# ClustVarLV: an R package for the clustering of variables around latent variables

Evelyne Vigneau
Sensometrics and Chemometrics Laboratory, ONIRIS, Nantes

AgroStat 2016.
Lausanne, Switzerland,
March 21-24 2016

## Introduction

- The clustering of variables is a strategy for deciphering the underlying structure of a data set.
- The Clustering of variables around Latent Variables (CLV) method[1] makes it possible to identify homogeneous groups of variables and, simultaneously, a latent variable in each group. It has been implemented in the ClustVarLV R package[2].
- The main functionalities of this package are illustrated by considering a sensory analysis study[3] of 12 varieties of apple from South Hemisphere, described using 43 sensory attributes. They were also assessed by a panel of 60 consumers for their degree of liking (0-100).

## Identifying directional groups of sensory variables

```
R> library(ClustVarLV)
R> data(apples_sh)
R> resclv_senso <- CLV(X = apples_sh$senso,
        _ method = "directional", sX = TRUE)
R> plot(resclv_senso,type="dendrogram")
R> plot(resclv_senso,type="delta")
R> summary(resclv_senso, K = 4)
R> LVsenso<-get_comp(resclv_senso, K = 4)
R> plot_var(resclv_senso, K = 4, axeh = 1,
         _ axev = 2,label=TRUE, cex.lab=0.7)
```

