Clustering sparse des données mixtes avec le package R vimpclust

M. Chavent* J. Lacaille† A. Mourer‡ M. Olteanu§

19 avril 2021

Résumé

Cette présentation s'intéresse à la sélection de variables dans le contexte du clustering et plus précisément au clustering sparse des données mixtes (mélange de variables numériques et catégorielles). Nous illustrons la méthode introduite dans Chavent et al. [2020], qui combine un pré-traitement des variables catégorielles et une extension de l'algorithme des K-means sparse de Witten and Tibshirani [2010] au cas group-sparse. Cette méthode est implémentée dans le package `R vimpclust` disponible sur le `CRAN` [Mourer et al., 2020].
Mots-clefs : clustering sparse, K-means pondéré, données mixtes.

1 Introduction

La méthode des K-means sparse pour données mixtes est implémentée dans la fonction `sparsewkm` du package `vimpclust`. Nous reprenons dans cette présentation l'exemple de la vignette [1] qui porte sur les données HDdata [2]. Ces données décrivent 270 patients susceptibles d'être atteints d'une maladie du coeur à l'aide de treize variables, dont six sont numériques et sept sont catégorielles. On cherche une partition en deux classes des 270 patients tout en identifiant les variables les plus importantes pour le clustering.

2 Resultats

On applique la fonction `sparsewkm` à ce jeu de données.

```
res <- sparsewkm(X = HDdata, centers = 2)
```

Cette fonction maximise un critère de variance inter-classe pondéré et pénalisé avec la norme L_1 des poids des variables. La sparsité est obtenue en mettant à zéro les poids des variables qui ne sont pas importantes pour le clustering. La pénalité est contrôlée par un paramètre de régularisation λ que l'on fait varier entre 0 et 1 et qui permet de tracer le chemin de régularisation de chacune des 13 variables (voir Figure 1-a). Sur ce graphique, les variables catégorielles sont représentées par des lignes pointillées tandis que les variables numériques sont représentées par des lignes continues. Afin de choisir la valeur de λ, on trace également les pourcentages de variances expliquées par les partitions en fonction de λ.

```
plot(res, what="weights.features")
plot(res, what="expl.var")
```

*INRIA Bordeaux Sud-Ouest - CQFD team - France, `marie.chavent@u-bordeaux.fr`
†Safran Aircraft Engines - Datalab - Villaroche - France, `jerome.lacaille@safrangroup.com`
‡SAMM - EA 4543 Université Pantheon Sorbonne - France, `mourer.alex@gmail.com`
§CEREMADE, UMR 7534 Université Paris Dauphine PSL - France, `olteanu@ceremade.dauphine.fr`
1. http ://cran.r-project.org/web/packages/vimpclust/vignettes/sparsewkm.html
2. https ://archive.ics.uci.edu/ml/datasets/statlog+(heart)

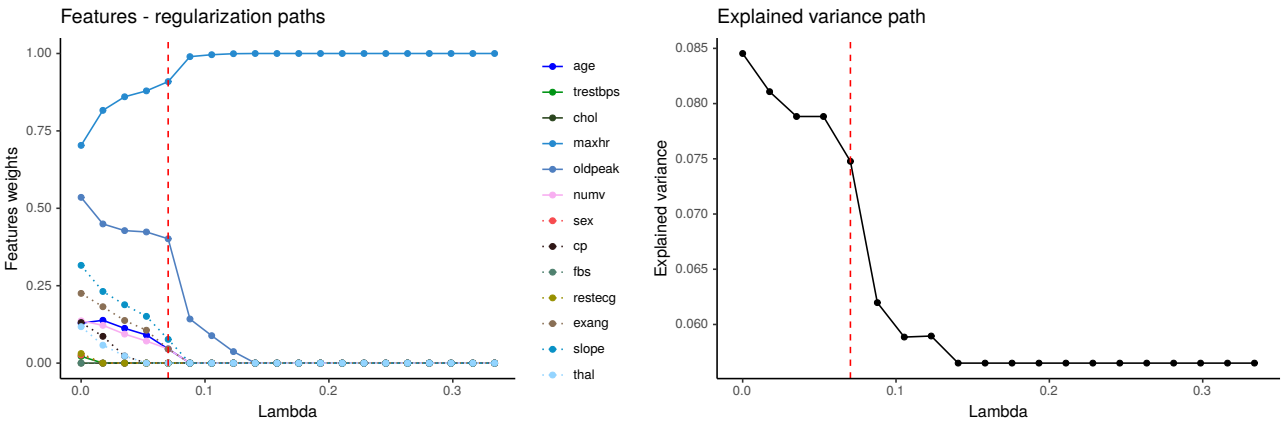

FIGURE 1 – À gauche (a) les chemins des poids des variables et à droite (b) le chemin des variances expliquées.

On observe qu'avec la valeur $\lambda = 0.07$ (représentée par une ligne pointillée verticale rouge) l'algorithme sélectionne 6 variables (4 numériques et deux catégorielles). La partition en deux classes obtenue avec ces 6 variables a un pourcentage d'inertie expliquée de 7.5% (ce qui n'est pas nécessairement petit pour une partition en 2 classes) soit une diminution de 1% par rapport à la partition obtenue avec les 13 variables. D'une manière générale, le but est de sélectionner le plus petit nombre de variables possible pour interpréter le clustering tout en conservant un maximum de la variance expliquée.

La table 1 montre que les 6 variables sélectionnées discriminent bien les classes. Ces résultats ont été obtenus avec la fonction `info_clust` du package.

Variables	Cluster 1	Cluster 2	Globale
maxhr	127.1	164.2	149.7
oldpeak	1.85	0.53	1.05
slope - lev. 1	15.1%	69.5%	48.1%
slope - lev. 2	73.6%	26.8%	45.2%
slope - lev. 3	11.3%	3.7%	6.7%
exeang - lev. 1	41.5%	83.5%	67.0%
exeang - lev. 2	58.5%	16.4%	33.0%
age	58.2	52.0	54.4
numv	1.03	0.43	0.67

TABLE 1 – Moyennes (variables numériques) et fréquences (modalités des variables catégorielles) par cluster et dans l'ensemble de la population pour les 6 variables sélectionnées triées par ordre décroissant d'importance.

Références

M. Chavent, J. Lacaille, A. Mourer, and M. Olteanu. Sparse k-means for mixed data via group-sparse clustering. *ESANN 2020 proceedings, i6doc.com publ., ISBN 978-2-87587-074-2*, 2020.

A. Mourer, M. Chavent, and Olteanu M. vimpclust : Variable importance in clustering. `http://cran.r-project.orgr/web/packages/vimpclust/`, 2020.

D. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 2010.