ClustVarLV: A package for the clustering of variables around latent variables

Evelyne Vigneau, Mingkun Chen, El Mostafa Qannari

sensométrie chimiométrie

Oniris
École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes Atlantique

Outline

- Context : the clustering of variables
- CLV method: data structure / types of groups
- Algorithms et main functions in « ClustVarLV »
- *Illustration 1:* psychological scales
- *Illustration* 2: preference mapping
- ClustVarLV et ClustOfVar
- Conclusion and perspectives

Le package **ClustVarLV**

Clustering of variables around Latent Variables

Documentation for package 'ClustVarLV' version 1.2

- DESCRIPTION file.

Help Pages

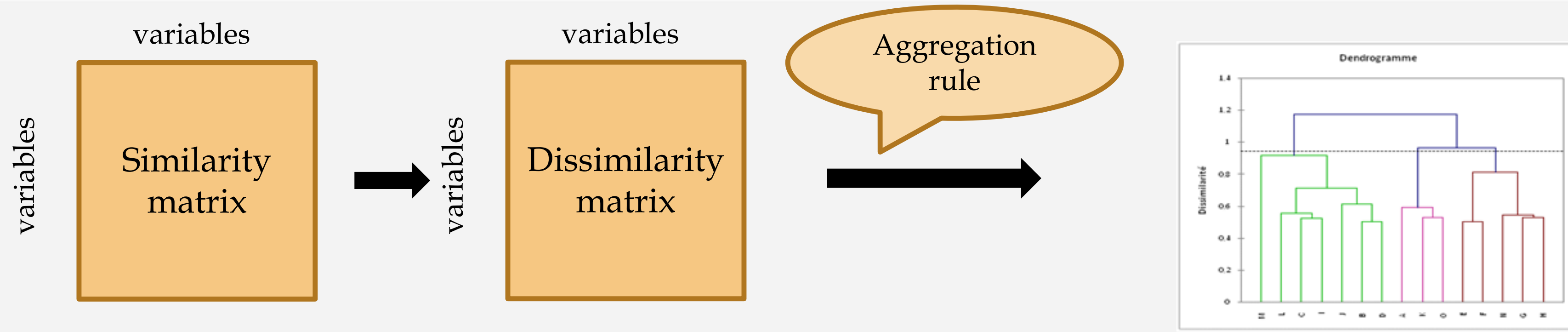
Main functions

apples_sh	apples from southern hemisphere data set
authen_NMR	Authentication data set/ NMR spectra
CLV	Hierarchical clustering of variables with consolidation
CLV_kmeans	K-means algorithm for the clustering of variables
descrip_gp	Description of the clusters of variables
gpmb_on_pc	Representation of the variables and their group membership
LCLV	L-CLV for L-shaped data
print.clv	Print the CLV results
print.clvkmeans	Print the CLV_kmeans results
print.lclv	Print the LCLV results

datasets

Useful functions

The clustering of variables



The clustering of variables

variables

variables

Similarity matrix

→

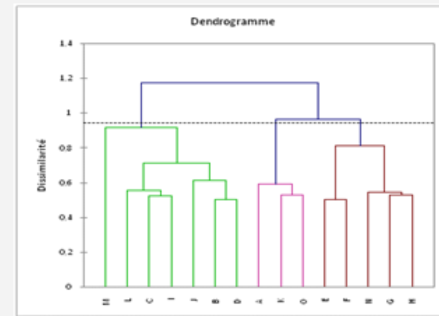
variables

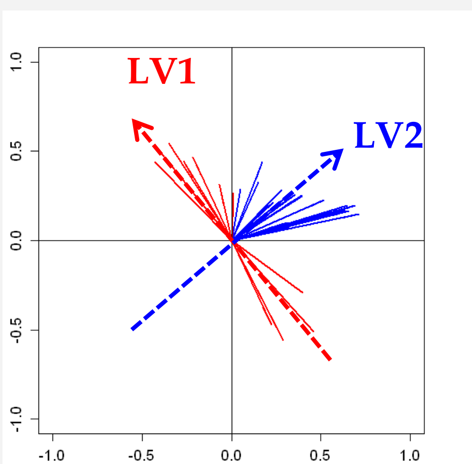
variables

Dissimilarity matrix

Aggregation rule

→





Factor analysis / exploratory approaches :
Identifying groups of variables
defined around Latent Variables (LV)

CLV (Clustering of variables around Latent Variables)
available on R

VARCLUS : procédure SAS/STAT

Highlighting the inter-correlations structure between the variables

- **Principal Components Analysis (PCA)**

  ⇨ analysis of the linear relationships between the variables and dimensionality reduction using the first Principal Components (PC).

- **Principal Components with rotation** (RC)

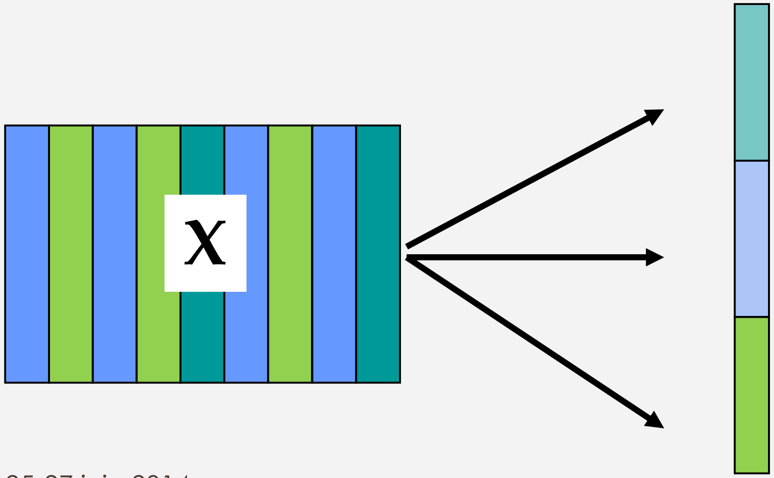
  ⇨ Linear combinations of the initial variables more easy to interpret than the PC.

