Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Preface Offprint

Preface

First published: September 2010
ISBN: 978-84-923846-8-6

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

In memory of Cas Troskie
(1936-2010)

Preface

Communicating and understanding data

In almost every area of research where numerical data are collected, databases and spreadsheets are being filled with tables of numbers and these data are being analyzed at various levels of statistical sophistication. Sometimes simple summary methods are used, such as calculating means and standard deviations of quantitative variables, or correlation coefficients between them, or counting category frequencies of discrete variables or frequencies in cross-tabulations. At the other end of the spectrum, advanced statistical modelling is performed, which depends on the researcher's preconceived ideas or hypotheses about the data, or often the analytical techniques that happen to be in the researcher's available software packages. These approaches, be they simple or advanced, generally convert the table of data into other numbers in an attempt to condense a lot of numerical data into a more palatable form, so that the substantive nature of the information can be understood and communicated. In the process, information is necessarily lost, but it is tacitly assumed that such information is of little or no relevance.

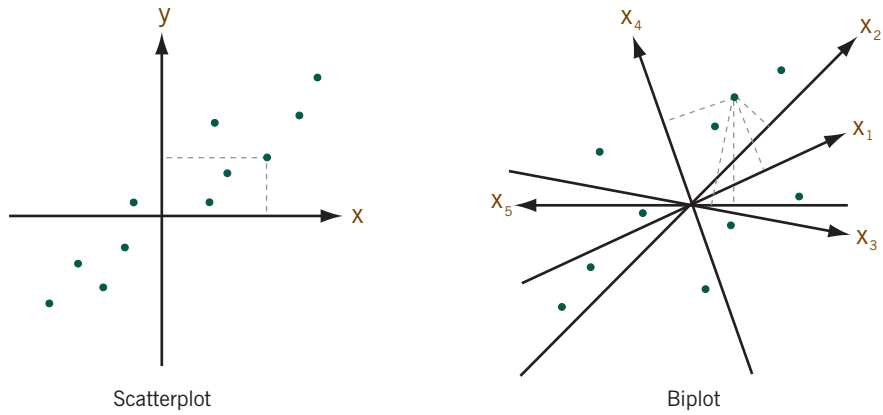
Graphical display of data

Graphical methods for understanding and interpreting data are another form of statistical data analysis; for example, a histogram of a quantitative variable or a bar chart of the categories of a discrete variable. These are usually much more informative than their corresponding numerical summaries—for a pair of quantitative variables, for example, a correlation is a very coarse summary of the data content, whereas a simple scatterplot of one variable against the other tells the whole story about the data. However, graphical representations appear to be limited in their ability to display all the data in large tables at the same time, where many variables are interacting with one another.

The biplot: a scatterplot of many variables

This book deals with an approach to statistical graphics for large tables of data which is intended to pack as much of the data content as possible into an easily digestible display. This methodology, called the *biplot*, is a generalization of the scatterplot of two variables to the case of many variables. While a simple scatterplot of two variables has two perpendicular axes, conventionally dubbed the horizontal x-axis and vertical y-axis, biplots have as many axes as there are variables, and these can take any orientation in the display (see Exhibit 0.1). In a scatterplot the cases, represented by points, can be projected perpendicularly onto the axes to read off their values on the two variables, and similarly in a biplot the cases

Exhibit 0.1:
A simple scatterplot of two variables, and a biplot of many variables. Green dots represent "cases" and axes represent "variables", labelled in brown



can be projected perpendicularly onto all of the axes to read off their values on all the variables, as shown in Exhibit 0.1. But, whereas in a scatterplot the values of the variables can be read off exactly, this is generally impossible in a biplot, where they will be represented only approximately. In fact, the biplot display is in a space of *reduced dimensionality*, usually two-dimensional, compared to the true dimensionality of the data. The biplot capitalizes on correlations between variables in reducing the dimensionality—for example, variables x and y in the scatterplot of Exhibit 0.1 appear to have high positive correlation and would be represented in a biplot in approximately the same orientation, like x_1 and x_2 in the biplot of Exhibit 0.1, where projections of the points onto these two axes would give a similar lining-up of the data values. On the other hand, the observation that variables x_3 and x_5 point in opposite directions probably indicates a high negative correlation. The analytical part of the biplot is then to find what the configuration of points and the orientations of the axes should be in this reduced space in order to approximate the data as closely as possible.

