Sparse and group-sparse clustering for mixed data

An illustration of the `vimpclust` package

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

13 juin 2022

1. Inria Bordeaux Sud-Ouest, Équipe ASTRAL - IMB, UMR CNRS 5251, U. Bordeaux

2. SAMM, EA 4543 - Université Paris 1 Panthéon-Sorbonne

3. Safran Aircraft Engines - Datalab - Villaroche

4. CEREMADE, UMR 7534 - Université Paris Dauphine PSL

Introduction

Introduction

1. **Contexte**
   1.1 Données tabulaires :
   - $\mathbf{X}$ une matrice $n \times p$
   - n observations $\mathbf{x}_i$
   - p variables $\mathbf{x}^j$

   1.2 Souvent des données mixtes

   1.3 Possiblement un grand nombre de variables

2. **Objectifs**
   2.1 Production d'une structure de classification - groupes d'individus (clusters):
   - K clusters des n observations
   - $\{C_1, \ldots, C_K\}$

   2.2 Faire de la sélection de variables sur des données mixtes

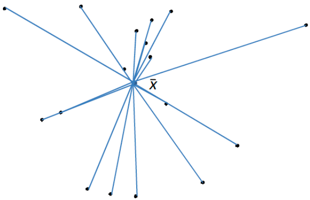
3. **Motivations**
   3.1 Les clusters sous-jacents ne diffèrent que suivant certaines variables

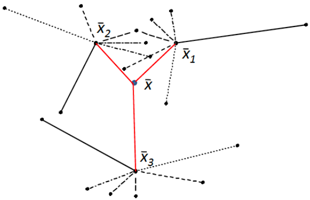
   3.2 Prendre en compte le pouvoir discriminatif de chaque variable

   3.3 Améliorer l'interprétabilité

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

$$\underbrace{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(x_i^j-\bar{x}^j)^2}_{\text{Inertie totale: } t_j}=\underbrace{\sum_{j=1}^{p}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i^j-\bar{x}_k^j)^2}_{\text{Inertie intra: } v_j}+\underbrace{\sum_{j=1}^{p}\frac{1}{n}\sum_{k=1}^{K}n_k(\bar{x}_k^j-\bar{x}^j)^2}_{\text{Inertie inter: } b_j}$$





Minimiser l'inertie intra (homogénéité) $\underset{C_1,\dots,C_K}{\text{minimiser}} \sum_{j=1}^{p} \mathbf{v_j}$

$\Leftrightarrow$

Maximiser l'inertie inter (séparabilité) $\underset{C_1,\dots,C_K}{\text{maximiser}} \sum_{j=1}^{p} \mathbf{b_j}$

$\Leftrightarrow$

L'algorithme des K-means

K-means sparses et groupes de variables

Des K-means aux K-means sparses

K-means[1]: Maximiser la **variance inter-classes par variable** $\underset{C_1,\ldots,C_K}{\text{maximiser}} \sum_{j=1}^{p} \mathbf{b_j}$

critère **pondéré** et **pénalisé**:

$$\sum_{j=1}^{p} \mathbf{w_j b_j} - \lambda \mathbf{h}(\mathbf{w}) = \mathbf{w}^{\intercal}\mathbf{b} - \lambda \mathbf{h}(\mathbf{w})$$

- $\mathbf{w} = (w_1, \ldots, w_p)^{\intercal}$ poids des variables;
- $\mathbf{b} = (b_1, \ldots, b_p)^{\intercal}$ variances inter-classes ;
- $\lambda \geqslant 0$ paramètre de pénalisation et h une fonction croissante des poids.

Les K-means avec sélection de variables, **sparse** (**weighted**) **K-means** [Witten and Tibshirani, 2010]:

$$\underset{C_1,\ldots,C_K,\mathbf{w}}{\text{maximiser}} \ \mathbf{w}^{\intercal}\mathbf{b} - \lambda \|\mathbf{w}\|_1 \ \text{s.c.} \ \|\mathbf{w}\|_2^2 \leqslant 1, w_j \geqslant 0 \ \forall j$$

[1] [Forgy, 1965, MacQueen et al., 1967, Hartigan and Wong, 1979, Lloyd, 1982]

Pénalisation par groupes de variables

p variables divisées en L groupes connus:

- $\mathbf{X} = \left[\mathbf{X}^1 | \dots | \mathbf{X}^L\right]$;
- $\mathbf{X}^l \in \mathbb{R}^{n \times p_l}$;
- $p_1 + \ldots + p_L = p$;
- $\mathbf{b}^\top = (\mathbf{b}_1, \dots, \mathbf{b}_L)$;
- $\mathbf{w}^\top = (\mathbf{w}_1, \dots, \mathbf{w}_L)$

pénalité de groupes ℓ_1 (régression Yuan and Lin [2006]):

