Sparse k-means for Mixed Data via Group-Sparse Clustering

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




1. INRIA Bordeaux Sud-Ouest
CQFD team - 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

ESANN

Problem

- **Objective**: perform **clustering** and **variable selection** on **mixed data** (numerical + categorical).
- We extend the Sparse k-means framework [Witten and Tibshirani, 2010], by adding a pre-processing + L_1-group penalty.

Overview

- **Pre-processing**: Transform each categorical feature into dummy variables and thus define a natural group structure.
- **Methodology**: The specific L_1-group penalty allows us to select variables by group, forcing the model to select or discriminate the entire group.
- **Advantages**: **It improves the clustering on mixed data and its interpretability**.

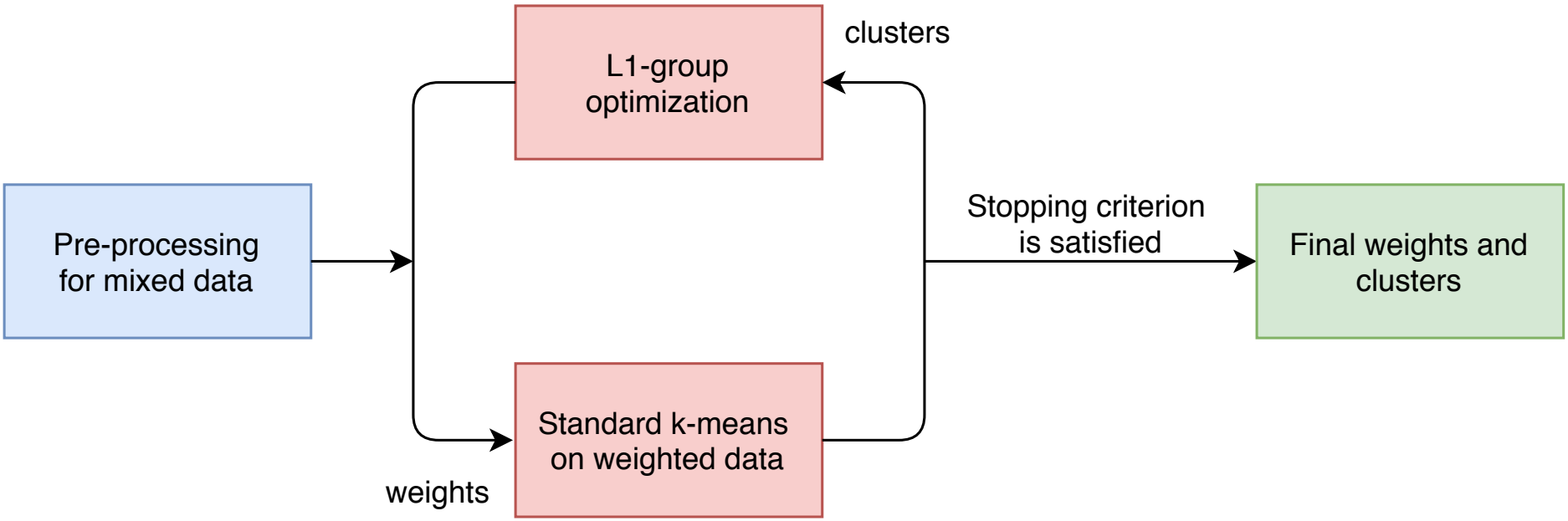
Pre-processing

- **X**: numerical + categorical variables.
- **Transform each categorical feature into dummy variables and thus define a natural group structure**.
- Tranformation:
  - numerical variables: scaled to zero mean + unit variance
  - dummy variables: centered + normalized by $1/\sqrt{\frac{n}{n_{l,j}}}$, where $n_{l,j}$ is the number of input data taking the j-th value of the lth feature [Chavent et al., 2014].

Mathematical formulation

- $\mathbf{X} = \left[\mathbf{X}^1|...|\mathbf{X}^L\right]$.
- $\overline{X}_{j,k} = \frac{1}{n_k} \sum_{i \in C_k} X_{i,j}$.
- $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{i,j}$.
- Beetween-class variance:
  - $\sum_{k=1}^{K} \frac{n_k}{n} \sum_{j=1}^{p} (\overline{X}_{j,k} - \overline{X}_j)^2 = \sum_{j=1}^{p} b_j$.
- L_1-group penalty (regression framework: [Yuan and Lin, 2006]):
  - $\|\mathbf{w}\|_{1,group} = \sum_{\ell=1}^{L} \|\mathbf{w}_\ell\|_2$.
- Penalized criterion:
  - $\max_{\mathbf{w}, C_1, ..., C_K} \mathbf{w}^T \mathbf{b} - \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\mathbf{w}_\ell\|_2$,
  - where the following constraints: $\mathbf{w} \geq 0$, $\|\mathbf{w}\|_2 \leq 1$ are needed to not be reduced to a trivial solution.

