Toulouse School of Economics

CoDaWork2022

Toulouse, June 27 - July 1, 2022

Short Abstracts of the 9th International Workshop on Compositional Data Analysis

Editors:

Ch. Thomas-Agnan, V. Pawlowsky-Glahn

ISBN: 978-84-947240-3-9

Editorial: Asociación para Datos Composicionales

Título: Abstract of the 9th International Workshop on Compositional Data Analysis (CoDaWork2022)

Subtítulo: 28 June - 1 July 2022, Toulouse, France

Editado por: Thomas-Agnan, Christine ; Pawlowsky-Glahn, Vera

Formato del producto: Digital: descarga y online

Detalle Formato: PDF; Con posibilidad de lectura sin conexión

Idioma de publicación: Inglés

Número de edición: 1

Fecha de edición: 28/06/2022

País de edición: España

Ward R-mode clustering method for Compositional Data

V. Di Donato[1], V. Pawlowsky-Glahn[2], J.J. Egozcue[3], J.A. Martín-Fernández[2]
[1]Università degli Studi di Napoli Federico II, Naples, Italy; *valedido@unina.it*
[2]University of Girona, Girona, Spain
[3]Technical University of Catalonia, Barcelona, Spain

Abstract

R-mode Cluster Analysis (CA) is commonly adopted in paleoecological studies to identify groups of species sharing a similar behaviour. We present an approach to R-mode CA in a consistent manner with compositional data (CoDa) properties and complementing Ward's method (WM - Ward, 1963) and clr-biplots (Aitchison and Greenacre, 2001). To comply with the CoDa approach, it is necessary to recall some simple concepts:

a) the variation matrix (Aitchison, 1986) for a D-parts CoDa set is a square matrix expressed as: $\mathbf{T}_{D,D}$ with $t_{ij} = var[\ln(X_i/X_j)]$. Diagonal elements of $\mathbf{T}$ are zeros;

b) in the most common application of the WM the criterion for choosing the pair of clusters to merge at each step is the minimization of the within groups sum of squares; and

c) in the clr-biplots, the length of the links between two clr-variable points $\mathrm{clr}(X_i)$ and $\mathrm{clr}(X_j)$ approximates the standard deviation of the logratio of the involved variables which is proportional to the Aitchison distance between the parts, $d_a(X_i, X_j)$ (Martín-Fernández et al., 2018).

It is possible to consider an R-mode clustering of the loadings of the clr-biplot based on WM algorithm to highlight affinities among the parts of the compositions. In fact, by analogy, the WM applied on the loadings of the biplot minimizes at each step the squares of the length between variable points and the centroid of the cluster to which they belong. Importantly, the sum of these lengths represents an approximation of the total variance of the subcomposition, coherently with the definition of variation matrix given above. In this ongoing work, we explore the relationship of an R-mode clustering, with a factor analysis model, usually expressed as $\mathbf{Y}=\mathbf{F}\mathbf{L}^T+\mathbf{E}$ (where $\mathbf{Y}$ is the centred data matrix, $\mathbf{F}$ the matrix of factor scores, $\mathbf{L}$ the matrix of factor loadings and $\mathbf{E}$ a matrix of residuals or error terms). In a similar manner, the clustering of the biplot loadings focuses on the shared variability of the parts leaving apart their singularities.

Key words: Variable selection, Discriminant analysis, Principal Balances

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. 416 pp. Reprinted in 2003 by Blackburn Press.

Aitchison, J. and Greenacre, M. (2002). Biplots of Compositional Data. *J R Stat Soc Ser C (Applied Stat)*, 51:375–392.

Martín-Fernández J.A., V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosona-Delgado, (2018). Advances in principal balances for compositional data, *Mathematical Geosciences*, 50(3):273–298.

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236–244.