finn
Adrian

Contents

	Page
<i>Preface</i>	xv
1 Introduction	1
1.1 Cluster Analysis	1
1.2 Classification	4
1.3 Examples	7
1.4 Software	12
1.5 Organization of the Book	13
1.6 Bibliographic Notes	14
2 Model-based Clustering: Basic Ideas	15
2.1 Finite Mixture Models	15
2.2 Geometrically Constrained Multivariate Normal Mixture Models	18
2.3 Estimation by Maximum Likelihood	23
2.4 Initializing the EM Algorithm	31
2.5 Examples with Known Number of Clusters	39
2.6 Choosing the Number of Clusters and the Clustering Model	46
2.7 Illustrative Analyses	60
2.8 Who Invented Model-based Clustering?	71
2.9 Bibliographic Notes	75
3 Dealing with Difficulties	79
3.1 Outliers	79
3.2 Dealing with Degeneracies: Bayesian Regularization	92
3.3 Non-Gaussian Mixture Components and Merging	97
3.4 Bibliographic Notes	105
4 Model-based Classification	109
4.1 Classification in the Probabilistic Framework	109
4.2 Parameter Estimation	113
4.3 Parsimonious Classification Models	114
4.4 Multinomial Classification	119
4.5 Variable Selection	124
4.6 Mixture Discriminant Analysis	126
4.7 Model Assessment and Selection	127

viii	<i>Contents</i>	
5	Semi-supervised Clustering and Classification	134
5.1	Semi-supervised Classification	134
5.2	Semi-supervised Clustering	141
5.3	Supervised Classification with Uncertain Labels	144
5.4	Novelty Detection: Supervised Classification with Unobserved Classes	154
5.5	Bibliographic Notes	160
6	Discrete Data Clustering	163
6.1	Example Data	163
6.2	The Latent Class Model for Categorical Data	165
6.3	Model-based Clustering for Ordinal and Mixed Type Data	185
6.4	Model-based Clustering of Count Data	190
6.5	Bibliographic Notes	197
7	Variable Selection	199
7.1	Continuous Variable Selection for Model-based Clustering	199
7.2	Continuous Variable Regularization for Model-based Clustering	208
7.3	Continuous Variable Selection for Model-based Classification	210
7.4	Categorical Variable Selection Methods for Model-based Clustering	211
7.5	Bibliographic Notes	215
8	High-dimensional Data	217
8.1	From Multivariate to High-dimensional Data	217
8.2	The Curse of Dimensionality	221
8.3	Earlier Approaches for Dealing with High-dimensional Data	227
8.4	Subspace Methods for Clustering and Classification	238
8.5	Bibliographic Notes	257
9	Non-Gaussian Model-based Clustering	259
9.1	Multivariate <i>t</i> -Distribution	259
9.2	Skew-normal Distribution	267
9.3	Skew- <i>t</i> Distribution	270
9.4	Box–Cox Transformed Mixtures	278
9.5	Generalized Hyperbolic Distribution	282
9.6	Example: Old Faithful Data	285
9.7	Example: Flow Cytometry	287
9.8	Bibliographic Notes	288
10	Network Data	292
10.1	Introduction	292
10.2	Example Data	294
10.3	Stochastic Block Model	298
10.4	Mixed Membership Stochastic Block Model	304
10.5	Latent Space Models	312
10.6	Stochastic Topic Block Model	320
10.7	Bibliographic Notes	329

<i>Contents</i>	ix
11 Model-based Clustering with Covariates	331
11.1 Examples	331
11.2 Mixture of Experts Model	333
11.3 Model Assessment	339
11.4 Software	339
11.5 Results	340
11.6 Discussion	348
11.7 Bibliographic Notes	349
12 Other Topics	351
12.1 Model-based Clustering of Functional Data	351
12.2 Model-based Clustering of Texts	363
12.3 Model-based Clustering for Image Analysis	368
12.4 Model-based Co-clustering	373
12.5 Bibliographic Notes	382
<i>List of R Packages</i>	384
<i>Bibliography</i>	386
<i>Author Index</i>	415
<i>Subject Index</i>	423

