Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n1p134

Dimensionality reduction by clustering of variables while setting aside atypical variables
By Vigneau, Chen

Published: 26 April 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 01, April 2016, 134-153
DOI: 10.1285/i20705948v9n1p134

Dimensionality reduction by clustering of variables while setting aside atypical variables

E.Vigneau*[a,b] and M. Chen[a]

[a]*Sensometrics and Chemometrics Laboratory, Oniris, National College of Veterinary Medicine, Food Science and Engineering Nantes-Atlantic, Nantes, F-44322, France*
[b]*INRA, Nantes, F-44316, France*

Published: 26 April 2016

Clustering of variables is one possible approach for reducing the dimensionality of a dataset. However, all the variables are usually assigned to one of the clusters, even the scattered variables associated with atypical or noise information. The presence of this type of information could obscure the interpretation of the latent variables associated with the clusters, or even give rise to artificial clusters. We propose two strategies to address this problem. The first is a "$K + 1$" strategy, which consists of introducing an additional group of variables, called the "noise cluster" for simplicity. The second is based on the definition of sparse latent variables. Both strategies result in refined clusters for the identification of more relevant latent variables.

keywords: dimensionality reduction, clustering of variables, noise cluster, sparse latent variables.

1 Introduction

The aim of clustering variables is to divide the set of the observed variables into homogeneous and distinct groups so as to identify the underlying structure in the measurement space. In explorative data analysis, it is a helpful tool for reducing the dimensionality of the data and interpreting complex problems more easily.

However, for high-dimensional datasets, clustering may lead to unsatisfactory results. A large number of measured variables does not necessarily mean that all the information

*Corresponding author: evelyne.vigneau@oniris-nantes.fr.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

is useful. On the contrary, atypical or noise information may lead to rather inconsistent clusters, and thus make it difficult to interpret the latent variables revealed by these clusters. When the aim is to identify the structure of groups among the variables, strategies for discarding scattered variables may be helpful. For instance, the identification of homogeneous spectral ranges, the detection of biological pathways, the extraction of segments in a panel of consumers or the development and validation of scales in social science are all domains where the main interest is to reveal latent dimensions associated with clusters of observed items, or features.

In order to clarify our purpose, let us consider a simple example with two groups of variables and several scattered variables. The variables in a group are highly correlated regardless of whether their correlations are positive or negative. The correlation structure of the whole set of variables, on the basis of the first two principal components, is shown on the left-hand side of Figure 1. All the variables identified in gray on the right-hand side of Figure 1 are expected to be set aside because they do not fit the structure defined by the two main groups. The same principle would be applied to all the variables that are isolated, scattered or random. For simplicity, all these variables are referred to as atypical or noise variables in the following. These terms are to be considered in a broad sense in relation to the true, but unknown, underlying structure.

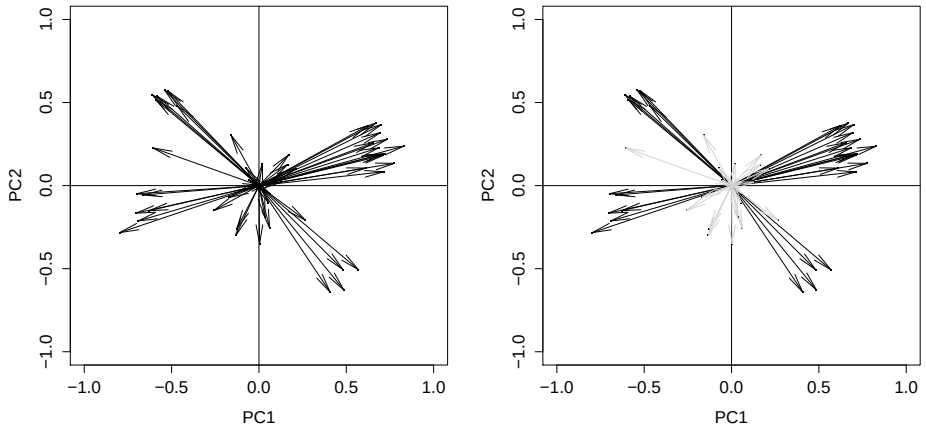

Figure 1: Correlation structure for set of variables including two clusters of highly correlated variables and several scattered variables. The expectation is to set aside the variables in grey on the right-hand side plot

For the clustering of variables, we consider here the Clustering of variables around Latent Variables (CLV) method proposed by Vigneau and Qannari (2003). This method aims to identify, simultaneously, a partition of the variables into groups as unidimensional as possible, and a set of latent variables each of which is associated with one of the groups. The CLV latent variables are synthetic variables that, like the principal components in Principal Components Analysis (PCA) or the rotated components in Factor Analysis,

enable a large dataset to be analyzed. However, the CLV latent variables are easier to interpret than the principal components because each of them is a linear combination of only those variables belonging to the same cluster. In fact, there is a wide range of methods for the clustering of variables. The most common are hierarchical approaches based on a similarity measure between the variables, such as the correlation coefficient, the squared correlation coefficient and Kendall's correlation coefficient, among others. The CLV method is conceptually close to the well-known VARCLUS procedure in the SAS software (Sarle, 1990). Diametrical Clustering (Dhillon et al., 2003), applied to gene expression data, and the weighted-variances clustering recently proposed by Enki et al. (2012) for constructing interpretable principal components are two proposals that also show some similarity with the CLV approach. Regarding the partitioning solution, the above-mentioned methods are crisp strategies where each variable is assigned to exactly one cluster. Thus, one might expect that atypical or noise variables make the groups obtained more difficult to interpret or may even give rise to artificial groups.

Fuzzy strategies, in which each entity can be allocated to several clusters with a membership between 0 and 1, are often proposed as alternatives to crisp clustering. In fact, an atypical variable is expected to have modest membership values for several clusters. In the context of the clustering of genes, Berget et al. (2005) suggested modifying the criterion used for fuzzy clustering by adding a penalty term to identify relevant clusters and a rest cluster. In the sensory analysis field, Dahl and Naes (2009) also adapted the fuzzy clustering approach by considering two clusters, a good cluster and a noise cluster, with the aim of identifying the outlying assessors (associated with the noise cluster). Both these interesting works relied on clustering criteria based on Euclidean distances between the entities to be clustered. In the CLV method, the entities considered are variables and similarity indices based on covariance (or correlation) are used.

In this work, we aim to modify the CLV strategy in order to set aside or down-weight the atypical or noise variables. We considered two approaches to achieve this goal. The first, called a "$K + 1$" strategy, is based on a modification of CLV criteria by adding a specific term so that an additional cluster (the "noise cluster") can be introduced, as in Dave (1991) and Berget et al. (2005). Thus, those variables without a clear affiliation to any of the main clusters will be put aside and assigned to this additional entity. The second approach concerns the CLV latent variables, with a sparsity constraint added, in the same vein as Sparse Principal Component Analysis (Zou et al., 2006). Thus, the variables with a low contribution to their associated latent variable will be given a zero loading in the sparse CLV latent variables.

This paper is organized as follows. Section 2.1 gives a brief overview of the CLV criteria and the algorithms used for their maximization. The "$K + 1$" and "Sparse LV" strategies are described in Sections 2.2 and 2.3, respectively. In Section 3, the proposed strategies are applied to a simulated set of data in order to explore their behavior, compared to the standard CLV approach. Finally, we consider a real case study where the aim was to reveal the sparse latent dimensions in a large multidisciplinary survey on food perception, and nutritional and health status in elderly people.

