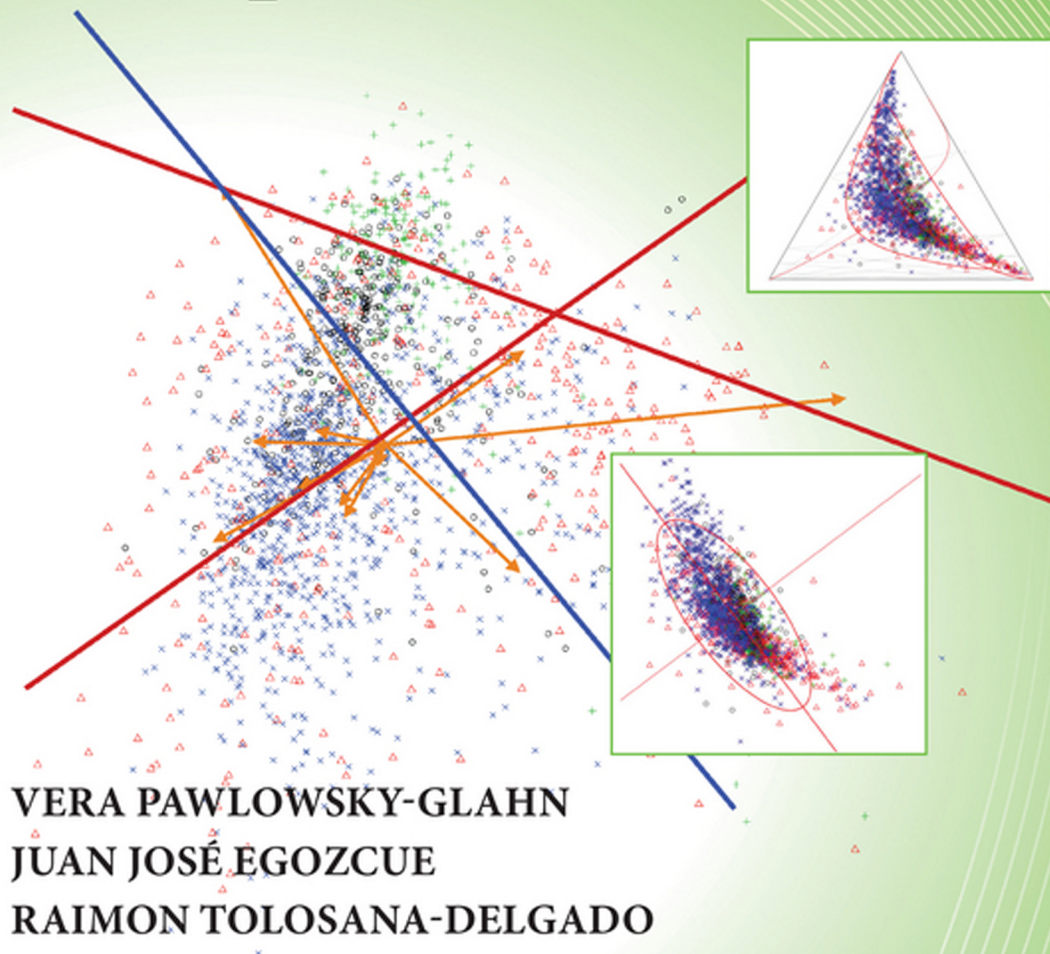


Modeling and Analysis of Compositional Data



VERA PAWLOWSKY-GLAHN
JUAN JOSÉ EGOZCUE
RAIMON TOLOSANA-DELGADO

STATISTICS IN PRACTICE

WILEY

Modeling and Analysis of Compositional Data

STATISTICS IN PRACTICE

Series Advisors

Human and Biological Sciences

Stephen Senn

CRP-Santé, Luxembourg

Earth and Environmental Sciences

Marian Scott

University of Glasgow, UK

Industry, Commerce and Finance

Wolfgang Jank

University of Maryland, USA

Founding Editor

Vic Barnett

Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceuticals; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Modeling and Analysis of Compositional Data

Vera Pawlowsky-Glahn

University of Girona, Spain

Juan José Egozcue

Technical University of Catalonia, Spain

Raimon Tolosana-Delgado

*Helmholtz Institut Freiberg for Ressources
Technology, Germany*

WILEY

This edition first published 2015
© 2015 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Pawlowsky-Glahn, Vera.

Modelling and analysis of compositional data / Vera Pawlowsky-Glahn, Juan José Egozcue, Raimon Tolosana-Delgado.

pages cm

Includes bibliographical references and indexes.

ISBN 978-1-118-44306-4 (cloth)

1. Multivariate analysis. 2. Mathematical statistics. 3. Geometric analysis. I. Egozcue, Juan José, 1950- II. Tolosana-Delgado, Raimon. III. Title.

