Weighted consensus clustering for multiblock data

Ndèye Niang [1] Ouattara Mory [2]

[1]Conservatoire National des Arts et Métiers
CNAM, Paris

[2]Université Nangui Abrogoua
UNA, Abidjan

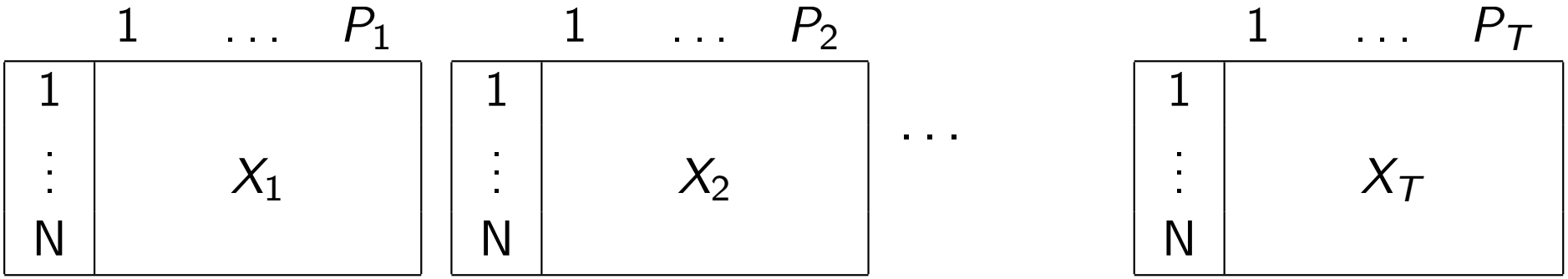
SFC, 2019

Table of Contents

Table of contents

Context

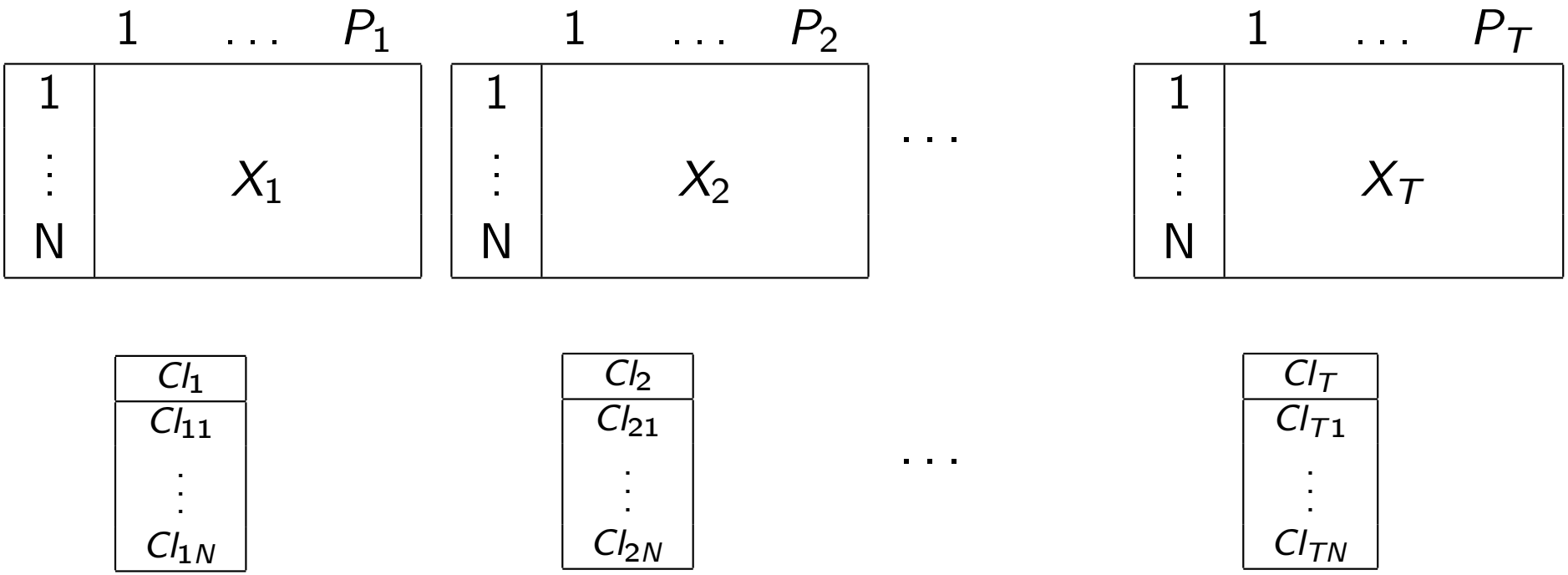
Multiblock data:
Large number of variables organized in homogeneous and meaningful blocks $[\mathbf{X}_1, \ldots, \mathbf{X}_T]$ describing N individuals



