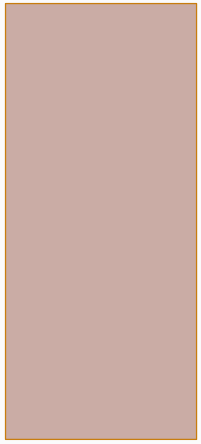
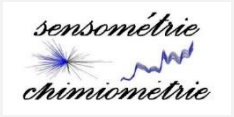
# ClustVarLV:
# Un package pour la classification de variables autour de variables latentes

Evelyne Vigneau, Mingkun Chen, El Mostafa Qannari






# Plan

- Contexte : la classification de variables
- La méthode CLV : structure de données / types de groupes
- Algorithmes et fonctions principales du package ClustVarLV
- *Exemple* : analyse exploratoire d'échelles de mesure
- *Exemple* : cartographie des préférences
- ClustVarLV et ClustOfVar
- Conclusion et perspectives

# Le package **ClustVarLV**

**Clustering of variables around Latent Variables**

Documentation for package 'ClustVarLV' version 1.2

- DESCRIPTION file.

## Help Pages

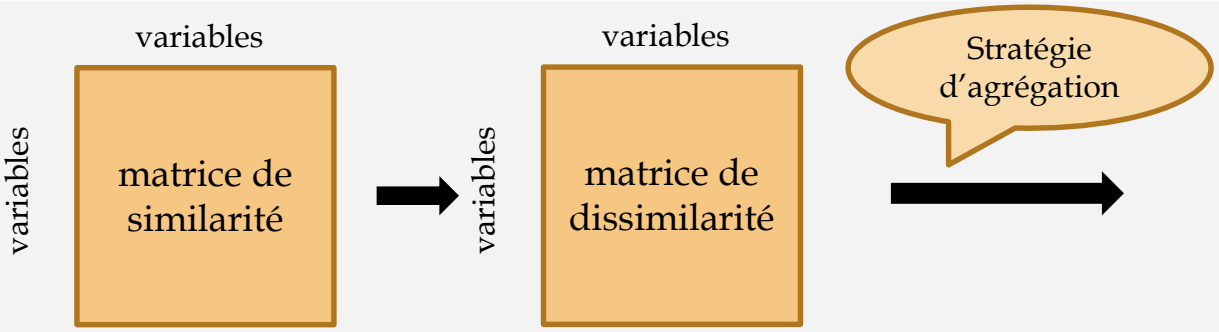
**fonctions principales**

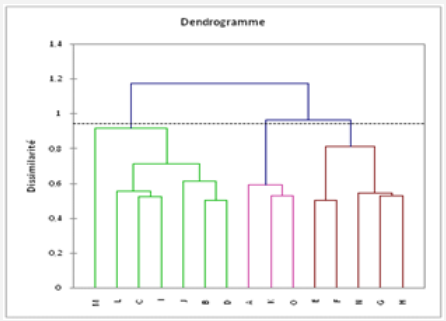
| | |
|---|---|
| apples_sh | apples from southern hemisphere data set |
| authen_NMR | Authentication data set/ NMR spectra |
| CLV | Hierarchical clustering of variables with consolidation |
| CLV_kmeans | K-means algorithm for the clustering of variables |
| descrip_gp | Description of the clusters 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 |

**fichiers d'exemples**

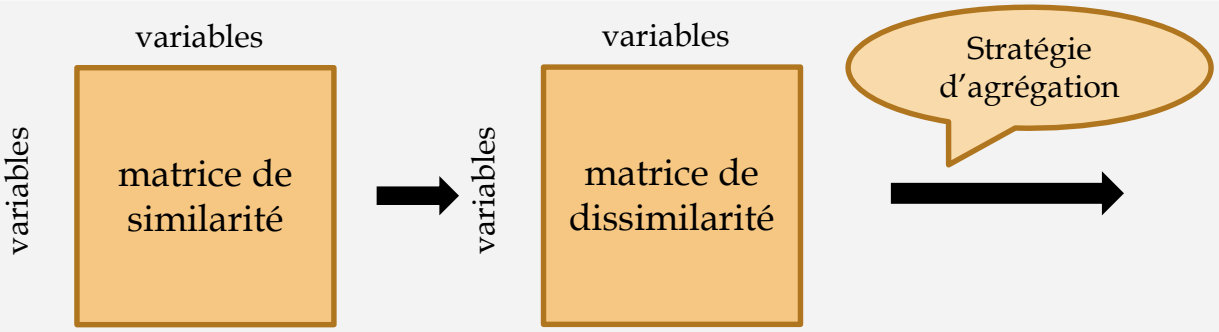
**aide à l'analyse des résultats**

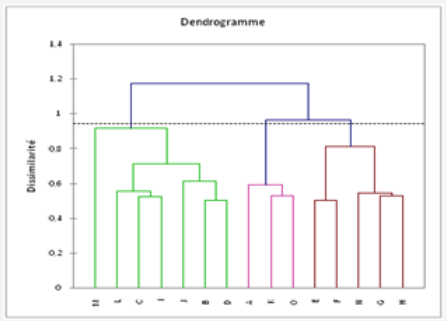
# La classification de variables

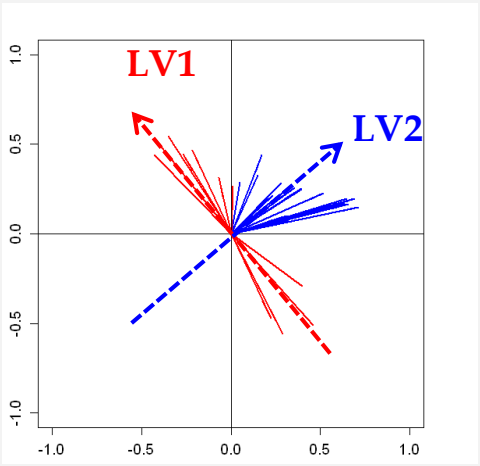


# La classification de variables




approche factorielle :
faire des groupes de variables
organisés autour de variables latentes (LV)

