CORRESPONDENCE ANALYSIS AND DATA CODING WITH JAVA AND R

Fionn Murtagh

Chapman & Hall/CRC, 2005.

MULTIDIMENSIONAL NONLINEAR DATA ANALYSIS

Shizuhiko Nishisato

Chapman & Hall/CRC, 2006.

Apart from the book on Multiple Correspondence Analysis, reviewed above, Chapman & Hall/CRC have published two other books recently that fall directly in the same area of Statistics, by Fionn Murtagh and Shizuhiko Nishisato respectively. While treating methods which rely on the same mathematical theory, these two books could not be more different, hence this comparative review.

Fionn Murtagh was a doctoral student of Jean-Paul Benzécri, and his book is imbued with Benzécri's teachings, ideas and philosophy. In my view this is the first English-language text that brings across the flavour of Benzécri's contributions to data analysis. Almost all statisticians would think of correspondence analysis (CA) as just another multivariate technique in their toolbox, which gives interesting visualizations of specific sets of categorical data. For Benzécri, however, CA is a core method of multivariate analysis, which can expose structure in any data that is suitably transformed, or *coded* – the method is really central to his thinking, as it was to the sociologist Pierre Bourdieu who popularized it amongst social scientists as a way to define social space. After 33 years of working in this field since my own doctoral studies with Benzécri in Paris, my opinion is that Benzécri is closer to being right than wrong: I often say to students that if they were stranded on a desert island with only one computer program, then it should be CA, because they could answer most of the questions they might have about a data set. The method provides a framework for investigating discrete and continuous structure, associations and relationships between variables.

Murtagh's book is packed with examples of the way data can be coded to feed into the CA algorithm, which leads to visualizations of the structure in the coded data.

Literally, what goes in comes out – this could be the motto of the book. After an historical introduction (Chapter 1), very much from the "Benzécrean" perspective, the mathematics of CA and cluster analysis is given. Here there is an unfortunate excursion (in my opinion), but just for 10 pages, into the over-complicated tensor notation used by Benzécri in his books, incomprehensible to most readers. This notation was something I had to learn to endure personally in my own studies in Paris and, in retrospect, I cannot understand why mathematical elegance should shroud what are basically simple mathematical concepts, much more easily understood using pragmatic matrix-vector notation. Anyway, fortunately for the readers, the rest of the book uses straightforward scalar notation and on some rare ocasions a matrix and a vector can be spotted (this is in strange contrast to the R program code throughout the book, which is necessarily constructed on data matrices and row and column vectors).

Chapter 3 is the core of the book, treating the various types of data coding, for example disjunctive coding (i.e., dummy variables), fuzzy coding (a less strict form of disjunctive coding, used for continuous-scale variables to conserve more information) and doubling (creating two variables for each original one, representing two extreme poles).

Chapter 4 includes five case studies, all of which involve data originally on continous scales – this demonstrates that this book is quite different from the the usual publications on CA.

Chapter 5 is exclusively devoted to longitudinal and textual analysis, which is fitting since all the early development of CA by Benzécri was in a linguistic context. As Murtagh says (page 161): "the way in which textual and document analysis is carried out in the correspondence analysis approach is quite different from other contemporary approaches". Linguists might not all agree, but at least here they will find this approach excellently presented and illustrated here.

One of the main features of the book is the support for practical application provided in the form of R code, which is also available on the author's website `www.correspondances.info`. In summary, this book is highly recommended as a text explaining Benzécri's ideas and giving many examples of the application of CA in non-standard situations. On the negative side, related work by other researchers is not even referred to, for example the work of Nishisato (e.g., Nishisato 1994), the Leiden group (e.g., Gifi 1990) and the two edited books by Greenacre & Blasius (1993) and Blasius & Greenacre (1998). The non-citing of Gifi (1990) is probably the most serious, since the work of Jan de Leeuw and co-workers is just as innovative in using the optimal scaling ideas inherent in CA as the basis for generalizing multivariate methods to categorical data, which fits very closely the general Benzécrean philosophy.

Shizuhiko Nishisato's book *Multidimensional Nonlinear Data Analysis* is so different from Murtagh's that, apart from the appearance of some maps of data with respect to component axes, one might think that this was a completely different subject. The first sentence, however, tells us that “multidimensional nonlinear data analysis” (MUNDA) is a family of methods for quantifying categorical data” and that “this procedure covers such methods as correspondence analysis, dual scaling, homogeneity analysis, quantification theory, optimal scaling and the method of reciprocal averages.” This book looks like a second edition of the book on dual scaling by Nishisato (1994) but now dual scaling is apparently subsumed in MUNDA. Later on page 51 it is stated that “correspondence analysis (is) one of the many names of MUNDA”. This rather confusing classification serves to confirm that, apart from a few academic nuances, all the methods mentioned above are really the same, and differ mainly for historical-cultural reasons.

Putting this aspect aside, Nishisato does an excellent job of putting the historical record straight, right up to the present time (see Chapter 3-Historical Overview, 17 pages long). In reading the rest of the book I was interested to see that Nishisato has now recognized all the geometric concepts inherent in the correspondence analysis approach, although his description of chi-square distances will leave readers baffled as to what it really is. I could find no definition of this distance except for an erroneous set of formulas on page 79 which expressed it in terms of the solution “weights” (or coordinates) rather than on the original data.

Subsequent chapters treat different data types and how the method deals with them. Chapter 6 deals with the analysis of “incidence data”, alias simple correspondence analysis. Chapter 7 deals with the analysis of “multiple choice data”, alias multiple correspondence analysis. Chapter 8 deals with “sorting data”, which is analyzed just like “multiple choice data”.

Chapter 9 is on “forced classification of incidence data”, an idea which is inherent in the method known as canonical correspondence analysis (CCA). The “forcing” is the same as the “constraining” or “restricting” of the solution to be linearly related to an external set of continuous or dummy variables, which splits the space into constrained and unconstrained parts (Nishisato calls the unconstrained dimensions the “conditional components”). This methodology is routinely used by ecologists and published in R software such as the vegan package by Jari Oksanen. In the introduction Nishisato apologizes for not treating CCA in his book, which is a pity since CCA includes forced classification if the constraining variables are categorical.

Chapters 10 to 12 are on the analysis of “dominance data”, i.e. paired comparisons, rank-order data and successive categories (or ratings) data. The analysis of these types of data can be performed equivalently using the doubling coding prior to CA. This fact is not referred to explicitly by Nishisato, but he does say (page 76) that “researchers in correspondence analysis...have recently derived formulations of correspondence analysis for dominance data as well, thus diminishing the initial differences between dual scaling and correspondence analysis, except for some details.”

The main problem with Nishisato's book is that there is no reference to computing and how to perform the methods that he describes. Theory is explained but not the algorithms for finding solutions, neither is software referred to or commented on. The main benefit of this book is the extensive reference list and Nishisato's comprehensive treatment of the history of this area of multivariate analysis.

Michael Greenacre
Universitat Pompeu Fabra

References

Blasius, J. and Greenacre, M. J. (1998). *Visualization of Categorical Data*. Academic Press, San Diego.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Greenacre, M. J. and Blasius, J. (1993). *Correspondence Analysis in the Social Sciences*. Academic Press, London.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Lawrence Erlbaum, New Jersey.