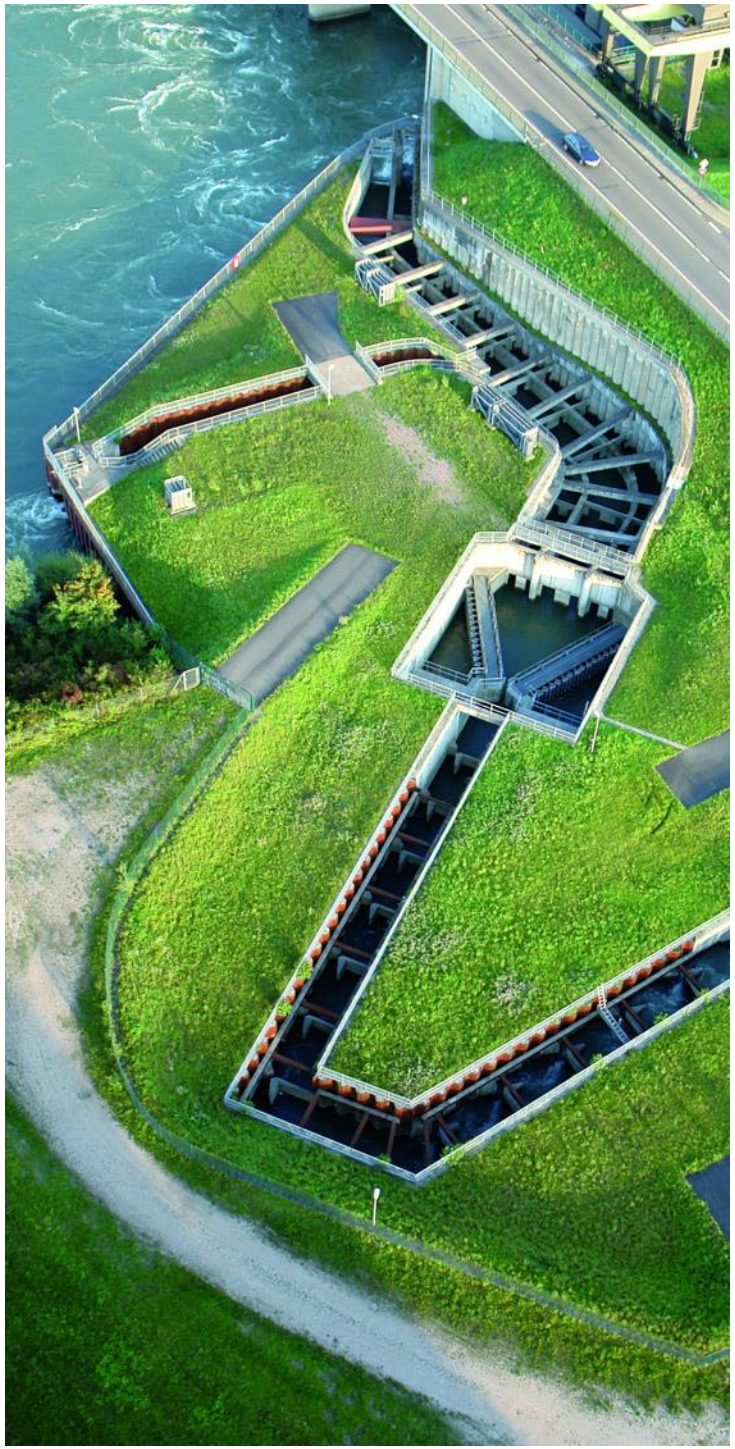
Recherche et Développement

Classification de variables : Une approche dynamique en grande dimension

Christian Derquenne



XXVIèmes Rencontres de la SFC – 5 septembre 2019

Outline

Context and goal

Dynamic approach for clustering variables

Clustering variables in high dimension

Application on simulated data

Contributions, applications, further researches

Context and goal

Importance of understanding the behavior of data …

→ Visualization (plot, scatterplots, map, …)

→ Discovering of structure (anomalies detection, clustering, …)

… before more complex exploitation of data !!!

→ Modelling, forecasting, …

Frame of this presentation:

→ **Discovery of structure**: clustering of variables (linear or not linear relationships)

→ **Usually uncontrolled statistically according** to the data structure

→ **Dynamic approach developed for HAC, DAC and NHNP** allowing to manage the number of groups and their quality [Derquenne (2016, 2017)]

→ Extension of the **dynamic approach in high dimension** [Derquenne (2018,2019)]

Outline

Context and goal

Dynamic approach for clustering variables

Clustering variables in high dimension

Application on simulated data

Contributions, applications, further researches

Dynamic approach

Independence *vs* Unidimensionality

« **Good** » **cluster** = compact cluster
→ Correlated variables + unidimensionality

Unidimensionality ⇒ Dependency between variables within group

Dependency between variables within group ⇏ Unidimensionality

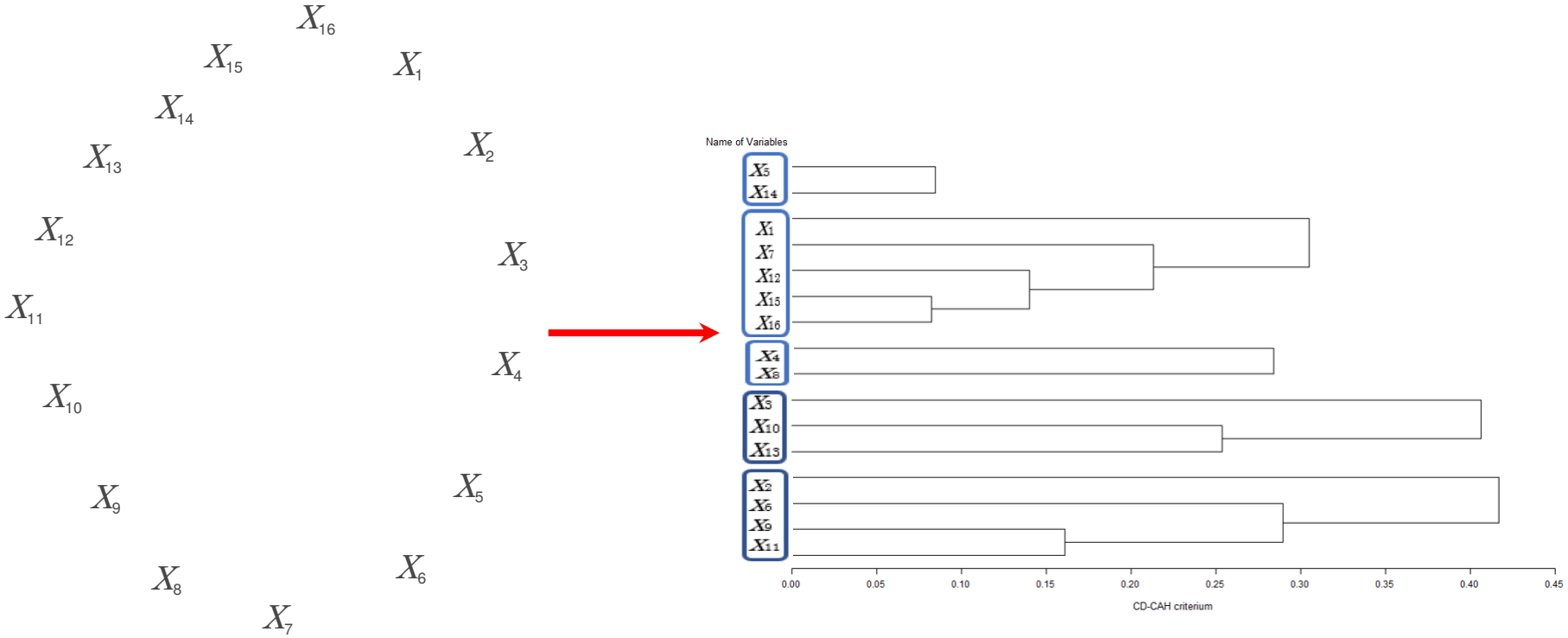
Unidimensionnality / Independence	$\lambda_2 \leq 1$	$\lambda_2 > 1$
$\rho = 0$	**Unidim**	**n.a**
$\rho \neq 0$	**OK**	**Multidim**