- **CLV approach**

  ⇨ dimensionality reduction (K latent variables (LV) associated with groups of variables).

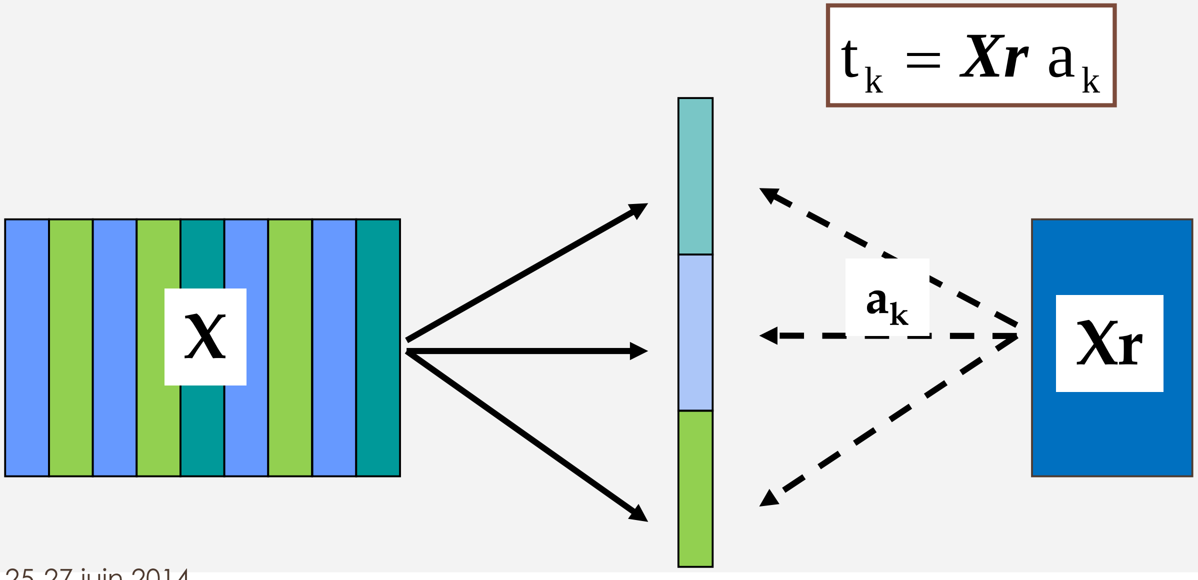
  ⇨ easier interpretation (each LV is a linear combinayion of the variables belonging to the associated group).

