Clustering, données mixtes et sélection de variables: le package-R `vimpclust` [1]

Marie Chavent[1], Jerome Lacaille[2], Alex Mourer[1,2,3], and Madalina Olteanu[4]

[1]Inria bdx sud ouest équipe ASTRAL IMB, UMR CNRS 5251, Université de bordeaux - France
[2]Safran Aircraft Engines - Datalab - Villaroche - France
[3]SAMM - EA 4543 - Université Pantheon Sorbonne - France
[4]CEREMADE, UMR 7534 - Université Paris Dauphine PSL - France

Rencontres R
13 juillet 2021

1. https://cran.r-project.org/web/packages/vimpclust/index.html

1. **Contexte**

   1.1 Souvent des données mixtes

   1.2 Les clusters sous-jacents ne diffèrent que suivant certaines variables

   1.3 Possiblement un grand nombre de variables

2. **Objectifs**

   2.1 Production d'une structure de classification : groupes d'individus (clusters)

   2.2 Faire de la sélection de variables sur des données mixtes

3. **Motivations**

   3.1 Prendre en compte le pouvoir discriminatif de chaque variable

   3.2 Améliorer l'interprétabilité

1. **Clustering sur des données mixtes**

2. **Clustering avec sélection de variables sur des données mixtes**

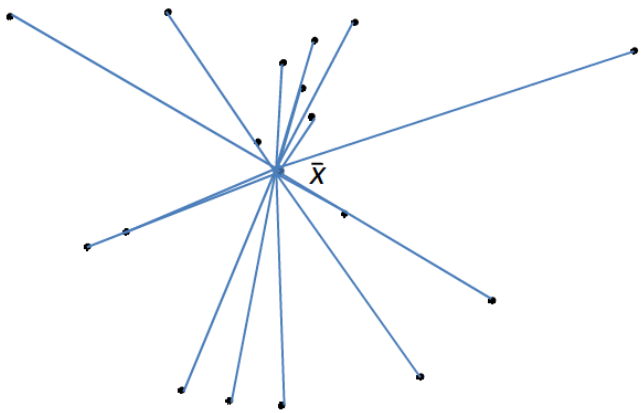
3. **Clustering avec sélection de groupes de variables sur des données mixtes**

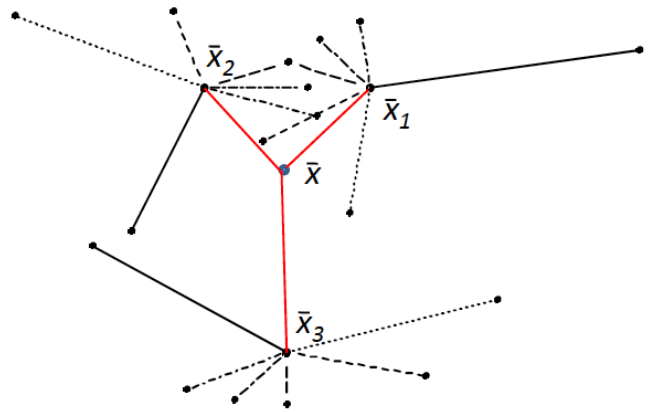
Une partition en K classes des individus est un ensemble de classes non vides, deux à deux disjointes et dont la réunion est l'ensemble des individus.

On notera $P_K = (C_1, \ldots, C_k, \ldots, C_K)$.

L'homogénéité et **la séparation** des classes avec **un seul critère**.

$$\underbrace{\sum_{i=1}^{n} d^2(x_{i.}, \bar{x})}_{\textbf{Inertie totale}} = \underbrace{\sum_{k=1}^{K} \sum_{i \in C_k} d^2(x_{i.}, \bar{x}_k)}_{\textbf{Inertie intra}} + \underbrace{\sum_{k=1}^{K} \mu_k d^2(\bar{x}_k, \bar{x})}_{\textbf{Inertie inter}}$$





Minimiser l'inertie intra (rendre homogène les classes)

$\Leftrightarrow$

Maximiser l'inertie inter (séparer des classes)

Le K-means minimise l'inertie intra et donc maximise l'inertie inter.