Aim:
Clustering individuals taking into account the variables block structure

Context

Agglomerate the observations for each block separately



Aggregate these contributory partitions into a consensus partition

Table of contents

Consensus of Partitions

Several representations of a partition π^t:

- A qualitative variable whose categories are the clusters labels
- Disjonctive matrix H_t of dimension $N \times k_t$
- Co-association matrix of dimension $N \times N$, pairwise similarity

Cl_t

1	1
2	2
3	2
4	1
5	3
6	3

H_t

	H_{t1}	H_{t2}	H_{t3}
1	1	0	0
2	0	1	0
3	0	1	0
4	1	0	0
5	0	0	1
6	0	0	1

W_t

	1	2	3	4	5	6
1	1	0	0	1	0	0
2	0	1	1	0	0	0
3	0	1	1	0	0	0
4	1	0	0	1	0	0
5	0	0	0	0	1	1
6	0	0	0	0	1	1

Consensus de partitions

Consensus clustering - Ensemble cluster Strehl (2002), Li (2008)

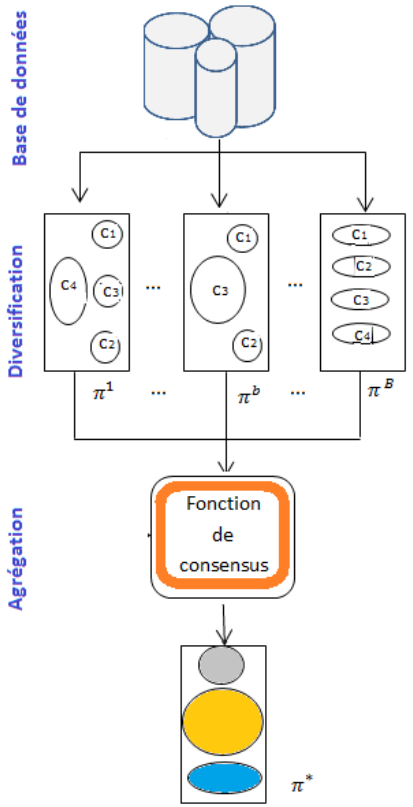
1. **Different partitions**

$$\Pi = \{\pi^t,\ t = 1, \ldots, T\}$$

π^t a partition from block t

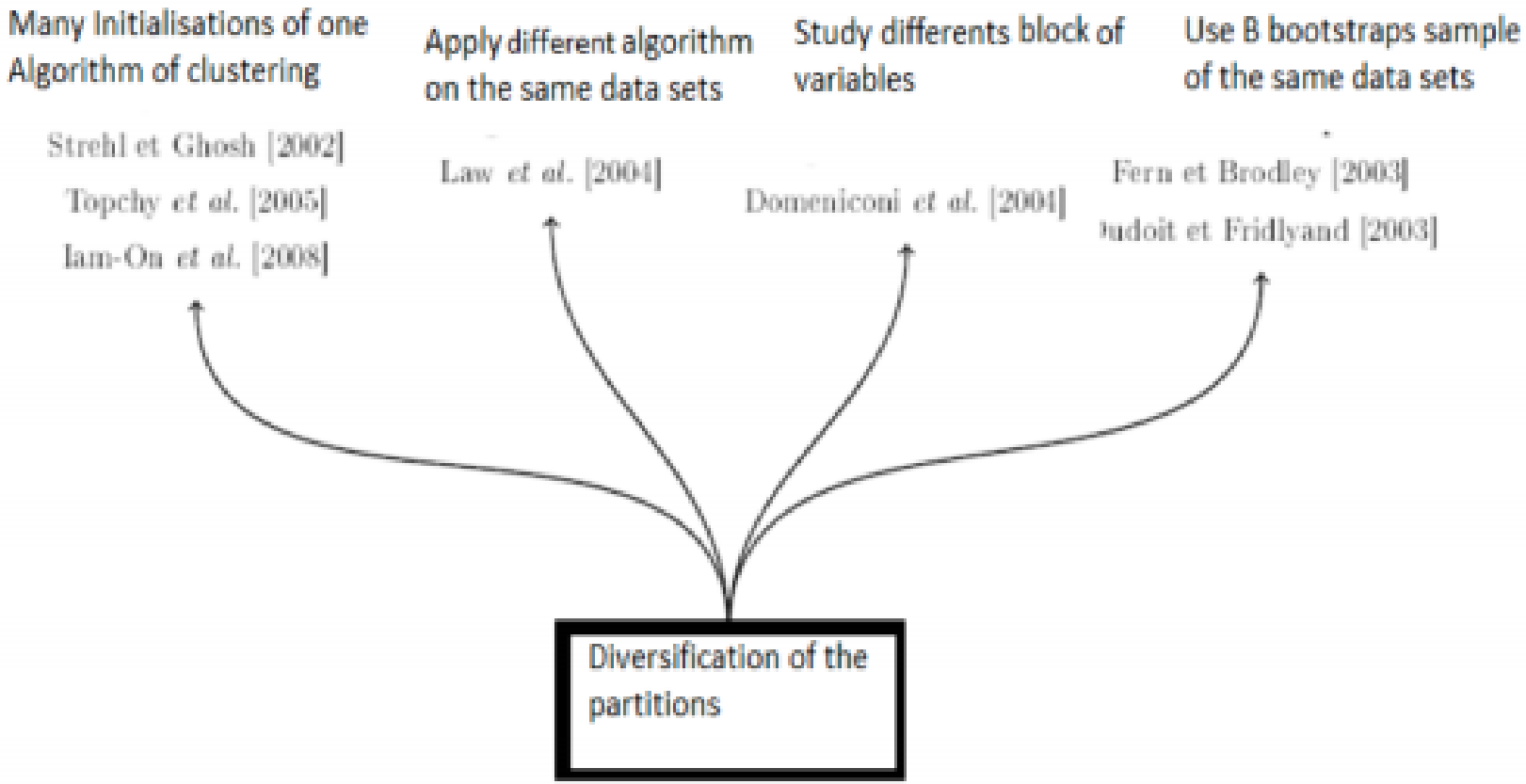
2. **Agregation of partitions**

Find the partition π^* summarizing the best the partitions of Π through a consensus function