2 Methods

2.1 The CLV method

Let us consider a set of p variables observed on n individuals. We denote $\mathbf{x_j} = (x_{1j}, x_{2j}, \ldots, x_{nj})^t \in \mathbb{R}^n$, the vector of observations for the j^{th} variable. All the observed variables $\mathbf{x_j}(j = 1, \ldots, p)$ are assumed to be centered. In addition, the user may choose to standardize them, or not, to a unit variance. Given a number of clusters, K, the objective of the CLV method is to seek a partition of the observed variables into K groups $(G_1, G_2, \ldots, G_K)$ and K latent variables, $(\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_K})$ associated with each group respectively, so as to maximize the internal coherence of the groups. Two types of criteria are considered, which define two types of groups: (i) "directional groups" in which positively or negatively correlated variables will be merged together, regardless of the sign of the correlation coefficients, (ii) "local groups" in which only positively correlated variables will be associated in the same group.
For directional groups, the CLV criterion to be maximized is T, defined as :

$$T = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov}^2 \left(\mathbf{x_j}, \mathbf{c_k}\right), \text{ with } \mathrm{var}(\mathbf{c_k}) = 1 \tag{1}$$

For local groups, the clustering criterion is S, defined as :

$$S = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov} \left(\mathbf{x_j}, \mathbf{c_k}\right), \text{ with } \mathrm{var}(\mathbf{c_k}) = 1 \tag{2}$$

In (1) and (2), $\delta_{kj} = 1$ if the j^{th} variable belongs to group G_k, and $\delta_{kj} = 0$ otherwise. $\mathrm{cov}\left(\mathbf{x_j}, \mathbf{c_k}\right)$ stands for the covariance between the variable $\mathbf{x_j}$ and the latent variable $\mathbf{c_k}$ and $\mathrm{var}(\mathbf{c_k})$ is the variance of $\mathbf{c_k}$.

A partitioning algorithm is used for the optimization of T or S. It consists of two alternating steps: the assignment step, during which the δ_{kj} are defined given the latent variables $\mathbf{c_k}$, for $k = 1, \ldots, K$, and the estimation step of the latent variables, given the partition of the variables. More precisely, this algorithm is conceived as follows:

1. Initialization step. A partition into K clusters is generated at random or selected from the nested partitions obtained with a hierarchical algorithm. When random initialization is considered, several random starts (say 100) are performed and the solution leading to the largest value of the clustering criterion, T (Eq.1) or S (Eq.2), is selected. For a non-random initialization, a reasonable choice is to run beforehand an ascendant hierarchical clustering based on the considered clustering criterion. This technique proceeds sequentially from a stage where each variable is a group by itself to a stage in which all the variables are merged together. At a given level, the two clusters of variables resulting in the smallest decrease of the clustering criterion are merged together (Vigneau and Qannari, 2003). Even if the construction of a hierarchy requires additional computational resources, when the

number of variables is not too large (less than a few hundred, say), this way of initializing the alternating optimization algorithm is recommended as it offers a relevant initial solution.

2. Estimation step. In a cluster G_k $(k = 1, \ldots, K)$, the latent variable $\mathbf{c_k}$ is defined
   - for directional groups, as the first standardized principal component of the matrix $\mathbf{X_k}$ formed by the variables belonging to G_k,
   - for local groups, as the standardized mean variable of the variables in G_k.
3. Assignment step. Each variable $\mathbf{x_j}$ $(j = 1, \ldots, p)$ is considered in turn. $\mathbf{x_j}$ is assigned to the cluster G_k for which its squared covariance coefficient, for directional groups, or its covariance coefficient, for local groups, with $\mathbf{c_k}$ is higher than for every other group latent variable. More formally,
   - for directional groups:

$$\delta_{kj} = 1 \text{ if } \max_l \left\{\text{cov}^2\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\} = \text{cov}^2\left(\mathbf{x_j}, \mathbf{c_k}\right) \tag{3}$$

   - for local groups:

$$\delta_{kj} = 1 \text{ if } \max_l \left\{\text{cov}\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\} = \text{cov}\left(\mathbf{x_j}, \mathbf{c_k}\right) \tag{4}$$

4. Repeat 2 and 3 until the stability of the partition is achieved. Usually, at the convergence, there is strictly no change in the group's memberships between two iterations, and the clustering criterion will not then change any more. However, in order to manage the most complex cases, with very fuzzy structure, the stopping rule may be defined so that the variation of the clustering criterion will be less than a very small threshold (say, 1E-5). In addition, the maximum number of iterations allowed for steps 2-3 is fixed, by default, to 20. If after 20 iterations, the convergence criterion is not reached, the procedure is stopped. In practise, this last stopping rule is almost never used. For real case studies, like the one discussed in Section 3.2, two or three iterations are usually sufficient, even if the initialization is based on a random partition.

With this algorithm, all the variables are assigned to one, and only one, group. Ideally, the atypical or noise variables should be discarded. To address this question, two modified procedures were investigated.

2.2 "$K + 1$" strategy

This strategy consists in introducing an additional cluster for handling the atypical or noise variables in the clustering. Based on the concept introduced by Dave (1991), this additional cluster, also named "noise" cluster, can be represented by a prototypical variable that is expected to have the same correlation with all the observed variables $\mathbf{x_j}$ $(j = 1, \ldots, p)$. The CLV criteria are consequently updated and a fixed parameter,

ρ, representing the common correlation coefficient associated with the definition of the "noise" cluster protoype, is introduced. According to the type of groups sought, this consists of maximizing:
for directional groups, a new criterion T defined as:

$$T_{new} = \sum_{k=1}^{K}\sum_{j=1}^{p} \delta_{kj}\ \mathrm{cov}^2\left(\mathbf{x_j}, \mathbf{c_k}\right) + \sum_{j=1}^{p}(1 - \sum_{k} \delta_{kj})\ \rho^2\ \mathrm{var}\left(\mathbf{x_j}\right) \tag{5}$$

and for local groups, a new criterion S:

$$S_{new} = \sum_{k=1}^{K}\sum_{j=1}^{p} \delta_{kj}\ \mathrm{cov}\left(\mathbf{x_j}, \mathbf{c_k}\right) + \sum_{j=1}^{p}(1 - \sum_{k} \delta_{kj})\ \rho\ \sqrt{\mathrm{var}\left(\mathbf{x_j}\right)} \tag{6}$$

with $\mathrm{var}(\mathbf{c_k}) = 1$. In (5) and (6), the last term holds for the contribution of the additional "noise cluster" in the criterion. Namely, if the variable $\mathbf{x_j}$ belongs to one of the main groups G_k $(k = 1, \ldots, K)$, we have $\delta_{kj} = 1$ for a given k and $(1 - \sum_k \delta_{kj}) = 0$. On the contrary, if the variable $\mathbf{x_j}$ does not belong to any main groups, $(1 - \sum_k \delta_{kj}) = 1$. The number of variables that will be assigned to the "noise cluster" depends on the parameter ρ. If ρ is chosen to be very small, most variables will be assigned to one of the groups G_k rather than to the "noise cluster". If ρ is large, then a high number of variables will be allocated to the "noise cluster". ρ is a tuning parameter to be chosen between 0 and 1. Choosing $\rho = 0$ leads to the basic CLV criteria (eqs.(1) and (2)). When $\rho = 1$, all variables will be assigned to the "noise cluster".

The criteria T_{new} and S_{new} can be maximized by using the same type of algorithm as that described in Sect.2.1, except with a modification in the assignment step, which is constructed as follows:
Assignment step. A variable $\mathbf{x_j}$ will be assigned to cluster G_k in which the (squared) covariance between $\mathbf{x_j}$ and $\mathbf{c_k}$ is greater than with other latent variables, except if this value is too small in relation to the value of the ρ parameter. More formally, we have:

- for directional groups,

$$\begin{cases} \delta_{kj} = 0\ \forall k & \text{if} \max\left\{\max_l\left\{\mathrm{cov}^2\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\}, \rho^2\mathrm{var}(\mathbf{x_j})\right\} = \rho^2\mathrm{var}(\mathbf{x_j}) \\ \delta_{kj} = 1 & \text{if} \max\left\{\max_l\left\{\mathrm{cov}^2\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\}, \rho^2\mathrm{var}(\mathbf{x_j})\right\} = \mathrm{cov}^2\left(\mathbf{x_j}, \mathbf{c_k}\right) \end{cases}$$

