Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XX ciclo

Alma Mater Studiorum - Università di Bologna

Clustering of variables around latent components: an application in consumer science

Isabella Endrizzi

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2008

Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XX ciclo

Clustering of variables around latent components: an application in consumer science

Isabella Endrizzi

Coordinator:
Prof. Daniela Cocchi

Tutor:
Prof. Angela Montanari

Co-tutors:
Dr. Daniela G. Calò
Prof. Evelyne Vigneau
Ms. Flavia Gasperi

Dipartimento di Scienze Statistiche "P. Fortunati"

Marzo 2008

Acknowledgements

I would like to say thank you to the many people who directly and indirectly contributed to this work.

First of all, I would like to thank Angela Montanari, my supervisor, for her support and advice, and Daniela Calò, for her useful suggestions and kindness.

My period of study at the ENITIAA of Nantes was crucial for my research so I would like to thank all the people who make it possible. A special thank to Evelyne Vigneau, who guided me with her knowledge and her devotion to the research during my stay and also, for her fundamental contribution and comments on my writing.

I am very grateful to Chantal Gilbert for giving me the chance to learn and grow not only from a professional point of view during my training placement at the CCFRA, and for still being a kind and good friend.

Furthermore, I would like to express my deep gratitude to Flavia Gasperi and my colleagues of IASMA for their encouraging support in the hard moments and their precious suggestions.

Finally, I would like to thank all the people who I had close, especially my husband, my parents and my sister, for their patience, encouragement and love that sustained me in these three years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The present work aims at the development of a strategy for clustering consumers on the basis of their preference on products, be they food, beverages, fragrances or household products. The analysis of preferences has assumed nowadays an extreme importance in business: in order to obtain a competitive advantage, companies have realized that knowing consumer preferences and habits is very important as much as analysing sensory characteristics of a product. It is well known that the investigation of the existing relationships among sensory, instrumental and preference data is fundamental to develop new products, market researches and quality evaluations. But

the information on consumers' background should not be overlooked because it is valuable to target different groups of consumers.

In the latest years many scientists studied this subject trying to answer few main issues as "Are there clusters of consumers?", "How the clustering of consumers can be interpreted in terms of preferences or consumption?", "How the consumers' preferences relate to the physical/chemical measurements or to sensory attributes of the products?" and "Is there a connection between the preferences and the socio-demographic variables or the consumer habits?".

Although the strong interest that the research has in this field, explaining consumer preference and relating it to the quality of the product and the attitudes of consumers is not a straightforward work.

First of all it is very important, but difficult, to acquire good preference data on consumers and also difficult to interpret these data in the light of products and consumer characteristics. In order that the research may be useful in practice, the number of consumers has to be sufficiently large to provide statistically stable results. Furthermore, consumer panel has to be representative of the population of consumers and its response must to reflect real preference on the products. In order to verify the relevance and reliability of consumer response, one should ideally have at disposal repeated measurements in consecutive time periods. Moreover, the products tested have to be sufficiently many and different from each other to be relevant for the type of product under investigation, but not so many or so different that the respondents become confused (Thybo *et al*., 2003).

Last but not least, in preference studies the evaluated products are considered as the statistical units, while hedonic ratings of each consumers are located as columns in

the preference data table. Due to the fact that it is impossible to submit to the judges a too high number of stimuli, according to one of basic rules in sensory analysis, the number of statistical subjects is commonly extremely small in comparison with the number of the subjects. Therefore, the resulting matrices are in the shape of a flat rectangle and variables are often strongly intercorrelated and multicollinearity may be a problem.

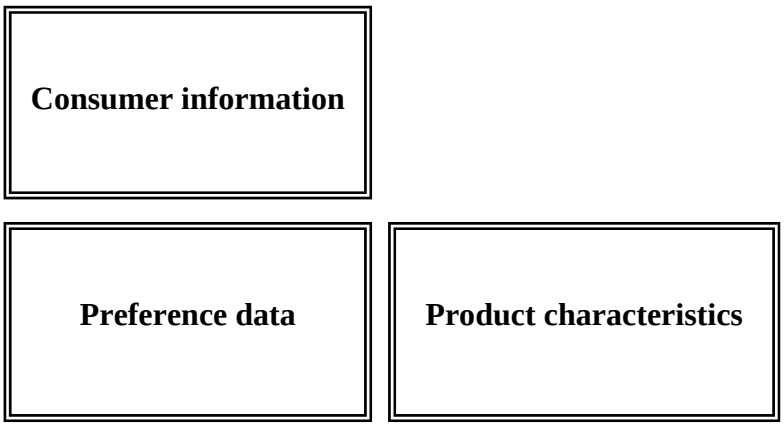
The consumer preference may be then defined and measured following the aim behind the study, depending on the scale of measurements: traditionally it takes the form of a paired preference test, preference ranking or hedonic scaling (Greenhoff and MacFie, 1994).

Once the scale of measurements and the kind of test to submit to the consumers are chosen, two alternative ways can be followed according to whether preference is supposed to be driven by product characteristics or consumer descriptors (Squillacciotti, 2004).

In this study we refer to a situation where a number of products are submitted to a consumer panel for evaluation and then a segmentation of the consumers is identified on the basis of preference data. In the first case, mentioned above, the judgements are subsequently linked to product descriptors in order to find out which product characteristics consumers prefer and which they reject. In other words, the researcher is looking for relevant descriptors to explain consumer preference that will ensure an appreciation of the product. Product characteristics may include physical aspects, chemical composition, sensory attributes as evaluated by a panel of trained judges and/or any other measurable features.

In the alternative approach, preference of consumers is supposed to be driven by individual characteristics, whose selection will depend on the goal of the study. In general consumer background includes socio-demographics information, purchase behaviour, use and consumption, habits and opinions regarding products under investigation. In this case the researcher mostly wants to identify and describe groups of consumers showing similar preference and characteristics.

Figure 1.1 – L-shape structure



Each of the approaches, briefly described above, is in many case reasonable, but often incomplete. Furthermore there is not an unique product able to satisfy all consumers in the same way: consumers have different tastes and preference and such differences can derive from several concurring causes. Then, it is very likely that preference simultaneously depends on product characteristics, on consumer background and on their interaction.

The inclusion in the study of all the cited elements (hedonic rating on a certain number of products given by a consumer panel, product characteristics and consumer information), gives a particular data structure, where three data matrices take place. This structure, called L-shape, consists in a central matrix containing preference data, on its

left a second matrix with product descriptors is situated and a third matrix containing consumer background positioned above the first one (figure 1.1). In L-structured data the three matrices have different dimensions, in fact, row and columns tables share with the central matrix just one dimension each.

The goal of the present work is to define a strategy that automatically provide a consumer segmentation based on preference data taking into account all external information about consumers and products at disposal, i.e. row and columns descriptors of central matrix simultaneously. In other words, a method where each of the three matrices has an active role in the clustering of consumers, getting homogeneous groups from all points of view: preference, products and consumer characteristics.

1.1 Work plan

This work is carried out under a special agreement between the Department of Statistics "Paolo Fortunati" of Bologna University and the Agri-food Quality Department of S. Michele all'Adige Institute (IASMA) with the aim to optimize application and study of statistical methodologies in a scientific research context in sensory analysis and consumer science area. The data used here were collected and provided by IASMA as part of multidisciplinary studies which aim at improving quality and marketability of agri-food products.

The statistical procedure proposed in this thesis has been developed with the support of Sensometrics and Chemometrics Department of ENITIAA (L'Ecole

Nationale d'Ingènieurs des Techniques des Industries Agricoles et Alimentaires) of Nantes in order to study a new consumer segmentation technique for L-structured data.

Chapter 2 will be devoted to the description of Cluster analysis around Latent Variables (CLV) (Vigneau and Qannari, 2003), a method that directly provides a segmentation of the panel of consumers and, in each segment, identifies preference directions through latent components. This approach can be extended taking into account external data and giving, for each group, a single model that links preference scores to product characteristics. A particular attention will be given to the description of all the cases included in the algorithm.

Being the aim of this work the definition of a technique for the clustering of L-shape data, Chapter 3 will begin by giving a description of methodologies born to treat this kind of data structure, with a particular attention to two-step PLS (Kubberød *et al.*, 2002; Thybo *et al.*, 2003; Lengard and Kermit, 2006; Esposito Vinzi *et al.*, 2007) which the procedures here proposed were drown inspiration from. New procedures, which assign an active role to Z data, will be presented. In particular, two step L-CLV, which aims to use CLV algorithm in the case of L-structured data taking into account all the information at disposal, will be described.

Chapter 4 will show the results from the applications of some of the described methods to data from preference studies on food products. In the first part of the chapter traditional approaches and two-way analyses will be apply to the dataset (first assuming preference is driven by products characteristics and then taking into account just consumer descriptors). In the second part two step L-CLV will be applied in order to show the different results obtained when all the three matrices have an active role in the

clustering. Some final comments are contained in the conclusion (Chapter 5) with a brief introduction to further developments.

Chapter 2

Cluster analysis of variables around latent components

2.1 Introduction

Cluster analysis is the art of finding groups in data. The classification of similar objects into groups is an important human activity. In everyday life, this is part of a learning process (a child learns to distinguish between cats and dogs, between tables and chairs, between men and women) and in science classification played an essential role (one only has to think to extensive classifications of plants, animals, etc.). In marketing, it is often attempted to identify market segments, that is, groups of consumers with similar needs (Kaufman and Rousseeuw, 1990).

The majority of classification methods are conceived to identify groups of units. Nevertheless, clustering of variables is a useful tool as soon as the observations are not the main subject of the analysis. Grouping variables may be worthwhile in many practical situations, such as preference studies, whenever a panel of consumers is asked to evaluate a certain number of products and a special attention is focussed on variables (here referred to consumers).

Often, the methods to cluster variables consist to define a matrix of distance (or dissimilarity for ordinal variables) measures (Gower, 1985; Everitt *et al.*, 1974). Many clustering techniques like Hierarchical Cluster Analysis (HCA) method can be applied on such a matrix. In articles from Qannari *et al.* (1997, 1998), Abdballah and Saporta (1998) and Soffritti (1999), procedures of clustering of variables based on a definition of similarity (or dissimilarity) measures between variables have been discussed.

Alternative methods are available to allow variable clustering without a distance matrix has been defined. The VARCLUS procedure (SAS Institute Inc., 1999) implemented in SAS software is one of them. This procedure is closely related to principal component analysis and can be used as an alternative method for eliminating redundant dimensions. This variable clustering will find groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters. If the second eigenvalue for the cluster is greater than a specific threshold, the cluster is split into different dimensions then a reassignment phase of the variables to the clusters is performed.

Another strategy of analysis was developed by Vigneau *et al.* (2001). Their approach, called CLV (Clustering around Latent Variables), consists in clustering variables around latent components, more precisely the aim is to simultaneously

determine K clusters of variables and K latent components such that the variables in each group are as correlated as possible with the corresponding latent component. The solution is given by running an iterative partitioning algorithm which involves at the first step the choice of K starting groups. In order to choose an appropriate number of groups and determine an initial partition, the authors suggest to apply a hierarchical cluster analysis. The partitioning algorithm is carried out using the cluster from the hierarchical analysis as starting point. In this way, the advantage of the hierarchical method is complemented by the non hierarchical one to improve the results by allowing the switching of cluster membership (Hair *et al.*, 1992). Furthermore, hierarchical cluster analysis has the same justification as the partition algorithm: both aim to maximize the same criterion. Determined latent variables, which are respectively associated with each group, may be considered as components of principal component analysis or of PLS-regression, with the difference that CLV components are non orthogonal but more easily interpretable. In the following paragraph this method will be described in detail.

2.2 CLV approach

In order to render the method adaptable to different contexts, two parameters are considered by the authors (Vigneau & Qannari, 2006). The first parameter is related to the kind of groups of variables that one would seek, two cases are taken into account:

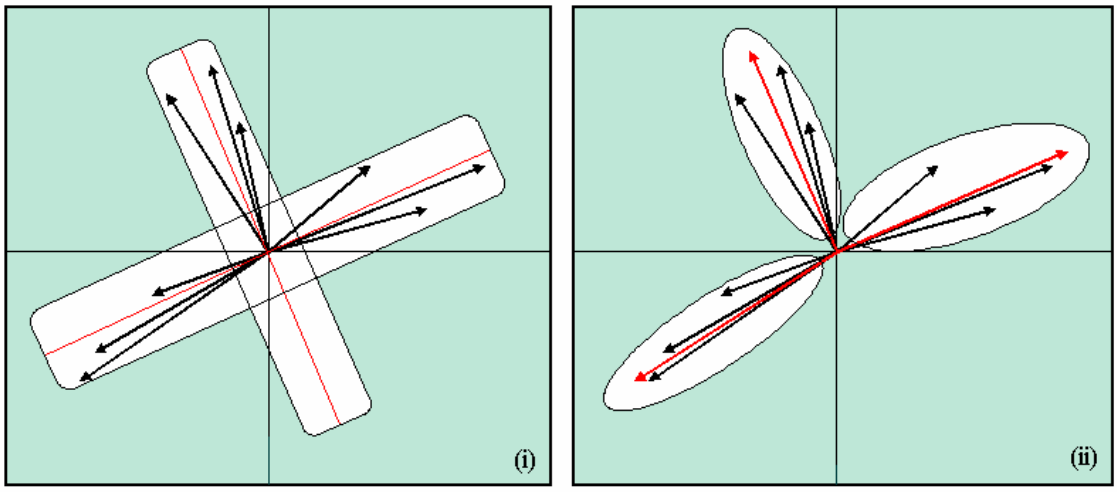
i) The first case is applied to a situation where the aim is to lump together correlated variables regardless the sign of correlation

coefficients. The CLV latent components are similar to factors after oblique rotation. In this case each group of variables is built around an axis which represent the latent component of the group;

ii) In the second case, two negatively correlated variables are considered as dissimilar and so cluster of variables is locally defined;

Both the cases described above are depicted in figure 2.1.

Figure 2.1 – Cases considered in CLV method: when positive and negative correlations imply agreement (i) and when negative correlations show disagreement (ii).



A second parameter allows to take into account in the segmentation a table with external variables linked to the rows of initial data matrix. Whenever external variables, in addition to variables which have to be clustered, are available it could be worthwhile to express the latent component of each group as a linear combination of these external information. In this case it could be possible to interpret group structure in the light of added information, giving, in each segment, a single model which relates preference scores to the characteristics of the products. This situation could be considered as an alternative to the most common method used to relate the preferences of a panel of consumers to external data: External Preference Mapping analysis (PrefMap), which

gives a separated model for each consumer (Greenhoff and MacFie, 1994). In these type of studies, the goal is to identify groups of consumers which express similar preference on a certain number of products and to explain this preference in each segment using chemical or sensory characteristics of the products.

2.2.1 Negative correlation means agreement

Consider the case where the aim is to lump together correlated variables regardless of the sign of the correlation coefficients. In order to reach this purpose the authors seek to determine K (supposed to be fixed) clusters of variables and K latent components. Consider a panel of *p* consumers and the hedonic scores, $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_p$ they have attributed to *n* products. These variables are assumed to be centred, but not necessarily standardized. Let be $G_1,G_2,\ldots,G_k$ the K clusters of variables and $\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_k$ the K latent components associated respectively with the K groups. The aim is to minimize the quantity:

$$E = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\left\|y_j - \alpha_{kj}c_k\right\|^2 \qquad (1)$$

where $\delta_{kj} = 1$ if the j^{th} variable belongs to G_k and $\delta_{kj} = 0$ otherwise, and α_{kj} is a real number that have to be determined.

For a fixed partition, developing E considering its partial derivative with respect to α_{kj}, it is easy to verify that optimal values of α_{kj} are:

$$\alpha_{kj} = \frac{c_k' y_j}{c_k' c_k}$$

Replacing this value in expression (1) we find that minimizing E it is equivalent to maximising:

$$T = n\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}Cov^2\left(y_{j,}c_k^*\right)=\frac{1}{n}\sum_{k=1}^{K}c_k^{*'}Y_kY_k^{'}c_k^* \text{ with } c_k^*=\frac{c_k}{\sqrt{c_k^{'}c_k}} \quad (2)$$

where Y_k is the matrix, the columns of which consist of the variables belonging to the group G_k.

A solution of this problem is given by an iterative algorithm in the course of which the variables are allowed to move in and out of the groups at the different stages of the algorithm achieving at each step an increase of the criterion. This algorithm runs as follows:

Step1. The algorithm starts with K groups of variables which could be determined by random or, preferably, by a hierarchical clustering method.

Step2. In the group G_k, the latent component c_k^* is defined as the first standardized eigenvalue of $Y_kY_k^{'}$, or in other words, the first standardized principal component of Y_k.

Step3. At each step of the algorithm new groups of variables are created by assigning a variable to a group whether its squared coefficient of covariance with the latent component of this group is higher than with any other latent component.

In further steps, the process starting from step 2 goes on iteratively until stability is achieved. The definition of latent components of the groups is comparable to principal components of Y, with the difference that PCA factors are two by two orthogonal. On the other hand, CLV components are easier to interpret because each component is a linear combination of the variables which belonging to the group and which have the same preference direction.

As mentioned above, the authors suggest to complete the algorithm with a hierarchical clustering based on the same criterion T (eq. 2), that has just been described, in order to obtain few benefits: this added algorithm helps to choose the number of clusters, defining an initial partition to be used in step1 of the previous algorithm. The hierarchical procedure is an agglomerative technique based on the same criterion that is used for the partitioning algorithm.

Consider the criterion written as follows:

$$T = \sum_{k=1}^{K} \lambda_1^{(k)} \tag{3}$$

where $\lambda_1^{(k)}$ is the largest eigenvalue of the matrix $\frac{1}{n} Y_k Y_k^{'}$ or equivalently of $\frac{1}{n} Y_k^{'} Y_k$ which is the matrix of covariance among variables of group G_k. This form of criterion T suggests the following hierarchical procedure:

Initially, each consumer identify a cluster and in the final step, all the consumers are merged in a unique group. T is then equal to:

$$T_0 = \sum_{j=1}^{p} \mathrm{var}\left(y_j\right)$$

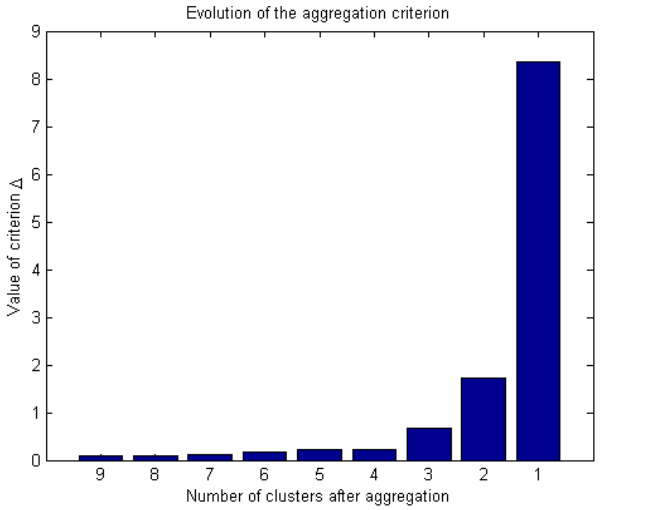
Passing from stage i to stage (i+1) means merging together two clusters of variables, A and B, that leads to a decrease of criterion T given by:

$$\Delta = T_{i-1} - T_i = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$$

where $\lambda_1^{(A)}$, $\lambda_1^{(B)}$ and $\lambda_1^{(A \cup B)}$ are the first eigenvalues associated to the matrix of covariance among variables in groups A, B and A∪B respectively. The authors veryfied that $\lambda_1^{(A \cup B)} \leq \lambda_1^{(A)} + \lambda_1^{(B)}$ which shows that the criterion decreases when two groups are merged. Therefore, the rule adopted to merge the two groups of consumers results in the

smallest decrease of the criterion, in order to preserve it as large as possible after each step.

Figure 2.2 – An example of an evolution of criterion: it turns out that when passing from five to four clusters the criterion did not change significantly, but the loss in the quality of the partition is more important when passing from four to three clusters.



The plot of the evolution criterion $\Delta=T_{i-1} - T_i$ in the course of the hierarchical algorithm is a very helpful tool in order to decide how many segments are present in the panel. If Δ jumps when passing from a solution with K segments to a solution with K-1 segments by merging two clusters, then this should be considered as an indication that no similar clusters are been merged. Therefore a partition with K clusters should be retained. An example of the plot of evolution of the aggregation criterion in the course of the hierarchy in given in figure 2.2. The partition obtained by cutting the hierarchical tree in K groups is used as initial segmentation for the partitioning algorithm.

2.2.2 Negative correlation means disagreement

In this paragraph we consider the situation where the aim of the researcher is to take into consideration both the strength and the sign of correlation between variables. For instance, suppose that *p* consumers are asked to rate their acceptability of *n* products. A negative covariance between the scores of two consumers highlights their different views of the products. The procedure, this time, is based on the individuation of K groups of variables $G_1,G_2,\ldots,G_k$ and K latent components $c_1,c_2,\ldots,c_k$ such that the quantity Q is minimized:

$$Q = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\left\|y_j - c_k\right\|^2 \qquad (4)$$

It is clear from Q criterion that the classification procedure is similar to the k-means algorithm (MacQueen, 1967). In a fixed partition, the minimum of criterion Q is expected to be $c_k = \bar{y}_k$, where $\bar{y}_k$ represents the centroid or the mean of the group G_k.

At the stage *i* of the hierarchical ascending algorithm the criterion Q will be equal to:

$$Q_i = \frac{1}{n}\sum_{k=1}^{K_i}\sum_{j=1}^{p}\delta_{kj}\left\|y_j\right\|^2 - \frac{1}{n}\sum_{k=1}^{K_i}\sum_{j=1}^{p}\delta_{kj}\left\|\bar{y}_k\right\|^2 = \sum_{j=1}^{p}p_k Var(y_j) - \sum_{k=1}^{K_i}p_k Var(\bar{y}_k) \qquad (5)$$

where p_k is the number of variable belonging to group G_k.

Calling $W_i = \sum_{k=1}^{K_i}p_k Var(\bar{y}_k)$, it can be demonstrated that if two groups, A and B, were lump together from stage *i* to (*i+1*), the criterion Q increase, and consequently W decrease of Δ:

$$\Delta = W_i - W_{i+1}$$

$$= p_A Var(\bar{y}_A) + p_B Var(\bar{y}_B) - (p_A + p_B)Var(\bar{y}_{A\cup B})$$

$$= \frac{p_A p_B}{p_A + p_B} Var(\bar{y}_A - \bar{y}_B)$$

where $\bar{y}_A$, $\bar{y}_B$, $\bar{y}_{A \cup B}$ represent the centroids of groups A, B and A∪B respectively and p_A (p_B) is the number of variables collected in the group G_A (G_B). Here as well the merging of two groups leads to a decrease of criterion; the rule adopted to merge two groups of consumers A and B will result in the smallest decrease of the criterion W (Ward's criterion).