CLV method for various data structures

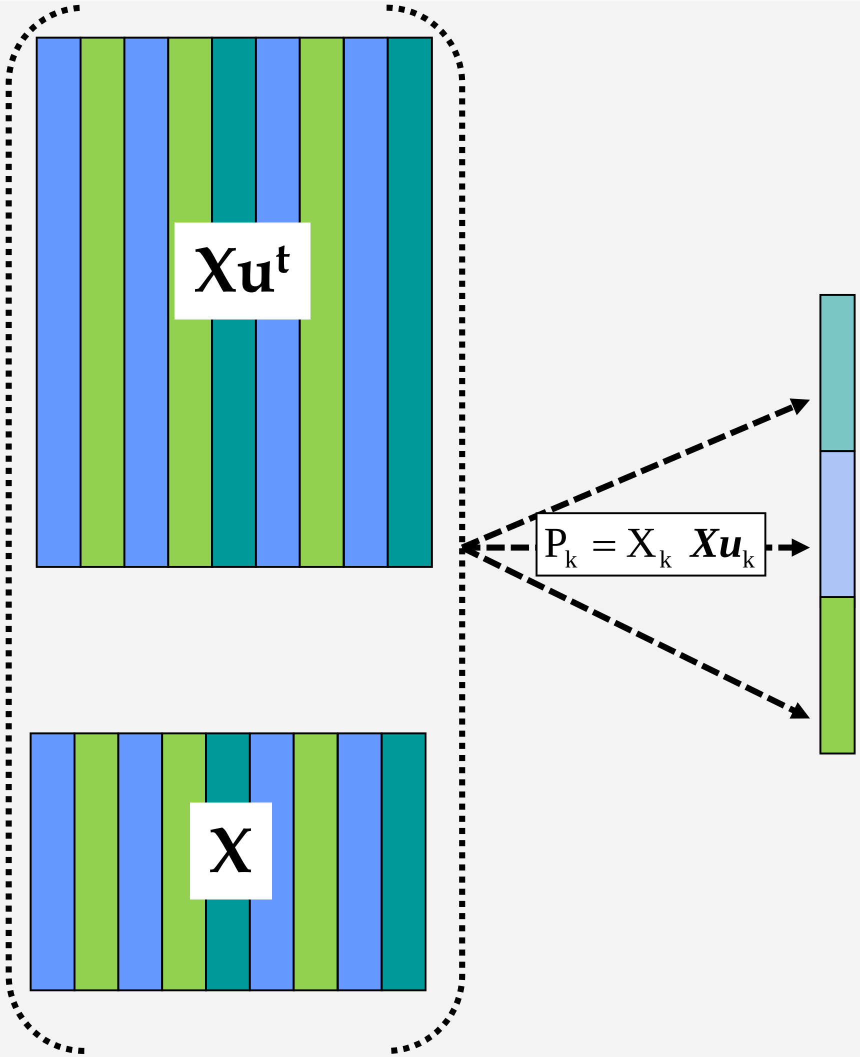


$\boldsymbol{c_k}$: latent variable of group G_k

CLV method for various data structures

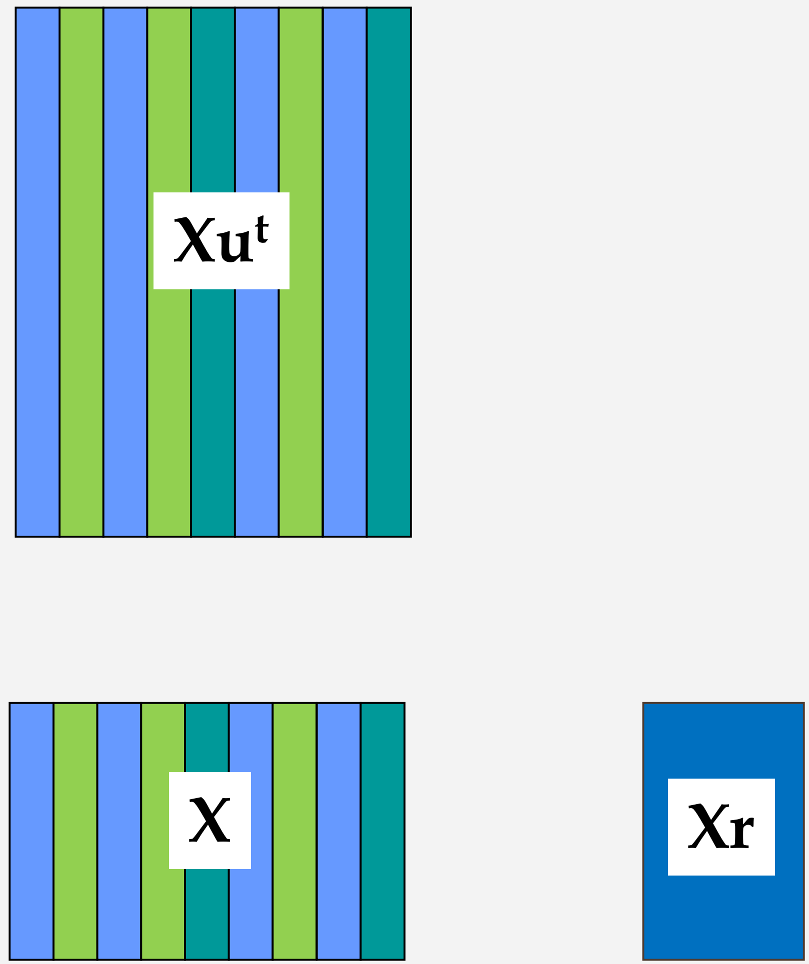


CLV method for various data structures



$\tau_k = P_k\ u_k$

CLV method for various data structures (L-shaped data)



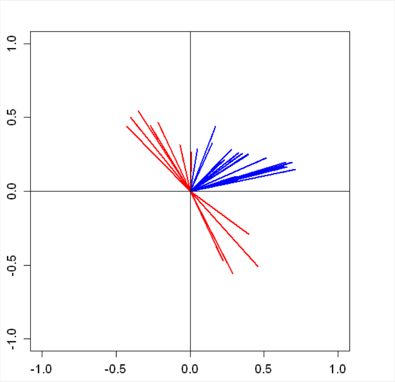
$$\tau_k = P_k \, u_k$$

$$t_k = Xr \, a_k$$

CLV method: two types of groups

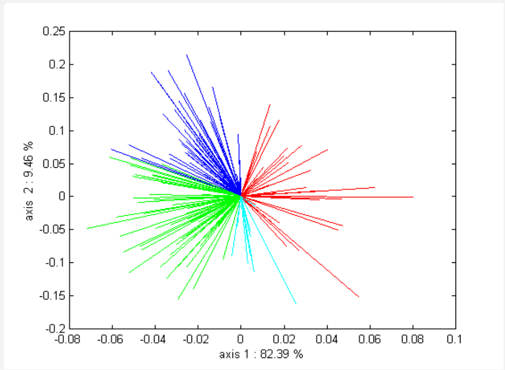
Directional groups

High positive or negative correlations ⇨ agreement



Local groups

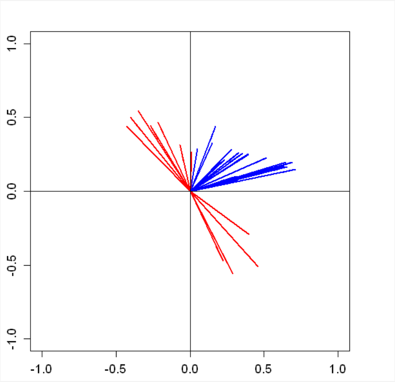
High positive correlations ⇨ agreement
High negative correlations ⇨ disagrrement



