Selection of ratios in compositional data analysis

Michael Greenacre
Department of Economics and Business
Universitat Pompeu Fabra
Barcelona

Compositional data analysis is the branch of statistics specialized in the collection, analysis, visualization and interpretation of compositional data, i.e. data that measure parts of a whole, usually expressed as proportions adding up to 1 or percentages summing to 100%. Such data are more frequent in the natural sciences, e.g. chemistry and geology, but also occur in the economic and social sciences, e.g. time budgets, percentages of categorical responses in a survey, or proportions of expenditure on a set of budget items.

The first part of this talk is concerned with the special features of compositional data and why such data should not be treated in the classical way, e.g. by computing means, corrrelation coefficients, doing regression and principal component analysis (PCA), etc... Log-ratio analysis (LRA) is introduced as the multivariate analogue of PCA that respects the measurement peculiarities of compositional data. LRA has also been defined independently as so-called "spectral mapping". In passing, I also demonstrate the very close connection between LRA and CA (correspondence analysis).

The second part is concerned with the selection of a small set of ratios that can effectively replace the original data set, and which can indeed be treated as "regular" statistical variables for univariate summarization, multivariate analysis and modelling.

Some online literature resources

- A working paper on this topic, which is currently under review at *Mathematical Geosciences*, can be found at:

https://econ-papers.upf.edu/ca/paper.php?id=1554

- A chapter on compositional data analysis is included in my book *Multivariate Analysis of Ecological Data*, which is online for free download at

www.multivariatestatistics.org

The direct link to the PDF of this chapter is:

http://www.fbbva.es/TLFU/dat/greenacre_maed_ch14.pdf