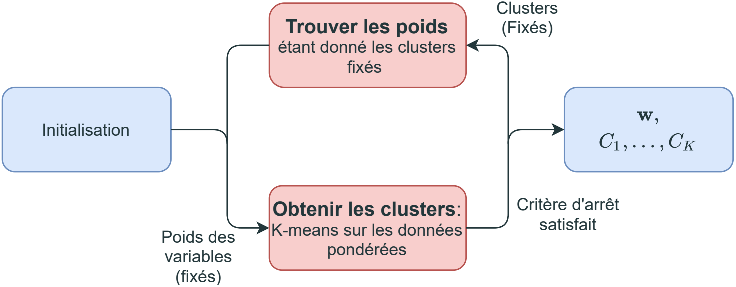
$$h(\mathbf{w}) = \|\mathbf{w}\|_{1,group} = \sum_{l=1}^{L} \sqrt{p_l}\|\mathbf{w}_l\|_2$$

Nouveau problème d'optimisation - **Group-Sparse K-means** Chavent et al. [2020]

$$\underset{C_1,\dots,C_K,\mathbf{w}}{\text{maximiser}}\ \mathbf{w}^T\mathbf{b} - \lambda \sum_{l=1}^{L} \sqrt{p_l}\|\mathbf{w}_l\|_2 \text{ s.c. } \|\mathbf{w}\|_2^2 \leq 1,\ w_j \geqslant 0\ \forall j$$

Group-Sparse K-means

$$\underset{C_1,\ldots,C_K,\mathbf{w}}{\text{maximiser}}\ \mathbf{w}^T\mathbf{b} - \lambda \sum_{l=1}^{L} \sqrt{p_l}\|\mathbf{w}_l\|_2 \ \text{s.c.}\ \|\mathbf{w}\|_2^2 \leq 1,\ w_j \geqslant 0\ \forall j$$



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

Application à des données mixtes

sparse K-means pour données mixtes

Données mixtes: $[\mathbf{X_1}, \mathbf{X_2}]$

$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{x}^3$	$\mathbf{x}^4$
1.926	2.731	a_1	a_2
1.926	2.500	a_1	b_2
2.000	2.679	b_1	a_2
1.963	1.680	b_1	b_2
2.071	3.036	a_1	b_2
...	...	...	...

$\underbrace{\mathbf{x}^1 \ \mathbf{x}^2}_{\mathbf{X_1}} \quad \underbrace{\mathbf{x}^3 \ \mathbf{x}^4}_{\mathbf{X_2}}$

étape 1

Données numériques : $[\mathbf{X_1}, \mathbf{G}]$

$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{a}_2$	$\mathbf{b}_2$
1.926	2.731	1	0	1	0
1.926	2.500	1	0	0	1
2.000	2.679	0	1	1	0
1.963	1.680	0	1	0	1
2.071	3.036	1	0	0	1
...	...	...	...	...	...

$\underbrace{\mathbf{x}^1 \ \mathbf{x}^2}_{\mathbf{X_1}} \quad \underbrace{\mathbf{a}_1 \ \mathbf{b}_1 \ \mathbf{a}_2 \ \mathbf{b}_2}_{\mathbf{G}}$

étape 2

Fréquences des modalités et variance des variables quantitatives : $\mathbf{M}$

$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{a}_2$	$\mathbf{b}_2$
σ_1^2	σ_2^2	$\frac{n_{3,1}}{n}$	$\frac{n_{3,2}}{n}$	$\frac{n_{4,1}}{n}$	$\frac{n_{4,2}}{n}$

étape 3

Données recodées

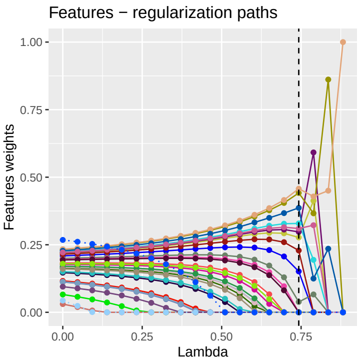
$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{a}_2$	$\mathbf{b}_2$
$\frac{1.926}{\sqrt{\sigma_1^2}}$	$\frac{2.731}{\sqrt{\sigma_2^2}}$	$1/\sqrt{\frac{n_{3,1}}{n}}$	0	$1/\sqrt{\frac{n_{4,1}}{n}}$	0
$\frac{1.926}{\sqrt{\sigma_1^2}}$	$\frac{2.500}{\sqrt{\sigma_2^2}}$	$1/\sqrt{\frac{n_{3,1}}{n}}$	0	0	$1/\sqrt{\frac{n_{4,2}}{n}}$
$\frac{2.000}{\sqrt{\sigma_1^2}}$	$\frac{2.679}{\sqrt{\sigma_2^2}}$	0	$1/\sqrt{\frac{n_{3,2}}{n}}$	$1/\sqrt{\frac{n_{4,1}}{n}}$	0
$\frac{1.963}{\sqrt{\sigma_1^2}}$	$\frac{1.680}{\sqrt{\sigma_2^2}}$	0	$1/\sqrt{\frac{n_{3,2}}{n}}$	0	$1/\sqrt{\frac{n_{4,2}}{n}}$
$\frac{2.071}{\sqrt{\sigma_1^2}}$	$\frac{3.036}{\sqrt{\sigma_2^2}}$	$1/\sqrt{\frac{n_{3,1}}{n}}$	0	0	$1/\sqrt{\frac{n_{4,2}}{n}}$
...	...	...	...	...	...