Checked by the eigenvalue Saporta test [1999] (*p*-value for the 2sd eigenvalue > α_U given)

Checked by the correlation linear test (*p*-value of the highest correlation < α_C given)

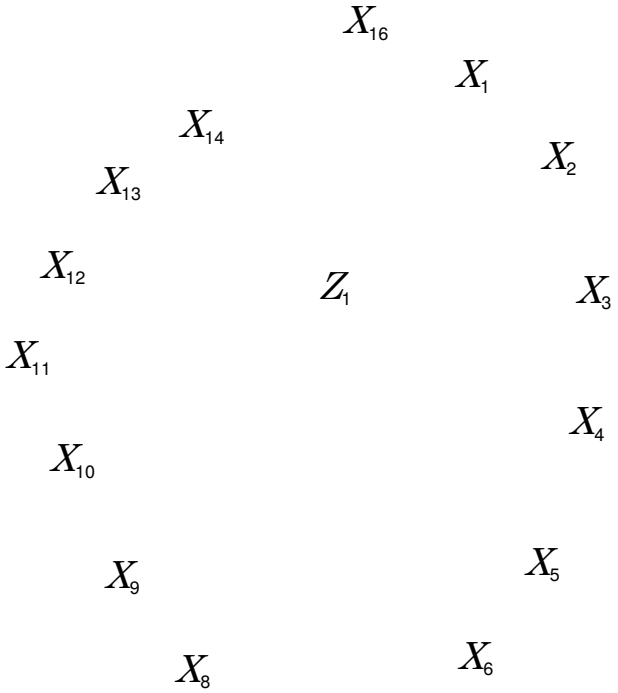
Dynamic approach

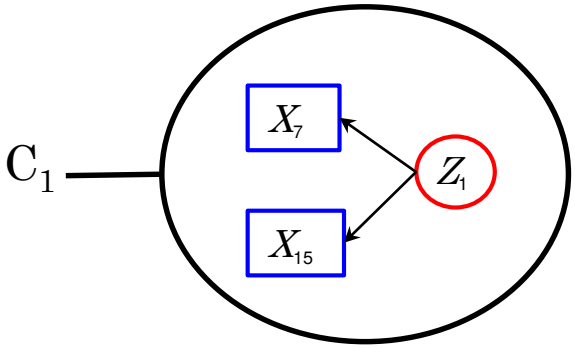
Revisited Hierarchical Ascending Clustering (1)



Dynamic approach

Revisited Hierarchical Ascending Clustering (2)



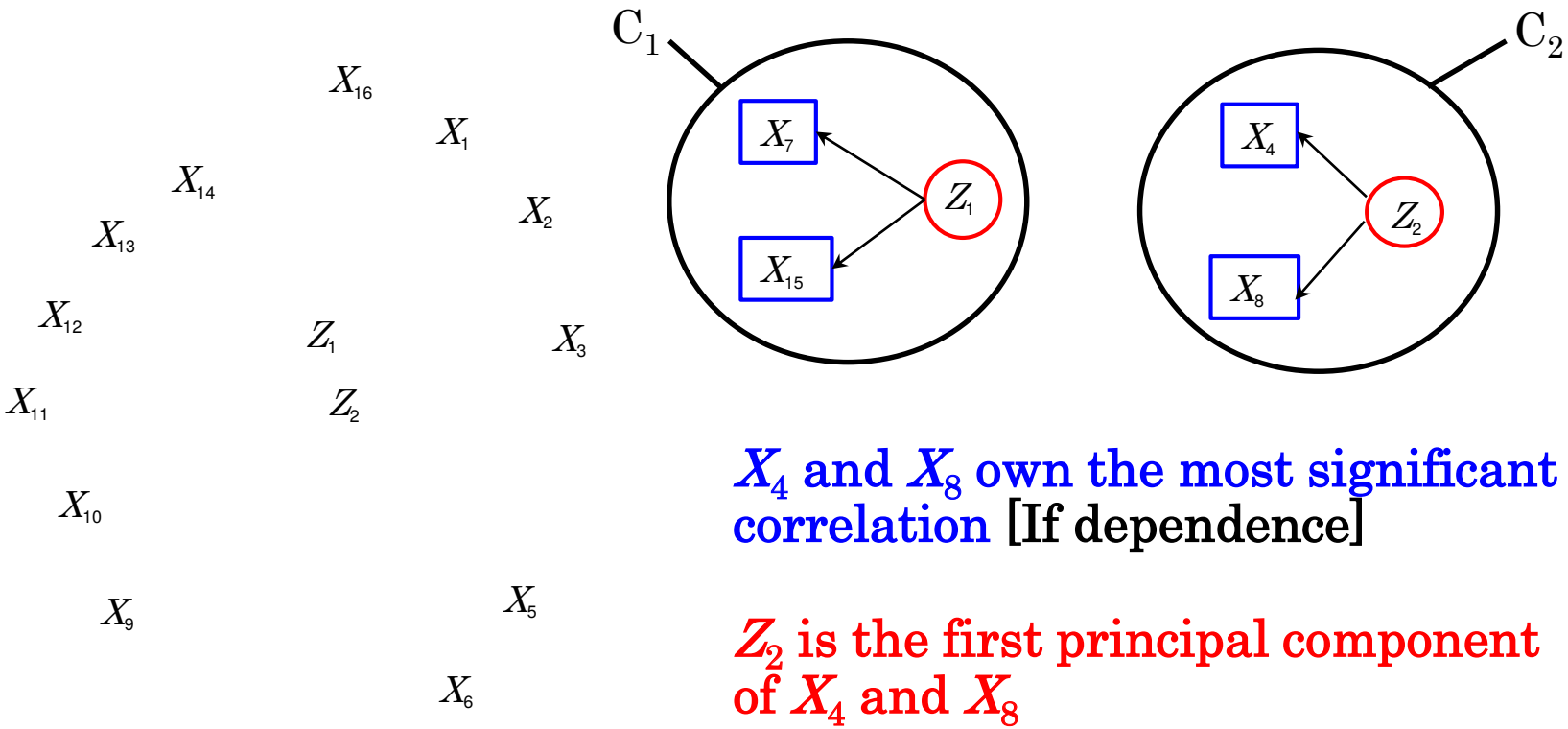


X_7 and X_{15} own the most significant correlation [If dependence]

