SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
BEYOND BIG DATA: SHAPING THE FUTURE
VIRTUAL • JUNE 2–4, 2021

Cluster analysis in a multi-block setting

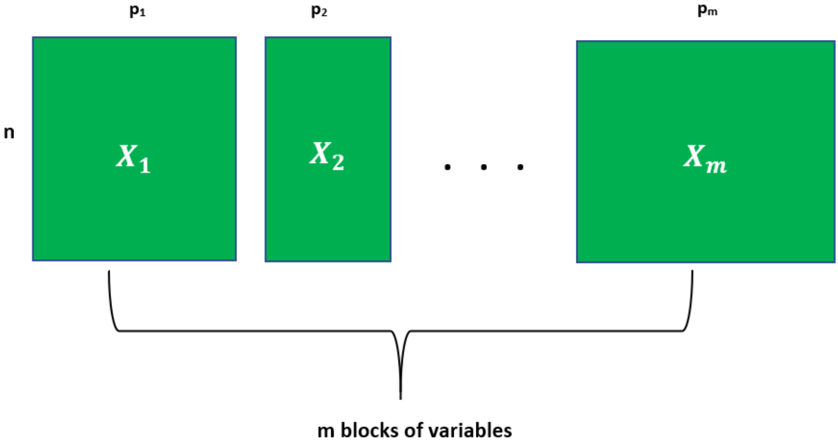
Fabien Llobell, El Mostafa Qannari

XLSTAT by Addinsoft

Oniris ÉCOLE NATIONALE VÉTÉRINAIRE, AGROALIMENTAIRE ET DE L'ALIMENTATION

fllobell@xlstat.com

Data structure



When do we have this structure?

- When there are measurements of different types.
  Examples: measurements on the vegetation of a country, its wealth, the health of residents...

  $\Longrightarrow$ One block by measurement type.

When do we have this structure?

- When there is a repetition of measurements. Example: the weather of each day.

  $\Longrightarrow$ One block by day.

When do we have this structure?

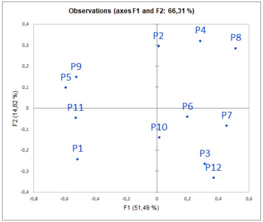
- When measurements are made by different people. Example: sensory data: each participant gives their opinion on their perception of the products.

  $\Longrightarrow$ One block by participant.

Existing exploratory analysis of multi-block data

Aim: Build a map of observations. Some of the proposed methods:

- STATIS
- Generalized Procrustes Analysis
- Multiple Factor Analysis

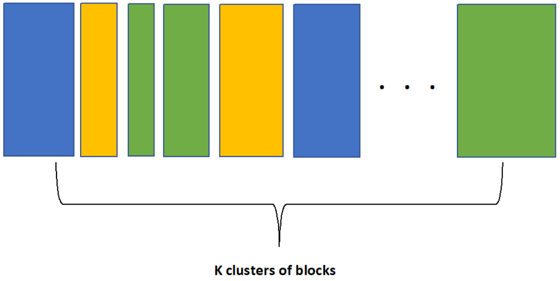


Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994)
Gower, J. C. (1975)
Pagès, J. (2005)

Existing exploratory analysis of multi-block data

Aim: Cluster analysis of the blocks:

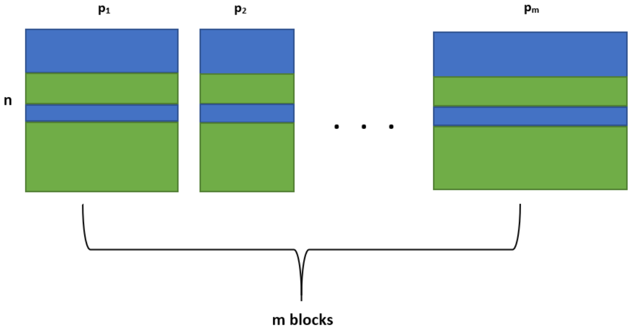
- CLUSTATIS



Llobell, F., & Qannari, E. M. (2020)

