ClustOfVar: an *R* package for the clustering of variables

Marie Chavent[1,2,⋆], **Vanessa Kuentz**[3], **Benoît Liquet**[4], **Jérôme Saracco**[1,2]

1. IMB, University of Bordeaux, France

2. CQFD team, INRIA Bordeaux Sud-Ouest, France

3. CEMAGREF, UR ADBX, France

4. ISPED, University of Bordeaux, France

⋆Contact author: marie.chavent@u-bordeaux2.fr

Keywords: Mixture of quantitative and qualitative variables, hierarchical clustering of variables, k-means clustering of variables, dimension reduction.

Clustering of variables is as a way to arrange variables into homogeneous clusters i.e. groups of variables which are strongly related to each other and thus bring the same information. Clustering of variables can then be useful for dimension reduction and variable selection. 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 **ClustOfVar** package has then been developped specifically for that purpose. The homogeneity criterion of a cluster is the sum of correlation ratios (for qualitative variables) and squared correlations (for quantitative variables) to a synthetic variable, summarizing "as good as possible" the variables in the cluster. This synthetic variable is the first principal component obtained with the PCAMIX method. 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 methodologies are illustrated on real datasets.

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.

Dhillon, I.S., Marcotte, E.M. and Roshan, U. (3003). Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

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

Pagès, J. (2004). Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, **52**(4), 93-111.

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