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

Evelyne Vigneau

Sensometrics and Chemometrics Laboratory, ONIRIS , INRA, Nantes
E-mail :*evelyne.vigneau@oniris-nantes.fr*

Abstract
The Clustering of variables around Latent Variables (CLV) method has been implemented in an R package: ClustVarLV. The main functionalities of this package are described with a specific focus on sensory and preference data analysis.

Keywords: Clustering of variables, R package

1. Introduction

Variables clustering makes it possible to identify the underlying structure in the measurement space. We consider specifically the CLV, or Classification of variables around Latent Variables, method (Vigneau & Qannari, 2003).

By dividing the set of observed variables into homogeneous and distinct groups, this method offers a straightforward alternative to Principal Components rotation approaches (e.g. Varimax, Promax, …). More precisely, the CLV approach aims to identify a perfect simple structure, in the sense that each variable has exactly one non-zero loading for one latent variable. The CLV latent variables may be used for reducing the dimensionality of the data and more easily interpreting complex problems. Unlike Principal Components, the CLV components are not necessarily orthogonal. They are not designed to take account of as much total variance as possible, but they may be more relevant in terms of interpretation.

The CLV method has been implemented in the `ClustVarLV` R package (Vigneau & Chen, 2015b, Vigneau et al., 2015c). This package includes two main functions: `CLV()` and `CLV_kmeans()`. The `CLV()` function performs an agglomerative hierarchical algorithm followed by a consolidation step performed on the highest levels of the hierarchy. The number of solutions considered for the consolidation can be chosen by the user (parameter *nmax*, equal to 20 by default). The consolidation is based on an alternated optimization algorithm, *i.e.* a k-means partitioning procedure, which is initialized by cutting the dendrogram at the required level. Alternatively, the user may choose to use the `CLV_kmeans()` function which is typically a partitioning algorithm for clustering the variables into a given number, *K*, of clusters. It involves either repeated random initializations or an initial partition of the variables supplied by the user. This second function may be useful when the number of variables is larger than a thousand because in this case the hierarchical procedure is likely to be time consuming. However, when the number of variables does not exceed several hundred, the dendrogram obtained with the `CLV()` function provides a useful tool for choosing an appropriate number, *K*, of groups of variables.
These functions allow the user to perform the clustering of *p* quantitative variables, measured on *n* observations, gathered in the data matrix X (*n* x *p*), as described in Section 2. When external information is available, either on the observations or on the variables, additional parameters make it possible to take into account of this information while defining the clusters of the X-variables (see Section 3). Finally,

Section 4 addresses the use of a specific input parameter included in the `CLV_kmeans()` function. This parameter has been introduced to allow putting aside some of the variables, likely to be atypical or noise variables, in order to identify more compact and reliable groups of variables.

2. Clustering of variables using the `ClustVarLV` package

Firstly, we consider the case where only the block of the X-variables is available. In this simple situation, two objectives for the clustering of the variables may be distinguished (Figure 1):

- When the aim is to group correlated variables in the same cluster, whatever the sign of the correlation coefficient, each cluster is to be defined directionally around a new axis. In this case, the algorithm aims to maximize:

$$T = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \; cov(\boldsymbol{x_j}, \boldsymbol{c_k})^2 \text{ with } var(\boldsymbol{c_k}) = 1$$

where K is the number of clusters, $\boldsymbol{x_j}$ (j=1,..., p) is the j^{th} variable to be clustered, $\boldsymbol{c_k}$ (k = 1,...,K) is the latent variable associated with cluster G_k and δ_{kj} reflects a crisp membership (δ_{kj} = 1 if the j^{th} variable belongs to cluster G_k and δ_{kj} = 0, otherwise).
It is noteworthy that, for a given partition of the variables, the optimum value of T is obtained when the latent variable $\boldsymbol{c_k}$ within each cluster is the first normalized principal component of $\mathbf{X_k}$, the matrix formed by the variables belonging to the k^{th} cluster.

- When the aim is to separate variables that are highly but negatively correlated, each cluster must be defined locally around a latent variable that has the same orientation as the variables in the cluster. In this case, the criterion to be maximized is:

$$S = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \; cov(\boldsymbol{x_j}, \boldsymbol{c_k}) \quad \text{with } var(\boldsymbol{c_k}) = 1$$

It can easily be shown that, for a given partition of the variables, the optimum value of S is obtained when the latent variable $\boldsymbol{c_k}$ within each cluster is proportional to the mean of the variables of this cluster.