Our aim

Cluster analysis of the observations by taking account the multi-block structure:



Existing method

Niang and Ouattara (2019) proposed to use a consensus clustering technique which consist in:

- Perform a clustering of observations within each block
- Choose of a partition in each block
- Set up a consensus partition (by STATIS method)

Drawbacks of such a clustering strategy

- Choosing the number of clusters in each m block (if m is large this can be problematic and time consuming)
- Computation time: m clustering algorithms + 1 algorithm to find the consensus

$\Longrightarrow$We propose a clustering method directly based on the blocks of variables.

Introduction

Method

Illustration

The case of Check-All-That-Apply data
- Illustration
- Discussion: Benefits over the standard method

Conclusion

Preprocessing

- If within a block there are variables on different scales, it is better to standardize the variables of the block.
- Set all the blocks on an equal footing:
  Standardize each block by dividing it by its Frobenius norm:
  $X_l = \frac{X_l}{||X_l||} = \frac{X_l}{\sqrt{trace(X_l X_l^\top)}}$

Minimization criterion

$$D_K = \sum_{k=1}^{K} \sum_{l=1}^{m} \sum_{i \in G_k} ||x_{il} - c_l^{(k)}||^2$$

x_{il}: Observation i in block l
$c_l^{(k)}$: Centroid of cluster G_k in block l
K: Number of clusters

Hierarchical algorithm

$\longrightarrow$ First step: Each observation is a cluster.

$\longrightarrow$ Each intermediate step:
Aggregate the 2 clusters associated with the smallest increase of D_K

$\longrightarrow$ Last step: All the observations are in the same cluster

Partitioning algorithm

We can improve the clustering quality by performing a "consolidation":

$\Longrightarrow$ Use the hierarchical result as initial partition

$\Longrightarrow$ In each block, compute the distance between the observations and the cluster centroids

$\Longrightarrow$ For each observation, sum the distances with the centroids of each block and assign the observation to the nearest cluster.

$\Longrightarrow$ Run the two last steps until convergence

Property

$$\sum_{l=1}^{m} ||X_l - c_l||^2 = D_K + \sum_{k=1}^{K}\sum_{l=1}^{m} n_k ||c_l - c_l^{(k)}||^2$$
$$= D_K + B_K$$

n_k: Number of observations in cluster G_k.
c_l: Centroid of block l
$c_l^{(k)}$: Centroid of cluster G_k in block l

Index

For each block, compute the Between clusters variation/ Total variation:

$$I_l = \frac{\sum_{k=1}^{K} n_k ||c_l - c_l^{(k)}||^2}{||X_l - c_l||^2}$$

Choice of the number of clusters: use of Hartigan index

$$
\begin{aligned}
H(K) &= \left(\frac{\text{Within-clusters variation}_K}{\text{Within-clusters variation}_{K+1}} - 1\right)(n - K - 1) \\
&= \left(\frac{D_K}{D_{K+1}} - 1\right)(n - K - 1)
\end{aligned}
$$

where n is the number of observations and D_K is the criterion with K clusters

Decision: K associated with the maximum of difference between $H(K-1) - H(K)$

Hartigan, J. A. (1975)

Choice of the number of clusters: use of Calinski-Harabasz index

$$CH(K) = \frac{\text{Between-clusters variation}_K \times (n-K)}{\text{Within-clusters variation}_K \times (K-1)} = \frac{\mathrm{B}_K \times (n-K)}{\mathrm{D}_K \times (K-1)}$$

where D_K is the criterion with K clusters and B_K the Between-clusters variation with K clusters

Decision: K associated with the maximum of $CH(K)$

Caliński, T., & Harabasz, J. (1974).

