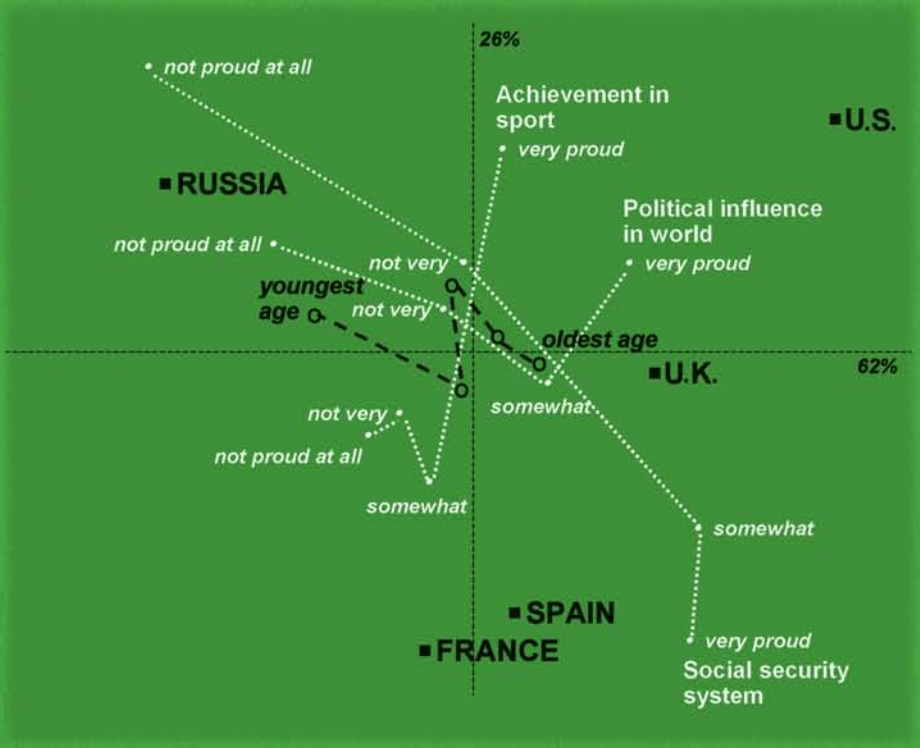
Statistics in the Social and Behavioral Sciences Series

Multiple Correspondence Analysis and Related Methods



Edited by
Michael Greenacre and Jörg Blasius

Statistics in the Social and Behavioral Sciences Series

Multiple Correspondence Analysis and Related Methods

Chapman & Hall/CRC

Statistics in the Social and Behavioral Sciences Series

Aims and scope

Large and complex datasets are becoming prevalent in the social and behavioral sciences and statistical methods are crucial for the analysis and interpretation of such data. This series aims to capture new developments in statistical methodology with particular relevance to applications in the social and behavioral sciences. It seeks to promote appropriate use of statistical methods in these applied sciences by publishing a broad range of reference works, textbooks and handbooks.

The scope of the series is wide, including applications of statistical methodology in sociology, psychology, economics, education, marketing research, political science, criminology, public policy, demography, survey methodology and official statistics. The titles included in the series are designed to appeal to applied statisticians, as well as students, researchers and practitioners from the above disciplines. The inclusion of real examples and case studies is therefore essential.

Proposals for the series should be submitted directly to:

Chapman & Hall/CRC
Taylor and Francis Group
Informa
24-25 Blades Court
Deodar Road
London SW15 2NU, UK

Statistics in the Social and Behavioral Sciences Series

Multiple Correspondence Analysis and Related Methods

Edited by

Michael Greenacre and Jörg Blasius

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-628-5 (Hardcover)
International Standard Book Number-13: 978-1-58488-628-0 (Hardcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

In fond memory of Ruben Gabriel

Kuno Ruben Gabriel, 1929–2003

Preface

This book is aimed at a wide spectrum of researchers and students who are concerned with the analysis of tabular data, chiefly data measured on categorical scales. In other words, all social scientists who work with empirical data will benefit from this book, as well as most environmental scientists, market researchers, psychologists, and archaeologists, to name but a few, where applications of correspondence analysis already abound.

The idea for this book grew out of the international conference on correspondence analysis and related methods (CARME 2003), held at the Universitat Pompeu Fabra in Barcelona from 29 June to 2 July 2003. A total of 88 scientific papers were delivered at the conference, attended by 175 researchers from 18 countries, testifying to the ever-increasing interest in this multipurpose, multicultural, and multidisciplinary statistical method. The extension of correspondence analysis to more than two variables, called multiple correspondence analysis (MCA), and various methods related to it were so prominent at this meeting that we decided to embark on this book project. The idea was to gather experts in the field and to assemble and edit a single text that encompassed the subject, especially the different approaches taken by researchers from different statistical "schools."

For the record, this is the third time we have embarked on a project like this. The first book, *Correspondence Analysis in the Social Sciences* (Greenacre and Blasius 1994), was edited after the first conference organized in Cologne in 1991 and contained various methodological and applied chapters, the latter written mostly by sociologists, in an attempt to show the usefulness of correspondence analysis in exploring social science data. The second book, *Visualization of Categorical Data* (Blasius and Greenacre 1998), which was edited after the second conference organized in Cologne in 1995, broadened the content to all methods that have as their goal the graphical display of categorical data. Once again, both statisticians and social science researchers contributed to the book,

which has been as successful as the first one in communicating a new field of multidisciplinary research to a wider audience. The present book, *Multiple Correspondence Analysis and Related Methods*, carries on this tradition, giving a state-of-the-art description of this new field of research in a self-contained textbook format.

A total of 40 authors have contributed to this book, and the editing process was by no means a simple task. As in the two previous books, our objective has been to produce a unified text, with unified presentation and notation. Cross-referencing between chapters was introduced, and a common reference list and index was established. In addition, we have included several introductory chapters so that readers with little experience in the field can be gently introduced to the subject. In our selection of chapters, we tried to be inclusive as well as exhaustive—inclusive of the different cultural origins of the subject's development and exhaustive of the methodological and applications fields, covering the whole subject and a wide variety of application areas. Another goal was to make the book as practically oriented as possible and to include details about software and computational aspects; most chapters have a "software note" at the end, and there is an appendix at the end of the book that summarizes the computational steps of the basic method and some related ones.

Before we move on to the acknowledgments, we feel a note of sadness that our dear friend, Ruben Gabriel, so prominent in this area of methodology, passed away a month before the CARME 2003 conference. Ruben was to have been one of the keynote speakers and a contributor to this book, so it was fitting that we agreed to dedicate this book to his memory and include him in this way in our project. To quote from the obituary written by Jack Hall, Ruben's colleague at the Department of Statistics, University of Rochester:

> Ruben Gabriel was born in Germany. His family fled to France in the 1930's and then to Palestine where he was raised on a Kibbutz. He studied at the London School of Economics and at the Hebrew University of Jerusalem, earning a PhD in Demography in 1957. He was on the faculty of the Department of Statistics at Hebrew University for many years, including a term as chair, and joined the University of Rochester as Professor of Statistics in 1975, serving as chair from 1981 to 1989 and retiring in 1997. While at the University, he also served as Professor of Biostatistics in the Medical Center, collaborating on medical research with faculty in many departments.
>
> Ruben had a distinguished statistical research career, with 150 scientific publications, and was honored as a Fellow of the American Statistical

Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. Of special note was his "biplot," a graphical data analytic tool to assist in the understanding of the structure of an array of data—say of several variables on each of several units (e.g., persons)—now widely used in data analysis in many fields.

The CARME 2003 conference and this book would not have been possible without the generous support of the Fundación BBVA in Madrid. In the first instance, we would like to thank this organization and especially its director, Rafael Pardo (who also contributes to the book), for their interest in fostering research on correspondence analysis, both on the theoretical and applied levels. We also thank the other sponsors of the conference: the Spanish Ministry of Science and Technology, grant MCYT2096; IdesCAT (the Catalan Statistical Institute); DURSI (the Department of Universities, Research and Information Society of the Catalan government); and the Faculty of Economic Sciences of the Universitat Pompeu Fabra.

Then, to all our authors: you have been very patient with us, and we hope that we have been patient with you! It was a long process, but we hope that you will appreciate the fruits of your labors and share together in the success of this venture. Our respective institutions—the Department of Economics and Business at Pompeu Fabra University, Barcelona, and the Seminar für Soziologie at the University of Bonn—have given us the academic freedom to dedicate many hours to this task that is of our own choosing, and we thank them for that. We also thank Martin Blankenstein (University of Bonn) for preparing the reference list of this book and Andreas Mühlichen (University of Bonn) for assistance in preparing a large number of figures. For their work on the CARME 2003 conference, we acknowledge the assistance of Anna Cuxart, Clara Riba, Frederic Udina, Robert Diez, and Friederika Priemer, and we thank all our colleagues and family for providing heartfelt support.

Finally, we thank Chapman & Hall and editor Rob Calver in London for supporting the publication of this book, and Mimi Williams, project editor, Florida, for handling the complex task of producing this multi-authored book in such a cooperative and friendly way.

Michael Greenacre and Jörg Blasius

Barcelona and Bonn

About the authors

Elena Abascal Fernández is a professor of statistics at the Public University of Navarra, Pamplona, Spain. Her interests are mainly in multivariate data analysis, descriptive factorial analysis, and classification techniques and their application to marketing research and to the analysis of survey data.

Address: Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadía, E-31006 Pamplona, Spain.

Email: eabascal@unavarra.es.

Carolyn J. Anderson is an associate professor in the Departments of Educational Psychology, Psychology, and Statistics at the University of Illinois at Urbana-Champaign. Her major interests are in the analysis of multivariate categorical data and latent variable models.

Address: University of Illinois, 1310 South Sixth Street, MC-708, Champaign, IL, 61820.

Email: cja@uiuc.edu.

Web address: http://www.psych.uiuc.edu/~canderso.

Matthias C. Angermeyer is professor and, since 1995, head of the Department of Psychiatry, University of Leipzig. His main research interests are epidemiology of mental disorders, public attitudes toward people with mental illness, and evaluation of mental health services.

Address: University of Leipzig, Department of Psychiatry, Johannisallee 20/1, 04317 Leipzig, Germany.

Mónica Bécue-Bertaut is a professor in the Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Spain. Her interests are primarily in textual statistical methods and text mining tools and their application to open-ended responses, especially to multilingual responses. She is coauthor (with Ludovic Lebart and André Salem) of the book *Análisis Estadístico de Textos*.

Address: Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Edifici FME, c/ Pau Gargallo, 5, E-08028 Barcelona, Spain.

Email: Monica.Becue@upc.es.

Antonio Blázquez-Zaballos is an assistant professor in the Departamento de Estadística, Universidad de Salamanca, Spain. His main research interest is in multivariate data analysis, particularly in the generalizations of biplots to mixed data types. Recent work includes contributions to the detection of genotype-environment interaction.

Address: Departamento de Estadística, Universidad de Salamanca, C/ Espejo s/n, 37007, Salamanca, Spain.

E-mail: abz@usal.es.

Jörg Blasius is a professor of sociology at the Institute for Political Science and Sociology, Bonn, Germany. His research interests are mainly in exploratory data analysis, data collection methods, sociology of lifestyles, and urban sociology. Together with Michael Greenacre, he has coedited two books on correspondence analysis and visualizing of categorical data.

Address: University of Bonn, Institute for Political Science and Sociology, Seminar of Sociology, Lennéstr. 27, 53113 Bonn, Germany.

Email: jblasius@uni-bonn.de.

Web address: http://www.soziologie.uni-bonn.de/blasius1.htm.

Stéphanie Bougeard is a researcher in the veterinary epidemiological team of the French Agency for Food Safety (AFSSA). Her research interests are in factorial methods for qualitative data. This work is a part of her Ph.D. thesis.

Address: Department of Epidemiology, French Agency for Food Safety (AFSSA), Les Croix, BP53, F-22 440 Ploufragan, France.

Email: s.bougeard@afssa.fr.

Henri Caussinus is an emeritus professor, Laboratory of Statistics and Probabilities, Université Paul Sabatier, Toulouse, France. One of his main interests is the use of probabilistic models for exploratory data analysis, with special emphasis on graphical displays and on the choice of their relevant dimension.

Address: Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex 04, France.

Email: Caussinus@cict.fr.

Jan de Leeuw is Distinguished Professor and Chair of Statistics at the University of California at Los Angeles. His interests are mainly in multivariate analysis, optimization, and computational statistics.

Address: Department of Statistics, University of California at Los Angeles, Box 951552, Los Angeles, CA 90095-1664.

E-mail: deleeuw@stat.ucla.edu.

Web address: http://gifi.stat.ucla.edu.

