Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Offprint

Preface

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

Preface

The world around us – and especially the biological world – is inherently multidimensional. Biological diversity is the product of the interaction between many species, be they marine, plant or animal life, and of the many limiting factors that characterize the environment in which the species live. The environment itself is a complex mix of natural and man-induced parameters: for example, meteorological parameters such as temperature or rainfall, physical parameters such as soil composition or sea depth, and chemical parameters such as level of carbon dioxide or heavy metal pollution.

The properties and patterns we focus on in ecology and environmental biology consist of these many covarying components. Evolutionary ecology has shown that phenotypic traits, with their functional implications, tend to covary due to correlational selection and trade-offs. Community ecology has uncovered gradients in community composition. Much of this biological variation is organized along axes of environmental heterogeneity, consisting of several correlated physical and chemical characteristics. Spectra of functional traits, ecological and environmental gradients, all imply correlated properties that will respond collectively to natural and human perturbations. Small wonder that scientific inference in these fields must rely on statistical tools that help discern structure in datasets with many variables (i.e., multivariate data). These methods are comprehensively referred to as multivariate analysis, or multivariate statistics, the topic of this book. Multivariate analysis uses relationships between variables to *order* the objects of study according to their collective properties, that is to highlight spectra and gradients, and to *classify* the objects of study, that is to group species or ecosystems in distinct classes each containing entities with similar properties.

Although multivariate analysis is widely applied in ecology and environmental biology, also thanks to statistical software that makes the variety of methods more accessible, its concepts, potentials and limitations are not always transparent to practitioners. A scattered methodological literature, heterogeneous terminology, and paucity of introductory texts sufficiently comprehensive to provide a methodological overview and synthesis, are partly responsible for this. Another reason is the fact that biologists receive a formal quantitative training often limited to univariate, parametric statistics (regression and analysis of variance), with some

exposure to statistical modelling. In order to provide a training opportunity that could compensate for this, we collaborated on an introductory, intensive workshop in multivariate analysis of ecological data, generously supported and hosted several times by the BBVA Foundation in Madrid, Spain. The material for the workshop, consisting of lectures and practical sessions (R being our choice of software for the daily practicals) developed out of a graduate and postgraduate course at the University of Tromsø, Norway, now in its tenth year. Further intensive courses for professional ecologists and biologists were given at research institutions and universities in Iceland, Norway, United Kingdom, Italy and South Africa.

The aim of the material, developed for the various teaching and training purposes, refined, expanded and organized in this book, was always to provide the practitioner with the necessary tools to (i) choose the appropriate method for a given problem and data, (ii) implement the method correctly with the help of a computer, (iii) interpret the results with regard to the question asked, and (iv) clearly communicate the results and interpretation with the help of graphical illustrations. The last point about the importance of publishing quantitative results has been an emphasis of ours. As the ecologist Robert MacArthur has put it, "you have a choice, you can either keep up with the literature or you can contribute to it". For your scientific contribution to be effective, quantitative results and their interpretation must be presented in an understandable and accessible way.

The book, aimed at graduate and post-graduate students and professional biologists, is organized in a series of topics, labelled as parts consisting of multiple chapters, reflecting the sequence of lectures of our courses. The background for understanding multivariate methods and their applications is presented in the first introductory part, summarizing the character of ecological data and reviewing multivariate methods. The second part defines the basic concepts of distance and correlation measures for multivariate data, measuring inter-sample and inter-variable relationships. Initial approaches to analysing multivariate data are given in the third part, in the form of clustering and multidimensional scaling, both of which visualize these relationships in a fairly simple way. The fourth part introduces the core concept of the biplot, which explains how a complete data set can be explored using well-known ideas of linear regression and geometry, leading up to the method of principal component analysis. The fifth part is devoted to correspondence analysis and the related method of log-ratio analysis, and ending with canonical correspondence analysis, one of the key methodologies in ecology, which attempts to relate multivariate biological responses to multivariate environmental predictors. The sixth part is dedicated to aids to interpretation of results, statistical inference, and modelling, including an introduction to permutation testing and bootstrapping for the difficult problem of hypothesis testing in the multivariate context. Throughout the book the methods are illustrated using

small to medium-sized data sets. The seventh and last part of the main text of the book consists of two case studies that apply the above multivariate techniques to larger data sets, where the reader can see the challenge for analysis, interpretation and communication when dealing with large studies and complex designs. Finally, three appendices are provided on theoretical, bibliographical and computational aspects. All the analyses presented in the book can be replicated using the R scripts and data sets that are available on the website www.multivariatestatistics.org.

All the above topics contain material accessible to graduate students and practitioners with a basic statistical training. To make the material accessible we relied on the more visual (geometric) and intuitive aspects of the subjects. But chapters, and sections within chapters, do vary with regard to technicality and difficulty. A suggestion for the reader unfamiliar with multivariate methods is to first focus on the more general, accessible sections of the book, respecting the suggested sequence of parts and chapters, and wait before dwelling into the deeper, more technical layers of explanation upon a second reading. With some basic exposure to multivariate methods, the text can also be used as a handbook, with individual core chapters covering the rationales, strengths and weaknesses of various standard methods of interest, and providing illustrations (based on case studies) of appropriate visualization and pertinent interpretation of results.

Our most sincere gratitude goes to the colleagues and institutions that have hosted and helped organize our workshops and courses. First and foremost, we thank the BBVA Foundation and its director, Prof. Rafael Pardo, for continual support and assistance from their dedicated and friendly staff, who are now helping us further to publish the book that summarizes it all. A special thanks goes to the University of Tromsø, which has helped us maintain our course over a prolonged period. Many other institutions have provided help during the planning and running of our intensive courses. The Marine Research Institutes of Iceland and Norway, the University of Lancaster, UK, the Universities of Stellenbosch and Potchefstroom in South Africa, the Universities of Parma and Ancona, Italy, and the Italian Ecological Society.

Many colleagues have helped, directly or indirectly, with the preparation of this book: Giampaolo Rossetti, for his hospitality and support in Parma; Michaela Aschan, Maria Fossheim, Magnus Wiedmann, Grégoire Certain, Benjamin Planque, Andrej Dolgov, Edda Johannesen and Lis Lindal Jørgensen, our co-workers on the Barents Sea Ecosystem Resilience project *(BarEcoRe)* of the Norwegian Research Council; Paul Renaud, Sabine Cochrane, Michael Carroll, Reinhold Fieler and Salve Dahle, colleagues from Akvaplan-niva in Tromsø; Janne Søreide, Eva Leu, Anette Wold and Stig Falk-Petersen, for interesting discussions on fatty acid compositional data; and our families, for their patience and support.

Beginning with our first course at the University of Tromsø in 2004 and our first workshop at the BBVA Foundation in 2005, which developed into courses and workshops in six countries, we have had the privilege of a continuous exposure to insightful comments and friendly exchange with over 500 attendants sharing our passion for science and the environment. To all of them go our sincere gratitude and hope for a long and rewarding career. To us remains the intense satisfaction of interacting with motivated and enthusiastic people. We hope for more, thanks to this book.

Michael Greenacre
Raul Primicerio
Barcelona and Tromsø
November 2013