Données mixtes : 21 vins de Loire, 31 variables, 2 quali (Le, S., Josse, J. Husson, F. 2008)

TABLE – Données mixtes

	Label	Soil	Bitterness	Smooth	. . .
2EL	Saumur	Env1	1.926	2.731	. . .
1CHA	Saumur	Env1	1.926	2.500	. . .
1FON	Bourgueuil	Env1	2.000	2.679	. . .
1VAU	Chinon	Env2	1.963	1.680	. . .
1DAM	Saumur	Reference	2.071	3.036	. . .
. . .	. . .	. . .	. . .	. . .	. . .

TABLE – Données numériques

	Saumur	Bourgueuil	Chinon	Reference	Env1	Env2	Env4	Bitterness	Smooth	
2EL	1	0	0	0	1	0	0	1.926	2.731	. . .
1CHA	1	0	0	0	1	0	0	1.926	2.500	. . .
1FON	0	1	0	0	1	0	0	2.000	2.679	. . .
1VAU	0	0	1	0	0	1	0	1.963	1.680	. . .
1DAM	1	0	0	1	0	0	0	2.071	3.036	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

TABLE – Fréquences des modalités et variance des variables quantitatives

Saumur	Bourgueuil	Chinon	Reference	Env1	Env2	Env4	Bitterness	Smooth	. . .
0.524	0.286	0.19	0.333	0.333	0.238	0.095	0.061	0.51	. . .

TABLE – Données recodées

	Saumur	Bourgueuil	Chinon	Reference	Env1	Env2	Env4	Bitterness	Smooth	
2EL	$\frac{1}{\sqrt{0.524}}$	0	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	$\frac{1.926}{\sqrt{0.061}}$	$\frac{2.731}{\sqrt{0.51}}$	. . .
1CHA	$\frac{1}{\sqrt{0.524}}$	0	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	$\frac{1.926}{\sqrt{0.061}}$	$\frac{2.500}{\sqrt{0.51}}$	. . .
1FON	0	$\frac{1}{\sqrt{0.286}}$	0	0	1	0	0	$\frac{2.000}{\sqrt{0.061}}$	$\frac{2.679}{\sqrt{0.51}}$	. . .
1VAU	0	0	$\frac{1}{\sqrt{0.19}}$	0	0	$\frac{1}{\sqrt{0.238}}$	0	$\frac{1.963}{\sqrt{0.061}}$	$\frac{1.680}{\sqrt{0.51}}$	. . .
1DAM	$\frac{1}{\sqrt{0.524}}$	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	0	$\frac{2.071}{\sqrt{0.061}}$	$\frac{3.036}{\sqrt{0.51}}$	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

1. **Clustering sur des données mixtes**

2. **Clustering avec sélection de variables sur des données mixtes**

3. **Clustering avec sélection de groupes de variables sur des données mixtes**

Inertie inter-classe d'une partition :

$$\begin{aligned}\mathcal{B}(\mathbf{X}, C_1, \ldots, C_K) &= \sum_{k=1}^{K} \mu_k d^2(\bar{x}_k, \bar{x}) \\ &= \sum_{j=1}^{p} \underbrace{\sum_{k=1}^{K} \mu_k (\bar{x}_k^j - \bar{x}^j)^2}_{\mathcal{B}(\mathbf{x}^j, C_1, \ldots, C_K)}\end{aligned}$$

On note :

$$b_j = \mathcal{B}(\mathbf{x}^j, C_1, \ldots, C_K).$$

Inertie inter-classe pondérée :