- for local groups,

$$\begin{cases} \delta_{kj} = 0\ \forall k & \text{if} \max\left\{\max_l\left\{\mathrm{cov}\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\}, \rho\sqrt{\mathrm{var}(\mathbf{x_j})}\right\} = \rho\sqrt{\mathrm{var}(\mathbf{x_j})} \\ \delta_{kj} = 1 & \text{if} \max\left\{\max_l\left\{\mathrm{cov}\left(\mathbf{x_j}, \mathbf{c_l}\right)\right\}, \rho\sqrt{\mathrm{var}(\mathbf{x_j})}\right\} = \mathrm{cov}\left(\mathbf{x_j}, \mathbf{c_k}\right) \end{cases}$$

By the construction of criteria T_{new} (5) and S_{new} (6), the tuning parameter ρ is analogous to a correlation coefficient. As it will be shown in the illustrative section (sub-sections 3.1 and 3.2), various values of ρ, between 0 and 1, can be investigated. The curve of the evolution of the number of variables that are set aside as a function

of the correlation threshold can provide a guide to the choice of a specific value for ρ. Without screening a range of values for ρ, cutoff values for the item's loadings (or correlation coefficient between a manifest variable and a latent variable) commonly used in Exploratory Factor Analysis may be considered. Even if this question remains controversial, setting the cutoff at 0.4 seems to be the lowest acceptable threshold, whereas 0.6 or 0.7 would be the upper limit (Matsunaga, 2010). Taking account the statistical significance of the item's loadings according to the sample size, Hair et al. (2010) published a table (Chapter 3, Table 2 in the reference) showing, for instance, that a loading of 0.4 could be considered as significant ($\alpha = 5\%$) for a minimum number of 200 observations.

2.3 "Sparse LV" strategy

Let us consider the CLV clustering with the criterion in (1) for directional groups. For a partition of K clusters, the assumption is that the atypical or noise variables will be randomly assigned to one of the main groups G_k $(k = 1, \ldots, K)$. However, it is expected that their contribution to the latent variable $\mathbf{c_k}$ (the first principal component of $\mathbf{X_k}$ in this case) will be small. Methods such as the Sparse PCA (SPCA) (Zou et al., 2006), SCoTLASS (Jolliffe et al., 2003) and sPCA-rSVD (Shen and Huang, 2008) have been proposed in order to improve the interpretation of principal components (PCs) by using modified PCs with possible zero loadings producing sparse loading vectors. The same rationale can be adopted for the CLV latent variables, which can be modified so that zero loadings will be given to the variables that do not fit to the main groups. We denote $\mathbf{c_k^*}$, $(k = 1, \ldots, K)$ the sparse latent variables. They are defined, according to the iterative algorithm proposed by Shen and Huang (2008), as follows:

1. For a group G_k $(k = 1, \ldots, K)$, which contains p_k variables, let $\boldsymbol{\alpha_k}$ be the loading vector associated with the first principal component $\mathbf{c_k}$ of $\mathbf{X_k}$.

2. Calculate a sparse loading vector $\boldsymbol{\beta_k} = (\beta_{1k}, \ldots, \beta_{jk}, \ldots, \beta_{p_k k})^T$. For each variable $\mathbf{x_j}$ belonging to this group, its loading value β_{jk} is defined as

$$\beta_{jk} = \left(|\mathrm{cov}\,(\mathbf{x_j}, \mathbf{c_k})| - \rho \sqrt{\mathrm{var}\,(\mathbf{x_j})} \right)_+ \mathrm{Sign}\,(\mathrm{cov}\,(\mathbf{x_j}, \mathbf{c_k})) \tag{7}$$

   This corresponds to the soft-thresholding (or Lasso) penalty (Tibshirani, 1996; Shen and Huang, 2008). In the expression in (7), the symbol $(\,)_+$ means that the value of the difference is retained if it is positive, it is equal to 0 otherwise. More precisely, we have:

$$b = (|a| - \Delta)_+ \mathrm{Sign}(a) = \begin{cases} a - \Delta & \text{if} \quad a \geq \Delta \\ a + \Delta & \text{if} \quad a \leq -\Delta \\ 0 & \text{if} \quad |a| < \Delta \end{cases}$$

3. Update the loading vector $\boldsymbol{\alpha_k}$ in order to maximize ${\boldsymbol{\alpha_k}}^t{\mathbf{X_k}}^t\mathbf{X_k}\boldsymbol{\beta_k}$, subject to the constraint ${\boldsymbol{\alpha_k}}^t\boldsymbol{\alpha_k} = 1$. In other words, set $\boldsymbol{\alpha_k} = {\mathbf{X_k}}^t\mathbf{X_k}\boldsymbol{\beta_k}/\|{\mathbf{X_k}}^t\mathbf{X_k}\boldsymbol{\beta_k}\|$. Update $\mathbf{c_k} = \mathbf{X_k}\boldsymbol{\alpha_k}/\|\mathbf{X_k}\boldsymbol{\alpha_k}\|$.

4. Repeat steps 2 and 3 until convergence of the modified loadings vector $\boldsymbol{\beta_k}$. The convergence criterion is defined here on the L1-norm of the difference of two successive normalized loadings vectors. As previously, the maximum number of iterations allowed is fixed to 20 (by default). In our experience, this level is very rarely reached. For instance, in the case study discussed in Section 3.2, whatever the value of ρ, the highest number of iterations required before the stabilization of the sparse loadings vector, in a group of variables, was eight.

5. The sparse loading vector and modified latent variable in group G_k are $\mathbf{v_k} = \boldsymbol{\beta_k}/\|\boldsymbol{\beta_k}\|$ and $\mathbf{c_k^*} = \mathbf{X_k}\mathbf{v_k}/\|\mathbf{X_k}\mathbf{v_k}\|$, respectively.

For local groups, a similar strategy is adopted but with CLV latent variables being the mean variable of the measured variables in a group. The algorithm for defining sparse LV is constructed as follows :

1. For group G_k $(k = 1, \ldots, K)$, let $\boldsymbol{\alpha_k} = \left(\frac{1}{p_k}, \frac{1}{p_k}, \ldots, \frac{1}{p_k}\right)^T$ be the loading vector associated with the standardized mean variable $\mathbf{c_k}$ in G_k.

2. Calculate a sparse loading vector $\boldsymbol{\beta_k}$ with (7), in the same way as for directional groups.

3. For a fixed vector $\boldsymbol{\beta_k}$, update $\boldsymbol{\alpha_k} = \boldsymbol{\beta_k}/(\sum_j \beta_{jk})$ and $\mathbf{c_k} = \mathbf{X_k}\boldsymbol{\alpha_k}/\|\mathbf{X_k}\boldsymbol{\alpha_k}\|$.

4. Repeat steps 2 and 3 until convergence of $\boldsymbol{\beta_k}$.

5. The sparse loading vector and the modified latent variable in group G_k are $\mathbf{v_k} = \boldsymbol{\beta_k}/(\sum_j \beta_{jk})$, and $\mathbf{c_k^*} = \mathbf{X_k}\mathbf{v_k}/\|\mathbf{X_k}\mathbf{v_k}\|$, respectively.

It can be noticed that the thresholding parameter used for the "Sparse LV" strategy is the same as the assignment thresholding adopted for the "$K+1$" strategy. As a matter of fact, this tuning parameter, ρ, is a threshold value of a correlation coefficient. Its value is to be chosen between 0 and 1.

From a practical point of view, an R package, named `ClustVarLV` (Vigneau and Chen, 2015), for the clustering of variables using the CLV method is available on the CRAN website. The new developments presented here have been implemented in the `CLV_kmeans` procedure.

3 Applications

3.1 Simulated dataset

The purpose of working with a simulated dataset was to evaluate the ability of both strategies to identify variables generated completely at random, in addition to a set

of variables organized into K groups. Four groups of variables, of different size, were generated around prototype variables more or less correlated to each others.