Z_1 is the first principal component of X_7 and X_{15}

Dynamic approach

Revisited Hierarchical Ascending Clustering (3)

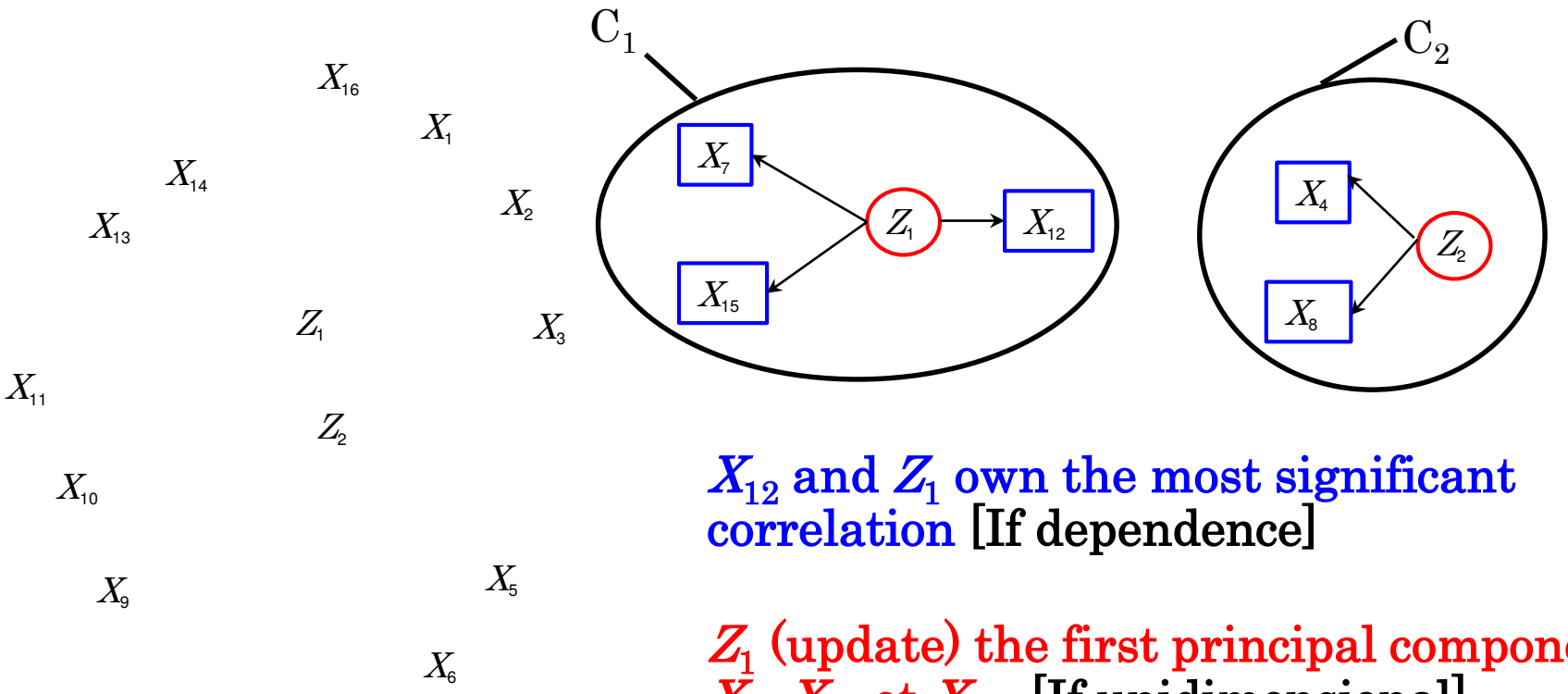


X_4 and X_8 own the most significant correlation [If dependence]

Z_2 is the first principal component of X_4 and X_8

Dynamic approach

Revisited Hierarchical Ascending Clustering (4)

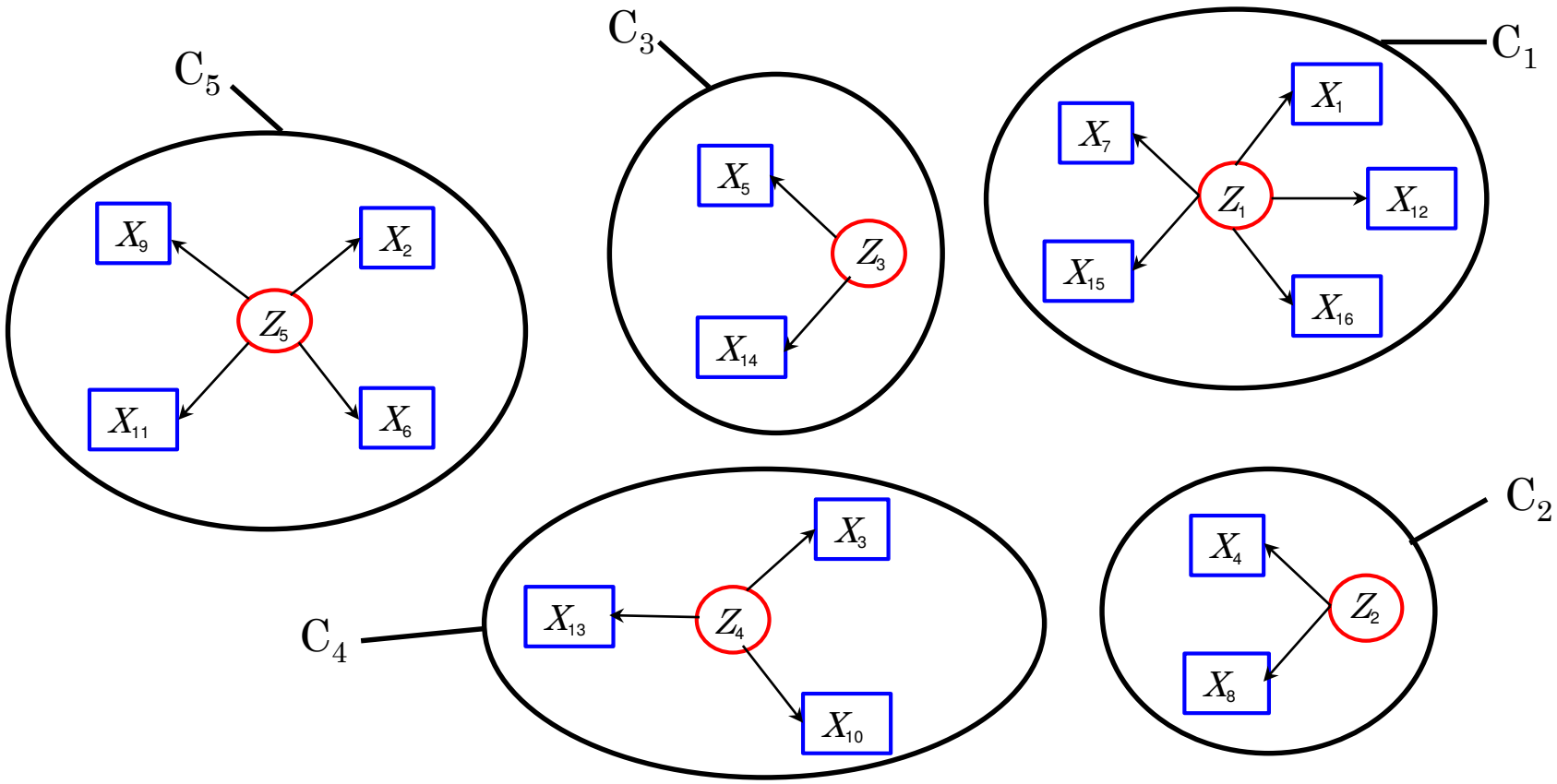


X_{12} and Z_1 own the most significant correlation [If dependence]