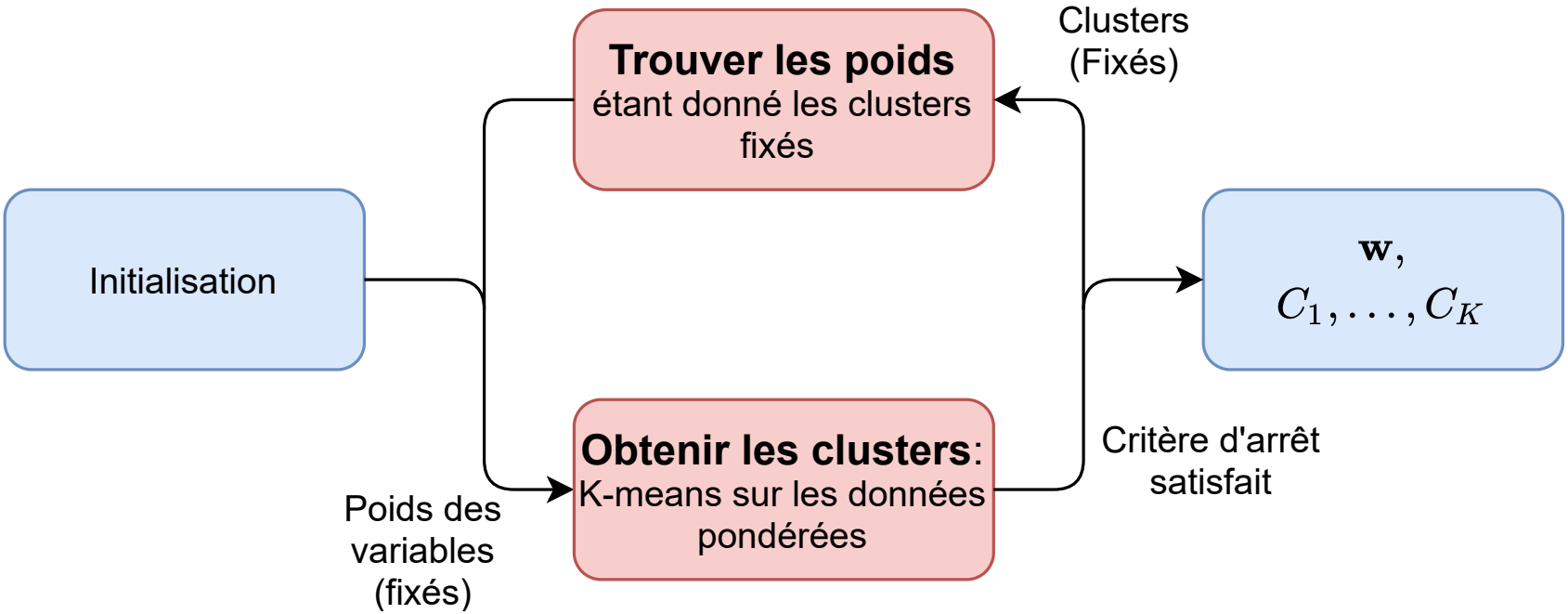
$$\begin{aligned}\mathcal{B}(\mathbf{X}, C_1, \ldots, C_K, \mathbf{w}) &= \sum_{j=1}^{p} w_j b_j \\ &= \mathbf{w}^T \mathbf{b}\end{aligned}$$

où $\mathbf{w}^T = (w_1, \ldots, w_j, \ldots, w_p)$ avec w_j le poids de la variable $\mathbf{x}^j$ et $\mathbf{b}^T = (b_1, \ldots, b_p)$ vecteur de variance inter-classe

Sparse K-means (Witten & Tibshirani, 2010)

$$\max_{\mathbf{w}, C_1, \ldots, C_K} \mathbf{w}^T \mathbf{b} - \lambda \|\mathbf{w}\|_1 \tag{1}$$

Algorithme itératif pour K et λ fixé :



$$\mathbf{w}^T \mathbf{b} = \sum_{j=1}^{p} \sum_{k=1}^{K} \mu_k (\sqrt{w_j} \times \bar{x}_k^j - \sqrt{w_j} \times \bar{x}^j)^2$$