Data description

Life conditions in 540 cities and villages of Gironde (South West of France)

3 blocks of variables:

- Housing (3 variables)
- Employment (9 variables)
- Environment (4 variables)

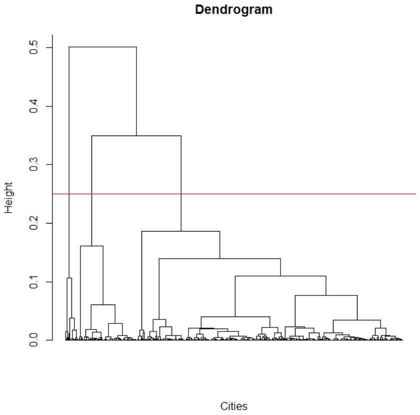
Example: the 4 variables of environment are building, water, vegetation, agriculture. Each variable represents the percentage of land (*i.e.* building land, water land, ...)

Importance of standardization of each block

$\Longrightarrow$ Different scales in the various blocks

$\Longrightarrow$ Different number of variables

Hierarchical algorithm results



Hartigan index suggestion: 3 clusters
Calinski-Harabasz index suggestion: 2 clusters

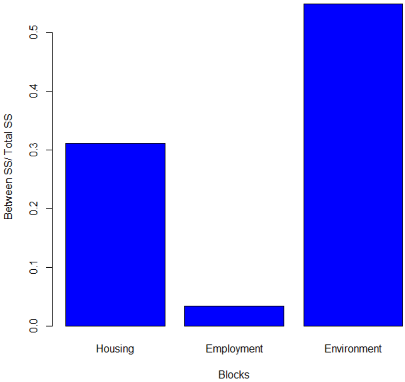
Partitioning algorithm consolidation

$\Longrightarrow$ Initialisation by the hierarchical results in 3 clusters

$\Longrightarrow$ 6% of communes change of cluster

$\Longrightarrow$ The minimization criterion decreases by 3%.

Indices: Between clusters variation/ Total variation



Check-All-That-Apply (CATA) data

Each subject is asked to check the attributes related to each of the given products:

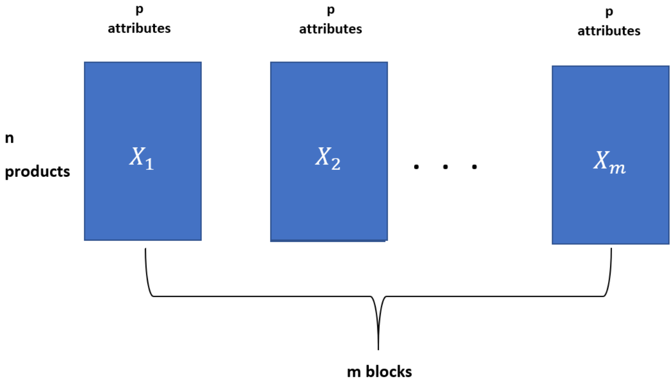
Please, check all the words or phrases which best describe this product:

☐ Sweet	☐ Bitter
☐ Bland	☐ Dry
☐ Sour	☐ Firm
☐ Chewy	☐ Crunchy
☐ Juicy	☐ Mealy
☐ Floral	☐ Soft
☐ Hard	☐ Off flavour

$\Longrightarrow$ One block per subject

Meyners et al., 2013

Data structure



Binary data:

$\Longrightarrow$ 1: Attribute checked

$\Longrightarrow$ 0: Attribute not checked

Data description

- 9 beers
- 15 attributes
- 76 subjects

Attributes: Situations in which the subjects could see themselves drinking the beer: At a party, at a BBQ, while watching TV, at rugby, at fine dining...

Giacalone et al., 2015

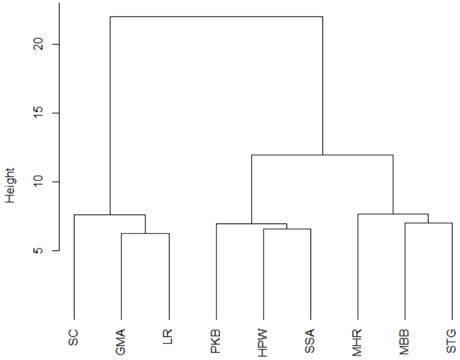
Importance of standardization

Same scale, same number of variables...