We simulated 100 datasets having p=85 variables and n=50 observations. They contained four groups of 30, 20, 10 and 5 variables, respectively, and 20 variables generated independently. Four prototype variables, $\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{Z_3}, \mathbf{Z_4}$, were generated from a multivariate Gaussian distribution with a mean vector $\mu = (0, 0, 0, 0)$ and a specified covariance matrix Σ:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.3 & 0.2 \\ 0 & 0.3 & 1 & 0.1 \\ 0 & 0.2 & 0.1 & 1 \end{pmatrix} \tag{8}$$

The variables in each group and the random variables were generated as follows:

$$\begin{aligned} \mathbf{x_j} &= \omega_j \mathbf{Z_1} + \boldsymbol{\epsilon_j}, \quad j = 1, \ldots, 30, \\ \mathbf{x_j} &= \omega_j \mathbf{Z_2} + \boldsymbol{\epsilon_j}, \quad j = 31, \ldots, 50, \\ \mathbf{x_j} &= \omega_j \mathbf{Z_3} + \boldsymbol{\epsilon_j}, \quad j = 51, \ldots, 60, \\ \mathbf{x_j} &= \omega_j \mathbf{Z_4} + \boldsymbol{\epsilon_j}, \quad j = 61, \ldots, 65, \\ \mathbf{x_j} &\sim \mathcal{N}(0, 1), \qquad j = 66, \ldots, 85 \end{aligned}$$

where $\omega_j \in \{+1, -1\}$ is used to generate randomly positive and negative correlations between each simulated variable and its prototype. $\boldsymbol{\epsilon_j} \sim \mathcal{N}(0, \sigma)$ with σ set at 0.8.

$K = 4$	CLV				CLV "$K+1$"					CLV "Sparse LV"			
$\rho = 0.5$	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_A	C_1	C_2	C_3	C_4
G_1	30	0	0	0	30	0	0	0	0	30	0	0	0
G_2	0	20	0	0	0	20	0	0	0	0	20	0	0
G_3	0	0	10	0	0	0	10	0	0	0	0	10	0
G_4	0	0	0	5	0	0	0	5	0	0	0	0	5
isolated	4.77	4.44	4.97	5.82	0.01	0.03	0.03	0.09	19.84	0.01(5.12)	0.03(4.74)	0.02(5.06)	0.06(4.96)

Table 1: Cross-tabulation (averaged over 100 repetitions) displaying the memberships of the variables to their true groups (G_1, G_2, G_3, G_4 and isolated variables) and the clusters obtained for a partition into K=4 clusters, according to the strategy used. For the "$K+1$" strategy, C_A represents the additional "noise cluster". For the "Sparse LV" strategy, the numbers in parentheses correspond to the number of variables with a zero loading. For both proposed strategies ρ was set to 0.5.

Table 1 summarizes the clustering results obtained when looking for $K = 4$ directional groups, using the basic CLV algorithm (see Section 2.1), the "K+1" strategy (see Section 2.2), or the "Sparse LV" strategy (see Section 2.3) with $\rho = 0.5$. The true groups were denoted G_1 to G_4 and the clusters of variables obtained were denoted C_1 to C_4.

If the number of clusters was correctly chosen, *i.e.* $K = 4$, (Table 1), we observed that the basic CLV method always retrieved the structure based on the four groups of variables. This was expected because the simulated pattern was very simple and clear.

However, the four clusters also contained more or less the same number of isolated variables. In fact, these variables were randomly distributed across the four clusters. Both alternative strategies also retrieved the true underlying groups. Moreover, the isolated variables were satisfactorily identified. They were assigned to the additional cluster, C_A, with the "$K+1$" strategy, or their loading was set to zero with the "Sparse LV" strategy (numbers given in parentheses).

$K=3$	CLV			CLV "$K+1$"				CLV "Sparse LV"		
$\rho=0.5$	C_1	C_2	C_3	C_1	C_2	C_3	C_A	C_1	C_2	C_3
G_1	30	0	0	30	0	0	0	30	0	0
G_2	0	20	0	0	20	0	0	0	20	0
G_3	0	0	10	0	0	10	0	0	0	10
G_4	0.42	3.41	1.17	0	0	0	5	(0.69)	(3.17)	(1.14)
isolated	6.76	6.34	6.90	0.01	0.02	0.03	19.94	0.01(7.05)	0.02(6.29)	0.02(6.61)

$K=5$	CLV					CLV "$K+1$"						CLV "Sparse LV"				
$\rho=0.5$	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	C_A	C_1	C_2	C_3	C_4	C_5
G_1	30	0	0	0	0	30	0	0	0	0	0	30	0	0	0	0
G_2	0	20	0	0	0	0	20	0	0	0	0	0	20	0	0	0
G_3	0	0	10	0	0	0	0	10	0	0	0	0	0	10	0	0
G_4	0	0	0	5	0	0	0	0	5	0	0	0	0	0	2	0
isolated	2.7	2.34	2.7	3.04	9.22	0.01	0.03	0.03	1.2	3.53	16.58	0.01(3.43)	0.03(3.29)	0.02(3.56)	0.06(3.19)	2.27(4.14)

Table 2: Cross-tabulation (averaged over 100 repetitions) displaying the memberships of the variables to their true groups and the clusters obtained when K=3 or K=5 is sought ($\rho = 0.5$).

Table 2 shows the partitions obtained when the number of clusters was incorrectly chosen, namely when K was set to 3 or 5. With $K = 3$ (top part of Table 2), using the basic CLV approach, the variables of the smallest group, *i.e.* G_4, were aggregated into the other three clusters, even though the covariances between $\mathbf{Z}_4$ and the other prototype variables were small (maximum 0.3). In fact, the variables generated around $\mathbf{Z}_4$ were treated similarly to the isolated variables: they were randomly allocated to one of the three clusters. In contrast, the "$K + 1$" strategy gave an interesting and more reasonable result: the existence of the small group of variables, G_4, was not identified but the other groups were correctly highlighted. With the "Sparse LV" strategy, those variables that came from G_4 obtained a zero loading on the sparse latent variables of their current cluster. When the number of clusters K was chosen as 5 instead of 4 (bottom part of Table 2), the basic CLV approach led to a fifth cluster selecting some of the isolated variables (on average, 9.22 out of 20). The true groups, G_1 to G_4, were perfectly retrieved and the rest of the isolated variables were distributed at random in the clusters C_1 to C_4. For the "$K + 1$" and "Sparse LV" strategies, a fifth cluster had a relatively small size (on average 3.53 variables or 2.27 with a non-null loading). This change from a large cluster, using the basic CLV method, to a very small cluster, when the refining strategies were used, could be an indication of the irrelevance of a cluster.

Regarding the impact of the tuning parameter ρ, three criteria were computed: (i) the proportion of True Positives (TP), *i.e.* the correct assignments for variables belonging to one of the four true groups; (ii) the proportion of True Negatives (TN), *i.e.* the correct assignments, or loading, for the isolated variables; (iii) the percentage of all the variables assigned to the "noise cluster" C_A, or having a zero loading. Figure 2 shows

the performance of both strategies with different values of parameter ρ, from 0 to 1, by steps of 0.01. The results were almost the same whichever strategy was used. As expected, the percentage of the variables set aside, or with a zero loading (third criterion) increased with ρ. Two stages can be identified in this simulation setting. The first stage, in the interval $[0.35, 0.7]$, corresponds to cases in which all, or almost all, the isolated variables were put aside. Within this interval, the solutions were almost perfect. The second stage, reached for large value of $\rho > 0.9$, corresponds to situations where all the variables were treated as noise. If the number of clusters was incorrectly chosen, the number of errors increased (results not shown here), with a slight decrease in sensitivity when the number of clusters, K, was too small and, on the contrary, a greater decrease in specificity when K was too large. More precisely, when $K = 2$ or 3, the number of variables wrongly assigned to the "noise cluster" may be large (compared with the case of $K = 4$), even for relatively small values of ρ (0.2-0.6), because some of the variables belonging to a true group are put aside, especially those coming from a small group. On the contrary, when $K = 5$ or 6, the isolated variables are not all assigned to the "noise cluster", even for relatively large values of ρ (0.5-0.8) because they may lead to small artificial clusters.