CLV (Clustering of variables around Latent Variables)
programmée sous matlab, sous R

VARCLUS : procédure SAS/STAT

# Mise en évidence de la structure des variables

- **Analyse en Composantes Principales (ACP)**

  ⇨ exploration des relations entre variables et réduction de la dimension des données sur la base des premières composantes principales (PC).

- **Composantes Principales avec rotation** (RC)

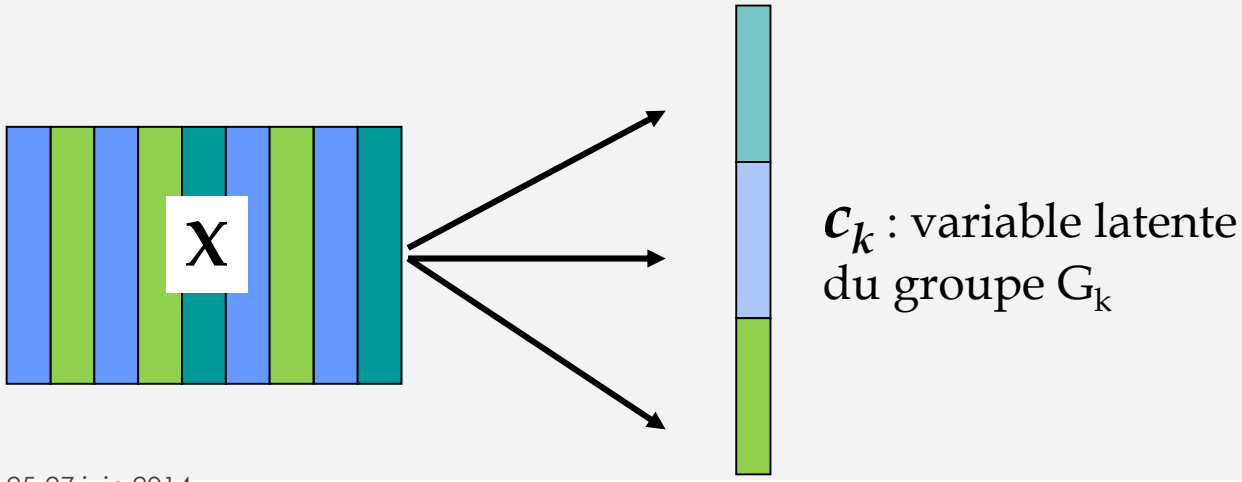
  ⇨ combinaisons linéaires des variables de départ plus facilement interprétables que les PC.

- **Approche CLV**

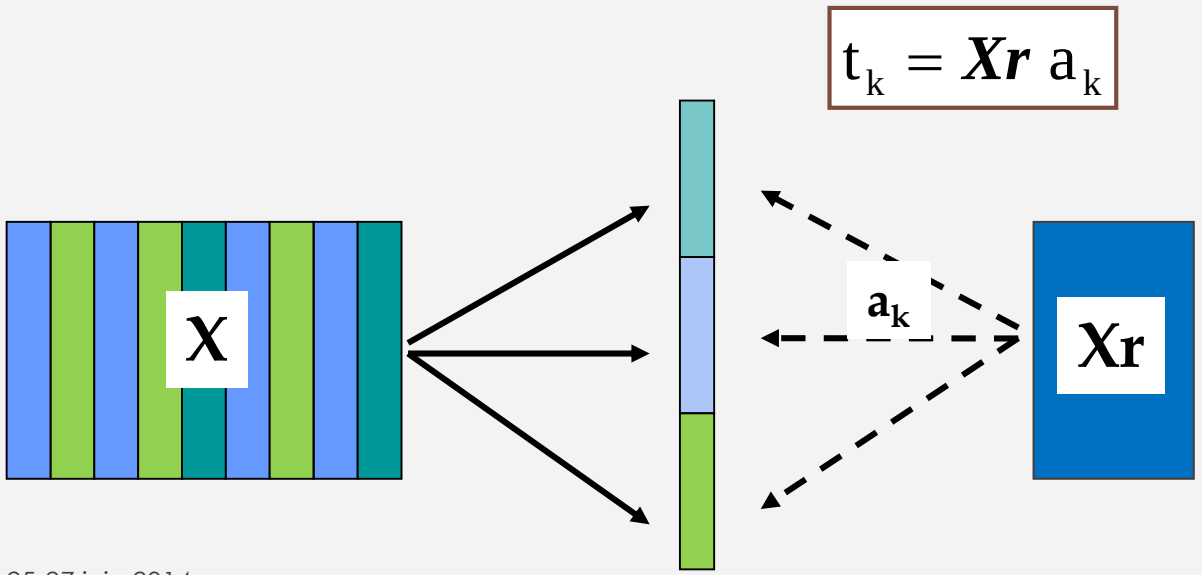
  ⇨ réduction de la dimension (K variables latentes (LV) de groupes)

