Geometric Data Analysis

Geometric Data Analysis

From Correspondence Analysis to Structured Data Analysis

Brigitte Le Roux

MAPS 5 (CNRS) Department of Mathematics and Computer Science,
Université René Descartes, Paris, France

and

Henry Rouanet

CRIP 5 Department of Mathematics and Computer Science,
Université René Descartes, Paris, France

KLUWER ACADEMIC PUBLISHERS

NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 1-4020-2236-0
Print ISBN: 1-4020-2235-2

Created in the United States of America

Visit Springer's eBookstore at: http://ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

Contents

Foreword

Geometric Data Analysis (GDA) is the name I have proposed to designate the approach to Multivariate Statistics initiated by Benzécri as Correspondence Analysis, an approach that has become more used and appreciated over the years.

After numerous working sessions with Brigitte Le Roux and Henry Rouanet, both in Paris and in Stanford, it was evident that they were highly qualified to write a reference book about GDA that should meet the following two requirements: first, present in full the formalization of GDA in terms of the structures of linear algebra, which is an essential part of the mathematical foundations; and second, show how conventional statistical methods applicable to structured data analysis, i.e., analysis of variance and statistical inference, can be used in conjunction with GDA.

The richness of the actual content of the book they have written far exceeds these requirements. For example, Chapter 9, Research Case Studies, is nearly a book in itself. It presents the methodology in action with three well chosen extensive applications, one from medicine, one from political science, and one from education. The authors have taken time and effort to make this book accessible to a wide audience of practicing scientists. The mathematical framework is carefully explained. It is an important and much needed contribution to the statistical use of geometric ideas in the description and analysis of scientific data.

PATRICK SUPPES

Stanford, California
February, 2004

Preface

In our computer age, all research areas are replete with massive and complex data sets. Statistical packages offer myriads of methods for processing "multivariate data". The problem has now become: Which statistical method to choose, to make sense of data in the most meaningful way?

To say that "reality is multidimensional" is a truism. Yet, statistical thinking remains permeated with an ideology for which — alleging that "everything that exists, exists in some amount" — doing scientific work means quantifying phenomena. The achievements of this approach often fall short of promises. Indeed, the "reduction to unidimensionality" is sometimes so futile that it leads some good minds to the wholesale rejection of any statistical analysis, as reflected in sentences like this: "Intelligence is multidimensional, therefore it cannot be measured."

Beyond the opposition "quality" vs "quantity", there is *geometry,* whose objects (points, lines, planes, geometric figures) may be described by numbers, but are not reducible to numbers. Geometric thinking in statistics, with the idea that for transmitting information, a good picture may be more efficient than lots of numbers, is probably as old as statistics itself, and is historically traceable with the advent of scatter diagrams, charts and pictorial representations of statistical results. In the computer age, to meet the multidimensionality challenge, a more elegant way than a sterile retreat to a "qualitative approach" is offered by "l'Analyse des Données": the approach of multivariate statistics that Jean-Paul Benzécri, the geometer–statistician, initiated in the 1960s, and that we call *Geometric Data Analysis* (GDA)[1].

To cope with multivariate data, GDA consists in modeling data sets as *clouds of points* in multidimensional Euclidean spaces, and in basing the interpretation of data on the clouds of points. Clouds of points are

[1] The name *"Geometric Data Analysis",* which marks the unique thrust of the approach, was suggested to us by Patrick Suppes.

not ready–made geometric objects, they are constructed from data tables, and the construction is based on the mathematical structures of abstract linear algebra. The formalization of these structures is an integral part of the approach; properly speaking, GDA is the *formal–geometric approach* of multivariate data analysis. At the same time, *clouds of points are not mere graphical displays,* like temperature charts (where coordinate scales may be changed arbitrarily); they have a well–defined distance scale, like geographic maps.

Why a new book? Since the 1970s, Geometric Data Analysis has enjoyed a sustainable success in France, where "Analyse des Données" is taught both in statistics and in applied research departments, from biometry to economics and social sciences. In the international scientific community, Correspondence Analysis (CA) (the "leading case" of GDA), has been appreciated more and more widely over the years. The phrase *Correspondence Analysis* is now well–rooted, and CA is renowned as a powerful method for visualizing data. Yet GDA, as a comprehensive set of methods for multivariate statistics, remains largely to be discovered, both from the theoretical and practical viewpoints. Accordingly, the following topics have been emphasized in this book.

- *Formalization,* which is the most valuable guide at the crucial stages of the construction of clouds and of the determination of principal axes.
- *Aids to interpretation,* which are indispensable constituents of GDA.
- *Multiple Correspondence Analysis,* which is so efficient for analyzing large questionnaires.
- *Structured Data Analysis,* a synthesis of GDA and analysis of variance.
- *Integration of statistical inference* into GDA.
- *Full size research studies* (the largest chapter of the book), detailing the strategy of data analysis.

This book should thus provide a reference text to all those who use or/and teach Multivariate Statistics, as well as to mathematics students interested in applications, and applied science students specialized in statistical analysis.

The *mathematical prerequisites* are essentially some acquaintance with linear algebra; the specific background gathered in the Mathematical Bases chapter should render the book self–contained in this respect. There are *no statistical inference prerequisites!* Inference procedures are only used in the Research Case Studies chapter, and their principles recalled in the preceding chapter.

About the Authors and Acknowledgements

Brigitte Le Roux is Maître de Conférences at the Laboratoire de Mathématiques Appliquées de Paris 5 (MAP5), Université René Descartes and CNRS, Paris. E–mail: `lerb@math-info.univ-paris5.fr`.
Henry Rouanet is guest researcher at the Centre de Recherches en Informatique de Paris 5 (CRIP5), Université René Descartes, Paris. E–mail: `rouanet@math-info.univ-paris5.fr`.

The authors gratefully acknowledge the support of the laboratories (MAP5 and CRIP5) to which they belong.

This book has profited from our teaching experience over the years at the Department of 'Mathématiques & Informatique' of our University; we especially thank Dominique Seret, the Head of Department, who supported the book project by arranging a six–month sabbatical leave for B. Le Roux.

We are also grateful to our mathematician colleagues, especially Pierre Cazes (University Paris–Dauphine), who scrutinized the entire manuscript, and Bernard Bru (University René Descartes), for many discussions.

Above all, our gratitude goes to the experts thanks to whom we made so much progress in the theory and practice of Geometric Data Analysis: the late Pierre Bourdieu (Collège de France), and Frédéric Lebaron (Centre de Sociologie Européenne); Jean Chiche and Pascal Perrineau (Sciences Politiques, CEVIPOF); Werner Ackermann (Centre de Sociologie des Organisations); Jeanine Louis–Sylvestre and Michèle Chabert (Laboratoire de Physiologie du comportement alimentaire, EPHE).

We also wish to thank, for their remarks, Philippe Bonnet, Jean–Luc Durand and Pierre Vrignaud. Special thanks are due to Geneviève Vincent (Language Center of our University), for assistance in English.

Our utmost gratitude goes to Patrick Suppes (Stanford University), whose energetic encouragements have been the efficient cause of this book; the EPGY case study in the book is a reflection of the collaborative work undertaken with him.

We thank Jean–Paul Benzécri who, during our friendly meetings in the Loire Valley, expressed his blessings for our enterprise.

Last but not least, we thank Paul Roos, James Finlay, Inge Hardon and Anneke Pot of Kluwer, for their helpful attention and smiling patience.

BRIGITTE LE ROUX & HENRY ROUANET

Paris
December 28, 2003

[1] The book has been composed in LaTeX; our thanks go to AsTeX association and especially to Michel Lavaud (University of Orléans).