2.2.3 When external data are taken into account

Both strategies, described in paragraphs 2.2.1 and 2.2.2, can be extended to the case when it is available a table, X, which contains the characterisation of the products according to *q* external variables (Vigneau and Qannari, 2002). From an exploratory point of view, this extension allows to explain a partition of variables with the help of added variables linked with them. From a predictive point of view, this extension gives a single prediction model for each segment of variables. In other words, the problem is concerned with clustering a set of variables and at the same time exploring how this clustering may be explain by external information.

2.2.3.1 Extension: negative correlation means agreement

In addition to the variables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$, we consider a data set X consisting of *q* external variables, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$, which refer to the same n units. The authors impose that

the latent component of each group is expressed as a linear combination of external variables. The following quantity has to be minimized:

$$E^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\left\|y_j - \alpha_{kj}c_k\right\|^2 \text{ under the constraint that } c_k = Xb_k \quad (6)$$

It can be verified that this is equal to maximize:

$$T^* = \frac{1}{n}\sum_{k=1}^{K}\frac{b_k' X' Y_k Y_k' X b_k}{b_k' X' X b_k} \quad (7)$$

For a fixed partition in K groups, the optimal value is given by:

$$T^* = \sum_{k=1}^{K}\nu_1^{(k)}$$

where $\nu_1^{(k)}$ is the largest characteristic value of $\frac{1}{n}(X'X)^{-1}X'Y_kY_k'X$. This optimal value is expected to be in the form of vector b_k which is a characteristic value associated to $\nu_1^{(k)}$. The latent component $c_k = Xb_k$, in the group Gk, is the first component of Principal Component Analysis with Instrumental Variables technique (PCAIV) performed of Y_k on the block of external variables X (see for instance Sabatier, 1987; Sabatier *et al*., 1989).

The solution of this problem of maximization could become difficult in the presence of collinearity between the external variables X. Because of the inversion of matrix X'X , the uncertainty associated with the estimation of coefficient vectors b_k could be very large. In fact, it could be impossible to estimate the vectors of coefficients b_k whenever the number of observations is less than the number of external variables.

There is an alternative that consists of a direct extension of T criterion taking into account external variables.

The new criterion $\widetilde{T}$ will be maximized:

$$\widetilde{T} = n\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}Cov^2\left(y_j, c_k\right) \text{ under the constraints } c_k = Xa_k \text{ and } a_k'a_k = 1 \qquad (8)$$

then:

$$\widetilde{T} = \frac{1}{n}\sum_{k=1}^{K} a_k' X' Y_k Y_k' X a_k$$

The maximization of $\widetilde{T}$ under the considered constraints leads to a partition similar to that obtained maximizing T, except that in this case, the latent component of group G_k is given by $c_k = Ya_k$ where a_k is the first eigenvalue of the matrix $\frac{1}{n}X'Y_kY_k'X$ associated to the largest eigenvector $\mu_1^{(k)}$. In this case the component c_k, in the group G_k, is the first principal component of PLS2 regression (briefly described in paragraph 3.2.1) of the block of variables Y_k on the block of variables X.

The partitioning algorithm may be completed by a hierarchical clustering algorithm according to the same rationale as discussed above, and, once again, two groups of consumers A and B will merged whether the decrease of the criterion will be as small as possible.

2.2.3.2 Extension: negative correlation means disagreement

As CLV depending on criterion E, could be adapted to take into account external variables, the approach could be extend to the case where the sign of correlation is important (criterion Q). Let us consider the maximization of the following criterion:

$$Q^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\left\|y_j - c_k\right\|^2 \qquad (9)$$

If K, the number of the groups, is fixed, in each group G_k, the latent component c_k is defined as:

$c_k = Xb_k$ with $b_k = (X'X)^{-1}X'\bar{y}_k$

where X is the matrix of external variables centred and eventually standardized.

Then, c_k is obtained from the multiple linear regression of $\bar{y}_k$ on X, and for a fixed partition in K groups, the minimum value of Q* will be equal to:

$$Q^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\left\|y_j - c_k\right\|^2$$

$$= \frac{1}{n}\sum_{j=1}^{p} y'_j y_j - \frac{2}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj} y'_j c_k + \frac{1}{n}\sum_{k=1}^{K} p_k c'_k c_k$$

$$= \frac{1}{n}\sum_{j=1}^{p} y'_j y_j - \frac{2}{n}\sum_{k=1}^{K}\sum_{j=1}^{p} p_k \bar{y}'_k X(X'X)^{-1}X'\bar{y}_k + \frac{1}{n}\sum_{k=1}^{K} p_k \bar{y}'_k X(X'X)^{-1}X'\bar{y}_k$$

then:

$$Q^* = \sum_{j=1}^{p} Var(yj) - W^* \text{ where } W^* = \sum_{k=1}^{K_i} p_k Cov(\bar{y}_{k,} Xb_k) \qquad (10)$$

It may be noticed that $Cov(\bar{y}_{k,} Xb_k) = \frac{1}{n}\bar{y}'_k X(X'X)^{-1}X'\bar{y}_k =$

$= \frac{1}{n}\bar{y}'_k X(X'X)^{-1}X'X(X'X)^{-1}X'\bar{y}_k = \frac{1}{n}(X(X'X)^{-1}X'\bar{y}_k)'(X(X'X)^{-1}X'\bar{y}_k)$ that can be written:

$$W^* = \sum_{k=1}^{K_i} p_k Var(P_X(\bar{y}_k)) \qquad (11)$$

where $P_X = X(X'X)^{-1}X'$ is the projection on the space generated by X.

This is the well known multiple regression model which can lead to very instable prediction when the predictors are strongly correlated between themselves. An alternative criterion will maximize:

$$\widetilde{S} = \sum_{k=1}^{K}\sum_{j=1}^{p} \delta_{kj} Cov\left(y_{j,} c_k\right) \text{ under the constraints } c_k = Xa_k \text{ and } a_k^{'} a_k = 1 \qquad (12)$$

The problem of optimization is solved as above, except that the latent components are defined as follows:

$$c_k = Xa_k \text{ where } a_k = \frac{X^{'}\overline{y}_k}{\sqrt{\overline{y}_k^{'} XX^{'}\overline{y}_k}}$$

In a group G_k the latent component c_k is then the first component of PLS1 regression (briefly described in paragraph 3.2.1) of the centroid variable of the group $\overline{y}_k$ on X. Once again the authors use an initial hierarchical algorithm based on the same criterion $\widetilde{S}$.

2.2.4 Comparison with VARCLUS procedure

As mentioned above, VARCLUS procedure and CLV approach are two alternative methods which allow clustering of variables without a distance matrix has been defined. These two methods are very similar. VARCLUS, as well as CLV, answers both the issues i) and ii) raised at the beginning of this chapter without, however, considering the situation where external variables are available. The procedure implemented in SAS attempts to divide a set of variables into nonoverlapping clusters in such a way that each cluster can be interpreted as essentially unidimensional.

Associated with each cluster is a linear combination of the variables in the cluster, which may be either the first principal component or the centroid component. In the case i) the procedure goes to maximize the sum across clusters of the variation accounted for by the cluster components. This criterion is the same as T criterion (eq.2) except that VARCLUS algorithm is a descendent hierarchical algorithm while the CLV hierarchy is ascendant. At that point groups, which have the smallest percentage of variation explained by its cluster component or the largest eigenvalue associated with the second principal component (it depends which options are specified) are split into two clusters. How to split the cluster is given by finding the first two principal components, performing a quartimax rotation on the eigenvectors, and assigning each variable to the rotated component with which it has the higher squared correlation. A second phase involves a search algorithm in which each variable in turn is tested to see if assigning it to a different cluster increases the amount of variance explained. If a variable is reassigned during the search phase, the components of the two clusters involved are recomputed before the next variable is tested. This consolidation phase is quite complicated and it needs a lot of computations. In CLV the consolidation phase is placed at the end of partition algorithm that allows an easier program, a consolidation of the groups and an increase of criterion T.

In the case ii) variable partition is built around centroids defined as the unweighted average of the variables of the group. Nevertheless, the unidimentional criterion, which the choice of the groups that have to be split is based on, is not pertinent if the researcher is looking for local groups of positive correlated variables. In fact, a group could be almost unidimentional even if lumps together negative correlated variables. Furthermore, the first phase of the reassignment procedure using squared

correlation between variables and centroids of groups is questionable. For this reason, it is limited to one iteration, for second phase benefit, which aims to iteratively increase the amount of variance explained. Despite the difference between the two procedures, VARCLUS with CENTROID option and CLV in the case ii) are based on the same criterion, i.e. the weighted sum of variances of group centriods. But, the case when external data are available is not considered in VARCLUS procedure.

2.2.5 Summary of CLV cases

CLV method, whose algorithm is synthesized in figure 2.3, includes several options in classification of variables. It stands out because of the kind of method (METHOD), which depends of the meaning assigned to the negative correlation among the variables: agreement or disagreement. The option where external variables are included in the clustering is regulated by VARX parameter. A third dichotomic parameter (COLI) allows the practitioner to take into account a possible problem of collinearity among the external variables.

The summary of different cases considered in CLV algorithm is shown in table 2.1.

Figure 2.3 –The three steps of the CLV algorithm.

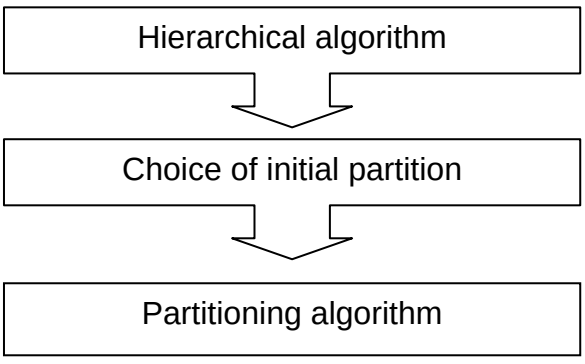


Table 2.1 – The cases considered in CLV algorithm according to three parameters (METHOD, VARX and COLI), c_k indicates the latent component in group G_k, $\mathbf{Y_k}$ matrix which consists of variables of that group and $\bar{y}_k$ centroid of group k (Vigneau & Qannari, 2006).

		METHOD=1 Strong positive or negative correlation means agreement	METHOD=2 Strong negative correlation means disagreement
VARX Not available		Criterion: E (eq. 1) o T(eq. 2) $\mathbf{c_k}$: collinear to 1st principal component of $\mathbf{Y_k}$	Criterion: Q (eq. 4) $\mathbf{c_k}$: centroid (or average) of group G_k
VARX Available	COLI=0 No problems of collinearity among VARX	Criterion: E*(eq. 6) o T*(eq.7) $\mathbf{c_k}$: 1st PCAIV component of $\mathbf{Y_k}$ on $\mathbf{X}$	Criterion: Q*(eq. 9) $\mathbf{c_k}$: estimated with linear regression of $\bar{y}_k$ on $\mathbf{X}$.
	COLI=1 Collinearity problems among VARX	Criterion: $\widetilde{T}$ (eq. 8) $\mathbf{c_k}$: 1st component of PLS2 regression of $\mathbf{Y_k}$ on $\mathbf{X}$.	Criterion: $\widetilde{S}$ (eq. 12) $\mathbf{c_k}$: 1st component of PLS1 regression of $\bar{y}_k$ on $\mathbf{X}$.

It is to point out that in the rest of this work we will refer only to the method 2, which is the most suitable to treat data from a preference test, where *p* consumers are asked to evaluate *n* products and so in the case in which negative correlation between the ratings of two consumers highlights their different opinion about the products.

2.3 Determination of the number of groups

In the previous paragraphs, the number of groups was supposed known, but that it is not true. In CLV approach the number of groups can be identified through looking at the plot which shows the evolution of the aggregation criterion (Vigneau *et al.*, 2001). In some cases, when that figure does not permit to unequivocally identify the number of the groups, it could be worthwhile to arrange a tool based on an automatic procedure. This procedure, proposed by Sahmer (2003) and Sahmer *et*

al. (2006) according to T criterion, is easily adaptable in the case which aims to determine the number of clusters in the case ii).

2.3.1 Permutation test

The main rationale of this procedure is based on the comparison of $\Delta\widehat{Q}$ values at each step of the hierarchical algorithm with the correspondent values obtained from a simulation where variables are not more correlated. The method develops through the following steps:

1. An hierarchical classification on the columns of Y data table is performed and $\Delta\widehat{Q}^{(i)}$ is the value of the aggregation criterion which corresponds to the step at the end of which there are *i* groups;
2. The following procedure is performed B times (for instance 100 times):
   (a) A random permutation of values located in each column of Y is undertaken, independently from other columns. Let indicate Y_{perm} as resultant matrix. The variances of Y_{perm}'s variables are equal to those of variables of Y. Nevertheless, correlation among Y_{perm}'s variables corresponds to a correlation among not correlated variables.
   (b) Hierarchical classification of Y_{perm}'s columns is undertaken. $\Delta\widehat{Q}^{(i)}_{perm}$ represents the value of aggregation criterion at the step where there are *i* groups.

3. In each step *i* of hierarchical classification, the 5th percentile $q_{0.05}^{(i)}$ of B values of $\Delta Q_{perm}^{(i)}$ is computed.
4. The choice of the right number of groups is performed according to the following rules:
   (a) At the first step of the classification, i.e. the step where for the first time two variables are lumped together in a group, if the value of $\Delta\widehat{Q}^{(p-1)}$ is larger than $q_{0.05}^{(p-1)}$, we have to conclude that observed data are not different from those obtained with not correlated variables. There are not groups;
   (b) At the last step of the classification, i.e. the step at the end of which the researcher decide that there are only one group or more, if $\Delta\widehat{Q}^{(1)}$ is less than $q_{0.05}^{(1)}$, we have to conclude that there is just one group of variables, otherwise we pass to the next step.
   (c) Whether $\Delta\widehat{Q}^{(2)}$ is less than $q_{0.05}^{(2)}$ (passing from a solution with 3 groups to 2 groups), we decide for the solution with 2 groups, otherwise the procedure goes on as long as $\Delta\widehat{Q}^{(i)}$ is less than $q_{0.05}^{(i)}$. If this situation is repeated for *i* = K times we will conclude that there are K groups.

For more complex cases, that take into account the availability of external variables in the presence of collinearity problems or not, the permutation algorithm and the choice of the right number of clusters remain unchanged except that it will be involved the variations of the criterion pertinent to the case under investigation.

2.4 Goodness of clustering

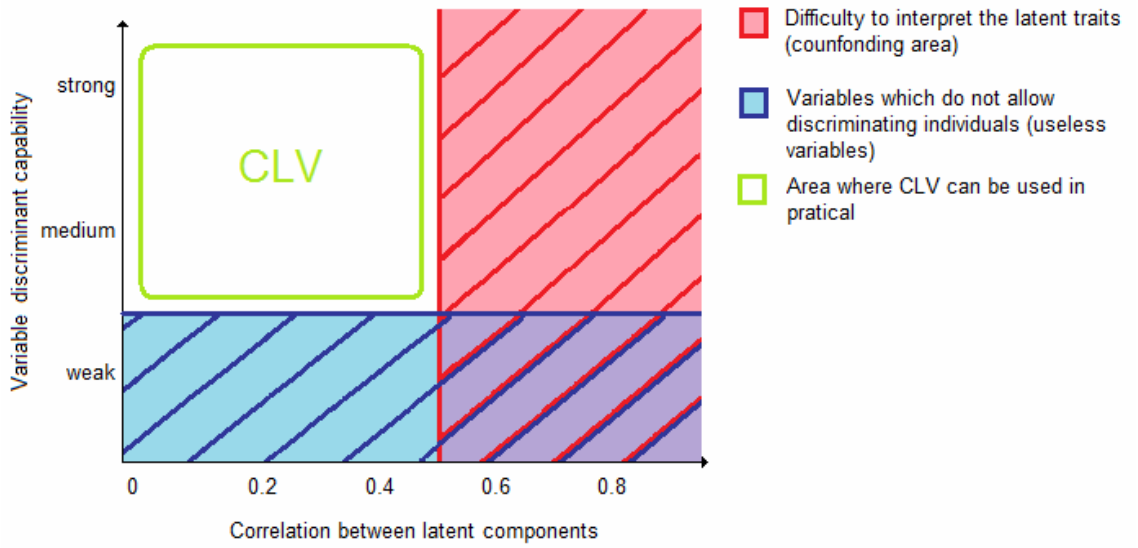
In order to measure the goodness of clustering, there are two main factors influencing the results of the method, which have to be considered by the researcher.

First of all, the results can be conditioned by the discriminating power of the variables. If the variables measured by the researcher are useless in terms of discriminant capability the resulting consumer clustering cannot be excellent.

Secondly, in order to have clusters easy to interpret, estimated latent components associated to each group have to be as uncorrelated as possible to each other, in this way they will describe very different preference directions.

Hardouin JB. (2006), in a work where he compared different methods for clustering quality of life items, suggested an area where CLV can be used in practice with variable discriminating power at least "medium" and correlation among latent components not greater than 0.5 (figure 2.4).

Figure 2.4 –Quality of the CLV method (Hardouin, 2006).



In the case where negative correlation indicates disagreement (method 2), a global index of clustering goodness in each group should be concerned with the relationship between variables, belonging to group k, and the associated latent component c_k. In order to measure this relationship we propose the RV coefficient, described in details in the next paragraph.

2.4.1 RV coefficient

The RV coefficient was introduced by Escoufier (1973) as a measure of similarity between squared symmetric positive semi-definite matrices and as a theoretical tool to analyze multivariate techniques. This coefficient, sometimes called *vector* or *matrix* correlation is similar to the correlation coefficient and it takes values between 0 (no similarity) and +1 (maximum similarity).

In order to compare rectangular matrices using the RV coefficient the first step is to transform them into square matrices. The RV coefficient between (n by p) matrix A and (n by m) matrix B, which columns of both of them are supposed to be centred, is defined by:

$$RV = \frac{trace\{AA'BB'\}}{\sqrt{(trace\{AA'AA'\})(trace\{BB'BB'\})}} \tag{13}$$

An obvious practical problem is to be able to perform statistical testing on the value of a given coefficient. In particular it is often important to be able to decide if a value of coefficient could have been obtained by chance alone. Such statistical approaches have focused on permutation tests, performed on the entries of each column of A and B. Interestingly, work by Kazi-Aoual *et al.* (1995, see also Schlich, 1996) has

shown that the mean and the variance of the permutation test distribution can be computed directly from AA' and BB'.

The first step is to derive an index of the dimensionality of the matrices, denoted β_A:

$$\beta_A = \frac{[trace\{AA'\}]^2}{trace\{AA'AA'\}} \tag{14}$$

The mean of the set of permutated coefficients between A and B is then equal to:

$$E(RV) = \frac{\sqrt{\beta_A \beta_B}}{n-1} \tag{15}$$

The case of variance is more complex and involves computing three quantities for each matrix. The first quantity is denoted δ_A, it is equal to:

$$\delta_A = \frac{\sum_n a_{nn}^2}{trace\{AA'AA'\}} \quad \text{where } a \text{ is the generic element of matrix AA'} \tag{16}$$

The second quantity is denoted α_A and is defined as $\alpha_A = n - 1 - \beta_A$. The third one is denoted C_A and is defined as:

$$C_A = \frac{(n-1)[n(n+1)\delta_A - (n-1)(\beta_A + 2)]}{\alpha_A (n-3)} \tag{17}$$

With these notations, the variance of the permuted coefficients is obtained as:

$$V(RV) = \alpha_A \alpha_B \frac{2n(n-1) + (n-3)C_A C_B}{n(n+1)(n-2)(n-1)^3} \tag{18}$$

The sampling distribution of the permuted coefficients is relatively similar to a normal distribution and therefore we can use a Z criterion to perform null hypothesis testing used to test the null hypothesis that the observed value of RV was due to chance.

$$Z(RV) = \frac{RV - E(RV)}{\sqrt{V(RV)}} \tag{19}$$

One can expect this value to be roughly greater than 2 if the agreement between A and B is better than what can be by chance (Schlich, 1996).

In order to know how much the global preference direction, represented by c_k, summarise the individual preference directions, we calculate the RV coefficient for the group k replacing A with Y_k (the matrix which columns are k vectors containing hedonic scores on n products given by the p_k consumers collected in the group k) and B with c_k in equation (13)

$$RV_k = \frac{trace\{Y_k Y_k' c_k c_k'\}}{\sqrt{(trace\{Y_k Y_k' Y_k Y_k'\})(trace\{c_k c_k' c_k c_k'\})}}$$

where the greater $trace\{Y_k Y_k' c_k c_k'\}$ is, the more similar are Y_k and c_k in terms of product distances, while the greater is $trace\{Y_k Y_k' Y_k Y_k'\}$, the more different the products are for the group of consumers collected in G_k.

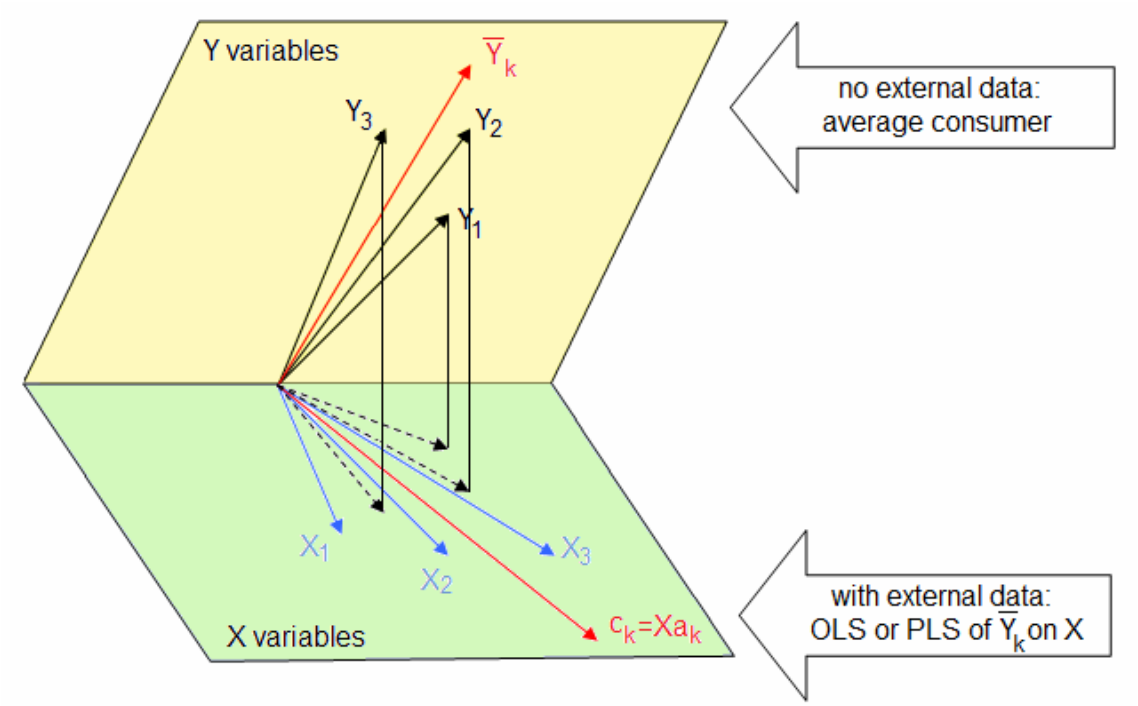
When the clustering is performed on the base of criterion Q (eq.4), i.e. the case where external variables are not available, the latent component of each group is the centroid of the group (representing the G_k average consumer) built on Y plane.