Group-sparse K-means (Chavent, Lacaille, Mourer, & Olteanu, 2020)

Les variables sont divisées en L groupes connus à priori :

$$\mathbf{X} = \left[\mathbf{X}^1|...|\mathbf{X}^L\right]$$
$$\mathbf{b}^T = (\mathbf{b}_1, \ldots, \mathbf{b}_L)$$
$$\mathbf{w}^T = (\mathbf{w}_1, \ldots, \mathbf{w}_L)$$

$$\max_{\mathbf{w}, C_1, \ldots, C_K} \mathbf{w}^T \mathbf{b} - \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\mathbf{w}_\ell\|_2 \tag{2}$$

Pour $C_1, ..., C_K$ fixé :

$$\mathbf{w}^\star = \begin{cases} \frac{\tilde{S}(\mathbf{b},\lambda)}{\|\tilde{S}(\mathbf{b},\lambda)\|_2} & \text{si } \tilde{S}(\mathbf{b},\lambda) \neq 0 \\ 0 & \text{si } \tilde{S}(\mathbf{b},\lambda) = 0 \end{cases} \tag{3}$$

où

$$\tilde{S}(\mathbf{b},\lambda) = (\tilde{s}(\mathbf{b}_1,\lambda), \ldots, \tilde{s}(\mathbf{b}_L,\lambda))^T$$

et

$$\tilde{s}(\mathbf{b}_\ell,\lambda) = \mathbf{b}_\ell \times \max(1 - \frac{\lambda\sqrt{p_\ell}}{\|\mathbf{b}_\ell\|_2}, 0)$$