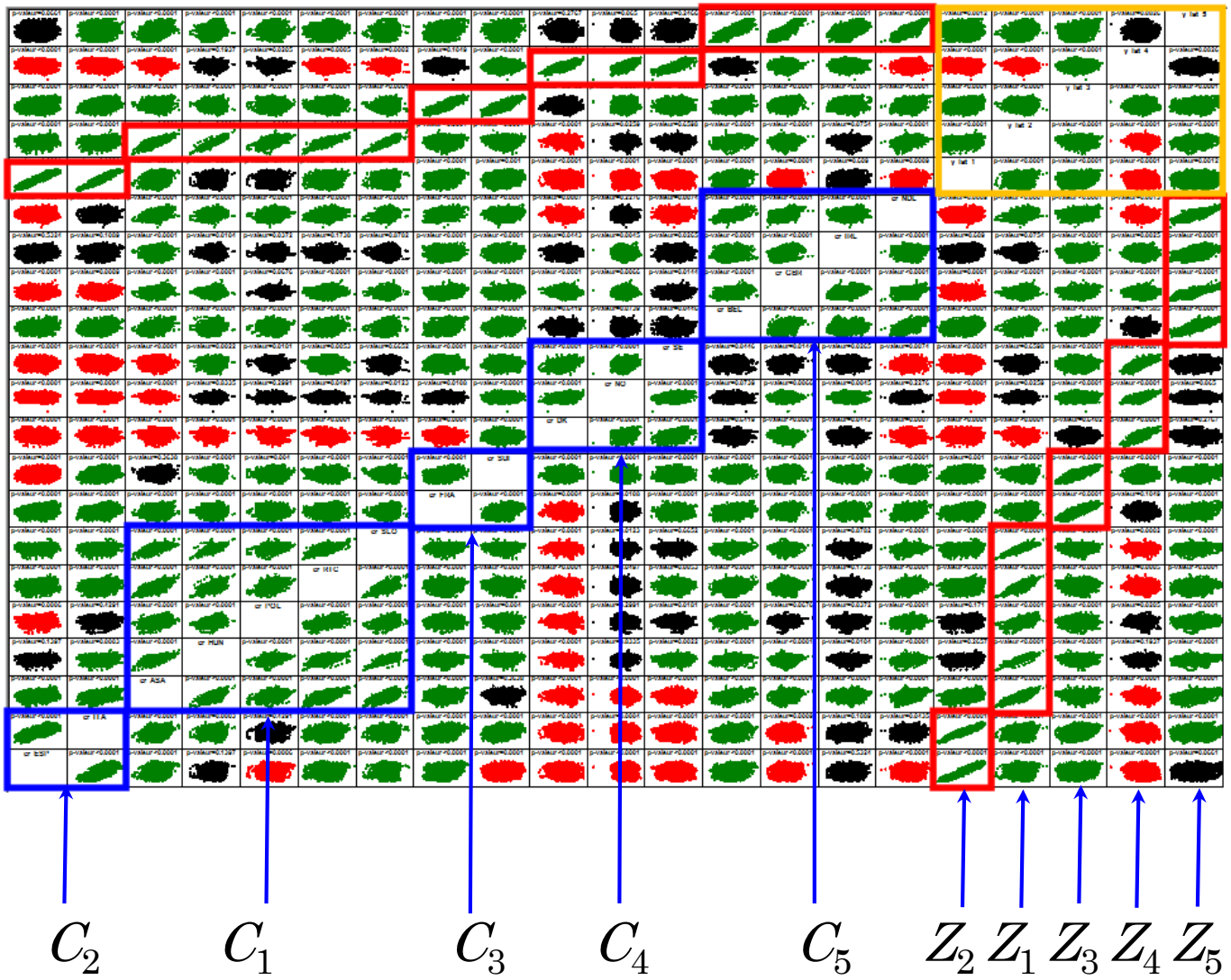
Z_1 (update) the first principal component X_7, X_{12} et X_{15} [If unidimensional]

Dynamic approach

Revisited Hierarchical Ascending Clustering (5)



Dynamic approach



5 clusters

5 latent variables

Dynamic approach

Revisited Hierarchical Ascending Clustering (6)

This **dynamic approach** (linear or nonlinear relationships between variables) [Derquenne, (2016, 2017)] offers an **important gain with respect to other approaches** : VARCLUS [Sarle, (1990)], ClustOfVar [Chavent et al., (2011)], CLV [Vigneau et al., (2003, 2016)], Dirichlet process [Chen et al., (2016)] and Ward or maximum link on dissimilarities matrix **in terms of detection of number of clusters and contents of clusters** (simulation studies)

Outline

Context and goal

Dynamic approach for clustering variables

Clustering variables in high dimension

Application on simulated data

Contributions, applications, further researches

Clustering variables in high dimension

Problem, postulat and proposition (1)

→The most methods of clustering well-perform when the number n (<10000) of individuals and the number p (<100) of variables have a reasonable size

→ If the dataset contains a lot variables and/or a lot of individuals then the CPU time and the memory size of computer will be affected

→ How to take into account high dimension (big data) when we apply a classical statistical method ?

Problem, postulat and proposition (2)

→ **Postulat 1**: if a method provides good results on one sample coming from to an entire dataset (big dataset), there is not a reason that it gives bad results on other samples of this same big dataset

→ **Postulat 2**: if we combine all results of several samples coming from the entire dataset with aid of a good process, the aggregated results must be comparable to the global result if we had applied the statistical method on the overall dataset

→ **Postulat 3**: it should be noted, however, that this process works as long as the data division process and the aggregation strategy are correct

→ This process is based on the « **Divide and Conquer Principle** » **(DCP)**

Clustering variables in high dimension

Divide and Conquer Principle (DCP)