CLV method: two types of groups

Directional groups

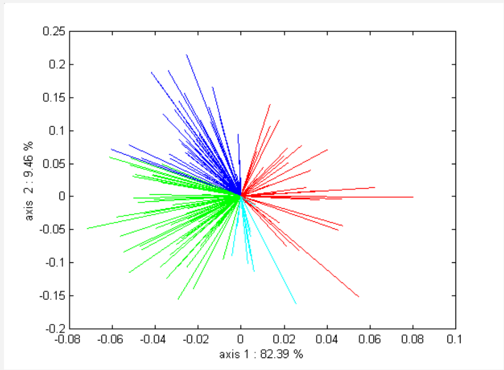
High positive or negative correlations ⇨ agreement



method = 1

Local groups

High positive correlations ⇨ agreement
High negative correlations ⇨ disagrrement



method = 2

```
> CLV(X, method= 2)
```

CLV method: two types of groups

Directionnal groups
`method = 1`

Local groups
`method = 2`

Maximization of

nb of groups

$$T = n\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\operatorname{cov}^2(\boldsymbol{x}_j,\boldsymbol{c}_k)$$

Latent Variable

Group's membership indicator

$$S = \sqrt{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\operatorname{cov}(\boldsymbol{x}_j,\boldsymbol{c}_k)$$

avec $\boldsymbol{c}_k{}'\boldsymbol{c}_k = 1$

Algorithm (1)

Partitioning algorithm

➓ **Initialization :** user's choice (…) or
at random (`nstart`)

➊ **Estimation of the LV**

`method=1`, matrix **X** : $\mathbf{c_k}$ (k=1,…K) is the first standardized principal component of X_k

`method=2`, matrix **X** : $\mathbf{c_k}$ (k=1,…K) is proportional to the averaged variable $\overline{x}_k$

➋ **Assignment step**

cas `method=1`, matrix **X** : $\delta_{kj} = 1 \text{ if } \max_{l=1,\ldots,K} \left\{ \mathrm{cov}^2(x_j, c_l) \right\} = \mathrm{cov}^2(x_j, c_k)$

cas `method=2`, matrix **X** : $\delta_{kj} = 1 \text{ if } \max_{l=1,\ldots,K} \left\{ \mathrm{cov}(x_j, c_l) \right\} = \mathrm{cov}(x_j, c_k)$

until convergence

Function (1)

Partitioning algorithm

```
>  CLV_kmeans(X, method=1 , sX=TRUE, init= K, nstart=100)
```

data matrix (n x p)

type of groups

sX=TRUE / FALSE
standardization of the variables, or not

- if `init` is a scalar, say K : nb de groups in the partition
- if `init` is a vector of p integers $\in\{1,\dots,K\}$: initial partition

nb of repetitions of the algorithm. (`nstart`=1 if initialization by a partition given by the user)

<u>Outputs</u> :

⇨ partition into K groups (if `nstart`>1, optimal partition among the `nstart` solutions is given)

⇨ Latent variables for each group of variables (not standardized)

+ value of the criterion at convergence, nb of iterations before convergence, summary for the `nstart` solutions

Algorithm (2)

Ascendant hierarchical algorithm

- At the beginning (step 1) : each variable is a group by itself (K=p)
- At the end (step p) : all the variables are in the same group (K=1)

illustration for *method=1*

- At step j value of the criterion T_j partition : {A, B, ….}

$\Downarrow$ $\Downarrow$

- At step j+1 value of the criterion $T_{j+1} < T_j$ partition : {$A \cup B$, …}

aggregation criterion: $\Delta T_j = (T_j - T_{j+1}) \quad > 0$

Rule : at each step, j, the two groups, A et B, for which ΔT_j is minimized are merged together (loss of within-group coherence as small as possible)

Advantages :

- Initialization of the partitioning algorithm
- Help for choosing the number of groups, K, on the basis of the variations of ΔT_j

Function (2)

Ascendant hierarchical algorithm with consolidation by the *k-means* algorithm

```
>  CLV(X, method=1 , sX=TRUE, nmax= 20, graph=TRUE)
```

Maximal size of the partition for which a *k-means* consolidation is performed (20, by default).

TRUE by default
⇨ dendrogram
⇨ graph showing the evolution of the aggregation criterion

Outputs :

⇨ partitions into 1, 2, 3, …, `nmax` groups before consolidation (by cutting the dendrogram) **and** after consolidation (*k-means*).

⇨ Latent variables for each group associated to each partition.

⇨ detailed results of the hierarchy.

Functions (3)

The same functions are used
with or without external variables

Example (available with the package) :

```
> data(apples_sh)
# local groups with external variables Xr
> resclvYX <- CLV_kmeans (X = apples_sh$pref,
                   Xr = apples_sh$senso, method = 2,
                   sX = FALSE, sXr = TRUE, graph = TRUE)
```

Illustration 1 : exploratory analysis for psychological scales

- **AUPALESENS project** (France, 2010-2014)
  ‘‘Making eating more enjoyable for seniors to promote healthy aging and prevent malnutrition”

- n=559 subjects (>65 ans)

- Pluridisciplinary questionnaire … only considered here
  scales used for assessing psychological behaviour (5-points Likert scale)