Model

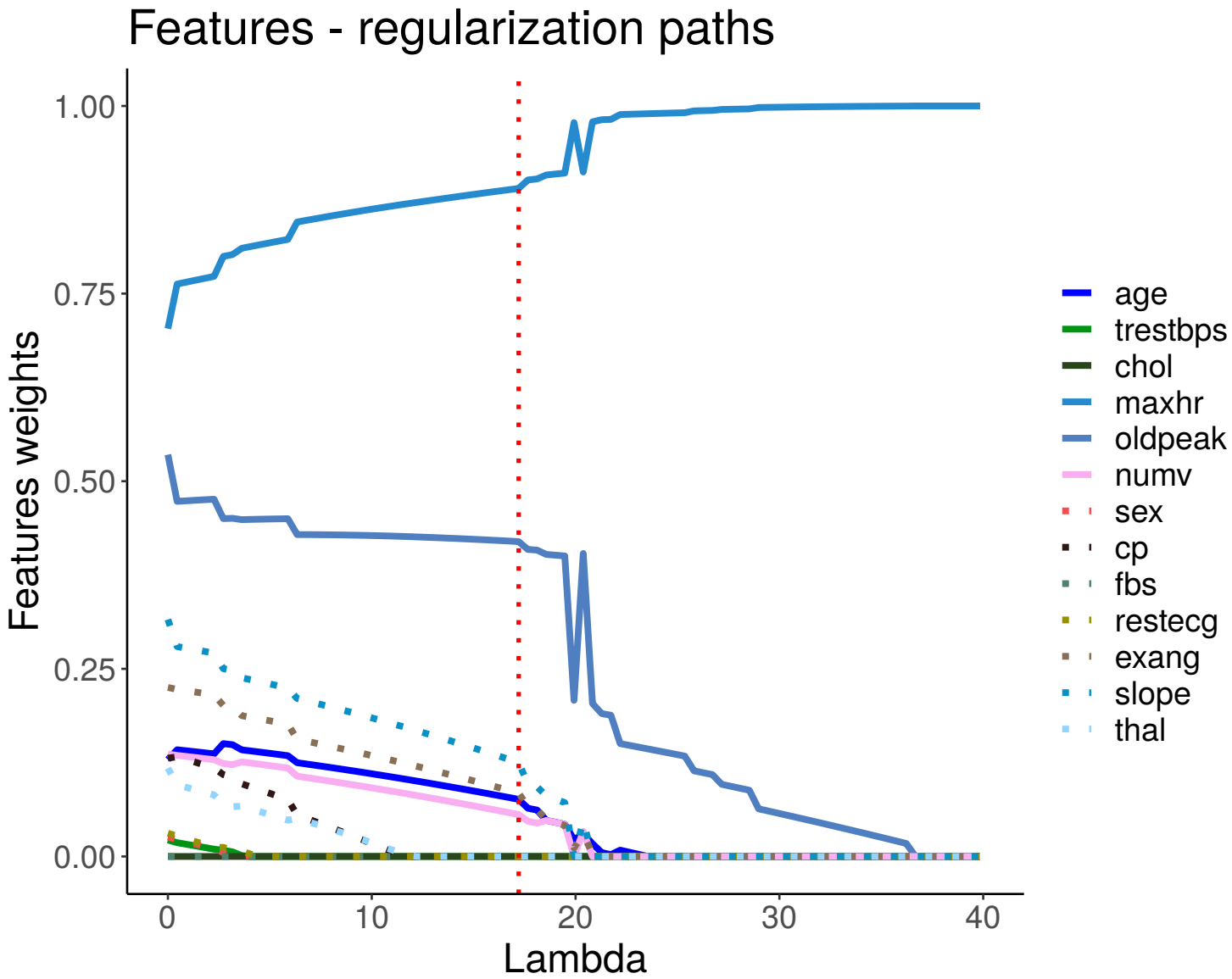

L1-group optimization
clusters
Pre-processing for mixed data
Stopping criterion is satisfied
Final weights and clusters
Standard k-means on weighted data
weights

Application: Data

- Heart dataset[Sta,]
  - $n = 270$.
  - **six numerical features**.
  - **seven categorical features**.
  - One binary control variable. We used this knowledge to select the number of clusters $k = 2$.