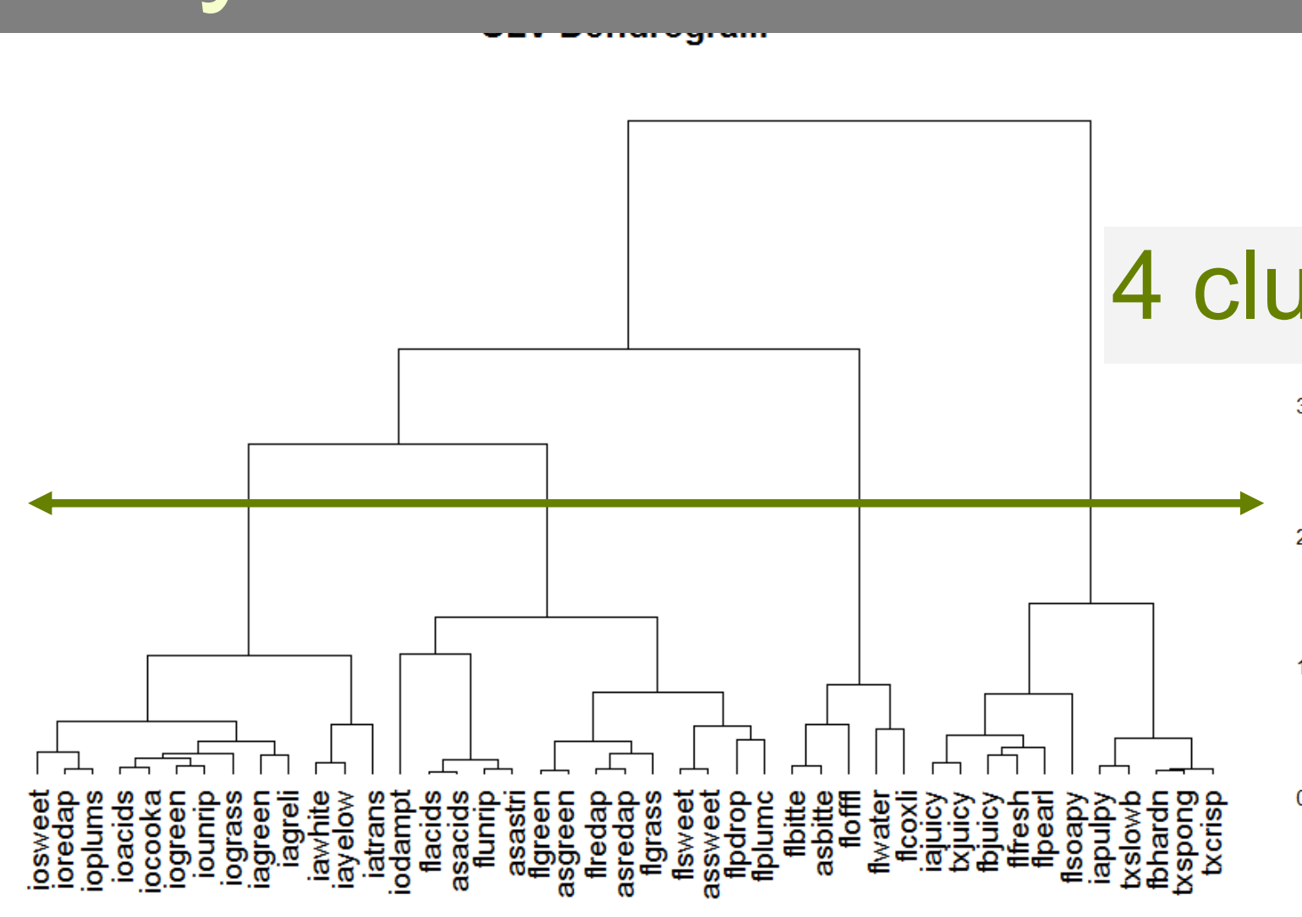
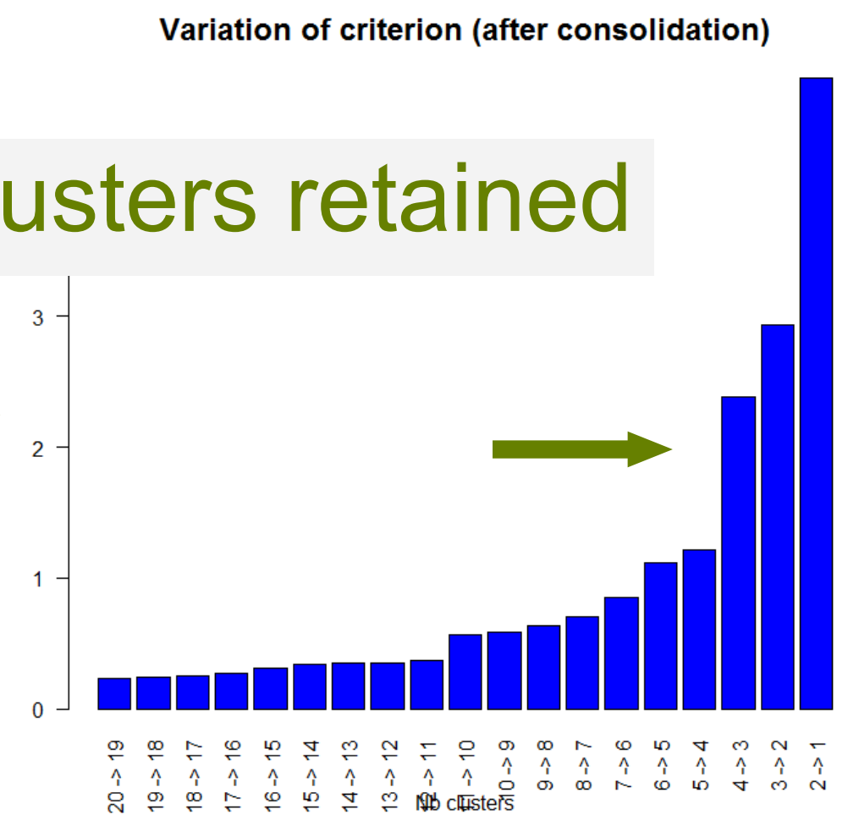


Fig.1 : Dendrogram of the 43 sensory attributes (directional CLV criterion)



4 clusters retained

Fig.2 : Variation of the clustering criterion.