  ⇨ interprétation simplifiée (chaque LV est combinaison linéaire des variables du groupe correspondant)

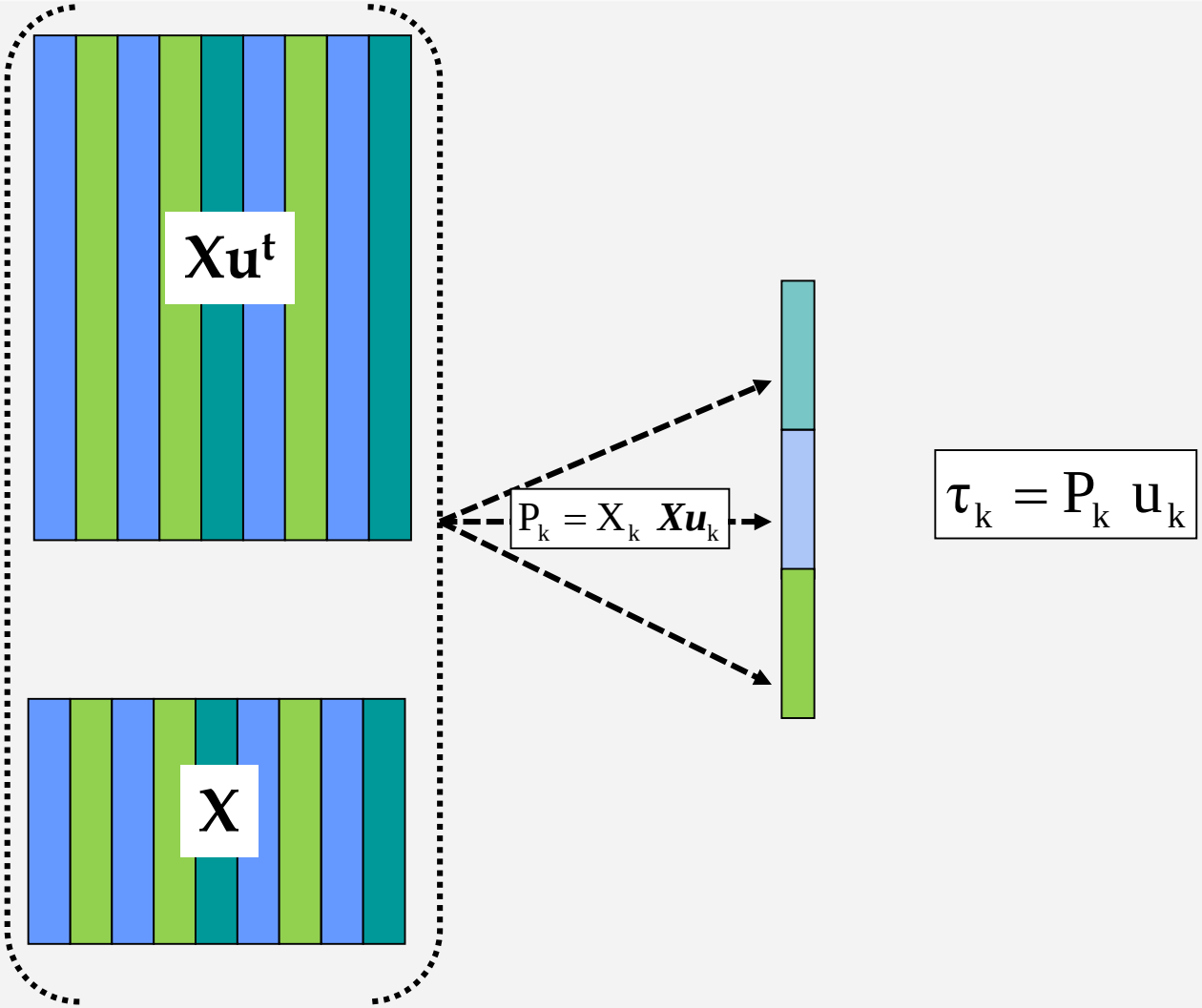
# La méthode CLV pour différentes structures de données



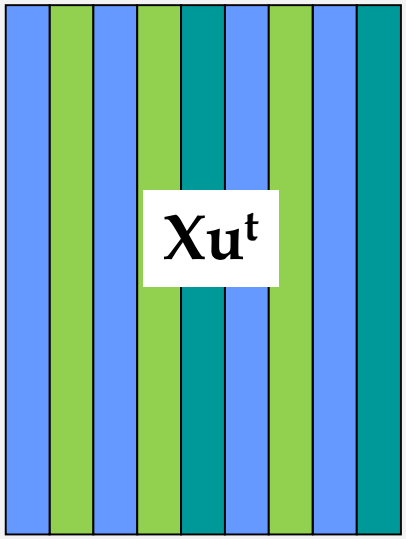
# La méthode CLV pour différentes structures de données



# La méthode CLV pour différentes structures de données



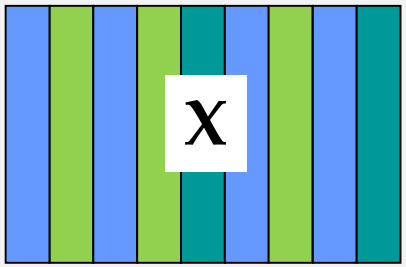
# La méthode CLV pour différentes structures de données (en L)

