A comparison of some methods for clustering of variables of mixed types

N'Dèye Niang, Mory Ouattara, Gilbert Saporta

A comparison of some methods for clustering of variables of mixed types

Ndèye Niang[1], **Mory Ouattara**[2], **Gilbert Saporta**[3],
[1] Cédric-CNAM, Paris, France, ndeye.niang_keita@cnam.fr
[2] Université de San Pedro, Côte d'Ivoire, ouattara.mory@usp.edu.ci
[3] Cédric-CNAM, Paris, France, gilbert.saporta@cnam.fr

We compare several old and recent methods for clustering a set of qualitative and quantitative variables.

Keywords: clustering of variables, mixed data, PCA, RV coefficient

---

The simultaneous treatment of a mixture of J quantitative variables $\mathbf{x}_j$ and Q qualitative variables $\tilde{\mathbf{x}}_q$ with m_q categories, whether in factorial analysis or clustering, is often based on the determination of one or more global or local (i.e. per class) synthetic variables optimizing the following criterion introduced in 1977 by Tenenhaus [6], reused by Escofier (1979), then Saporta [5], Kiers (1991) under the name of PCAMIX, and Pagès (2004):

$$\max_{\mathbf{c}} \left(\sum_{j=1}^{J} r^2 \left(\mathbf{c}, \mathbf{x}_j\right) + \sum_{q=1}^{Q} \eta^2 \left(\mathbf{c}, \tilde{\mathbf{x}}_q\right) \right) \tag{1}$$

where r^2 is the squared Pearson correlation coefficient between two quantitative variables and η^2 the squared correlation ratio between a quantitative and a qualitative variable. Both coefficients are equal to the proportion of variance of a dependent variable explained by an independent one.
The ClustOfVar algorithm [1] uses criterion (1) to perform a clustering of a set of variables of different nature around latent components in each group, extending the method of Vigneau and Qannari [7] introduced for exclusively quantitative variables.
Clustering variables around components is an interesting alternative to direct algorithms that start from the table of similarities, dissimilarities or distances between all variables, because it simultaneously optimizes the clustering and the representation of classes by a component as in a clusterwise approach.
A key issue is to use consistent and comparable similarity measures in the three cases : a pair of quantitative variables, a pair of categorical variables and a pair consisting in a quantitative variable and a categorical one. The association coefficients r^2 and η^2 are in common use, while various solutions have been proposed for the case of two categorical variables: chi-squared and its derivatives such as T^2, which is the square of the Tschuprow coefficient, or the largest eigenvalue of the Correspondence Analysis matrix derived from the cross-tabulation of two categorical variables [1].

Coefficients associated with categorical variables are not, however, comparable with each other or with r^2 because their distributions depend on their number of categories. In criterion (1) a qualitative variable plays a greater role the higher its number of categories m_q. The Escoufier RV coefficients [4] between tables generated by each quantitative variable and tables of indicators of the categories of the qualitative variables make it possible to define Euclidean similarities equal, according to the cases, to r^2,$\frac{\eta^2}{\sqrt{m_q-1}}$ or T^2 [3].

We can then perform hierarchical clustering with Ward's algorithm or k-means partitioning, either directly on the similarity matrix, or on the coordinates obtained by the Torgerson formula. This elegant but somewhat forgotten solution still suffers from a flaw: dividing by the square root of the degree of freedom does not completely correct the effect of the number of categories. For this, it may be wise to use as dissimilarity the p-value of the independence test in the spirit of the *likelihood linkage algorithm* [2]. However Euclidean properties are lost.

In addition, when the number of observations is very large, all p-values are close to zero (*paradox of large samples*) and are no longer usable. We propose to replace them by the corresponding fractiles of the standard normal distribution in the spirit of the *test values* used in the SPAD software. The larger the fractile, the greater the association between two variables. These different approaches are compared on real data sets.

References

[1] M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco. ClustOfVar: An R package for the Clustering of Variables. *Journal of Statistical Software*, 50(13):1–16, 2012.

[2] F. Costa Nicolau and H. Bacelar-Nicolau. Some trends in the classification of variables. In C. Hayashi, editor, *Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan*, pages 89–98. Springer, 1998.

[3] E.M. Qannari, E. Vigneau, and Ph. Courcoux. Une nouvelle distance entre variables. Application en classification. *Revue de Statistique Appliquée*, 46(2):21–32, 1998.

[4] P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3):257–265, 1976.

[5] G. Saporta. Simultaneous analysis of qualitative and quantitative data. In *Atti della XXXV Riunione Scientifica, Societa Italiana di Statistica, Padova, Italy*, volume 1, pages 62–72, 1990.

[6] M. Tenenhaus. Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, 25(2):39–56, 1977.

[7] E. Vigneau and E.M. Qannari. Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, 32(4):1131–1150, 2003.