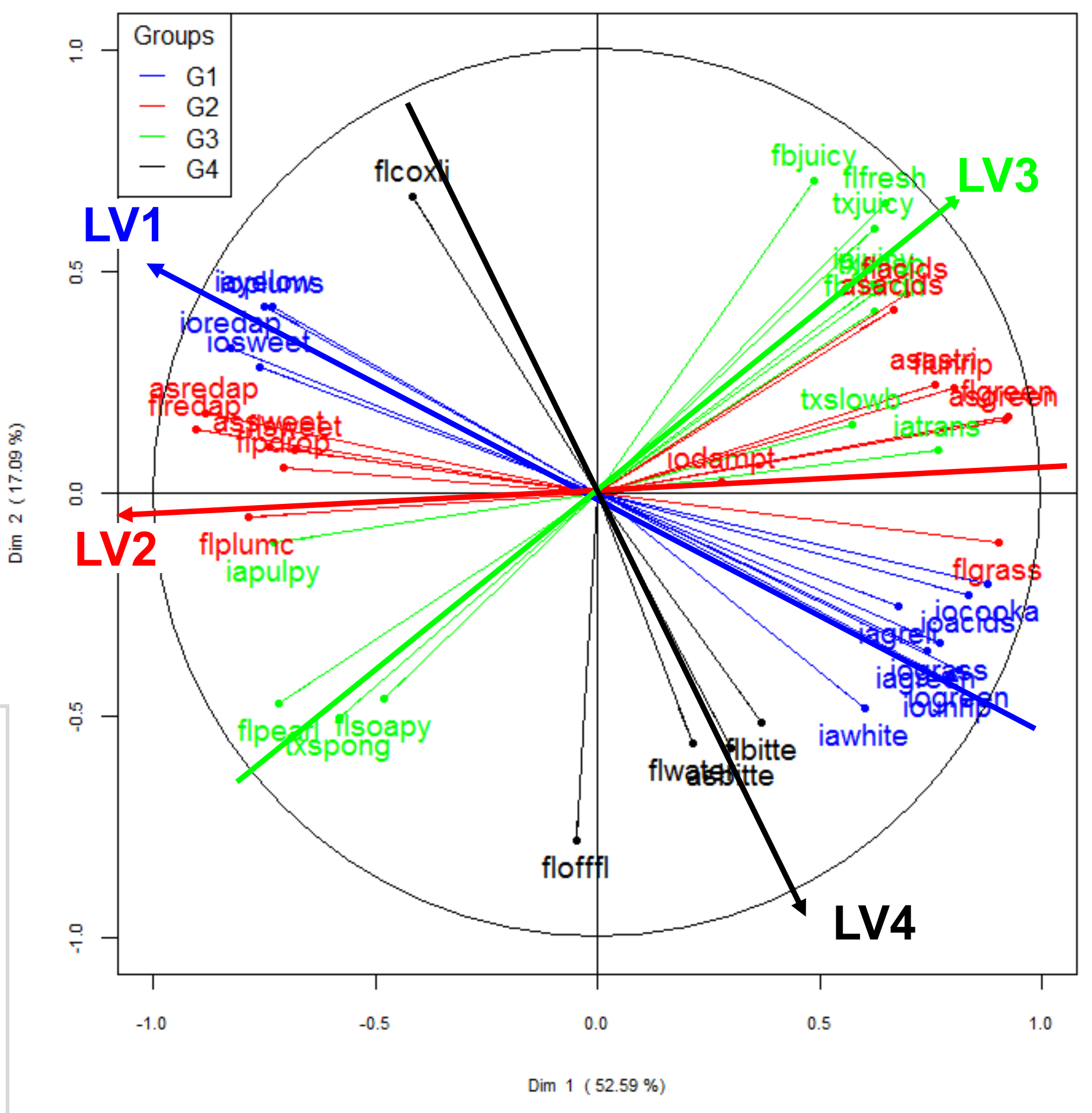


Fig.3 : Mapping of the sensory attributes( PCA) with the 4 groups of variables highlighted.

Four sensory latent dimensions (« LVsenso ») were highlighted :

- LV1 : internal odor and color of the apples (12 attributes)
- LV2 : flavor, from green to red apples (14 attributes)
- LV3 : mainly texture attributes (12 attributes)
- LV4 : mainly bitterness

```
Size of the clusters
 1  2  3  4
12 14 12  5

% of var explained within
   G1        G2        G3        G4
   83.5%     73.4%     73.4%     72.9%
% of the total var. explained by the LV: 76.2%

G1          cor in group  |cor|next group
iogreen           0.98               0.74
ioredap          -0.97               0.80
ioacids           0.96               0.74
iounrip           0.96               0.68
iocooka           0.96               0.81
...
```

Fig.4 : extract of the output of summary function.

## Segmenting a panel of consumers while taking account of sensory external information.

- The latent variable in each cluster of variables can be constrained to be a linear combination of the external variables[1].
- Each segment of consumers is described by a latent variable and a vector of loadings highlighting its drivers of liking.

```
R>resclv_segextC <- CLV(X=apples_sh$pref,
        _ Xr=cbind(LVsenso,LVsenso^2),
        _ method="local", sX=FALSE, sXr=TRUE)
R> plot(resclv_segext,type="delta")
R> plot_var(resclv_segextC , K=2)
R> load2G <- get_load(resclv_segext, K=2)
```

- Local groups of variables/ consumers are sought.
- As external variables, the latent sensory dimensions and their squared effects are considered.
- Two segments of consumers are retained.