Expanded Contents

	Page
<i>Preface</i>	xv
1 Introduction	1
1.1 Cluster Analysis	1
1.1.1 From Grouping to Clustering	1
1.1.2 Model-based Clustering	3
1.2 Classification	4
1.2.1 From Taxonomy to Machine Learning	4
1.2.2 Model-based Discriminant Analysis	6
1.3 Examples	7
1.4 Software	12
1.5 Organization of the Book	13
1.6 Bibliographic Notes	14
2 Model-based Clustering: Basic Ideas	15
2.1 Finite Mixture Models	15
2.2 Geometrically Constrained Multivariate Normal Mixture Models	18
2.3 Estimation by Maximum Likelihood	23
2.4 Initializing the EM Algorithm	31
2.4.1 Initialization by Hierarchical Model-based Clustering	33
2.4.2 Initialization Using the smallEM Strategy	36
2.5 Examples with Known Number of Clusters	39
2.6 Choosing the Number of Clusters and the Clustering Model	46
2.7 Illustrative Analyses	60
2.7.1 Wine Varieties	60
2.7.2 Craniometric Analysis	65
2.8 Who Invented Model-based Clustering?	71
2.9 Bibliographic Notes	75
3 Dealing with Difficulties	79
3.1 Outliers	79
3.1.1 Outliers in Model-based Clustering	79
3.1.2 Mixture Modeling with a Uniform Component for Outliers	81
3.1.3 Trimming Data with tclust	88
3.2 Dealing with Degeneracies: Bayesian Regularization	92
3.3 Non-Gaussian Mixture Components and Merging	97
3.4 Bibliographic Notes	105

	<i>Expanded Contents</i>	xi
4	Model-based Classification	109
4.1	Classification in the Probabilistic Framework	109
4.1.1	Generative or Predictive Approach	110
4.1.2	An Introductory Example	111
4.2	Parameter Estimation	113
4.3	Parsimonious Classification Models	114
4.3.1	Gaussian Classification with EDDA	114
4.3.2	Regularized Discriminant Analysis	115
4.4	Multinomial Classification	119
4.4.1	The Conditional Independence Model	119
4.4.2	An Illustration	123
4.5	Variable Selection	124
4.6	Mixture Discriminant Analysis	126
4.7	Model Assessment and Selection	127
4.7.1	The Cross-validated Error Rate	129
4.7.2	Model Selection and Assessing the Error Rate	131
4.7.3	Penalized Log-likelihood Criteria	133
5	Semi-supervised Clustering and Classification	134
5.1	Semi-supervised Classification	134
5.1.1	Estimating the Model Parameters through the EM Algorithm	135
5.1.2	A First Experimental Comparison	136
5.1.3	Model Selection Criteria for Semi-supervised Classification	138
5.2	Semi-supervised Clustering	141
5.2.1	Incorporating Must-link Constraints	143
5.2.2	Incorporating Cannot-link Constraints	144
5.3	Supervised Classification with Uncertain Labels	144
5.3.1	The Label Noise Problem	145
5.3.2	A Model-based Approach for the Binary Case	146
5.3.3	A Model-based Approach for the Multi-class Case	150
5.4	Novelty Detection: Supervised Classification with Unobserved Classes	154
5.4.1	A Transductive Model-based Approach	155
5.4.2	An Inductive Model-based Approach	157
5.5	Bibliographic Notes	160
6	Discrete Data Clustering	163
6.1	Example Data	163
6.2	The Latent Class Model for Categorical Data	165
6.2.1	Maximum Likelihood Estimation	167
6.2.2	Parsimonious Latent Class Models	169
6.2.3	The Latent Class Model as a Cluster Analysis Tool	171
6.2.4	Model Selection	172
6.2.5	Illustration on the Carcinoma Data Set	174
6.2.6	Illustration on the Credit Data Set	178
6.2.7	Bayesian Inference	180
6.3	Model-based Clustering for Ordinal and Mixed Type Data	185
6.3.1	Ordinal Data	185
6.3.2	Mixed Data	186
6.3.3	The ClustMD Model	186