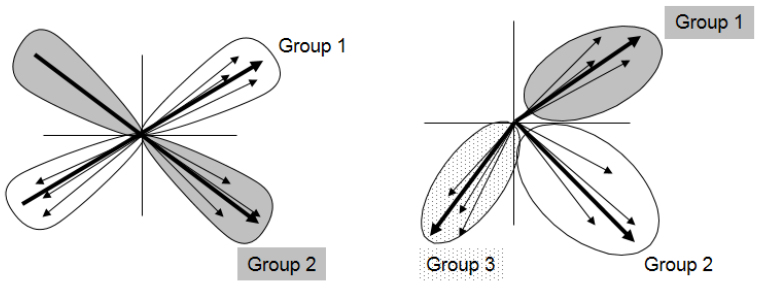

Figure 1: Two types of clusters in CLV. On the left side: directional groups, on the right side: local groups. Arrows indicate variables and bold arrows indicate latent components associated with the various groups

Let us consider a case study (Daillant-Spinnler *et al.*, 1996) dealing with the sensory analysis of 12 varieties of apple from South Africa and New Zealand. The apples were peeled and quartered, and then assessed according to the 43 sensory attributes. The same varieties were also assessed by a panel of 60 consumers who were asked to indicate their degree of liking (between 0-100) for each sample. This dataset is available in the `ClustVarLV` R package:

```
R> library(ClustVarLV)
R> data(apples_sh)
```

For the purpose of reducing the complexity in the sensory space by identifying synthetic sensory latent variables, the clustering of the 43 sensory attributes was undertaken. As directional groups of variables were sought, the following code may be used:

```
R> resclv_senso <- CLV(X = apples_sh$senso, method = "directional", sX = TRUE)
R> plot(resclv_senso,type="dendrogram")
R> plot(resclv_senso,type="delta")
```

`CLV()` function performs the hierarchical clustering of the sensory attributes around directional latent variables, and uses the solutions, from *K*=2 to *K*=20 (the default), obtained by cutting the dendrogram in order to consolidate partitions of different sizes. The `plot()` function provides helpful graphical displays for the choice of the number of clusters to be retained for further investigation. Herein, four clusters were retained.

The detailed description of the groups of variables, the four latent variables and graphical displays showing the group membership of the variables on the basis of selected Principal Components, may be obtained by:

```
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)
```

If we consider now the consumers' preference data, the aim is to identify groups, or segments, of consumers with similar preferences. Since consumers with opposed directions of preference are to be separated into distinct segments, local groups are sought. Herein two segments of 42 and 18 consumers, respectively, were identified and described:

```
R> resclv_pref <- CLV(X = apples_sh$pref, method = "local", sX = FALSE)
R> plot(resclv_pref,type="dendrogram")
R> plot(resclv_pref,type="delta")
R> plot_var(resclv_pref, K = 2, axeh = 1, axev = 2)
R> data_biplot(X = apples_sh$pref,sX = FALSE, axeh = 1, axev = 2,cex.lab=0.7)
R> get_partition(resclv_pref,K=2)
R> get_comp(resclv_pref,K=2)
```

3. Clustering of variables while taking account of external information.

The CLV criteria can easily be adapted to the case where external information is available. Suppose, for instance, that in addition to the X-variables to be clustered, the observations are described by a second block of variables, denoted **Xr** (r stands for additional information collected on the rows of the matrix **X**). In this case, each latent variable, $\boldsymbol{c_k}$ (*k*=1,…,*K*), associated to each group of variables is defined under the constraint that:

$$\boldsymbol{c_k} = \mathbf{Xr}\,\boldsymbol{a_k}\ with\ \boldsymbol{a_k'a_k} = 1$$

It can be shown (Vigneau & Qannari, 2003) that the solutions of the optimization problems are obtained when $\boldsymbol{c_k}$ is the first component of a Partial Least Squares (PLS) regression of the group matrix $\mathbf{X_k}$ on the external matrix **Xr**, in the case of directional groups, or the first component of a PLS regression of the centroid variable $\overline{\boldsymbol{x_k}}$ on the external matrix **Xr**, in the case of local groups.

