Analysis of Mixed Data: Methods & Applications

Edited by

Alexander R. de Leon
Keumhee Carrière Chough

To my family
— A. R. L.

To Nari and Hanna
— K. C. C.

Preface

Multivariate data of mixed types occur frequently in many fields of science and social science. The analysis of such data has created new challenges that have made it necessary to develop new statistical techniques and methodologies. Statisticians, working in collaboration with biologists, economists, epidemiologists, social scientists, and many others, have met these challenges with many remarkable advances over the past two decades. These include, among others, applications of mixed models to mixed outcomes in clustered and longitudinal studies, advances in dependence modeling of mixed data via copula and graphical models, extensions of Bayesian methods to mixed data settings, and adaptations of entropy- and divergence-based association measures to mixed outcomes.

Despite the attention researchers have given to mixed data analysis in recent years, there has been no single book that focused purely on this important topic. A close scrutiny of the literature reveals the following textbooks and monographs that contain discussions, albeit mostly brief, of mixed data analysis:

- H. Goldstein (2011). *Multilevel Statistical Models*. 4th ed.. Wiley & Sons, Inc.
- D. M. Berridge and R. Crouchley (2011). *Multivariate Generalized Linear Mixed Models in R*. Chapman & Hall/CRC.
- L. Wu (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC.
- C. E. McCulloch, S. R. Searle, and J. M. Neuhaus (2008). *Generalized, Linear, and Mixed Models*. 2nd ed.. Wiley & Sons, Inc.
- M. J. Daniels and J. W. Hogan (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall/CRC.
- P. X.-K. Song (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.
- G. Molenberghs and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer.
- A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC.
- R. J. Little and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. 2nd ed.. Wiley & Sons, Inc.
- J. L. Schafer (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- D. R. Cox and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretations*. Chapman & Hall.
- G. A. F. Seber (1984). *Multivariate Observations*. Wiley & Sons, Inc.

In addition, the following edited volumes devote separate chapters to mixed data:

- G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.) (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC.
- M. Aerts, H. Geys, G. Molenberghs, and L. M. Ryan (Eds.) (2002). *Topics in Modelling of Clustered Data*. Chapman & Hall/CRC.

However, since these books only provide snapshots of contemporary developments in mixed data analysis, there is thus a need for an authoritative book on mixed data analysis that traces important developments, systematizes terminology and methodologies, and gives an overview of applications. Our intention in producing this book is to show the depth and diversity of current research in the field, bringing together the work of as many researchers as possible, thus providing synthesis as well as development of directions for future research.

The twelve chapters in this book were written by leading researchers who have made important and sustained contributions to mixed data analysis, and who were selected with a view to covering as much ground as possible in this broad area. While each chapter is self-contained and can be read independently of the others, we have informally organized them into groups to facilitate smooth thematic transitions. With the technical nature of the subject and this being an edited volume with numerous chapter authors, we have endeavored to maintain a certain degree of clarity and harmony in the presentation style. We believe that we have achieved this goal for the following reasons:

- The book was carefully and thoroughly edited for smooth readability and seamless transitions between chapters. All the chapters follow a common structure, with an introduction and a concluding summary, and include illustrative examples, many drawn from real-life case studies in developmental toxicology, economics, medicine and health, marketing, and genetics.
- We have included ample cross-references between chapters to enable readers to connect the book's various topics and research strands and to facilitate self-study.
- We have, as much as possible, unified notations, table formats, and terminologies across chapters. In particular, we have adopted, whenever possible, a common set of notations for mathematical and statistical quantities, such as vectors and matrices (both random and fixed) as well as distributions.
- To facilitate easy referencing by readers, we have come up with a combined index as well as a single up-to-date bibliography for the entire book. The references collected at the end of the book provide the most comprehensive, most current, and most complete list of published material on mixed data analysis.
- As a unique feature of the book, we have included an introductory chapter that provides a "wide angle" overview and comprehensive survey of mixed data analysis. The chapter contains useful background material that should prepare readers for the rest of the book.