xii	<i>Expanded Contents</i>	
	6.3.4 Illustration of ClustMD: Prostate Cancer Data	187
6.4	Model-based Clustering of Count Data	190
	6.4.1 Poisson Mixture Model	191
	6.4.2 Illustration: Vélib Data Set	194
6.5	Bibliographic Notes	197
7	Variable Selection	199
7.1	Continuous Variable Selection for Model-based Clustering	199
	7.1.1 Clustering and Noisy Variables Approach	200
	7.1.2 Clustering, Redundant and Noisy Variables Approach	200
	7.1.3 Numerical Experiments	203
7.2	Continuous Variable Regularization for Model-based Clustering	208
	7.2.1 Combining Regularization and Variable Selection	209
7.3	Continuous Variable Selection for Model-based Classification	210
7.4	Categorical Variable Selection Methods for Model-based Clustering	211
	7.4.1 Stepwise Procedures	212
	7.4.2 A Bayesian Procedure	212
	7.4.3 An Illustration	214
7.5	Bibliographic Notes	215
8	High-dimensional Data	217
8.1	From Multivariate to High-dimensional Data	217
8.2	The Curse of Dimensionality	221
	8.2.1 The Curse of Dimensionality in Model-based Clustering and Classification	223
	8.2.2 The Blessing of Dimensionality in Model-based Clustering and Classification	225
8.3	Earlier Approaches for Dealing with High-dimensional Data	227
	8.3.1 Unsupervised Dimension Reduction	228
	8.3.2 The Dangers of Unsupervised Dimension Reduction	230
	8.3.3 Supervised Dimension Reduction for Classification	231
	8.3.4 Regularization	236
	8.3.5 Constrained Models	237
8.4	Subspace Methods for Clustering and Classification	238
	8.4.1 Mixture of Factor Analyzers (MFA)	238
	8.4.2 Extensions of the MFA Model	241
	8.4.3 Parsimonious Gaussian Mixture Models (PGMM)	244
	8.4.4 Mixture of High-dimensional GMMs (HD-GMM)	247
	8.4.5 The Discriminative Latent Mixture (DLM) Models	251
	8.4.6 Variable Selection by Penalization of the Loadings	254
8.5	Bibliographic Notes	257
9	Non-Gaussian Model-based Clustering	259
9.1	Multivariate t -Distribution	259
9.2	Skew-normal Distribution	267
9.3	Skew- t Distribution	270
	9.3.1 Restricted Skew- t Distribution	273
	9.3.2 Unrestricted Skew- t Distribution	275
9.4	Box-Cox Transformed Mixtures	278
9.5	Generalized Hyperbolic Distribution	282
9.6	Example: Old Faithful Data	285

	<i>Expanded Contents</i>	xiii
9.7	Example: Flow Cytometry	287
9.8	Bibliographic Notes	288
10	Network Data	292
10.1	Introduction	292
10.2	Example Data	294
10.2.1	Sampson’s Monk Data	295
10.2.2	Zachary’s Karate Club	296
10.2.3	AIDS Blogs	296
10.2.4	French Political Blogs	297
10.2.5	Lazega Lawyers	298
10.3	Stochastic Block Model	298
10.3.1	Inference	299
10.3.2	Application	301
10.4	Mixed Membership Stochastic Block Model	304
10.4.1	Inference	305
10.4.2	Application	306
10.5	Latent Space Models	312
10.5.1	The Distance Model and the Projection Model	313
10.5.2	The Latent Position Cluster Model	314
10.5.3	The Sender and Receiver Random Effects	315
10.5.4	The Mixture of Experts Latent Position Cluster Model	315
10.5.5	Inference	316
10.5.6	Application	317
10.6	Stochastic Topic Block Model	320
10.6.1	Context and Notation	322
10.6.2	The STBM Model	323
10.6.3	Links with Other Models and Inference	326
10.6.4	Application: Enron E-mail Network	326
10.7	Bibliographic Notes	329
11	Model-based Clustering with Covariates	331
11.1	Examples	331
11.1.1	CO ₂ and Gross National Product	331
11.1.2	Australian Institute of Sport (AIS)	331
11.1.3	Italian Wine	332
11.2	Mixture of Experts Model	333
11.2.1	Inference	337
11.3	Model Assessment	339
11.4	Software	339
11.4.1	flexmix	339
11.4.2	mixtools	340
11.4.3	MoEClust	340
11.4.4	Other	340
11.5	Results	340
11.5.1	CO ₂ and GNP Data	340
11.5.2	Australian Institute of Sport	341
11.5.3	Italian Wine	343
11.6	Discussion	348