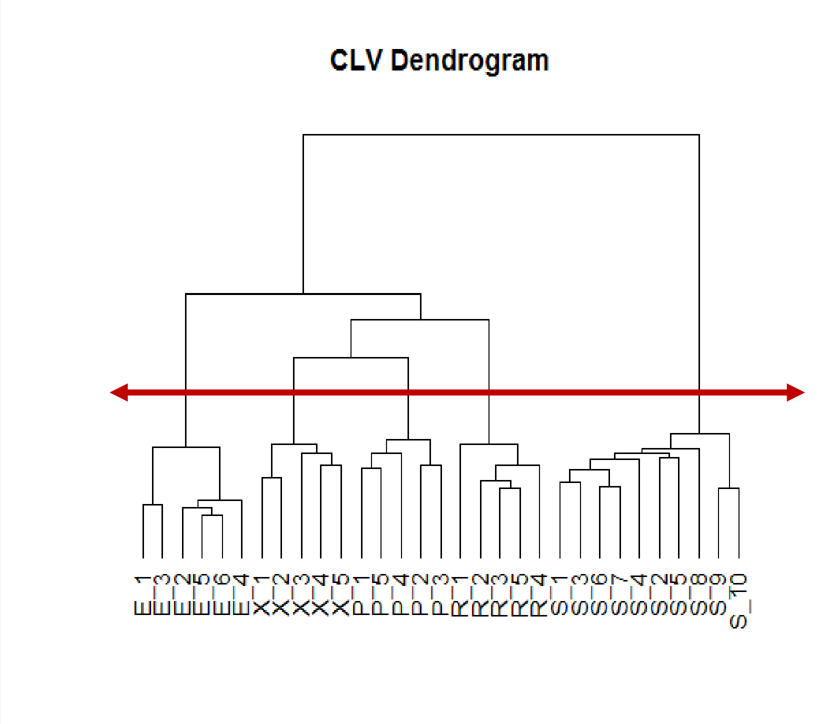
**Bailly, Maitre, Amand, Hervé, Alaphilippe (2012). Appetite, 59(853-858)*

Eating behaviour (based on DEBQ)

- « Emotional eating » (E) : 6 items
- « eXternal eating » (X) : 5
- « Restrained eating» (R) : 5 items
- « Food enjoyment» (P) : 5 items
- « Self estime» (S) : 10 items

Illustration 1

```
> load("AUPA_psycho.rda")
> X<-AUPA_psycho
> dim(X)
[1] 559  31
> res.clv<-CLV(X,method=1,sX=TRUE,graph=TRUE)
```



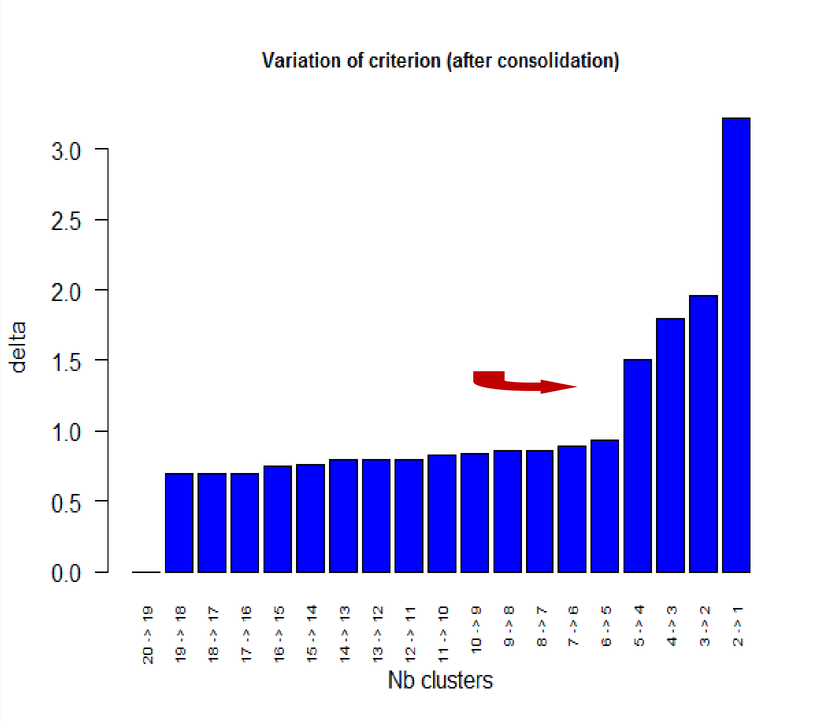


Illustration 1

```
> descrip_gp(res.clv,X,K=5)

$number  1   2   3   4   5
         6   5   5   5  10

$prop_within
 Group.1 Group.2 Group.3 Group.4 Group.5
 0.6036  0.4077  0.4653   0.388  0.3614

$prop_tot    0.4368

$cormatrix
      Comp1 Comp2 Comp3 Comp4 Comp5
Comp1  1.00  0.36  0.27  0.08  0.20
Comp2  0.36  1.00  0.23  0.23  0.11
Comp3  0.27  0.23  1.00  0.14  0.05
Comp4  0.08  0.23  0.14  1.00 -0.16
Comp5  0.20  0.11  0.05 -0.16  1.00
```