M. Purificación Galindo Villardón is a professor and head in the Departamento de Estadística, Universidad de Salamanca, Spain. Her main research interest is in multivariate data analysis, particularly the application of biplot and correspondence analysis to clinical and environmental data. Recent work includes contributions to three-way nonsymmetrical correspondence analysis.

Address: Departamento de Estadística, Universidad de Salamanca, C/ Espejo s/n, 37007, Salamanca, Spain.

E-mail: pgalindo@usal.es.

Ignacio García Lautre is an associate professor of statistics at the Public University of Navarra, Pamplona, Spain. His interests are mainly in multivariate data analysis, descriptive factorial analysis, classification techniques, and their application to economic data and, more generally, to the analysis of survey data.

Address: Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadía, E-31006 Pamplona, Spain.

Email: nacho@unavarra.es.

Beatriz Goitisolo is a professor of statistics and econometrics at the University of Basque Country, Bilbao, Spain. Her interests are mainly in multivariate data analysis in general and in particular in the analysis of several contingency tables and their applications to social sciences.

Address: Departamento de Economía Aplicada III (Econometría y Estadística), Facultad de Ciencias Económicas y Empresariales, Avda Lehendakari Agirre 83, E-48015 Bilbao, Spain.

Email: Beatriz. Goitisolo@ehu.es.

John C. Gower, formerly head of the Biomathematics Division, Rothamsted Experimental Station, is currently a professor of statistics at the Open University, U.K. His research interests are mainly in exploratory multivariate data analysis and visualization, especially biplot methodology, Procrustes analysis, multidimensional scaling, analysis of asymmetry, and classification methods.

Address: Department of Statistics, Walton Hall, The Open University, Milton Keynes, MK7 6AA, U.K.

Email: j.c.gower@open.ac.uk.

Web address: http://statistics.open.ac.uk/staff/jg1.html.

Michael Greenacre is a professor of statistics in the Department of Economics and Business, Pompeu Fabra University in Barcelona. His research interests are mainly in exploratory analysis and visualization of large data sets, and applications of multivariate analysis in sociology and marine ecology. He has published two books on correspondence analysis and co-edited two more with Jörg Blasius.

Address: Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, E-08005 Barcelona, Spain.

Email: michael@upf.es.

Web address: http://www.econ.upf.es/~michael.

Patrick J.F. Groenen is a professor of statistics at the Econometric Institute, Erasmus University Rotterdam, the Netherlands. His research interests are exploratory multivariate analysis, optimization, visualization, and multidimensional scaling. Together with Ingwer Borg, he has written a textbook on multidimensional scaling.

Address: Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738 DR Rotterdam, the Netherlands.

E-mail: groenen@few.eur.nl.

Mohamed Hanafi is a researcher at the École Nationale des Ingénieurs des Industries Agricoles et Alimentaires (ENITIAA). His interests are primarily in multivariate analysis with applications in sensory analysis and chemometrics.

Address: ENITIAA-INRA Unité de Sensométrie et Chimiométrie, Rue de la Géraudière, BP 82225, F-44322 Nantes Cedex 03, France.

Email: hanafi@enitiaa-nantes.fr.

Willem J. Heiser is a professor and head of the Section of Psychometrics and Research Methodology, Department of Psychology, at Leiden University. His interests are multivariate analysis techniques, multidimensional scaling, optimal scaling, classification methods, and the history of statistics. He has been the president of the Psychometric Society, 2003–2004 and is the present editor of the *Journal of Classification*.

Address: Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

Email: heiser@fsw.leidenuniv.nl.

Heungsun Hwang is an assistant professor of marketing, HEC Montréal, Montréal, Canada. His recent interests include generalizations of growth curve models and correspondence analysis to capture subject heterogeneity.

Address: Department of Marketing, HEC Montréal, 3000 Chemin de la Côte Ste Catherine, Montréal, Québec, H3T 2A7, Canada.

Email: heungsun.hwang@hec.ca.

Alex J. Koning is an assistant professor at the Econometric Institute, Erasmus University Rotterdam, the Netherlands. His research interests include goodness-of-fit tests, statistical quality control, and nonparametric statistics.

Address: Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738 DR Rotterdam, the Netherlands.

E-mail: koning@few.eur.nl.

Pieter M. Kroonenberg occupies the chair "Multivariate Analysis, in particular of three-way data" in the Department of Education and Child Studies, Leiden University, the Netherlands. His major interest is in three-mode analysis in all its facets, but he is also interested in other multivariate data-analytic methods and applying such methods to data from different fields. He is presently completing a book on the practice of three-mode analysis.

Address: Department of Education, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, the Netherlands.

Email: kroonenb@fsw.leidenuniv.nl.

Web address: http://three-mode.leidenuniv.nl.

M. Isabel Landaluce Calvo is a professor of statistics at the University of Burgos, Burgos, Spain. Her interests are mainly in multivariate data analysis, descriptive factorial analysis, and classification techniques and their application to marketing research and to the analysis of survey data.

Address: Departamento de Economía Aplicada, Universidad de Burgos, Plaza Infanta Doña Elena s/n, E-09001 Burgos, Spain.

Email: iland@ubu.es.

Ludovic Lebart is a senior researcher at the Centre National de la Recherche Scientifique and a professor at the Ecole Nationale Supérieure des Télécommunications in Paris. His research interests are the exploratory analysis of qualitative and textual data. He has

coauthored several books on descriptive multivariate statistics, survey methodology, and exploratory analysis of textual data.

Address: Ecole Nationale Supérieure des Télécommunications, 46 rue Barrault, 75013 Paris, France.

Email: lebart@enst.fr.

Web address: http://www.lebart.org.

Herbert Matschinger is, since 1996, a senior researcher in the Department of Psychiatry, University of Leipzig. Previously, he was a researcher in the Department of Psychiatric Sociology at the Central Institute of Mental Health in Mannheim. His main research interests are generalized canonical analysis, mixture modeling, IRT modeling, and analysis of repeated measurements.

Address: University of Leipzig, Department of Psychiatry, Johannisallee 20/1, 04317 Leipzig, Germany.

Email: math@medizin.uni-leipzig.de.

Oleg Nenadić is a research assistant at the Institute for Statistics and Econometrics, University of Göttingen. His interests are mainly in computational statistics, multivariate analysis, and visualization.

Address: Georg-August-Universität Göttingen, Institut für Statistik und Ökonometrie, Platz der Göttinger Sieben 5, 37075 Göttingen, Germany.

Email: onenadi@uni-goettingen.de.

Web address: http:// www.statoek.wiso.uni-goettingen.de/.

Ndèye Niang is an assistant professor of statistics at the Institut d'Informatique d'Entreprise, the engineering school in computer science of CNAM in Paris and a member of the data analysis group of CEDRIC, the computer science research team of CNAM. Her interests are in multivariate analysis and applications to quality control.

Address: Chaire de Statistique Appliquée, CNAM, 292 rue Saint Martin, F-75141 Paris Cedex 03, France.

Email: niang@cnam.fr.

Shizuhiko Nishisato is a professor emeritus, University of Toronto, Canada. He is a former President of the Psychometric Society, a former editor of *Psychometrika*, and a fellow of the American Statistical Association. He coined the term "dual scaling," the subject of his lifelong work.

Address: OISE/University of Toronto, 252 Bloor Street West, Toronto, Ontario, M5S 1V6, Canada.

Email: snishisato@oise.utoronto.ca.

Hicham Noçairi is a Ph.D. student at Ecole Nationale des Ingénieurs des Industries Agricoles et Alimentaires (ENITIAA) in Nantes, France. His interests are primarily in discrimination methods in presence of multicollinearity among predictors, with applications in chemometrics.

Address: ENITIAA-INRA Unité de Sensométrie et Chimiométrie, Rue de la Géraudière, BP 82225, F-44322 Nantes Cedex 03, France.

Email: nocairi@enitiaa-nantes.fr.

Jérôme Pagès is a professor and head of the Laboratory of Applied Mathematics at Agrocampus Rennes in Rennes, France. His interest is primarily in exploratory data analysis, especially in the treatment of multiple tables. In collaboration with Brigitte Escofier, he published a book about simple and multiple factor analysis, which is a classic one in France.

Address: Agrocampus Rennes, 65 rue de Saint-Brieuc, CS 84215, F-35042 Rennes Cedex, France.

Email: jerome.pages@agrocampus-rennes.fr.

Rafael Pardo is director of the Fundación BBVA in Madrid and professor of research at the CSIC (Spanish National Council for Scientific Research). His current areas of research are scientific and environmental culture in late modern societies, social capital, and research methodology.

Address: Fundacion BBVA, Paseo de Recoletos 10, E-28001 Madrid, Spain.

Email: rpardoa@fbbva.es.

El-Mostafa Qannari is a professor at École Nationale des Ingénieurs des Industries Agricoles et Alimentaires (ENITIAA) in Nantes, France. He is also head of a research unit affiliated with INRA (Institut de Recherche Agronomique). His interests are primarily in multivariate analysis with applications in sensory analysis and chemometrics.

Address: ENITIAA-INRA Unité de Sensométrie et Chimiométrie, Rue de la Géraudière, BP 82225, F-44322 Nantes Cedex 03, France.

Email: qannari@enitiaa-nantes.fr.

Henry Rouanet is a guest researcher at the Centre de Recherche Informatique de Paris 5 (CRIP5), Université René Descartes, Paris. His main interests are analysis of variance and Bayesian inference. He has coauthored several books about statistics and geometric data analysis.

Address: UFR Math-Info, Université René Descartes, 45 rue des Saints-Pères, F-75270 Paris Cedex 06, France.
Email: rouanet@math-info.univ-paris5.fr.
Web address: http://www.math-info.univ-paris5.fr/~rouanet.

Anne Ruiz-Gazen is an associate professor, GREMAQ, Université des Sciences Sociales and Laboratory of Statistics and Probabilities, Université Paul Sabatier, Toulouse, France. Her main research interests are robust statistics and multivariate data analysis.

Address: GREMAQ, Université Toulouse I, 21 allée de Brienne, F-31000 Toulouse, France.
Email: ruiz@cict.fr.

Gilbert Saporta is a professor and head of the chair of Applied Statistics at CNAM-Paris (Conservatoire National des Arts et Métiers). He is responsible of the data analysis group of CEDRIC, the computer science research team of CNAM. His interests are in applied multivariate analysis, scoring techniques, and time-dependent data. He has been president of SFdS, the French statistical society.

Address: Chaire de Statistique Appliquée, CNAM, 292 rue Saint Martin, F-75141 Paris Cedex 03, France.
Email: saporta@cnam.fr.
Web address: http://cedric.cnam.fr/~saporta.

Yoshio Takane is a professor of psychology at McGill University, Montréal, Canada. He is a past president of the Psychometric Society. His recent interests are primarily in the development of methods for structured analysis of multivariate data, and artificial neural network simulations.

Address: Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montréal, Québec, H3A 1B1, Canada.
Email: takane@ takane2.psych.mcgill.ca.
Web address: http://takane.brinkster.net/Yoshio/.

Victor Thiessen is a professor in the Department of Sociology and Social Anthropology, Dalhousie University in Halifax, Nova Scotia. He has published articles on survey methodology as well as on the educational and occupational aspirations of youth. His current investigations focus on (a) the various pathways along which young Canadians navigate their way from schooling to employment and (b) the effects of information and communication technologies in schools and at home on educational attainments.

Address: Department. of Sociology and Social Anthropology, Dalhousie University, Halifax, NS, B3H 4P9, Canada.

Email: victor. thiessen@dal.ca.

Anna Torres-Lacomba is an associate professor of marketing research at the Universidad Carlos III, Madrid, Spain. Her interests are mainly in multivariate data analysis applied to marketing problems, specially the use of correspondence analysis for measuring preferences and brand image.

Address: Departament d'Economia i Empresa, Pompeu Fabra University, Ramon Trias Fargas 25–27, E-08005 Barcelona, Spain.

Email: anna.torres@upf.edu

Wijbrandt van Schuur is an associate professor in the Sociology Department of the University of Groningen and a member of the sociological graduate school ICS. His research interests are the development and application of measurement models, especially nonparametric IRT models such as the Mokken models, unidimensional unfolding, and the circumplex. His substantive interests are in the areas of political and cultural sociology.

Address: Department of Sociology, University of Groningen, Grote Rozenstraat 31, NL-9712 TG Groningen, the Netherlands.

Email: h.van.schuur@ppsw.rug.nl.

José L. Vicente-Villardón is a professor in the Departamento de Estadística, Universidad de Salamanca, Spain. His main research interest is in multivariate data analysis, especially biplot, related methods, and integration of several data matrices. Recent work is related to biplots based on generalized linear models and three-way generalizations of canonical correspondence analysis.

Address: Departamento de Estadística, Universidad de Salamanca, C/ Espejo s/n, 37007, Salamanca, Spain.