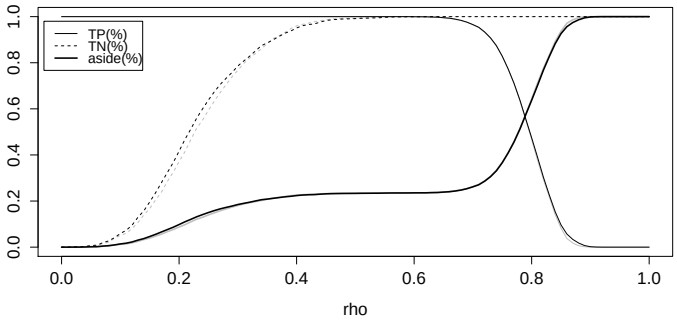

Figure 2: Change in percentage of TP, TN and variables set aside, according to the tuning parameter ρ. The results of the "K+1" strategy are given in black, and those for the "Sparse LV" strategy in gray.

3.2 Survey of food perception, and nutritional and health status in the elderly.

We now consider a real dataset collected during a French cross-sectional survey that aimed to characterize elderly people with variables relative to nutrition and health, sensory, psychological and sociological factors, eating behavior, food preferences and attitudes. A large, multidisciplinary, questionnaire was submitted to 559 participants older than 65 years (65-99 years old, 387 women, 172 men) (Maitre, 2014).

One of the objectives was to identify how the variables pertaining to different domains

could be associated, so that underlying latent dimensions would be highlighted. Herein, we consider 48 quantitative variables. Some of them are individual items, others are constructs already published or specifically developed in this survey. The exhaustive list of the variables is provided in the Appendix. In what follows, all the variables were standardized, to have the same unit variance.

First, the CLV method, for directional groups, was applied. The `CLV()` procedure, available in the `ClustVarLV` package (Vigneau and Chen, 2015), starts with a hierarchical clustering. The dendrogram obtained is shown on the left-hand side of Figure 3. Within the `CLV()` procedure, this dendrogram is cut at different levels, from a partition into one cluster to a partition into 20 clusters (by default), and the optimization algorithm described in section 2.1 is performed. The graph of the variation of the clustering criterion T (Eq.1) is shown on the right-hand side of Figure 3. This graph is helpful for choosing an appropriate number of clusters. As the clustering criterion jumps when passing from two to one cluster, a partition into two clusters could be considered. At a more detailed level, a change could also be observed when passing from four to three clusters, but not from three to two clusters. Herein the solution of four clusters was retained. It should be noted that in this case study, as in many others, *e.g.* in consumer preference studies (Vigneau et al., 2014) or for gene expression data (Bryan, 2004), natural clusters do not exist and classic measures of clustering strength and validity, like silhouette (Rousseeuw, 1987) or gap statistics (Tibshirani et al., 2001) are of little use. As mentioned in Bryan (2004), "in the absence of isolation, the exercise of grouping is variously referred to as dissection or segmentation and is considerably more difficult to define and motivate statistically than the natural clustering".

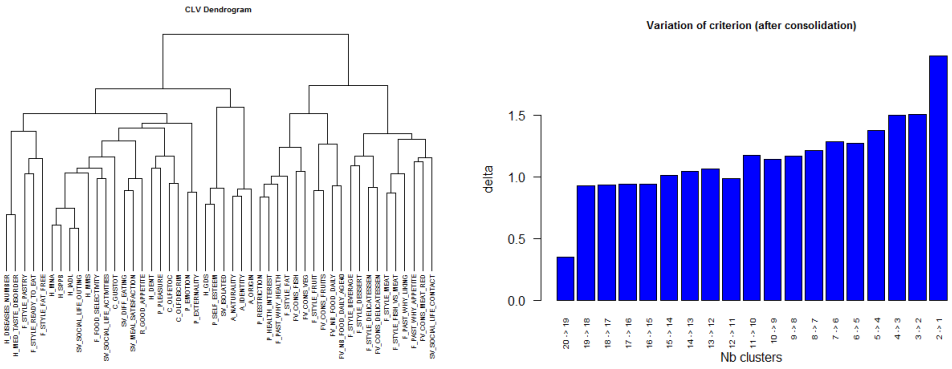

Figure 3: Graphs obtained by the clustering into directional groups of questionnaire variables. On the left side, the dendrogram of the hierarchical clustering stage; on the right side, the variation of the clustering criterion after consolidation of the partitions by means of the partitioning algorithm.

- The first cluster is made of 22 variables mainly related to health and mobility indicators. Most of the variables related to health status have large correlation coefficients with the latent variable of this cluster (larger than 0.7). However,

we can observe that some variables have been assigned to this cluster even if their relationship with the latent variable of this cluster is weak (some correlation coefficients with the first cluster latent variable are of 0.18 or -0.12).
- The second cluster groups 6 indicators including the three variables of the Authenticity scale, the Geriatric Depression Scale (H_GDS) and the feeling of being isolated (SV_ISOLATED).
- Eleven variables are in the third cluster, which is mainly characterized, positively, by the psychological sub-scales for health interest, restricted food behavior and high frequencies of consumption of fruit, fish and vegetables.
- The fourth cluster includes 9 variables related to food preferences and habits, characterizing an eater style with meat, delicatessen, accompanied with a glass of wine (here, French elderly people were concerned).

Compared with the partition into two clusters, it was observed that the first and second clusters, related to physical and psychological health, respectively, were aggregated. The third and fourth clusters, related to food behavior and food preferences, respectively, were merged into the same cluster in the partition into two clusters.

The proposed strategies of CLV clustering, *i.e.* the "K+1" strategy and the "Sparse LV" strategy, were also applied to this dataset. These are part of an alternated optimization algorithm (available using the `CLV_kmeans()` function of the `ClustVarLV` package) which requires an intialization step. For both strategies, as for the classic CLV procedure, the initial partition was defined by cutting the dendrogram previously obtained (shown in Figure 3). Various values of the tuning parameter, ρ, from 0 to 1 by steps of 0.05, were investigated. Figure 4 shows the number of variables set aside with the "K+1" strategy and the number of variables having a zero loading with the "Sparse LV" strategy, as a function of the parameter ρ.

Concerning the "K+1" strategy, if the value of ρ was greater than the largest correlation coefficient, in absolute value, between a variable within a cluster and the associated latent variable, then the involved cluster was emptied. This was the case for the fourth cluster with $\rho = 0.65$. Indeed, in the initial partition, the largest correlation coefficient with the latent variable of this cluster was 0.646. If $\rho = 0.80$, all the clusters were emptied. From a practical point of view, choosing a value of ρ greater than 0.65 does not seem to be a sound choice (87.5% of the variables were put aside and interpretable latent dimensions were lost). On the contrary, if $\rho = 0.20$, the solution was nearly the same as with the classic CLV procedure. By considering $\rho = 0.4$, one third of the variables was discarded overall. The first cluster is made of 12 items *i.e.* 55% of the variables that were in this cluster with the classic CLV solution; the second cluster remains the same with 6 items; the third cluster has 8 items *i.e.* 73% of the variables remain; the fourth cluster has 6 items *i.e.* 67% of the previous ones. The first part of Table 3 indicates the memberships of the variables as well as their correlation coefficient with the associated latent variable (LV^{K+1}). Among the variables that were set aside, the dental status (H_DENT) was a binary indicator (complete set of teeth *vs* missing teeth) which was probably not precise enough. The chemosensory ability variables (C_OLFETOC, C_OLFDISCRIM, C_GUSTOT) were not retained. It would appear that, even if the decline in

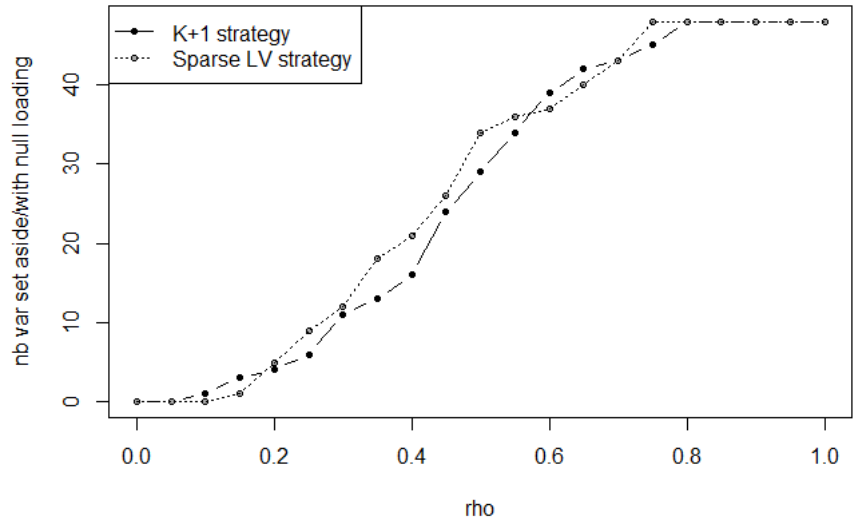

Figure 4: Number of variables set aside with the "K+1" strategy or having a zero loading with the "Sparse LV" strategy, as a function of the tuning parameter ρ.