Within-group variability explained by the Latent Variable of the group

Total variability explained by the 5 Latent Variables

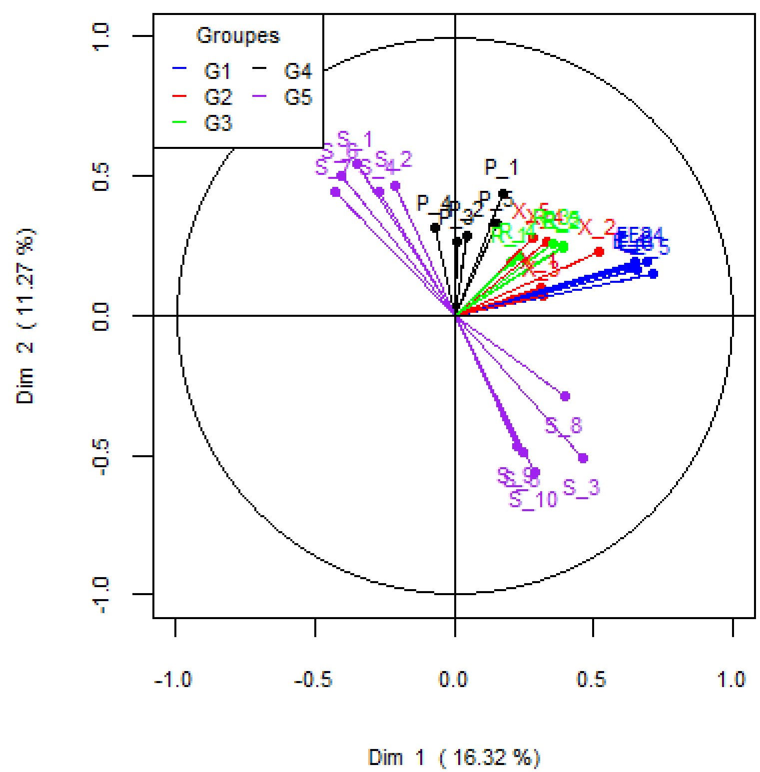
Correlation matrix betwwen the Latent Variables

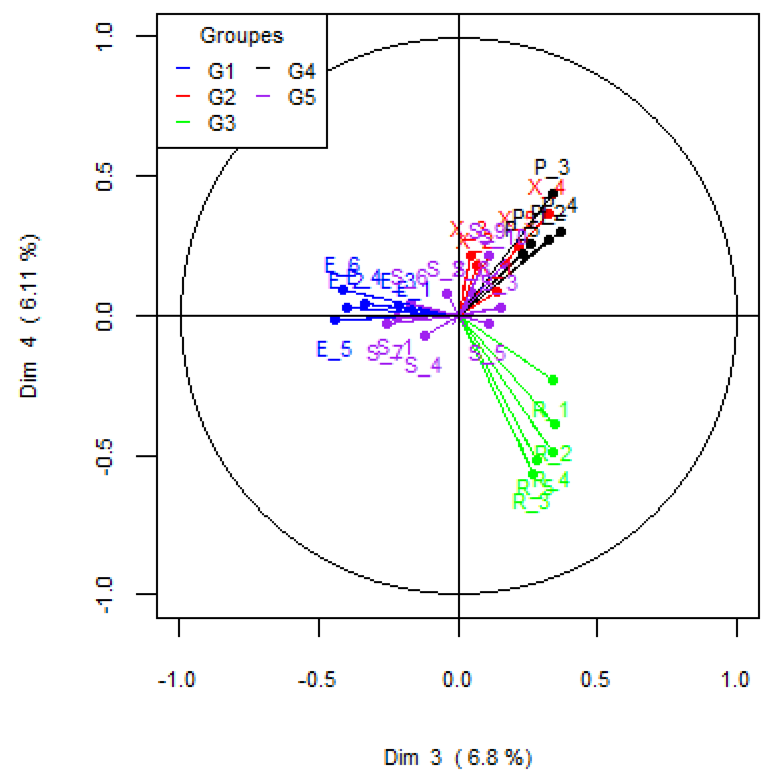
Illustration 1

```
> descrip_gp(res.clv,X,K=5)

$groups[[1]]
    cor in group cor next group
E_5         0.85           0.25
E_4         0.80           0.34
E_6         0.80           0.25
E_2         0.79           0.25
E_3         0.73           0.31
E_1         0.68           0.29

$groups[[2]]
    cor in group cor next group
X_2         0.76           0.38
X_4         0.67           0.30
X_5         0.65           0.19
X_1         0.58           0.17
X_3         0.51           0.22

$groups[[3]]
    cor in group cor next group
R_5         0.77           0.25
R_3         0.76           0.21
R_2         0.71           0.23
R_4         0.66           0.11
R_1         0.47           0.14

$groups[[4]]
    cor in group cor next group
P_1         0.72           0.18
P_3         0.63           0.14
P_2         0.61           0.10
P_4         0.58          -0.14
P_5         0.57           0.19

$groups[[5]]
     cor in group cor next group
S_3          0.70           0.21
S_1         -0.68          -0.10
S_6         -0.66           0.17
S_7         -0.65          -0.17
S_10         0.65           0.07
S_5          0.55          -0.12
S_4         -0.53           0.10
S_9          0.53          -0.10
S_2         -0.51           0.14
S_8          0.49           0.23
```

Illustration 1: exploratory analysis of the scales

```
> gpmb_on_pc(res.clv,X,K=5,axeh=1,axev=2,label=TRUE)
> gpmb_on_pc(res.clv,X,K=5,axeh=3,axev=4,label=TRUE)
```





The groups of variables perfectly coincide with the underlying psychological scales

Illustration 2: preference mapping of apple using L-CLV

Xu^t

Consumers questionnaire

- Frequency of consumption,
- Apple cultivars known
- Important sensory attributes,
- Modalities of consumption (peeled/during meal/ …)
- Purchase criteria
- ….
- Age, gender, professional activity….