As with any similar collection, there are bound to be omissions of some topics. While we strived for a good balance between theory and applications, it is impossible to include all topics in a single volume. For example, mixed-data graphical models are not discussed, apart from a brief mention in Chapter 1. Nevertheless, the book's technical level along with the many examples and case studies should make the book appeal to a broad audience. We believe that it would be a valuable resource to methodologically interested as well as subject matter-motivated researchers. These include graduate students, applied statisticians, biostatisticians, and researchers in subject matter areas like medicine, health, genetics, and epidemiology, among many others. The book should be an excellent supplement to the textbooks and monographs enumerated above.

This book was completed with considerable help from several people. Our deepest thanks go to all the contributors, for their patience, enthusiasm, and expertise. Likewise, we are indebted to the following reviewers for their input and comments: Dipankar Bandyopadhyay (Division of Biostatistics, University of Minnesota, Minneapolis, MN), Kenneth A. Bollen (Department of Sociology, University of North Carolina, Chapel Hill, NC), Claudia Czado (Zentrum Mathematik, Technische Universität München, Munich, Germany), Ruzong Fan (Biostatistics and Bioinformatics Branch, National Institute of Child Health and Human Development, Bethesda, MD), Garrett Fitzmaurice (Department of Biostatistics, Harvard School of Public Health, Boston, MA), Helena Geys (Center for Statistics, Hasselt University, Diepenbeek, Belgium), Harry Joe (Department of Statistics, University of British Columbia, Vancouver, BC), Stefan Lang (Department of Statistics, Universität Innsbruck, Innsbruck, Austria), Huilin Li (Department of Environmental Medicine, New York University, New York, NY), Mingliang Li (Department of Economics, SUNY at Buffalo, Buffalo, NY), Irini Moustaki (Department of Statistics, London School of Economics, London, UK),

Raffaella Piccarreta (Department of Decision Sciences, Università Bocconi, Milan, Italy), David S. Siroky (School of Politics & Global Studies, Arizona State University, Tempe, AZ), and Qiaohao Zhu (Cross Cancer Institute, Edmonton, AB). We thank as well Michael J. Daniels, Christel Faes, Jian Kang, Helga Wagner, and David M. Zimmer, for doubling as reviewers.

Special thanks go to Beilei Wu, for her invaluable help at crucial stages, Yin Li, for computational assistance, and Hanna Carrière, for the lovely cover art. We are also grateful to Wiley & Sons, Inc., for permission to reproduce Figure 5.1 in Chapter 5.

This book was made possible by funding from the Natural Science and Engineering Research Council (NSERC) of Canada. Part of the work was done in Australia while Keumhee C. Chough was visiting Prof. Richard Higgins at the University of Melbourne. The contributors also acknowledge financial support from various research organizations and agencies: NSERC and Le Fonds québécois de la recherche sur la nature et les technologies, for grants to Denis Larocque and François Bellavance; National Institute on Drug Abuse, for Grants 5K05DA000017 and 5P01DA001070 to Peter M. Bentler, who also acknowledges a financial interest in EQS and its distributor Multivariate Software; Japan Science and Technology Agency, for PRESTO 3709 award to Takahiro Hoshino; Tsinghua-Yue-Yuen Medical Science Fund and Natural Science Foundation of China, for Grant 11071137 to Ying Yang; Portuguese National Foundation for Science and Technology, for Grant PTDC/SAU-ESA/100841/2008 to Armando Teixeira-Pinto; National Institutes of Health, for Grant NIH CA85295 to Michael J. Daniels.

It was a pleasure to work with John Kimmel and his colleagues in the production department at CRC/Chapman & Hall. John's commitment to and encouragement of this project from first to last has been remarkable. We thank him for giving us the opportunity to work on this book.

Finally, we express our gratitude to all our family and friends. Keumhee, in particular, is especially grateful to Jean Chough and James Osadczuk, for their loving support throughout the book production.