In that case RV_k coefficient values will be greater than those calculated for groups of consumers performed using criteria Q* or $\widetilde{S}$, where latent components are built on external variable plane (figure 1.5).

Furthermore, the magnitude of a RV coefficient depends on the number of observations and for this reason, the number of the product have to be enough in order to compute a statistical test of RV significance. In addition, RV coefficient depends also

on the number of variables in the groups, therefore tiny groups of consumers show generally coefficients less than those showed by more numerous groups .

Figure 2.5 –Geometry of CLV in method 2. Orange plane represent the plane of Y variables (black arrows), while the green one represent that of external variables (blue arrows). Dotted arrows are Y variables projections, while red arrows are the latent components.



Chapter 3

CLV for L-structured data

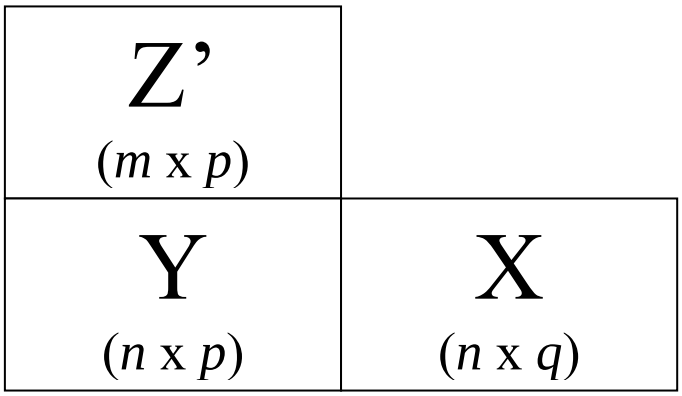
3.1 Introduction

The particular data structure object of the methods described in this chapter is commonly called L-shape structure, because of the shape that the involved data matrices take when are positioned one next to each other. Such a structure has had a large use in the field of preference studies, where the response of a consumer panel in terms of overall liking is traditionally evaluated on a certain number of products. Information from such a study are generally collected in a (n x p) Y matrix, where statistical units are the n products, while the p variables are vectors containing the preference ratings for

each judge. The *n* products are described through variables which have a very different nature: often they give chemical composition, physical characteristics, information about ingredients or sensory evaluations performed by a panel of trained judges or all of this together. These product characteristics are collected in a (n x q) X matrix.

Data structure of our interest includes a further (p x m) Z matrix, which collects consumers background. Generally, it consists of socio-demographic indications, variables able to describe purchase and consumption behaviour of the products of interest, and/or opinions correlated to them. The data structure here described is depicted in figure 3.1.

Figure 3.1 – The three matrices located in a L-shape structure.



The main goal of a preference study is commonly to identify groups of consumers showing similar preferences. In order to reach this aim the researcher can consider two different approaches according to the role assigned to Z matrix. The first approach, once the relationship between Y and X has been defined, consists in finding clusters of consumers considering that persons collected in the same group appreciate the same product characteristics. These groups are then described relating them to the columns of Z. A second way to analyze consumer preference consists in defining

segments of consumers on the basis of their characteristics before establishing any relationship with hedonic notes. Then, for each segment, a model that links evaluation scores to product characteristics is undertaken. Both these approaches could be considered incomplete because they build segment of consumers using partial information, i.e. assigning a role exclusively descriptive to at least one of the external matrices (X and Z): in the first case, clusters are going to be inhomogeneous in terms of consumer descriptors (Z), while, in the second case, they will be nonuniform in terms of product characteristics (X).

A solution consists in giving an active role to both the external matrices in order to obtain homogeneous segments of consumers from product characteristics point of view as well as consumer profile.

Methods described in this chapter are developed in the framework of preference studies and are applied whenever a L data structure is involved. In the next paragraph PLS regression methods are briefly described, giving particular attention to two-step L-PLS regression models available in the literature, while in the second part of the chapter new adaptations of CLV approach are proposed, in order to obtain an automatic consumer segmentation in a L-structure context.

3.2 Analysis on L-structured data in PLS-R

In the last few years, many researchers have been dealt with complex data structures such as those described above, especially after the introduction of PLS regression by S.Wold *et al.* (1983). Since then, many developments have been

proposed, mainly in the field of chemometrics, cosmetics, biology as well as agro-alimentary industry like L-PLSR (Martens *et al.*, 2005) and exo-LPLS regression (Saebø *et al.*, 2008), both born to model three-block data. The extensions of PLS regression method, illustrated in the following paragraph, inspired the procedure, proposed in the second part of this.

3.2.1 Overview of Partial Least Squares Regression

Herman Wold first formalized the basic idea behind partial least squares with two papers. In the first work, published in 1966, he introduced NIPALS (Non Linear Iterative Partial Least Squares) algorithm in order to implement principal component analysis with missing data. In 1975, he published a second work where a method for modelling the relationship among several blocks of variables observed on the same number of units was defined. From this method, called PLS approach, all the other PLS techniques derive as particular cases. A large diffusion of this method in the scientific community has happened in 1983 after the introduction of PLS regression, as we know it today, by S. Wold *et al.*.

Partial least squares regression is a technique for modelling relations between two blocks of variables by means of latent variables and it consists in an adaptation of Wold's multi-blocks PLS approach. In the field of chemometrics (Geladi and Kowaski, 1986) and sensometrics (Martens and Naes, 1989), this technique is a standard multivariate tool for obtaining multivariate models which relate a set of independent

variables X with a set of dependent variables Y. This paragraph introduces a brief description of its main concepts.

Let X be a matrix with n samples and q predictors and Y a matrix with p response variables for the same number of samples as X. Usually, variables in X and Y are centred, subtracting the mean, and scaled dividing by standard deviation (Geladi and Kowaski, 1986). The goal of PLS regression is to predict Y from X and to describe their common structure. When Y is a vector and X is full rank, this goal could be accomplished using ordinary multiple regression. When the number of predictors is large compared to the number of observations, X is likely to be singular and the regression approach is not reliable because of multicollinearity (Wold, S. *et al*., 1984). Several approaches have been developed to resolve this problem. One of them is to eliminate some predictors (using, for instance, step-wise methods), another one, called Principal Component Regression (Kendall, 1957) is to perform principal component analysis of the X matrix and then use the principal components of X as regressors for modelling Y. The orthogonality of the components eliminates the multicollinearity problem. But, the problem of choosing an optimum subset of predictors remains. By contrast, PLS regression finds components from X that are also relevant for Y building a model as follows:

$$Y = XB + E$$

where B is a (q x p) matrix of regression coefficients and E is a (n x p) matrix of residuals.

Specifically, PLS regression searches for a set of components that performs a simultaneous decomposition of X and Y with the constraint that these components

explain as much as possible of covariance between X and Y. Then it follows a regression step where decomposition of X is used to predict Y.

So, the independent variables are decomposed as X = TP' where T'T is a diagonal matrix. The decomposition of X will be exact if the number of latent variables T is equal to the rank of X. By analogy with PCA T is called the score matrix, and P the loading matrix. Likewise, Y is estimated as $\hat{Y} = TC'$.

The latent vectors could be chosen in a lot of different way and in order to specify T, additional conditions are required. For PLS regression this amounts to finding two sets of weights w and c in order to create a linear combination of the columns of X and Y such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors t=Xw and u=Yc with the constraints that w'w=1, c'c=1 and t'u be maximal. From here, we deduce that PLS regression is a compromise between a Canonical Correlation Analysis of X and Y, a PCA of X and oblique PCA of Y. When the first latent vector is found, it is subtracted from both X and Y and the procedure is re-iterated until X becomes a null matrix.

It is to be noticed that, whenever Y consists in just one variable, this regression is called PLS1 otherwise it is called PLS2 or two-block PLS. Details on theory and algorithms can be found in most textbooks covering multivariate statistics, like Martens and Naes (1989), Tenenhaus (1998) and Naes *et al.* (2002) among many others. While the general PLS regression is limited to handling only two sets of data at time, in the next section, extensions based on a two-steps procedure are briefly described.

3.2.2 Two-step L-PLSR

Since X (*n* x *q*), Y (*n* x *p*) and Z (*p* x *m*) have different dimensions, the variables of Z cannot be modelled with X and Y by regression over *n* objects in a two-way PLS model. In recent years, many models have been proposed in literature to estimate relationships among matrices in a L structure, some of which are based on simultaneous or alternated singular value decompositions. Here we focus on two-step modelling alternatives.

Kubberød *et al.* (2002) proposed a two-step approach for using Z-information folding the information from Z with Y' in order to give the Z-information a dimension *n*, common with X and Y. Authors first estimated reduced-rank regression coefficient matrix $B'_{Z,Y}$ (*n* x *m*) from the mean-centred model $Y' \cong Z\ B_{Z,Y}$ by PLS-R, and secondly regressed both Y and $B'_{Z,Y}$ on X based on the linear model $[Y|\ B'_{Z,Y}] \cong X\ B_X$, in another reduced-rank PLSR step.

Thybo *et al.* (2003) used as well a two-step method to identify relationships among the three tables, but in order to simplify the analysis, they replaced the regression coefficient matrix by the correlation coefficient matrix $R_{Y,Z}$ (*n* x *m*) between the *n* rows in Y and the m columns in Z, correlated over *p* elements. A two-blocks PLS regression is performed between $[X|\ R_{Y,Z}]$ and Y.

A procedure, which can be considered as a mixture of Kubberød *et al.*(2002) and Thybo *et al.* (2003) proposals, was applied in Lengard and Kermit (2006). Instead of using regression coefficients from a prior PLS regression, these authors used correlation coefficients to represent the relationships between Z and Y' $R_{Z,Y}$ (*m* x *n*), considering the model $[Y|\ R'_{Z,Y}] \cong X\ B_X$.

Esposito Vinzi *et al.* (2007) proposed a double regular PLS-R, where in the first step a PLS2 regression of Y on X is performed according to $Y \cong T_X C'_X$ (where T_X components are built in order to maximized the variance accounted for in X and in the relation with Y), while in the second step loading vector matrix (p x A_x) C_X (where A_x is the number of components involved in the regression of Y on X) is regressed on the consumer descriptors contained in Z, according to the model $C_X \cong T_Z C'_Z$ (where C'_Z is the loading vector matrix in the regression of C_X on Z). In author opinion, there is not difference, under methodological point of view, in starting by regressing Y on X or on Z (as Kubberød *et al.*, 2002, have done it). From practical point of view their approach in preferable because product descriptors usually have a stronger predictivity on preferences than consumer characteristics. Final model is expressed as $Y \cong T_X C_Z T'_Z$ that includes information regarding interaction between X and Z on Y variability.

These two-step PLS regressions present a number of interesting advantages. First of all, being based on regular PLS regressions, provide parameter estimation when there are missing data in the original matrices. These models can be applied to tables with more variables than individuals and in case of multicollinearity. Data visualisation is based on typical PLS graphical representation and, in the case where a reasonable number of variables are involved, that allows an immediate interpretation. Regarding the interpretation of the results it can be used all PLS standards tools of evaluating predictivity or quality of the model.

Despite of the clear advantages, described methods are built around a quite cumbersome two-step procedure: the inclusion in the model of consumer background Z matrix, which usually presents a considerable amount of variables, gives a graphical

representation difficult to interpret. Furthermore, models which include correlation matrix have complicated mean-centring properties.

A non secondary problem consists in defining groups of consumers. PLS approaches do not include automatic clustering methods and for this aim classical cluster analysis are generally applied, once the PLS-R is estimated. In Esposito Vinzi *et al.* (2007), in order to identify homogeneous groups of consumers taking into account information from X and Z, the segmentation is performed applying k-means cluster analysis on T_Z columns.

In Tenenhaus *et al.* (2005), instead, consumer segmentation is undertaken by dividing the space, where products, products characteristics and consumers were depicted, in four areas bordered by the two bisectors in order to obtain groups correlated to PLS components. This solution does not take into account that the number of groups could be different from four and that latent variables much correlated to the groups could be nonorthogonal.

3.3 A new proposal based on CLV approach: active role of Z

Classification of consumers could be performed only on the hedonic ratings contained in Y matrix. In this way, the obtained groups will be differentiated on the basis of their preference expressed to the tested products. Clustering could be also undertaken in the case when external variables are available, being these products characteristics or consumer background. When the researcher is interested in clustering consumers with respect to their preference taking into account product descriptors, the

clusters will be slightly different that those obtained using only Y data. An alternative approach, often used in marketing research, is to seek segments of consumers which have similar socio-demographic characteristics or/and analogue behaviours with respect to the products under investigation.

In next paragraphs a way to fold Z information with Y in order to perform CLV approach in two-way analysis (3.3.1) and L structured data context (3.3.2) are suggested.

3.3.1 Two-way CLV on Y taking into account Z information

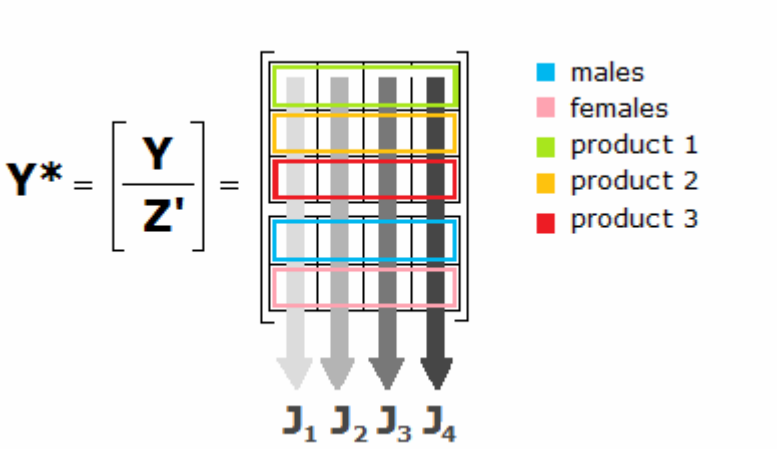
In CLV approach, as proposed by the authors, a classification around latent components can be performed on preference data only, or taking into account external variables. In this second case, the technique is extended to a two-way analysis, where table X with product descriptors is involved in building the clusters. The alternative two-way analysis, where hedonic notes as well as individuals background are both taken into account in order to find out groups of consumers showing similar preference and behaviours, is not considered in the algorithm proposed by Vigneau and Qannari (2003).

Here, we suggest a way to cluster consumers taking into account their profile in terms of information collected by a questionnaire. Survey responses are commonly qualitative data, and since many multivariate statistical techniques require another way to represent qualitative concepts to make sense, they are traditionally converted in dummy variables. For this reason, from now on we consider Z a matrix containing

binary variables, where 1 indicates the presence of a given characteristic and 0 otherwise.

Our proposal consists in building a new matrix obtained putting (n x p) Y matrix of zero-mean hedonic scores and the transpose of (p x m) Z matrix together. This new matrix, called Y*, has (n + m) rows and p columns. In this way, the columns of Y* are vectors comprising preference evaluations and questionnaire responses ascribable to each individual as shown in figure 3.2.

Figure 3.2 – Y* matrix representation in the case where 4 consumers, characterized by gender, evaluated 3 products.



Then, clustering of the columns of Y* (judges) is performed using CLV algorithm (according to Q criterion, with no further centring operation) achieving groups which show different taste and different characteristics like: age, gender, purchase behaviour, frequency of consumption and so on. Standard tools provided by CLV method for choosing the right number of groups can also be used. The values of latent components c_k in group G_k give scores of the n products and the m judge descriptors and so a descriptive profile for each group can be identified. If m is a very

large integer, description of groups in terms of Z information may be complex and a previous selection could be needed.

3.3.2 Two-step L-CLV

In paragraphs 3.2, it was highlighted how PLS methods, especially recent developments in partial least square regression, are pioneering techniques in searching existing relationships between L structured data. As previously mentioned, these techniques solve frequent problems in real data-set as multicollinearity, high number of variables in comparison with units number and missing data. Despite of this, these methods present very complex procedures which need for introducing external algorithms in order to perform consumer segmentation.

CLV approach, in the case described in paragraph 2.2.3.2, should be a more useful and complete tool for automatically defining homogeneous groups of consumers taking into account information from both Y and X matrices. This approach does not consider the case when a Z matrix, containing individual characteristics, is available. Here we present a procedure that provides a segmentation of consumers involving Z information. This procedure, inspired by few papers on two-step L-PLS consists in two fundamental steps: in the first stage an ordinary matrix multiplication between preference data (n x p) Y matrix and dummy transformation of individual descriptors (p x m) Z matrix is computed. This product, denoted by YZ, for each pair i and j with $i=1,2,\ldots,n$ and $j=1,2,\ldots,m$ and where $r=1,2,\ldots,p$ identifies consumers, is given by:

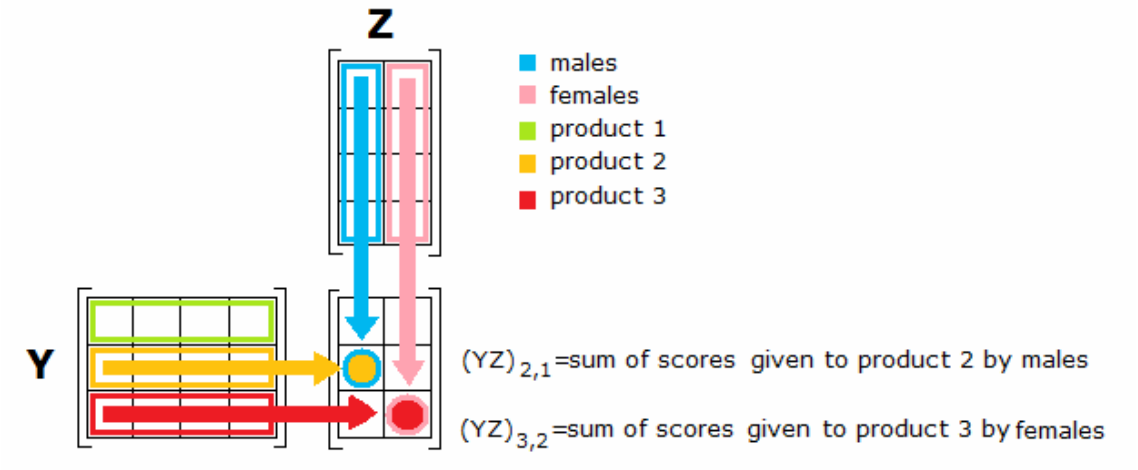
$$(YZ)_{ij} = \sum_{r=1}^{p} y_{ir} z_{rj} = y_{i1} z_{1j} + y_{i2} z_{2j} + \cdots + y_{ip} z_{pj}$$

where the generic element $(YZ)_{ij}$ is the sum of hedonic scores of product *i* over all the consumers which present j^{th} modality of Z. In figure 3.3 YZ matrix is depicted in the case where Z collects only gender information about 4 consumers who evaluated 3 products.

The rationale behind this calculation is to capture information from Z and to incorporate them with Y, obtaining a new table with the *right* dimensions. It is to be noticed that, this operation create an information loss in terms of Y and Z variability. This information loss problem could be partially solved by constructing a new larger matrix $\tilde{Y} = [Y|YZ]$, obtained with placing side by side the hedonic ratings Y and the product between Y and consumer attitudes Z. However, information about Z variability is lost.

In the second step, the original CLV algorithm is performed on $\tilde{Y}$ utilizing X matrix as external variables.

Figure 3.3 – YZ matrix representation in the case where 4 consumers, characterized by gender, evaluated 3 products.



This procedure, called two-step L-CLV, thanks to the similarity with two-step L-PLS models, presents a number of interesting advantages. First of all, it does not need to implement a new algorithm but it suits the procedure that maximizes Q* or $\widetilde{S}$ criterion and so enjoys the same advantages:

- provides an automatic segmentation of the consumers;
- builds nonorthogonal latent components which are associated to the groups;
- the number of groups is suggested by the evolution of aggregation criterion diagram;
- multicollinearity problem can be managed by maximizing $\widetilde{S}$ criterion;

Furthermore, for each identified segment, a single model, which relates preference evaluations and consumer descriptors with product characteristics is estimated. Finally, clustering is performed on the *m* columns of YZ as well as the *p* columns of Y (consumers); in this way, clusters of consumers are obtained with a direct indication regarding to their characteristics. Group assignment of Z variables does not merely reflect the frequency that different categories assume in each group. In fact, information from Z take part actively in the clustering and its role is not only descriptive. Therefore, for each group it is given a list of consumer descriptors associated to the profile of products characteristic of each group. A clearer explanation to how Z variables are assigned to the groups will be given in the next chapter through a real data-set application.

In the practical use of the most common statistical techniques based on latent components, it is of interest to interpret the estimated factors. As in PCA factor loadings are conceived as indices of relationship of variables to the factors, here correlation between each group latent component c_k and its associated consumer descriptors can

give an idea about the importance that the variables have in the interpretation of each group.

Once the correlation between columns of YZ and each latent component is computed, the problem is, of course, to identify when an index of relationship can be defined *significant*. In lieu of a statistical definition of significance, it is reasonably easy after experience in a particular research area to define what should be considered a minimum adequate value for a single variable on a factor. According to Overall and Klett (1972) we propose a critical value of 0.35, as a threshold, where values less then it have not to be taken into account in the interpretation phase.

Chapter 4

Applications in consumer liking for food products

4.1 Introduction

In the present chapter we will show some applications of the methodologies described in the previous chapters on two datasets from consumer tests performed on berry fruit juices and on typical cheeses, respectively.

The structure of available data follows closely the L-structure described in Chapter 3. A central matrix, containing the consumer preference scores on a certain number of products, is described by two matrices containing external information

respectively on products and on consumers. Two datasets will be used to illustrate our new proposal (described in paragraph 3.3.2). A broader discussion of the two datasets is given in paragraphs 4.2 and 4.3.