Email: villardon@usal.es.

Matthijs J. Warrens is a Ph.D. student in the Section of Psychometrics and Research Methodology, Department of Psychology, Leiden University. His interests are methods for optimal scaling and item-response theory.

Address: Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, the Netherlands.

Email: Warrens@fsw.leidenuniv.nl.

Amaya Zárraga is a professor of statistics and data analysis at the University of Basque Country, Bilbao, Spain. Her interests are mainly in multivariate data analysis and in particular the analysis of several contingency tables and their applications to social sciences.

Address: Departamento de Economía Aplicada III (Econometría y Estadística), Facultad de Ciencias Económicas y Empresariales, Avda Lehendakari Agirre 83, E-48015 Bilbao, Spain.

Email: amaya.zarraga@ehu.es.

Table of Contents

Section I Introduction..... 1

Chapter 1 Correspondence Analysis and Related Methods in Practice..... 3
Jörg Blasius and Michael Greenacre

Chapter 2 From Simple to Multiple Correspondence Analysis..... 41
Michael Greenacre

Chapter 3 Divided by a Common Language: Analyzing and Visualizing Two-Way Arrays..... 77
John C. Gower

Chapter 4 Nonlinear Principal Component Analysis and Related Techniques..... 107
Jan de Leeuw

Section II Multiple Correspondence Analysis..... 135

Chapter 5 The Geometric Analysis of Structured Individuals × Variables Tables..... 137
Henry Rouanet

Chapter 6 Correlational Structure of Multiple-Choice Data as Viewed from Dual Scaling..... 161
Shizuhiko Nishisato

Chapter 7 Validation Techniques in Multiple Correspondence Analysis 179
Ludovic Lebart

Chapter 8 Multiple Correspondence Analysis of Subsets of Response Categories 197
Michael Greenacre and Rafael Pardo

Chapter 9 Scaling Unidimensional Models with Multiple Correspondence Analysis 219
Matthijs J. Warrens and Willem J. Heiser

Chapter 10 The Unfolding Fallacy Unveiled: Visualizing Structures of Dichotomous Unidimensional Item–Response–Theory Data by Multiple Correspondence Analysis 237
Wijbrandt van Schuur and Jörg Blasius

Chapter 11 Regularized Multiple Correspondence Analysis 259
Yoshio Takane and Heungsun Hwang

Section III Analysis of Sets of Tables 281

Chapter 12 The Evaluation of "Don't Know" Responses by Generalized Canonical Analysis 283
Herbert Matschinger and Matthias C. Angermeyer

Chapter 13 Multiple Factor Analysis for Contingency Tables 299
Jérôme Pagès and Mónica Bécue-Bertaut

Chapter 14 Simultaneous Analysis: A Joint Study of Several Contingency Tables with Different Margins 327
Amaya Zárraga and Beatriz Goitisolo

Chapter 15 Multiple Factor Analysis of Mixed Tables of Metric and Categorical Data 351
Elena Abascal, Ignacio García Lautre, and M. Isabel Landaluce

Section IV MCA and Classification 369

Chapter 16 Correspondence Analysis and Classification 371
Gilbert Saporta and Ndèye Niang

Chapter 17 Multiblock Canonical Correlation Analysis for Categorical Variables: Application to Epidemiological Data 393
Stéphanie Bougeard, Mohamed Hanafi, Hicham Noçairi, and El-Mostafa Qannari

Chapter 18 Projection-Pursuit Approach for Categorical Data 405
Henri Caussinus and Anne Ruiz-Gazen

Section V Related Methods 419

Chapter 19 Correspondence Analysis and Categorical Conjoint Measurement 421
Anna Torres-Lacomba

Chapter 20 A Three-Step Approach to Assessing the Behavior of Survey Items in Cross-National Research 433
Jörg Blasius and Victor Thiessen

Chapter 21 Additive and Multiplicative Models for Three-Way Contingency Tables: Darroch (1974) Revisited 455
Pieter M. Kroonenberg and Carolyn J. Anderson

Chapter 22 A New Model for Visualizing Interactions in Analysis of Variance 487
Patrick J.F. Groenen and Alex J. Koning

Chapter 23 Logistic Biplots 503
José L. Vicente-Villardón, M. Purificación Galindo-Villardón, and Antonio Blázquez-Zaballos

Appendix Computation of Multiple Correspondence Analysis, with Code in R 523
Oleg Nenadić and Michael Greenacre

References 553

Index 575

SECTION I

Introduction

CHAPTER 1

Correspondence Analysis and Related Methods in Practice

Jörg Blasius and Michael Greenacre

CONTENTS

1.1 Introduction 4
1.2 A simple example 7
1.3 Basic method 12
1.4 Concepts of correspondence analysis 14
1.4.1 Profiles, average profiles, and masses 14
1.4.2 Chi-square statistic and total inertia 16
1.4.3 Chi-square distances 18
1.4.4 Reduction of dimensionality 19
1.4.5 Contributions to inertia 19
1.4.6 Reconstruction of the data 21
1.5 Stacked tables 21
1.6 Multiple correspondence analysis 27
1.7 Categorical principal component analysis 30
1.8 Active and supplementary variables 31
1.9 Multiway data 32
1.10 Content of the book 33
1.10.1 Introduction 33
1.10.2 Multiple correspondence analysis 34
1.10.3 Analysis of sets of tables 36
1.10.4 Multiple correspondence analysis and classification 38
1.10.5 Related methods 39

1.1 Introduction

Correspondence analysis as we know it today has come a long way in the 30 years since the publication of Benzécri's seminal work, *Analyse des Données* (Benzécri et al. 1973) and, shortly thereafter, Hill's paper on applied statistics, "Correspondence analysis: a neglected multivariate method" (Hill 1974), which drew the English-speaking world's attention to the existence of this technique. In a bibliography of publications on correspondence analysis, Beh (2004) documents an almost exponential increase in articles on the subject (Figure 1.1). However, this bibliography focused mostly on methodological journals, and it does not include the explosion of applications of correspondence analysis in fields as diverse as archaeology, linguistics, marketing research, psychology, sociology, education, geology, ecology, and medicine—indeed, in all the physical, social, human, and biological sciences.

Correspondence analysis (CA) is an exploratory multivariate technique for the graphical and numerical analysis of almost any data matrix with nonnegative entries, but it principally involves tables of frequencies or counts. It can be extended to analyze presence/absence data, rankings and preferences, paired comparison data, multiresponse tables, multiway tables, and square transition tables, among others. Because it is oriented toward categorical data, it can be used to analyze almost any type of tabular data after suitable data transformation, or recoding, as exemplified by a recent book by Murtagh (2005).

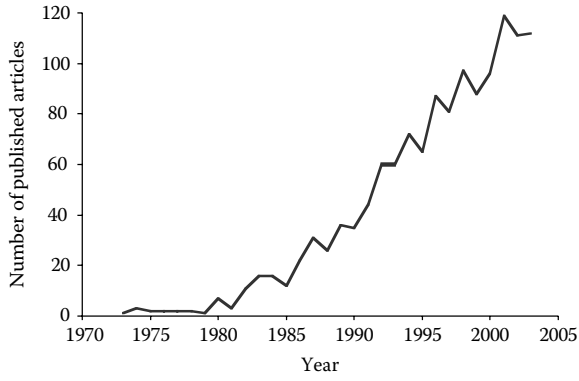


Figure 1.1 Growth in publications on correspondence analysis in selected journals. (Based on papers in 66 statistical journals and 22 books, thus representing a lower bound on the number of articles published. From Beh 2004).

There are several different ways of defining and thinking about CA, which is evident in the rediscovery of the method by so many different authors in the last century. We like to think of CA as a type of principal component analysis (PCA) of categorical data, where we consider the geometric definition of PCA rather than its statistical one. Similar to PCA, the rows or columns of a data matrix are assumed to be points in a high-dimensional Euclidean space, and the method aims to redefine the dimensions of the space so that the principal dimensions capture the most variance possible, allowing for lower-dimensional descriptions of the data. The fact that CA analyzes categorical data rather than metric data opens up the PCA style of data analysis to a world of new possibilities. Categorical data abound in almost all areas of research, especially in the social sciences, where most kinds of survey data are collected by means of nominally or ordinally scaled categorical items.

The history of correspondence analysis can be traced back to Hirschfeld (1935) (later changing his name to Hartley), who gave an algebraic formulation of the correlation between rows and columns of a contingency table. Fisher (1940) used the same ideas in the framework of discriminant analysis and is also regarded as one of the "founding fathers" of the technique. Independent from this approach, Guttman (1941) developed a method for constructing scales for categorical data, where he treated the general case for more than two qualitative variables. Since we are concentrating mostly on the multiple form of CA in this book, Louis Guttman should be credited as being the originator of the ideas behind present-day multiple correspondence analysis (MCA); he even used the term "principal components" and "chi-square distance" in his description of the method (Guttman 1950a,b). In the early 1950s Hayashi (1950, 1952, 1954) built on Guttman's ideas to create a method that he called "quantification of qualitative data," and was later followed in this tradition by the "dual scaling" ideas of Nishisato (1980, 1994). Apart from these brief historical remarks, we direct the interested reader to various texts where the history of CA is described, notably de Leeuw (1973), Nishisato (1980), Greenacre (1984), and Gifi (1990).

Our personal approach, due to our belief that the geometric approach has the most benefits, is to follow in great part the ideas of Jean-Paul Benzécri. Benzécri, a mathematician and linguist, developed CA and MCA in France in the 1960s and 1970s as part of a philosophy that placed the data firmly at the center of attention of the researcher. According to Benzécri, the data are king, not the model one might want to propose for them: one of his famous principles

states that "the model should follow the data, not the inverse." He gathered around him an influential team in France who made a large contribution to the early development of MCA—notably, Ludovic Lebart and Brigitte Escofier, to mention but two of the original key coworkers. Benzécri's original ideas are worthy of consideration because they do represent an extreme, counterbalancing the excessive attention paid to confirmatory modeling in statistics. Pretty much at the same time, and finding the right balance between Benzécri's ideas and statistical practice, one of the most important data analysis schools was developing in the Netherlands, inspired by Jan de Leeuw in Leiden and including Willem Heiser, Jacqueline Meulman, and Pieter Kroonenberg, to name but a few. De Leeuw engendered a style and culture of research on MCA and related methods that remain the strongest and most important at the present time (for an excellent review, see Michailidis and de Leeuw 1998). This group is best known for their collectively authored book under the nom de plume of Gifi, published internationally in 1990, but existing in other editions published in Leiden since 1981. In their approach, MCA (called homogeneity analysis) is the central analytical tool that is used to embed categorical data, through optimal scaling of the categories, into the interval-scale-based world of classical multivariate statistical analysis.

In the English-speaking world, interest in CA accelerated with the publication of textbooks by Lebart et al. and Greenacre, both published in 1984. In the late 1980s, several CA procedures were included in the leading statistical software packages of the time, notably SPSS, BMDP, and SAS. At the same time, the number of applications significantly increased, in the social sciences especially, influenced by the work of the French sociologist Pierre Bourdieu (see Rouanet et al. 2000), which has been translated into many languages. With these applications and with the further developments of the method, the number of yearly publications in this area increased steeply. In his overview of publications in the field of CA/MCA, Beh (2004) reports just one publication for 1973, 16 for 1983, 60 for 1993, and 112 for 2003 (the last year of his report), as shown in Figure 1.1.

In this book we have 23 chapters on the topic of MCA, starting from the basics and leading into state-of-the-art chapters on methodology, each with applications to data. There are data from social science research (Chapters 1, 2, 3, 7, 8, 22, and the Appendix), education (Chapters 4 and 5), health (Chapter 6), item responses in psychology (Chapters 9 and 10), marketing data and product preference (Chapters 11 and 19), health sciences (Chapter 12), food preferences (Chapter 13),

urban research (Chapter 14), elections in political science (Chapter 15), credit scoring (Chapter 16), epidemiology of animal health (Chapter 17), animal classification (Chapter 18), international comparisons (Chapter 20), psychological experiments (Chapter 21), and microarray studies in genetics (Chapter 23).

In this first and introductory chapter, we will start with a simple example of CA applied to a small two-way contingency table, explain the basics of the method, and then introduce MCA and related methods, ending up with a short overview of the rest of the book's contents.