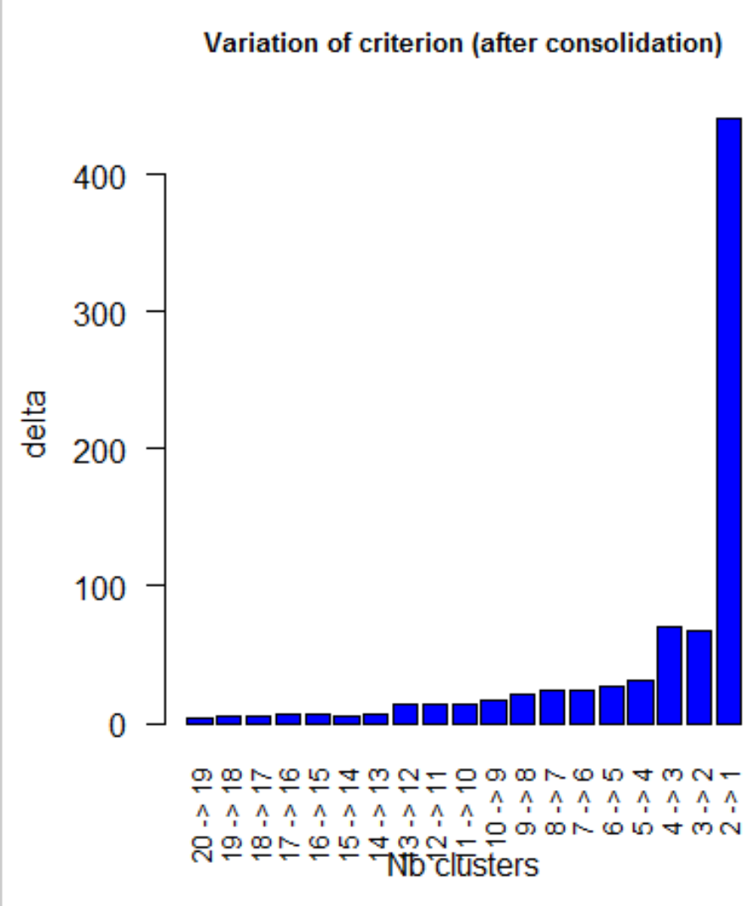


Fig.5 : Variation of the clustering criterion.