4.2 Interberry dataset

The data are part of INTERBERRY project, a multidisciplinary study which aims at improving the soft fruit quality and marketability, carried on by IASMA (Agricultural Institute of S. Michele all'Adige – Italy) in the last three years. In the frame of this general aim innovative processed products based on berry fruit has been studied in order to develop a new line of healthy drinks with appealing sensory characteristics and nutritional advantages. In order to meet the needs of modern consumer that increasingly buy *ready to eat* food to save time without giving up pleasure and nutrients intake of a healthy diet, 25 juice prototypes obtained without any pasteurization treatment were designed and selected through a focus group of 10 people involved in several aspects of juice production: marketing, product development and research. Each one of the five berry fruits under investigation (strawberry, raspberry, blackberry, redcurrant and blueberry) was proposed in 5 different formulations (table 4.1): fresh squeezed berry fruit (20%) was mixed with each of the 5 different base juices (apple, orange, blood orange, pineapple and pomegranate) in order to find out which base juice could enhance the sensory peculiarity of each berry fruit.

Apple, orange and pineapple juices were chosen as the most widespread commercial products, while pomegranate juice was used because of the growing interest in products with a high content of antioxidants.

These juices were firstly analyzed in terms of chemical compositional parameters, and secondly they were assessed by a consumer panel in terms of overall liking. Information about consumer characteristics and habits regarding fruit and fruit juices purchase and consumption were also collected. Three tables are thus obtained, which are described in details in the following paragraphs.

Table 4.1 - Mixes evaluated in the 5 sessions. The codes assigned to each juice are given in parentheses.

bases 80% / berries 20%	Pineapple	Orange	Pomegranate	Apple	Blood Orange
Strawberry	Test 1 (1P)	Test 1 (1O)	Test 1 (1PG)	Test 1 (1A)	Test 1 (1BO)
Raspberry	Test 2 (2P)	Test 2 (2O)	Test 2 (2PG)	Test 2 (2A)	Test 2 (2BO)
Blackberry	Test 3 (3P)	Test 3 (3O)	Test 3 (3PG)	Test 3 (3A)	Test 3 (3BO)
Red Currant	Test 4 (4P)	Test 4 (4O)	Test 4 (4PG)	Test 4 (4A)	Test 4 (4BO)
Blueberry	Test 5 (5P)	Test 5 (5O)	Test 5 (5PG)	Test 5 (5A)	Test 5 (5BO)

4.2.1 Preference data

A sample of 72 consumers was recruited among the staff and the students of IASMA among people which like berry fruit. The panel was asked to score the 25 juices in 5 sessions (table 4.1). In each session 5 blind-coded products were presented in a randomized order. Consumers were asked, at first, to rank the juices accordingly to their overall liking and, subsequently, to rate their acceptance for each juice on a nine-point hedonic scale (to 0 indicating extremely disliking and 9 extremely liking). Data are contained in a (25 x 72) table which from now on we will refer to as Y (Figure 4.1 (a)).

4.2.2 Products characteristics

Two types of product descriptors were recorded: 5 chemical parameters related to consumer perception quality aspects [ch_1 … ch_5], and 2 qualitative design variables, where one describe the type of base juice added and the other one the type of berry juice added. Design variables basically reflect the composition of the juices and were coded as dummy variables, in this way 10 design dichotomic variables were used: 5 modalities each [Q_1=$base_1$…$base_5$; Q_2=$berry_1$…$berry_5$]. The list of product characteristics with a brief description is given in table 4.2.

In order to give the same weight to each variable and to have more comparable loadings each quantitative variable ch_i was centred and standardized by sqrt(ch_i' ch_i), while each Q_i qualitative variables were not centred but globally weighted in order to have trace(Qi'Qi)=1. These data are contained in (25 x 15) table Y (Figure 4.1 (b)).

Table 4.2 – List of product descriptors (interberry data).

Variable name	Variable description
ch_1	Sugar content (sum of fructose, sucrose and glucose concentration):g/Kg
ch_2	Malic acid content):g/Kg
ch_3	Citric acid content):g/Kg
ch_4	Ascorbic acid content):g/Kg
ch_5	Total amount of polyphenols):mg/Kg
$base_1$	It is equal to 1 whether apple (A) is contained in juice mix, 0 otherwise
$base_2$	It is equal to 1 whether blood orange BO) is contained in juice mix, 0 otherwise
$base_3$	It is equal to 1 whether orange (O) is contained in juice mix, 0 otherwise
$base_4$	It is equal to 1 whether pineapple (P) is contained in juice mix, 0 otherwise
$base_5$	It is equal to 1 whether pomegranate (PG) is contained in juice mix, 0 otherwise
$berry_1$	It is equal to 1 whether strawberry (S) is contained in juice mix, 0 otherwise
$berry_2$	It is equal to 1 whether raspberry (R) is contained in juice mix, 0 otherwise
$berry_3$	It is equal to 1 whether blackberry (Blk) is contained in juice mix, 0 otherwise
$berry_4$	It is equal to 1 whether red currant (RC) is contained in juice mix, 0 otherwise
$berry_5$	It is equal to 1 whether blueberry (Blu) is contained in juice mix, 0 otherwise

4.2.3 Consumer information

At the end of each test, consumers were asked to fill a questionnaire concerning information considered relevant for consumer description and for the explanation of their purchase choices. Consumer descriptor matrix, referred to as Z, contains 179 qualitative variables and may be synthesised as follows:

- 7 socio-demographic variables, among which sex, age class, school level, etc., aiming at giving an idea of the *familiar environment* of the interviewed consumers;
- 38 variables [fruit_1 … fruit_38] describing fruit consumption and purchase habits;
- 37 variables [juice_1 … juice_37] describing juice consumption and purchase habits;
- 41 variables [bf_1 … bf_41] giving the description of berry fruit consumption and purchase habits;
- 26 variables which collect food neophobia scale scores, impressions on new foods, exotic foods, ready to eat foods and familiar foods [Fns_1 … Fns_10; NF_1 … NF_4; EF_1 … EF_4; RE_1 … RE_4; FF_1 … FF_4];
- 30 variables [H_1 … H_30] collecting consumer knowledge about healthy diet and antioxidants in general.

A detailed list of variables collected by the questionnaires is given in the appendix A.

These data are contained in (72 x 179) table Z (Figure 4.1 (c)).

Figure 4.1 – Schematic representation of interberry data

(a) preference data matrix **Y**

	Consumer preference scores		
	C1	…	C72
Prod1	Y (25 x 72)		
Prod2			
…			
Prod25			

(b) product characteristics matrix **X**

	Chemical compounds	**Design variables**	
	ch_1 … ch_5	$base_1$…$base_5$	$berry_1$…$berry_5$
Prod1	X (25 x 15)		
Prod2			
…			
Prod25			

(c) consumer information matrix **Z**

	Socio-demographic descriptors	**Fruit habits**	**Juice habits**	**Food familiarity**	**Berry fruits**	**Health**
	sex … child	fruit1…fruit38	juice1…juice37	Fns1…FF4	Bf1…Bf41	H1…H30
C1	Z (72 x 179)					
C2						
…						
C72						

4.2.4 Pre-treatment of data

A number of pre-treatments were performed on Z before using it in order to *filling in* missing data, transform qualitative variables in dummies and select a subset of them.

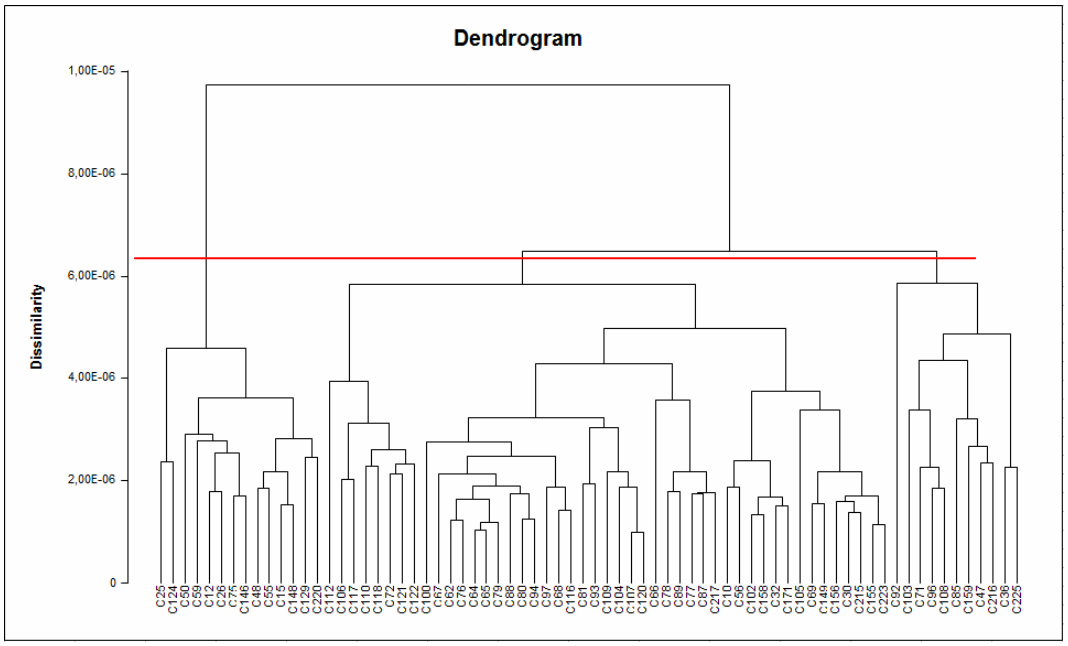
Significant advances have been made in the last 15 years regarding methodologies which handle treatment of missing data and discussions of these kind of problems can be found in Graham *et al.* (2003), Schafer & Graham (2002), Graham & Hofer (2000), Little & Rubin (1987) and Schafer (1997), to name few. Here the replacement of missing values was performed with values of similar complete cases. A subset of Z variables without missing data was selected and on this subset HCA with χ^2 distance and Ward's method was performed in order to exhibit proximities between consumers. On this basis, for each consumer who had missing data the answers of the closest consumer were used for filling them.

Typically, dummy variables are used in analysis of qualitative data, such as survey responses, because many models require another way to represent qualitative concepts to make sense. Here we performed the conversion of categorical consumer descriptors to dummy variables obtaining a new table $\widetilde{Z}$ (72 x 983).

$\widetilde{Z}$ data set tends to be too large especially after dummy transformation where number of variables grows at an exponential rate. In order to facilitate the task, the data set must be summarized somehow. Procedures for discarding or selecting variables based on statistical criterion have been proposed by Jolliffe (1972), McCabe (1984), Krzanowski (1987), Al-Kandari and Jolliffe (2001) or Guo *et al.* (2002) among many others.

Since, we aim to identify groups of consumers with a different background, the $\widetilde{Z}$ variables, most discriminant in that sense, were chosen. The selection of these discriminant variables was performed carrying out the following procedure. First of all, a HCA classification of consumers in terms of their characteristics, habits and opinions with χ^2 distance and Ward's method was performed on $\widetilde{Z}$ table. The resulting dendrogram, which suggests a solution with 3 clusters, is given in figure 4.2.

Figure 4.2 – HCA dendrogram which classifies 72 consumers (interberry data) according to their socio-demographic characteristics and consumption habits.



To characterize the 3 clusters, χ^2 test was used between the qualitative variable that indicates the belonging to the groups and each original Z variables. All the variables which showed a significant p-value (less than 0.05) in the χ^2 test for 3 groups were chosen as important. Then, a new Z table, containing only 132 modalities, will be considered.

4.3 CHEESE dataset

The dataset here considered is part of a study performed by IASMA on some of the most important cheese of the dairy production of Trentino. The main issues of this research are the characterisation of these typical products and the definition of key quality attributes to support, on one hand, quality control and on the other, the definition of typicality specifications necessary for their valorisation as PDO candidates (Protected Designation of Origin). The data here analysed are related to three typical cheeses from Trentino: *Puzzone di Moena* (PU), *Nostrano di Primiero* (PR) and *Nostrano di Campitello* (CA). These cheeses belong to the generic cheese typology *Nostrani trentini* (Gasperi *et al.*, 2004), and more precisely they are traditional cheeses from valleys of Fiemme, Fassa, and Primiero, called *washed-rind* cheeses. They are strictly made with raw bovine milk. During the ripening, these cheeses are literally wetted with salted water until the rind becomes yellow-ochre coloured. This treatment stimulates formation of micro-organisms responsible for a second fermentation, which considerably contribute to the flavours of the final product.

In order to investigate the possibility to support the request for a new PDO certification, these cheeses were firstly described through a sensory Quantitative Descriptive Analysis (Stone *et al.*, 1974), and secondly they were assessed by a consumer panel in terms of overall liking. Information about consumer characteristics and habits regarding cheese purchase and consumption were also collected. Three tables are thus obtained, which are described in details in the following.

4.3.1 Preference data

IASMA organized a central location test, aimed to evaluate consumer overall liking on typical cheeses and to introduce them to a larger audience in order to support typical production area. The three cheeses were proposed in an anonymous way, following a different random order for each consumer in order to minimize the presentation order effect on hedonic ratings and to avoid influencing consumers one another. A panel of 316 mountain cheese lovers was asked to score the three cheeses according to their degree of global appreciation on a 9 points scale (to 0 indicating extremely disliking and 9 extremely liking). Data are contained in a (3 x 316) table which from now on we will refer to as Y (Figure 4.3 (a)).

4.3.2 Sensory characteristics

The sensory panel comprised eight assessors selected for their ability to recognize, describe and quantify basic taste, simply odours and texture properties following the procedure for selection and training of hard and semi-hard cheese sensory assessors described by Gallerani *et al.* (2000).

The panel developed according to consensus method a profile protocol for QDA (Gasperi *et al*., 2004) containing 35 attributes listed in table 4.3. The intensity of each attribute was evaluated on 100 mm unstructured scale, anchored at each extreme. Six samples, two for each product, were presented in the session in an order balanced for assessor, product and presentation. The two replicates were mediated and panel means

Table 4.3 - List of sensory attributes measured by a descriptive panel: sensory definitions (anchor on a 100 mm unstructured scale) and codes (the first letter of the Code column identifies attribute group: V_ visual appearance, X_texture, T_taste, P_physical sensation in mouth, O_odour (by smelling) and A_aroma (by tasting).
* evaluated on a photo of a 10 cm x 5 cm sector of inner surface of cheese (by scanner)
** evaluated on the cheese sample

Visual appearance	Instruction: look the image of sample (*) or the real sample (**) and evaluate the following attributes	
Eyes, diameter	Average diameter of the eyes * (0 mm → 10 mm)	V_Diam
Eyes, number	Average percentage of surface related to the eyes on the total surface of the imagine * (0% → 25%)	V_Numb
Eyes, shape irregularity	Irregularity in the shape of eyes * (0=circle → high)	V_Shap
Eyes, distribution irregularity	Irregularity in the distribution of eyes in the surface* (low → high)	V_Dist
Colour intensity	Intensity of yellow competent in the surface colour ** (low → high)	V_Colo
Odour (By Smelling)	Smell the sample opening the sealed pot and after breaking of sample cuboid into two parts and evaluate the following attribute	
Milk	The odour associated with fresh cream (not perceptible → high)	O_Milk
Sour Milk	The odour associated with sour milk products (not fat yogurt) (not perceptible → high)	O_Sour
Boiled milk	The odour associated with boiled whole milk (boiled for 2 min) (not perceptible → high)	O_Boil
Rind	The odour associated with hard cheese rind (24 months ripened Parmigiano Reggiano) (not perceptible → high)	O_Rind
Cow	The odour associated with cattle (straw layer from cattle) (not perceptible → high)	O_Cow
Hard-boiled egg	The odour associated with the hard boiled egg white (egg boiled for 15 min) (not perceptible → high)	O_Egg
Ammonia	The odour associated with ammonia (fresh milk aromatized with 500 mg/l of NH_4OH) (not perceptible → high)	O_Ammo
Butyric	The odour associated with butyric acid (fresh milk aromatized with 250 mg/l of butyric acid) (not perceptible → high)	O_Buty
Texture	Instruction: Manipulate the sample (manual texture) or chew it (oral texture) and evaluate the following attribute:	
Surface moisture	The degree to witch the surface cube feels moist to the touch (between thumb and forefinger) (low/dry → high/wet)	X_Mois
Elasticity	The rapidity of recovering initial thickness after a deforming pressure with the thumb on the cube (low → high)	X_Elas
Deformability	The degree to with the sample cube can be compressed without breaking (after a deforming pressure on the sample cube with the thumb) (low → high)	X_Defo
Surface roughness on fracture	The degree to witch the inner surface (after break of the sample parallelepiped into two parts) appears rough or smooth (low/smooth → high/rough)	X_Roug
Firmness	The force requires to bite with molar the sample cube (first bite) (low/soft → high/hard)	X_Firm
Solubility	The ability of sample after 4 acts of chewing to melt in mouth (low → high)	X_Solu
Friability	The degree to which the sample break in particles (number of fragments after 4 acts of chewing) (low → high)	X_Fria
Granules	The degree of the consistency of partides in the mouth after 8 acts of chewing (low/mealy → high/grainy)	X_Gran
Adhesivity	The force required to unstick from teeth the chewed cheese (after 8 acts of chewing) (low → high)	X_Adhe
Taste and tactile sensation	Instruction: taste the sample and evaluate the following attribute:	
Sweet	The basic taste sensation related to fructose added in ricotta cheese (30 g/kg) (low → high)	T_Swee
Salty	The basic taste sensation related to sodium chloride added in ricotta cheese (6 g/kg) (low → high)	T_Salt
Acid	The basic taste sensation related to lactic acid added in ricotta cheese (5 g/kg) (low → high)	T_Acid
Bitter	The basic taste sensation related to caffeine added in ricotta cheese (0,2 g/kg) (low → high)	T_Bitt
Pungent	The burning sensation of tongue and mouth surface related to capsaicin added in ricotta cheese (20 mg/kg) (low → high)	P_Pung
Aroma (by tasting)	Instruction: taste the sample and evaluate the following attribute	
Milk	The flavour associated with fresh cream/fresh milk (pasteurized cream 35% fat) (not perceptible → high)	A_Milk
Sour Milk	The flavour associated with sour milk products (not fat natural yogurt) (not perceptible → high)	A_Sour
Rind	The odour associated with hard cheese rind (24 months ripened Parmigiano Reggiano) (not perceptible → high)	A_Rind
Cow	The flavour associate with cattle (straw layer from cattle) (not perceptible → high)	A_Cow
Hard-Boiled Egg	The flavour associated with hard boiled egg white (egg boiled for 15 min) (not perceptible → high)	A_Egg
Ammonia	The flavour associated with ammonia (fresh milk aromatized with 500 mg/kg of NH_4OH) (not perceptible → high)	A_Ammo
Butyric	The flavour associated with butyric acid (fresh milk aromatized with 250 mg/kg of butyric acid) (not perceptible → high)	A_Buty
Fruity	The flavour associated with different fruit (fresh milk aromatized with 500 mg/kg of ethyl hexanoate) (not perceptible → high)	A_Frui

were extracted as the last step. The final table comprises three products (rows) described by 35 attributes (columns) and from now on we will refer to as X (Figure 4.3 (b)).

4.3.3 Consumer background

At the end of the test, consumers were asked to fill a questionnaire concerning information considered relevant to investigate consumer behaviours and its purchase choices.

Figure 4.3 – Schematic representation of cheese database.

(a) preference data matrix **Y**

	Consumer preference scores		
	C1	…	C316
Prod1	Y (3 x 316)		
Prod2			
Prod3			

(b) product characteristics matrix **X**

	Visual	**Texture**	**Taste**	**Sensation**	**Odour**	**Aroma**
	V_diam…V_colo	X_mois…X_adhe	T_swee…T_bitt	P_pung	O_milk…O_buty	A_milk…A_frui
Prod1	X (3 x 35)					
Prod2						
Prod3						

(c) consumer information matrix **Z**

	Socio-demographic descriptors	**Cheese consumption and purchase habits**
	sex, age, school, members, province	dairy1…dairy13
C1	Z (316 x 18)	
C2		
…		
C335		

Consumer descriptor matrix, referred to as Z, contains 18 qualitative variables that may be synthesised as follows:

- 5 socio-demographic variables: sex, age class, school level, number of family members and province of origin;
- 13 variables [dairy1 … dairy13] describing consumption, purchase habits and opinions on cheese;

A detailed list of variables collected by the questionnaires is given in the appendix A. Data are contained in a (3 x 316) table Z (Figure 4.3 (c)). As we have done before, we performed the conversion of categorical consumer descriptors to dummy variables obtaining a new Z table (316 x 95).

4.4 Traditional approaches on Interberry data

Conjoint use of multidimensional preference mapping and cluster analysis is the most common approach of dealing with data from a preference study. Depending on the data analysed, two modes of preference mapping exit, known as *internal* and *external* analysis (Carroll, 1972). With Internal Preference Mapping (IPM) the objective is to achieve a multidimensional representation of the stimuli (products) based solely on the preference data, whilst in External Preference Mapping (EPM) the aim is to relate product acceptability to a multidimensional representation of products derived by chemical or sensory data.

The aim of this paragraph is to show the application of these traditional approaches on data presented in paragraph 4.3.1 in the case when preference is supposed to be driven by product characteristics (paragraph 4.4.1) and when it is supposed to depend on individual descriptors (paragraph 4.4.2).

4.4.1 External Preference Mapping

In order to understand product attributes that influence consumer preference, sensory analysts commonly follow a three step procedure, called External Preference Mapping (EPM):

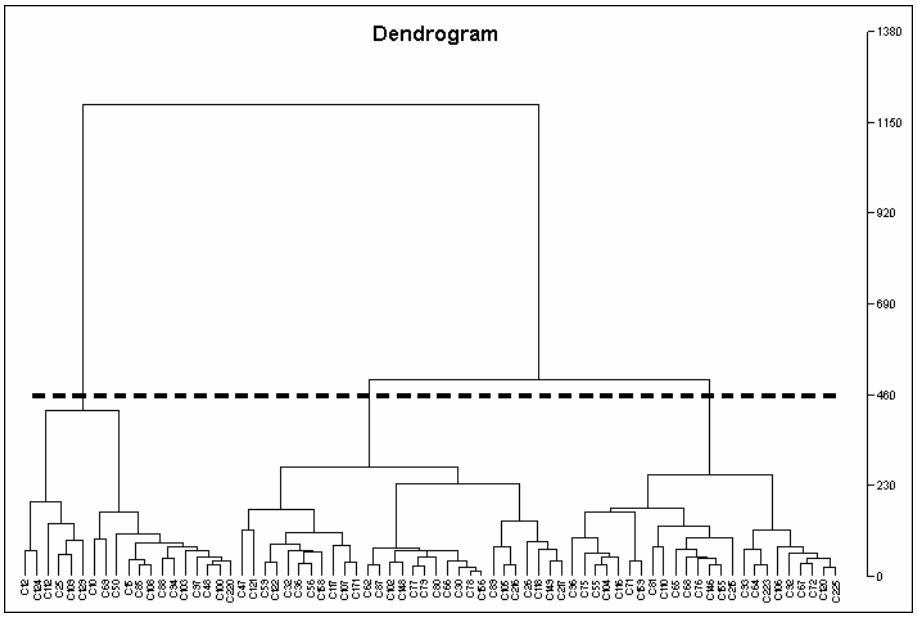
a) Firstly, in order to identify segments, a cluster analysis on preference data is computed on the basis of their global liking of the products;

b) The second step consists in mapping the products on the basis of their characteristics. For instance, this can be obtained by running a PCA on matrix X;

c) Thirdly, preference data for each individual are regressed onto principal components found in step b).