1.2 A simple example

Correspondence analysis is an exploratory method based on well-known geometrical paradigms. To provide an initial illustration of the method, we use data from the most recently available survey from the International Social Survey Program (ISSP 2003). The original respondent-level data are available at the "Zentralarchiv für Empirische Sozialforschung" (Central Archive for Empirical Social Research) in Cologne, Germany (http://www.gesis.de). We start with a simple cross-tabulation of how respondents reacted to the statement, "When my country does well in international sports, it makes me proud to be {Country Nationality}." We chose respondents from five specific countries—U.K., U.S., Russia, Spain, and France—because their respective cities (London, New York, Moscow, Madrid, and Paris) were involved in the final decision for the summer Olympics in 2012. Respondents could give one of five possible responses to the above question—(1) "agree strongly," (2) "agree," (3) "neither agree nor disagree," (4) "disagree," (5) "disagree strongly"—as well as various "nonresponses." Table 1.1 shows the frequencies, and Table 1.2 the (column) percentages

Table 1.1 Frequencies of the cross-table "international sports × country."

	U.K.	U.S.	Russia	Spain	France	Total
Agree strongly	230	400	1010	201	365	2206
Agree	329	471	530	639	478	2447
Neither nor	177	237	141	208	305	1068
Disagree	34	28	21	72	50	205
Disagree strongly	6	12	11	14	97	140
Total	776	1148	1713	1134	1295	6066

Table 1.2 Column percentages of the cross-table "international sports × country."

	U.K.	U.S.	Russia	Spain	France	Average
Agree strongly	29.6	34.8	59.0	17.7	28.2	36.4
Agree	42.4	41.0	30.9	56.4	36.9	40.3
Neither nor	22.8	20.6	8.2	18.3	23.6	17.6
Disagree	4.4	2.4	1.2	6.5	3.9	3.4
Disagree strongly	0.8	1.1	0.6	1.2	7.5	2.3
Total	100.0	100.0	100.0	100.0	100.0	100.0

for each country. To keep our explanation simple at this stage and to avoid discussion about the handling of missing data, we have restricted attention to those respondents with no missing data for this question as well as for the other variables used in extended analyses of these data later in this chapter. These additional variables are on other aspects of national identity as well as several demographic variables: sex, age, marital status, and education. The topic of missing data is treated in Chapters 8 and 12 of this book.

Table 1.1 shows sample sizes ranging from 776 for U.K. to 1713 for Russia. In all countries, most people either "agree strongly" or "agree" with the statement on international sports. However, there are some differences between the countries: whereas in Russia most people "agree strongly," in Spain there are fewer than 20% giving this response. On the other hand, in France there is the largest share of respondents (7.5%) who "disagree strongly" with the statement. Calculating the chi-square statistic for testing independence on this table produces a value of $\chi^2 = 879.3$ (16 degrees of freedom), which is highly significant (P-value close to 0); Cramer's V measure of association is $V = 0.190$.

Analyzing Table 1.1 by CA gives a map of the pattern of association between countries and response categories, shown in Figure 1.2. This graphical representation, called the *symmetric map*, plots the *principal coordinates* of the rows and columns (to be explained more fully below). This two-dimensional map is not an exact representation of the data because it would require four dimensions to represent this 5×5 table perfectly. (We say the *dimensionality* of the table is four.) The two-dimensional map in Figure 1.2 accounts for 95.6% of the total "variance" in the table, where the measure of variance is closely related to the chi-square statistic. The objective of CA is to represent the maximum possible variance in a map of few dimensions, usually two dimensions. In this case there is only a small (4.4%) proportion of variance that is

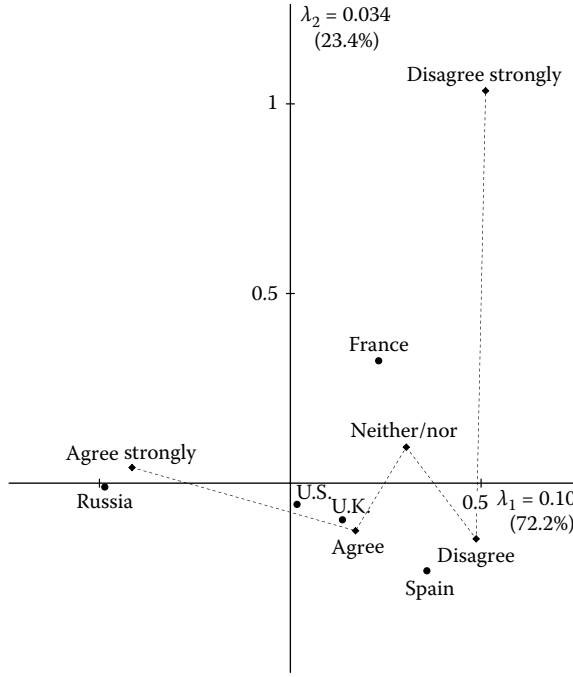


Figure 1.2 Symmetric correspondence analysis map of Table 1.1.

not represented here and that is effectively discarded because it is unlikely to be of interest. This strategy is identical in spirit to the coefficient of determination in linear regression, where we say that the predictors in a regression model explain a certain percentage of variance, with the remainder unexplained and relegated to "residual" or "error" variance. Each orthogonal axis in CA accounts for a separate part of variance, similar to uncorrelated predictors in a regression model: in Figure 1.2 the first axis explains 72.2%, the second an additional 23.4%.

Interpretation of the map consists in inspecting how the categories lie relative to one another and how the countries are spread out relative to the categories. The first (horizontal) dimension reflects a clear subdivision of the responses toward "international sports," with the category "agree strongly" on the left and the other four categories on the right. Furthermore, all categories retain their original order along the first dimension, although the intercategory distances are different: for example, "disagree" and "disagree strongly" are very close to each other along this dimension, while "agree strongly" is relatively far from "agree"

(projecting the categories perpendicularly onto the horizontal axis). The first dimension can be interpreted as "level of pride toward achievements in international sport," especially focused on the contrast between "strong agreement" and the other response categories. As for the countries, we see Russia on the left, opposing the other countries on the right; thus of these five nations, the Russians feel most proud when Russia is doing well in international sports. At the opposite right-hand side of this axis, we see that the French and Spanish are the least proud of the five nations in this respect, but there is also a big difference between these two along the second (vertical) dimension.

The second dimension mainly reflects the outlying position of "disagree strongly" as well as France compared with the other categories and countries. As we already noted by inspecting Table 1.2, 7.5% of the French chose this category compared with approximately 1% respondents from the other countries. This contrast is so strong that it accounts for most of the 23.4% of the variance along this second dimension.

In Table 1.2 the U.S. and U.K. have very similar response patterns, which are not much different from the overall, or average, pattern. Geometrically, this is depicted by these two countries lying close to each other, toward the origin of the map. The average response pattern is given in the last column of Table 1.2, which is the percentage responses for the whole data set.

All features in the map are relative in the sense that we can see that the Spanish are less proud of sporting achievement than the British, for example, but the map does not tell us how much less. The only point that is represented perfectly in the map is the center, or origin, which coincides with the average for the data set at hand, so we can judge in the map how the countries deviate from the average. To get an idea of absolute scale, an alternative way of representing the table is the so-called *asymmetric map*, where one set of points, usually the "describing variable" (the rows in this case), is depicted in *standard coordinates* (Figure 1.3), and the other set, the "variable being described" (the columns here), is depicted in principal coordinates as before (Greenacre and Blasius, 1994: Preface). In this map the country points, still in principal coordinates, are in the same positions as in Figure 1.2, but the category points are now reference points in the space depicting a 100% response in each of the five respective categories (notice the change in scale of Figure 1.3 compared with Figure 1.2). Now the "strong agreement" point, for example, represents a fictitious country where all respondents "strongly agreed." Thus we can see that Russia, for example, deviates from the average toward strong agreement, but now we can also judge how far away Russia is from being a country with 100% strong agreement.

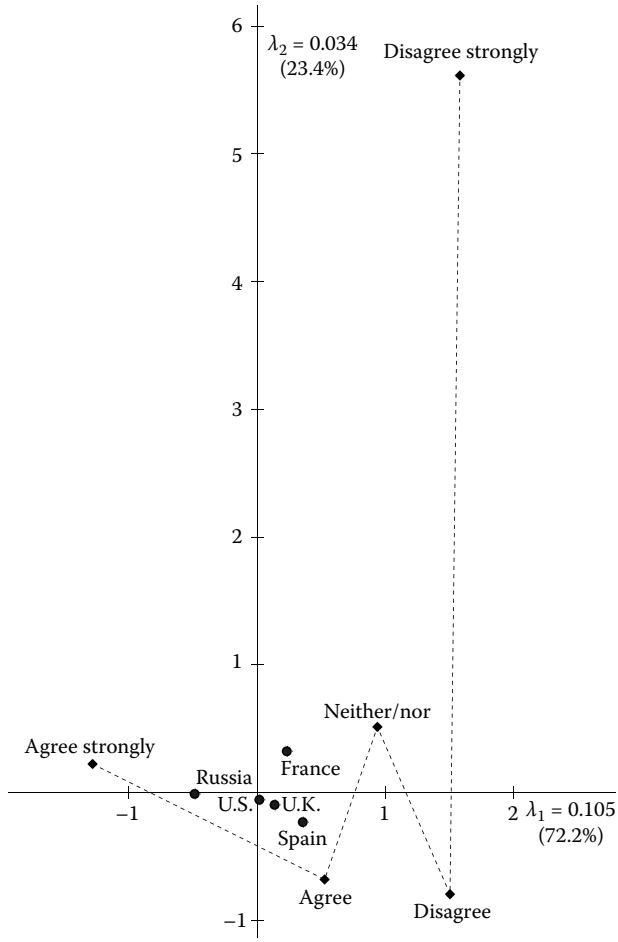


Figure 1.3 Asymmetric correspondence analysis map of Table 1.1.

Because the overall level of strong disagreement is low, this category is very far away from the countries in Figure 1.3, but France is closer to the strong disagreement "pole" than the others. Although enlightening as to the absolute level of variation existing among the countries, the asymmetric map is generally not preferred; the outlying positions of the set of points in standard coordinates (the response categories in Figure 1.3) force the other set of points (the countries) into a bunch at the center of the map. In the symmetric map of Figure 1.2, the category points—represented in principal coordinates—can be seen to lie in

positions along the axes, which are a scaled-down version of those in Figure 1.3. This fact, further explained in the next sections, underpins our interpretation of the symmetric map, which is the preferred map in CA.

1.3 Basic method

There are several different, but mathematically equivalent, ways to define CA. Because our approach is chiefly graphical and in the French tradition of Benzécri et al. (1973), we see CA as an adaptation to categorical data of PCA, which is a method for identifying dimensions explaining maximum variance in metric data. Both methods are based on decompositions of centered and normalized matrices, using either the eigenvalue-eigenvector decomposition of a square symmetric matrix or the singular-value decomposition (SVD) of a rectangular matrix. As in Greenacre (1984), we present the theory in terms of the SVD, which is an approach that is better equipped to show the relationship between row and column solutions. Similar to PCA, CA provides eigenvalues that are squared singular values (called principal inertias in CA), percentages of explained variance (percentages of inertia), factor loadings (correlations with principal axes), and communalities (percentages of explained inertia for individual rows or columns). In PCA, visualizations of the results are also made, but they are less common than in the CA/MCA framework, where the map is the central output for the interpretation.

For the presentation of the basic CA algorithm, we use a simple cross-table, or contingency table, of two variables with I rows and J columns, denoted by $\mathbf{N}$, with elements n_{ij}. As a first step, the *correspondence matrix* $\mathbf{P}$ is calculated with elements $p_{ij} = n_{ij}/n$, where n is the grand total of $\mathbf{N}$, the sample size in this case. CA analyzes the matrix $\mathbf{P}$ and is not concerned with the sample size n unless aspects of statistical inference such as confidence intervals are of interest (see Chapter 7). Corresponding to each element p_{ij} of $\mathbf{P}$ is a row sum $p_{i\cdot}$ ($= n_{i\cdot}/n$) and column sum $p_{\cdot j}$ ($= n_{\cdot j}/n$), denoted by r_i and c_j respectively. These marginal relative frequencies, called *masses*, play dual roles in CA, serving to center and to normalize the correspondence matrix.

