Tri-PLS for compositional data

Gallo M.[a*], Trendafilov N.[b] , Todorov V.[c] and Martín-Fernández J.A.[d]

[a]Department of Human and Social Sciences, University of Naples-L'Orientale,
P.zza S.Giovanni 30, Naples (It-80134), Italy
[b]Department of Mathematics and Statistics, The Open University,
Walton Hall, Milton Keynes MK7 6AA, UK
[c]United Nations Industrial Development Organization,
Wagramerstr 5 P.O. Box 300 A-1400, Vienna, Austria
[d] Departament IMAE, Universitat de Girona,
Campus Montilivi, Edif. P-IV, Girona (E-17071), Kingdom of Spain

Keywords: Compositions, Aitchison geometry, Log ratio, NPLS, trilinear models

Introduction

Compositional data (CoDa, [1] and [2]) consist of vectors of positive values summing to a unit, or in general to some fixed constant. They can often be found in many disciplines and appear as proportions, percentages, concentrations, absolute and relative frequencies. Unfortunatly, the constant-sum constraint that characterizes compositions is frequently disregarded or improperly incorporated into statistical modeling and a misleading interpretation of the results is given. Due to these specifications, several difficulties arise when dealing with CoDa. The first word of warning came already in 1897 from Karl Pearson who showed the dangers of underestimating spurious correlations.

There are several approaches to incorporate CoDa into statistical modeling when it is not realistic to assume a multinomial distribution of the data. Based on the log-ratio transformations, Aitchison [1] proposed preprocessing the compositional data by means of log-ratio transformations, and successively analyzing them in a straightforward way by 'traditional' methods. Following Aitchison's approach, the high dimensionality of CoDa in many scientific fields has encouraged the use of bilinear and trilinear decomposition models. Thus, in attempts to find adequate low-dimensional descriptions of compositional variability, CoDa are collected into two or three-way arrays ([3], [4], [5], [6], [7]). On the other side, Hinkle and Rayens [8] examined the problems that potentially occur when one performs a partial least squares (PLS) on compositional data.

The principal goal of this talk is to extend the PLS regression to three-way compositional data, following the approach proposed by Bro [9] and Bro and al. [10]. Both Candecomp/Parafac (CP - [11] [12]) and Tucker3 [13] models can be viewed as latent variables models extending principal component analysis to three-way data set. However, the most fundamental properties of PCA cannot be extended to these two models. PCA is an optimal representation of a two-way array with respect to the criteria of best low-rank approximation in least squares sense and the best approximation of the data within a joint low-dimensional subspace, while Tucker3 is only the best approximation of a three-way array within a joint low-dimensional subspace and CP is the best low-rank approximation in a least squares sense.

The proposed extension of PLS to three-way compositional data is illustrated on real data sets and a software implementation will be available in the R package rrcovHD.

*Corresponding author. E-mail: mgallo@unior.it, nickolay.trendafilov@open.ac.uk, v.todorov@unido.org, josepantoni.martin@udg.edu.

References

[1] J. Aitchison, *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability, 1986.

[2] J. Aitchison, C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn, "Logratio analysis and compositional distance," *Mathematical Geology* **32**(3), pp. 271–275, 2000.

[3] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 139–177, 1982.

[4] J. Aitchison and M. Greenacre, "Biplots of compositional data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **51**(4), pp. 375–392, 2002.

[5] M. Gallo, "Tucker3 model for compositional data," *Commun. Statist. Theor. Meth.* **in press**, 2013.

[6] M. Engle, M. Gallo, K. Schroeder, N. Geboy, and Z. J.W., "Three-way compositional analysis of water quality monitoring data," *Environmental and Ecological Statistics* , pp. 1–17, 2013.

[7] M. Gallo and A. Buccianti, "Weighted principal component analysis for compositional data: application example for the water chemistry of the arno river (tuscany, central italy)," *Environmetrics* , 2013.

[8] J. Hinkle and W. Rayens, "Partial least squares and compositional data: problems and alternatives," *Chemometrics and intelligent laboratory systems* **30**(1), pp. 159–172, 1995.

[9] R. Bro, "Multiway calibration. multilinear pls," *J. Chemometrics* **10**(1), pp. 47–61, 1996.

[10] R. Bro, A. Smilde, and S. de Jong, "On the difference between low-rank and subspace approximation: improved model for multi-linear pls regression," *Chemometrics and Intelligent Laboratory Systems* **58**(1), pp. 3–13, 2001.

[11] J. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young? decomposition," *Psychometrika* **45**, pp. 283–319, 1970.

[12] R. Harshman, "Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers Phonet* , 1970.

[13] L. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika* **31**(3), pp. 279–311, 1966.