sensory capacities varies among the elderly (Sulmont-Rossé et al., 2015), the relationships between the measured sensory abilities and the latent dimensions exhibited herein were rather weak. Among the psychological dimensions considered, related to eating behavior and attitude, the external eating and emotional eating behavior scales seemed to be less useful that the restriction eating behavior scale and the health interest scale. Finally, it may be pointed out that not all the proposed eater style questions would be retained in further studies.

When the "Sparse LV" strategy was used, the number of variables that obtained a zero loading regarding the latent variables of the clusters is given in Figure 4. Compared with the "K+1" strategy, the degree of sparsity is larger with the "Sparse LV" strategy for intermediate values of the tuning parameter ρ. As previously, we considered a value of 0.4 for ρ. The correlation coefficients of the variables with a non-null loading regarding the sparse CLV latent variables (LV^{sp}) are given in the second part of Table 3. In a given column, the empty cells correspond to variables which were assigned either to another cluster or to the considered cluster but obtained a zero loading. One of the main differences with the solution from the "K+1" strategy concerns the second cluster in which the Authenticity scales have a zero loading. Thus, the correlation coefficient between the CLV latent variables associated with the second cluster ($LV2^{K+1}$ and $LV2^{sp}$) is only 0.81. This difference between the results of both strategies, for intermediate levels of ρ, could be explained by the iterative soft thresholding step involved in the CLV with the "Sparse LV" algorithm (see section 2.3, Eq. 7). In this case study, it can be observed

	CLV "K+1" ($\rho = 0.4$)				CLV "SparseLV" ($\rho = 0.4$)				Sparse PCA (17,5,4,3)				PCA				PCA, rotation Promax			
	$LV1^{K+1}$	$LV2^{K+1}$	$LV3^{K+1}$	$LV4^{K+1}$	$LV1^{sp}$	$LV2^{sp}$	$LV3^{sp}$	$LV4^{sp}$	sPC1	sPC2	sPC3	sPC4	PC1	PC2	PC3	PC4	RC1	RC2	RC3	RC4
H_DISEASE_NB	-0.53				-0.47				-0.49				0.44	*-0.15*	*0.13*	0.42	*-0.72*	*0.31*	*0.02*	*-0.10*
H_MNA	0.80				0.81				0.84				0.73	*0.30*	*-0.03*	*-0.10*	0.71	*0.10*	*0.18*	*0.16*
H_SPPB	0.80				0.85				0.82				0.71	*0.14*	*0.02*	*-0.33*	0.83	*-0.08*	*-0.02*	*0.01*
H_IADL	0.73				0.79				0.73				0.68	*-0.12*	*0.07*	*-0.33*	0.72	*0.01*	*-0.25*	*-0.11*
H_DENT													*-0.32*	*-0.14*	*-0.16*	*0.02*	*-0.29*	*-0.10*	*-0.13*	*0.08*
H_GDS		0.71				0.86			-0.62				-0.59	*-0.21*	*0.34*	*-0.10*	-0.42	*-0.22*	*-0.09*	-0.46
H_MMS	0.44				0.39				0.40				0.45	*0.04*	*0.14*	*-0.03*	*0.35*	*0.18*	*0.01*	*-0.08*
H_MED_TASTE_DISORDER	-0.49				-0.43				-0.43				-0.40	*-0.09*	*0.03*	0.43	-0.68	*0.30*	*0.06*	*0.02*
A_NATURALITY		-0.56											*0.19*	*-0.06*	*-0.31*	*0.26*	*-0.09*	*0.30*	*-0.09*	*-0.36*
A_IDENTITY		-0.60							0.35				0.43	*-0.05*	*-0.27*	*0.16*	*0.16*	*0.31*	*-0.12*	*0.33*
A_ORIGIN		-0.58							0.29				*0.37*	*-0.06*	*-0.29*	*0.04*	*0.22*	*0.16*	*-0.15*	*0.30*
C_OLFETOC									0.34				0.40	*0.02*	*0.00*	*-0.02*	*0.31*	*0.14*	*-0.04*	*0.04*
C_OLFDISCRIM													*0.19*	*0.02*	*-0.04*	*0.06*	*0.09*	*0.12*	*0.00*	*0.08*
C_GUSTOT													*0.21*	*0.07*	*0.04*	*-0.24*	*0.37*	*-0.17*	*0.01*	*-0.06*
P_RESTRICTION			0.66				0.80					0.74	0.49	*-0.29*	*0.20*	*0.20*	*0.08*	0.55	*-0.25*	*-0.16*
P_EMOTION											0.77		*0.30*	*-0.17*	0.54	*-0.13*	*0.26*	*0.15*	*-0.12*	-0.55
P_EXTERNALITY											0.57		*0.21*	*0.06*	0.54	*0.02*	*0.13*	*0.19*	*0.15*	-0.46
P_PLEASURE													*0.30*	*0.29*	*0.16*	*0.28*	*0.06*	*0.35*	*0.34*	*0.03*
P_SELF_ESTEEM		-0.53				-0.72					-0.64		*0.19*	*0.34*	-0.46	*0.18*	*0.11*	*0.06*	*0.34*	0.59
P_HEALTH_INTEREST			0.72				0.75					0.91	0.43	*-0.25*	*-0.08*	0.49	*-0.19*	0.75	*-0.20*	*0.18*
F_STYLE_BEVERAGE				0.45						0.25			*-0.01*	0.50	*-0.12*	*-0.05*	*0.20*	*-0.24*	0.44	*0.24*
F_STYLE_MEAT				0.66				-0.81		0.80			*-0.02*	0.54	*0.12*	*0.16*	*0.02*	*0.00*	0.58	*0.08*
F_STYLE_FRUIT			0.45										*0.32*	*-0.20*	*0.06*	*0.15*	*0.04*	*0.36*	*-0.19*	*-0.04*
F_STYLE_FAT			-0.55				-0.49						*-0.35*	*0.29*	*0.04*	*-0.20*	*0.01*	-0.44	*0.28*	*-0.05*
F_STYLE_PASTRY													*-0.14*	*0.19*	*0.26*	*0.12*	*-0.15*	*0.05*	*0.27*	*-0.17*
F_STYLE_FAT_FREE													*0.00*	*-0.08*	*0.10*	*0.34*	*-0.31*	*0.38*	*0.02*	*-0.02*
F_STYLE_READY_TO_EAT													*-0.26*	*0.20*	*0.13*	*0.17*	*-0.27*	*0.03*	*0.29*	*-0.04*
F_STYLE_FISH_VS_MEAT				-0.60				0.79		-0.66			*0.17*	-0.42	*-0.13*	*0.04*	*-0.04*	*0.22*	-0.44	*0.04*
F_STYLE_DESSERT													*0.10*	*-0.17*	*0.21*	*0.02*	*-0.01*	*0.16*	*-0.13*	*-0.23*
F_STYLE_DELICATESSEN				0.61				-0.29		0.55			*-0.15*	0.56	*0.14*	*0.10*	*-0.01*	*-0.12*	0.60	*0.03*
F_PAST_WHY_APPETITE													*-0.23*	*0.08*	*-0.14*	*-0.04*	*-0.10*	*-0.19*	*0.06*	*0.12*
F_PAST_WHY_HEALTH			0.57				0.64		0.32			0.42	0.44	*-0.20*	*0.10*	*0.10*	*0.17*	*0.37*	*-0.20*	*-0.07*
F_PAST_WHY_LIKING				-0.56				0.53		-0.56			*0.20*	*-0.39*	*-0.12*	*-0.15*	*0.14*	*0.04*	*-0.45*	*-0.01*
FV_FOOD_SELECTIVITY	-0.46				-0.38				-0.40				-0.44	*-0.14*	*-0.17*	*-0.17*	*-0.21*	*-0.34*	*-0.16*	*0.02*
FV_CONS_MEAT_RED													*0.00*	*0.25*	*-0.05*	*0.17*	*-0.07*	*0.08*	*0.27*	*0.16*
FV_CONS_DELICATESSEN				0.51			-0.49						*-0.33*	0.45	*-0.05*	*0.00*	*-0.08*	*-0.30*	0.46	*0.13*
FV_CONS_FISH			0.41										*0.19*	*-0.16*	*-0.07*	*0.35*	*-0.21*	0.47	*-0.12*	*0.13*
FV_CONS_FRUITS			0.43										*0.32*	*-0.09*	*-0.04*	*0.20*	*0.03*	*0.36*	*-0.08*	*0.11*
FV_CONS_VEG													*0.18*	*0.03*	*0.13*	*0.33*	*-0.15*	0.42	*0.11*	*0.00*
FV_NB_FOOD_DAILY													*0.15*	*0.17*	*0.32*	*0.25*	*-0.06*	*0.33*	*0.27*	*-0.16*
FV_NB_FOOD_DAILY_40			0.40										*0.29*	*0.05*	*0.18*	*0.23*	*0.02*	*0.38*	*0.10*	*-0.06*
SV_ISOLATED		0.60				0.71					0.47		*-0.30*	*-0.14*	0.49	*-0.07*	*-0.22*	*-0.06*	*-0.02*	-0.54
SV_DIFF_EATING	-0.46				-0.38				-0.37				*-0.39*	*-0.20*	*-0.08*	*-0.03*	*-0.32*	*-0.14*	*-0.17*	*-0.03*
SV_MEAL_SATISFACTION	0.51				0.45				0.48				0.50	*0.14*	*-0.17*	*0.14*	*0.28*	*0.28*	*0.07*	*0.29*
SV_SOCIAL_CONTACT													*-0.08*	*0.01*	*-0.25*	*0.16*	*-0.18*	*0.08*	*0.00*	*0.27*
SV_SOCIAL_ACTIVITIES	0.51				0.48				0.48				0.50	*-0.02*	*0.08*	*-0.09*	0.42	*0.14*	*-0.08*	*-0.04*
SV_SOCIAL_OUTING	0.78				0.83				0.82				0.73	*0.09*	*-0.02*	*-0.24*	0.75	*0.03*	*-0.05*	*0.07*
R_GOOD_APPETITE	0.45				0.35				0.39				0.42	*0.25*	*0.24*	*0.14*	*0.26*	*0.28*	*0.27*	*-0.07*