Under the null hypothesis of independence, the expected values of the relative frequencies p_{ij} are the products $r_i c_j$ of the masses. Centering involves calculating differences $(p_{ij} - r_i c_j)$ between observed and expected relative frequencies, and normalization involves dividing these differences by the square roots of $r_i c_j$, leading to a matrix of

standardized residuals $s_{ij} = (p_{ij} - r_i c_j) / \sqrt{r_i c_j}$. In matrix notation this is written as:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-1/2} \tag{1.1}$$

where $\mathbf{r}$ and $\mathbf{c}$ are vectors of row and column masses, and $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices with these masses on the respective diagonals. The sum of squared elements of the matrix of standardized residuals, $\sum_i \sum_j s_{ij}^2 = \text{trace}(\mathbf{S}\mathbf{S}^\mathsf{T})$, is called the *total inertia* and is the amount that quantifies the total variance in the cross-table. Because the standardized residuals in $\mathbf{S}$ resemble those in the calculation of the chi-square statistic, χ^2, apart from the division by n to convert original frequencies to relative ones, we have the following simple relationship:

$$\text{total inertia} = \chi^2/n \tag{1.2}$$

The association structure in the matrix $\mathbf{S}$ is revealed using the SVD:

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} \tag{1.3}$$

where $\mathbf{\Sigma}$ is the diagonal matrix with singular values in descending order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_S > 0$, where S is the rank of matrix $\mathbf{S}$. The columns of $\mathbf{U}$, called *left singular vectors*, and those of $\mathbf{V}$, the *right singular vectors*, are orthonormal: $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$. The connection between the SVD and the eigenvalue decomposition can be seen in the following:

$$\mathbf{S}^\mathsf{T}\mathbf{S} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$$

$$\mathbf{S}\mathbf{S}^\mathsf{T} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$$

showing that the right singular vectors of $\mathbf{S}$ correspond to the eigenvectors of $\mathbf{S}^\mathsf{T}\mathbf{S}$, the left singular vectors correspond to the eigenvectors of $\mathbf{S}\mathbf{S}^\mathsf{T}$, and the squared singular values σ^2 in $\mathbf{\Sigma}^2$ correspond to the eigenvalues λ of $\mathbf{S}^\mathsf{T}\mathbf{S}$ or $\mathbf{S}\mathbf{S}^\mathsf{T}$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. Within the context of CA, these eigenvalues are termed *principal inertias*, and their sum $\sum_s \lambda_s$ is equal to the total inertia since $\text{trace}(\mathbf{S}\mathbf{S}^\mathsf{T}) = \text{trace}(\mathbf{S}^\mathsf{T}\mathbf{S}) = \text{trace}(\mathbf{\Sigma}^2) = \text{trace}(\mathbf{\Lambda})$.

The SVD provides all the results we need to make CA maps. The principal and standard coordinates can be calculated for the row and column categories:

$$\text{principal coordinates of rows:} \quad \mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Sigma} \tag{1.4}$$

$$\text{standard coordinates of rows:} \quad \mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U} \tag{1.5}$$

$$\text{principal coordinates of columns:} \quad \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Sigma} \tag{1.6}$$

$$\text{standard coordinates of columns:} \quad \mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V} \tag{1.7}$$

For a two-dimensional map we would use either (a) the first two columns of the coordinates matrices **F** and **G** for the symmetric map, (b) **A** and **G** for the asymmetric map of the columns (called "column principal" in SPSS, for example), or (c) **F** and **B** for the asymmetric map of the rows ("row principal"). The proportion of inertia explained would be $(\sigma_1^2 + \sigma_2^2)/\sum_s \sigma_s^2$, i.e., $(\lambda_1 + \lambda_2)/\sum_s \lambda_s$. For further details about the computation of CA, see Blasius and Greenacre (1994).

1.4 Concepts of correspondence analysis

In this section we summarize the main concepts underlying CA, mostly geometric concepts. Each concept will be illustrated briefly in the context of the CA of the "international sports" example of Table 1.1 and Figure 1.2 and Figure 1.3.

1.4.1 Profiles, average profiles, and masses

As mentioned previously, CA is based on relative values; the sample size is not important for the construction of the map. The data table can be expressed as proportions (or percentages) relative to the row or column margins. Table 1.2 illustrates the latter possibility, showing for each country the percentage responses across the categories of the variable "international sport": for example, "U.K.: agree strongly" is 230/776 = 0.296 (or 29.6%). The columns containing the relative frequencies for the single countries are called *profiles*, in this case *column profiles*. Furthermore, we calculate average profiles as the row or column margins relative to the grand total: for example, the profile value for "All countries: agree strongly" is 2206/6066 = 0.364 (or 36.4%). This *average column profile* is given in the last column of Table 1.2. In the

CA map, the average column profile is represented at the origin where the axes cross.

Table 1.2 shows that Russia has a profile value in the category "agree strongly" clearly above average (0.590 > 0.364), whereas the respective value of Spain is clearly below average (0.177 < 0.364). Under the condition that the two countries are well represented in the map, which is very likely, as 95.6% of the variation is explained by the first two dimensions, Russia should be associated strongly with "agree strongly," whereas Spain should be just the opposite. That this is true has already been seen in Figure 1.3. Comparing the column profiles with the average column profile as well as with one another gives a first understanding of which columns (countries) should be located close to one another and which should be separated.

Table 1.2 shows the column profiles and the average column profile. In the same way, we could calculate the *row profiles*, expressing the frequencies in each row of Table 1.1 relative to their corresponding row total. Further, we can compute the *average row profile*, i.e., the column sums (or column totals, see the last row in Table 1.1) divided by the sample size. We could then compare the elements of the row profiles with their corresponding elements in the average row profile, which gives a first understanding of which row categories are in the same part of the map and which are relatively distinct from one another. In the map of the row profiles, the origin again reflects the position of the average row profile.

The distinction between the presentation of the rows, i.e., discussing the differences between the row profiles compared with the average row profile, and the presentation of the column profiles, i.e., discussing the differences between the column profiles compared with the average column profile, is also reflected in the two possibilities of visualizing the data using asymmetric maps. Figure 1.3 shows the column profiles in a map spanned by the rows. The categories of the rows are expressed in terms of "artificial countries," for example, the position of "agree strongly" reflects a hypothetical "country" containing only respondents all strongly agreeing with the statement. We also could calculate an asymmetric map in which the profiles of the rows are visualized in a space spanned by the columns. Because the differences between the countries are of central interest, this possibility of visualization is less meaningful. Finally, and most often done, row and column profiles are shown together (symmetric map). However, there is a close relation between standard and principal coordinates because there are only scale-factor differences between

them along the axes. The standard coordinates can be obtained by dividing the principal coordinates by the singular values (i.e., by the square roots of the principal inertias), for example for the rows: $\mathbf{A} = \mathbf{F}\boldsymbol{\Sigma}^{-1}$ (see Equation 1.4 to Equation 1.7).

In addition to being elements of the average profile, the marginal values described previously (denoted by r_i and c_j in Section 1.3) are also used as weights in CA to give more or less importance to the display of the individual row and column profiles. For this reason they are called masses, and their dual role in the analysis will become apparent when we explain the concepts of total variance and distance in the next two sections.

1.4.2 Chi-square statistic and total inertia

A common way of describing the relationships in contingency tables is the chi-square statistic, which tests for significant associations between rows and columns. The chi-square statistic is defined as the sum of squared deviations between observed and expected frequencies, divided by the expected frequencies, where the expected frequencies are those calculated under the independence model:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \quad \text{where } \hat{n}_{ij} = n_{i\cdot} \times n_j / n, \{i = 1, 2, \ldots, I; j = 1, 2, \ldots, J\}$$

Repeating this calculation for relative frequencies p_{ij}, we obtain the chi-square statistic divided by the grand total n of the table:

$$\frac{\chi^2}{n} = \sum_i \sum_j \frac{(p_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}}, \quad \text{where } \hat{p}_{ij} = r_i \times c_j \tag{1.8}$$

This is exactly the total inertia defined in Section 1.3, the sum of squared standardized residuals in the matrix $\mathbf{S}$ of Equation 1.1. Calculating the chi-square value for Table 1.1 gives $\chi^2 = 879.3$ and a total inertia of $\chi^2/n = 0.145$. Comparing the total inertia with other solutions taken from literature, the value is relatively high, which means that there is a relatively large amount of variation in the data or, equivalently, there are relatively large differences between the countries in terms of their national pride concerning international sports.

Table 1.3 Chi-square components of Table 1.1.

	U.K.	U.S.	Russia	Spain	France	Total
Agree strongly	9.66	0.73	240.46	108.36	23.83	383.04
Agree	0.81	0.13	37.52	72.05	3.77	114.28
Neither nor	11.93	6.02	85.52	0.35	26.00	129.82
Disagree	2.31	3.00	23.51	29.59	0.89	59.30
Disagree strongly	7.92	7.93	20.60	5.66	150.70	192.81
Total	32.63	17.81	407.61	216.01	205.19	879.25

Note: The contributions to total inertia are these values divided by $n = 6066$.

Table 1.3 shows the chi-square components for each cell, showing how much each column (country), each row (response category for "international sports"), and each cell (country × response category) contribute to the deviation from independence. The larger the deviation, the larger the contribution to chi-square (equivalently, to total inertia), and the larger will be the influence on the geometric orientation of the axes in CA. The largest deviations from independence are due to Russia, with a contribution of 46.4% (407.61/879.25). Within "Russia" the response "agree strongly" has a chi-square component of 240.46, i.e., 59.0% of the inertia of Russia and 27.3% of the total inertia. This explains why the first and most important principal axis in the CA of Table 1.1 showed Russia and "agree strongly" clearly separated from other rows and columns. In contrast, U.K. and especially the U.S. have very little contribution to total inertia (3.7% and 2.0%, respectively), i.e., they are close to the average. A comparison of the three column profiles of U.K., U.S., and Russia with the average column profile (Table 1.2) supports these findings: the differences between the respective profile elements are large for Russia and small for U.K. and U.S. Furthermore, the contributions to total inertia are also reflected in the map: U.K. and U.S. are relatively close to the origin, and Russia is relatively far away.

With respect to the response categories, the strongest impacts on total inertia are from "agree strongly" and "disagree strongly": both are relatively far away from the origin. However, although "agree strongly" has a higher contribution to total inertia (383.04/879.25 = 0.436 or 43.6%) than "disagree strongly" (21.9%), the profile of the latter is farther away from the origin (see Figure 1.2). This is due to the different masses of the two response categories, as will be explained in more detail in Section 1.4.5.

1.4.3 Chi-square distances

In the CA of a contingency table such as Table 1.1, distances between any pair of row profiles or between any pair of column profiles become apparent. From Equation 1.8, we can rewrite the total inertia as

$$\frac{\chi^2}{n} = \sum_i \sum_j c_j \frac{(p_{ij}/c_j - r_i)^2}{r_i}$$

where p_{ij}/c_j is an element of the jth column profile: $p_{ij}/c_j = n_{ij}/n_{\cdot j}$; and r_i is the corresponding element of the average column profile: $r_i = n_{i\cdot}/n$. This shows that the total inertia can be interpreted geometrically as the *weighted sum of squared distances* of the column profiles to the average profile: the weight of the column profile is its mass c_j, and the squared distance is a Euclidean-type distance where each squared difference is divided by the corresponding average value r_i. For this reason this distance is known as the chi-square (χ^2) distance. For example, from Table 1.2, the χ^2 distance between U.K. and the "average country" is:

$$d_{\text{UK,O}} = \sqrt{\frac{(.296-.364)^2}{.364} + \frac{(.424-.403)^2}{.403} + \frac{(.228-.176)^2}{.176} + \frac{(.044-.034)^2}{.034} + \frac{(.008-.023)^2}{.023}}$$

$$= 0.205$$

In a similar fashion, the χ^2 distance between U.K. and the U.S. is:

$$d_{\text{UK,US}} = \sqrt{\frac{(.296-.348)^2}{.364} + \frac{(.424-.410)^2}{.403} + \frac{(.228-.206)^2}{.176} + \frac{(.044-.024)^2}{.034} + \frac{(.008-.011)^2}{.023}}$$

$$= 0.151$$

In Figure 1.2 we can see that the positions of the origin, U.K., and the U.S., closely agree with these interpoint χ^2 distances. Similar distance calculations can be made between the row profiles and their average and between pairs of row profiles.

In the CA solution, the χ^2 distances between profiles are visualized as ordinary Euclidean distances. For example, the (squared) χ^2 distance between two columns j and j' is exactly equal to the (squared) Euclidean distance $\sum_s (g_{js} - g_{j's})^2$ between the points in principal coordinates in the full S-dimensional space.

1.4.4 Reduction of dimensionality

As in PCA and in other data reduction techniques, as few dimensions as possible are used for interpretation. Due to limitations in the graphical display, in CA/MCA usually only planar maps are used, showing pairs of dimensions at a time, although three-dimensional graphical displays are becoming easier to use. (See, for example, Rovan 1994, and the RGL package in the R language presented in the computational appendix of this book.) The determination of the number of dimensions to be interpreted can be performed in various ways, similar to PCA: (a) consider all those with eigenvalues that explain more than average inertia, (b) examine a "scree plot" of the eigenvalues to identify the "elbow" in the descending sequence, or (c) use the application-based method of including all dimensions that have a coherent substantive interpretation (see also Blasius 1994). The issue of eigenvalues and their percentages of explained inertia is more problematic in MCA, a subject that will be treated in detail in Chapter 2.