This book fills in all the details, both theoretical and practical, about this highly useful idea in data visualization and answers the following questions:

- How are the case points positioned in the display?
- How are the different directions of the variable axes determined?
- In what sense is the biplot an optimal representation of the data and what part (and how much) of the data is not displayed?
- How is the biplot interpreted?

Furthermore, these questions are answered for a variety of data types: quantitative data on interval and ratio scales, count data, frequency data, zero-one data and multi-category discrete data. It is the distinction between these various data types that defines most chapters of the book.

As in my book *Correspondence Analysis in Practice* (2nd edition), this book is divided into short chapters for ease of self-learning or for teaching. It is written with a didactic purpose and is aimed at the widest possible audience in all research areas where data are collected in tabular form.

After an introduction to the basic idea of biplots in Chapter 1, Chapters 2 and 3 treat the situation when there is an existing scatterplot of cases on axes defined by a pair of "explanatory" variables, onto which other variables, regarded as "responses" are added based on regression or generalized linear models. Here readers will find what is perhaps the most illuminating property of biplots, namely that an arrow representing a variable is actually indicating a regression plane and points in the direction of steepest ascent on that plane, with its contours, or isolines, at right-angles to this direction.

Chapter 4 treats the positioning of points in a display by multidimensional scaling (MDS), introducing the concept of a space in which the points lie according to a measure of their interpoint distances. Variables are then added to the configuration of points using the regression step explained before.

Chapter 5 is a more technical chapter in which the fundamental result underlying the theory and computation of biplots is explained: the singular value decomposition, or SVD. This decomposition provides the coordinates of the points and vectors in a biplot with respect to dimensions that are ordered from the most to the least important, so that we can select the reduced-dimensional space of our choice (usually two- or three-dimensional) that retains the major part of the original data.

Chapter 6 explains and illustrates the simplest version of the biplot of a cases-by-variables matrix in the context of principal component analysis (PCA), where the variables are measured on what are considered to be interval scales. Here the issue of biplot scaling is treated for the first time.

Chapter 7 deals with the lesser-known topic of log-ratio analysis (LRA), which deserves much more attention by data analysts. The variables in this case are all measured on the same scale, which is a positive, multiplicative scale, also called a ratio scale. All pairs of ratios within rows and within columns are of interest, on a logarithmic scale. The LRA biplot shows the rows and the columns as points, but it is really the vectors connecting pairs of rows or pairs of columns that are interpreted, since these link vectors depict the log-ratios.

Chapter 8 treats biplots in correspondence analysis (CA), which competes with log-ratio analysis as a method for analyzing data on a common ratio scale, espe-

cially count data. In contrast to LRA, CA easily handles zero values, which are common in many research areas where count data are collected, for example linguistics, archaeology and ecology.

Chapters 9 and 10 consider biplots in the display of large sets of multivariate categorical data, typically from questionnaire surveys, using variations of CA. Here there are two approaches: one is to consider associations between two different sets of variables, which are cross-tabulated and then concatenated (Chapter 9), while the other considers associations within a single set of variables (Chapter 10) —the latter case, called multiple correspondence analysis (MCA), has become one of the standard tools for interpreting survey data in the social sciences.

Chapter 11 focuses on discriminant analysis (DA) of grouped data, where the biplot displays group differences rather than differences between individual cases. Each group is represented by its mean point, or centroid, and it is these centroids that are optimally displayed in the biplot, along with the variables that contribute to their separation.