Methods of step b) and c) together can be also called PCR (Principal Component Regression; briefly introduced in paragraph 3.2.1), consisted in running a MLR (Multiple Linear Regression) on the principal components of X and modelling each Y-variable individually.

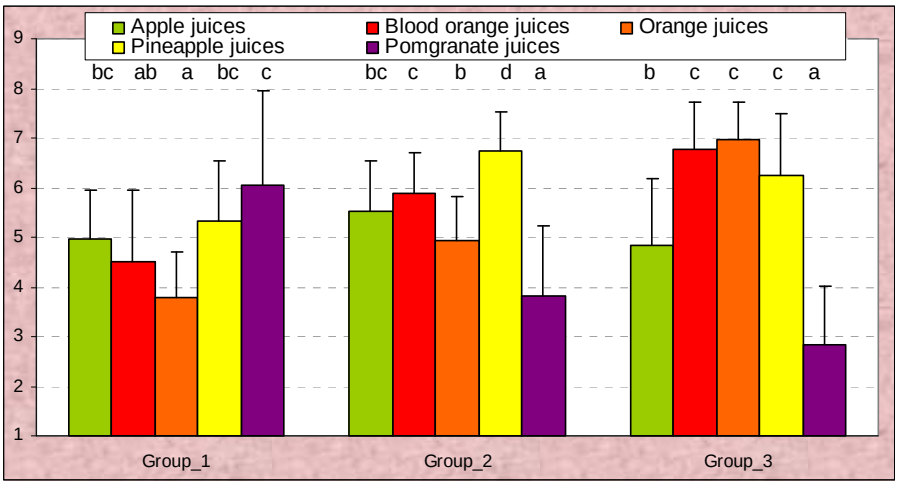
Figure 4.4 – Dendrogram of HCA classification of the 72 consumers according to their liking scores.



In this paragraph, the results of this procedure are given. First a HCA classification of consumers in terms of their liking rates on the products with Euclidean distance and Ward's method is performed on table Y. The resulting dendrogram, which suggests a solution with 3 clusters of 19, 29 and 24 consumers respectively, is given in figure 4.4.

The description of the 3 groups in terms of appreciated and not appreciated juices is given in figure 4.5. This description was obtained performing ANOVA in each group of consumers over preference scores of products with the same base fruit component. Subsequently, Tukey's HSD multi-comparison test was undertaken.

Figure 4.5 – Acceptability average scores and standard deviations for each group and base fruit component and multi-comparison test (bars with different letters are significantly different ($p<0,05$)).



Group 1 overall prefers juices based on pomegranate and dislikes mixes with orange. Group 2 and group 3 do not like pomegranate mixes but people collected in group 2 overall prefer pineapple juices, while people of the third group like mixes based on orange and blood orange juices. First two principal components from X, which explain 72% of the global variance, were used and results of PCR are given in figures 4.6 and 4.7. Juices based on orange and blood orange are rich in citric acid, mixes based on pomegranate have an high content of polyphenols and sugars, apple juices are rich in

malic and ascorbic acid, while pineapple mixes have a moderate content of all of these chemical compounds. Then, if the practitioner wants to know how the groups are built in terms of individual descriptors, χ^2 test will be used between the qualitative variable that indicates the belonging to the group and each individual characteristic, socio-demographic for instance.

Figure 4.6 – Bi-plot of PCR procedure.

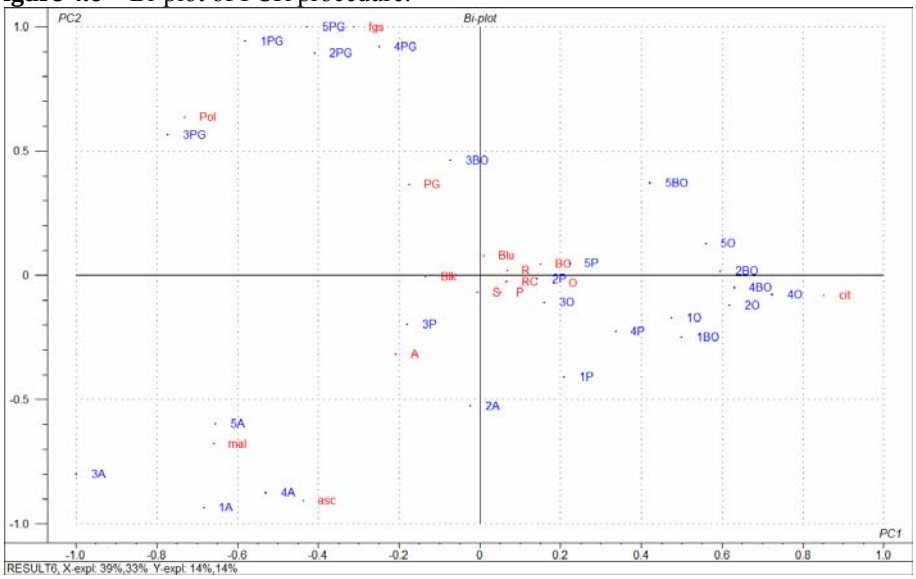
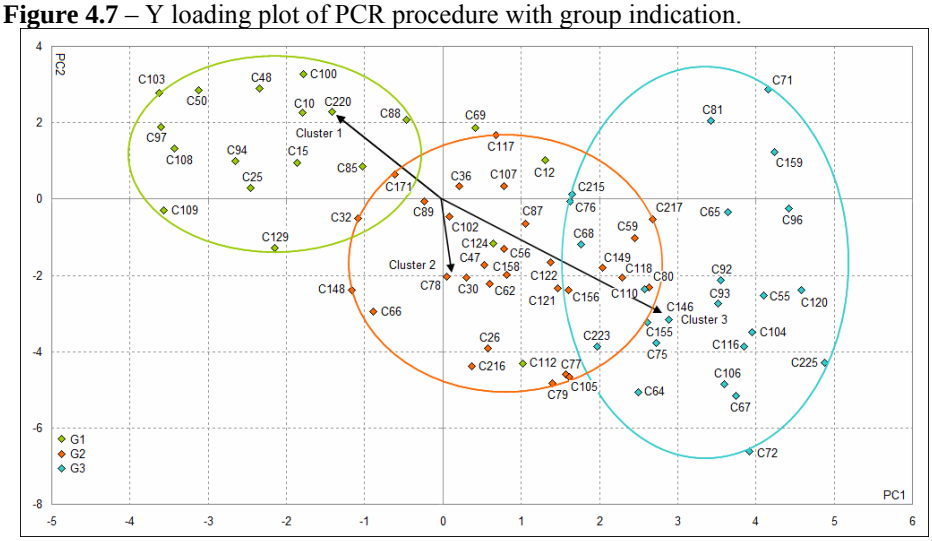


Figure 4.7 – Y loading plot of PCR procedure with group indication.



Only 3 descriptors show a significant p-value ($p<0.05$) characterizing the third group as the youngest one, with 92% of students and 100% of not married people (table

4.4). Consumer classification has been done just on preference data, therefore it is not surprising that group description in terms of socio-demographic characteristics is very poor.

Table 4.4 – Group description in terms of socio-demographic characteristics.

Descriptor	χ^2 p-value	G1 (n=19)	G2 (n=29)	G3 (n=24)
child	0.1915			
sex	0.4796			
age_cod	0.0157	53% < than 20 years 10% between 21-35 37% > than 36 years	52% < than 20 years 38% between 21-35 10% > than 36 years	71% < than 20 years 4% between 21-35 25% > than 36 years
status	0.0156	68% not married	82% not married	100% not married
school	0.5836			
smoke	0.7313			
staff	0.0285	58% students	65% students	92% students

4.4.2. Internal Preference Mapping on groups of consumers with similar socio-demographic characteristics

In analyzing consumer preference, assigning a main role to Z matrix consists in seeking groups of consumers in terms of individual descriptors before establishing any relationship with hedonic rates. Then, consumer preference of each segment is analyzed. With this aim, an IPM could be performed in each group in order to understand if groups with different socio-demographic characteristics have also different tastes.

IPM is basically a PCA of the matrix of hedonic scores across the products and the consumers (the matrix that we have called Y), which is carried out on data centred by consumer. The preference map is the resulting bi-plot which allows a graphical interpretation of individual preference.

A HCA classification of consumers on a subset of Z containing just individual demographic characteristics (called Z*) is performed. The resulting dendrogram that

suggest a solution with 2 segments of 39 and 33 consumers respectively, is given in figure 4.8.

The characterisation of the two groups in terms of individual descriptors is carried out using χ^2 test (table 4.5). For all the seven descriptors the two groups show a significant different distribution (p-value < 0.05), identifying two different demographic segments.

Figure 4.8 – HCA dendrogram from consumer classification on *Z** table.

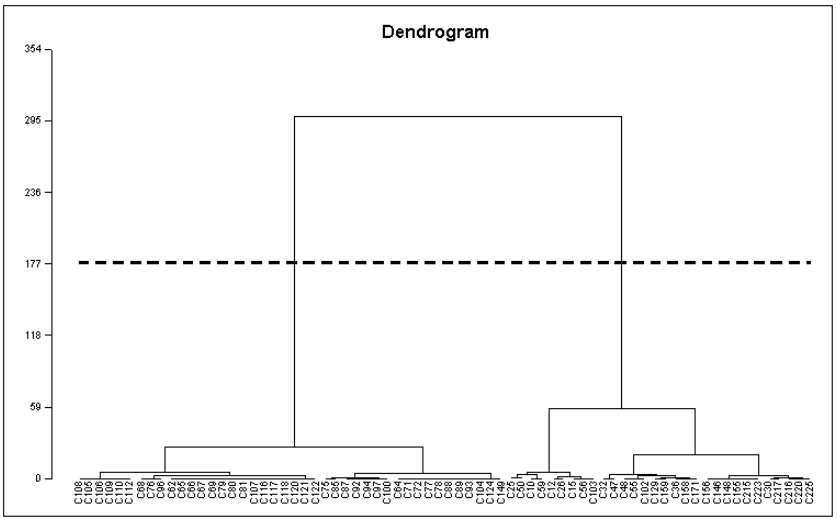


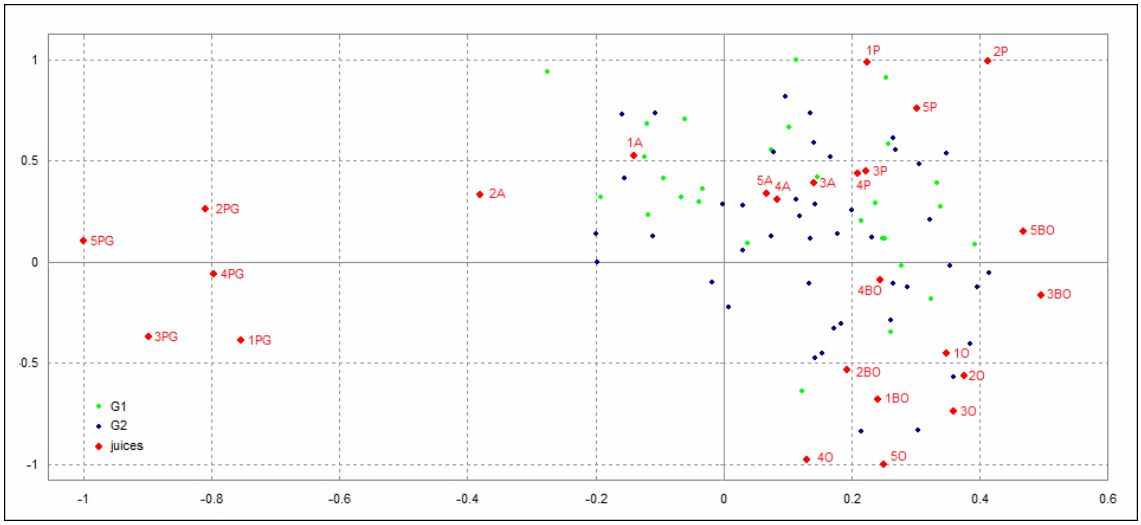
Table 4.5 – Group description in terms of individual characteristics.

Descriptor	χ^2 p-value	G1 (n=39)	G2 (n=33)
child	0.0015	100% without children	34% with at least 1 child
sex	0.0041	89% males	42% females
age_cod	< 0.0001	100% less than 20 years old	91% more than 21 years old
status	0.0003	100% not married	33% married
school	< 0.0001	100% lower educational level	36% school levels 3 and 4
smoke	0.0039	100% no smokers	24% smokers
staff	< 0.0001	100% students	61% IASMA staff

In figure 4.9 the preference map is given, where consumers of the two groups are depicted in two different colours (G1=green, G2= blue) and products are red-coloured.

It is clear that, groups identified on Z* data do not show different preference directions. In fact, socio-demographic clusters are nonuniform in terms of consumer preference and assigning an active role to both the matrices is needed to obtain more interpretable groups.

Figure 4.9 – Internal Preference Mapping: consumers of G1 are depicted in green, those of G2 in blue and products are red-coloured.



4.5 Two-way analyses on Interberry data

Here, the two alternative hypotheses, managed with classical approaches in the previous paragraph, will be dealt with CLV method. In paragraph 4.5.1 we will take the example of the first case, very common in sensory analysis, where information about the products are available together with preference data, while in the paragraph 4.5.2 we

will go through the second case more frequent in market research, where together with preference data there are also consumer characteristics.

4.5.1 Consumer preference clustering taking into account product information

The aim of this kind of analysis is to identify, in a panel of consumers asked to evaluate a certain number of products, subgroups of individuals which prefer similar products that have similar characteristics. Data used in this application are Y and X matrices described in paragraph 4.2. On this data-set CLV approach based on $\widetilde{S}$ criterion was applied. The evolution of $\widetilde{S}$ criterion (figure 4.10) clearly shows a partition in 4 groups. The results obtained with the permutation test are also displayed in figure 4.11 giving us the opportunity to explain once again how to read the graph and which partition has to be retained to the permutation test point of view.

At the first step of the classification, when the first two variables are collected in one group (in this case, when there are 71 clusters), the value of the aggregation criterion (red point in the graph) has to be lower than the 5^{th} percentile (blue in the graph) in order to conclude that observed data are different from those obtained with not correlated variables.

Once it has verified that there are groups in the data, we will conclude that there are k groups if the step k will be the first stage (reading the graph from left to right) where the value of aggregation criterion is less than the 5^{th} percentile. In this case a solution with 4 clusters has to be retained.

Before the execution of the partition algorithm these 4 groups contain 17, 13, 22 and 20 consumers respectively and the criterion is equal to 21.78. After the consolidation, the first group (G1) contains 18 consumers, the second (G2) 13, the third (G3) 18 and the last (G4) has 23 people, while the criterion is 21.81.

Figure 4.10 – Evolution of the aggregation criterion $\widetilde{S}$ when CLV approach was applied on Y matrix and X was considered as matrix of external variables.

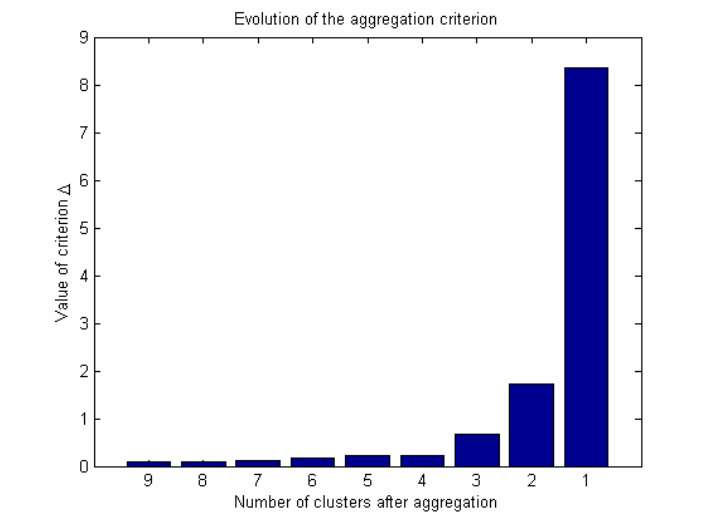
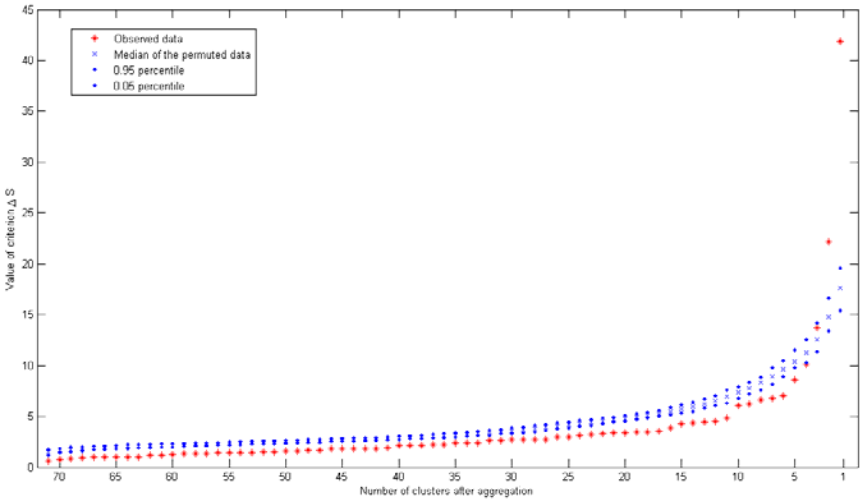


Figure 4.11 – Comparison of criterion values with those obtained in the permutation procedure.



Correlation matrix of group latent components, given in table 4.6 , shows how the first three latent variables describe different preference directions, while G4

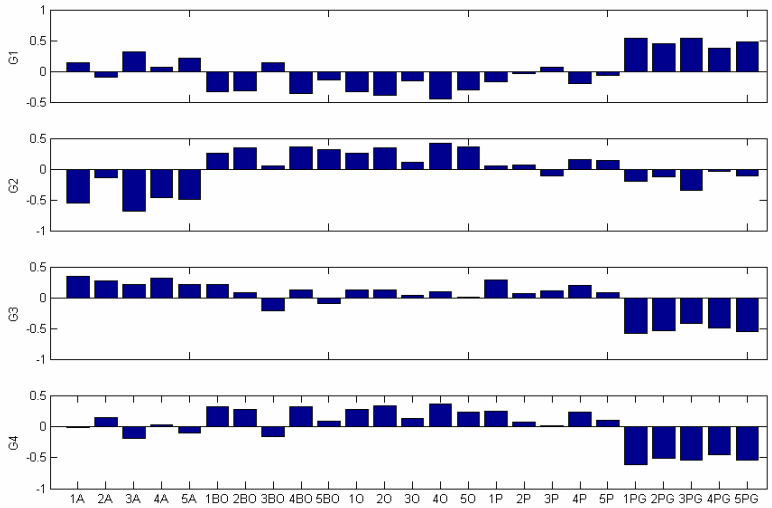
preference profile is negative correlated to G1 latent component and positively correlated to G2 and G3 latent components. RV and normalized RV coefficients (ZRV), for each group *k*, betweenY_k and c_k, are also given in table 4.4. The RV coefficients are all significant indicating that the RV observed values are not due to chance.

Table 4.6 – Indices of goodness of clustering .

	G1	G2	G3	G4
RV_k	0.557	0.432	0.548	0.674
ZRV_k	9.231	6.708	8.981	11.331
Correlation matrix				
	c1	**c2**	**c3**	**c4**
c1	1			
c2	-0.748	1		
c3	-0.666	0.022	1	
c4	-0.976	0.601	0.810	1

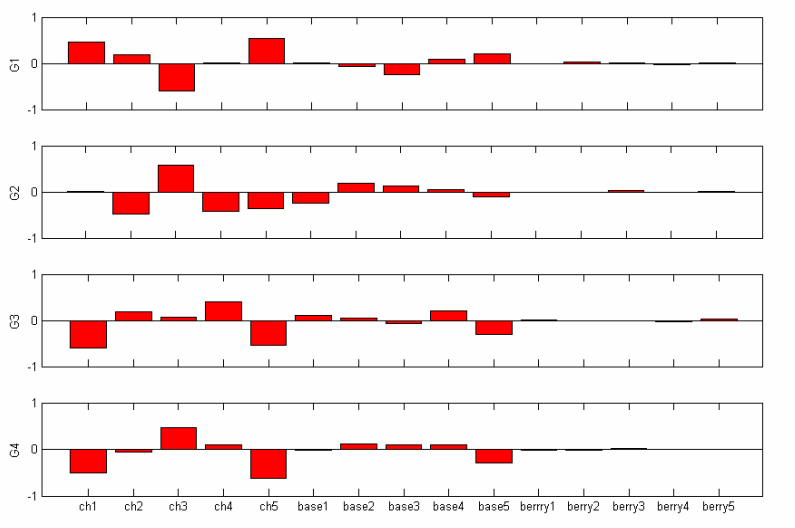
The description of the four groups is given in figures 4.12 and 4.13, which show values of latent components associated to the groups and values of loadings associated to the external variables respectively. The first group is essentially characterized by the fact that consumers belonging to it prefer juices based on pomegranate (PG), while do not appreciate those based on orange (O and BO), opposite to the others clusters.

Figure 4.12 – Latent variables c_k associated to the 4 groups.



Consumers in group 2 do not like juices based on apple (A) but they prefer those made of orange and are neutral towards pomegranate juices. The remaining two groups do not like absolutely juices based on pomegranate but, G3 consumers prefer juices with apple and pineapple (P) while those people collected in G4 appreciate mixes with orange.

Figure 4.13 – Loadings a_k associated to the **X** variables.