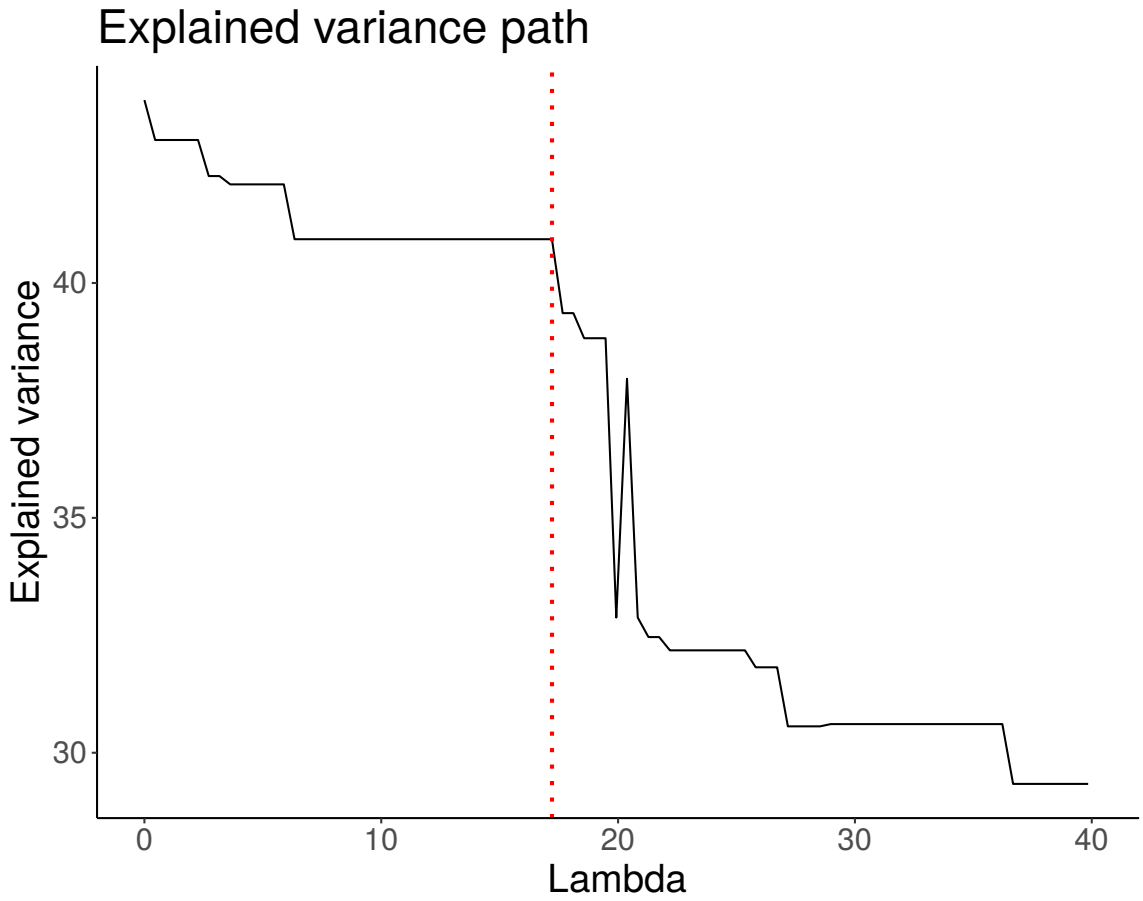
Features weights given λ

$\underline{k = 2}$



Choice of λ

A good choice for λ should preserve a **high percentage of variance explained** by the clustering, while **discarding a large number of features**.



Results

Features are ordered by the decreasing norm of their weights.

Feature	Cluster 1	Cluster 2	Overall statistics
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

Future works

- R-package: https://github.com/chavent/vimpclust
- Criterion to find λ and k based on clustering stability [Mourer et al., 2020].

References

Heart disease data set.
`https://archive.ics.uci.edu/ml/datasets/statlog+(heart)`.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014).
Multivariate analysis of mixed data: The pcamixdata r package.
arXiv preprint arXiv:1411.4911.

Mourer, A., Forest, F., Lebbah, M., Azzag, H., and Lacaille, J. (2020).
Selecting the number of clusters k with a stability trade-off: an internal validation criterion.
arXiv preprint arXiv:2006.08530.

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

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