Table 3: Correlation coefficients of each variable with the components defined using various strategies: the "K+1" strategy of CLV ($K = 4$, $\rho = 0.4$), the "Sparse LV" strategy of CLV ($K = 4$, $\rho = 0.4$), the Sparse PCA ($K = 4$, # sparse coeffs.: 17, 5, 4, 3), the classic PCA ($K = 4$) and the oblique rotated components ($K = 4$, Promax). An empty cell indicates that the variable is not assigned to the cluster or has a zero loading in the component. For the two last methods, the correlation coefficients less than 0.4, in absolute value, are in italic.

that choosing a slightly lower tuning parameter for the "Sparse LV" strategy than for the "K+1" strategy led to very comparable solutions. For instance with $\rho = 0.3$, the latent variables obtained with the "Sparse LV" strategy were highly correlated with the corresponding latent variables obtained with $\rho = 0.4$ for the "K+1" strategy.

By comparison, the third part of Table 3 shows the results obtained using the sparse PCA (Zou et al., 2006) for the first fourth principal components. Defining the value for the penalty parameter for each component is a tricky problem. As an alternative the `spca()` procedure included in the R package `elasticnet` makes it possible to define the number of loadings to be kept. Herein, for each component, this was chosen as the number of variables whose correlation with the ordinary Principal Component (PC) was greater than 0.4, in absolute value (given in the fourth part of Table 3). Thus, the sparse Principal Components (sPC) have 17, 5, 4 and 3 non-null loadings, respectively. It can be noticed that the variables associated with the first sPC are mainly those of the first cluster (related to physical health and mobility), but also include some of the second cluster (related to psychological health), namely H_GDS, A_IDENTITY and H_ORIGIN but not H_NATURALITY. The other variables that were assigned to the second cluster, P_SELF_ESTEEM and SV_ISOLATED, are retrieved on the third sPC associated with eating behavior items that were not retained previously. The second sPC has similarities with the fourth cluster (related to food preferences), while the last sPC involves important variables of the third cluster (related to food behavior). Let us note that one variable, F_PAST_WHY_HEALTH (indicating a change in food consumption for health reasons), has a non-null loading for two sparse PCs. This cannot occur with the CLV approach, whatever the strategy used, because it is based on a crisp partition of the variables.

The CLV approach, as well as the sparse PCA method, provide a useful tool to define simple and interpretable latent components, unlike Principal Components. Indeed, each of them is a linear combination of all the variables. In order to exhibit simple structures better in a dataset, rotation of the Principal Components has also been suggested (Jolliffe, 2002). The last part of Table 3 gives the results of the Promax (oblique rotation) transformation of the first four PCs. By highlighting the variables that have a correlation coefficient greater than 0.4, in absolute value, with the Rotated Components (RC), it appears that there is a concordance with the variables in the clusters obtained with the "K+1" strategy of CLV. In fact, the correlation coefficients between the latent variables of the clusters and the corresponding Rotated Components vary from 0.70 to 0.96. Nevertheless, the sparsity property inherent in the CLV latent variables enables an easier interpretation of the extracted components.

4 Conclusion

We have proposed two modifications of the CLV method, a "$K + 1$" strategy and a "Sparse LV" strategy, to take into account atypical or noise variables. These variables can be pure white noise information but, within the context of clustering, this notion also applies to isolated or scattered variables, which do not fit the underlying group structure well. With the "$K + 1$" strategy, an additional cluster ("noise cluster") is defined and

is assumed to contain all these atypical variables. With the "Sparse LV" strategy, the loadings of the variables that are not well-associated with the latent variables of each cluster are set to zero. From an algorithmic point of view, both approaches correspond to a modification of the assignment step or the latent variables estimation step, respectively, of a Clustering around Latent variables (CLV) procedure. The tuning parameter used in both strategies (ρ) represents a correlation threshold and can easily be chosen between 0 and 1.

Both approaches perform well, and in a similar manner, on the basis of a simulated dataset. They also show a logical behavior when the number of clusters is incorrectly chosen. For more complex situations, like in the real case study presented here, both approaches may lead to solutions that can differ slightly for the same value of the tuning parameter. However, herein, the four clusters defined, as well their associated latent variables, highlight the same latent traits associated with physical health, psychological health, food behavior and food preferences, respectively. The Principal Components Analysis (PCA), the sparse PCA or the Rotated Principal Components also make it possible to reveal these dimensions, more or less efficiently.

The "$K + 1$" and the"Sparse LV" strategies seem to be similar, but they reflect two different facets of the clustering of variables performed with the CLV method: (i) the identification of clusters of variables as homogeneous as possible, and (ii) the reduction in the dimensionality using the cluster latent variables. According to the first facet, the "$K + 1$" strategy leads to the identification of a "noise cluster" in addition to the main clusters. This approach is more straightforward if the main objective of the user is the segmentation of the set of variables. However, if the CLV method is applied as an alternative to a factor analysis, in order to extract latent variables associated with groups of correlated variables, then the "Sparse LV" strategy is recommended.

