ClustVarLV : un package pour la classification de variables autour de variables latentes

E. Vigneau, **M. Chen** et **E.M. Qannari**

Unité de Sensométrie et Chimiométrie
Oniris, site de la Géraudière
44322 Nantes
evelyne.vigneau@oniris-nantes.fr
mingkun.chen@oniris-nantes.fr
elmostafa.qannari@oniris-nantes.fr

Mots clefs : Classification de variables, variables latentes, données externes.

La méthode de Classification de variables autour de Variables Latentes (CLV) a été initiée il y a déjà une dizaine d'années [1] et a régulièrement été enrichie pour pouvoir répondre à des problématiques diverses dans différents domaines d'application : analyse sensorielle, études consommateurs, spectroscopie vibrationnelle, analyse de questionnaires en psychologie, sociologie ou dans le domaine de la santé...

Une des spécificités de l'approche CLV est qu'elle s'apparente, d'un côté, à la famille des techniques de classification, en produisant une partition de variables, et, d'un autre côté, elle s'apparente aux méthodes factorielles d'analyse de données puisque chaque groupe de variables est construit de sorte à être le plus unidimensionnel possible et qu'il est représenté par une variable latente qui lui est propre. Ainsi, cette approche s'avère être une méthode efficace de réduction de la dimensionnalité dans un espace de grande dimension et d'identification de structures simples (à l'instar des techniques de rotation telles que la rotation Varimax).

Conçue comme un problème d'optimisation d'un critère basé sur l'association de chacune des variables avec la variable latente du groupe auquel elle appartient, l'approche CLV est d'une conception simple et peut aisément être adaptée à des situations plus ou moins complexes. Ainsi, lorsque l'on dispose de plusieurs ensembles d'information en complément du bloc des variables à classer, les variables latentes de groupes peuvent être définies sous contrainte, afin de tenir compte des données externes.

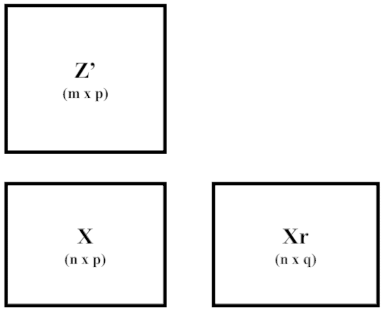

Figure 1: Organisation de données en L: les variables du bloc **X** sont les variables à classer, les blocs **Xr** et **Z** contiennent des informations externes à prendre en compte dans le processus de classification.

En considérant une structure de données en L, comme celle illustrée en figure 1, il est possible

de réaliser la classification des variables du bloc central (**X**) de sorte que les variables latentes des groupes puissent être expliquées par les variables externes contenues dans le bloc **Xr**. Dans le domaine de l'évaluation sensorielle, la segmentation des consommateurs en fonction de leur appréciation pour les produits testés (bloc **X** dans lequel les n lignes correspondent aux produits et les p colonnes aux consommateurs) peut ainsi directement être expliquée par les caractéristiques sensorielles ou physico-chimiques des produits (bloc **Xr**). Simultanément, si le bloc d'information **Z** est disponible, la classification des variables de **X** peut être réalisée en tenant compte des informations complémentaires disponibles sur les variables elle-mêmes. Ce bloc **Z** peut contenir, par exemple, des informations socio-démographiques, d'usages et attitudes, collectées auprès des consommateurs de l'étude.

Un package nommé **ClustOfVar** [2] permet d'ores et déjà de réaliser une classification de variables fondée sur un des critères de l'approche CLV mais en le généralisant pour pouvoir prendre en compte un mélange de variables quantitatives et qualitatives. La structure des algorithmes mis en oeuvre est identique à celle présentée dans [1]. Deux options sont fixées dans `ClustOfVar`: les groupes de variables formés sont forcément bi-directionnels (les variables dans un groupe sont fortement corrélées, positivement ou négativement), les variables quantitatives sont forcément normées. Ces choix sont bien souvent justifiés mais dans certains domaines (dont les études de préférence) ils ne sont pas forcément adaptés.

Le package **ClustVarLV** [3] permet à l'utilisateur de réaliser la classification de variables autour de variables latentes, avec des groupes bi-directionnels ou locaux, en tenant compte d'informations externes disponibles sur les observations et/ou les variables. La fonction principale du package, `CLV()`, combine deux démarches de classification : un algorithme ascendant hiérarchique et un algorithme de consolidation par partitionnement. Elle est complétée par une fonction de partitionnement "pur" (avec initialisation aléatoire répétée), `CLV_kmeans()`, plus rapide lorsque le nombre de variables est très important, et la fonction `LCLV()` dans le cas spécifique des données organisées en L. Des fonctions complémentaires de représentation graphique et de description des groupes de variables y sont également intégrées.

Une illustration des fonctionnalités du package **ClustVarLV** sera présentée sur la base d'une étude sensorielle hédonique avec une structure de données en L. Une rapide comparaison des packages **ClustOfVar** et **ClustVarLV** dans le cas de variables mixtes sera proposée.

Références

[1] Vigneau, E., Qannari, E. M. (2003). Clustering of variables around Latent Variables. *Comm. Stat. - Simul. Comput.*, **32**(4), 1131-1150.
[2] Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012). ClustOfvar : An R package for the clustering of variables. *Journal of Statistical Software*, **50**(13), 1-16.
[3] Vigneau, E., Chen, M. (2014). ClustVarLV : Clustering of variables around Latent Variables, R package version 1.2. http://cran.r-project.org/web/packages/ClustVarLV.