xiv	<i>Expanded Contents</i>	
11.7	Bibliographic Notes	349
12	Other Topics	351
12.1	Model-based Clustering of Functional Data	351
12.1.1	Model-based Approaches for Functional Clustering	353
12.1.2	The fclust Method	354
12.1.3	The funFEM Method	356
12.1.4	The funHDDC Method for Multivariate Functional Data	358
12.2	Model-based Clustering of Texts	363
12.2.1	Statistical Models for Texts	364
12.2.2	Latent Dirichlet Allocation	365
12.2.3	Application to Text Clustering	366
12.3	Model-based Clustering for Image Analysis	368
12.3.1	Image Segmentation	368
12.3.2	Image Denoising	370
12.3.3	Inpainting Damaged Images	372
12.4	Model-based Co-clustering	373
12.4.1	The Latent Block Model	375
12.4.2	Estimating LBM Parameters	376
12.4.3	Model Selection	379
12.4.4	An Illustration	380
12.5	Bibliographic Notes	382
	<i>List of R Packages</i>	384
	<i>Bibliography</i>	386
	<i>Author Index</i>	415
	<i>Subject Index</i>	423

Preface

About this book

The century that is ours is shaping up to be the century of the data revolution. Our numerical world is creating masses of data every day and the volume of generated data is estimated to be doubling every two years. This wealth of available data offers hope for exploitation that may lead to great advances in areas such as health, science, transportation and defense. However, manipulating, analyzing and extracting information from those data is made difficult by the volume and nature (high-dimensional data, networks, time series, etc) of modern data.

Within the broad field of statistical and machine learning, model-based techniques for clustering and classification have a central position for anyone interested in exploiting those data. This textbook focuses on the recent developments in model-based clustering and classification while providing a comprehensive introduction to the field. It is aimed at advanced undergraduates, graduates or first-year Ph.D. students in data science, as well as researchers and practitioners. It assumes no previous knowledge of clustering and classification concepts. A basic knowledge of multivariate calculus, linear algebra and probability and statistics is needed.

The book is supported by extensive examples on data, with 72 listings of code mobilizing more than 30 software packages, that can be run by the reader. The chosen language for codes is the R software, which is one of the most popular languages for data science. It is an excellent tool for data science since the most recent statistical learning techniques are provided on the R platform (named CRAN). Using R is probably the best way to be directly connected to current research in statistics and data science through the packages provided by researchers.

The book is accompanied by a dedicated R package (the `MBCbook` package) that can be directly downloaded from CRAN within the R software or at the following address: <https://cran.r-project.org/package=MBCbook>. We also encourage the reader to visit the book website for the latest information: <http://math.unice.fr/~cbouveyr/MBCbook/>.

This book could be used as one of the texts for a graduate or advanced undergraduate course in multivariate analysis or machine learning. Chapters 1 and 2, and optionally a selection of later chapters, could be used for this purpose. The book as a whole could also be used as the main text for a one-quarter or

one-semester course in cluster analysis or unsupervised learning, focusing on the model-based approach.

Acknowledgements

This book is a truly collaborative effort, and the four authors have contributed equally. Each of us has contributed to each of the chapters.

We would like to thank Chris Fraley for initially developing the `mclust` software and later R package, starting in 1991. This software was of extraordinary quality from the beginning, and without it this whole field would never have developed as it did. Luca Scrucca took over the package in 2007, and has enhanced it in many ways, so we also owe a lot to his work. We would also like to thank the developers and maintainers of `Rmixmod` software: Florent Langrognet, Rémi Lebrete, Christian Poli, Serge Iovleff, Anwuli Echenim and Benjamin Auder.

