

# Simultaneous component and cluster analysis: the between and the within approaches.

Roberto Rocci<sup>1</sup> and Maurizio Vichi<sup>2</sup>

<sup>1</sup>Dept. SEFeMQ, Univ. of “Tor Vergata”, Rome. e-mail: roberto.rocci@uniroma2.it

<sup>2</sup>Dept. of Statistics, Univ. “L aSapienza”, Rome. e-mail: maurizio.vichi@uniroma1.it

**Keywords:** cluster analysis, component analysis, three-way data.

## Abstract

A relevant methodology to summarize three-way data is three-mode factor analysis (T3) (Tucker, 1966) where units, variables and occasions are reduced into a small number of components. Although in multivariate analyses variables and occasions are often summarized by factorial methodologies, units are more frequently partitioned into a few homogeneous classes by a clustering technique. This can be obtained by using the “*sequential approach*” which can be considered a natural extension of the well-known two-way “*tandem analysis*”. It consists of applying firstly a T3 model that identifies a reduced number of components for variables and occasions on which a clustering of units is then performed. This approach can be useful when a large number of variables and occasions is available, especially if some of them do not contain relevant information. However, it has some drawbacks because components identified by T3 are optimal in reconstructing the total variability of the data, but not necessarily for the classification. To avoid problems connected with the sequential approach, the classification and the reduction can be integrated in a unique model. In this way, the components are extracted according to the classification, selecting those variables and occasions that most contribute to identify the optimal partition. Some authors have proposed integrated methodologies for two-way data (e.g., De Soete & Carroll, 1994; Vichi & Kiers, 2001). Following this line of research, this paper discusses and compares two new methodologies for simultaneous classification and variables-occasions reduction. The basic idea underlying the two methods is to identify components for variables and occasions that maximize the between or the within variability of the optimal partition. The comparison is performed from a theoretical point of view and by analysing simulated and real data.

## References

- De Soete, G., & Carroll, J.D. (1994). k-means clustering in a low-dimensional Euclidean space, in: E. Diday et al. (Eds), *New Approaches in Classification and Data Analysis*, Springer, Heidelberg, 212-219.
- Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.
- Vichi, M., & Kiers, H.A.L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.