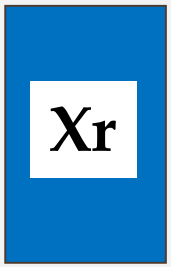



$$\tau_k = P_k \; u_k$$

$$\updownarrow$$

$$t_k = Xr \; a_k$$



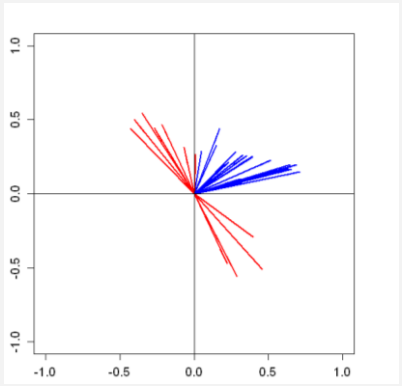



# La méthode CLV : type de groupes

## Deux cas

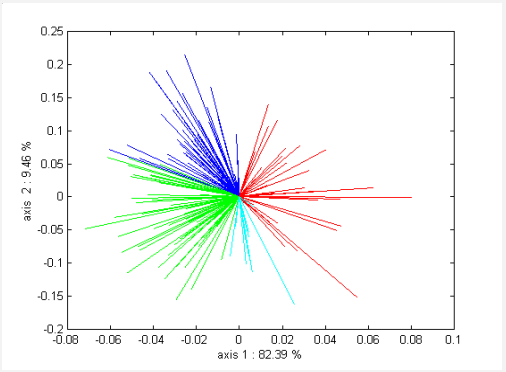
### Groupes directionnels

Fortes corrélations positives ou négatives ⇨ accord




### Groupes locaux

Fortes corrélations positives ⇨ accord
Fortes corrélations négatives ⇨ désaccord

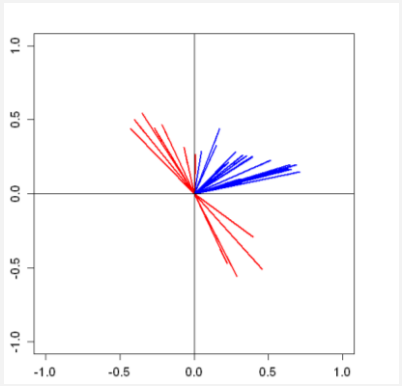


# La méthode CLV types de groupes

## Deux cas

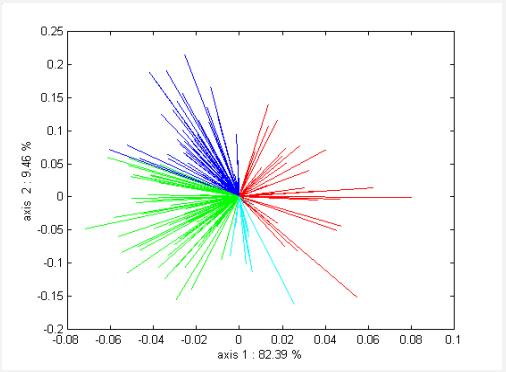
### Groupes directionnels

Fortes corrélations positives ou négatives ⇨ accord




### Groupes locaux

Fortes corrélations positives ⇨ accord
Fortes corrélations négatives ⇨ désaccord




**method = 1** — **method = 2**

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

# La méthode CLV : type de groupes

Groupes directionnels
**method = 1**

Groupes locaux
**method = 2**

Maximisation de

nb de groupes

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

variable latente

indicateur d'appartenance

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

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

# Algorithmes

## Algorithme de partitionnement (*k-means*)

**0 Initialisation :** raisonnée ( …) ou
au hasard (`nstart` fois)

**1 Etape d'estimation des variables latentes**

cas `method=1`, matrice X : $\mathbf{c_k}$ (k=1,…K) est la 1ére composante normée de $\mathbf{X_k}$

cas `method=2`, matrice X : $\mathbf{c_k}$ (k=1,…K) est proportionnelle à la variable moyenne $\bar{x}_k$

**2 Etape d'affectation des variables**

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

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

*jusqu'à convergence*

# Fonctions

## Algorithme de partitionnement (*k-means*)

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

matrice de données (n x *p*)

Choix du type de groupes

sX=TRUE / FALSE
Variables standardisées ou non

nb de fois que l'algo est répété.
(`nstart`=1 si initialisation par une partition)

- si `init` est un entier, *K* : nb de groupes de la partition
- si `init` est un vecteur de *p* entiers $\in\{1,...,K\}$ : partition initiale

Outputs :

⇨ partition en *K* groupes (si `nstart`>1, partition optimale parmi les `nstart` solutions )

⇨ variables latentes de chaque groupe (non normées)

+ valeur finale du critère, nb d'itérations avant convergence, historique des répétitions

# Algorithmes

## Algorithme ascendant hiérarchique

- Au départ (étape 1) : chaque variable forme un groupe ($K$=$p$)
- À la fin (étape $p$) : toutes les variables sont dans le même groupe ($K$=1)

*illustration pour* `method=1`

- A une étape $j$ donnée — valeur du critère $T_j$ — partition : {A, B, ….}

⬇ ⬇

- A l'étape suivante — valeur du critère $T_{j+1} < T_j$ — partition : {A ∪ B, …}

**critère d'agrégation** : $\Delta T_j = (T_j - T_{j+1}) \quad > 0$

**Stratégie** : à chaque étape, $j$, agrégation des deux groupes, A et B, qui minimisent $\Delta T_j$ (limiter la perte de cohérence intra-groupes)

**Intérêts :**

- Initialisation de l'algorithme de partitionnement
- Choix du nombre de groupes, $K$, sur la base de l'évolution des $\Delta T_j$

# Algorithmes

## Algorithme ascendant hiérarchique avec consolidation par l'algorithme *k-means*

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

nombre maximum de partitions pour lesquelles une consolidation sera effectuée (par défaut 20).