Chapter 12 is a variant of the dimension-reduction theme where the optimal reduced space is obtained subject to constraints on the solution in terms of additional explanatory variables. This idea has applications in all the versions of the biplot treated in the book: MDS, PCA, LRA, CA, MCA and DA. In the PCA context this constrained form is known as redundancy analysis (RDA) and in the CA context as canonical correspondence analysis (CCA).

Throughout the book there are illustrations of biplots in many different research contexts. It then concludes with three more detailed case studies in the biomedical, social and environmental sciences respectively:

- Analysis of a large data set in cancer research based on gene-expression arrays —using the PCA biplot and the DA biplot (or centroid biplot) to distinguish between four cancer types (Chapter 13).
- Analysis of data on several thousand respondents in a questionnaire survey on attitudes to working women—using CA and MCA biplots, and also constrained biplots to study the effect of different response categories (Chapter 14).
- Analysis of morphological and diet data of a sample of fish, to identify possible components of the diet that are related to the fish's morphology—using the LRA biplot with constraints (Chapter 15).

Computations in the R language

The biplots reported and discussed in the book are all computed in the open-source R environment and the Computational Appendix explains several of these analyses by commenting on the R code used.

Finally, the book concludes with a bibliography for further reading and online resources, a glossary of terms, and an epilogue in which some of my personal opinions are expressed about this area of statistics.

Readers will find a supporting website for this book at:

http://www.fbbva.es
http://www.multivariatestatistics.org

containing additional material such as the glossary and summaries of the material in Spanish, and the complete script file of the R code.

Dedication to Cas Troskie

This book is appropriately dedicated to Prof. Cas Troskie, former head of the Department of Statistical Sciences at the University of Cape Town (UCT), South Africa, and a maestro of theoretical and applied multivariate analysis. Cas was one of the most influential people in my statistical career. In fact when he visited me in Barcelona on several occasions I always introduced him as the reason behind my decision to do statistics as a major in my initial Bachelor of Science studies at UCT. As early as 1969, aged 33 and the youngest department head on the UCT campus, he was encouraging students like myself to write computer programs and put decks of punched cards into card readers linked to the university computer and wait expectantly for printouts to emerge with the results. He had a singular faith in principal components of a data set, which prepared me for my subsequent studies in France on correspondence analysis. I am not alone in being affected by his dynamic personality and sharp intelligence, since he inspired dozens of Masters and PhD theses, leaving a huge legacy to the statistical community, not only in South Africa but worldwide. One of his theoretical papers, co-authored with one of his PhD students, has been cited often in the electrical engineering literature and has made a significant impact in the design of MIMO (multiple input multiple output) wireless communications systems, which will form the cornerstone of most future wireless technologies.

Acknowledgements

This book owes its publishing to the BBVA Foundation and its Director, Prof. Rafael Pardo. One of the visions of the Foundation is to disseminate advanced educational material in a form that is easily accessible to students and researchers worldwide; hence this series of manuals on applicable research, attractively produced, distributed online for free and complemented by a supporting website with additional online material. For an academic it is like a dream come true to have such an outlet and I express my gratitude and appreciation to Prof. Pardo for including me in this wonderful project. Thanks are also due to the Foundation's publications director, Cathrin Scupin, for her continuing co-operation and support throughout the publishing process. Then there is the fantastic production team at Rubes Edi-

torial in Barcelona, Jaume Estruch, Núria Gibert and Imma Rullo, to whom I am equally grateful—they are responsible for the physical aspects of this book, expert copy-editing of the manuscript, and the design of the supporting website. Thanks are due to the Pompeu Fabra University in Barcelona, and for partial funding by the Spanish Ministry of Science and Technology grants MTM2008-00642 and MTM2009-09063. Finally, there are many friends who have supported me in this project—too many to list individually, but they know who they are!

So, if you have this book in your hands or are seeing this online, I wish you good reading, good learning and especially good biplotting!

Michael Greenacre
Barcelona, July 2010