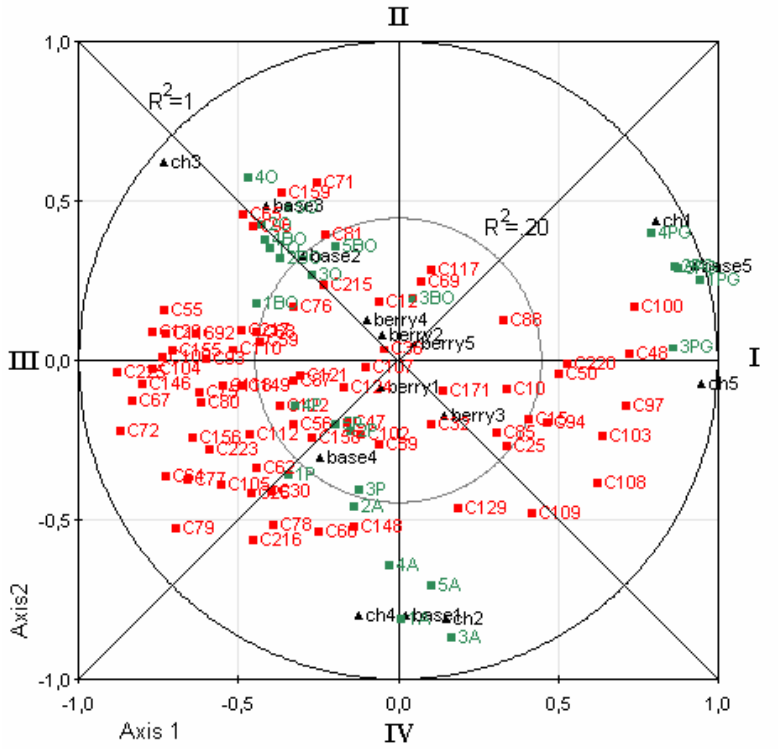
In figure 4.13 it can be seen that preference is not driven by the presence of berry juice: design variables, which indicate juice composition in terms of berry fruits (berry1, … , berry5), show loadings close to zero. Consumers in the first group like sweet (ch1) juices rich in antioxidants (ch5), the second group prefer sour products rich in citric acid (ch3) as orange juices, while the third group appreciate juices with a moderate content of ascorbic (ch4) and malic acid (ch2) like those based on apple and pineapple. Finally, the fourth group appreciate acid and not sweet juices.

4.5.1.1 A comparison with classical two-way PLS

In this paragraph, a two-way PLS regression estimated by SIMCA 11 (Umetrics, 2005) will be presented according to the procedure described in Tenenhaus *et al.* (2005), work chosen for the interesting way they clustered consumers. This application aims to highlight differences and similarities of the two techniques, even if they have very different objectives indeed.

In the berry juice example, the PLS regression of hedonic scores of the 72 consumers on the 15 standardized (as described in paragraph 4.2) chemical and design variables, using the first two components, leads to products*product descriptors*hedonic scores map shown in figure 4.14. This is the correlation circle between products, product descriptors and judgements variables and the first two components (t_1,t_2).

Figure 4.14 – Correlation circle of product (green), product characteristics (black) and judgments (red) variables with (t1,t2) and construction of clustering by means of the two bisectors (roman digits indicate the four groups).



The quality of this regression is given by the global R^2 between Y and (t_1,t_2), which is equal to 0.33 and by R^2 resulting from cross-validation, which amounts to 0.20.

In figure 4.14, some consumers are located in the centre of the graph showing that their preference cannot be explained by the model. In Tenenhaus *et al.* (2005) these consumers were not taken into account in the clustering: they decided, somewhat arbitrarily, to eliminate from the analysis those consumers j that have a $R^2(Y_j;t_1,t_2) = corr(Y_j;t_1)^2 + corr(Y_j;t_2)^2 \leq 0.20$. With the same criterion, also less explicative product descriptors have been eliminated. In our example, 23 consumers (32% of the panel) have to be excluded by further analyses. In addition, the design variables describing juice composition in terms of berry fruit (S, R, Blk, RC, Blu) and presence of blood orange (BO) and pineapple (P) are positioned in the centre of the circle. That suggests their information play no role on preference aspect of juices, as found with CLV (figure 4.13).

Subsequently, clustering of the remaining 49 consumers is performed according to Tenenhaus *et al.* procedure: the plan of figure 4.14 is divided into four areas bordered by two bisectors obtaining four groups, where the weight of each judge is proportional to its correlation with the PLS components. Then, PLS regression has been rerun for each group.

It is clear that this very simple approach does not aim to identify homogeneous groups of consumers in terms of hedonic scores and product characteristic, but it consists in describing the PLS components.

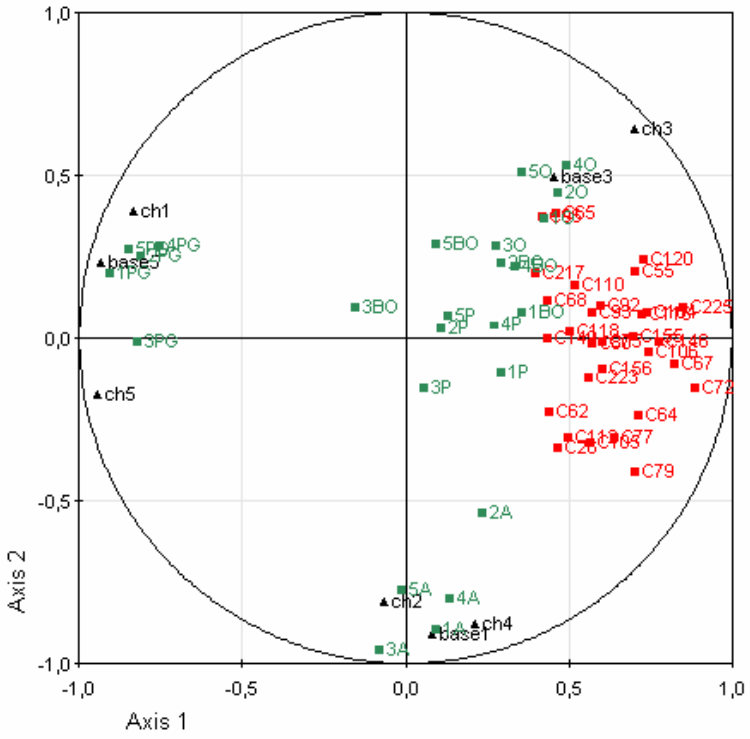
The cross-tabulation of the partitions obtained by using CLV or Tenenhaus *et al*. approach (table 4.7) clearly shows the difference between the two strategies.

Table 4.7 – Cross tabulation of group assignment in CLV and bisectors method in PLS. G_{EXCL} is the group of excluded consumers

		bisectors method in PLS					
		G I	G II	G III	G IV	G_{EXCL}	Tot
CLV method	G1	9	0	0	2	7	18
	G2	0	3	3	0	7	13
	G3	0	0	8	5	5	18
	G4	0	0	19	0	4	23
	Tot	9	3	30	7	23	72

Nevertheless, group construction according to Tenenhaus *et al*. provides clusters where PLS components represent more or less an average of the standardized assessments of the consumers in the group k and where corr(Y_j;t_k) are very close and positive. Therefore, PLS components can be considered as a typical consumer that in someway summarize each group. This concept coincide with the idea, behind CLV, that group latent components can be seen as local preference dimensions.

Figure 4.15 – Analysis of the group III, correlation circle of product (green), product characteristics (black) and judgments (red) variables with (t1,t2).



By way of example, the results of the analysis of the third group will be presented. In figure 4.15 the variable map in the PLS regression of the 30 consumers on the 8 remained product characteristics, is depicted.

The quality of the model with two components is noticeably higher than the global one: R^2 is equal to 0.433 and the cross-validated index is equal to 0.327. This group prefers juices based on orange and blood orange, which as everyone knows are rich in citric acid (ch3) and it rejects juices based on pomegranate, characterized by a high content in polyphenols (ch5) and sugars (ch1). These information can be taken out of score and loading plots (figure 4.16 and 4.17 respectively) as we do in CLV approach.

Figure 4.16 – Score plot of two-way PLS (interberry data).

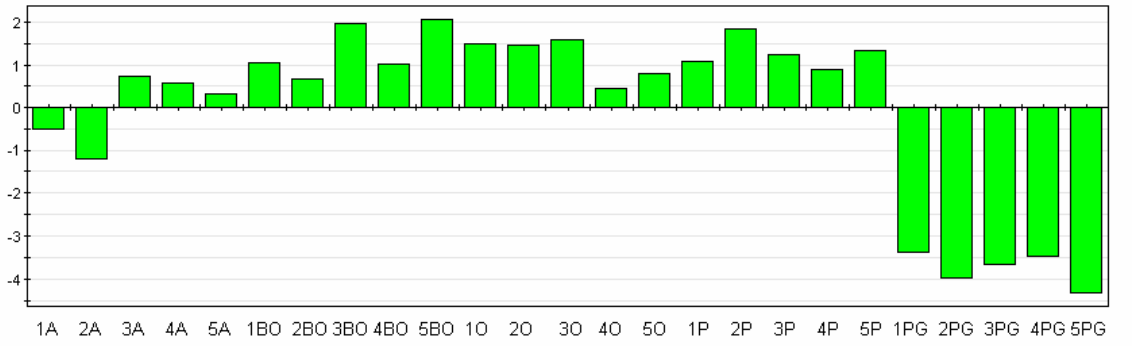
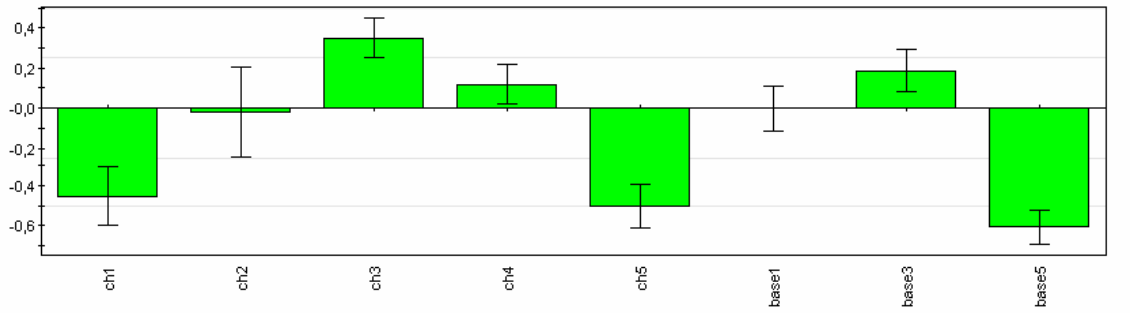


Figure 4.17 – Loading plot of two-way PLS (interberry data).

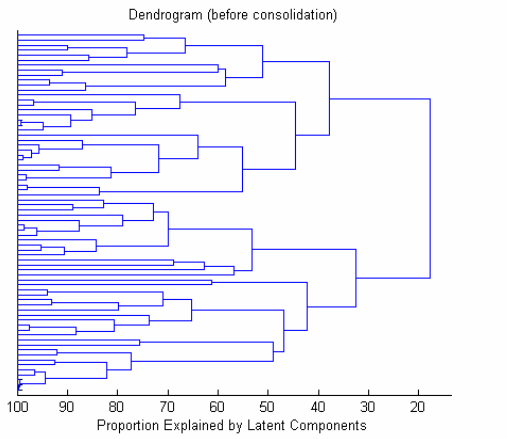


4.5.2 Consumer preference clustering taking into account consumer descriptors

Seeking segments of consumers similar in preference and in demographic characteristics, consumption and opinion towards a specific kind of products is generally performed in the framework of market research, in order to adopt an appropriate marketing strategy. In the following application we used data, introduced in paragraph 4.2: preference data Y matrix (preliminary centred) and a subset of $\widetilde{Z}$ containing just individual demographic characteristics (called Z*) in order to simplify method illustration. To $Y* = \left[\frac{Y}{Z*'}\right]$ was applied CLV method in the case when negative correlation means disagreement and without taking into account external variables according to Q criterion.

The dendrogram (figure 4.18), which illustrate the arrangement of the groups produced by the hierarchical clustering algorithm, clearly shows a partition in to two groups. Before the execution of the partitioning algorithm these two groups contain 39 and 33 consumers respectively and the value of the criterion is 57.18.

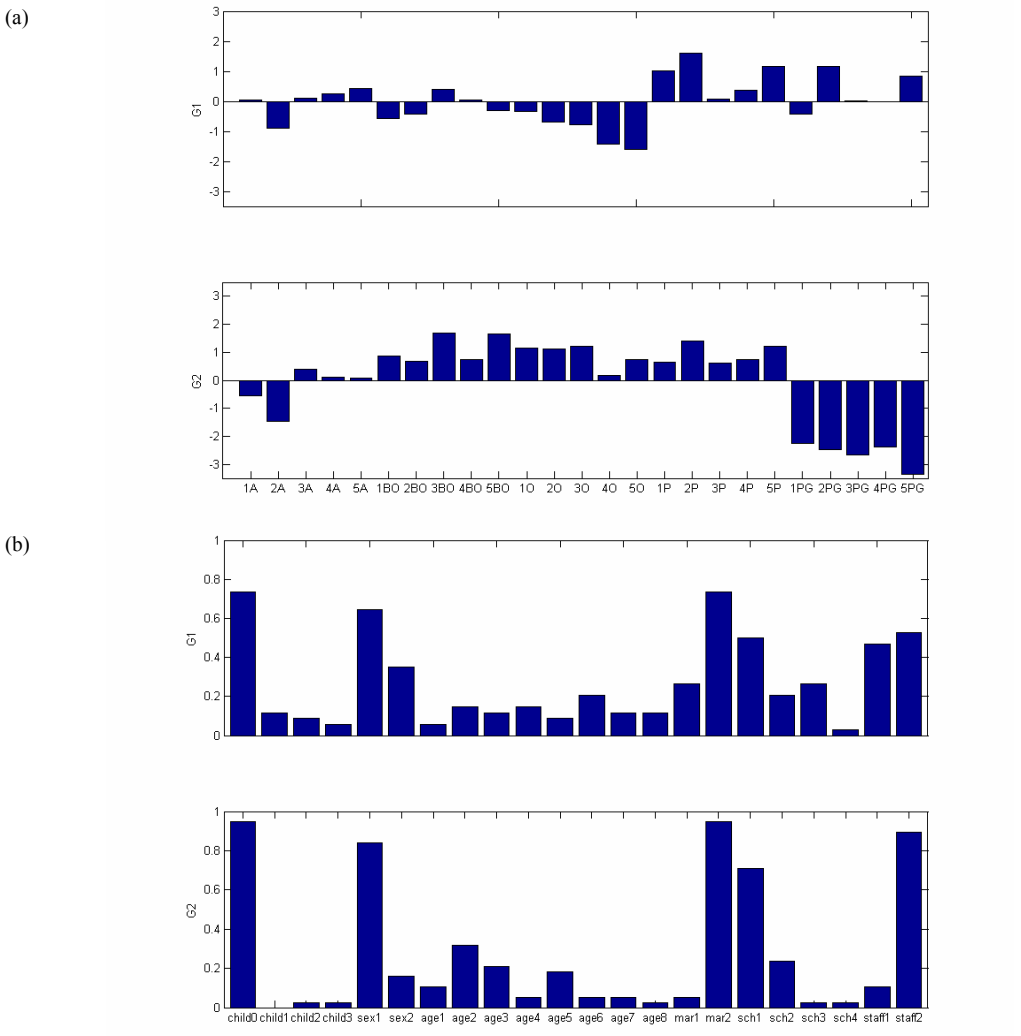
Figure 4.18 – Dendrogram of CLV ascendant hierarchical classification performed on Y*



After consolidation, the first group (G1) has 34 consumers (47%) while the second (G2) 38 (53%) and the criterion is equal to 60.20. The two latent components, which explain 34.14% of the global variance, describe two different profile (r = -0.017) and they are well associated to the individual consumer characteristics collected in the groups (RV_1=0.661 and RV_2=0.943, both significant).

Latent variables associated to the groups are depicted in figure 4.19: on the left side there is product description (a) and on the right the profile for each group in terms of consumer demographic characteristics is shown (b).

Figure 4.19 – Latent components c_k associated to the 2 groups: (a) product descriptors and (b) demographic characteristics.



The first group does not appreciate juices based on orange and is almost neutral towards those with blood orange as main ingredient, while it prefers some mixes based on pineapple and pomegranate.

The second group strongly rejects all the mixes with pomegranate and few based on apple, while it likes orange and blood orange juices.

Regarding demographic characteristics, the first group presents more married people (mar1) with at least one child (child1 – child3) than the second cluster and then more people that are more than 26 (age6 - age8). This group collects both students (staff2) and members of IASMA staff (staff1), consequently it is the most cultured cluster (sch3).

Second group is composed of the majority of male students of secondary school, aged from 14 to 20 years (age1 – age5), then, it is not surprising that they are not married and that the school level is less than that in G1.

It could be pointed out that, the c_k coefficient of a given modality of Z is assessed as the proportion of consumers in group G_k who have *chosen* this modality. The sum of these coefficients (proportions) in a group *k*, for all the modalities coding an item in the questionnaires, is one.

4.6 Analysis for L-structured data

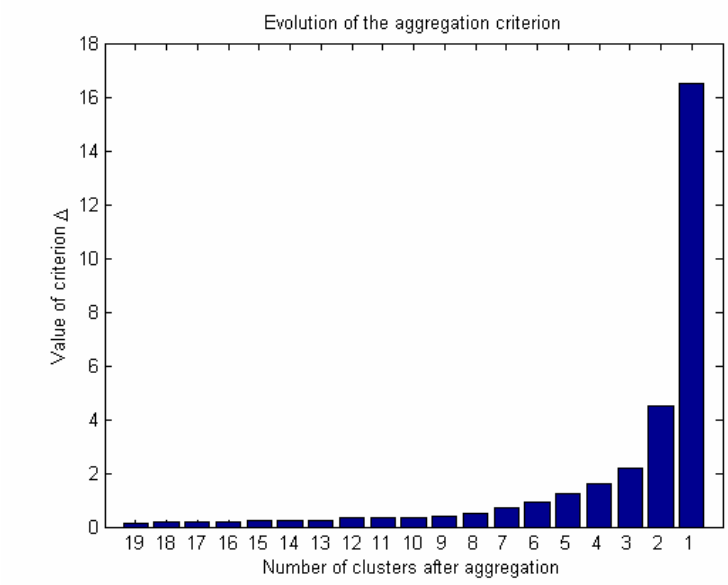
Here a new procedure, applicable to CLV approach and described in the previous chapter (paragraph 3.3), is illustrated using a real data-set in order to demonstrate that is possible to cluster a panel of consumers taking into account all the information at disposal.

At the first step, the matrix product between the matrix containing preference scores Y and $\widetilde{Z}$ matrix of consumer characteristics is computed. In this $Y\widetilde{Z}$ matrix, each element is the sum of the hedonic ratings for a given product (row) evaluated by all the consumers which belong to a specific $\widetilde{Z}$ category (column). In the second step CLV approach based on $\widetilde{S}$ criterion, is performed on $\widetilde{Y}$, obtained from the juxtaposition of Y and $Y\widetilde{Z}$, and taking into account X matrix of product descriptors as external data.

4.6.1 Interberry data

The evolution of the aggregation criterion $\Delta = \widetilde{S}_{i-1} - \widetilde{S}_i$ in the course of the hierarchy is given in figure 4.20. At each step, it results that the criterion $\widetilde{S}$ does not change to much passing from seven to three clusters solution, but the loss in the quality of the partition is more important passing from three to two clusters.

Figure 4.20 – Evolution of evolution criterion $\Delta = \widetilde{S}_{i-1} - \widetilde{S}_i$



This loss is grater when the two last groups are merged together. Therefore, a partition into three clusters is retained.

The partition algorithm converges after only three iterations to a solution in three segments of 21, 20 and 31 consumers and 16, 40 and 77 $\widetilde{YZ}$ variables respectively. The criterion is equal to 481.1 for a global variation accounted for of 97%.

The global variance of the CLV classification is computed like percent ratio between the value of the criterion at step k and the initial criterion value.A way to assess whether a subset of variables reflects a common feature of relevance to the problem under study is to calculate the sum of squares of the Y explained by the first PLS1 component from PLS regression of $\bar{y}_k$ on X (last line in table 4.9).

In the table 4.8 information about the three groups are partially collected, while in table 4.9 is given the correlation matrix of group latent components, where it can be seen how similar are the latent components of G2 and G3.

Table 4.8 – Description of the three groups, where nj indicates the number of consumers and nv the number of $\widetilde{YZ}$ characteristics. A partial list of $\widetilde{Z}$ characteristics is also reported.

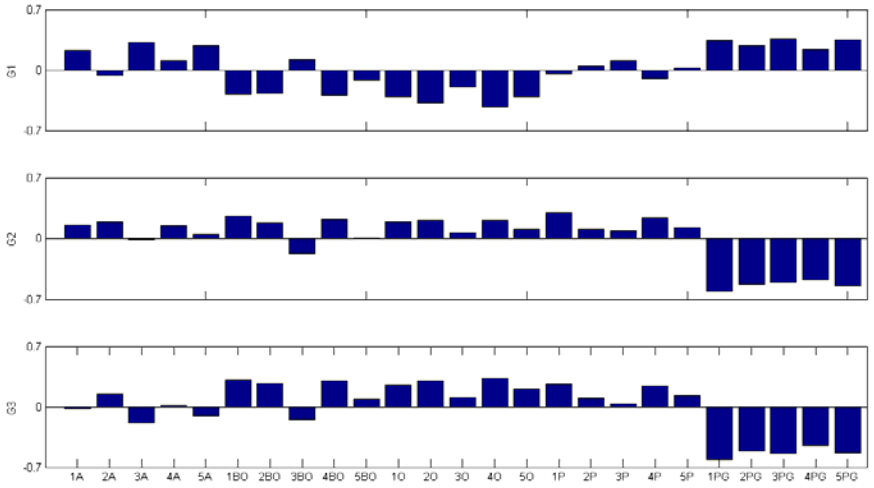
	G1	**G2**	**G3**
nj	21	20	31
nv	16	40	77
list of consumer characteristics	status_1	age_5	staff_2
	child_2	school_2	age_2
	child_4	fruit8_4	age_3
	fruit3_1	fruit3_2	school_1
	RE_1_5	juice8_1	fruit10_3
	Fns10_9	juice9_1	juice1_2
	staff_1	juice11_1	juice2_2
	age_7	juice15_1	juice26_3
	school_3	juice17_1	H14_3
	Bf5_5	juice23_1	H19_1
	…	…	…

Regarding the goodness of clustering the RV_k coefficients indicate a good association between the preferences of the consumers in the group (Y part information in $\widetilde{Y}$) and the relative latent factor (RV_1=0.556, RV_2=0.510 and RV_3=0.634 all significant) but the coefficients, which evaluate the association to the variables derived from consumer descriptors ($Y\widetilde{Z}$ part information in $\widetilde{Y}$), are not so good especially for the smallest group (RV_1=0.110 (not significant), RV_2=0.347 and RV_3=0.451).

Table 4.9 – Correlation matrix of group latent components.

	c_1	c_2	c_3
c_1	1		
c_2	-0.726	1	
c_3	-0.878	0.961	1
$R^2(Y)$	0.461	0.549	0.632

Figure 4.21 – Latent components c_k associated to the 3 groups.

