Journal of Statistical Software

April 2011, Volume 40, Book Review 2. *http://www.jstatsoft.org/*

Reviewer: Gary Evans
University of California, Los Angeles

Exploratory Multivariate Analysis by Example Using R

François Husson, Sébastien Lê, Jérôme Pagès
Chapman & Hall/CRC Press, Boca Raton, FL, 2011.
ISBN 978-1439835807. 240 pp. USD 79.95 (P).
`http://factominer.free.fr/book/`

As its title suggests, this is an R demonstration book in the vein of, for example, Faraway (2005) on linear models. By working through books such as these, if they are done well, one can quickly become interested and conversant, even fairly adept, in the particular field of statistics being studied and its modern applications. Since this is intended as an undergraduate text for non-statistician scientists, it is fairer to hold it to the former standard, by which it can be considered highly successful.

The book is divided into four lengthy chapters, each one dealing with one of the fundamental techniques of multivariate analysis (MVA): principal component analysis (PCA), correspondence analysis (CA), multiple correspondence analysis (MCA), and clustering. The approach is to provide just enough mathematics to remove some of the black-box feeling of the methods. Then, the emphasis is on data characterization, method selection, output generation and interpretation, much like the well-known introductory book by Manly (2005) – though somewhat more detailed. The authors use their R package, **FactoMineR** (Lê, Josse, and Husson 2008), to perform the data analyses.

The chapter on PCA nicely carries out the game plan. The `PCA` function in **FactoMineR** produces two plots: vectors of variables on a correlation circle and the cloud of individual points (essentially the constituents of the traditional biplot). A screeplot is also available and simple commands allow one to obtain numerical PCA output such as loadings, component variances, variable-to-component correlations, etc. A particular strength of **FactoMineR** for PCA is in allowing the inclusion of supplementary (or illustrative) variables (categorical or quantitative) and individuals. Overall, between the book and the **FactoMineR** website, the package is well-presented and it compares favorably with the standard `prcomp`/`biplot` approach in R.

Interpretation being the essence of MVA, the authors devote a good deal of attention to showing the reader how to extract meaning from both the graphic (geometric) and numeric `PCA` output and how to connect the two. They start with a small dataset (orange juices) to allow the reader to quickly grasp the results. The technique is then applied to other larger datasets (decathlon, European city temperatures, chicken genetic expression) which, chosen to be well-suited to PCA, yield illuminating graphics. In the analyses of these, very effective use is made of supplementary elements to highlight features of the data, and all results are clearly and persuasively explained.

Chapter 2 on CA follows basically the same pattern. After a fairly thorough discussion of contingency tables, the mathematics of CA is only tersely presented, essentially by analogy to PCA. Another interesting, and again fairly small, data set (French women's attitudes on working women circa 1970) is then studied. The analysis is highly readable and some thought-provoking subtleties are discovered in the data, but we must make a small stylistic criticism. We are 27 pages into this chapter before we see any R code. In a software demonstration book such as this, such an approach is actually a little distracting. More effective is the more typical approach of presenting the code with the table or figure it produces, as in the Faraway books, for example. The code in this case involves the **FactoMineR** function, `CA`. It is similar to the R function `ca` from package **ca** (Nenadić and Greenacre 2007), though its graphics are a bit more user-friendly. The chapter finishes with analyses of 3 fairly complex data sets (Olympic medals, Loire Valley wines, French mortality data) in which CA is used in conjunction with more standard statistics to impressive instructional effect.

In Chapter 3, MCA is presented as an extension of CA using indicator matrices. Some sparse-matrix mathematics is used to compute formulas for inertias and distances, hinting at how the indicator matrix facilitates the development of the theory of MCA. A little more mathematics like this in the other chapters perhaps could have given the reader a deeper appreciation for these methods and alertness to some of the caveats the authors often emphasize. Unlike Chapters 1 and 2, there is no small data set here so the undergraduate reader, when finished, may not feel as comfortable with MCA as with PCA and CA. The chapter does, however, feature a very rich data set on genetically modified food organisms (GMO) (based on a survey designed and conducted by the authors), but the authors' analyses of this and the other sets in this chapter are a bit more cursory than in the prior two. Without a doubt, the **FactoMineR** function, `MCA`, is a valuable addition to one's R toolkit, but this chapter would need some supplementation in an effective course on MCA.