Sparse K-means pour données mixtes

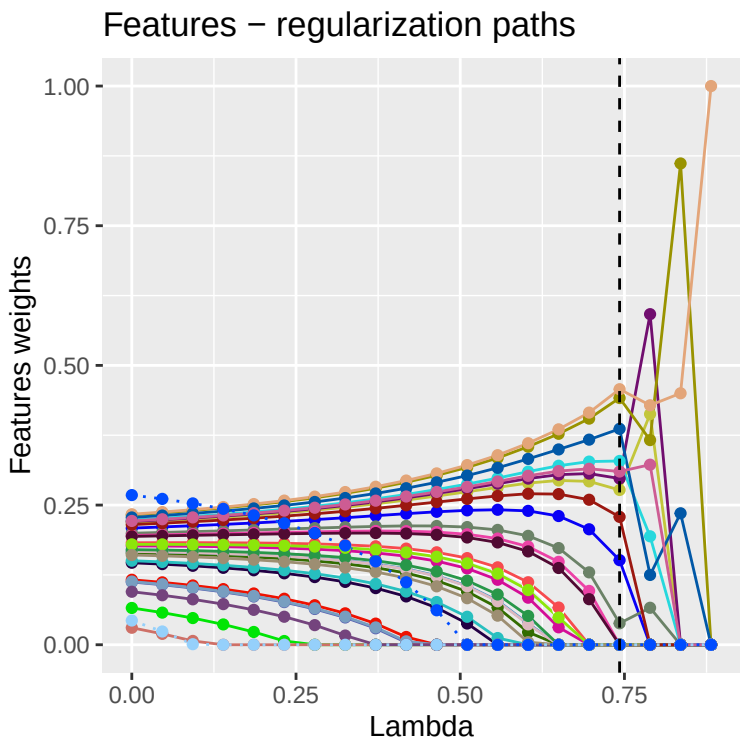
TABLE – Données recodées

	Saumur	Bourgueuil	Chinon	Reference	Env1	Env2	Env4	Bitterness	Smooth	
2EL	$\frac{1}{\sqrt{0.524}}$	0	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	$\frac{1.926}{\sqrt{0.061}}$	$\frac{2.731}{\sqrt{0.51}}$	...
1CHA	$\frac{1}{\sqrt{0.524}}$	0	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	$\frac{1.926}{\sqrt{0.061}}$	$\frac{2.500}{\sqrt{0.51}}$	...
1FON	0	$\frac{1}{\sqrt{0.286}}$	0	0	1	0	0	$\frac{2.000}{\sqrt{0.061}}$	$\frac{2.679}{\sqrt{0.51}}$	...
1VAU	0	0	$\frac{1}{\sqrt{0.19}}$	0	0	$\frac{1}{\sqrt{0.238}}$	0	$\frac{1.963}{\sqrt{0.061}}$	$\frac{1.680}{\sqrt{0.51}}$	...
1DAM	$\frac{1}{\sqrt{0.524}}$	0	0	$\frac{1}{\sqrt{0.333}}$	0	0	0	$\frac{2.071}{\sqrt{0.061}}$	$\frac{3.036}{\sqrt{0.51}}$	...
...	...	...	...	...	...	...	...	...	...	...

- Structuration naturelle des colonnes en 31 groupes (de taille 3, 4, 1, ..., 1).
- Sparse K-means pour sélectionner les groupes i.e. les variables qualitatives ou quantitatives.

Package R vimpclust (vignette)

```
res <- sparsewkm(X = wine, centers = 4)
plot(res, what="weights.features")
plot(res, what="expl.var")
```



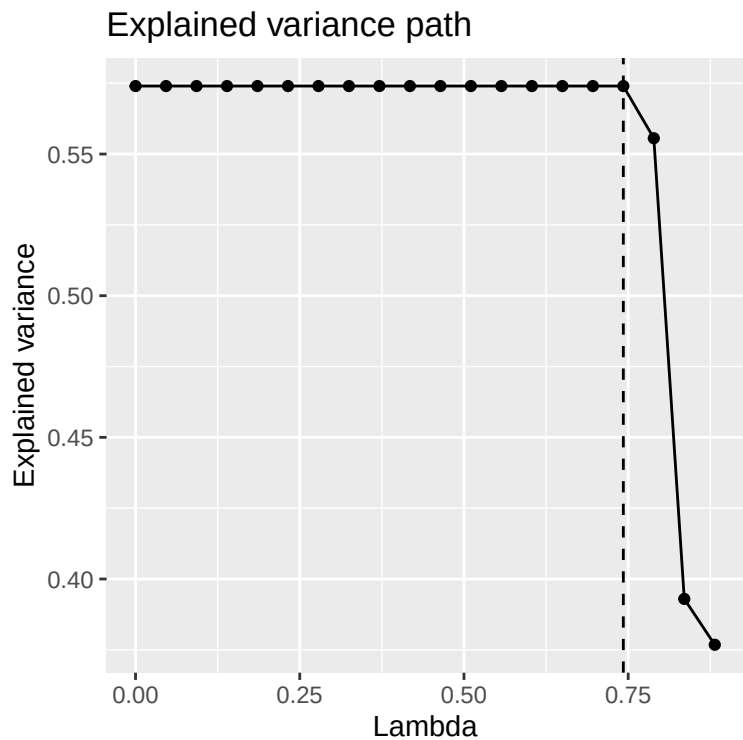


TABLE – 10 variables sélectionnées

Aroma.quality.before.shaking	Quality.of.odour	Balance	Intensity	Overall.quality
Surface.feeling	Aroma.quality	Smooth	Harmony	Typical

1. **Clustering sur des données mixtes**

2. **Clustering avec sélection de variables sur des données mixtes**

3. **Clustering avec sélection de groupes de variables sur des données mixtes**

Clustering de variables mixtes (Chavent, Kuentz, Liquet & Sarraco, 2012)

