Clustering of variables around latent variables

Marie Chavent[1,2,*], **Vanessa Kuentz**[1,2], , **Benoit Liquet**[3], **Jérôme Saracco**[1,2]

1. IMB, University of Bordeaux, France
2. CQFD team, INRIA Bordeaux Sud-Ouest, France
3. ISPED, University of Bordeaux, France

*Contact author: marie.chavent@math.u-bordeaux1.fr

Keywords: Mixture of quantitative and qualitative variables, iterative relocation algorithm, hierarchical clustering.

Clustering of variables is studied as a way to arrange variables into homogeneous clusters, thereby organizing data into meaningful structures. Once the variables are clustered into groups such that variables are similar to the other variables belonging to their cluster, the selection of a subset of variables is possible. Several specific methods have been developed for the clustering of numerical variables. However concerning qualitative variables or mixtures of quantitative and qualitative variables, much less methods have been proposed. The homogeneity criterion of a cluster of variables is defined here as the sum of the correlation ratio (for qualitative variables) and the square correlation (for quantitative variables) between the variables and a latent variable, which is in all cases a numerical variable. The latent variable maximizing this homogeneity criterion of a cluster is obtained with PCAMIX. Two algorithms for the clustering of variables are proposed: iterative relocation algorithm, ascendant hierarchical clustering. We also propose a bootstrap approach in order to determine suitable numbers of clusters. The proposed methodology is illustrated with simulated and real datasets via R codes.

References

Chavent M, Kuentz V., Saracco J. (2009). A Partitioning Method for the clustering of Categorical variables. In Classification as a Tool for Research, Hermann Locarek-Junge, Claus Weihs (Eds), Springer, Proceedings of the IFCS'2009, Dresden.

Kiers, H.A.L., (1991). Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197–212.

Vigneau, E. and Qannari, E.M., (2003). Clustering of Variables Around Latent Components, *Communications in Statistics - Simulation and Computation*, **32**(4), 1131–1150.