TRUE par défaut
⇨ dendrogramme
⇨ graphique de l'évolution du critère d'agrégation

Outputs :

⇨ partitions en 1, 2, 3, …, `nmax` groupes avant consolidation (coupure du dendrogramme) **et** après consolidation (*k-means*).

⇨ variables latentes de chaque groupe pour ces partitions

⇨ tableau des résultats détaillés de la hiérarchie

# Fonctions

On utilise les mêmes fonctions
avec ou sans prise en compte de variables externes

*Exemple disponible dans le 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)
```

# Exemple n°1 : analyse exploratoire d'échelles de mesure

- **Projet AUPALESENS** (2010-2014)

  ''Améliorer le plaisir à manger des personnes âgées pour prévenir les risques de dénutrition et maintenir un bon état de santé'

- n=559 personnes (>65 ans)

- Questionnaire pluridisciplinaire …extraction des

  **questions relatives au comportement psychologique** (5-points Likert scale)

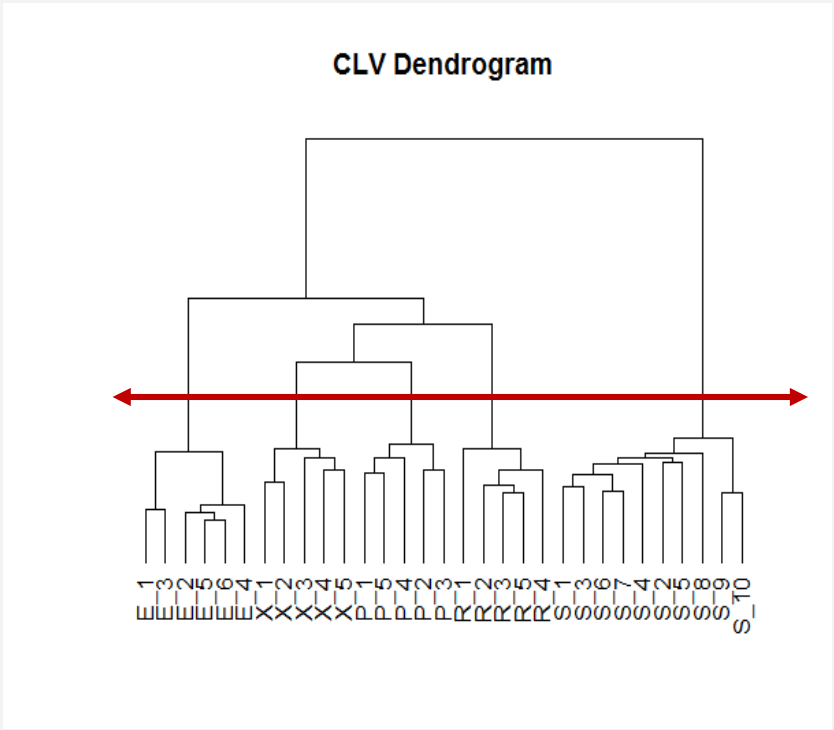
**Bailly, Maitre, Amand, Hervé, Alaphilippe (2012). Appetite, 59(853-858)*

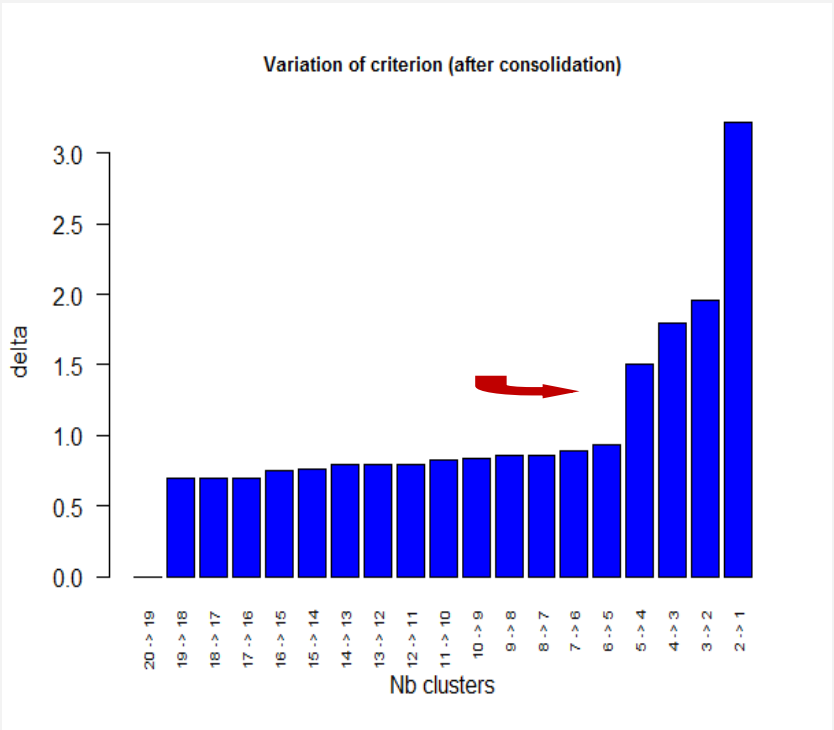
comportement alimentaire

- « Emotion » (E) : 6 items
- « eXternalité » (X) : 5
- « Restriction » ( R) : 5 items
- « Plaisir pour l'alimentation» (P) : 5 items
- « estime de Soi» (S) : 10 items

# Exemple n°1

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



# Exemple 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
```