(i) The entire dataset is splitted in S samples

(ii) Each sample is processed in parallel

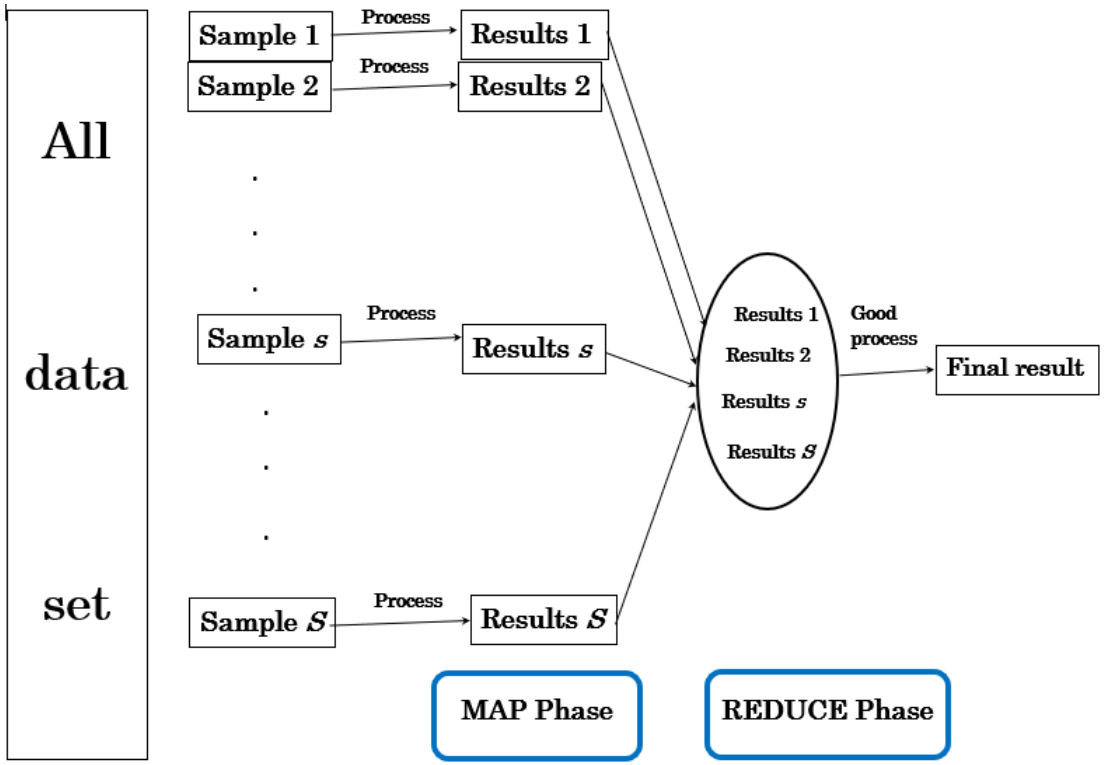
(iii) The results are combined and they are processed

(iv) The final result is obtained

DCP is based on **MapReduce**

→ Generic method to deal with massive datasets stored a distributed file system

→ Developped by Google™ [Dean et al., 2004]



Clustering variables in high dimension

Application to dynamic approach clustering variables

Let's be $X_1, \ldots, X_j, \ldots, X_p$, p (>> 100) numeric variables, such as $X_j \in \mathbb{R}^n$ where n (>> 10000) is the number of individuals contained in the entire dataset E

Let's be $E_1, \ldots, E_s, \ldots, E_S$, S random samples without replacement of n individuals such as $E = \bigcup_{s=1}^{S} E_s$ where $\text{card}(E_s) = n_s$ and $\sum_{s=1}^{S} n_s = n$

Each sample E_s is cut in L random samples without replacement of p variables, $Q_1, \ldots, Q_l, \ldots, Q_L$ where $\text{card}(Q_l) = p_l$ and $\sum_{l=1}^{L} p_l = p$

Remarks: The cutting of L random samples of variables can be different or not for each sample of individuals: E_s

T_{sl} corresponds to a sub-dataset of n_s individuals and p_l variables

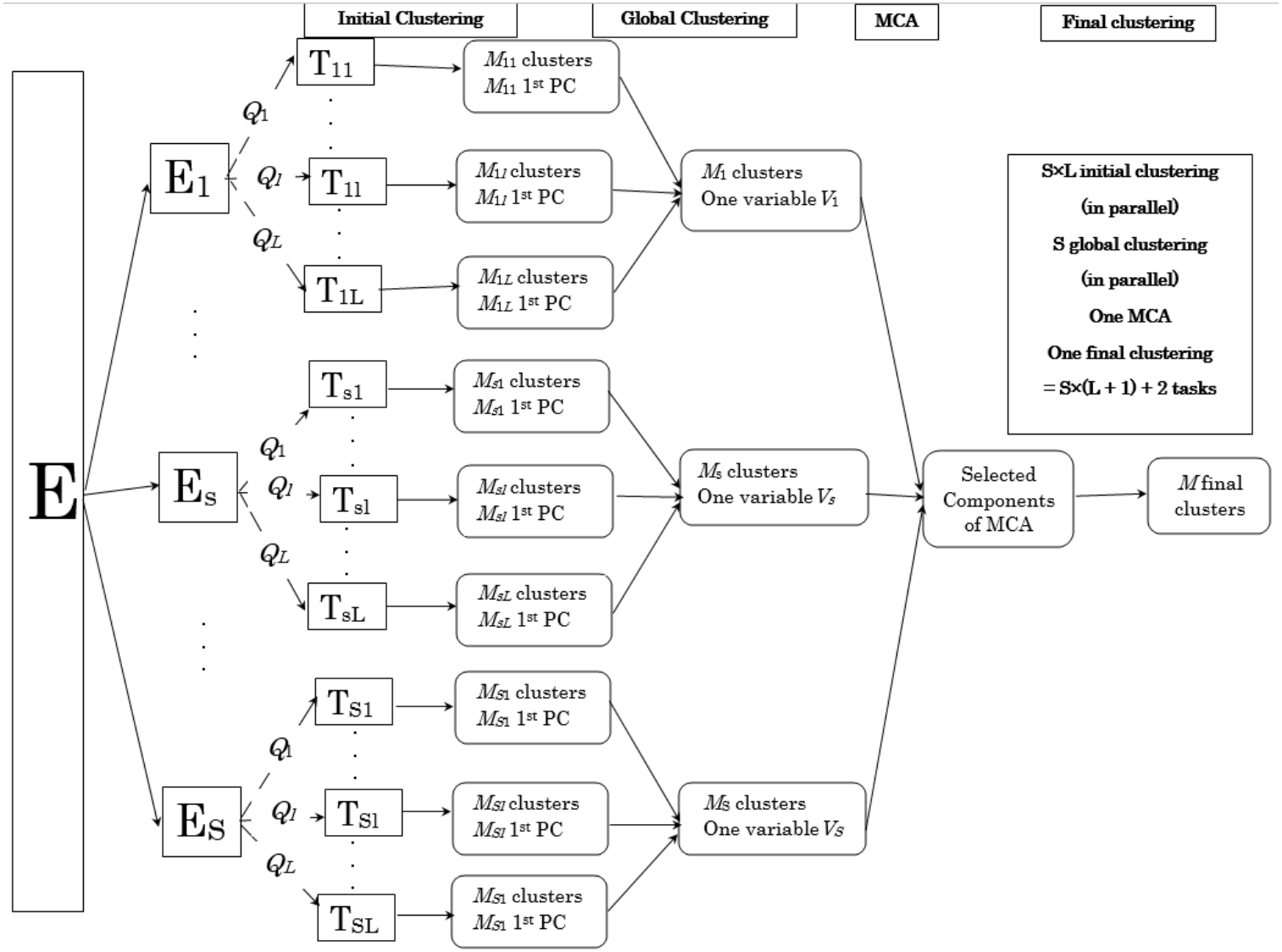
Clustering variables in high dimension

Application to dynamic approach clustering variables

Variables / Individuals	Q_1	...	Q_l	...	Q_L	
E_1	T_{11}	...		...		$\}\ n_1$
⋮		⋱				⋮
E_s		...	T_{sl}	...		$\}\ n_s$
⋮				⋱		⋮
E_S		...		...	T_{SL}	$\}\ n_S$
	$\leftrightarrow$ p_1	...	$\leftrightarrow$ p_l	...	$\leftrightarrow$ p_L	

