

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 Osadcuk, for their loving support throughout the book production.

Alexander R. de Leon

Calgary, Alberta

Keumhee C. Chough

Edmonton, Alberta



Contents

Preface	iii
Editors	xiii
Contributors	xv
List of Figures	xix
List of Tables	xxi
1 Analysis of mixed data: An overview	1
<i>by Alexander R. de Leon and Keumhee Carrière Chough</i>	
1.1 Introduction	1
1.2 Early developments in mixed data analysis	1
1.2.1 Early models	2
1.2.2 Multivariate methods for mixed data	5
1.3 Issues in joint analysis of mixed outcomes	6
1.3.1 Joint models for mixed outcomes	6
1.3.1.1 Direct approaches	7
1.3.1.2 Indirect approaches	7
1.3.2 Estimation	9
1.4 Highlights of book	10
2 Combining univariate and multivariate random forests for enhancing predictions of mixed outcomes	13
<i>by Abdessamad Dine, Denis Larocque and François Bellavance</i>	
2.1 Introduction	13
2.2 Predictions from univariate and multivariate random forests	14
2.2.1 Data setting	14
2.2.2 Basic tree building	15
2.2.3 Multivariate tree for mixed outcomes	15
2.2.4 Random forests	16
2.2.5 Description of predictive models	16
2.3 Simulation study	18
2.3.1 Simulation design	18
2.3.2 Results	19
2.4 Discussion	29
3 Joint tests for mixed traits in genetic association studies	31

<i>by Minjung Kwak, Gang Zheng and Colin O. Wu</i>	
3.1	Introduction 31
3.2	Analysis of binary or quantitative traits 32
3.3	Joint analysis of mixed traits 34
3.4	Application 39
3.5	Discussion 40
4	Bias in factor score regression and a simple solution 43
<i>by Takahiro Hoshino and Peter M. Bentler</i>	
4.1	Introduction 43
4.2	Model 45
4.3	Bias of estimates due to estimated factor scores: Factor analysis model 47
4.4	Proposed estimation method 49
4.4.1	Theoretical justification 50
4.4.1.1	When observed variables are continuous 50
4.4.1.2	General case 50
4.4.2	Alternative estimation method 50
4.5	Simulation studies 51
4.5.1	Study 1: Two-dimensional factor analysis model 51
4.5.2	Study 2: Two-level factor analysis model 51
4.5.3	Study 3: Factor analysis and two-parameter logistic models 54
4.5.4	Study 4: Factor analysis and nominal response models 54
4.6	Application 56
4.7	Theoretical details 59
4.7.1	Generalized least squares as the fourth step 59
4.7.2	Sketch of proof of Proposition 3 59
4.8	Discussion 60
5	Joint modeling of mixed count and continuous longitudinal data 63
<i>by Jian Kang and Ying Yang</i>	
5.1	Introduction 63
5.2	Complete data model 64
5.2.1	GEE approach 64
5.2.2	Modeling correlations between outcomes 65
5.2.3	Varying-coefficients model 66
5.2.4	Fully parametric model 67
5.3	Handling missing data problem 68
5.3.1	Parametric missing data model 68
5.3.2	Imputation 69
5.4	Application 73
5.4.1	Complete data model 73
5.4.1.1	GEE approach 73
5.4.1.2	Varying-coefficients model 75
5.4.1.3	Fully parametric model 75
5.4.2	Missing data model 76
5.4.2.1	Fully parametric model 77
5.4.2.2	Imputation 77
5.5	Discussion 78

6	Factorization and latent variable models for joint analysis of binary and continuous outcomes	81
	<i>by Armando Teixeira-Pinto and Jaroslaw Harezlak</i>	
6.1	Introduction	81
6.2	Clinical trial on bare-metal and drug-eluting stents	82
6.3	Separate analyses	82
6.4	Factorization models for binary and continuous outcomes	84
6.5	Latent variable models for binary and continuous outcomes	86
6.6	Software	89
6.7	Discussion	91
7	Regression models for analyzing clustered binary and continuous outcomes under the assumption of exchangeability	93
	<i>by E. Olusegun George, Dale Bowman and Qi An</i>	
7.1	Introduction	93
7.2	Distribution theory and likelihood representation	95
7.3	Parametric models	96
	7.3.1 Model descriptions	96
	7.3.2 Estimation of parameters	98
7.4	Application to DEHP data	98
	7.4.1 Data summary	98
	7.4.2 Estimates and model comparisons	100
7.5	Litter-specific joint quantitative risk assessment	102
7.6	Discussion	106
8	Random effects models for joint analysis of repeatedly measured discrete and continuous outcomes	109
	<i>by Ralitza Gueorguieva</i>	
8.1	Introduction	109
8.2	Models	111
	8.2.1 Multivariate GLMMs	112
	8.2.2 Correlated probit model	113
8.3	Estimation and inference	114
	8.3.1 Estimation via Gauss-Hermite quadrature	115
	8.3.2 Estimation via the EM algorithm	116
	8.3.3 Inference	118
8.4	Applications	118
	8.4.1 Developmental toxicity study in mice	118
	8.4.2 Depression clinical trial	120
8.5	Discussion	121
9	Hierarchical modeling of endpoints of different types with generalized linear mixed models	125
	<i>by Christel Faes</i>	
9.1	Introduction	125
9.2	Multivariate multi-level models	126
9.3	Special cases	127
	9.3.1 Marginal model	127
	9.3.2 Shared random effects model	128