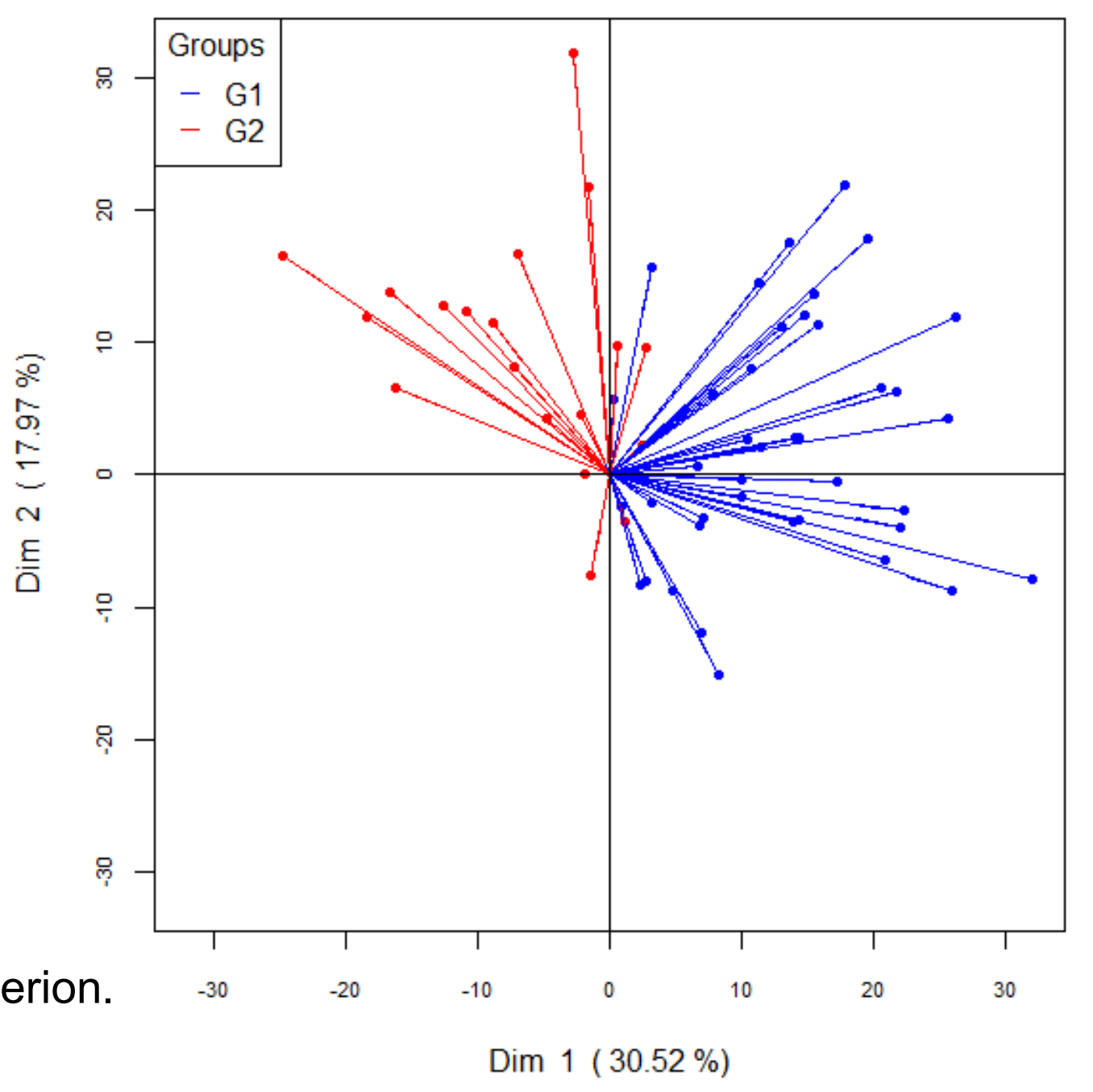


Fig.6 : preference mapping with the two segments.

