Compositional data analysis, correspondence analysis, and the log-ratio connection

Michael Greenacre

(Universitat Pompeu Fabra, Barcelona)

Compositional data analysis (CoDa) is concerned with analysing data tables that have the property of *closure*, i.e. the samples, usually rows of the table, have non-negative values that sum to a constant, usually 1 or 100%. This is often the case in chemistry and the geosciences.

Correspondence analysis (CA), popular in the social and environmental sciences, is concerned with analysing data tables of counts, but also "relativizes" the data by considering the counts relative to the row and column sums of the table -- these relativized rows (or columns) are called *profiles*, also with non-negative values adding up to 1 - thus, correspondence analysis is also a method of compositional data analysis.

Thanks to the work of John Aitchison (1986) there developed a CoDa school based on the **log-ratio** transformation of the data. This transformation also relativizes the values in compositional data set, but not relative to the margins, but rather by looking at all the pairwise ratios, on a logarithmic scale. A principal component analysis type approach to CoDa was developed, summarized by Aitchison and Greenacre (2002).

These two approaches to analysing similar data sets, CoDa and CA, were unified by Greenacre (2009), who showed that they were members of the same family of methods, thanks to the Box-Cox transformation on data ratios, which converges to the log-transformation.

In this talk, I will explain this interesting log-ratio connection, demonstrate the implications for data analysis and emphasize how the connection with CA inspires the idea of weighting the components in CoDa. During the talk I will mention the interesting history behind these two approaches, including the historical connection to *spectral mapping*, a method not well-known in the statistical world, which was developed in the pharmaceutical industry by Paul Lewi already in the 1970s for analysing chemical spectra in drug research (Lewi 1976, Greenacre and Lewi 2009).

For background reading as an introduction to this talk, see chapter 7 of Greenacre (2010), freely downloadable from www.multivariatestatistics.org.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. Reprinted in 2003 by the Blackburn Press.

Greenacre, M. (2009). Power transformations in correspondence analysis. *Computational Statistics and Data Analysis* 53, 3107–3116.

Greenacre, M. (2010). Biplots in Practice. BBVA Foundation, Madrid. URL: www.multivariatestatistics.org.

Greenacre, M. and Aitchison, J. (2002). *Biplots of compositional data*. *Applied Statistics* 51, 375-392.

Greenacre, M. and Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio scale measurements. Journal of Classification 26, 29–54.

Lewi, P. (1976). Spectral mapping: a technique for classifying biological activity profiles of chemical compounds. Arzneimittel Forschung 26, 1295-1300.