Consensus of Partitions - Step 1

Step 1: Diversification for several partitions from



Consensus of Partitions - Step 2

Step 2: Agglomeration of contributory partitions

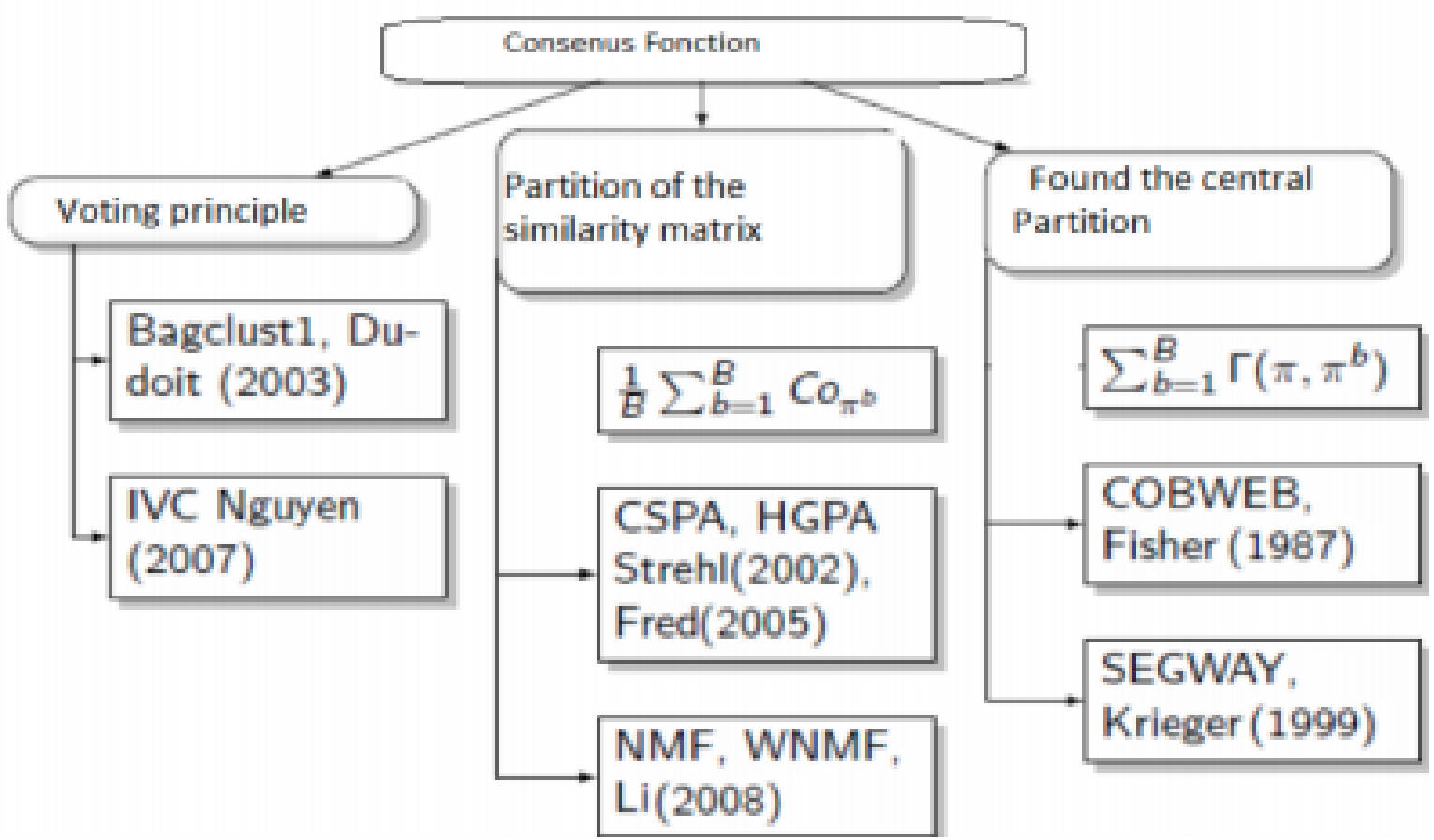
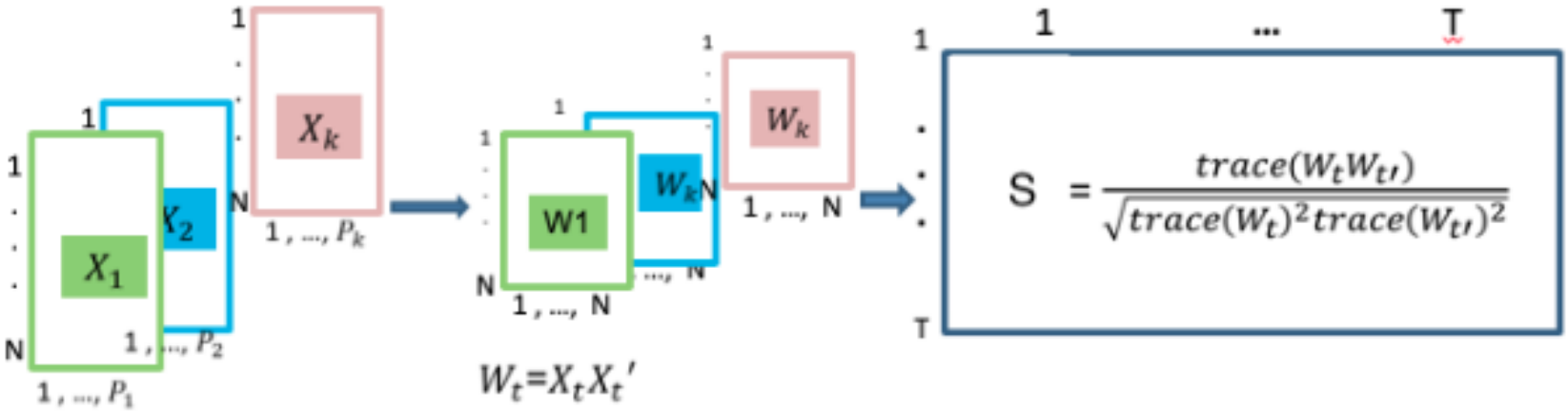


