Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computational Statistics and Data Analysis 53 (2009) 3249–3251

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Book review

Brigitte Le Roux, Henry Rouanet. Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Kluwer Academic Publishers, Dordrecht, Boston, London (2005). 475 pp., (Hardcover) USD $ 171.00; €155.00, ISBN: 978-1-4020-2235-7

The subtitle of *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis* distinguishes this book from *Geometric Data Analysis* by Michael Kirby (Kirby, 2001) with subtitle *An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. Although both books deal with vector spaces and dimensionality reduction, the Kirby book is devoted to the most efficient methods of pattern analysis using wavelet decomposition. Henry Rouanet and Brigitte Le Roux as world leading specialists in the French "Analyse des Données" statistical tradition, founded by Jean-Paul Benzécri in the 1960s and the 1970s, view geometric data analysis (GDA) as the analysis of multivariate data sets represented as clouds of points in Euclidean space. Thus, their *Geometric Data Analysis* is based upon Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA) and versions of Euclidean Classification.

The core of GDA for Le Roux and Rouanet is outlined in their introductory chapter as the Euclidean cloud of data points originating from contingency tables or any individuals × variables multivariate data matrix together with the determination of principal axes and Euclidean classifications. They provide a unified conceptual framework for GDA including the role of structuring (grouping) variables and statistical inference methods. In GDA, "the model follows the data, not the reverse" according to the scheme: Data Table → Clouds of points in Euclidean space → Structured data → Inductive reasoning (Statistical Inference).

CA is introduced in Chapter 2 in a formal geometric way by explaining the mathematical structures underlying the method. Matrix notation is only used as shorthand and the reader is warned that a matrix approach to statistics, ignoring abstract linear algebra, is "simply not powerful enough to cope with geometric structures". It follows that a thorough understanding of abstract linear algebra is a prerequisite for a complete appreciation of the book. The necessary concepts of abstract linear algebra, multidimensional geometry and spectral decomposition are provided in Chapter 10, making the book self-contained in this respect. Chapter 2 should therefore be studied in tandem with Chapter 10. The basic notation used in the book together with necessary measure theory concepts are also introduced in Chapter 2. The distinction made between measures (functions where grouping of units entails summing of values) and variables (functions where grouping units entails averaging values) is fundamental to the construction of CA as is the concept of density of a measure. A subscript-superscript notation is introduced, maintaining the distinction between variables and measures. CA is presented in terms of a contingency table where the latter is regarded as the statistical paradigm of a measure over a Cartesian product. Two clouds of points – one associated with the rows of the contingency table and the other with the columns – are constructed and then these clouds are searched for their principal axes and variables leading to geometric representations of the clouds. Several aids to interpretation of CA – contributions of axes and points to the variance of the cloud; contributions of points to the variance of an axis; contributions of deviations to the variance of an axis; relative contributions and qualities of representation – are formally derived. A detailed example of the CA of a small contingency table helps the reader to a full understanding of the mathematical constructs underlying the method. The chapter is concluded by extensive exercises covering the material presented in the chapter together with some new results. Practitioners of CA would benefit from the detailed solutions given.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2008.01.012

The Euclidean cloud, the central object of GDA, is studied in detail in Chapter 3. The basic descriptive statistics for a cloud of points like its mean point, inertia, sum of squares, variance and contributions are defined. This is followed by a discussion of the orthogonal projection of a cloud onto a subspace and its breakdown into fitted and residual clouds. Attention is then directed to the partitioning of a cloud into subclouds leading to the concepts of between and within clouds together with their associated variances. Linear mappings are reviewed before tackling the problem of determining the principal directions and variables of the cloud together with issues like principal breakdown of distances, passage formulae and contributions. Specific analysis of the cloud, *i.e.* determining principal axes of a cloud under the constraint that they must belong to given prespecified subspaces receives also attention. A section is devoted to principal or inertia hyperellipsoids, including concentration hyperellipsoids. Euclidean classification – a form of cluster analysis with variance as aggregation index and in harmony with the mathematical structures of GDA – is introduced. Finally, it is shown how to choose a Cartesian framework to go from points to numbers in order to find matrix expressions for the direction and eigenvalues equations as well as the passage and reconstitution formulae that followed from abstract linear algebra. Working through the exercises at the end of the chapter and comparing your solutions to the extensive solutions offered, helps to obtain a sound understanding of the material.