Vigneau, Charles, Chen (2013). *Food Quality and Preference*, 22(4), 83-92

X

produits

hedonic test

224 regular apple consumers
31 apples varieties

Liking scores on a 9-points scale

Xr

Sensory descriptive analysis

15 assessors, 15 attributes

Crunchy
Juicy
Fondant

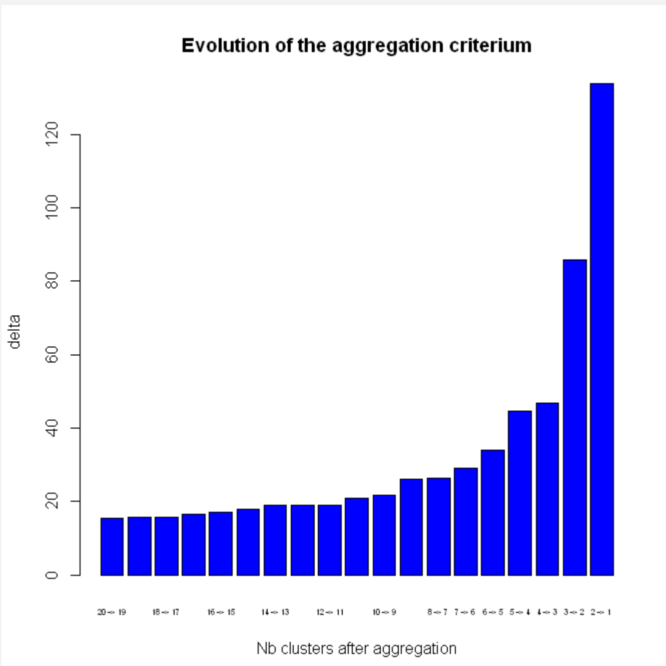
Sweet
Acid

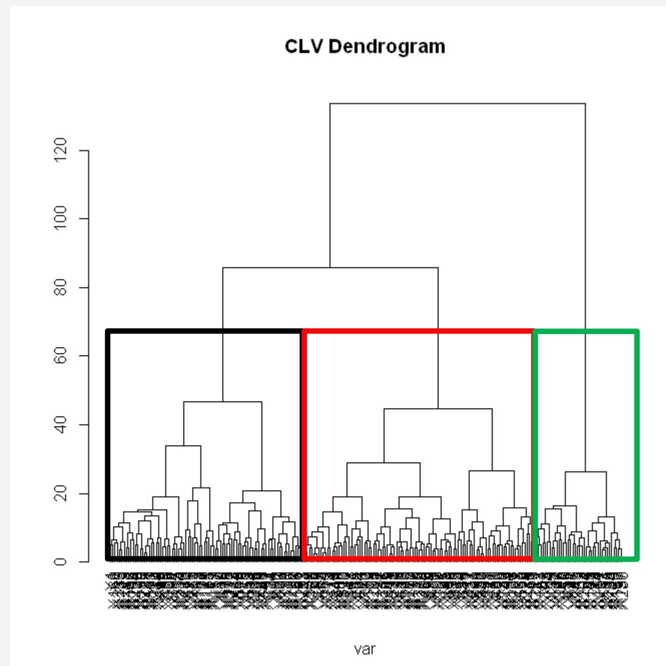
Odour intensity
Aroma intensity

A_Pineapple/Banana
A_Sweet/Rose
A_Woody/Earthy
A_Rustic
A_Lemon
A_White flowers
A_Ripe fruit
A_Green

Illustration 2

```
> resL<-LCLV(X=pref, Xr=senso, Xu=questions,
                        _ sX=TRUE,sXr=TRUE,sXu=FALSE,graph=TRUE)
```

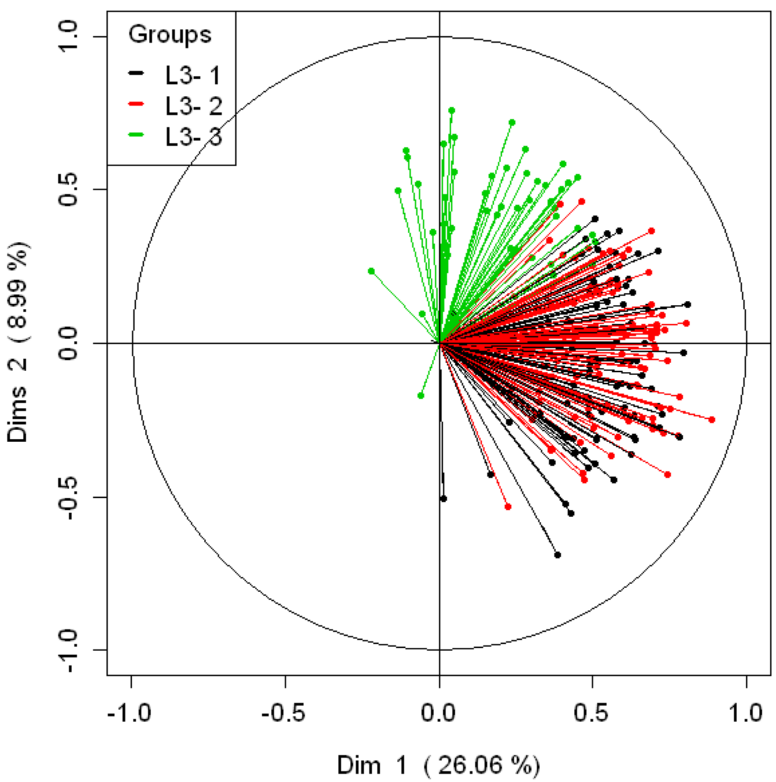




Segment L3-1	82 consumers	(37%)
Segment L3-2	96 consumers	(43%)
Segment L3-3	46 consumers	(20%)

Illustration 2

```
> gpmb_on_pc(resL,X=pref,K=3,axeh=1,axev=2,label=FALSE)
> gpmb_on_pc(resL,X=pref,K=3,axeh=2,axev=3,label=FALSE)
```



