Classification de variables : Une approche dynamique en grande dimension

Christian Derquenne
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

<table>
<thead>
<tr>
<th>Independence</th>
<th>( \lambda_2 \leq 1 )</th>
<th>( \lambda_2 &gt; 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho = 0 )</td>
<td>Unidim</td>
<td>n.a</td>
</tr>
<tr>
<td>( \rho \neq 0 )</td>
<td>OK</td>
<td>Multidim</td>
</tr>
</tbody>
</table>

Checked by the eigenvalue Saporta test [1999] (\( p \)-value for the 2\textsuperscript{sd} eigenvalue > \( \alpha_U \) given)

Checked by the correlation linear test (\( p \)-value of the highest correlation < \( \alpha_C \) 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)

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

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

Revisited Hierarchical Ascending Clustering (5)
Dynamic approach

5 clusters

5 latent variables

$C_2$  $C_1$  $C_3$  $C_4$  $C_5$  $Z_2$  $Z_1$  $Z_3$  $Z_4$  $Z_5$
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?
Clustering variables in high dimension

Problem, postulat and proposition (2)

→ **Postulat 1**: if a method provides good results on one sample coming from 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 split into $S$ 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 in a distributed file system
→ Developed by Google™ [Dean et al., 2004]
Clustering variables in high dimension

Application to dynamic approach clustering variables

Let’s be $X_1, \ldots, X_j, \ldots, X_p$, $p \gg 100$ numeric variables, such as $X_j \in \mathbb{R}^n$ where $n \gg 10000$ 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

<table>
<thead>
<tr>
<th>Variables</th>
<th>$Q_1$</th>
<th>$\ldots$</th>
<th>$Q_l$</th>
<th>$\ldots$</th>
<th>$Q_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>$T_{11}$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$E_s$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$T_{sl}$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$E_s$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$T_{sl}$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>$\ldots$</td>
<td>$p_l$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$p_L$</td>
</tr>
</tbody>
</table>
Clustering variables in high dimension

General process

Initial Clustering

Global Clustering

MCA

Final clustering

$E$

$E_1$

$E_s$

$S \times L$ initial clustering (in parallel)

$S$ global clustering (in parallel)

One MCA

One final clustering

$= S \times (L + 1) + 2$ tasks

Selected Components of MCA

$M$ final clusters
Clustering variables in high dimension

Application to dynamic approach clustering variables

Detailed process (1):

(i) Apply a initial clustering variables method on $T_{sl}$
   $\rightarrow$ we obtain $M_{sl}$ clusters and $M_{sl}$ associated first principal components: $Y_{1}^{(sl)}, \ldots, Y_{M_{sl}}^{(sl)}$

The process (i) is made in parallel on each $T_{sl}$ of $E_s$ [MAP Phase]
   $\rightarrow$ then we obtain: $Y_{1}^{(s1)}, \ldots, Y_{M_{s1}}^{(s1)}, \ldots, Y_{1}^{(sl)}, \ldots, Y_{M_{sl}}^{(sl)}, \ldots, Y_{M_{sL}}^{(sL)}, \ldots, Y_{M_{sL}}^{(sL)}$ first principal components

(ii) Apply again a global clustering variables on the previous first principal components
   $\rightarrow$ we obtain $M_s$ new clusters $(C_{1}^{(s)}, \ldots, C_{k}^{(s)}, \ldots, C_{M_s}^{(s)})$ and $M_s$ associated new first principal
   components: $Z_{1}^{(s)}, \ldots, Z_{k}^{(s)}, \ldots, Z_{M_s}^{(s)}$
   $\rightarrow$ then each initial variable $X_j \in C_{k}^{(s)}$
   $\rightarrow$ let’s be $V_s \in \{1, \ldots k, \ldots, M_s\}$, a new variable containing the cluster numbers of
   $(C_{1}^{(s)}, \ldots, C_{k}^{(s)}, \ldots, C_{M_s}^{(s)})$ 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 = $U_1, \ldots, U_r, \ldots, U_R$

<table>
<thead>
<tr>
<th>Variables</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$v_{11}$</td>
<td>$v_{12}$</td>
<td>$v_{13}$</td>
<td>$v_{s1}$</td>
</tr>
<tr>
<td>$X_j$</td>
<td>$v_{j1}$</td>
<td>$v_{j2}$</td>
<td>$v_{j3}$</td>
<td>$v_{sj}$</td>
</tr>
<tr>
<td>$X_p$</td>
<td>$v_{p1}$</td>
<td>$v_{p2}$</td>
<td>$v_{p3}$</td>
<td>$v_{sp}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principal components</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
</tr>
<tr>
<td>$U_r$</td>
</tr>
</tbody>
</table>