In Chapter 4 principal component analysis (PCA) is considered as a GDA method by going from the numbers provided by an individuals × variables data matrix to a Euclidean cloud of points. This geometric approach to PCA enables researchers to study both the associations between variables and the proximities between the individual points (again with abstract linear algebra providing the mathematical structure of the PCA methodology). Both simple PCA (the analysis of covariances) and standard PCA (the analysis of correlations) are discussed before attention is directed to a biweighted protocol of measures. The extensive exercises with detailed solutions at the end of the chapter contain a wealth of information both for the theorist and the practitioner of PCA.

MCA is introduced by considering the language of a questionnaire (in standard form) where individuals must choose for each question one and only one response from a set of alternatives (modalities). Chapter 5 is therefore of particular importance for practitioners facing the problem of analysing questionnaire data. MCA provides a geometric model for such data by representing individuals by points and summarizing the relations between the categorized variables. As such, MCA is regarded as the counterpart for PCA for categorized variables. By defining a multiple correspondence measure, it is shown how to derive the principal axes and variables, the cloud of individuals and the cloud of modality points for an MCA. The merits of constructing concentration ellipses are emphasized. Introducing supplementary points and questions are dealt with together with optimal codings. In particular, two kinds of specific MCA are presented — restriction to a subset of modalities and analysis of a subcloud of individuals. Practitioners of MCA most certainly will find the detailed example with the many aids to interpretation extremely informative as well as the exercises where detailed solutions are discussed.

A short Chapter 6 deals with structured GDA — the inclusion of structuring (or grouping) variables describing the basic sets indexing the rows and columns of the input table. Since similar grouping variables are found in (multiple) analysis of variance, structured GDA can be thought of as a synthesis between GDA and analysis of variance. The experimental paradigm is catered for as well as observational data, underlining the wide applicability of GDA techniques in practice. A section is devoted to the “Analysis of comparisons”. Contrasts, nesting and crossing structures, additive and interaction clouds are also discussed.

The question about the sensitivity to data perturbations of the methods discussed so far for determination of the principal axes and variables of a Euclidean cloud, is studied in Chapter 7. Four such stability problems are considered: the effect of coding according to a partition, the influence of a group of points, the effect of a change in metric and the influence of a variable and/or a modality. These problems are approached by comparing the analysis of a reference Euclidean cloud to that of a cloud modified according to the particular stability problem concerned. Comparisons are in terms of the (geometric) structures provided by abstract linear algebra.

Le Roux and Rouanet are firm believers that inference is to be preceded by descriptive procedures (*i.e.* procedures that do not depend on sample size) leading to descriptive conclusions. For them inference procedures (significance tests, confidence intervals) attempt to extend descriptive conclusions. They do depend on sample size and lead to inductive conclusions. Current statistical inference in multivariate analysis is discussed in the first sections of Chapter 8. This is followed by a review of elementary univariate procedures in the line of inductive data analysis. Combinatorial inference is introduced as an extension of traditional inference; it is regarded as “the first stage of statistical inference, as the straight extension of descriptive statistics”. Of particular interest is the proposal of combinatorial procedures for nonrandom problems like typicality of a cloud, homogeneity and nonassociation tests.

Sections are devoted to permutation tests and Bayesian procedures, respectively. Inductive geometric procedures are discussed in a section under the heading "Inductive geometric data analysis" that includes geometric statistical modeling, permutation modelling and Bayesian data analysis. Finally, guidelines for inductive analysis are provided. The authors convincingly show that modern GDA methodology includes powerful inferential tools.

GDA as presented in the first eight chapters, is applied to three research case studies (Parkinson's disease study, political space study and education programme for gifted youth study) in Chapter 9 that almost forms a book on its own. Each study comprises an individuals × variables table together with structuring factors on the individuals. This chapter provides in depth training how to perform a GDA — from describing the initial clouds to a full inductive geometric analysis. Perhaps researchers in the applied sciences should start with this chapter — it should benefit their understanding of GDA greatly.

The book contains information on available software to perform GDA.

In summary, Le Roux and Rouanet have presented a comprehensive stand-alone work on GDA that might well turn out to become a cornerstone of GDA in the English literature. I would like to recommend this book most strongly to statisticians and users of statistics. The notation is challenging but the time spent in mastering it is worthwhile. The book deserves five stars for its formal treatment of GDA in terms of abstract linear algebra; its extension of the limits of GDA by highlighting inductive geometric analysis; its in-depth analysed examples.

References

Kirby, M., 2001. Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns. John Wiley & Sons, Inc., New York, NY.

Niël J. le Roux
Department of Statistics and Actuarial Science,
University of Stellenbosch, South Africa
E-mail address: njlr@sun.ac.za.

Available online 3 February 2008