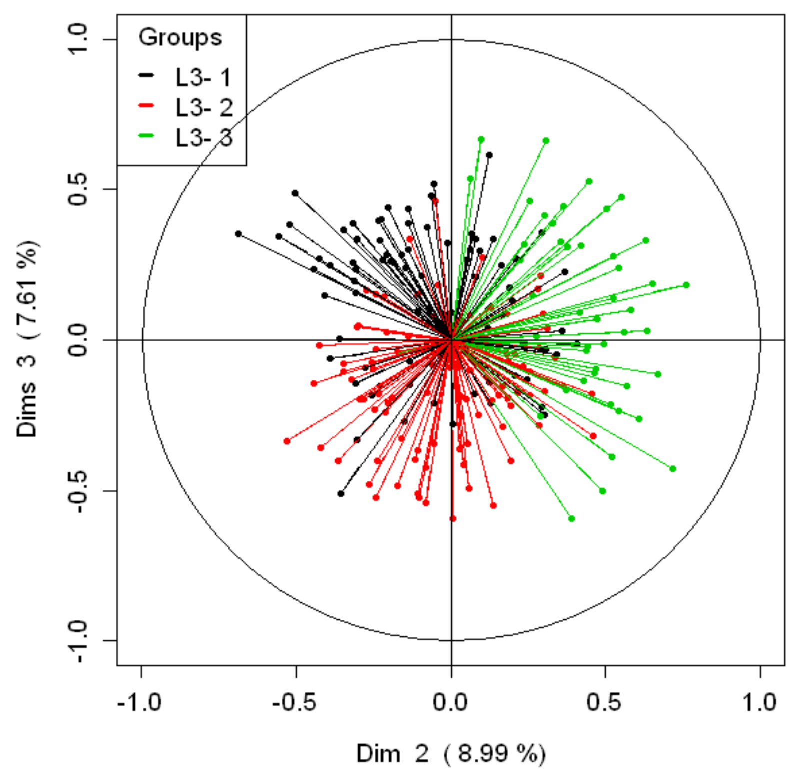


Illustration 2

Interpretation of the segmentation of consumers panel

❖ According to the sensory *drivers*

loadings (a_k) associated with the variables in Xr

- Consumers in the segments **1** and **2** appreciated the juicy and sweet varieties of apple, with « ananas/banana » aroma.
- Consumers in the segment **3** appreciated more fondant apples, with « rustic » and « ripe fruit » aroma. They dislike acifity and « green » aroma in apples.

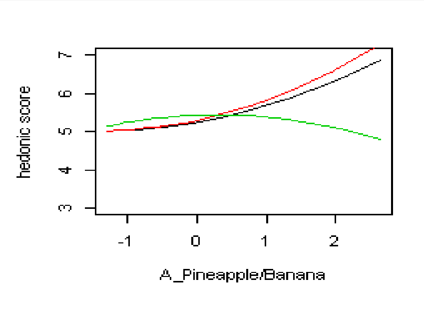


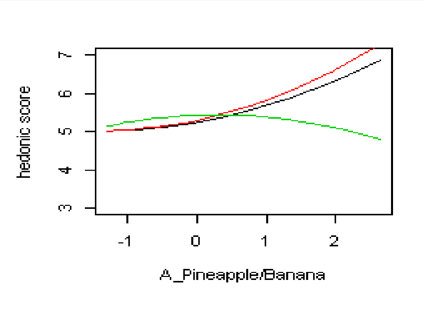
Illustration 2

Interpretation of the segmentation of consumers panel

❖ According to the sensory *drivers*

loadings (a_k) associated with the variables in Xr

- Consumers in the segments **1** and **2** appreciated the juicy and sweet varieties of apple, with « ananas/banana » aroma.
- Consumers in the segment **3** appreciated more fondant apples, with « rustic » and « ripe fruit » aroma. They dislike acifity and « green » aroma in apples.



❖ According to the Usage & Attitude items and the socio-demographic characteristics of the consumers

loadings (u_k) associated with the variables in Xu

- Segment **1** : mainly, the youngest in the panel
- Segment **2** et **3** : in majority, > 40 years old

are attentive to appearance, color, packaging cultivar, origin.

….

ClustVarLV et ClustOfVar

Both based on the CLV approach
Similar algorithms (hierarchical and k-means)

Type of groups	
directional or local	directional, only
Standardization	
choice	quantitative variables are standardized
Categorical variables	
data coding with dummy variables, clustering of the modalities	*integrated* clustering criterion updated
Variables externes	
integrated, associated with the obs. and/or the variables	-

Conclusion et perspectives

ClustVarLV : clustering of variables

… but not only that:

- data dimensionality reduction (latent variables)
- CLV components easier to understand

Conclusion et perspectives

ClustVarLV : clustering of variables

… but not only that:

- data dimensionality reduction (latent variables)
- CLV components easier to understand

Many different areas of application: sensory analysis and consumer's preference analysis, chemometry (IR, RMN spectroscopy), omic- data, psychometry, satisfaction questionnaires …

Conclusion et perspectives

ClustVarLV : clustering of variables

… but not only that:

- data dimensionality reduction (latent variables)
- CLV components easier to understand

Many different areas of application: sensory analysis and consumer's preference analysis, chemometry (IR, RMN spectroscopy), omic- data, psychometry, satisfaction questionnaires …

Developpments in progress

- « discarding » the atypical variables / the variables which are not well associated with the group's structure in the dataset.
- Supervised clustering of variables (by taking into account of a response variable)