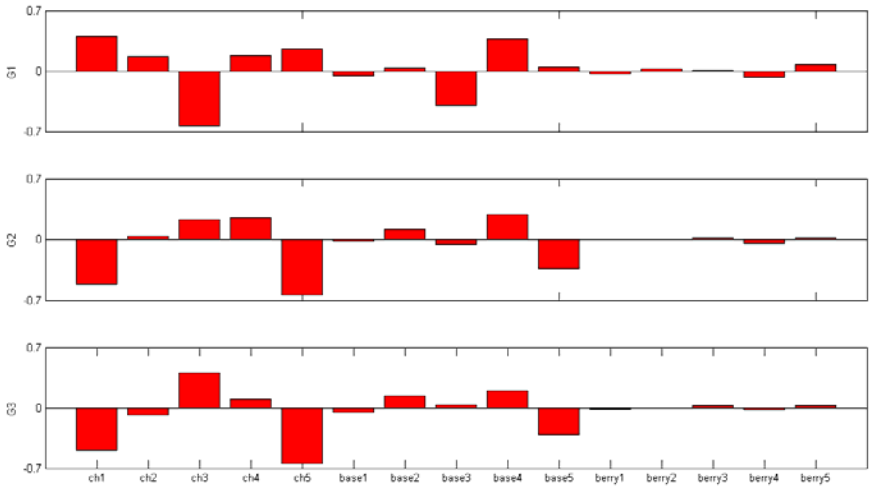


Thanks to the assignment in each segments of a certain number of $\widetilde{Z}$ categories, it is possible to give a description for each segment in terms of socio-demographics descriptors, purchase behaviours, habits and opinion as well as the product characterization achieved using latent components (figure 4.21) and associated X loadings (figure 4.22).

Consumers in the first group do not reject juices based on pomegranate (PG) nor apple (A), while they do not like mixes with orange of blood orange (O, BO). In fact, these people do not appreciate products rich in citric acid (ch3) as orange juices are. Latent components c_2 and c_3 seem to describe two similar preference directions. G2 and G3, on the contrary of G1, do not reject orange juices while absolutely do not like pomegranate mixes that are rich in sugar content (ch1) and antioxidants (ch5).

However, juices based on apple, characterized by an higher, content of malic acid (ch2) are more appreciated by consumers in the second group than in the third. There is not difference among the 3 groups in terms of liking of juices based on pineapple (base 4), well accepted by everyone. Furthermore, loadings of berry fruit design variables are close to zero, which means that there is not a berry fruit effect. It is not surprising that preference is not driven by the ingredient that makes up just the 20% of the juice formulation.

Figure 4.22 – Loadings a_k associated to the **X** variables.



Regarding to the individual descriptors the three groups are well separated. To the first group sixteen characteristics are been assigned, but only four of them are

enough correlated to G1 latent component ($r \geq 0.35$). Correlation coefficients between Z characteristics (via $\widetilde{YZ}$ variables) of each group and the associated latent component are given in appendix B. The consumers, associated to the product profile characteristic of this group, are married people (status_1) with at least one child (child_2 and child_4), that do not eat fruit as snack (fruit3_1) but regularly eat ready to eat salad (RE_1_5) and like to try new ethnic food (Fns10_9). They are members of the IASMA staff (staff_1), aged from 36 to 45 (age_7), with a high school level (school_3), who go to the supermarket themselves, where they buy berry fruits (Bf5_5). Older consumers are more accepting of unusual taste of juices based on pomegranate, maybe because there are exposed to a wide variety of different aromas and flavours over time compared to younger consumers (Luckow and Delahunty, 2004). In addition, group 1 contain a large part of people with high educational level, that has been shown to be related to a low food neophobia and an aptitude for not common flavours (Tuorila *et al.*, 2001).

Forty modalities are associated to the second group, thirty-six of them are enough correlated ($r \geq 0.35$) to latent component c_2. Consumers, which prefer products with a low content of sugars and polyphenols and a moderate content of citric and ascorbic acid, are aged from 21 to 25 years (age_cod_5) and show a middle school level (school2). They often buy fruit at the market (fruit8_4), occasionally eat it as snack (fruit3_2). They are not great juice drinkers, in fact, they drink less than one glass per week at the main meals (juice8_1, juice9_1, juice11_1, juice11_2, juice15_1 and juice17_1). However, they prefer enriched juices to the natural ones (juice23_1), especially those based on orange (juice32_7). They try new food only in special occasions (Fns6_9) and eat regularly just familiar food (FF_1_5). In addition, they often buy drinks based on strawberry (Bf26_4) even if they think that among the soft fruits

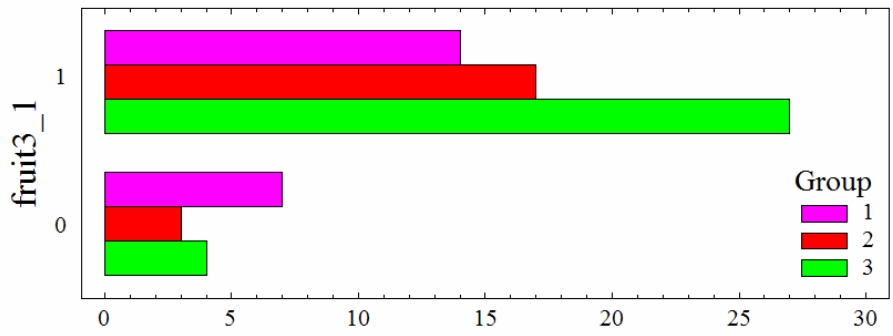
strawberries are the most “chemical” (H4_1) and they think that berry fruits have to be sweet (Bf40_9). They admit they do not know what free radical, antocyanins and antioxidants in general are (H17_2, H18_2, H20_2).

To the third group seventy-seven characteristics are been assigned and only one of them is not enough correlated to latent component c_3 (juice14_2). The consumers, associated to the product profile characteristic of this group, are students (staff_2). Consequently, it is a young group, consisting of teenagers (age_cod2 and 3) with a lower school level (school_1) in comparison with the other two groups. They are “easy” consumers that do not do the shopping themselves, even if they say that they very often buy fruit directly from the producers (fruit10_3). Actually, they are students in an agricultural institute, generally attended by farmers’ children, therefore they generally consume their own products. They are not aware of health aspects: they do not read the label (juice1_2 and juice2_2) when they purchase a juice and do not think that a juice has to be healthy (juice26_3). Furthermore, they do not think that fruit colour is related to nutritional aspects (H14_3), even if they say that they know what polyphenols (H19_1) and antocyanins (H20_1) are. For this reason, we can say that their notions about antioxidants are in general poor. Less known fruit (fruit26_1) are less appreciated by this group as shown by preference on tested juices. Age play a significant role in the acceptance of food products, in effect, this group composed of younger people has never tried new food, like high digestive tolerance milk (NF_4_2) nor ethnic food like soja sauce (EF_2_1), couscous (EF_3_1) and avocado (EF_4_1). In addition, they always eat berry fruit as fruit salad (Bf13_5) but they never buy berry fruit jam (Bf21_1).

As mentioned in the previous chapter, when a consumer descriptor is associated to a group it does not mean that its frequency in that cluster is higher than that in the others, but that the specific preference profile for this descriptor is associated to the profile of products characteristic of that group. For instance, we can observe the assignment of fruit3_1, a binary variable that is equal to 1 when fruit is never eaten as snack and 0 otherwise. In figure 4.23 frequency distribution of this consumer descriptor over the three groups is given.

It can be seen that this characteristic, assigned to the first group, presents higher frequency in the third group. Group assignment of $\widetilde{Z}$ variables does not simply reflect the frequency that different categories assume in each group. In fact, information from $\widetilde{Z}$, which is involved in $Y\widetilde{Z}$, represents the preference profile associated to each modality of Z. Such a profile is estimated by the sum of the preference values given by all the consumers who have chosen the modality.

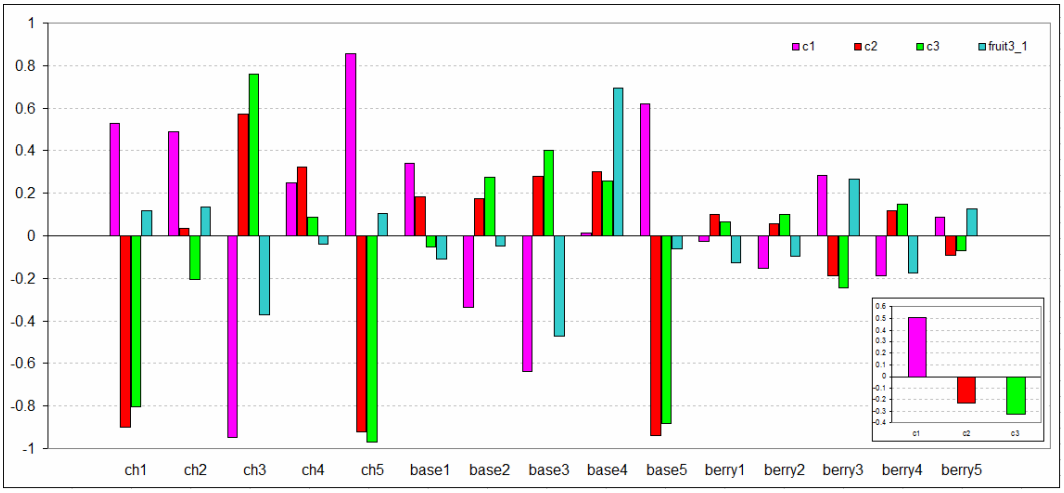
Figure 4.23 – Frequency distribution over the 3 groups for fruit3_1.



In figure 4.24 correlation between X variables and the preference profile of characteristic fruit3_1 (the column of $Y\widetilde{Z}$ that represent that modality) is depicted, as well as those between X variables and the latent components associated to each group (c_1,c_2 and c_3).

It can be seen that for all the X variables, except for ch4, base1 and base5, the preference profile of the descriptor follows the trend of c_1, which is the most correlated with fruit3_1 (square on the bottom right of the picture).

Figure 4.24 – Correlation between X variables and fruit3_1 (in blue) and group latent components (c_1 in pink ,c_2 in red and c_3 in green). Global correlation between fruit3_1 and each c_k is shown.



4.6.1.1 A comparison with two-step PLS for L-structured data

In this paragraph, a two-step PLS regression will be performed on interberry data in accordance with the procedure illustrated in Esposito Vinzi et al. (2007), briefly described in paragraph 3.2.2. In the first part of this method a PLS2 regression of Y (preference judgments) on X (product descriptors) is undertaken, as that performed in paragraph 4.5.1.1. Thirteen components are estimated and retained as significant, according to the two R^2 validation criteria implemented in SIMCA. Table 4.10 (a) contains cumulated $R^2(X)$ and $R^2(Y)$ for each component.

In order to increase $R^2(Y)$ on the first components a selection of Y variables was taken: in accordance with the criterion explained in paragraph 4.5.1.1, 23 consumers have been excluded, and a new model has been estimated. Exclusion of consumers with

a not predictable behaviour on the basis of our explanatory variables improved the model explicative power: the first two components now account for 44% of Y variability (table 4.10 (b)).

Table 4.10 - Cumulated X and Y R^2 for PLS-R step1: all Y-variables (a) and selected Y-variables (b).

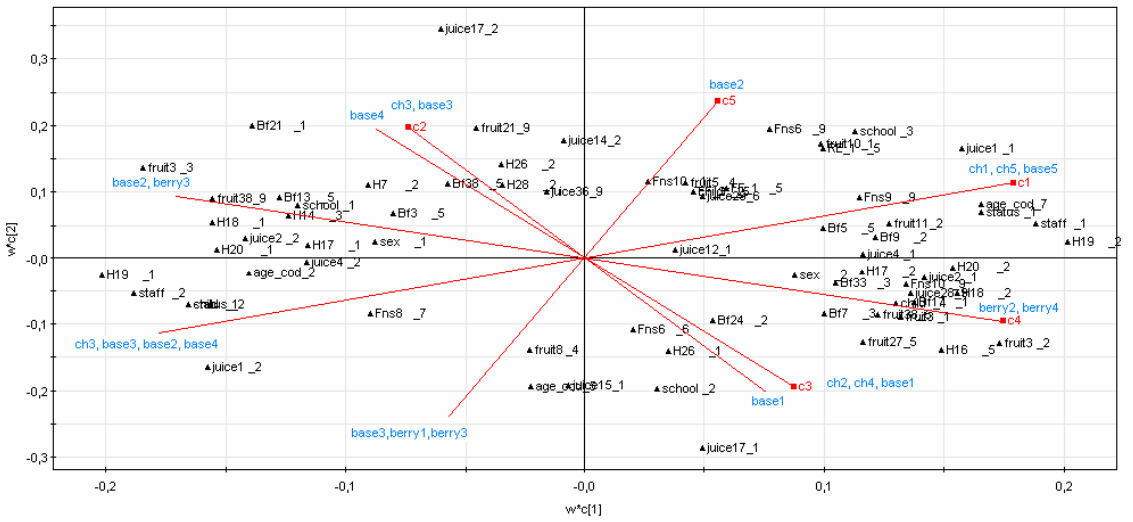
	All Y variables (a)		**Selected Y variables (b)**	
Components	**R^2X(cum)**	**R^2Y(cum)**	**R^2X(cum)**	**R^2Y(cum)**
1	0.223	0.254	0.222	0.345
2	0.431	0.328	0.434	0.436
3	0.535	0.405	0.536	0.513
4	0.624	0.452	0.624	0.564
5	0.717	0.482	0.714	0.590
6	0.805	0.506	0.799	0.613
7	0.890	0.528	0.887	0.631
8	0.973	0.548	0.973	0.648
9	0.987	0.608	0.987	0.690
10	0.995	0.644	0.995	0.720
11	0.998	0.687	0.998	0.753
12	1	0.717	1	0.775
13	1	0.749	1	0.800

In step 2 a PLS2 model has been undertaken, where the dependent variables are the first 5 regression coefficients from step 1 (C_x), while independent variables are consumer characteristics (Z). A first model was estimated, where all variables were retained. Just one component was significant. In order to improve $R^2(C_x)$, only 62 Z variables showing a VIP (Variable Importance in Projection) higher than 0.8 were kept in the model. A new regression was estimated, and two components were retained as

significant (R^2(Cx)cum = 0.277). Figure 4.25 shows the complete model on which variables characterising factors from step1 model is represented.

The first axis is characterized positively by H19_2, staff_1, age_7, status_1, fruit3_2 and juice1_1 and negatively by H19_1, staff_2, fruit3_3, status_2 and juice1_2.

Figure 4.25 – Global representation of step 2 model.



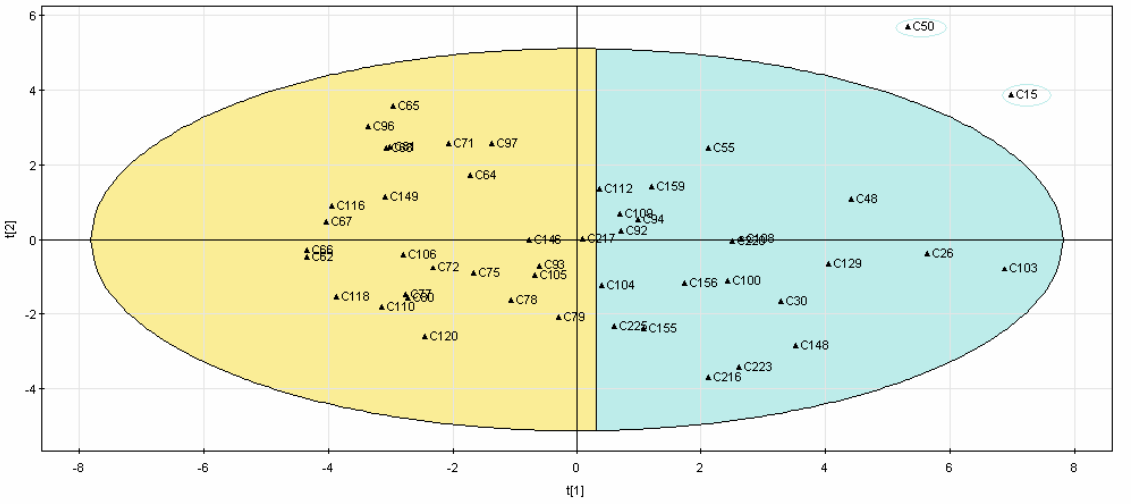
By looking at the real meaning of the variables, individuals on the right part of the graph are expected to be members of IASMA staff, married and aged from 36 to 45, that occasionally eat fruit as snack. In addition, they are “aware” consumers which read the label when they buy a juice even if they do not know polyphenols. People on the left side of axis 1 are expected to be not married students, that often eat fruit as snack, that do not read labels and that say to know polyphenols.

The second axis is characterized positively by juice17_2 and negatively by juice17_1 and then it can be interpreted as indicator of squeezed juices consumption: individuals on the upper side of the graph are expected to drink up to 3 glasses per week, while people on the lower side drink less then one glass per week.

The introduction of variables from step1 (blue in the figure 4.25) allows a global view of the phenomenon.

The graphical representation of the consumers on the same plane is shown in figure 4.26. Consumers are dispersed in a rather homogeneous way along all the map, and no cluster of individuals can be easily retrieved.

Figure 4.26 – Representation of consumers in step 2 model and group indication (G1 blue, G2 orange).



It may be remarked that there are two individuals C15 and C50, who may very likely be outliers, being outside Hotelling's T^2 ellipse. According to the clustering procedure followed by Esposito Vinzi et al. (2007), a *k-means* classification has been performed on the 48 consumers considered on the columns of T_z, where the number of groups were previously defined by an ascendant hierarchical classification. Two groups were individualized: G1 contains 23 consumers (blue in figure 4.20) while G2 collects 26 consumers (orange). Then new two-step PLS regression should be estimated on each group.

In comparison with CLV approach, this procedure seems quite cumbersome, and whenever the main goal of the researcher is consumer clustering it could be considered questionable to eliminate those consumers which behaviours cannot explain by the model.

Table 4.11 shows how the consumers are assigned to the groups in two-step CLV and k-means method in two-step PLS, respectively. First CLV group is split into two, one part goes in G1 of PLS method while the rest is collected in the group of excluded consumers. Second CLV group is split in three homogeneous parts, each of those is assigned to each PLS group. Finally, G2 in PLS method reflects the composition of the third CLV group, in fact three-fifths of it are contained in that group.

Table 4.11 – Cross tabulation of group assignment in 2-step CLV and k-means method in 2-step PLS. G_{EXCL} is the group of excluded consumers.

		k-means method in PLS			
		G1	G2	G_{EXCL}	Tot
CLV method	G1	10	1	10	21
	G2	7	7	6	20
	G3	6	18	7	31
	Tot	23	26	23	72

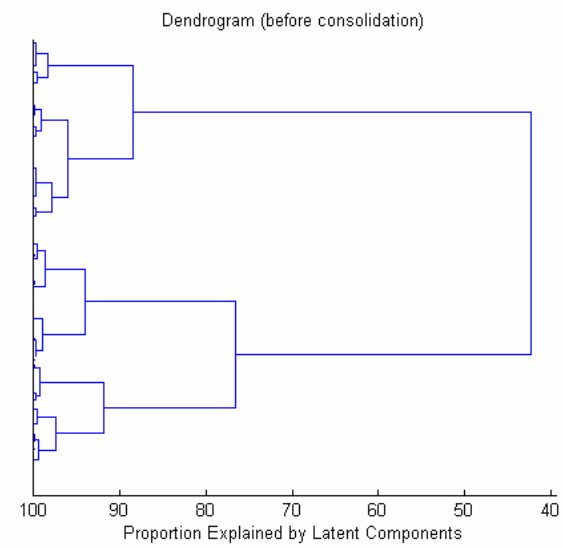
4.6.2 Cheese data

A second real data-set is here introduced to show the application of two-step L-CLV on matrices with a larger number of variables (consumers) but a very small number of observations (products). This example deals with sensory and preference data from a study on three cheeses (data described in paragraph 4.3), where also information on the 316 involved consumers are collected. As we have done above, the matrix product between the table containing preference scores Y and $\widetilde{Z}$ matrix of product characteristics is computed. Then, CLV approach based on $\widetilde{S}$ criterion, is performed on $\widetilde{Y}$, obtained from the juxtaposition of Y and $Y\widetilde{Z}$, and taking into account X matrix of product descriptors as external data.

The dendrogram (figure 4.27), which illustrate the arrangement of the groups produced by the hierarchical clustering algorithm, clearly shows a partition in to two groups.

The partition algorithm converges after only two iterations to a solution in two segments of 152 and 164 consumers and 50 and 45 $Y\tilde{Z}$ variables respectively. The criterion is equal to 4.01 for a global variation accounted for of 78.49%. In the table 4.12 information about the two groups are collected.

Figure 4.27 – Dendrogram of CLV ascendant hierarchical classification



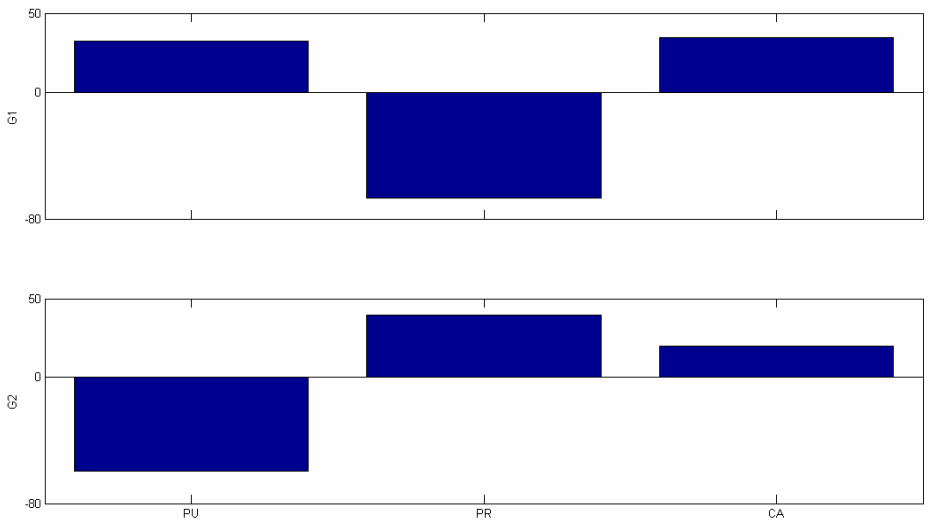
The description of the two groups is given in figures 4.28 and 4.29, which show values of latent components associated to the groups and values of loadings associated to the external variables respectively. The two latent components describe different preference directions ($r_{c_1c_2} = -0.638$) and the RV coefficients show that the latent factors well summarized the preference and consumer information contained in the groups (given in table 4.12).

People in the first group (G1) do not like *Nostrano di Primiero* (PR) but they appreciate *Puzzone di Moena* (PU) and *Campitello di Fassa* (CA) with almost the same strength. These two cheeses are characterized by big and numerous eyes, moisture, elasticity and deformability to the touch and solubility in mouth. They have a sweet taste and they are rich in odour and aroma of milk and odour of egg.