The authors would also like to thank the participants in the Working Group on Model-based Clustering, which has been gathering every year in the third week of July since 1994, first in Seattle and then since 2007 in different venues around Europe and North America. This is an extraordinary group of people from many countries, whose energy, interactions and intellectual generosity have inspired us every year and driven the field forward. The book owes a great deal to their insights.

Charles Bouveyron would like to thank in particular Stéphane Girard, Julien Jacques and Pierre Latouche, for very fruitful and friendly collaborations. Charles Bouveyron also thanks his coauthors on this topic for all the enjoyable collaborations: Laurent Bergé, Camille Brunet-Saumard, Etienne Côme, Marco Corneli, Julie Delon, Mathieu Fauvel, Antoine Houdard, Pierre-Alexandre Mattei, Cordelia Schmid, Amandine Schmutz and Rawya Zreik. He would like also to warmly thank his family, Nathalie, Alexis, Romain and Nathan, for their love and everyday support in the writing of this book.

Gilles Celeux would like to thank his old and dear friends Jean Diebolt and Gérard Govaert for the long and intensive collaborations. He also thanks his coauthors in the area Jean-Patrick Baudry, Halima Bensmail, Christophe Biernacki, Guillaume Bouchard, Vincent Brault, Stéphane Chrétien, Florence Forbes, Raphaël Gottardo, Christine Keribin, Jean-Michel Marin, Marie-Laure Martin-Magniette, Cathy Maugis-Rabusseau, Abdallah Mkhadri, Nathalie Peyrard, Andrea Rau, Christian P. Robert, Gilda Soromenho and Vincent Vandewalle for nice and fruitful collaborations. Finally, he would like to thank Mailys and Maya for their love.

Brendan Murphy would like to thank John Hartigan for introducing him to clustering. He would like to thank Claire Gormley, Paul McNicholas, Luca Scrucca and Michael Fop with whom he has collaborated extensively on model-based clustering projects over a number of years. He also would like to thank his students and coauthors for enjoyable collaborations on a wide range of model-based clustering and classification projects: Marco Alfò, Francesco Bartolucci, Nema Dean, Silvia D'Angelo, Gerard Downey, Bailey Fosdick, Nial Friel, Marie

Preface

xvii

Galligan, Isabella Gollini, Sen Hu, Neil Hurley, Dimitris Karlis, Donal Martin, Tyler McCormick, Aaron McDaid, Damien McParland, Keefe Murphy, Tin Lok James Ng, Adrian O'Hagan, Niamh Russell, Michael Salter-Townshend, Lucy Small, Deirdre Toher, Ted Westling, Arthur White and Jason Wyse. Finally, he would like to thank his family, Trish, Áine and Emer for their love and support.

Adrian Raftery thanks Fionn Murtagh, with whom he first encountered model-based clustering and wrote his first paper in the area in 1984, Chris Fraley for a long and very fruitful collaboration, and Luca Scrucca for another very successful collaboration. He warmly thanks his Ph.D. students who have worked with him on model-based clustering, namely Jeff Banfield, Russ Steele, Raphael Gottardo, Nema Dean, Derek Stanford and William Chad Young for their collaboration and all that he learned from them. He also thanks his other coauthors in the area, Jogesh Babu, Jean-Patrick Baudry, Halima Bensmail, Roger Bumgarner, Simon Byers, Jon Campbell, Abhijit Dasgupta, Mary Emond, Eric Feigelson, Florence Forbes, Diane Georgian-Smith, Ken Lo, Alejandro Murua, Nathalie Peyrard, Christian Robert, Larry Ruzzo, Jean-Luc Starck, Ka Yee Yeung, Naisyin Wang and Ron Wehrens for excellent collaborations.

Raftery would like to thank the Office of Naval Research and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD grants R01 HD054511 and R01 HD070936) for sustained research support without which this work could not have been carried out. He wrote part of the book during a fellowship year at the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford University in 2017–2018, which provided an ideal environment for the sustained thinking needed to complete a project of this kind. Finally he would like to thank his wife, Hana Ševčíková, for her love and support through this project.