External preference mapping is a domain in which the CLV approach with additional information on the observations has successfully been applied. In addition to clustering the consumers according to the similarity of their preference scores, the aim is also to segment the consumers while explaining the preferences by means of the sensory characteristics of the products. Thus, the segmentation and the modeling of the main directions of preference may be achieved simultaneously. If we consider again the

'apples_sh' dataset, parameter *Xr*, available for `CLV()` and `CLV-kmeans()` functions, can be used for taking account of the 43 sensory attributes as external block of information. Namely:

```
R> resclv_segext<-CLV(X=apples_sh$pref, Xr=apples_sh$senso, method="local",sX=FALSE,
sXr = TRUE)
```

Or alternatively, by taking into account the four sensory latent variables (see Section 3) and their quadratic effect:

```
R>resclv_segextC <- CLV(X = apples_sh$pref, Xr = cbind(LVsenso,LVsenso^2), method =
"local", sX = FALSE, sXr = TRUE)
```

The loadings of the sensory variables on the latent variable associated to the clusters of consumers can be extracted using the `get_load()` function.

When additional information is available on the variables, the CLV approach has also been adapted in order to take this information into account in the clustering process. This type of additional information may be data collected by means of questionnaires when the variables to be clustered are, for instance, consumers in preference studies. In other context, this may be an external "proximity" information between the variables, such as the spectral sequence of NMR or Near-Infrared data. If we denote by **Xu**, the matrix of the additional information on the variables, the rows in **Xu** are matched with the columns of the matrix **X**. The CLV approach consists in combining, in each cluster of variables, the X- and the Xu-information. Technically, the parameter *Xu*, available for `CLV()` and `CLV_kmeans()` functions, is to be used in order to define group latent variables that are contrained to be linear combinations of the external Xu-information.

Finally when the three data tables, X, Xr and Xu, are simultaneously available, so that the three blocks of data may be arranged in the form of an L, a specific function, named `L-CLV()`, has been included in the package `ClustVarLV`. In Vigneau *et al.* (2014), a case study illustrates this procedure for the segmentation of a panel of consumers, according to their likings (**X**), interpretable in terms of socio-demographic and behavioral parameters (given in **Xu**) and in relation with the sensory key-drivers (in **Xr**).

4. Clustering of variables while setting aside atypical and noise variables.

Recently, the CLV approach has also been updated in order to combine clustering of variables and the possibility to "omit" the variables which are not well associated with the highlighted group structure. Two strategies have been investigated (Vigneau & Chen, *in press*):

- The "K+1" strategy consists in introducing an additionnal group, called "noise cluster" for simplicity, for handling the atypical or noise variables. A variable $\boldsymbol{x_j}$ will be assigned to cluster G_k in which the (squared) covariance between $\boldsymbol{x_j}$ and $\boldsymbol{c_k}$ is greater than with other latent variables, except if this value is too small in comparison to the value of a pre-specified parameter, ρ. In this latter case, the variable $\boldsymbol{x_j}$ will be assigned to the "noise cluster".

- The "sparse LV" strategy consists in assigning a zero loading, *e.g.* $u_{jk}=0$, for the variable $\boldsymbol{x_j}$ to the latent variable $\boldsymbol{c_k}$ if the contribution of the variable $\boldsymbol{x_j}$ to $\boldsymbol{c_k}$ is small. This leads to sparse latent variables. For directional groups, the iterative soft thresholding algorithm, proposed by Shen & Huand (2008) for Sparse principal component analysis has been adapted. The thresholding procedure depends on a parameter, ρ, as for the "K+1" strategy.

The tuning parameter, ρ, is to be chosen between 0 and 1. It is analogous to a correlation coeffcient. If ρ =0, there will be no variable assigned to the « noise cluster », or, with a zero loading with respect to the latent variable in the cluster G_k in which the variable has been assigned. Contrariwise, if ρ is chosen close to 1, then the size of the "noise cluster" will be large, or, any loading within each cluster will tend to be null. A screening of a range of values, from 0 to 1 by predefined steps (0.1, 0.05 or 0.01), can provide a guide to the choice of a specific value for ρ.

The `CLV_kmeans()` function in the `ClustVarLV` package makes it possible to use these "cleaning" procedures. The choice of the strategy is indicated via the "*strategy*" parameter (with values "*none*", "*kplusone*" or "*sparselv*"). The threshold of correlation parameter, ρ, is defined via the "*rho*" parameter.