$MCA$ applied to the variables matrix and individuals matrix.
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

$\rightarrow M$ final clusters $G_1, ..., G_m, ..., G_M$ containing $p_1, ..., p_m, ..., p_M$ initial

variables $X_1, ..., X_j, ..., X_p$ with $\sum_{m=1}^{M} p_m = p$ on $n$ individuals

**Final clustering (G)**

<table>
<thead>
<tr>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{1,1}$</td>
</tr>
<tr>
<td>$X_{1,2}$</td>
</tr>
<tr>
<td>$X_{1,3}$</td>
</tr>
<tr>
<td>$X_{1,F}$</td>
</tr>
<tr>
<td>$X_{2,1}$</td>
</tr>
<tr>
<td>$X_{2,2}$</td>
</tr>
<tr>
<td>$X_{2,3}$</td>
</tr>
<tr>
<td>$X_{2,F}$</td>
</tr>
<tr>
<td>$X_{M,1}$</td>
</tr>
<tr>
<td>$X_{M,2}$</td>
</tr>
<tr>
<td>$X_{M,3}$</td>
</tr>
<tr>
<td>$X_{M,F}$</td>
</tr>
</tbody>
</table>

XXVIèmes Rencontres de la SFC – 5 septembre 2019
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

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

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

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 - \( \tilde{M} \) 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, \( \gamma \), V-Cramer, Tschuprow, %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

\[ n = 100,000 \text{ observations} - p = 1000 \text{ variables} - 9 \text{ clusters} \]

<table>
<thead>
<tr>
<th>Clusters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td>( X_1, X_{301}, ..., X_{800} )</td>
<td>( X_9, X_{101}, ..., X_{500} )</td>
<td>( X_{5}, X_{101}, ..., X_{1000} )</td>
<td>( X_2 )</td>
<td>( X_3 )</td>
<td>( X_4 )</td>
<td>( X_5 )</td>
<td>( X_6 )</td>
<td>( X_7 )</td>
</tr>
<tr>
<td>Simulation</td>
<td>( X_j \sim N(0,1) )</td>
<td>( \epsilon_j \sim N(0,1) )</td>
<td>( X_j = 2X_1 + \epsilon_j )</td>
<td>( j = 501, ..., 800 )</td>
<td>( X_j \sim N(0,1) )</td>
<td>( \epsilon_j \sim N(0,1) )</td>
<td>( X_j = 0.1X_3 + \epsilon_j )</td>
<td>( j = 801, ..., 1000 )</td>
<td>( X_j \sim N(0,1) )</td>
</tr>
</tbody>
</table>

Building of sub-datasets and process

\[ \rightarrow 10 \text{ random samples without replacement } E_s \text{ of individuals} (n_s = 10000) \]

\[ \rightarrow 10 \text{ random samples without replacement } Q_l \text{ of variables} (p_j = 100) \]

\[ \rightarrow 100 \text{ initial clustering in parallel} + 10 \text{ global clustering in parallel} + \text{ one MCA} + \text{ one final clustering} = 112 \text{ tasks} \]
Application on simulated data

Results (1)
Application on simulated data

Results (2)

\[ OCI = 0.3226 \quad ECI = 0.3177 \]

\[ Tchuprow = 0.6576 \quad V-Cramer = 0.7593 \]

\[ \text{Rand} = 0.9936 \quad \text{Jaccard} = 0.9829 \]

\[ \gamma = 0.9867 \quad \%GC = 79.4\% \]

<table>
<thead>
<tr>
<th>Obs</th>
<th>Est</th>
<th>1 (( X_1, X_{501} \ldots X_{500} ))</th>
<th>2 (( X_2 \ldots X_{500} ))</th>
<th>3 (( X_5, X_{501} \ldots X_{1000} ))</th>
<th>4 (( X_2 ))</th>
<th>5 (( X_3 ))</th>
<th>6 (( X_4 ))</th>
<th>7 (( X_5 ))</th>
<th>8 (( X_6 ))</th>
<th>9 (( X_8 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>301</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>492</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>185</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
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

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

In seconds
Outline

Context and goal

Dynamic approach for clustering variables

Clustering variables in high dimension

Application on simulated data

Contributions, applications, 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)


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.


Bibliographie (2)


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