la part de la variabilité intra-groupe expliquée par chaque variable latente

la part de la variabilité totale expliquée par les 5 variables latentes

matrice des corrélations entre variables latentes

# Exemple n°1

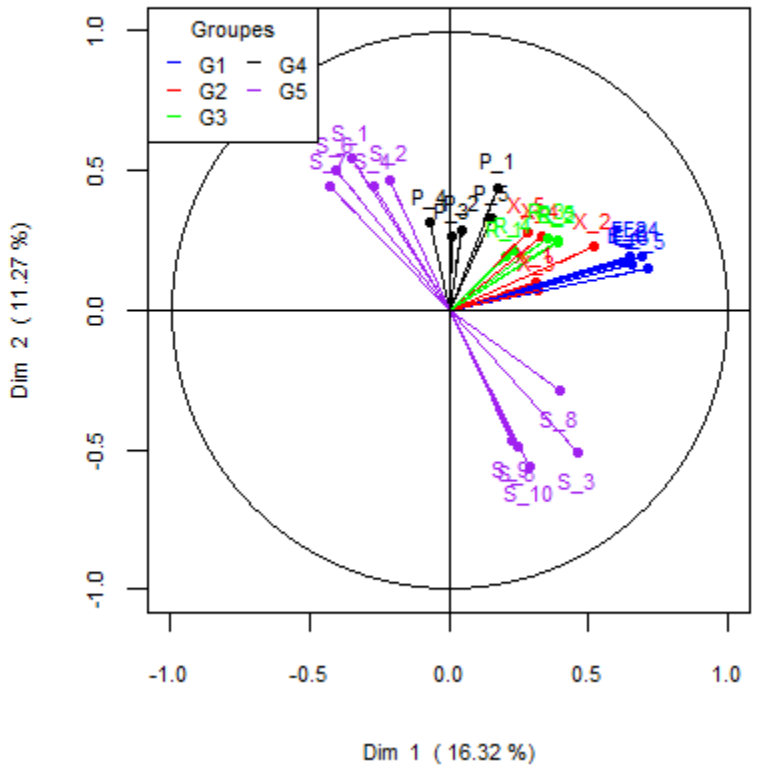
```
> descrip_gp(res.clv,X,K=5)
```

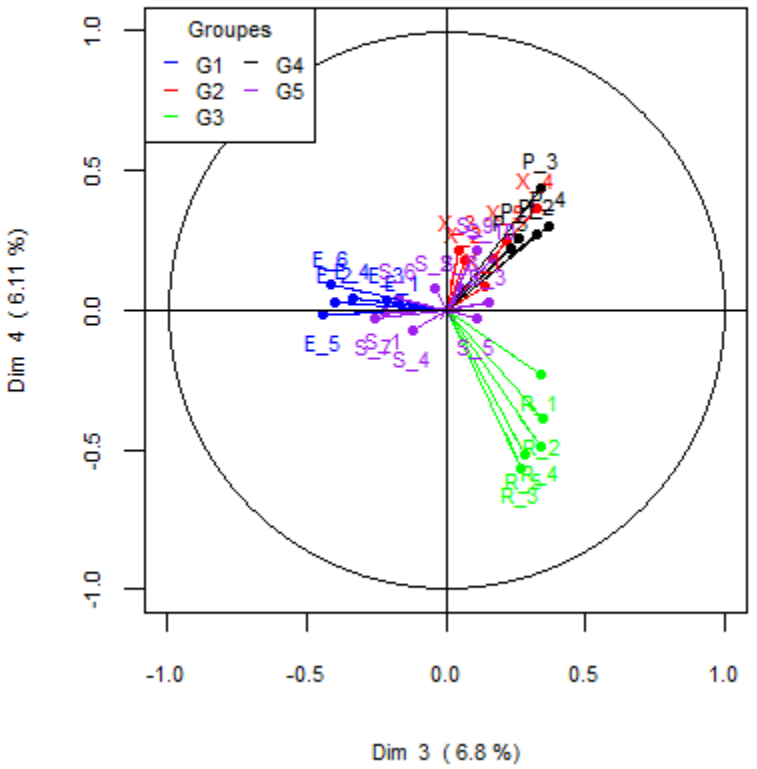
suite

```
$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
```

# Exemple n°1 : analyse exploratoire lors de la construction d'échelles de mesure

```
> 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)
```



# Exemple n°2 : cartographies de préférences de pommes, avec L-CLV

**Questionnaire consommateurs** **$Xu^t$**

- Fréquence de consommation,
- Connaissance des variétés de pomme
- Critères sensoriels importants,
- Manière de consommer (pelé, en quartier, …
- Critères d'achat (couleur, prix, …)
- ……
- Age, sexe, activité professionnelle….

consommateurs

produits

**Test hédonique** **X**

224 consommateurs réguliers
31 variétés de pommes

scores d'appréciation (échelle 9 points)