Clustering variables in high dimension

General process



Clustering variables in high dimension

Application to dynamic approach clustering variables

Detailed process (1):

(i) Apply a initial clustering variables method on T_{sl}

→ we obtain M_{sl} clusters and M_{sl} associated first principal components: $Y_1^{(sl)}, \dots, Y_{Msl}^{(sl)}$

The process (i) is made in parallel on each T_{sl} of E_s [MAP Phase]

→ then we obtain: $Y_1^{(s1)}, \dots, Y_{Ms1}^{(s1)}, \dots, Y_1^{(sl)}, \dots, Y_{Msl}^{(sl)}, \dots, Y_1^{(sL)}, \dots, Y_{MsL}^{(sL)}$ first principal components

(ii) Apply again a global clustering variables on the previous first principal components

→ we obtain M_s new clusters $(C_1^{(s)}, \dots, C_k^{(s)}, \dots, C_{Ms}^{(s)})$ and M_s associated new first principal components: $Z_1^{(s)}, \dots, Z_k^{(s)}, \dots, Z_{Ms}^{(s)}$

→ then each initial variable $X_j \in C_k^{(s)}$

→ let's be $V_s \in \{1, \dots k, \dots, M_s\}$, a new variable containing the cluster numbers of $(C_1^{(s)}, \dots, C_k^{(s)}, \dots, C_{Ms}^{(s)})$ for each initial variable X_j

The process (ii) is made in parallel on each E_s [REDUCE Phase]

Clustering variables in high dimension

Application to dynamic approach clustering variables

Detailed process (2):

(iii) Apply a Multiple Correspondence Analysis (MCA)

→ S variables = cluster numbers: V_1 ,…, V_s ,…, V_S

→ p individuals = names of initial variables: $X_1, \ldots, X_j, \ldots, X_p$

→ Principal components = $U_1, \ldots, U_r, \ldots, U_R$

Variables / Individuals	V_1	. . .	V_s	. . .	V_S
X_1	v_{11}	. . .	v_{s1}	. . .	v_{S1}
. . .		. . .			
X_j	v_{1j}	. . .	v_{sj}	. . .	v_{Sj}
. . .				. . .	
X_p	v_{1p}	. . .	v_{sp}	. . .	v_{Sp}

$\xrightarrow{MCA}$

Principal components / Individuals	U_1	. . .	U_r	. . .	U_R
X_1	u_{11}	. . .	u_{r1}	. . .	u_{R1}
. . .		. . .			
X_j	u_{1j}	. . .	u_{rj}	. . .	u_{Rj}
. . .				. . .	
X_p	u_{1p}	. . .	u_{rp}	. . .	u_{Rp}

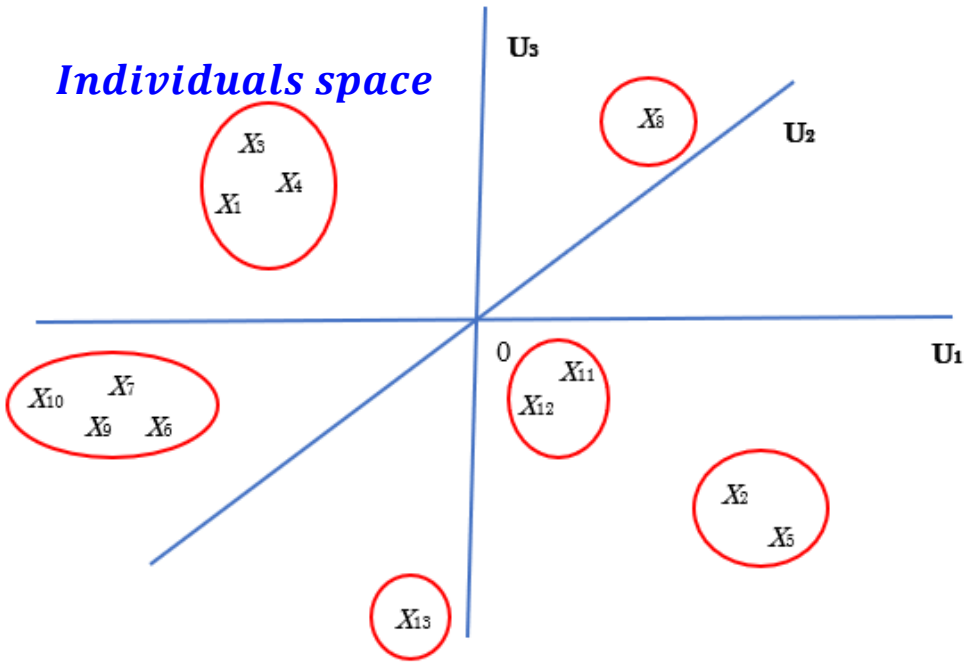
Clustering variables in high dimension

Application to dynamic approach clustering variables

Detailed process (3):

(iv) Apply a final clustering on « significant » principal components of MCA

→ M final clusters $G_1, \dots, G_m, \dots, G_M$ containing $p_1, \dots, p_m, \dots, p_M$ initial variables $X_1, \dots, X_j, \dots, X_p$ with $\sum_{m=1}^{M} p_m = p$ on n individuals



Final clustering →

Final cluster (G)	Variables
G_1	$X_{(1,1)}$ … $X_{(1,j)}$ … $X_{(1,p1)}$
. . .	. . .
G_m	$X_{(m,1)}$ … $X_{(m,j)}$ … $X_{(m,pm)}$
. . .	. . .
G_M	$X_{(M,1)}$ … $X_{(M,j)}$ … $X_{(M,pM)}$

Clustering variables in high dimension

Application to dynamic approach clustering variables

Detailed process (4):

(v) Evaluation of the proposed approach: *two levels*

First level : to evaluate the rebuilding quality of the observed clustering on entire dataset

→ Comparison between the global inertia of observed clustering (*OCI*) and the global inertia of estimated clustering (*ECI*) with the proposed method

$$OCI = \frac{1}{p}\sum_{m=1}^{\widetilde{M}}\sum_{X_j \in \tilde{G}_m} \rho^2\left(X_j, \tilde{Z}_m\right)$$

$$ECI = \frac{1}{p}\sum_{m=1}^{M}\sum_{X_j \in G_m} \rho^2\left(X_j, \hat{Z}_m\right)$$

where $\tilde{G}_m$ and G_m are respectively the observed and the estimated clusters

$\tilde{Z}_m$ and $\hat{Z}_m$ are associated first components - $\widetilde{M}$ and M (obs. and estimated number of clusters)

Second level : to evaluate the quality of the estimated clustering

→ Comparison between the contents of the observed and estimated clustering (Rand, Jaccard, γ, V-Cramer, Ttchuprow, %good classified)