1.4.5 Contributions to inertia

The principal coordinate positions of the row and column points, relative to their respective averages, are given by Equation 1.4 and Equation 1.6, respectively. Because chi-square distances are equal to distances between the points represented in principal coordinates in the full space, an alternative way to calculate total inertia and the inertial contributions of Section 1.4.2 is as a weighted sum of squares of principal coordinates. For example, multiplying the mass of the ith row (r_i) by the squared coordinate of the ith row on the sth dimension (f_{is}^2) gives the amount of inertia of row i on axis s. The sum of these inertias over all rows and over all dimensions, i.e., $\sum_i \sum_s r_i f_{is}^2$, gives the total inertia. The same holds for the columns, and we have the equality:

$$\text{total inertia} = \sum_i \sum_s r_i f_{is}^2 = \sum_j \sum_s c_j g_{js}^2 \tag{1.9}$$

Summing over points on single dimensions s gives the principal inertias λ_s: $\sum_i r_i f_{is}^2 = \sum_j c_j g_{js}^2 = \lambda_s$, $s = 1, 2, \ldots, S$, again recovering the total inertia as $\sum_s \lambda_s$. Summing over dimensions for a single point gives the inertia of the corresponding row or columns. For example, the inertia

of the ith row is $r_i \sum_s f_{is}^2$. Using these decompositions of inertia in terms of points and dimensions, we can compute:

1. The contribution from each row or each column to total inertia. This is the amount of variance each row or column contributes to the geometric model as a whole (as described in Section 1.4.2 in terms of chi-square, or inertia components).
2. Same as item 1, but with respect to single dimensions.
3. The contribution of each dimension to total inertia, i.e., the explained variance of each dimension.
4. The contribution of each dimension to the inertia of a row or column, i.e., the explained variance of each dimension to a point. The square roots of these values are often called *factor loadings* because they are also correlation coefficients between the point and the dimension.
5. The amount of explained variance of the first S^* dimensions to each row or to each column. These coefficients are called *qualities* in CA, known as *communalities* in PCA.

Together with the eigenvalues (principal inertias), these five coefficients constitute the standard numerical output in CA, as given by such statistical packages as SPSS and XLSTAT.

Mass plays a crucial role in the inertia contributions of each row and each column, and it must be considered when interpreting a CA map or, even earlier, when starting to perform a CA. In the given example, the contribution of "agree strongly" to total inertia is about twice that of "disagree strongly" (Table 1.3), but the mass of "agree strongly" is about 15 times higher than the mass of "disagree strongly" (Table 1.2), resulting in a shorter distance to the origin. In general, categories with low frequencies (i.e., with low masses) tend to be outlying because their distribution is often quite different from the average. In such cases these categories can contribute quite highly to the total inertia. For example, the fifth category "disagree strongly" has a low mass ($r_5 = 0.023$; see Table 1.2) but has a relatively high share of the total inertia (150.70/879.25 = 0.171; see Table 1.3) owing to its high profile value on France (97/140 = 0.693; see Table 1.1). For this reason, categories with very low relative frequencies should be carefully monitored in CA, and if they contribute too much to the solution, they should be combined with other categories in a substantively meaningful way.

1.4.6 Reconstruction of the data

In log-linear analysis, contingency tables can be reconstructed by means of interaction effects of different order. In CA, this reconstruction is performed by means of the margins and the bilinear decomposition inherent in the SVD. A common feature of both techniques is that the sparsest model is chosen. In log-linear analysis, this is the model with the fewest interaction effects; in CA, it is the model with the fewest dimensions. (A detailed description of the relationship between these models is given by van der Heijden et al. 1989, 1994; see also Goodman 1991.) Exact data reconstruction in CA can be obtained using the row and column principal coordinates and singular values on all dimensions:

$$\mathbf{P} = \mathbf{r}\mathbf{c}^{\mathsf{T}} + \mathbf{D}_r\mathbf{F}\boldsymbol{\Sigma}^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{D}_c \tag{1.10}$$

where $\mathbf{r}\mathbf{c}^{\mathsf{T}}$ is the matrix of expected relative frequencies under independence. Various alternative forms of this reconstruction formula are possible, for example in terms of row standard and column principal coordinates:

$$\mathbf{P} = \mathbf{r}\mathbf{c}^{\mathsf{T}} + \mathbf{D}_r\mathbf{A}\mathbf{G}^{\mathsf{T}}\mathbf{D}_c$$

since $\mathbf{A} = \mathbf{F}\boldsymbol{\Sigma}^{-1}$ (see Equation 1.4 and Equation 1.5).

In CA maps, where a reduced number S^* of dimensions are used (setting $\sigma_{S^*+1} = \sigma_{S^*+2} = \ldots = \sigma_S = 0$ in Equation 1.10), the data reconstruction is not exact, approximating the data as well as the percentage of inertia accounted for by the solution. The reconstruction formula shows how CA can be considered as a model for the table, with parameters fitted by weighted least squares.

1.5 Stacked tables

In the case of simple CA, described previously, the frequencies of a single contingency table are used as input information. In this section, we describe how CA can be used to visualize several contingency tables at a time.

As a first example of a stacked table, the variable of interest, "country," is cross-tabulated by several variables describing the countries, in this case several variables on national identity; the cross-tables are stacked one on top of each other, i.e., row-wise. One of the most famous applications of CA to such a table is given by Bourdieu (1979). In his book *La Distinction*, he describes the French population, differentiated by classes of occupation, using a large set of lifestyle indicators (for example, preferences in arts and music); see Blasius and Winkler (1989).

Table 1.4 Possibilities of stacked tables.

NI1 × Country	NI1 × Sex	NI1 × Mar.Status	NI1 × Edu.Level	NI1 × Age
NI2 × Country	NI2 × Sex	NI2 × Mar.Status	NI2 × Edu.Level	NI2 × Age
NI3 × Country	NI3 × Sex	NI3 × Mar.Status	NI3 × Edu.Level	NI3 × Age
NI4 × Country	NI4 × Sex	NI4 × Mar.Status	NI4 × Edu.Level	NI4 × Age
NI5 × Country	NI5 × Sex	NI5 × Mar.Status	NI5 × Edu.Level	NI5 × Age
NI6 × Country	NI6 × Sex	NI6 × Mar.Status	NI6 × Edu.Level	NI6 × Age
NI7 × Country	NI7 × Sex	NI7 × Mar.Status	NI7 × Edu.Level	NI7 × Age

We could also extend the set of tables columnwise by adding cross-tables of "international sports" with variables such as sex, marital status, educational level, and age group. In this case, one could describe which of the sociodemographic characteristics are most important to explain the responses toward "international sports." Table 1.4 shows the possible combinations of stacking tables for seven indicators on national identity (NI1 to NI7) with country and four sociodemographic indicators.

The simplest case is the CA on a single table, for example, NI1 × Country, which has been shown previously in Figure 1.2 and Figure 1.3. The next possibility is to add other indicators of national identity, as shown in the first column of Table 1.4. Another possibility is the columnwise stacking, as shown in the first row in Table 1.4. The third and most complex possibility is to analyze all 35 cross-tables as a single matrix input to CA. Before we analyze this complex table, we show the CA solution of the stacked table with country as column variable and seven statements toward national identity as row variables. These statements are (in the order as given in the questionnaire, the first letter indicating the variables in the forthcoming figures):

a. I would rather be a citizen of {Country} than of any other country in the world.
b. There are some things about {Country} today that make me feel ashamed of {Country}.
c. The world would be a better place if people from other countries were more like the {Country nationality}.
d. Generally speaking, {Country} is a better country than most other countries.
e. People should support their country even if the country is in the wrong.
f. When my country does well in international sports, it makes me proud to be {Country nationality}.
g. I am often less proud of {Country} than I would like to be.

All variables have five categories as previously: (1) "agree strongly," (2) "agree," (3) "neither nor," (4) "disagree," and (5) "disagree strongly" (the number indicating the category in the figures).

Compared with the CA on the simple table where the two-dimensional solution accounts for 95.6% of the variation, the two-dimensional solution of the stacked table accounts for only 72.0%, even though the dimensionality of the stacked table is also four. In Figure 1.4, four of the seven "agree strongly" responses (questions b, e, f, and g) are located on the negative side of the first dimension; the remaining three (questions a, c, and d) are on the negative part of dimension 2. For the countries, Russia is on the same side of dimension 1 as strong agreements to "there are some things about Russia today that make me feel ashamed," "people should support their country even if the country in the wrong," "… international sports …," and "I'm often less

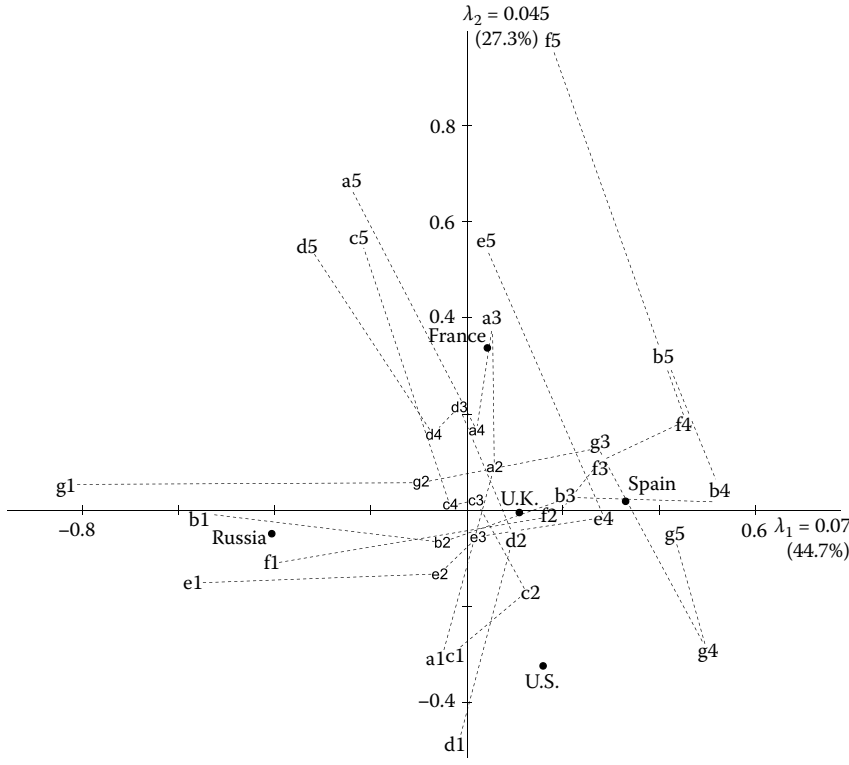


Figure 1.4 Symmetric CA map for stacked table of "national identity" indicators by "country."

proud of Russia than I would like to be," which suggests a mixture of pride toward achievement in international sports and a critical reflection toward their own country and their own attitudes. In contrast, the respondents from the U.S. agreed relatively strongly to statements that reflect national pride: "I would rather be a citizen of the U.S. than of any other country in the world," "the world would be a better place if people from other countries were more like Americans," and "generally speaking, the U.S. is a better country than most other countries."

The first dimension mirrors mainly a contrast between agreement and disagreement toward the statements b, e, f, and g, whereby the "disagree strongly" responses for e and f are located on the positive side of dimension 2. The second factor reflects the contrast between a strong agreement toward statements that suggest a kind of superiority of the home country and a strong disagreement toward most of the other statements on national identity (including the critical ones). It can be concluded that, of the five nations analyzed here, Americans are most proud of their country, whereas the French are the least proud. The British are close to the center, i.e., close to average. The Spanish are closest to some disagreement on national identity, but they relatively often avoid the choice of the strongest disagreement.

The first column of Table 1.5 shows the decomposition of inertia over the seven subtables of the stacked table. Total inertia of the stacked table is 0.1646, which can be shown to be the average of the inertias of

Table 1.5 Decomposition of inertias ($N = 6066$) for all cross-tables.

	Country	Sex	Marital Status	Educational Level	Age Groups	Average
Citizen of country	0.1690	0.0003	0.0258	0.0279	0.0544	0.0555
Feel ashamed of country	0.1898	0.0068	0.0096	0.0306	0.0098	0.0493
World would be better	0.0978	0.0029	0.0131	0.0502	0.0291	0.0386
Country is better than others	0.1289	0.0029	0.0093	0.0188	0.0204	0.0361
People should support country	0.1754	0.0009	0.0165	0.0377	0.0268	0.0515
Well in international sport	0.1450	0.0001	0.0114	0.0295	0.0108	0.0394
I am often less proud of country	0.2465	0.0031	0.0145	0.0155	0.0127	0.0585
Average	0.1646	0.0024	0.0143	0.0300	0.0234	0.0469

the seven subtables: (0.1690 + 0.1898 + ⋯ + 0.2465)/7 = 0.1646. The largest differences between the countries are in the responses toward statement g, "I am often less proud …," which corresponds to the subtable making the highest contribution to this average.