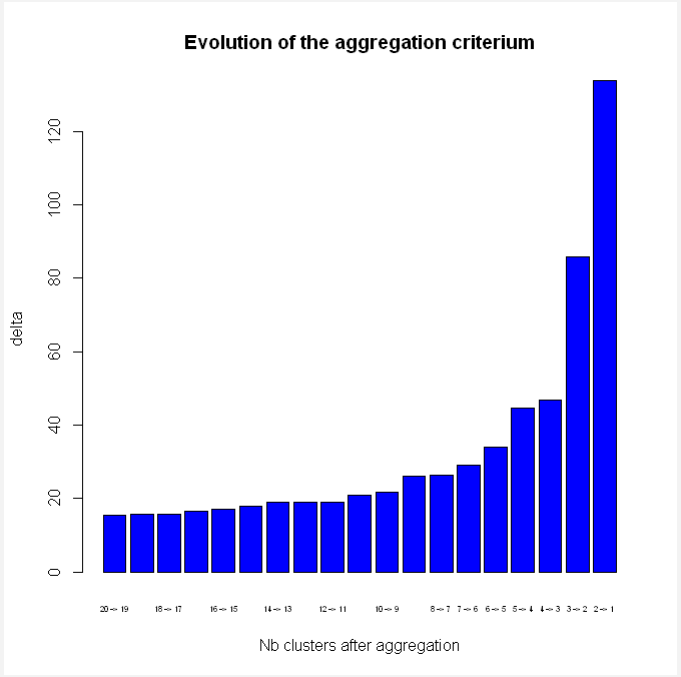
**Analyse sensorielles** **Xr**
15 juges, 15 descripteurs

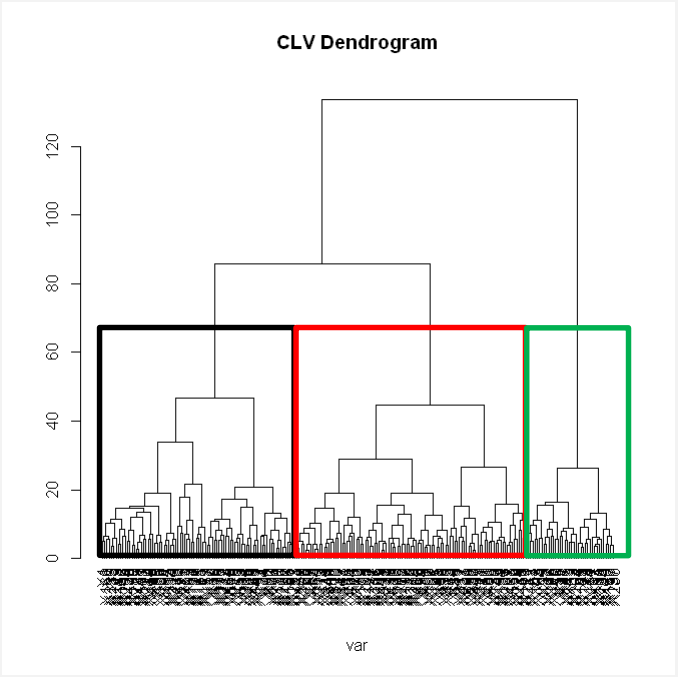
| | |
|---|---|
| Croquant | A_Ananas/Banane |
| Juteux | A_Sucré/Rose |
| Fondant | A_Boisé/Terre |
| | A_Rustique |
| Sucré | A_Citron |
| Acide | A_Fleurs blanches |
| | A_Fruit mûr |
| Intenisté odeur | A_Vert |
| Intensité arôme | |