Outline

Context and goal

Dynamic approach for clustering variables principle

Clustering variables in high dimension

Application on simulated data

Contributions, applications, further researches

Application on simulated data

Simulated data

→ n =100,000 observations – p = 1000 variables – 9 clusters

Clusters	1	2	3	4	5	6	7	8	9
Variables	$X_1, X_{501}, \ldots X_{800}$	$X_9, X_{10}, \ldots X_{500}$	$X_5, X_{801}, \ldots X_{1000}$	X_2	X_3	X_4	X_6	X_7	X_8
Simulation	$X_1 \to \mathcal{N}(0,1)$ $\varepsilon_t \to \mathcal{N}(0,1)$ $X_j = X_1 + 2\varepsilon_t$ $j = 501,\ldots,800$	$X_9 \to \mathcal{N}(0,1)$ $\varepsilon_t \to \mathcal{N}(0,1)$ $X_j = 2X_9 + 2\varepsilon_t$ $j = 10,\ldots,500$	$X_5 \to \mathcal{N}(0,1)$ $\varepsilon_t \to \mathcal{N}(0,1)$ $X_j = 0.1X_5 + \varepsilon_t$ $j = 801,\ldots,1000$	$X_2 \to \mathcal{N}(0,1)$	$X_3 \to \mathcal{N}(0,1)$	$X_4 \to \mathcal{N}(0,1)$	$X_6 \to \mathcal{N}(0,1)$	$X_7 \to \mathcal{N}(0,1)$	$X_8 \to \mathcal{N}(0,1)$

Building of sub-datasets and process

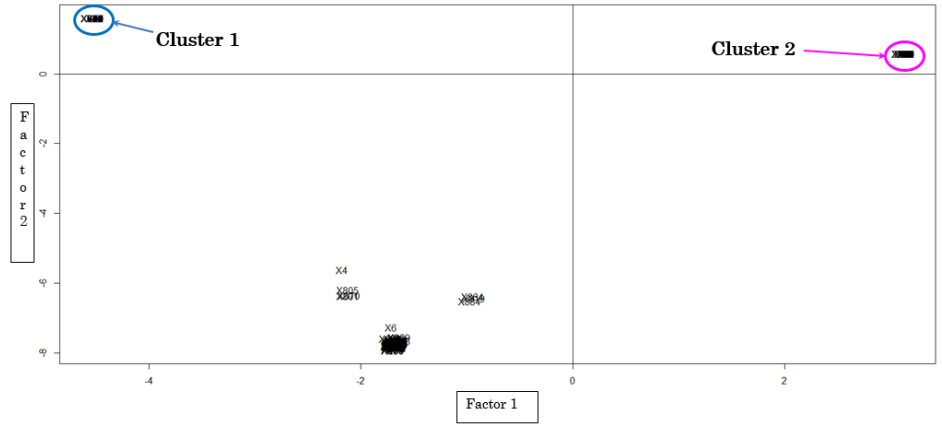
→ 10 random samples without replacement E_s of individuals (n_s = 10000)

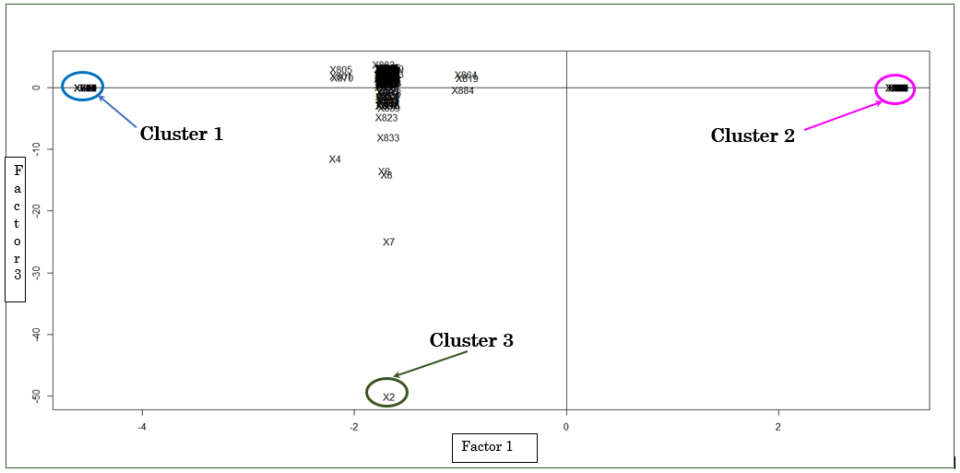
→ 10 random samples without replacement Q_l of variables (p_l = 100)

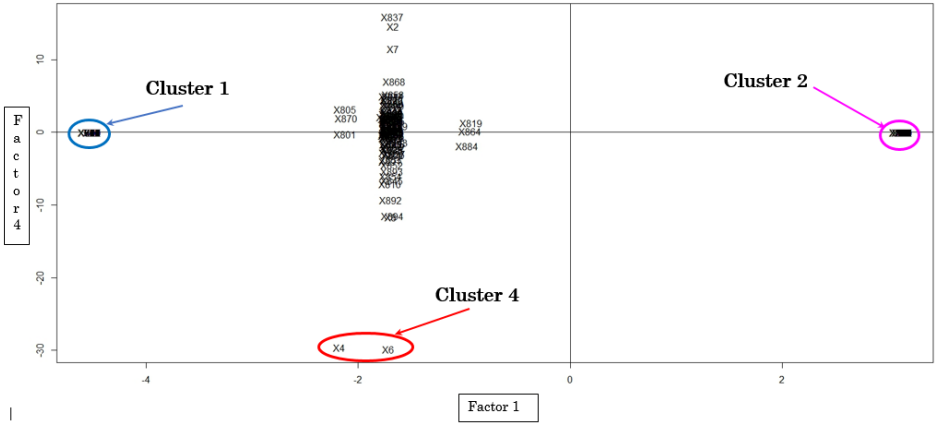
→ 100 initial clustering in parallel + 10 global clustering in parallel + one MCA + one final clustering = 112 tasks

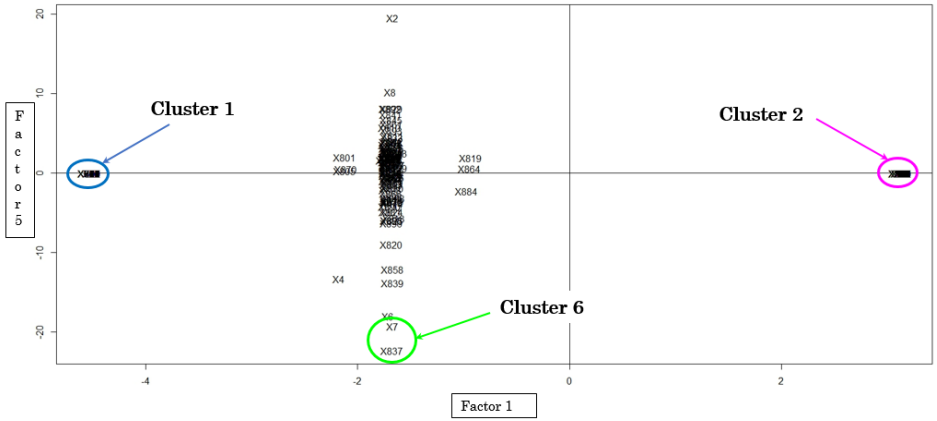
Application on simulated data

Results (1)