Table of contents

Proposed method: Consensus clustering based on RV index

RV index is a measure of the relationship between two sets of variables X_t and $X_{t'}$ (Robert & Escoufier, 1976).



RV index is non-negative and scaled between 0 and 1; the closer to 1, the more similar the matrices X_t and X_t'.

Consensus clustering based on RV index

1. Find W^* as a consensus matrix from the $W_t = H_t H_t'$
2. W^* is defined as a weighted sum of W_t, $W^* = \sum_t^T \alpha_t W_t$
3. α_t criterion:

$$\max_{\alpha_1,\ldots,\alpha_T} \sum_{t=1}^{T} RV(W_t, W^*) = \max_{\alpha_1,\ldots,\alpha_T} \sum_{t=1}^{T} RV(W_t, \sum_{t=1}^{T} \alpha_t W_t)$$

4. Agglomerative clustering of W^*

Consensus clustering based on RV index

Consensus matrix solution:

- the weights are the coordinates of the first standardized eigenvector α^1 of the $T \times T$ square matrix S whose elements are RV coefficients between every pair of indicator matrices (Lavit, 1984).

Summary

N individuals, T blocks of variables

Step 1 : Separate clustering of the blocks

Step 2 : Computation of S matrix of RV coefficients

Step 3 : Computation of the first eigenvector of S for W^*

Step 4 : Clustering of W^* to find the consensus of clustering

Comparison

CSPA	RVCONS	WNMF
$CO(z_i, z_j) = \frac{1}{M} \sum_{t=1}^{M} W_t(z_i, z_j)$	$w = \sum_{t=1}^{M} \alpha_t^1 w_t$	$\tilde{M} = \sum_{k=1}^{m} w_i M(P_k)$
Based on the CO matrix, **Agglomerative Clustering** allows to find the consensus of clustering	Based on the W matrix, **Agglomerative Clustering** allows to find the consensus of clustering	The minimization of $\min_{w,H} \| \tilde{M} - \tilde{H}\tilde{H}^T \|^2$ allows to find the weighted and consensus of clustering

We compare the performannces of these three algorithms on the data sets

Table of contents

Data

Evaluation of the performance of our algorithm on a synthetic data set (D1) and a real data set (DMU) with labelled individuals in order to have a reference partition.