This leanest chapter of the book is followed by a truly excellent one on clustering. Agglomerative clustering, Ward's method, Huygens' theorem, dendrograms, K-means, and the general philosophy of the method are all covered clearly, substantively and economically. The function, `HCPC` (hierarchical clustering on principal components), is fully demonstrated on two of the earlier data sets (city temperatures from Chapter 1 and tea from chapter 3). It produces a cluster plot, a 2-d dendrogram, and a 3-d dendrogram, the latter of which is especially useful in dealing with the fairly large data sets being studied. It also gives excellent numerical outputs which greatly aid in characterizing clusters and significant individuals. Clustering is tailor-made for the type of data analysis the authors so excel in so the reports here are very engaging. In short, this chapter is an example of what upper-division undergraduate writing should aspire to.

Some minor criticisms, which it is hoped will not obscure the gist of this review: As hinted at above, the book would be slightly stronger with a little more detailed presentation of the mathematics of MVA. In fairness to the authors, to not do so is a conscious decision on their part. Their aim is to give hands-on experience with these data analysis methods and, of course, the methods can be used without having thorough knowledge of the math behind them. But there is much to be gained by having some acquaintance with that knowledge. Perhaps an additional appendix in future editions could provide this. To be doubly fair, in their bibliography, the authors do cite some leading MVA works such as Joliffe (2002), Greenacre (1984) and Gifi (1990) (on PCA, CA, and MCA, respectively) so that the interested reader will be able to pursue the mathematics in more depth.

Two more small criticisms are, first, there are a few typos that slow the reader down at some key points, particularly in Chapters 2 and 3. A couple of them are in code, which can be problematic for beginners with R. They are readily detectable and should have been caught at some point. Second, in Chapter 3, the perfume dataset cannot be read in with the given command. The URL takes us to a data set in French which apparently has some missing separators.

Finally, there are no exercises in the book, nor on the authors' website. This is curious for an undergraduate text. There is some redundancy in the analyses carried out in each of the chapters, so perhaps one or two of the data sets could have been left to the reader to analyze with model write-ups available on the website. Of course, several sets of data and avenues for analyzing them are provided in the book so the enterprising reader can pursue these to gain experience with MVA. For example, using `HCPC`, I analyzed the GMO data from the MCA chapter with interesting results. About `HCPC` I should say that, when used on a couple of the data sets, it would produce only the 2-d dendrogram then an error message when the other graphics were sought. When it did this, it would not store the `HCPC` object to allow retrieval of any numerical output. I could not determine what caused this to happen. Another point about adding exercises is that they could be designed to promote greater learning of R than the reader will get from applying the one-line **FactoMineR** commands that execute the data analyses or reviewing the book's brief appendix on R.

These criticisms having been made in the interest of thoroughness, this enjoyable book, and the **FactoMineR** package, are highly recommended for an upper division undergraduate or beginning graduate level course in MVA. The acid test for such a work must be whether it is likely to spark an interest in students and prepare them adequately for more detailed, serious study of the subject and this book easily passes that test.

References

Faraway JJ (2005). *Linear Models with R.* Chapman & Hall, Boca Raton.

Gifi A (1990). *Nonlinear Multivariate Analysis.* John Wiley & Sons, New York.

Greenacre M (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Joliffe IT (2002). *Principal Component Analysis.* Springer-Verlag, New York.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R Package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18. URL `http://www.jstatsoft.org/v25/i01/`.

Manly BFJ (2005). *Multivariate Statistical Methods – A Primer.* 3rd edition. Chapman & Hall, Boca Raton.

Nenadić O, Greenacre M (2007). "Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The **ca** Package." *Journal of Statistical Software*, **20**(3), 1–13. URL http://www.jstatsoft.org/v20/i03/.

Reviewer:

Gary Evans
University of California, Los Angeles
Department of Statistics
Los Angeles, CA 90024, United States of America
E-mail: gary.evans@stat.ucla.edu
URL: http://directory.stat.ucla.edu/gary-evans

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 40, Book Review 2 *Published:* 2011-04-20
April 2011