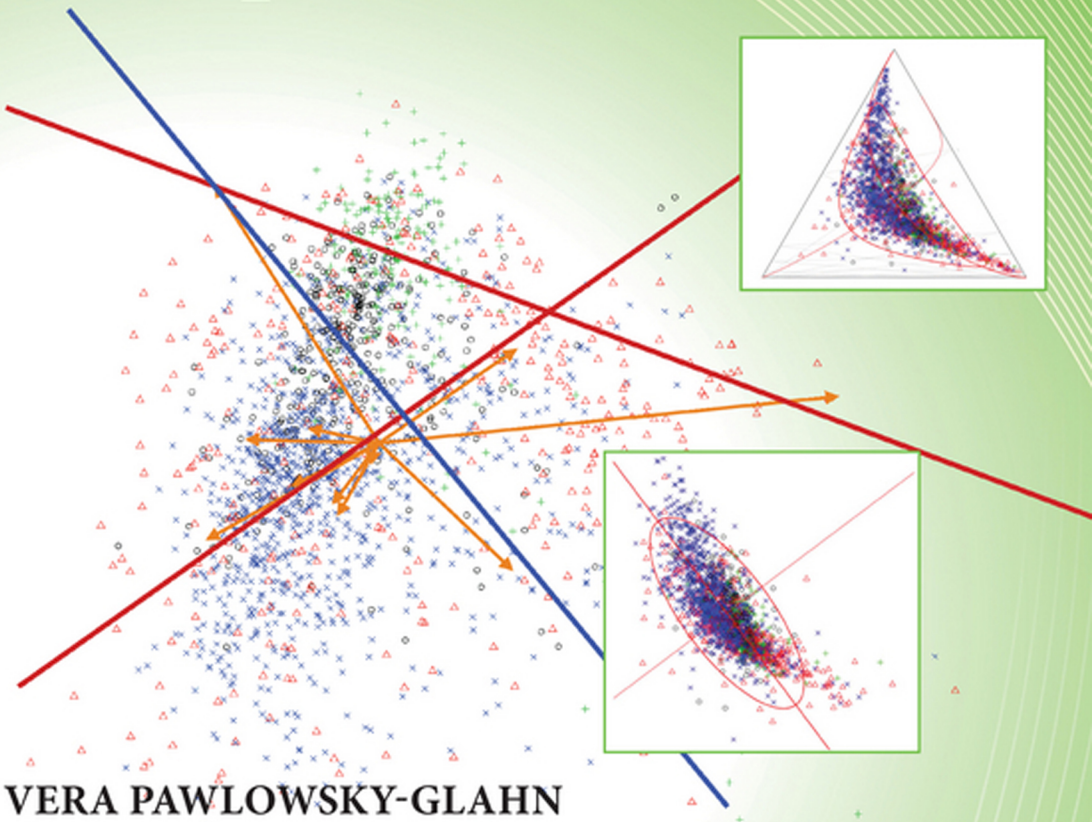
Modeling and Analysis of Compositional Data



VERA PAWLOWSKY-GLAHN
JUAN JOSÉ EGOZCUE
RAIMON TOLOSANA-DELGADO

STATISTICS IN PRACTICE

WILEY

Modeling and Analysis of Compositional Data

STATISTICS IN PRACTICE

Series Advisors

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 pharmaceutics; 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

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

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

1 2015

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

Albert Einstein

Eppur si muove

Galileo Galilei

Contents

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), Pawlowsky-Glahn and Egozcue (2001, 2002), Aitchison et al. (2002), Egozcue et al. (2003), Pawlowsky-Glahn (2003), Egozcue and Pawlowsky-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