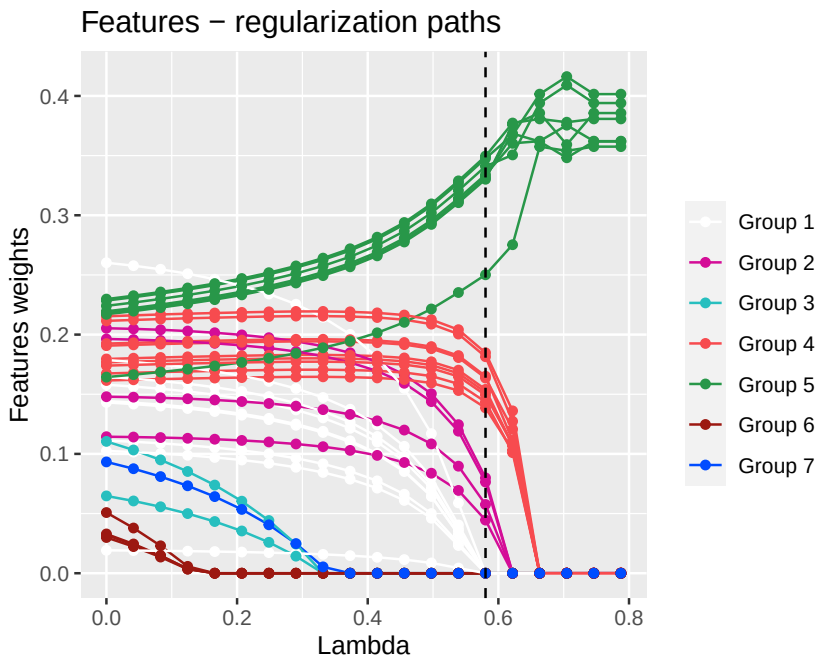
```
tree <- hclustvar(X.quanti = wine.quanti, X.quali = wine.quali)
plot(tree)
```

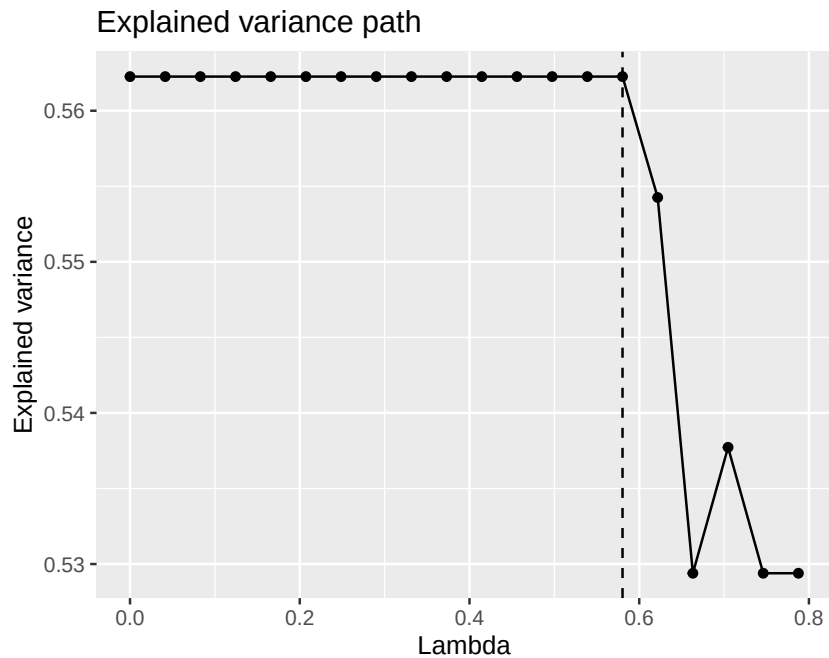

```
P7 <- cutreevar(obj = tree, k = 7)
```

$\Rightarrow$ 7 groupes de variables.

