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

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 Vladas 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

CAMBRIDGE
UNIVERSITY PRESS

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

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

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

Expanded Contents

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 Lebret, 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 Maïlys 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

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.