But some subjects tend to check a lot of attributes compared to others!

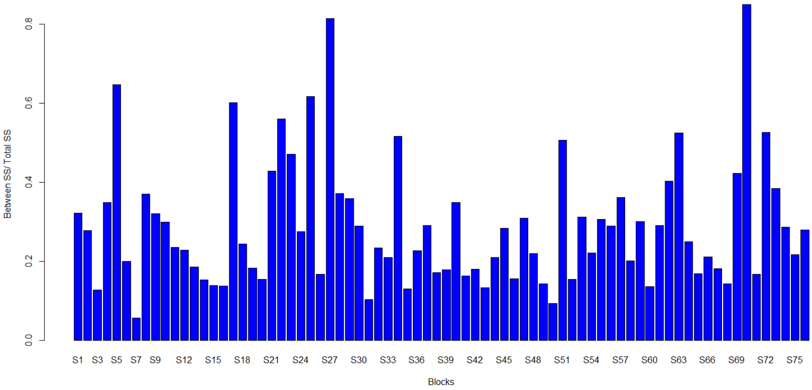
$\Longrightarrow$ The subjects must be put on an equal footing

Hierarchical algorithm results



$\Longrightarrow$ Cut in two clusters and use the partitioning algorithm (no changes)

Indices


Between SS/ Total SS
0.0
0.2
0.4
0.6
0.8
S1 S3 S5 S7 S9 S12 S15 S18 S21 S24 S27 S30 S33 S36 S39 S42 S45 S48 S51 S54 S57 S60 S63 S66 S69 S72 S75
Blocks

Usual approach to clustering products with CATA data

The usual method of clustering products in a CATA experiment:

- Compute the contingency table products $\times$ attributes
- Perform a Correspondence Analysis on this contingency table
- Use the CA axes to perform a cluster analysis

Toy example

	A1	A2	A3	A4
P1	1	1	0	0
P2	1	1	0	0
P3	0	0	1	1
P4	0	0	1	1

	A1	A2	A3	A4
P1	0	0	1	1
P2	0	0	1	1
P3	1	1	0	0
P4	1	1	0	0

	A1	A2	A3	A4
P1	0	0	1	1
P2	0	0	1	1
P3	1	1	0	0
P4	1	0	0	0

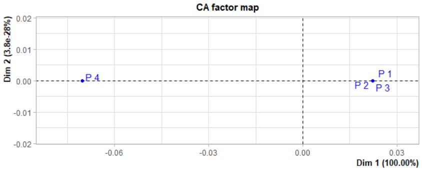
Subjects A, B and C

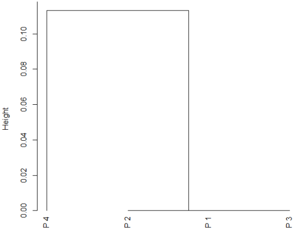
$\Longrightarrow$ 5 subjects A
$\Longrightarrow$ 4 subjects B
$\Longrightarrow$ 1 subject C

	A1	A2	A3	A4
P1	5	5	5	5
P2	5	5	5	5
P3	5	5	5	5
P4	5	4	5	5

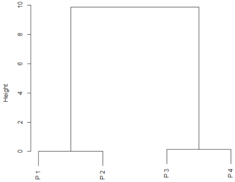
Contingency table

Clustering results





CA on contingency table



Our clustering method

Conclusion

- We have introduced a clustering method of observations in the case of data structured in several blocks of variables
- This method is based on an aggregation criterion similar to Ward's criterion.
- Two algorithms have been proposed
- An aid for choosing the number of clusters has been added
- A clustering quality index within each block has been introduced
- We have investigated the benefits of the method in the specific case of CATA data
- Perspectives: by taking account of the multiblock structure, we could take account of:
  - Specificities of some blocks (*e.g.* categorical variables)
  - Apply specific clustering strategies to some blocks