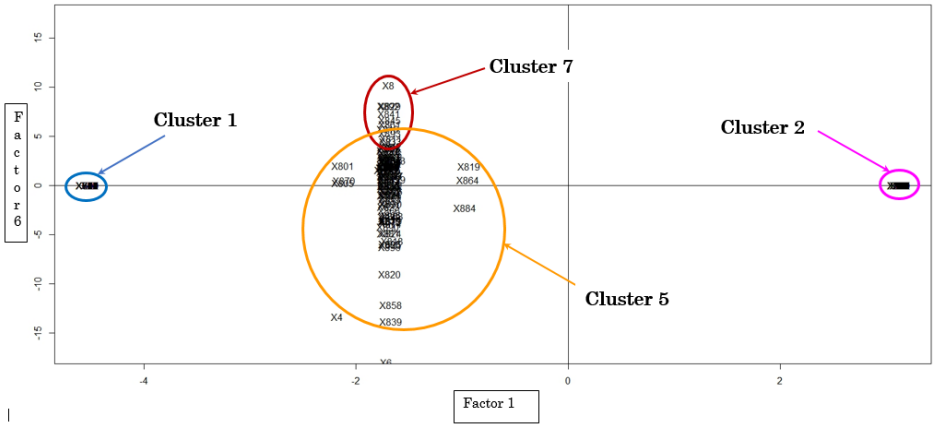






Application on simulated data

Results (2)



$OCI = 0{,}3226$ $ECI = 0{,}3177$

Tchuprow = 0.6576 V-Cramer = 0.7593

Rand = 0.9936 Jaccard = 0.9829

$\gamma = 0.9867$ %GC = 79.4%

Obs / Est	1 $(X_1, X_{501} \ldots X_{800})$	2 $(X_9 \ldots X_{500})$	3 $(X_5, X_{801} \ldots X_{1000})$	4 (X_2)	5 (X_3)	6 (X_4)	7 (X_6)	8 (X_7)	9 (X_8)
1	301	0	0	0	0	0	0	0	0
2	0	492	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	1	1	0	0
5	0	0	185	0	1	0	0	0	0
6	0	0	1	0	0	0	0	0	1
7	0	0	15	0	0	0	0	1	0

Application on simulated data

Performance

→ n =100,000 observations – p = 100 variables – 9 clusters

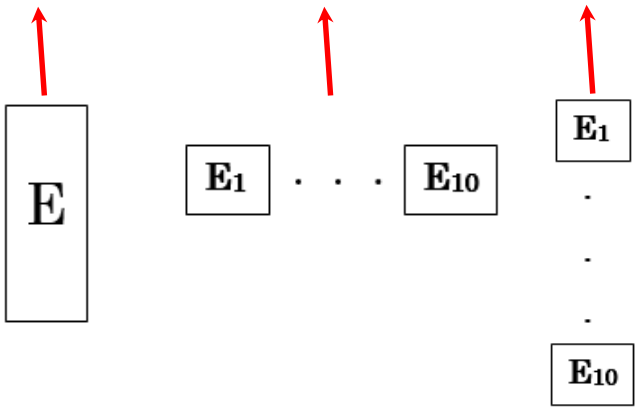
→ 10 random samples without replacement E_s of individuals (n_s = 10000)

→ 10 random samples without replacement Q_l of variables (p_l = 10)

→ 10 initial clustering in parallel + 10 global clustering in parallel + one MCA + one final clustering = 22 tasks

Computer	Entire dataset (ED)	MR sequential (MRS)	MR parallel (MRP)	ED/MRS	R = ED/MRP	Gain = 100(R/n-1)%
HP Z400 (2 cores)	309 ± 30	154 ± 5	103 ± 5	2.00	3.00	50.0
HP i7 (4 cores)	177 ± 8	112 ± 1	32 ± 1	1.25	5.53	176.5

In secondes



Outline

Context and goal

Dynamic approach for clustering variables

Clustering variables in high dimension

Application on simulated data

Contributions, applications, further researches

Contrib., applic. and further researches

(i) The **high dimension** method for clustering variables is based on the **dynamic approach** which already offered good quality of results (2016/17)

(ii) The extension of the clustering dynamic approach for high dimension **preserves the global inertia and the contents of clusters** (simulation study)

(iii) The clustering dynamic approach for high dimension provides **good performance in terms of CPU time**, even in case of sequential process

(iv) The dynamic approach are available in **SAS macros,** Ascending Hierarchical Clustering (AHC) for linear relationships and associated new high dimension approach are available **in R function,** nonlinear AHC and associated new high dimension approach are **in processing in R**

Contrib., applic. and further researches

(v) **Application on real data**

- Energy management (electric consumption, renewal energy, market prices, …)

(vi) Improving and further researches

- To integrate other clustering methods (CLV, ClustOfVar, …) in Divide and Conquer Principle
- To compare with other clustering methods developed for high dimension
- Using of other strategies than Ward's criteria and CPU time studies
- More simulation to validate the proposed approach in different cases (size and structure of clusters, level of correlation between variables, …) More simulations with powerful computers
- More complexe datasets
- Co-clustering in high dimension

Bibliographie (1)

Bühlmann P., Rütimann P., van de Geer S., and Zhang C-H, (2013): Correlated in regression: Clustering and sparse estimation. *Journal of Stat. Planning and Inference*, **143** (11), 1835-1858.

Chavent M., Kuentz V., Liquet B. et Saracco J., (2011): Classification de variables : le package ClustOfVar, *43ièmes Journées de Statistique*, Tunis, Tunisie.

Chen M., (2014): *Classification de variables autour de variables latentes avec filtrage de l'information : application à des données en grande dimension*, Thèse de doctorat, Université de Nantes, Ecole VENAM.

Chen Y. and Yang U., (2016): A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures, www.nature.com=scienticreports.

Dean, J. and Ghemawat, S. (2004): MapReduce: simplified data processing on large clusters. In Proceedings of Sixth Symposium on Operating System Design and Implementation (OSDI 2004).

Derquenne Ch., (2016): Classification de variables : une approche à double critères contrôlés dynamiques, *48ièmes Journées de Statistique,* Montpellier, France.

Bibliographie (2)

Derquenne Ch., (2016): Clustering variables based on a dynamic mixed criteria: application on the energy management, *ENBIS 16*, Sheffield, England.

Derquenne Ch., (2017): Classification de variables avec des relations non linéaires, *49ièmes Journées de Statistique, Avignon*, France.

Derquenne Ch., (2017): Clustering variables with nonlinear relationships based on polynomial transformation and a dynamic mixed criteria, *17th ASMDA,* London, England.

Derquenne Ch., (2018): A dynamic approach for clustering variables in high dimension, *23trd COMPSTAT,* Iasi, Romania.

Saporta G., (1999): Some Simple Rules for interpreting Outputs of Principal Components and Correspondence Analysis, *IXth International Symposium on ASMDA*, Lisbon, Portugal.

Takane Y., Young F.W. and de Leeuw J., (1978): The principal components of mixed measurement level multivariate data: A, alternative least squares method with optimal scaling features, *Psychometrika*.

Vigneau E. and Qannari E.M., (2003): Clustering of Variables Around Latent Components *Communications in Statistics - Simulation and Computation*, **32** (4), 1131-1150.