We now analyze the complete data matrix, extending the previous data set columnwise by stacking the sets of cross-tables with a number of sociodemographic characteristics. The final table is a supermatrix containing seven variables stacked row-wise and five variables stacked columnwise. The seven variables on national identity have been cross-tabulated with the following variables (with abbreviations as used in Figure 1.5):

Country: U.K., U.S., Russia, Spain, France — as already done
Sex: male (m), female (f)
Marital status: married (mar), widowed (wid), divorced (div), separated (sep), single (sin)
Education: no formal education (E1), lowest formal education (E2), above lowest formal education (E3), higher secondary (E4), above secondary (E5), university (E6)
Age groups: till 25 (A1), 26 to 35 (A2), 36 to 45 (A3), 46 to 55 (A4), 56 to 65 (A5), 66 and older (A6).

In total there are 24 sociodemographic categories, including the countries, so the resulting table of frequencies is 35×24. The CA map is given in Figure 1.5.

The dimensionality of the solution of the stacked table is determined by the minimum of the I rows and J columns minus the respective number of variables, i.e., $\min(I - Q_r; J - Q_c) = \min(35 - 7; 24 - 5) = 19$, where Q_r = number of row variables and Q_c = number of column variables. Total inertia of this supertable is 0.0469, which is again the average of the inertias of the 35 subtables (Table 1.5) (see Greenacre 1994). The first dimension explains 37.2% of the total variation, the second another 31.2%. Because the solution has 19 dimensions, 68.4% explanatory power for the first two dimensions is reasonable and is quite close to the previous solution. However, total inertia is—compared with the previous solution—relatively low because there are many cross-tables in the supermatrix with low levels of association.

Comparing the solutions of the row-wise stacked table with "country" as column variable and the supertable containing 35 cross-tables shows that, on the level of the responses toward national identity, there are almost no differences; the general structure of the response categories is the same. The same holds for the countries; they keep

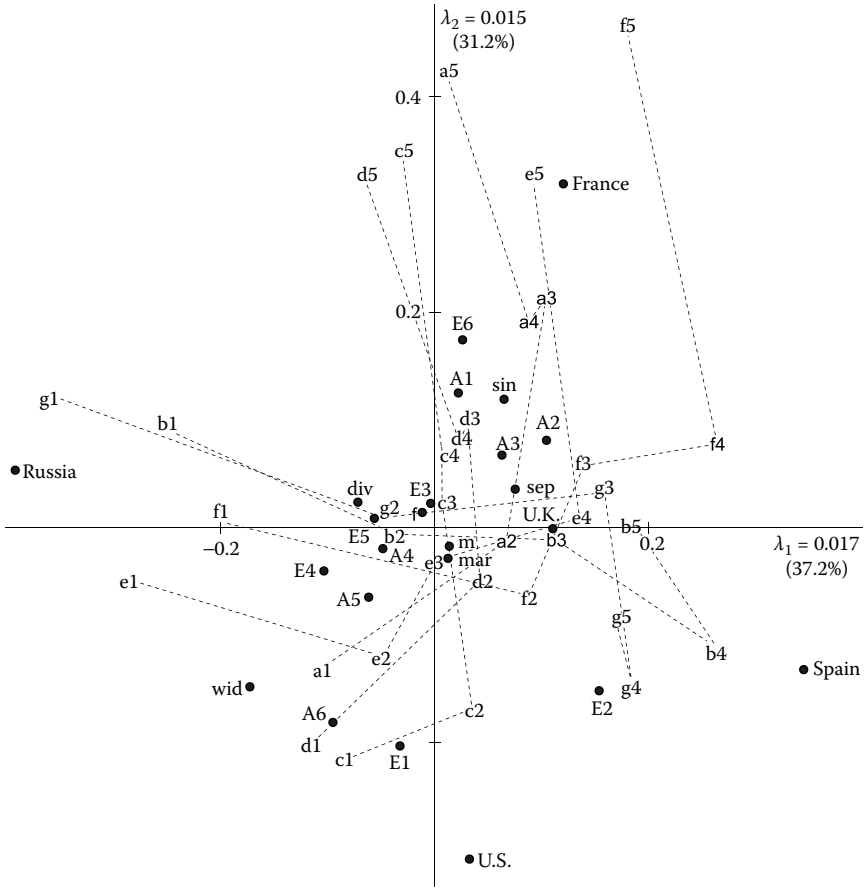


Figure 1.5 Symmetric CA map based on 35 tables, stacked row-wise and column-wise.

their relative positions in the two-dimensional space. In addition to the previous solution, one only can add some findings for the additional sociodemographic variables considered in Figure 1.5. For example, respondents 66 years and older (A6) and widowed persons (wid) are relatively often proud to live in the country where they live. Young people (A1), singles (sin), as well as respondents with a university degree (E6) have relatively often a low national identity. Furthermore, attitudes toward national identity seem to be uncorrelated with sex: there are almost no differences between males and females (compare their location close to the origin in Figure 1.5).

As already noted, in the complex table total inertia is the average value of the 35 subtables. Table 1.5 shows the inertia of all 35 tables as well as the average inertias of the sociodemographic characteristics and of the seven national identity indicators. It can be seen that the highest inertias belong to the cross-tabulations with country, i.e., the most variation in the data is caused by country differences. Further, there are almost no sex differences for the seven items on national identity, although there are some findings that might be worthwhile to report. For example, the association between "sex" and "international sport" is much smaller than the association between "sex" and "feel ashamed of country."

1.6 Multiple correspondence analysis

In the previous examples we analyzed the relation between two variables or between two different sets of variables. In this section, we are interested in the relationships within a set of variables, for example, the interrelationships between the statements on national identity. Thus, for example, we could find out if there is an association between a "strong agreement toward international sports" and a "strong agreement toward people should support their country." In the previous analysis of stacked tables, we could only see whether these categories had the same association with sociodemographic variables.

This new case, which is reminiscent of principal component analysis, involves all the cross-tables of a set of variables, such as the national identity indicators, with themselves. Assembling all these cross-tables into a square supermatrix of cross-tables, we obtain what is known in CA literature as the *Burt matrix*, which we denote by **C**. Alternatively, a data structure known as the *indicator matrix* can be constructed based on the original data. The indicator matrix, denoted by **Z**, is a respondents-by-categories table with as many rows as respondents (6066 in our example) and as many columns as response categories (35 for the seven national identity indicators). The elements of **Z** are zeros apart from ones in the positions to indicate the categories of response of each respondent (**Z** is often called a matrix of dummy variables). The Burt matrix is related quite simply to the indicator matrix as follows: $\mathbf{C} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z}$. If the usual CA algorithm is applied to an indicator matrix or to a Burt matrix, the method is called *multiple correspondence analysis* (MCA). In MCA there is no distinction

between describing variables and variables to be described, as is the case in simple CA of single or stacked tables. In MCA all variables have the same status. The relationship between the analyses of **C** and **Z** in MCA is discussed in depth in Chapter 2. In the following, we illustrate the method by analyzing the 6066×35 indicator matrix **Z** that codes the responses to the seven national identity questions. The graphical solution is given in Figure 1.6.

Inspecting Figure 1.6, we see that the first dimension contrasts the strong agreements and the strong disagreements (positive part) from the middle categories (negative part). With two exceptions (statements b and g), the second dimension contrasts the positive statements from the negative ones. Therefore, it can be seen as an overall dimension toward national identity, with a relatively high national identity in the positive part and a relatively low national identity in the negative part. The variables a, c, d, e, and f form a horseshoe, a typical structure we usually find in ordered categorical data (for more details, see Chapters 2 and 4). However, there are two points to be mentioned. First, neither item b, "there are some things …," nor item g, "I am often less proud…," fulfill this structure, and maybe even worse, the most-opposite categories "b1" and "b5" as well as "g1" and "g5" are close to each other. One reason for this finding might be that a significant number of respondents misunderstood the direction of the question, which would result in such a structure (Blasius and Thiessen 2001b). Another reason is that two dimensions are not sufficient to mirror the structure of these two variables adequately. In a higher-dimensional solution "b1" and "b5" as well as g1 and g5 might be far away from each other. Second, the horseshoe belongs to the first dimension, i.e., the second dimension is more important for substantive interpretation. This might be caused by the joint analysis of the data from five countries, where respondents in different countries may understand the single questions (slightly) differently (see Chapter 20).

Notice that we have not indicated the percentages of inertia explained in Figure 1.6. There are several methods of scaling the solution in MCA that change the fit (see Chapters 2 and 3), but the overall structure of the variable categories in the space remains the same and, hence, the substantive interpretation too. Using the indicator matrix as input, the measure of fit in terms of explained inertia is heavily underestimated. Various solutions to this problem are proposed in Chapter 2 (see Section 2.3.4).

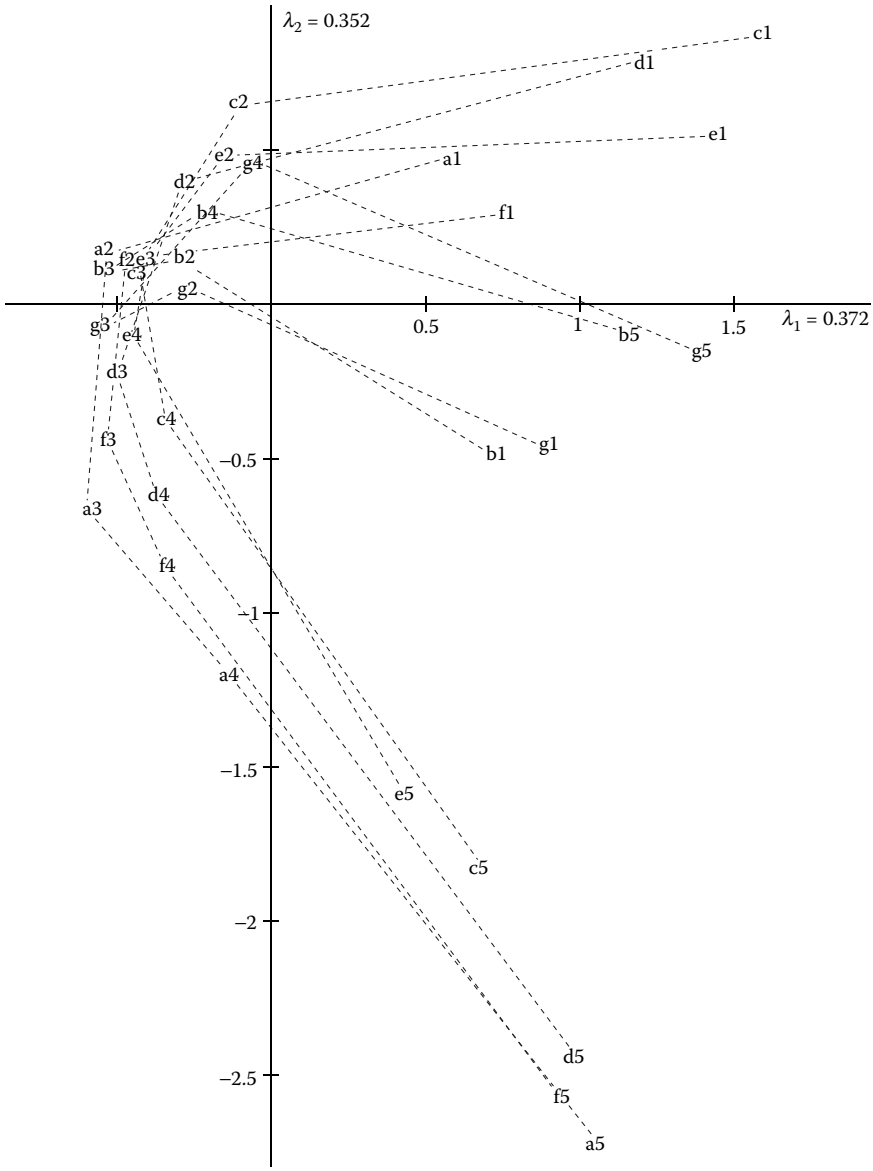


Figure 1.6 MCA map of indicators on national identity.

1.7 Categorical principal component analysis

In the previous sections we discussed CA and its extension to stacked tables and MCA. There are several situations in which the variable categories are ordered, for example, ordered from "agree strongly" to "disagree strongly," as in our example. Where MCA does not impose any constraints on the data, principal component analysis (PCA) has linear constraints, i.e., it assumes that the categories are ordered and that the distances between the categories are constant. Categorical principal component analysis (CatPCA), also known as nonlinear PCA (NLPCA), can be understood as a technique intermediate between (linear) PCA and (nonlinear) MCA. In its most utilized form, CatPCA takes into account the ordering of the categories but allows the intervals between categories to vary. CatPCA is the PCA of a data matrix of categorical data where the category values (1 to 5 in the example on national identity) are replaced by optimal scale values on each dimension. (For more details, see Chapter 4 as well as Gifi 1990 and Heiser and Meulman 1994.) The optimal scaling process allows order constraints to be imposed so that ordered categorical variables get increasing, or at least nondecreasing, quantifications in the low-dimensional solution space (usually two dimensional). When the ordering of the categories in CatPCA is not consistent with the implied ordering, this manifests itself in the form of tied optimal quantifications for two or more subsequent categories. Unlike classical PCA and unlike MCA, the number S^* of dimensions required must be specified in advance because the solutions are not nested.

