

Contents

1	Compositional Data as a Methodological Concept	1
1.1	What Are Compositional Data?	1
1.2	Introductory Problems	5
1.2.1	PhD Students Example	5
1.2.2	Beer Data Example	8
1.2.3	Geochemical Data Example.....	10
1.3	Principles of Compositional Data Analysis	11
1.4	Steps to a Concise Methodology	14
	References.....	15
2	Analyzing Compositional Data Using R	17
2.1	Brief Overview on Packages Related to Compositional Data Analysis.....	17
2.1.1	compositions	18
2.1.2	robCompositions.....	18
2.1.3	ggtern.....	21
2.1.4	zCompositions	21
2.1.5	mvoutlier, StatDA	21
2.1.6	CoDaPack	21
2.1.7	compositionsGUI	22
2.2	The Statistics Environment R.....	22
2.3	Basics in R.....	22
2.3.1	Installation of R and Updates	24
2.3.2	Install robCompositions	24
2.3.3	Help	25
2.3.4	The R Workspace and the Working Directory	26
2.3.5	Data Types	27
2.3.6	Generic Functions, Methods and Classes.....	32
	References.....	33

3	Geometrical Properties of Compositional Data	35
3.1	Motivation	35
3.2	Aitchison Geometry on the Simplex	40
3.3	Coordinate Representations of Compositions.....	43
3.3.1	Additive Logratio (alr) Coordinates	44
3.3.2	Centered Logratio (clr) Coefficients	45
3.3.3	Isometric Logratio (ilr) and Pivot Coordinates	48
3.3.4	Special Coordinate Systems: Generalization of Pivot Coordinates	52
3.3.5	Special Coordinate Systems: Symmetric Pivot Coordinates	54
3.3.6	Special Coordinate Systems: Balances	56
3.4	Examples	59
	References	67
4	Exploratory Data Analysis and Visualization	69
4.1	Descriptive Statistics of Compositional Data	69
4.2	Univariate Graphics	73
4.3	Bivariate Plotting	77
4.4	Multivariate Visualization	79
	References	82
5	First Steps for a Statistical Analysis	85
5.1	Distributions and Statistical Inference	85
5.1.1	Normality Testing	87
5.1.2	Statistical Inference in Coordinates.....	88
5.2	Classical and Robust Statistical Analysis	90
5.2.1	Univariate Location.....	91
5.2.2	Univariate Scale	91
5.2.3	Multivariate Location and Covariance.....	92
5.2.4	Center and Variation Matrix.....	93
5.3	Outlier Detection	94
5.3.1	Univariate Outliers	95
5.3.2	Multivariate Outliers	98
5.3.3	Interpretation of Multivariate Outliers.....	101
5.4	Example.....	103
	References	106
6	Cluster Analysis	107
6.1	Distance Measures and Dissimilarities	107
6.2	Hierarchical Clustering Methods	110
6.2.1	Agglomerative Clustering Algorithms.....	110
6.2.2	Tree Cutting.....	113
6.3	Partitioning Methods	114
6.4	Model-Based Clustering	117
6.5	Fuzzy Clustering	119

6.6	Clustering Parts: Q-Mode Clustering.....	119
6.7	Evaluation	122
6.8	Examples	124
	References.....	130
7	Principal Component Analysis	131
7.1	Introductory Remarks	131
7.2	Estimation of Principal Components	132
7.2.1	Estimation by SVD	132
7.2.2	Estimation by Decomposing the Covariance Matrix.....	135
7.3	Compositional Biplot	137
7.4	Examples	140
7.4.1	Representation of Principal Components in a Ternary Diagram	140
7.4.2	Example: Household Expenditures at EU Level	140
7.4.3	Example: Beer Data	143
7.4.4	Example with Two Different Compositions	144
7.4.5	Example for PCA Including External Non-compositional Variables	144
	References.....	148
8	Correlation Analysis	149
8.1	Correlation Measures	149
8.2	Relating Two Compositional Parts	151
8.3	Multiple Correlation	152
8.4	Correlation Between Groups of Compositional Parts	153
8.5	Examples	154
8.5.1	Example for Correlation Between Single Compositional Parts	154
8.5.2	Example for Multiple Correlation	157
8.5.3	Example for Correlation Between Groups of Compositional Parts	158
	References.....	162
9	Discriminant Analysis	163
9.1	Introductory Remarks	163
9.2	Bayes Discriminant Rule.....	165
9.3	Fisher Discriminant Rule	167
9.4	Examples	169
9.4.1	Example for LDA and QDA	169
9.4.2	Example for Fisher Discriminant Analysis.....	174
9.4.3	Example with Appropriate Evaluation of the Error Rate	174
	References.....	179

10	Regression Analysis	181
10.1	Introductory Remarks	181
10.2	Regression with Compositional Response	182
10.3	Regression with Compositional Covariates	186
10.3.1	Real Response	186
10.3.2	Compositional Response	188
10.4	Regression Within a Composition	189
10.5	Variable Selection	192
10.6	Robustness Issues	194
10.7	Examples	195
10.7.1	Example for Regression with Compositional Response	195
10.7.2	Example for Regression with Compositional Covariates and Real Response	197
10.7.3	Example for Regression with Compositional Covariates and Compositional Response	200
10.7.4	Example for Regression Within a Composition	201
	References	204
11	Methods for High-Dimensional Compositional Data	207
11.1	Specific Problems of High-Dimensional Compositions	207
11.2	Partial Least Squares for Regression and Classification	209
11.3	Marker Identification Using Pairwise Logratios	212
11.4	Principal Balances	215
11.5	Examples	216
11.5.1	Example for PLS for Two-Group Classification	216
11.5.2	Example for Marker Identification	220
	References	225
12	Compositional Tables	227
12.1	Motivation and Geometry	227
12.2	Independent and Interaction Parts of Compositional Tables	229
12.2.1	Decomposition of 2×2 Compositional Tables	231
12.2.2	Coordinate Representation of Compositional Tables	233
12.3	Extension to the General Case	234
12.4	Examples	236
12.4.1	Gender Based Cancer Data	237
12.4.2	Social Expenditures According to Funding Sources	239
	References	242
13	Preprocessing Issues	245
13.1	Specific Problems with Data Preprocessing of Compositions	245
13.2	Missing Values	248
13.2.1	k-Nearest Neighbor (knn) Imputation	249
13.2.2	Iterative Model-Based Imputation	252

13.3 Rounded and Count Zeros 254

 13.3.1 Rounded Zeros..... 255

 13.3.2 Count Zeros 259

13.4 Rounded Zeros in High-Dimensional Data 260

13.5 Structural Zeros 263

References..... 270

Software Versions Used in the Book 273

Bibliography 275

Index 277

Akaike information criterion
 Bayesian information criterion
 Backward stepwise
 Comprehensive R Archive Network
 Forward stepwise
 Point and click graphical user interface
 Linear discriminant analysis
 Least squares
 Maximum covariance estimator
 Principal component analysis
 Principal component analysis
 Principal component analysis
 Principal sum of squares
 Recursive binary partitioning
 Singular value decomposition
 Total sum squares