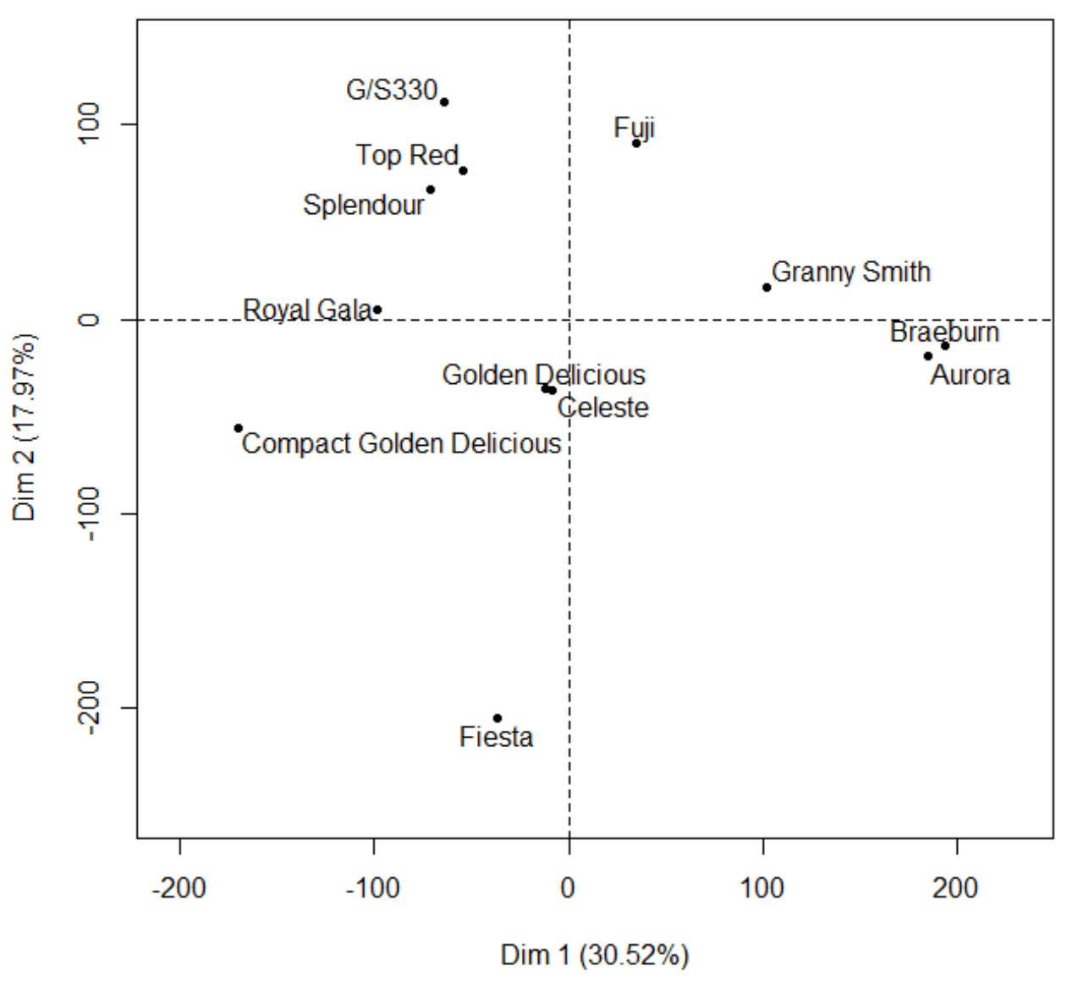


Fig.7 : Configuration of the products.

The first segment of consumers (68%) appreciated products with a crisp and juicy texture, the flavor of green -type apple.

The second segment (32%) is attracted by red-type apple.

Preference are mainly explained by only linear effects.

