Sparse Divisive Feature Clustering

Ndèye Niang-Keita, Mory Ouattara, Gilbert Saporta

11 December, 9:00 - 9:20

Sparse Divisive Feature Clustering

Ndèye Niang[1], **Mory Ouattara**[2], **Gilbert Saporta**[3]
[1] CEDRIC-CNAM, ndeye.niang_keita@cnam.fr
[2] SFA-UNA, ouattaramory.sfa@univ-na.ci
[3] CEDRIC-CNAM, gilbert.saporta@cnam.fr

We propose an approach based on a divisive algorithm for clustering variables in order to identify in a large data table underlying dimensions that are not necessarily orthogonal. The number of clusters does not have to be defined in advance. The clusters, which are as unidimensional as possible, are then represented in a parsimonious way by a small number of variables or components.

Keywords: Feature Clustering, Simple structure, Sparse principal component

Let us consider the context of unsupervised analysis of a large data set where variables are assumed to be structured in homogeneous blocks. Two situations may occur: either the features are divided into blocks defined beforehand (expert knowledge or "natural" clusters such as answers to questionnaires according to different themes) and multiblock methods such as STATIS or RGCCA are well adapted. We are interested here in the alternative case where the blocks are constructed from the data. Finding blocks of variables is linked to the objective of reducing the dimension of the features space in order to facilitate the interpretation. But also, more fundamentally, it is related to the objective of discovering simple structures as in factor analysis in the sense that each factor would be correlated with a small number of features and each feature would be correlated with few factors. It is then natural to look for clusters that are as unidimensional as possible. Each cluster will be summarized by a prototype or by a parsimonious linear combination of features.

Feature clustering solves problems that PCA cannot address because of dual orthogonality constraints on factors and components. The orthogonality constraints lead to optimal projections for units but usually not to simple structures for the components. Hence the use of orthogonal or oblique rotations or sparse PCA [4]. Sparse PCA facilitates the interpretation because each sparse component is related to few features, but the degree of sparsity depends on a parameter whose tuning remains problematic.
Compared to the clustering of individuals, variable clustering received much less attention in the literature. Proposed methods often simply copy usual methods of clustering of units which is intellectually unsatisfactory. Among specific methods for classifying features, let us mention the latent variables based CLV method [6], ClustOfVar package [1], the interpretable principal components based on [2] and the methods based on the likelihood of the link [3]. Previous methods are hierarchical or K-means like and suffer from well-known shortcomings: hierarchical methods are not adapted to the case of a large number of objects, K-means like methods assume that the number of clusters is fixed in advance.

The VARCLUS procedure of SAS software, which has never been the subject of scientific articles, presents several interests. It is a top-down hierarchical method that separates iteratively the features into two sub-clusters until there is only one eigenvalue larger than 1 in the PCA of each cluster. The condition on the eigenvalues gives VARCLUS two important advantages: the number of clusters is naturally obtained and the resulting clusters which are associated with a first large eigenvalue are unidimensional to some extent. Clustering features into unidimensional blocks is a simple and efficient way to search for so-called "oblique" factors. Each block can then be represented by a single component combining only its features and then necessarily sparse relatively to the number of variables. When the cluster size is still too large, a further simplification is needed.

We therefore propose a multi-step strategy. Firstly, VARCLUS is performed. This provides the optimal number of clusters and the associated partition is used as an initialization for CLV, which avoids the computational cost of the hierarchical agglomerative clustering or of the several random runs of an initial partition as proposed to get the number of clusters in CLV method. Secondly CLV is performed and noise (or isolated) variables are discarded in order to keep only relevant clusters in the spirit of [5]. For the last step of prototype determination: the prototype can be the first sparse principal component, its closest feature in terms of maximal correlation or the ≪ medoid ≫ feature.
All these different strategies are evaluated on simulated data and illustrated on real data.

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] D.G. Enki, N.T. Trendafilov, and I.T. Jolliffe. A clustering approach to interpretable principal components. *Journal of Applied Statistics*, 40(3):583–599, 2013.

[3] F. Nicolau and H. Bacelar-Nicolau. Some trends in the classification of variables. In *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.

[4] N. T. Trendafilov. From simple structure to sparse components: a review. *Computational Statistics*, 29(3):431–454, 2014.

[5] E. Vigneau and M. Chen. Dimensionality reduction by clustering of variables while setting aside atypical variables. *Electronic Journal of Applied Statistical Analysis*, 9(01):134–153, 2016.

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