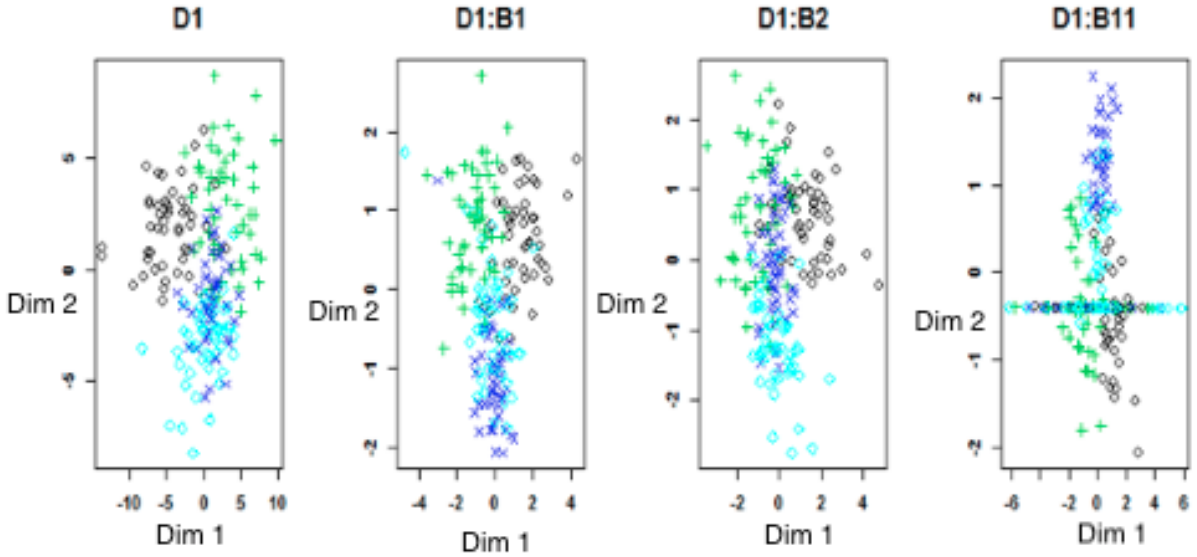
Data	# features	# instance	#classes	#blocks
D1	55	400	4	11
DMU	649	400	10	6

External criteria accuracy and adjusted Rand index to evaluate and compare the consensus clustering methods

Structure of D1 simulated data Set

Ten blocks highly correlated, RV between pair of blocks about 0.80

One more block with a lower RV coefficient equal to 0.01 with the first 10 blocks.



Result D1 (1): The mean (30 runs) Accuracy between the initial partitions and the reference partition

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	Labels
P1	1,00	0,83	0,85	0,78	0,73	0,77	0,82	0,80	0,82	0,74	0,51	0,76
P2	0,83	1,00	0,92	0,84	0,78	0,86	0,90	0,88	0,89	0,83	0,51	0,85
P3	0,85	0,92	1,00	0,87	0,79	0,88	0,91	0,90	0,91	0,84	0,53	0,86
P4	0,78	0,84	0,87	1,00	0,78	0,83	0,85	0,85	0,85	0,78	0,51	0,79
P5	0,73	0,78	0,79	0,78	1,00	0,76	0,80	0,77	0,78	0,73	0,53	0,81
P6	0,77	0,86	0,88	0,83	0,76	1,00	0,86	0,86	0,85	0,81	0,52	0,81
P7	0,82	0,90	0,91	0,85	0,80	0,86	1,00	0,87	0,88	0,84	0,53	0,85
P8	0,80	0,88	0,90	0,85	0,77	0,86	0,87	1,00	0,87	0,81	0,50	0,83
P9	0,82	0,89	0,91	0,85	0,78	0,85	0,88	0,87	1,00	0,82	0,51	0,84
P10	0,74	0,83	0,84	0,78	0,73	0,81	0,84	0,81	0,82	1,00	0,51	0,80
P11	0,51	0,51	0,53	0,51	0,53	0,52	0,53	0,50	0,51	0,51	1,00	0,52

Result D1 (2) : Accuracy and Adjusted Rand index between the initial partitions, the reference partition and the consensus partition

Method	Index	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	Label
CSPA	Acc	0.89	0.97	0.94	0.92	0.88	0.92	0.93	0.95	0.94	0.88	0.62	0.88
	AR	0.91	0.93	0.93	0.92	0.89	0.91	0.89	0.93	0.93	0.86	0.22	0.83
WNMF	Acc	0.80	0.84	0.83	0.80	0.77	0.81	0.83	0.83	0.82	0.79	0.70	0.85
	AR	0.89	0.90	0.90	0.89	0.86	0.88	0.86	0.90	0.90	0.88	0.25	0.80
RVCONS	Acc	0.89	0.97	0.94	0.91	0.88	0.92	0.96	0.95	0.94	0.87	0.62	**0.88**
	AR	0.92	0.93	0.94	0.92	0.89	0.91	0.89	0.93	0.93	0.86	0.22	**0.83**

- RVCONS consensus partition improves the accuracy and Adjusted Rand indices for the reference partition.