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

# Exemple n°2

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

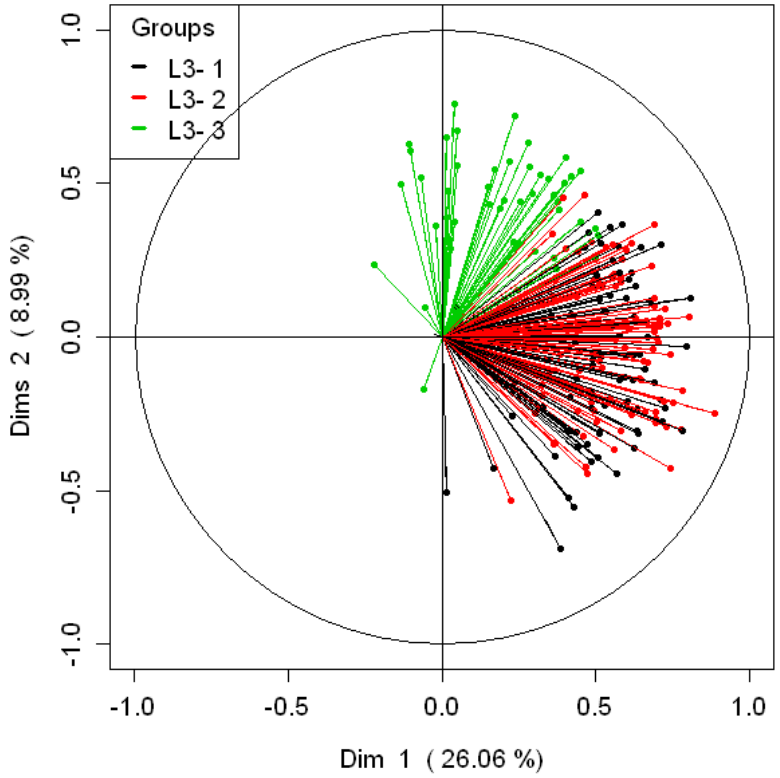


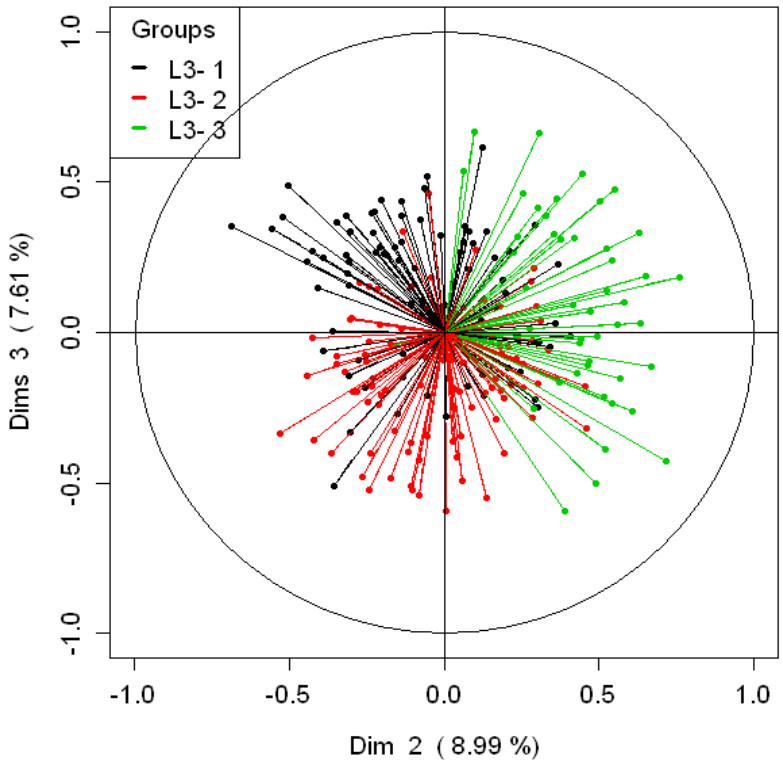




| | | |
|---|---|---|
| Segment L3-1 | 82 consommateurs | (37%) |
| Segment L3-2 | 96 consommateurs | (43%) |
| Segment L3-3 | 46 consommateurs | (20%) |

# Exemple n°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)
```



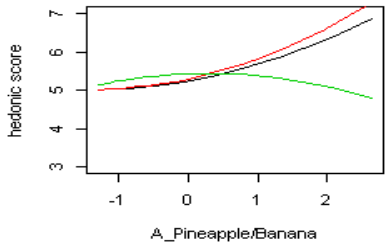
# Exemple n°2

## Segmentation du panel de consommateurs interprétable

❖ en fonction des *drivers* sensoriels

*loadings* ($a_k$) associés aux variables de Xr

- Les consommateurs des segments **1** et **2** apprécient les pommes juteuses, sucrées avec des arômes « ananas/banane »
- Les consommateurs du segment **3** apprécient les pommes fondantes, avec des arômes « rustique », «fruit mûr », rejettent l'acidité et les arômes « vert ».



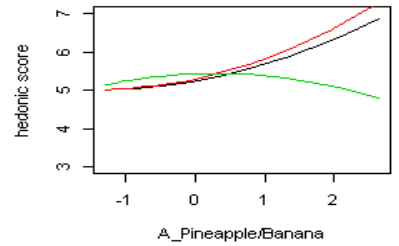
# Exemple n°2

## Segmentation du panel de consommateurs interprétable

❖ en fonction des *drivers* sensoriels

*loadings* ($a_k$) associés aux variables de Xr

- Les consommateurs des segments **1** et **2** apprécient les pommes juteuses, sucrées avec des arômes « ananas/banane »
- Les consommateurs du segment **3** apprécient les pommes fondantes, avec des arômes « rustique », «fruit mûr », rejettent l'acidité et les arômes « vert ».




❖ en termes de comportement et caractéristiques socio-démographiques des consommateurs

- Segment **1** : les plus jeunes principalement.
- Segment **2** et **3** : majorité > 40 ans

….

*loadings* ($u_k$) associés aux variables de Xu

font attention à
l'apparence, la couleur, le packaging
la variété, à l'origine.

# ClustVarLV et ClustOfVar

Basés sur l'approche CLV
algorithmes similaires (hiérarchique et k-means)

| | |
|---|---|
| **Type de groupes** | |
| directionnels ou locaux | directionnels |
| **Standardisation** | |
| au choix | Variables quantitatives normées |
| Variables qualitatives | |
| Avec codage disjonctif complet, classification des modalités | *intégrables*. adaptation du critère de classification |
| Variables externes | |
| *intégrables* pour les obs. et/ou les variables | - |

# Conclusion et perspectives

ClustVarLV : classification de variables

… mais pas seulement

- réduction de la dimensionnalité de données (var. latentes)
- interprétabilité des composantes CLV

# Conclusion et perspectives

ClustVarLV : classification de variables

… mais pas seulement

- réduction de la dimensionnalité de données (var. latentes)
- interprétabilité des composantes CLV

De nombreux domaines d'applications : analyse sensorielle et analyse de préférence, chimiométrie (infrarouge, RMN), données –omiques, psychométrie, questionnaires de satisfaction…

# Conclusion et perspectives

ClustVarLV : classification de variables

… mais pas seulement

- réduction de la dimensionnalité de données (var. latentes)
- interprétabilité des composantes CLV

De nombreux domaines d'applications : analyse sensorielle et analyse de préférence, chimiométrie (infrarouge, RMN), données –omiques, psychométrie, questionnaires de satisfaction…

Développements en cours

- Nettoyage des groupes des variables atypiques, peu associées à la structure de groupes du jeu de données.
- Classification supervisée de variables (orientée vers la prédiction d'une réponse)

Merci pour votre attention