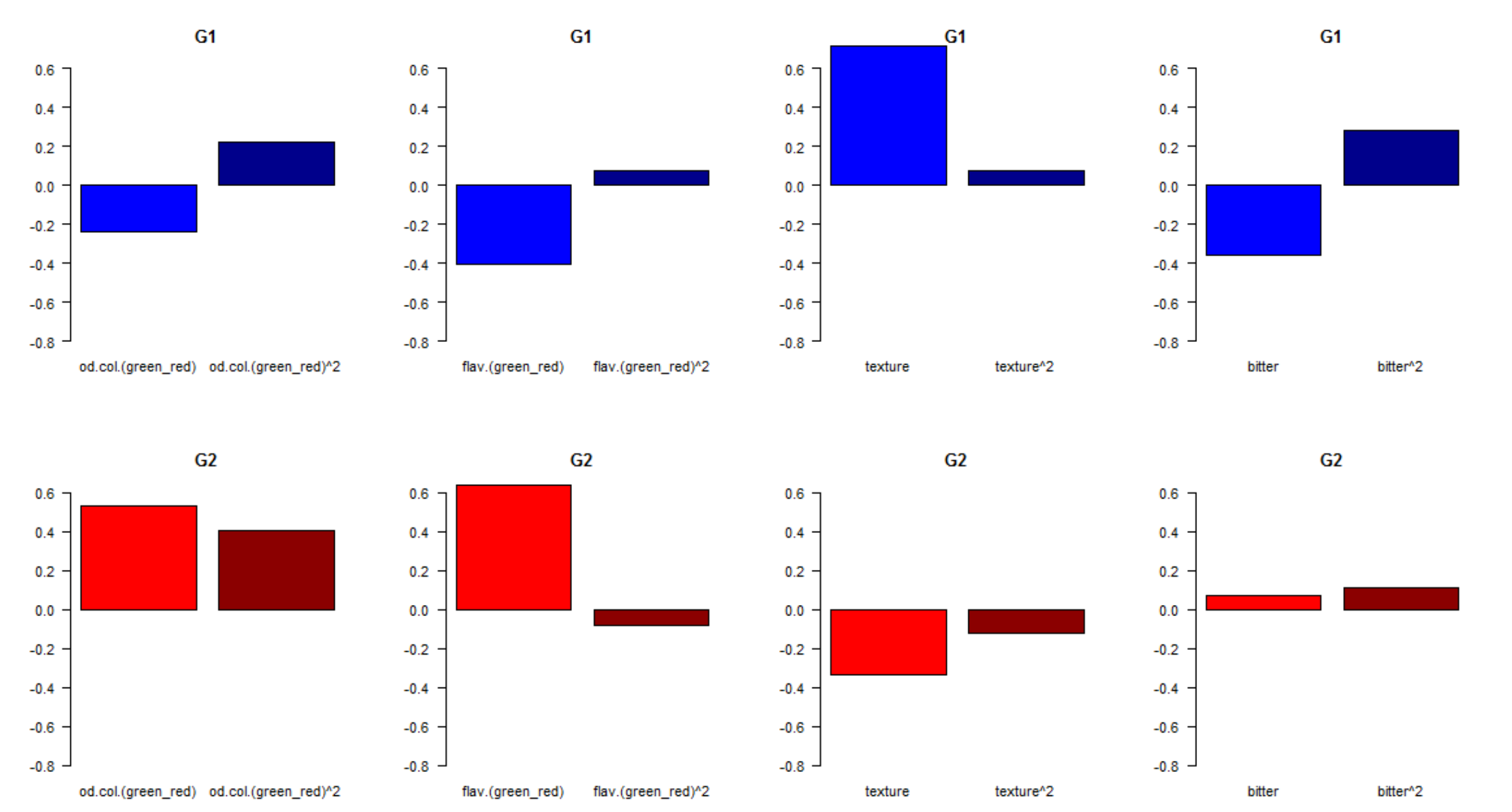


Fig.8 : Loadings in the segments of consumers regarding the sensory characteristics of the products.

## Segmenting a panel of consumers while setting aside atypical or noisy consumers.

The CLV_kmeans() function makes it possible to define clusters while setting aside the variables which are not well associated with the structure, using either the “K+1” or the “SparseLV” strategy [4].

```
R> r=0.25   # for instance
R> resclvkp1_pref<-
     _ CLV_kmeans(X=apples_sh$pref, clust=2,
     _ nstart=500, method=“local”,
     _ sX= FALSE, strategy="kplusone",rho=r)
R> nb_noise<-sum(
     _ get_partition(resclvkp1_pref==0)
R> plot_var(resclvkp1_pref,axeh=1, axev=2,
     _ label=FALSE, beside=TRUE)
```

- Two groups of variables/ consumers are sought in addition to a « noise cluster ».
- Several values of the thresholding parameter, $\rho$ (from 0 to 1, by 0.05), can be tested.
- The number of variables discarded in the noise cluster is plotted as a function of $\rho$ .