9.3.3	Correlated random effects model	128
9.3.4	Independent random effects and correlated errors model	129
9.4	Likelihood inference	130
9.4.1	Pseudo-likelihood estimation	130
9.5	Applications	131
9.5.1	Toxicology of ethylene glycol	131
9.5.1.1	Separate analysis of endpoints	132
9.5.1.2	Bivariate clustered data model	133
9.5.2	Health interview survey	134
9.5.2.1	Bivariate hierarchical model	135
9.5.2.2	Trivariate hierarchical model	136
9.6	Discussion	137
10	Joint analysis of mixed discrete and continuous outcomes via copula models	139
	<i>by Beilei Wu, Alexander R. de Leon and Niroshan Withanage</i>	
10.1	Introduction	139
10.2	Joint models via copulas	140
10.2.1	Case of independent data	141
10.2.2	Case of clustered data	143
10.3	Associations	144
10.3.1	Case of independent data	145
10.3.2	Case of clustered data	146
10.4	Likelihood estimation	147
10.5	Analysis of ethylene glycol toxicity data	148
10.5.1	Analysis ignoring clustering	149
10.5.2	Analysis accounting for clustering	151
10.6	Discussion	155
11	Analysis of mixed outcomes in econometrics: Applications in health economics	157
	<i>by David M. Zimmer</i>	
11.1	Introduction	157
11.2	Random effects models	159
11.2.1	Specification of joint distribution	159
11.2.2	Maximum likelihood estimation	160
11.2.3	Quasi-random number generation	160
11.2.4	Drawbacks of random effects models	161
11.3	Copula models	162
11.3.1	Copula basics	162
11.3.2	Specification of joint distribution	163
11.3.3	Maximum likelihood estimation	164
11.3.4	Drawbacks of copula models	164
11.4	Application to drug spending and health status	165
11.4.1	Data	165
11.4.2	Specification of marginal and joint distributions	166
11.4.3	Results	167
11.5	Application to nondrug spending and drug usage	169
11.5.1	Data	169
11.5.2	Specification of marginal and joint distributions	170
11.5.3	Results	170
11.6	Discussion	172

12 Sparse Bayesian modeling of mixed econometric data using data augmentation	173
<i>by Helga Wagner and Regina Tüchler</i>	
12.1 Introduction	173
12.2 Model specification	174
12.2.1 Data augmentation	175
12.2.2 Modeling dependence for cross-sectional data	176
12.2.3 Modeling dependence for longitudinal data	177
12.2.4 Prior distributions	178
12.3 Logit-normal model	179
12.3.1 Model specification	179
12.3.2 Priors	180
12.3.3 MCMC estimation	181
12.3.3.1 Sampling component indicators	181
12.3.3.2 Sampling variance and correlation parameters	181
12.3.3.3 Sampling latent utilities	182
12.3.3.4 Sampling indicator variables and regression coefficients	182
12.3.4 Simulation example	182
12.4 Modeling material deprivation and household income	183
12.5 Estimating consumer behavior from panel data	186
12.6 Discussion	187
13 Bayesian methods for the analysis of mixed categorical and continuous (incomplete) data	189
<i>by Michael J. Daniels and Jeremy T. Gaskins</i>	
13.1 Introduction	189
13.2 Examples	191
13.3 Characterizing dependence	194
13.4 (Informative) Priors	195
13.5 Incomplete responses	198
13.6 General computational issues	201
13.7 Analysis of examples	203
13.8 Discussion	208
Bibliography	209
Index	231



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 Gainesville, 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 Gainesville, 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
Jaroslav 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