Group-sparse K-means (vignette)

```
res <- groupsparsewkm(X, centers = 4, index = groupes)
plot(res, what = "weights.features")
plot(res, what="expl.var")
```





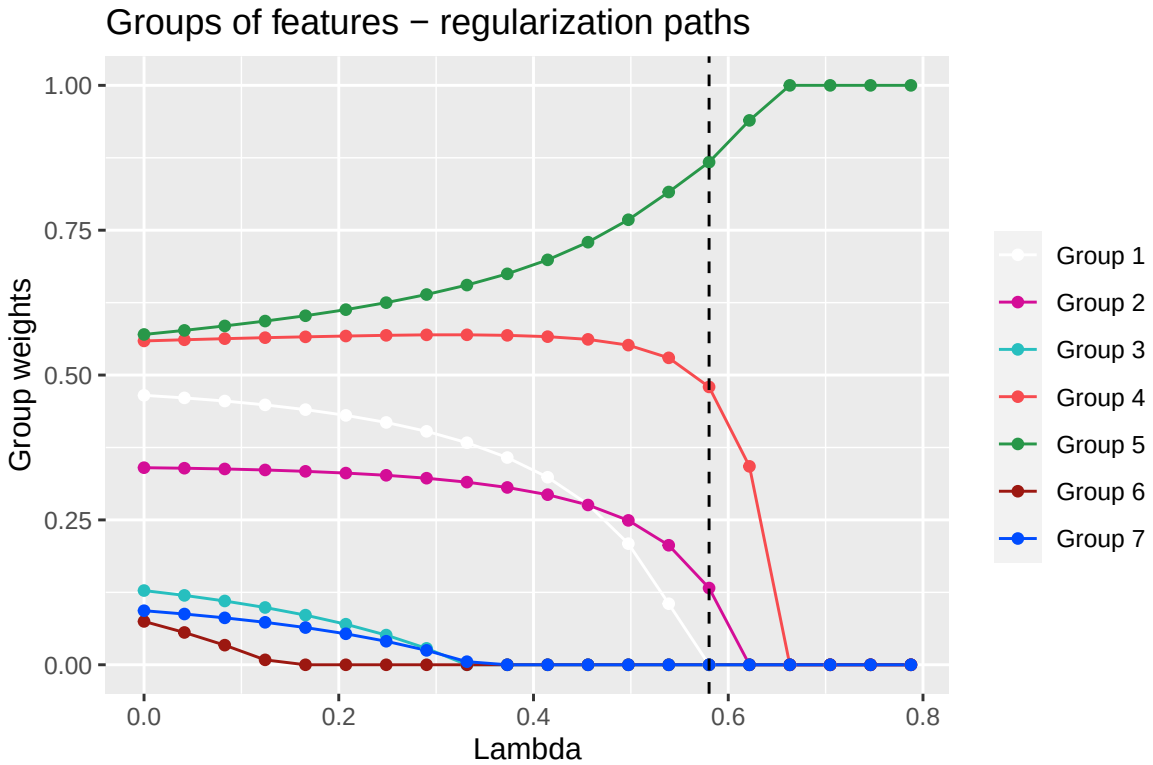


TABLE – 3 groupes sélectionnés

Group 2	Group 4	Group 5
Aroma.quality.before.shaking	Visual.intensity	Plante
Fruity.before.shaking	Nuance	Aroma.quality
Quality.of.odour	Surface.feeling	Balance
Fruity	Aroma.intensity	Smooth
	Aroma.persistency	Harmony
	Attack.intensity	Overall.quality
	Astringency	Typical
	Alcohol	
	Intensity	

References

code implémentant les exemples présentés aujourd'hui :
https ://github.com/MourerAlex/RencontresR2021

Le lien vers notre package R disponible sur le CRAN :
https ://cran.r-project.org/web/packages/vimpclust/index.html

Bibliographie


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


Chavent, M., Kuentz, V., Liquet B., Saracco, J. (2012), ClustOfVar : An R Package for the Clustering of Variables. *Journal of Statistical Software* 50, 1-16.


Witten, D.M., Tibshirani, R., (2010), A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490) :713-726.


Le, S., Josse, J. Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18.