QA278.P39 2015

519.5'35-dc23

2014043243

A catalogue record for this book is available from the British Library.

ISBN: 9781118443064

Set in 10.5/12.5pt, Times-Roman by Laserwords Private Limited, Chennai, India

We cannot solve our problems
with the same thinking
we used when we created them.

Albert Einstein

Eppur si muove

Galileo Galilei

Contents

Preface	xi
About the Authors	xv
Acknowledgments	xix
1 Introduction	1
2 Compositional Data and Their Sample Space	8
2.1 Basic concepts	8
2.2 Principles of compositional analysis	12
2.2.1 Scale invariance	12
2.2.2 Permutation invariance	15
2.2.3 Subcompositional coherence	16
2.3 Zeros, missing values, and other irregular components	16
2.3.1 Kinds of irregular components	16
2.3.2 Strategies to analyze irregular data	19
2.4 Exercises	21
3 The Aitchison Geometry	23
3.1 General comments	23
3.2 Vector space structure	24
3.3 Inner product, norm and distance	26
3.4 Geometric figures	28
3.5 Exercises	30
4 Coordinate Representation	32
4.1 Introduction	32
4.2 Compositional observations in real space	33
4.3 Generating systems	33
4.4 Orthonormal coordinates	36
4.5 Balances	38
4.6 Working on coordinates	43

4.7	Additive logratio coordinates (alr)	46
4.8	Orthogonal projections	48
4.9	Matrix operations in the simplex	54
4.9.1	Perturbation-linear combination of compositions	54
4.9.2	Linear transformations of S^D : endomorphisms	55
4.9.3	Other matrix transformations on S^D : nonlinear transformations	57
4.10	Coordinates leading to alternative Euclidean structures	59
4.11	Exercises	61
5	Exploratory Data Analysis	65
5.1	General remarks	65
5.2	Sample center, total variance, and variation matrix	66
5.3	Centering and scaling	68
5.4	The biplot: a graphical display	70
5.4.1	Construction of a biplot	70
5.4.2	Interpretation of a $2D$ compositional biplot	72
5.5	Exploratory analysis of coordinates	76
5.6	A geological example	79
5.7	Linear trends along principal components	85
5.8	A nutrition example	89
5.9	A political example	96
5.10	Exercises	100
6	Random Compositions	103
6.1	Sample space	103
6.1.1	Conventional approach to the sample space of compositions	105
6.1.2	A compositional approach to the sample space of compositions	106
6.1.3	Definitions related to random compositions	107
6.2	Variability and center	108
6.3	Probability distributions on the simplex	112
6.3.1	The normal distribution on the simplex	114
6.3.2	The Dirichlet distribution	121
6.3.3	Other distributions	127
6.4	Exercises	128
7	Statistical Inference	130
7.1	Point estimation of center and variability	130
7.2	Testing hypotheses on compositional normality	135
7.3	Testing hypotheses about two populations	136
7.4	Probability and confidence regions for normal data	142

7.5	Bayesian estimation with count data	144
7.6	Exercises	147
8	Linear Models	149
8.1	Linear regression with compositional response	150
8.2	Regression with compositional covariates	156
8.3	Analysis of variance with compositional response	160
8.4	Linear discrimination with compositional predictor	163
8.5	Exercises	165
9	Compositional Processes	172
9.1	Linear processes	173
9.2	Mixture processes	176
9.3	Settling processes	178
9.4	Simplicial derivative	183
9.5	Elementary differential equations	186
9.5.1	Constant derivative	187
9.5.2	Forced derivative	189
9.5.3	Complete first-order linear equation	194
9.5.4	Harmonic oscillator	200
9.6	Exercises	204
10	Epilogue	206
	References	211
	Appendix A Practical Recipes	222
A.1	Plotting a ternary diagram	222
A.2	Parameterization of an elliptic region	224
A.3	Matrix expressions of change of representation	226
	Appendix B Random Variables	228
B.1	Probability spaces and random variables	228
B.2	Description of probability	232
	List of Abbreviations and Symbols	234
	Author Index	237
	General Index	241

Preface

This book is an illustration of the adage collected by Thomas Fuller in *Gnomologia* (1732, Adage 560): *All things are difficult, before they are easy* and cited by John Aitchison (1986, Chapter 3). It has been a long way to arrive at this point, and there is still a long and not always easy way to go in the light of the insights presented here. Therefore, we dedicate this work to all those researchers who are not mainstream and have to struggle swimming against the tide.