A future research direction will be how to combine the proposed strategies with predictive modeling, to improve the prediction accuracy and enhance the model's interpretability. To date, the elimination of uninformative variables (Centner et al., 1996) and the selection of variables using the L1-penalty (Tibshirani, 1996) are the most common strategies for achieving both these goals. Both the "$K + 1$" and "Sparse LV" CLV strategies could be promising alternatives for revealing predictive and easy to interpret latent components. Another direction of investigation would be to introduce sparse clustering strategies in the context of the co-clustering (simultaneous clustering of the observations and the variables of a data set). Considering, in particular, the approach called CDPCA (Clustering and Disjoint Principal Component Analysis) (Vichi and Saporta, 2009), which includes the CLV method as particular case, robust clustering of the observations and the "Sparse LV" strategy" could be combined within the CDPCA framework.

References

Bailly, N., Maitre, I., Amand, M., Herve, C., and Alaphilippe, D. (2012). The dutch eating behaviour questionnaire (DEBQ). assessment of eating behaviour in an aging

french population. *Appetite*, 59:853–858.

Berget, I., Mevik, B.-H., Vebo, H., and Maes, T. (2005). A strategy for finding relevant clusters; with an application to microarray data. *Journal of Chemometrics*, 19:482–491.

Bryan, J. (2004). Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis*, 90:44–66.

Camus, S. (2004). Proposition d'échelle de mesure de l'authenticité perçue d'un produit alimentaire. *Recherche et Applications en Marketing*, 19:39–63.

Centner, V., Massart, D. L., de Noord, O. E., de Jong, S., Vandeginste, B. M., and Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Anal Chem*, 68(21):3851–3858.

Dahl, T. and Naes, T. (2009). Identifying outlying assessors in sensory profiling using fuzzy clustering and multi-block methodology. *Food Quality and Preference*, 20:287–294.

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664.

Dhillon, I. S., Marcotte, E. M., and Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619.

Enki, D. G., Trendafilov, N. T., and Jolliffe, I. T. (2012). A clustering approach to interpretable principal components. *Journal of Applied Statistics*, 40(3):583–599.

Folstein, M., Folstein, S., and McHugh, P. (1975). Mini-mental state. a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 13:189–198.

Guigoz, Y., Lauque, S., and Vellas, B. (2002). Identifying the elderly at risk for malnutrition: The mini nutritional assessment. *Clinics in Geriatric Medicine*, 59:737–757.

Guralnik, J., Simonsick, E., Ferrucci, L., Glynn, R., Berkman, L., Blazer, D., Scherr, P., and Wallace, R. (1994). A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *Journal of Gerontology*, 59:M85–M94.

Hair, J., Black, W., Babin, B., and Anderson, R. (2010). *Multivariate Data Analysis, 7th ed.* Prentice Hall.

Jolliffe, I. (2002). *Principal Component Analysis, 2nd ed.* Springer -Verlag, New York.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *J. Comput. Graph. Statist.*, 12:531–547.

Lawton, M. and Brody, E. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist*, 9:179–186.

Maitre, I. (2014). *Perceptions sensorielles et préférences alimentaires des seniors. Contribution au maintien du statut nutritionnel et à l'appréciation des produits.* PhD thesis, Univ. Angers (France).

Maitre, I., Van Wymelbeke, V., Amand, M., Vigneau, E., Issanchou, S., and Sulmont-Rossé, C. (2014). Food pickiness in the elderly: Relationship with dependency and

malnutrition. *Food Quality and Preference*, 32:145–151.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, 3(1):97–110.

Roininen, K., Lhteenmki, L., and Tuorila, H. (1999). Quantification of consumer attitudes to health and hedonic characteristics of foods. *Appetite*, 33:71–88.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.

Sarle, W. S. (1990). *SAS/STAT User's Guide: The Varclus Procedure. 4th Edition.* SAS Institute, Inc., Cary, NC, USA.

Sheikh, J. I. (1986). Geriatric depression scale (GDS) recent evidence and development of a shorter version. *Clinical Gerontologist*, 5:165–173.

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.

Sulmont-Rossé, C., Maitre, I., Amand, M., Symoneaux, R., and Van Wymelbeke, V. (2015). Evidence for different patterns of chemosensory alterations in the elderly population: Impact of age versus dependency. *Chemical Senses*, 40:153–164.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tibshirani, R., Guenther, W., and Hastie, T. (2001). Estimating the number of clusters in a dataset using the gap statistic. *J. Roy. Statist. Soc. Ser. B*, 63(2):411–423.

Vallieres, E. F. and Vallerand, R. J. (1990). Traduction et validation canadienne-française de l'chelle de l'estime de soi de rosenberg. *International Journal of Psychology*, 25:305–316.

Vichi, M. and Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53:3194–3208.

Vigneau, E., Charles, M., and Chen, M. (2014). External preference segmentation with additional information on consumers: A case study on apples. *Food Quality and Preference*, 32:83–92.

Vigneau, E. and Chen, M. (2015). *ClustVarLV: Clustering of variables around Latent Variables.* R package version 1.3.2.

Vigneau, E. and Qannari, E. M. (2003). Clustering of variables around latent components. *Comm. Stat. - Simul Comput.*, 32(4):1131–1150.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15:265–286.

Appendix

List of the variables in the "Survey of food perception, and nutritional and health status in the elderly" (section 3.2)

short name	name/description	reference
H_DISEASES_NB	number of diseases	
H_MNA	Mini Nutritional Assessment	(Guigoz et al., 2002)
H_SPPB	Short Physical Performance Battery	(Guralnik et al., 1994)
H_IADL	Instrumental Activities of Daily Living Scale	(Lawton and Brody, 1969)
H_DENT	dental status (0/1)	
H_GDS	Geriatric Depression Scale	(Sheikh, 1986)
H_MMS	Mini Mental Scale	(Folstein et al., 1975)
H_MED_TASTE	number of drugs inducing taste disorder	
A_NATURALITY	How much I agree that the foods I eat are natural	
A_IDENTITY	How much of the food I eat is an image of myself	Authenticity scales (Camus, 2004)
A_ORIGIN	How much I know about the origin of the foods I eat	
C_OLFETOC	olfactory identification	
C_OLFDISCRIM	olfactory discrimination	(Sulmont-Rossé et al., 2015)
C_GUSTOT	salt detection	
P_RESTRICTION	shortened French version of the Questionnaire	
P_EMOTION	Dutch Eating Behavior (DEBQ)	(Bailly et al., 2012)
P_EXTERNALITY		
P_PLEASURE	two subscales of the Health and Taste Attitude	(Roininen et al., 1999)
P_HEALTH_INTEREST	Questionnaire (HTAQ)	
P_SELF_ESTEEM	French version of the Rosenberg Self-Esteem Scale	(Vallieres and Vallerand, 1990)
F_STYLE_BEVERAGE		
F_STYLE_MEAT		
F_STYLE_FRUIT		
F_STYLE_FAT		
F_STYLE_PASTRY	Eater profiles (Likert scales)	(Maitre, 2014)
F_STYLE_FAT_FREE		
F_STYLE_READY_TO_EAT		
F_STYLE_FISH_VS_MEAT		
F_STYLE_DESSERT		
F_STYLE_DELICATESSEN		
F_PAST_WHY_APPETITE		
F_PAST_WHY_HEALTH	Changes in food consumption	
F_PAST_WHY_LIKING		
FV_FOOD_SELECTIVITY	Food selectivity	(Maitre et al., 2014)
FV_CONS_MEAT_RED		
FV_CONS_DELICATESSEN		
FV_CONS_FISH	Synthetic measures for frequency of consumption	
FV_CONS_FRUITS		
FV_CONS_VEG		
FV_NB_FOOD_DAILY	Nb food consumed daily, now	
FV_NB_FOOD_DAILY_40	at 40 years old	
SV_ISOLATED	feeling of being isolated,	
SV_DIFF_EATING	of having difficulties for eating,	
SV_MEAL_SATISFACTION	of being satisfied with their meals	
SV_SOCIAL_CONTACT		
SV_SOCIAL_ACTIVITIES		
SV_SOCIAL_OUTING	social life constructs	
R_GOOD_APPETITE	feeling of having a good appetite	