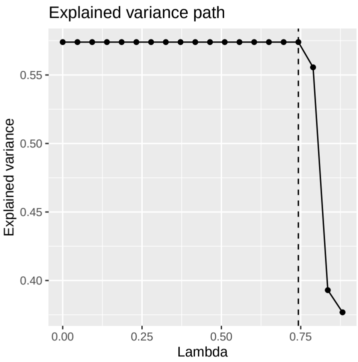
- Problème: sélection colonnes $\rightarrow$ sélection de modalités $\neq$ sélection de variables
- **Structuration naturelle des colonnes**
- But: sélection de variables $\rightarrow$ sélection de colonnes par groupes

Package R vimpclust (vignette)

Données mixtes : 21 vins de Loire, 31 variables, 2 catégorielles

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



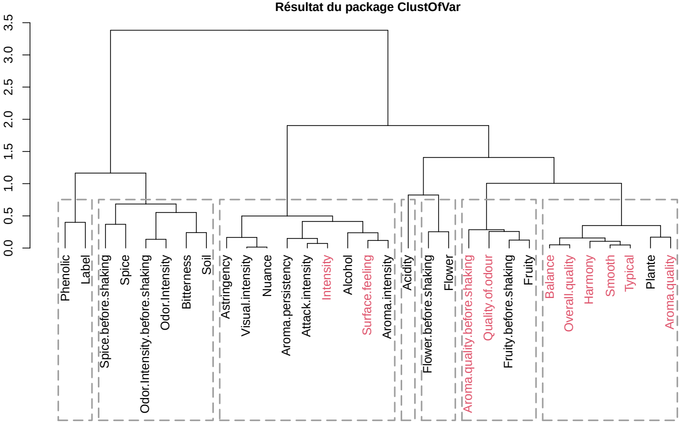


Aroma.quality.before.shaking	Quality.of.odour	Balance	Intensity	Overall.quality
Surface.feeling	Aroma.quality	Smooth	Harmony	Typical

Application à des groupes de variables (mixtes)

Clustering de variables mixtesChavent et al. [2011]

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

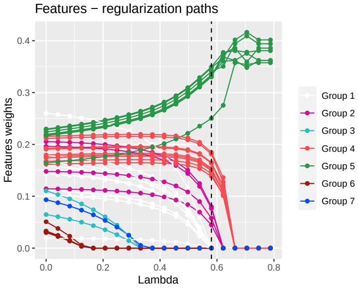


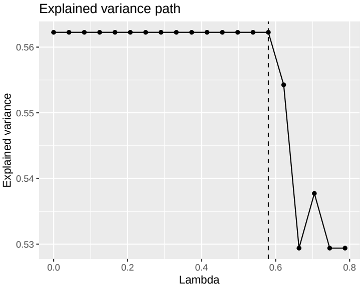

```
P7 <- cutreevar(obj = tree, k = 7)
```

⇒ 7 groupes de variables.

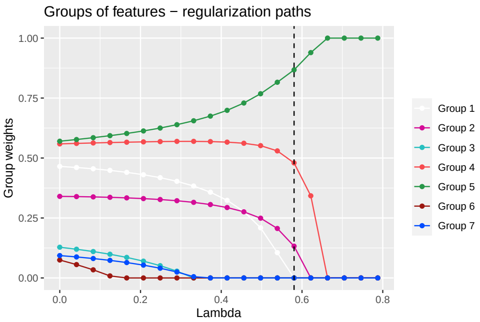
Group-sparse K -means (vignette)

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





Group-sparse K -means (vignette)



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/JDS2022

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

References

Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. ISSN 15579654. doi: 10.1109/TIT.1982.1056489.

Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490): 713–726, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Marie Chavent, Jerome Lacaille, Alex Mourer, and Madalina Olteanu. Sparse k-means for mixed data via group-sparse clustering. In M. Verleysen, editor, *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN): October 2-4, 2020*, pages 235–240, Online event, 2020. European Symposium on Artificial Neural Networks (ESANN), i6doc.com.

Marie Chavent, Vanessa Kuentz, Benoˆıt Liquet, and L Saracco. Clustofvar: An r package for the clustering of variables. *arXiv preprint arXiv:1112.0295*, 2011.