These pages are based on lecture notes originally prepared as support to a short course on compositional data analysis. The first version of the notes dates back to the year 2000. Their aim was to transmit the basic concepts and skills for simple applications, thus setting the premises for more advanced projects. The notes were updated over the years, reflecting the evolution of our knowledge about the geometry of the sample space of compositional data. The recognition of the role of the sample space and its algebraic-geometric structure has been essential in this process. This book reflects the state of the art at the beginning of the year 2014. Its aim is still to introduce the reader into the basic concepts underlying compositional data analysis, but it goes far beyond an introductory text, as it includes advanced geometrical and statistical modeling. One should also be aware that the theory presented here is a field of active research. Therefore, the learning process can just start with this book, and a study of the latest contributions presented at meetings and as articles in journals is strongly recommended.

The book relies heavily on the monograph “*The Statistical Analysis of Compositional Data*” by John Aitchison (1986) and on posterior fundamental developments that complement the theory developed there, mainly those by Aitchison (1997), Barceló-Vidal et al. (2001), Billheimer et al. (2001), Pawłowsky-Glahn and Egozcue (2001, 2002), Aitchison et al. (2002), Egozcue et al. (2003), Pawłowsky-Glahn (2003), Egozcue and Pawłowsky-Glahn (2005), and Mateu-Figueras et al. (2011). Specific literature for other aspects of compositional analysis is given in the corresponding chapters. Chapter 1 gives a brief overview of the history of these developments and presents some everyday examples to illustrate the need of compositional data analysis. Chapter 2 defines compositions and their characteristics and introduces their sample space,

the simplex. Zeros and other irregular components are addressed in Section 2.3. On the basis of these considerations, Chapter 3 presents the Aitchison geometry of the simplex, while Chapter 4 gathers several ways to represent compositional data within this geometry. These four chapters form the algebraic-geometric body of the book, the backbone of the rest of the material.

Chapter 5 deals with exploratory analysis techniques adapted to compositions. Chapter 6 covers some distribution models for random compositions, as well as some required elements of probability theory. In particular, the latter chapter includes the normal distribution on the simplex, essential for the following two chapters. They are devoted to advanced statistical modeling: Chapter 7 provides some tools for testing compositional hypotheses (numerically and graphically), while Chapter 8 focuses on linear models, including regression, analysis of variance, and discriminant analysis. The last two chapters give an overview of what lies beyond this book: Chapter 9 outlines several compositional models besides the linear model, while the epilogue (Chapter 10) summarizes the ongoing and open aspects of research, as well as further topics, too specific to deserve longer attention in a general-purpose book.

Readers should take into account that, for a thorough understanding of compositional data analysis, a good knowledge in standard univariate statistics, basic linear algebra, and calculus, complemented with an introduction to applied multivariate statistical analysis, is a must. The specific subjects of interest in multivariate statistics, developed under the assumptions that the sample space is the real space with the usual Euclidean geometry, can be learned in parallel from standard textbooks, for instance, Krzanowski (1988) and Krzanowski and Marriott (1994) (in English), Fahrmeir and Hamerle (1984) (in German), or Peña (2002) (in Spanish). Thus, the intended audience goes from advanced students in applied sciences to practitioners, although the original lecture notes proved to be useful for statisticians and mathematicians as well. Newcomers to the field may find specially useful to start with Chapters 1–3, then read the first five sections of Chapter 4 and switch to Chapters 5 and 7 before finishing up Chapter 4. Applied practitioners already familiar with the basics of compositional data analysis should have a look at the notation and concepts in Chapters 4 and 6, before passing to the modeling Chapters 7–9. This book includes an extensive list of references, two appendices with practical recipes and some basic elements of random variables, a list of the symbols used in the book, and two indices; an author index and a general index. In the latter, pages in boldface indicate the point where the corresponding concept is defined.

Concerning notation, it is important to note that, to conform to the standard praxis of registering multivariate observations as a matrix where each row is an observation or data point and each column is a variate, vectors will be considered as row vectors (denoted by square brackets) to make the transfer from

theoretical concepts to practical computations easier. Furthermore, as a general rule, theoretical parameters will be denoted by either Latin or Greek letters and their estimators by the same letters with a hat.

Throughout the book, examples are introduced to illustrate the concepts presented. The end of each example is indicated with a diamond suit (\diamond).

Most chapters end with a list of exercises. They are formulated in such a way that many can be solved using an appropriate software. CoDaPack is a user friendly, cross-platform, freeware to facilitate this task, which can be downloaded from the web. Details about this package can be found in Thió-Henestrosa and Martín-Fernández (2005) or Thió-Henestrosa et al. (2005). Those interested in working with R (or S-plus) may use the packages “compositions” by Boogaart and Tolosana-Delgado (2005, 2013) in general or “robCompositions” by Templ et al. (2011) for robust compositional data analysis, as well as their common graphical user interface “compositionsGUI” by Eichler et al. (2013).

Vera Pawlowsky-Glahn
Juan José Egozcue
Raimon Tolosana-Delgado