Classification de variables avec possibilité de mettre à l'écart des variables atypiques ou de bruit. Implémentation dans le package ClustVarLV.

E. Vigneau & **E.M. Qannari**

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

Mots clefs : Classification de variables, *noise cluster*, *Sparse components*.

En général, les méthodes de classification ne permettent pas de tenir compte de la présence de données atypiques ou de bruit. Le processus de classification concerne tous les objets statistiques, même ceux qui n'ont pas de liens (proximités) avec aucun autre objet du jeu de données. Les méthodes de classification floue offrent la possibilité d'affecter ce type d'objets à différentes classes mais avec un degré d'appartenance faible. On adoptera des approches de classification "nette", mais en offrant la possibilité qu'un objet atypique ne soit affecté à aucune classe, ou, alternativement, que cet objet ne soit pas pris en compte dans la définition des noyaux des différentes classes. On se placera dans le cadre de la classification de variables, les objets statistiques considérés pouvant être des items de réponse d'un questionnaire; des caractéristiques sensorielles, biochimiques; ou encore des variables spectométriques ou chromatographiques quantifiées par des approches holistiques de type -omiques, *etc.*

La méthode de classification de variables autour de variables latentes (CLV) [1], implémentée dans le package `ClustVarLV` [2], a été adaptée afin de pouvoir mettre à l'écart les variables assimilables à du bruit ou atypiques en regard de la structure des corrélations au sein du jeu de données. A partir d'une partition initiale des variables, l'algorithme d'optimisation alternée de CLV consiste en deux étapes : (a) une étape d'estimation de la variable latente de chaque groupe, et (b) une étape de (ré-)affectation de chacune des variables au groupe avec lequel elle présente le lien le plus fort. Dans la méthode CLV, selon l'objectif poursuivi par l'utilisateur, ce lien est mesuré soit par la covariance, soit par la covariance au carré, entre la variable considérée et la composante latente du groupe. Deux stratégies différentes de modification de l'algorithme sont proposées.

La première stratégie, nommée "K+1", consiste à définir les indicatrices, $\delta_{jk} \in \{0, 1\}$, d'appartenance de chaque variable, j ($j = 1, .., p$), à un groupe, G_k ($k = 1, ...K$), telles que $(\sum_k \delta_{jk})$ soit égale à 1 pour les variables classées, ou, égale à 0 pour les variables mises à l'écart. De manière analogue à la proposition de Davé [3], un terme est ajouté aux critères considérés dans la démarche CLV pour la prise en compte de ce *"noise cluster"*. La modification de l'algorithme de classification CLV est apportée à l'étape d'affectation (b). L'affectation d'une variable au groupe auquel elle est la mieux associée aura lieu, ou la variable sera placée dans le *"noise cluster"* en fonction d'un paramètre, noté ρ. Par exemple, dans le cas où l'on cherche à définir des groupes directionnels (*i.e.* critère de classification basé sur le carré des covariances), on

aura :

$$\begin{cases} \delta_{kj} = 0 \ \forall k & \text{si } \max\left\{\max_l \left\{n \operatorname{cov}^2\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho^2 \operatorname{var}(\mathbf{x}_j)\right\} = \rho^2 \operatorname{var}(\mathbf{x}_j) \\ \delta_{kj} = 1 & \text{si } \max\left\{\max_l \left\{n \operatorname{cov}^2\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho^2 \operatorname{var}(\mathbf{x}_j)\right\} = n \operatorname{cov}^2\left(\mathbf{x}_j, \mathbf{c}_k\right) \end{cases}$$

Le paramètre ρ, compris entre 0 et 1, représente un seuil de corrélation. Le choix de ce paramètre détermine le nombre de variables qui seront écartées. Ce choix est toujours délicat mais, dans le cas de la classification de variables, la plage des valeurs possibles est bornée, ce qui n'est pas le cas dans la démarche proposée par Davé [3].

La seconde stratégie, nommée "Sparse LV", consiste en une modification de l'étape de détermination des variables latentes associées aux groupes (étape (a)). Si on considère, par exemple, le cas de la classification des variables en groupes directionnels, la maximisation du critère dans CLV conduit à définir chaque variable latente de groupe comme étant colinéaire à la première composante principale du sous-tableau formé des variables affectées à ce groupe. Afin d'améliorer la stabilité et l'interprétation de ces variables latentes, une démarche similaire à celle de l'ACP Sparse [4] a été adoptée. Une procédure itérative, avec une étape de seuillage "doux", est mise en oeuvre. Elle conduit à ce que les variables dont la corrélation avec la variable latente du groupe est relativement faible se voient attribuer un *loading* nul dans la combinaison linéaire qui génère la variable latente de groupe. Le seuillage est fondé sur l'introduction du paramètre ρ, comme pour la stratégie "K+1".

Il est possible de mettre en oeuvre, l'une ou l'autre de ces deux stratégies en utilisant la fonction `CLV_kmeans()` du package `ClustVarLV`. Deux exemples de mise en oeuvre de cette démarche de classification de variables, avec "nettoyage", seront brièvement présentés. Dans le contexte de la segmentation de consommateurs en évaluation sensorielle, la possibilité de mettre de côté des consommateurs dont les préférences ne s'accordent pas avec les grandes tendances au sein du panel, permet d'exhiber une segmentation plus interprétable. Dans le domaine des techniques analytiques à haut débit, telle que la RMN, qui permettent de collecter un grand nombre de points de mesure pour chaque échantillon, la classification CLV permet d'identifier des sous-ensembles de variables fortement corrélées. Les stratégies complémentaires proposées ici permettent, également, d'écarter les variables spectrales dont la variabilité serait assimilable à du bruit.

Références

[1] Vigneau, E., Qannari, E. M. (2003). Clustering of variables around Latent Variables. *Comm. Stat. - Simul. Comput.*, **32**(4), 1131-1150.
[2] Vigneau, E., Chen, M. (2015). ClustVarLV: Clustering of variables around Latent Variables. URL http://CRAN.R-project.org/package=ClustVarLV,R package version 1.3.2.
[3] Davé, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, **12**,(11), 657-664.
[4] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15, 265-286.