Result D1 (3): Consensus coefficients

Table	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}
RVCONS	0.10	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.04
WNMF	0.16	0.08	0.06	0.12	0.06	0.05	0.09	0.09	0.14	0.15	0.41

- RVCONS : equal weights for the 10 highly correlated blocks and a quite zero weight to the noisy block.
- WNMF: the most important weight to the last noisy block which has the lowest similarity with the reference partition.

Structure of DMU data Set : The RV coefficients

	B_1	B_2	B_3	B_4	B_5	B_6
B_1	1	0.985	0.706	0.312	0.967	0.987
B_2	0.985	1	0.678	0.433	0.935	0.965
B_3	0.706	0.678	1	0.199	0.764	0.736
B_4	0.312	0.433	0.199	1	0.273	0.272
B_5	0.967	0.935	0.764	0.273	1	0.979
B_6	0.987	0.965	0.736	0.272	0.979	1

Result DMU (1): The mean (30 runs) Accuracy between the initial partitions and the reference partition

	P1	P2	P3	P4	P5	P6	Label
P1	1	0,72	0,57	0,69	0,46	0,49	0,64
P2	0,72	1	0,71	0,84	0,53	0,53	0,86
P3	0,57	0,71	1	0,66	0,46	0,4	0,7
P4	0,69	0,84	0,66	1	0,53	0,5	0,81
P5	0,46	0,53	0,46	0,53	1	0,46	0,57
P6	0,49	0,53	0,4	0,5	0,46	1	0,51

Result DMU(2) : The mean (30 runs) Accuracy and Adjusted Rand Index of WNMF and RVCONS

Method	Index	P_1	P_2	P_3	P_4	P_5	P_6	Label
CSPA	Acc	0.87	0.93	0.90	0.94	0.86	0.85	0.81
	AR	0.46	0.55	0.71	0.71	0.35	0.31	0.71
WNMF	Acc	**0.87**	0.91	**0.92**	0.91	0.86	0.83	0.80
	AR	**0.47**	0.40	0.65	0.55	**0.47**	0.31	0.70
RVCONS	Acc	0.86	0.93	0.90	0.94	0.86	0.84	0.80
	AR	0.45	0.56	0.74	0.74	0.35	0.31	0.70

- The 3 consensus partitions have quite similar accuracy and Adjusted Rand indices for the reference partition.
- P_1, P_3 and P_5 have greater WNMF accuracy or Adjusted Rand indices than the RVCONS ones

Result DMU(3): Consensus coefficients

Table	W1	W2	W3	W4	W5	W6
RVCONS	0.15	0.19	0.17	0.19	0.16	0.15
WNMF	**0.24**	0.12	**0.26**	0.14	0.14	0.10

- RVCONS : weights for the 6 blocks according more or less to their level of accuracy
- WNMF: the most important weight to a block with one of the lowest accuracy and a block associated with one with the highest accuracy

Données CNL- OQAI

Data collected on several aspects of the dwellings and householders

Structure of the dwellings (Logement) X1	Type of households (Ménage) X2	Habits of inhabitants (Habitude) X3	Pollutants (Polluants) X4

Separately classify dwellings ie locally according to each block

Cl_1	Cl_2	Cl_3	Cl_4
Cl_{11}	Cl_{12}	Cl_{31}	Cl_{41}
.	.	.	.
.	.	.	.
.	.	.	.
Cl_{1N}	Cl_{2N}	Cl_{3N}	Cl_{4N}

Integrated analysis to have a global synthesis using all the available information.

Table of contents

Conclusion - Perspectives

- RVCONS method for clustering multiblock data based on the RV index and a simple eigenvector derivation to define a similarity matrix which used to re-cluster the individuals to find a consensus partition.
- The first results of its application on simulated data as well as on real data show better performances compare to WNMF.
- More formal evaluations on simulated data as well as real one from batch process monitoring.
- Study weights assignment step particulary in case of large number of blocks considering sparsity.

Weighted consensus clustering for multiblock data

Ndèye Niang [1] Ouattara Mory [2]

[1]Conservatoire National des Arts et Métiers
CNAM, Paris

[2]Université Nangui Abrogoua
UNA, Abidjan

SFC, 2019