[book review] The Analysis of Proximity Data. B. S. Everitt and S. Rabe-Hesketh, Arnold, London, 1997.

Gilbert Saporta

STATISTICS IN MEDICINE
Statist. Med. **18**, 491–496 (1999)

BOOK REVIEWS
Editor: Niels Keiding

1. *B. S. Everitt and S. Rabe-Hesketh*, The Analysis of Proximity Data.

2. *Chap T. Le*, Applied Survival Analysis.

3. *David Machin, Michael J. Campbell, Peter M. Fayers and Alain P. Y. Pinol*, Sample Size Tables for Clinical Studies

4. *B. P. Butler, M. G. Cox, S. L. R. Ellison and W. A. Hardcastle* (*eds*), Statistics Software Qualification: Reference Data Sets.

1. THE ANALYSIS OF PROXIMITY DATA. B. S. Everitt and S. Rabe-Hesketh, Arnold, London, 1997. No. of pages: ix + 178. Price: £35. ISBN 0-340-67776-7

This small book presents a synthesis of various methods that can be applied to proximity data, namely multi-dimensional scaling and tree representations which are presented as spatial and non-spatial models respectively.

By proximity data, one usually understands a square symmetric matrix of distances, dissimilarities or similarities between n objects, but in addition this book offers a chapter devoted to the cases of asymmetric and rectangular data.

Chapter 1 is a short introduction and chapter 2 presents the concepts of similarity, dissimilarity and distances with various examples of such measures.

Chapter 3 deals with the usual methods of metric and non-metric multi-dimensional scaling which aim to represent the data as a set of points in a low dimensional space. Many examples illustrate this part. Three methods are developed. The first is the classical Gower–Torgerson scaling, or principal co-ordinate analysis, which is dual to principal component analysis; interpoint distances (which should be Euclidean) are converted to scalar products with the centroid at the origin of the space. The second consists in fitting, in the least squares sense, distances in the representation space to dissimilarities. The last method (non-metric multi-dimensional scaling) allows monotonic transformations of the dissimilarities.

The presentation is clear. However, in my opinion, the fact that classical scaling leads to a representation where distances are approximated from below should have been emphasized here, although it is pointed out in Appendix A. The reader may not notice that in classical scaling the distances between points in the representation space are always less than the original distances (since it is a projection) which is not the case for the other techniques. I also regret that on p. 28 the authors do not give the formulae for transforming non-Euclidean distances or dissimilarities into Euclidean distances by adding a constant, despite their simplicity.

One major problem when one has got a spatial representation is to interpret the figures. Chapter 4 presents briefly how to find interpretable axes and devotes a large part to the comparison of configurations by the so-called 'Procrustes rotations' for two and more solutions.

Chapter 5 presents methods for three-way data where one has several square symmetric dissimilarity matrices between the same objects but collected in different circumstances (judges or occasions). The INDSCAL model of Carroll and Chang[1] is presented in detail as well as rotated models. The first one supposes that judges are weighting differently the underlying dimensions of a common representation; the second allows in addition rotation of axes. The chapter ends with the presentation of the inferential approach of Ramsay where a model is assumed, and the parameters (that is, co-ordinates) are estimated by maximum likelihood; this allows the possibility of drawing confidence regions for points. Everitt and Rabe-Hesketh give a summary of the controversy about this approach. As inferential problems are concerned, it is surprising that resampling techniques are omitted, since they provide a distribution-free way to

assessing the stability of the solutions. The paper by Weinberg, Carroll and Cohen, nearly 15 years old, is of a great interest and deserved a quotation.[2]

More original in a book form is chapter 6, which collects various results, so far only sparsely available in journals, about non-symmetric matrices and rectangular tables. Rectangular tables are usually of the kind judges × ratings or preferences. One can find a very useful presentation of unfolding, vector models, ideal points models etc. used in psychometric and marketing literature.

The last chapter deals with another kind of representation of proximity data – trees – but not only the hierarchical trees well known in cluster analysis, called here ultrametric trees, but also additive trees. Of course this part could have been developed in further detail but there already exists many books about cluster analysis, including one by Everitt.

Appendix A presents some mathematical material about distances in multivariate analysis which was not included in the first chapters. Actually these 6 pages are very short and as far as mathematical material is involved, I would have appreciated references to recent mathematical developments which seem to have been ignored. The reader should know that distances and dissimilarity come from an active field and, for instance, that a European network of British, French, Dutch and Portuguese researchers has produced many contributions during the period 1983–1992. The international meeting DISTANCIA'92 in Rennes gathered about 200 people and the book *Classification and Dissimilarity Analysis*, edited by Van Custem,[3] presents the main theoretical contributions of this network.

Appendix B will be very useful for the practitioner since it presents a very comprehensive and up to date survey of available software for multi-dimensional scaling with electronic addresses.

The reference list is impressive (more than 200 references) with, however, some omissions (as already mentioned).

Finally, considering the hope written in the preface that the book 'be a helpful introduction to this area both for research workers who are not primarily statisticians (…) and for applied statisticians interested in the underlying methodology'. I think that it is fulfilled for the first group but only partly for the second. The main strengths of this book are undoubtedly the clarity of the exposition and the large number of examples.

GILBERT SAPORTA
Chair of Applied Statistics
Conservatoire National des Arts et Métiers
292 rue Saint Martin
F75003 Paris
France

REFERENCES

1. Arabie, P., Carroll, J. D. and DeSarbo, W. *Three Way Scaling and Clustering*, Sage University Papers, 1987.
2. Weinberg, S. L., Carroll, J. D. and Cohen, H. 'Confidence regions for INDSCAL using the jackknife and bootstrap techniques', *Psychometrika*, **49**, 475–491 (1984).
3. Van Custem, B. (ed.). *Classification and Dissimilarity Analysis*, Lecture Notes in Statistics 93, Springer-Verlag, 1994.

---

2. APPLIED SURVIVAL ANALYSIS. Chap T. Le, Wiley, New York, 1997. No. of pages: xiii + 257. Price: £34.95. ISBN 0-471-17085-2

According to the preface, 'this book is intended to meet the need of practitioners and students in applied fields for a single, fairly thin volume covering major, updated methods in the analysis of survival data'. It is based on a course at the University of Minnesota, and 'is written for the training of beginning graduate students in biostatistics, epidemiology, and environmental health, as well as for professional statisticians and biomedical research workers'. A number of books with similar aims have appeared over the last few years, so the question is whether this book provides a useful addition to the existing literature.

The book is divided into four chapters. Basic concepts are introduced in chapter 1, chapters 2 and 3 deal with estimation and testing in non-parametric and parametric models, while Cox regression is discussed in chapter 4. The derivations only use standard algebra and elementary calculus, but basic knowledge of statistical theory is assumed (including maximum likelihood estimation and related tests). A number of exercises are