Alexander R. de Leon
Calgary, Alberta
Keumhee C. Chough
Edmonton, Alberta

Contents

Editors

Alexander R. de Leon is Associate Professor in the Department of Mathematics and Statistics at the University of Calgary. Originally from the Philippines, he obtained his BSc and MSc, both in Statistics, from the School of Statistics of the University of the Philippines. After a research studentship at Tokyo University of Science, he completed his PhD in Statistics in 2002 at the University of Alberta. His research interests include methods for analyzing correlated data, multivariate models and distances for mixed discrete and continuous outcomes, pseudo- and composite likelihood methods, copula modeling, assessment of diagnostic tests, statistical quality control, and statistical problems in medicine, particularly in ophthalmology.

Keumhee Carrière Chough is Professor of Statistics in the Department of Mathematical and Statistical Sciences at the University of Alberta. After completing her BSc in Agriculture from Seoul National University, in Seoul, Korea, she earned her MSc from the University of Manitoba, and her PhD in Statistics from the University of Wisconsin-Madison in 1989. Since 1996, she has been with the Department of Mathematical and Statistical Sciences, University of Alberta, after stints as Assistant Professor at the University of Iowa (1990–1992) and University of Manitoba (1992–1996). She was also the Director of the Statistics Consulting Center at the University of Iowa (1990–1992). Her research interests include design and analysis for repeated measures data, missing data methods, high dimensional data analysis methods, multivariate methods, designs for clinical trials, item response data, variable selection methods, and survival analysis. As well, she specializes in such biostatistical methods as small area variation analysis techniques with applications to health care utilization. She has been a Health Scientist funded through the Alberta Heritage Foundation for Medical Research (1996–2011). She is a Fellow of the American Statistical Association, the Institute of Health Economics, and the Manitoba Centre for Health Policy.

Contributors

Qi An
Department of Mathematical Sciences
University of Memphis
Memphis, TN, USA

François Bellavance
Department of Management Sciences
HEC Montréal
Montréal, QC, Canada

Peter M. Bentler
Departments of Psychology and Statistics
University of California-Los Angeles
Los Angeles, CA, USA

Dale Bowman
Department of Mathematical Sciences
University of Memphis
Memphis, TN, USA

Keumhee Carrière Chough
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB, Canada

Michael J. Daniels
Department of Statistics
University of Florida
Gainsville, FL, USA

Alexander R. de Leon
Department of Mathematics and Statistics
University of Calgary
Calgary, AB, Canada

Abdessamad Dine
Department of Management Sciences
HEC Montréal
Montréal, QC, Canada

Christel Faes
Interuniversity Institute for Biostatistics and Statistical Bioinformatics
Hasselt University
Diepenbeek, Belgium

Jeremy T. Gaskins
Department of Statistics
University of Florida
Gainsville, FL, USA

E. Olusegun George
Department of Mathematical Sciences
University of Memphis
Memphis, TN, USA

Ralitza Gueorguieva
School of Public Health
Yale University
New Haven, CT, USA

Jaroslaw Harezlak
Department of Biostatistics
Indiana University School of Medicine
Indianapolis, IN, USA

Takahiro Hoshino
Department of Economics and Business Administration
Nagoya University
Chikusa-ku, Nagoya, Japan

Jian Kang
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA, USA

Minjung Kwak
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD, USA

Denis Larocque
Department of Management Sciences
HEC Montréal
Montréal, QC, Canada

Armando Teixeira-Pinto
Department of Health Information and Decision Sciences
University of Porto
Porto, Portugal

Regina Tüchler
Department of Statistics
Austrian Federal Economic Chamber
Wien, Austria

Helga Wagner
Department of Applied Statistics and Econometrics
Johannes Kepler University
Linz, Austria

Niroshan Withanage
Department of Mathematics and Statistics
University of Calgary
Calgary, AB, Canada

Beilei Wu
Department of Mathematics and Statistics
University of Calgary
Calgary, AB, Canada

Colin O. Wu
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD, USA

Ying Yang
Department of Mathematical Sciences
Tsinghua University
Beijing, P. R. China

Gang Zheng
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD, USA

David M. Zimmer
Department of Economics
Western Kentucky University
Bowling Green, KY, USA