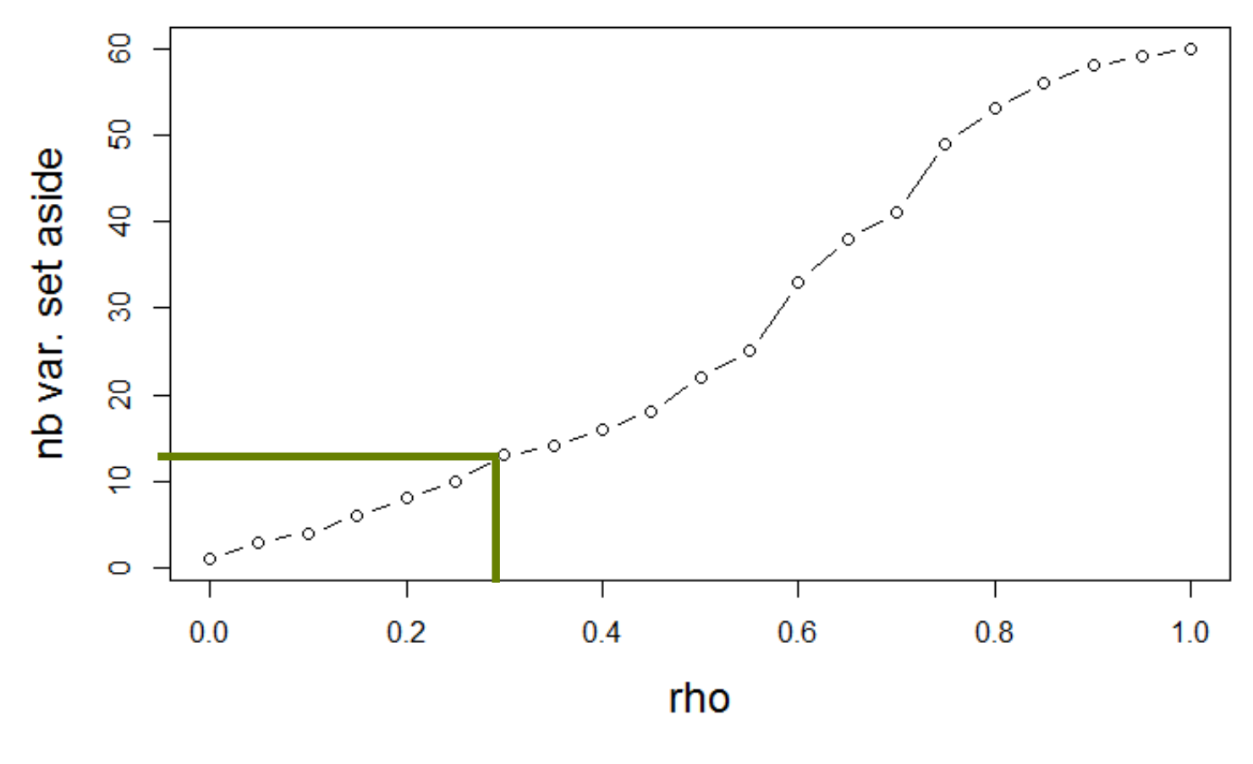


Fig.9: nb of consumers in the « noise cluster » as a function of $\rho$.

- With $\rho$ = 0.25, 10 consumers (17% of the panel) are set aside.
- They are plotted, alongside the consumers in both segments in Fig.10.
- Nine of them were previously in the segment G1, and one was in G2.

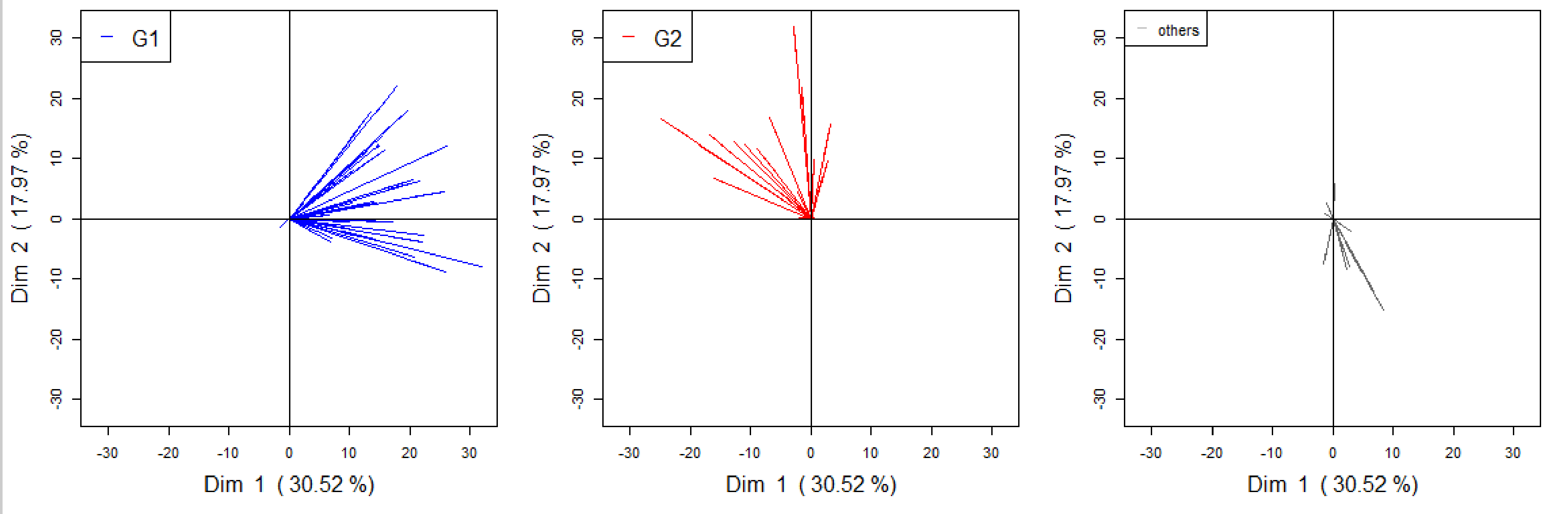


Fig.10 : Preference mapping with two segments in addition to a « noise cluster ».
The global configuration has been separated into 3 subplots for readability purpose

[1] Vigneau, E. & Qannari, E.M. (2003). Clustering of variables around latent component. *Communications in Statistics, Simulation & Computation*, 32, 1131–1150.

[2] Vigneau, E., Chen, M. & Qannari, E. M. (2015c). ClustVarLV: An R Package for the Clustering of Variables Around Latent Variables. *The R Journal*, 7(2), 134-148.

[3] Daillant-Spinnler, B., MacFie, H.J.H., Beyts, P.K. & Hedderley, D. (1996). Relationships between perceived sensory properties and major preference directions of 12 varieties of apples from the Southern Hemisphere. *Food Quality and Preference*, 7, 113–126.