Let us consider once again the consumers' preference data for the 12 varieties of apples. The clustering of the 60 consumers into *K*=2 groups can be performed in combination with the "K+1" strategy, as follows:

```
R> resclvkp1_pref <- CLV_kmeans(X = apples_sh$pref, clust=2, nstart=500, method =
"local", sX = TRUE, strategy="kplusone", rho=0.25)
```

According to the previous instruction, the k-means algorithm started from an initial partition into two clusters defined at random. However, the procedure was repeated 500 times (`nstart`) and the best partition (maximal value of the criterion *S*) was retained. By varying the value of the threshold ρ and by using `strategy="sparselv"`, instead of `strategy="kplusone"`, the number of variables (*i.e.* consumers) which were assigned to the noise cluster, or, with a zero loading, increases steadily with ρ as shown in Figure 2. This is commonly observed with preference data (Vigneau et al. 2016) for which there is no well separated clusters, but global tendencies of preference. As a rule of thumb, we may decide to set aside 10 to 20% of the consumers. In this case, ρ=0.25 should be selected.

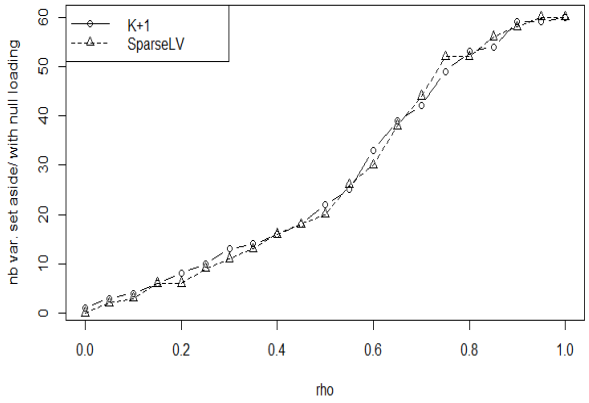

Figure 2: Number of variables (*i.e.* consumers) in the noise cluster, or, with a zero loading, as a function of the value of ρ and the strategy used.

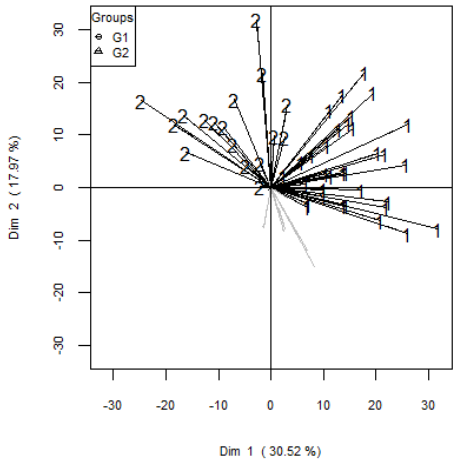

Figure 3: Internal preference mapping of the apples data with two groups of consumers highlighted, and some consumers set aside (gray lines).

The internal preference mapping, with the identification of the consumers in both clusters, as well as the consumers assigned to the "noise cluster" ("K+1" strategy) can be obtained by:

```
R> plot_var(resclvkp1_pref, axeh = 1, axev = 2,label=FALSE, v_symbol=TRUE)
```

The output is shown in Figure 3. It turns out that nine consumers have been excluded from the group G1 mainly because their directions of preference were not in good agreement with the other directions of preference of the consumers in G1. Only one consumer has been excluded from G2. This consumer had a direction of preference rather incoherent with those of the other consumers in G2.

Considering –omics data, the strategy of « cleaning » the groups of variables has been shown to be promising for the reduction of the number of variables in a pre-treatment step (Vigneau & Antignac, 2015a).

References

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.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6), 1015-1034.

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

Vigneau, E., Charles, M. & Chen M. (2014). External preference segmentation with additional information on consumers: A case study on apples. *Food Quality and Preference*, 22(4), 83–92.

Vigneau, E. & Antignac, J.-P. (2015a). *Exploratory strategy of metabolomic data based on variables clustering and noise cluster identification*. Chimiométrie 2015, Université de Genève, Suisse, 19-21 janvier 2015.

Vigneau, E. & Chen,M. (2015b). ClustVarLV: Clustering of Variables Around Latent Variables. URL https://CRAN.R-project.org/package=ClustVarLV. R package version 1.4.1.

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.

Vigneau, E., Qannari, E. M., Navez, B., Cottet, V. (2016). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. *Food Quality and Preference*, 47, 54–63.