The results of a CatPCA are mapped in the form of straight lines through the origin for the respective variables, with the response categories indicated on each vector (see Figure 1.7). The first dimension of the CatPCA to the seven indicators on national identity accounts for 34.6% of the variation, and the second accounts for another 23.3%; thus 57.9% of the variation is explained with the first two dimensions. The axes are labeled on the "agree strongly" responses, with the ticks showing the categories. Figure 1.7 shows that the distances between the successive categories are different. With respect to questions b and g, the last two categories ("disagree" and "disagree strongly") are tied (shown by an empty circle). In all questions, the largest difference is between "agree strongly" and "agree." As already shown in the MCA solution (Figure 1.6), questions b, "there are some things about …," and g, "I am often less proud …," seem to measure something different than the other five questions, appearing almost uncorrelated with them. Furthermore, questions a, c, and d, all three measuring pride toward country as part of

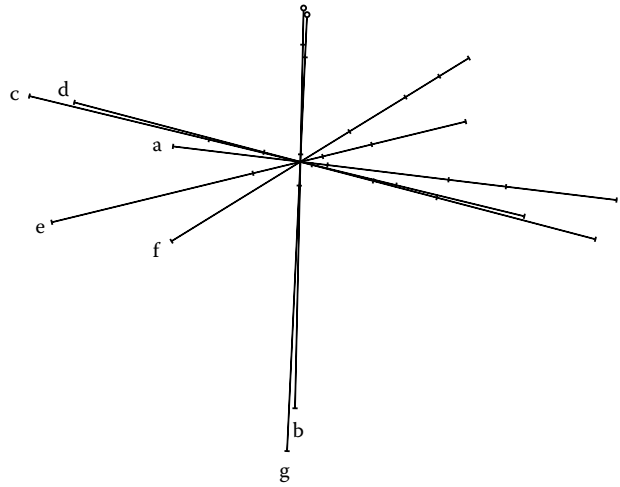


Figure 1.7 CatPCA map of indicators on national identity.

one's national identity, are very close to each other. More details on the theory and practice of CatPCA are given in Chapter 4.

1.8 Active and supplementary variables

For simple CA, MCA, and CatPCA, it is possible to project categories of additional variables on an already existing solution configuration. These additional variables are called supplementary variables, sometimes also referred to as "illustrative" or "passive" as opposed to the "active" variables of the analysis that determine the solution space. Supplementary variables have no influence on the geometric orientation of the axes; rather, they support and complement the interpretation of the configuration of active variable categories. One can think of supplementary points as additional points in the row or column profile spaces; these points have zero mass and thus play no role in the analysis apart from interpreting their positions.

To display supplementary points, we use the so-called *transition formulas* that relate row and column solutions, which are derived from the linear relationships between left and right singular vectors in an SVD. For example, to obtain the principal coordinates **F** of the row points from the principal coordinates **G** of the column points, we have

from Equation 1.1, Equation 1.3, Equation 1.4, and Equation 1.6 the following relationship:

$$\mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}}\mathbf{F}\boldsymbol{\Sigma}^{-1} \tag{1.11}$$

The matrix $\mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}}$ contains the column profiles of Table 1.2 (but written here as relative frequencies as row vectors), while $\mathbf{F}\boldsymbol{\Sigma}^{-1}$ is the matrix of row standard coordinates, denoted by $\mathbf{A}$ in Equation 1.5. This shows the barycentric relationship between rows and columns in CA: the column points (in principal coordinates) are weighted averages of the row points (in standard coordinates). This formula allows any additional column profile to be displayed on the map by computing its position as a weighted average of the row (standard) coordinates using its profile elements as weights. A similar transition formula can be obtained that allows supplementary row points to be displayed.

1.9 Multiway data

In CA of stacked tables and MCA, only two-way interaction effects are taken into account in determining the solution. These methods are often described as being *joint bivariate* in this sense. In some situations, however, higher-order interactions need to be taken into account. A common approach to analyzing multiway data is to code two or more variables interactively, which we illustrate in the context of the national identity data. We saw in Section 1.5 that there was little difference between average male and female attitudes. But there could exist some larger male–female differences in certain countries that are masked by the comparison of all males with all females. To visualize male–female differences for each country in terms of their attitudes to national identity—in other words a three-way interaction—we would create a new "interactive" variable called "country-gender" with 10 categories (5 countries × 2 genders) and then use this variable instead of the separate country and gender variables. Examples of interactive coding are given by Greenacre (1993a), who investigates whether the choice of a car depends on income group in combination with age group, as well as Carlier and Kroonenberg (1998), who examine the connection of region and profession at different points in time on the basis of data from official statistics. In Chapter 21 a data set with a four-way structure is analyzed, and two variables are interactively coded so that the data structure becomes three-way.

1.10 Content of the book

The aim of this book is to present introductory and state-of-the-art material on MCA and related techniques, both from a methodological and an applications perspective. Information on this topic is currently scattered, not all of it is in English, and the information has not been unified notationally or directly compared. Some literature is very statistical, while other works are only applications-oriented and lack important details on the methodology and general rules of interpretation. This volume is a compromise between statistical methodology and applications and is designed to explain the methodology to social scientists and other researchers in an intuitive, example-based way. At the same time, it provides statisticians and other methodologists with the theoretical background for understanding MCA and methods related to it. The book is subdivided into five parts: Introduction, Multiple Correspondence Analysis, Analysis of Sets of Tables, MCA and Classification, and Related Methods. Here we give a brief overview of these parts and the chapters comprising them.

1.10.1 Introduction

The first four chapters of the book contain the main concepts and the theoretical background of CA, MCA, and some of the related methods discussed and applied in the subsequent chapters. In these four chapters we provide an overview of these techniques and show how they are related to one another. In this first chapter we have already given an overview of CA of a single table and of stacked tables between two sets of variables, and a brief introduction to MCA. We have introduced the basic geometric concepts of profiles, masses, inertia, chi-square distances, and the reduction of dimensionality, all of which are important for the understanding of CA and the interpretation of CA maps. We also focused specifically on the measure of variance in the tables, i.e., the total inertia, and we showed how, in the multiple case, total inertia is equal to the average of the inertias of the tables constituting the stacked matrix. This is an aspect that is useful in the understanding of MCA.

In Chapter 2, Michael Greenacre discusses in detail some ways to generalize simple CA to MCA. He starts with the case of canonical correlation analysis, which leads to a solution that maximizes the correlation between the row variable and the column variable of a two-way contingency table. The extension to the multivariate case involves applying a more general definition of correlation among sets of variables,

leading to a version of MCA also known as homogeneity analysis. The geometric version of this is shown to reduce to the CA of the data coded as an indicator matrix or a Burt matrix. In both cases, a large amount of the total variance is induced purely by the data coding; hence, Greenacre proposes an adjusted version of MCA as the method of choice, where the coordinates have been rescaled to better estimate the fit of the solution. Joint correspondence analysis is also treated here, in which the effects of the main diagonal blocks in the Burt tables are excluded. The methodology is illustrated extensively using empirical data on attitudes toward science from the International Social Survey Program (ISSP).

Chapter 3, written by John Gower, aims to show similarities (and dissimilarities) between a number of techniques that are all concerned with two-way arrays. Two-way arrays can be of many different types: for example, tables of values on a single variable observed on a two-way classification, two-way contingency tables, square correlation matrices based on metric data, indicator matrices, and Burt tables. All techniques to analyze two-way arrays have in common either a decomposition in terms of simple structures such as main effects plus interactions or the singular-value decomposition. Gower discusses and compares principal component analysis, correspondence analysis, and multiple correspondence analysis. He also compares the fit measures for each of these techniques and the effect of scaling of the axes on the final solutions. For his empirical example he uses data from the ISSP on national identity.

Chapter 4, written by Jan de Leeuw, is a state-of-the-art description of nonlinear principal component analysis (NLPCA), also known as categorical principal component analysis. De Leeuw starts with an explanation of PCA and extends it to the nonlinear case, providing algorithmic details that are needed to understand the background of the method. Furthermore, he shows the relation with MCA (or homogeneity analysis, in the Dutch terminology) as well as with multiple regression, and an alternative way of performing NLPCA using a logistic approach. He demonstrates the methodology using mainly a Dutch data set on primary schoolchildren.

1.10.2 Multiple correspondence analysis

The next seven chapters are on multiple correspondence analysis, showing this methodology from several different points of view. Chapter 5 by Henry Rouanet gives the French view of PCA and MCA,

especially the role played by Jean-Paul Benzécri in the methodological development of these methods, and goes on to cite the influence of Pierre Bourdieu on their application. This approach to PCA and MCA, known as *analyse des données* in French, is called "geometric data analysis" (LeRoux and Rouanet 2004a). He starts with an explanation of PCA and extends it to MCA, where he gives the basic rules of interpretation. As examples, he uses two data sets, one from the judgment of basketball experts on high-level potential players and one from the Education Program for Gifted Youth.

In Chapter 6, Shizuhiko Nishisato discusses different aspects of the correlational structure of multichoice data from the view of dual scaling. Dual scaling of multivariate categorical data leads to the same solution as MCA, but it is part of a general framework of data scaling used in many different contexts. This chapter shows how the method can capture both linear and nonlinear associations between the variables. Further, Nishisato gives an overview on forced classification of dual scaling, which can be understood as a procedure for discriminant analysis for categorical data. As an empirical example, he uses a small data set from a health survey.

Chapter 7, written by Ludovic Lebart, is dedicated to validation techniques in MCA. One of the criticisms of MCA is that it does not involve techniques for statistical inference. However, there are several methods to validate the findings statistically, two of which are discussed in this chapter. The first is based on external validation, which involves external data, usually included as supplementary or passive variables, leading to cross-validation of the results. The other possibility is based on internal validation, using resampling techniques such as the bootstrap and other Monte Carlo methods. Lebart illustrates the different techniques using a British data set in which the respondents were asked about their standard of living and expectations for the future.

In Chapter 8, Michael Greenacre and Rafael Pardo discuss the application of subset correspondence analysis to the case of MCA. The idea here is to concentrate on some response categories of the variables only, excluding others from the solution. For example, missing responses on several questions can be analyzed alone, or substantive responses excluding missing values can be analyzed and mapped. In the former case, patterns of missing values can be explored on their own, focusing on their relationships with sociodemographic characteristics. In the latter case, the advantage of this method would be to keep all information that is available in the study and not lose any respondent data, as happens when applying listwise deletion of cases.

The authors demonstrate their methodology using ISSP data about attitudes toward women in the labor force.

In Chapter 9, Matthijs Warrens and Willem Heiser discuss the scaling of unidimensional models with MCA. The objective of this chapter is to determine what information on the parameters can be obtained from the application of MCA when exact unidimensional models are used as gauges, or benchmarks. The authors discuss eight possible models—where each model is either deterministic or probabilistic, dichotomous or polytomous, and monotonic or unimodal— and how these models are related to MCA. In the literature, these models are known under names such as Guttman and Rasch scales. The authors show the structure of these models and how they are generated, as well as the MCA solutions of simulated item-response data.

Chapter 10, written by Wijbrandt van Schuur and Jörg Blasius, has a similar purpose as Chapter 9 by Warrens and Heiser. Different item-response data—for example, dominance and cumulative data, proximity or unfolding data—give certain graphical patterns when mapped by MCA. This allows the authors to differentiate between unfolding and Rasch data, for example, on the basis of the results of an MCA. After discussing some of the typical patterns that these models provide, the authors apply MCA to two data sets that are labeled as dominance and unfolding data. Whereas data on religious beliefs from the Dutch World Value Survey form a dominance structure, as expected, there is no evidence that the unfolding data form an "unfolding structure."

In Chapter 11, Yoshio Takane and Heungsun Hwang discuss the topic of regularization in the MCA context. Regularization can be considered as an important and general way to supplement insufficient data by prior knowledge or to incorporate certain desirable properties in the estimates of parameters in the model. Because MCA does not always provide estimates that are on average closest to the population parameters, the authors propose an alternative estimation procedure for MCA, called regularized MCA. Using two small data sets taken from the literature, the authors compare the regularized solution with the ones obtained by MCA.

1.10.3 Analysis of sets of tables

Chapters 12 through 15 deal with different sets of data to be analyzed simultaneously. In Chapter 12, written by Herbert Matschinger and