Table 4.12 – Description of the two groups, where nj indicates the number of consumers and nv the number of $\widetilde{YZ}$ characteristics. A partial list of $\widetilde{Z}$ characteristics is also reported.

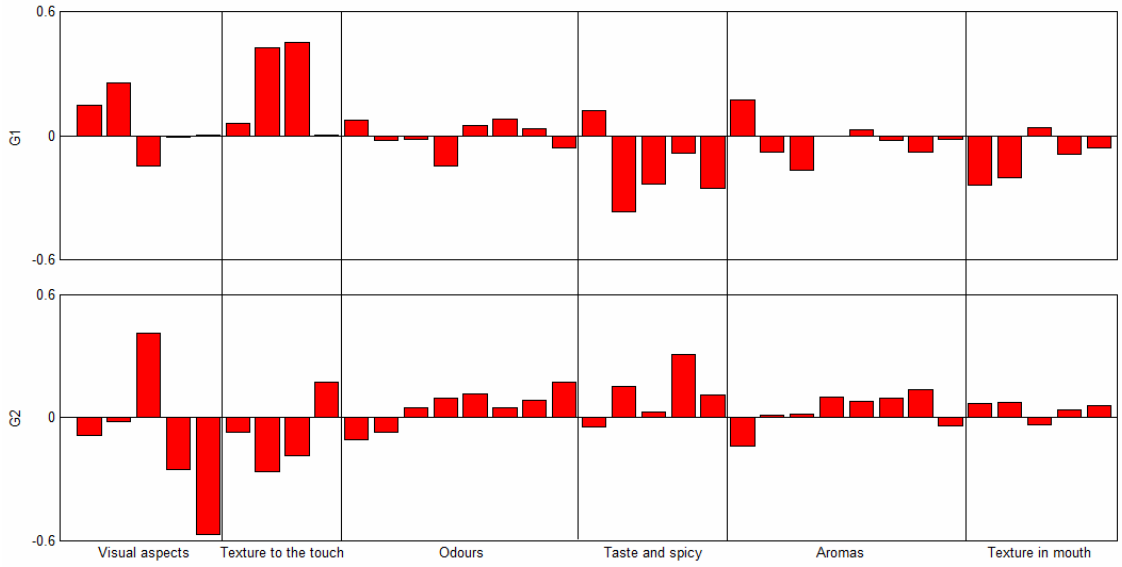
	G1	**G2**
nj	152	164
nv	50	45
RV_k	0.899	0.657
list of consumer descriptors	sex_2	sex_1
	age_1	age_4
	age_3	age_5
	age_8	age_6
	members_2	age_7
	members_4	edu_4
	members_5	prov_3
	members_6	weekfre_3
	prov_1	weekfre_4
	prov_2	weekfre_5
	…	…

Figure 4.28 – Latent components c_k associated to the 2 groups.



The second group (G2) does not like *Puzzone di Moena* but it prefers cheeses like *Nostrano di Primiero* with irregular-shaped eyes, firm, crumbly, granulous and adhesive in mouth and rough to the touch. It is characterized by stronger tastes (salty, bitter and spicy) with intensive odour of boiled-milk and odours and flavours of rind, cow, egg, ammonia and butyric.

Figure 4.29 – Loadings a_k associated to the **X** variables.



Also in terms of individual characteristics, the two groups show two different consumer profile (correlation of each variable with its associated latent component is given in detail in appendix B).

To the first group fifty modalities are been assigned, forty-seven of them are enough correlated with c_1. People collected in this group and then associated to its product profile, are female (sex_2) at any age (age_1, age_3 and age_8) from local provinces (prov_1 and prov_2), who live in families with at least 2 members (members_2, members_4, members_5 and members_6), with low school level (edu_2, edu_3). In this group are collected those people who eat cheese less then once a week (weekfre_1). They overall prefer fresh, young cheeses (pref1_2, pref1_4, pref2_2,

pref3_2 and pre3_4), that they generally buy in supermarket or in the dairies (where_1, where_3, where_5, where_6 and where_7). Maybe because they live in Trentino (a mountainous region in northern Italy) they show to know and eat quite regularly local cheeses (parm_2, parm_3 and parm_4; spr_1, spr_3 and spr_5; pri_2, pri_3 and pri_4; puz_2 and puz_4;cam_2, cam_3 and cam_4) except for *Vezzena* eaten less then once per month (vez_1 and vez_2), even if they declare that mountain cheeses are not better then those produced in lowland dairies (mount_2).

Forty-five modalities are associated to the second group, thirty-nine of them are enough correlated with the latent component c_2. Consumers, which do not like "Puzzone di Moena", are male (sex_1) aged between 40 and 79 years from different provinces of Italy (prov_3) and a middle school level (edu_4). They are great eaters of cheese (weekfre_3, weekfre_4 and weekfre_5), that they generally buy in specialist cheese shops (where_2 and where_4). They generally buy national cheeses (prod_2), in fact, they declare to consume local cheeses less than 2 times per month on average (pri_1, spr_2, puz_3, cam_1 and parm_1). That is not true for *Vezzena*, a mountain cheese well know out of region, eaten more then seven times per month (vez_3, vez_4 and vez_5). Maybe for this reason they declare that mountain cheeses are better than those produced in plain dairies (mount_1).

Chapter 5

Conclusion and perspectives

This chapter is dedicated to summarize the main aspects of this work, providing some concluding remarks and some suggestions for possible extensions to be realized.

This thesis aims to introduce a new method for consumer classification in the context of L-structured data, i.e. when three data table containing preference data (here called Y), products characteristics (X) and consumer descriptors (Z) are available. First we described in detail the use of a recent technique, CLV (Vigneau & Qannari, 2003), for variables clustering around latent components. This method allows clustering of variables without a distance matrix has been defined and it consists in automatically and simultaneously determining k clusters of variables and k latent components such that the variables in each group are as correlated as possible with the corresponding factor.

In CLV approach, as proposed by the authors, a classification around latent components can be performed on preference data only, or taking into account external variables, but it does not consider the case when a Z matrix, containing individual characteristics, is available. This thesis proposed a two-step procedure, inspired by few papers on two-step L-PLS, that provides a segmentation of consumers involving Z information. Moreover, correlation between latent components and RV coefficient, as measure of similarity between variables collected in a group and the associated latent factor, are proposed as indices of clustering goodness.

Finally, the new proposal is applied on two real data sets, coming from two consumer studies on food products conducted by the Agri-food Quality Department of the Agricultural Institute of S. Michele all'Adige (IASMA). Applying this procedure we verified that is possible to cluster a panel of consumers taking into account all the information at disposal.

Regarding future work, a new CLV algorithm for extracting clusters with a three-block input data could be searched for, using as starting point H. Martens' work (2005) on L-PLS-R. In this way two loading vectors will be estimated (one associated to X variables and the other one to Z variables) allowing a clearer group interpretation. In addition, the application of two-step CLV to fields different from those in which it was born, needs to be explored.

References

° Abdallah, H., Saporta, G. (1998) Classification d'un ensemble de variables qualitatives. *Revue de Statistique Appliquée* XLVI (4), 5-26.

° Al-Kandari N.M. & Jolliffe, I.L. (2001) Variable selection and interpretation of covariance principal components, *Communication in Statistics: Simulation and Computation*, 30, 339-354.

° Carroll, J.D. (1972) Individual differences and multidimensional scaling, in Shepard, R.N., Romney, A.K., & Nerlove, S.B. (Eds.) *Multidimensional scaling: Theory and Applications in the Behavioural Science*, Vol.1, Seminar Press, New York, 105-155.

° Escoufier, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, 751-760.

° Esposito Vinzi, V., Guinot, C. & Squillacciotti S. (2007) Two-step PLS Regression for L-Structured Data: an application in the cosmetics industry. *Statistical Methods and Applications*, Physica-Verlag, 16(2), 263-278.

° Everitt, B.S. (1974) *Cluster Analysis*. J. Wiley & Sons Publishers, London.

° Geladi, P., & Kowalski B. (1986) Partial least square regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.

° Gallerani, G., Gasperi, F., & Monetti, A. (2000) Judge selection for hard and semi-hard cheese sensory evaluation, *Food Quality and Preference*, 11 (), 465-474.

° Gasperi, F., Biasioli, F., Framondino, V., & Endrizzi, I. (2004) Ruolo dell'analisi sensoriale nella definizione delle caratteristiche dei prodotti tipici: l'esempio dei formaggi trentini/The role of sensory analysis in the characterization of traditional products: the case of study of cheese from Trentino. *Sci. Tecn. Latt.-Cas*, 55, 345-364.

° Gower, J.C. (1985) Measures of similarity, dissimilarity, and distance. In Kotz, S. & Johnson N.L. Encyclopaedia of Statistical Sciences, J.Wiley & Sons, New York, (5), pp 397-405.

° Graham, J.W., & Hofer, S.M. (2000) Multiple imputation in multivariate research. In Little, T.D., Schnabel, K.U., & Baumert, J. (Eds.) *Modelling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples*, Erlbaum, Hillsdale,NJ, pp 201-218.

° Graham, J.W., Cumsille, P.E., & Elek-Fisk, E. (2003) Methods for handling missing data. In Schinka, J.A. & Velicer, W.F. (Eds.) *Research Methods in Psychology*. Volume 2 of Handbook of Psychology (Weiner, I.B., Editor-in-chief), J. Wiley & Sons Publishers, New York, pp 87-114.

° Greenhoff, K., & MacFie, H.J.H. (1994) Preference mapping in practice. In *Measurements of food preferences*. Blackie academic & professional, London, 137-166.

° Guo Q. Wu, W., Massart, D.L., Boucon, C., & De Jong, S. (2002) Feature selection in principal component analysis of analytical data, *Chemometrics and Intelligen Laboratory Systems*, 61, 123-132.

° Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1992) *Multivariate Data Analysis, with Readings.* Maxwell Macmillian International, New York.

° Harduin, J.-B. (2006) Clustering of quality of life items around latent variables. Proceedings of *Intermediate Conference on "Statistical Latent Variables Models in Health Sciences"*, Perugia, Italy, 6 – 8 September.

° Jolliffe, I.T (1972) Discarding variables in a principal component analysis. I: Artificial data, *Applied Statistics*, 21, 160-173.

° Kaufman, L., & Rousseeuw, P.J. (1990) *Finding groups in data. An introduction to cluster analysis.* J. Wiley & Sons Publishers, London.

° Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J.-D. (1995) Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20, 643-656.

° Kendall, M.G. (1957) *A course in multivariate analysis*. Griffin, London

° Krzanowski W.J. (1987) Selection of variables to preserve multivariate data structure, using principal components, *Applied Statistics*, 36, 22-33.

° Kubberød, E., Ueland, Ø., Rødbotten, Westead, F., and Risvik E. (2002) Gender specific preferences and attitudes towards meat. *Food Quality and Preference*, 13 (5), 285-294.

° Lengard V., and Kermit, M. (2006) 3-Way and 3-block PLS regressions in consumer preference analysis. *Food Quality and Preference*, 17 (3-4), 234-242.

° Little, R.J.A., & Rubin, D.B. (2002) *Statistical analysis with missing data*, J. Wiley & Sons Publishers, New York.

° Luckow, T., & Delahunty, C. (2004) Which juice is "healthier"? A consumer study of probiotic non-diary juice drinks. *Food Quality and Preference*, 15 (7-8), 751-759.

° MacQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on mathematical statistics and probability*. Berkeley, University of California Press,1, 281-297.

° Martens, H., & Naes, T: (1989). *Multivariate Calibration*. J. Wiley & Sons Publishers, London.

° Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Høy, M., Westad, F., Thybo, A., & Martens, M. (2005) Regression of a matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. *Computational Statistics & Data Analysis*, 48 (1), 103-123.

° McCabe G.P. (1984) Principal variables, *Technometrics*, 26, 137-144.

° Naes, T., Isaksson, T., Fearn, T., & Davies T. (2002) *Multivariate Calibration and Classification.* NIR Publications, Chichester.

° Overall, J.E., & Klett, C.J. (1972) *Applied multivariate analysis*. McGraw-Hill, Inc, New York.

° Qannari, E.M., Vigneau, E., & Courcoux, P. (1998) Une nouvelle distance entre variables; application en classification. *Revue de Statistique Appliquée* XLVI (2), 21-32.

° Qannari, E.M., Vigneau, E., Luscon, P., Lefebrre, A.C., & Vey, F. (1997) Clustering of variables, application in consumer and sensory studies. *Food Quality and Preference*, 8 (5-6), 423-428.

° Sabatier, R. (1987) *Méthodes factorielles en analyse de données: approximations et prise en compte de variables concomittantes.* Thèse Doctorat Sèc. Univ. Sci. Techn. Languedoc, Montpellier, France.

° Sabatier, R., Lebreton, J. B.V., & Chessel, D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: Coppi, R., & Bolasco, S. (Eds.), *Multiway Data Analysis*, Elsevier Science Publishers North Holland, Amsterdam, pp 341-352.

° Saebø, S., Martens, M., & Martens, H. (2008) Three-block data modeling by endo- and exo-LPLS regression. In: Esposito Vinzi, V., Chin, W., Henseler, J., & Wang, H.

(Eds.) *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer-Verlag, Heidelberg.

° Sahmer, K. (2006) *Propriétés et extensions de la classification de variables autur de composantes latentes. Application an évaluation sensorielle.* Ph.D. Thesis, Univerisité Rennes II Haute Bretagne, Universität Dortmund.

° Sahmer, K., Vigneau, E., & Qannari, E.M. (2006) A cluster approach to analyze preference data: Choice of the number of clusters. *Food Quality and Preference*, 17 (3-4), 257-265.

° SAS/ STAT (1990) The VARCLUS procedure. User's Guide, Version 6, Cary, North Carolina: SAS Insitute Inc., (2), pp 1641-1659.

° Schafer, J.L. (1997) *Analysis of incomplete multivariate data*, Chapman and Hall, London.

° Schafer, J.L., & Graham, J.W. (2002) Missing data: Our view of state of art. *Psychological Methods*, 7(2), 147-177.

° Schlich, P. (1996) Defining and validating assessor compromises about product distances and attribute correlations. In Naes, T. & Risvik, E. (Ed.) *Mutlivariate analysis of data in sensory science*. New York: Elsevier.

° Soffritti, G. (1999) Hierarchical clustering of variables: a comparison among strategies of analysis. *Communications in statistics simulation and computation*, 28 (4), 977-999.

° Squillacciotti, S. (2004) *PLS modelling of mono- and multi-block L-structures: methodological contributions and applications*. Ph.D. Thesis, Università degli studi di Napoli "Federico II".

° Stone, H., Sidel, J., Oliver, S., Woolsey, A., & Singleton, R.C. (1974) Sensory evaluation by quantitative descriptive analysis. *Food Technology*, 28 (1), 24-34.

° Tenenhaus, M. (1998) La regression PLS. Éditions Technip, Paris.

° Tenenhaus, M., Pagès, J., Ambroisine, L., & Guinot, C. (2005) PLS methodology to study relationships between hedonic judgements and product characteristics. *Food Quality and Preference*, 16 (4), 315-325.

° Thybo, A.K., Kühn, B.F., & Martens H. (2003) Explaining Danish children's preferences for apples using instrumental, sensory and demographic/behavioural data. *Food Quality and Preference*, 15 (1), 53-63.

° Tuorila, H., Lähteenmäki, L., Pohjalainen, L., & Lotti, L. (2001) Food Neophobia among the Finns and related responses to familiar and unfamiliar foods. *Food Quality and Preference*, 12 (1), 29-37.

° Umetrics (2005) SIMCA-P and SIMCA-P+ 11 User Guide, Umeå, Sweden, UMETRICS AB.

° Vigneau E., & Qannari, E.M. (2002) Segmentation of consumers taking into account of external data. A clustering of variables approach. *Food Quality and Preference*, 13 (7-8), 515-521.

° Vigneau, E., & Qannari, E.M. (2003) Clustering of variables around latent components. *Communications in statistics simulation and computation*, 12 (4), 1131-1150.

° Vigneau, E., Qannari, E.M., Punter, P.H., & Knoops S. (2001) Segmentation of a panel of consumers using clustering of variables around latent directions of preference. *Food Quality and Preference*, 12 (5-7), 359-363.

° Vigneau, E., Qannari, E.M., Sahmer, K., & Ladiray, D. (2006) Classification de variables autour de composantes latentes. *Statistique Appliquée* (1), 27-45.

° Wold, H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiaah, P.H. (Eds.) *Multivariate analysis,* Academic Press, New York, pp 391-420.

° Wold, H. (1975) Path models with latent variables: The NIPALS approach. In Blalock, H.M. *et al.* (Eds.) *Quantitative Sociology: International perspectives on mathematical and statistical model building*, Academic Press, pp 307-357.

° Wold, H. (1985) Partial Least Squares. J.Wiley & Sons, New York.

° Wold, S., Albano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., & Sjöström, M. (1983) Pattern recognition: Finding and using regularities in multivariate data. In Martens, H. and Russwurm, H. jr (Eds.), *Food research and data analysis*. Applied Science Publishers, London, pp 147-188.

° Wold., S., Ruhe, A., Wold, H., & Dunn III, W.J. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverse. *SIAM Journal of Scientific and Statistical Computations*, 5, 735-743.

Appendix A

List of variables collected by the questionnaires in INTERBERRY and CHEESE case of study.

Table A.1 – Socio-demographic descriptors (interberry data)

variable name	variable description
child	n° of children (1=0;2=1;3=2;4=3 children)
sex	gender (1=male; 2=female)
age	age classes (1=14 years; 2=15; 3=16; 4=from 17 to 20; 5=from 21 to 25; 6=from26 to 35; 7=from 36 to 45; 8=46 and more)
status	marital status (1=married; 2=not married)
school	school level (from 1=lowest to 4=highest)
smoke	smoker (1=yes; 2=no)
staff	1=staff; 2=student

Table A.2 – Variables describing fruit consumption and habits (interberry data)

variable name	variable description
fruit1	fresh fruit portion per day (1=<1;2=2;3=3;4=4;5=5;6=more than 5)
fruit2	fruit at breakfast (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit3	fruit as snack (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit4	fruit for lunch or dinner (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit5	fruit at lunch or dinner (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit6	fruit bought at the supermarket (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit7	fruit bought at the F&V shop (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit8	fruit bought at the market (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit9	fruit bought at the discount (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit10	fruit bought from the producer (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit11	whole fruit (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit12	fruit salad (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit13	squeezed (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit14	cooked (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
fruit15	liking of golden delicious (from 1= "I do not like it at all" to 9="I like it very much")
fruit16	liking of granny smith (from 1= "I do not like it at all" to 9="I like it very much")
fruit17	liking of stark delicious (from 1= "I do not like it at all" to 9="I like it very much")
fruit18	liking of ripening of golden delicious (from 1="not ripe" to 9="too ripe")
fruit19	liking of apricot (from 1= "I do not like it at all" to 9="I like it very much")
fruit20	liking of pineapple (from 1= "I do not like it at all" to 9="I like it very much")
fruit21	liking of watermelon (from 1= "I do not like it at all" to 9="I like it very much")
fruit22	liking of orange (from 1= "I do not like it at all" to 9="I like it very much")
fruit23	liking of banana (from 1= "I do not like it at all" to 9="I like it very much")
fruit24	liking of cherry (from 1= "I do not like it at all" to 9="I like it very much")
fruit25	liking of strawberry (from 1= "I do not like it at all" to 9="I like it very much")
fruit26	liking of kiwi (from 1= "I do not like it at all" to 9="I like it very much")
fruit27	liking of raspberry (from 1= "I do not like it at all" to 9="I like it very much")
fruit28	liking of mandarin (from 1= "I do not like it at all" to 9="I like it very much")
fruit29	liking of apple (from 1= "I do not like it at all" to 9="I like it very much")
fruit30	liking of pomegranate (from 1= "I do not like it at all" to 9="I like it very much")
fruit31	liking of blueberry (from 1= "I do not like it at all" to 9="I like it very much")
fruit32	liking of blackberry (from 1= "I do not like it at all" to 9="I like it very much")
fruit33	liking of pear (from 1= "I do not like it at all" to 9="I like it very much")
fruit34	liking of peach (from 1= "I do not like it at all" to 9="I like it very much")
fruit35	liking of grapefruit (from 1= "I do not like it at all" to 9="I like it very much")
fruit36	liking of plum (from 1= "I do not like it at all" to 9="I like it very much")
fruit37	liking of currant (from 1= "I do not like it at all" to 9="I like it very much")
fruit38	liking of grape (from 1= "I do not like it at all" to 9="I like it very much")

Table A.3 – Variables describing juice consumption and purchase habits (interberry data)

variable name	variable description
juice1	I read the label (1=yes; 2=no)
juice2	if juice1=1 % of fruit (1=yes; 2=no)
juice3	if juice1=1 expiration date (1=yes; 2=no)
juice4	if juice1=1 ingredients (1=yes; 2=no)
juice5	if juice1=1 nutritional facts (1=yes; 2=no)
juice6	n° of glasses of fresh 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) at breakfast
juice7	n° of glasses of fresh 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) as snack
juice8	n° of glasses of fresh 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) at lunch/dinner
juice9	n° of glasses of UHT* 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) at breakfast
juice10	n° of glasses of UHT* 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) as snack
juice11	n° of glasses of UHT* 100% juices per week (1<1;2=1-3;3=4-7;4=more than 7) at lunch/dinner
juice12	n° of glasses of nectars per week (1<1;2=1-3;3=4-7;4=more than 7) at breakfast
juice13	n° of glasses of nectars per week (1<1;2=1-3;3=4-7;4=more than 7) as snack
juice14	n° of glasses of nectars per week (1<1;2=1-3;3=4-7;4=more than 7) at lunch/dinner
juice15	n° of glasses of squeezed juices per week (1=0;2=<1;3=1-3;4=4-7;5=more than 7) at breakfast
juice16	n° of glasses of squeezed juices per week(1<1;2=1-3;3=4-7;4=more than 7) as snack
juice17	n° of glasses of squeezed juices per week (1<1;2=1-3;3=4-7;4=more than 7) at lunch/dinner
juice18	n° of glasses of centrifuged juices per week (1<1;2=1-3;3=4-7;4=more than 7) at breakfast
juice19	n° of glasses of centrifuged juices per week (1<1;2=1-3;3=4-7;4=more than 7) as snack
juice20	n° of glasses of centrifuged juices per week (1<1;2=1-3;3=4-7;4=more than 7) at lunch/dinner
juice21	1=pulpy; 2=liquid
juice22	1=multi-fruit; 2=single fruit
juice23	1=enriched; 2=natural
juice24	a juice have to be refreshing (from 1=not important to 9=very important)
juice25	a juice have to be nutritious (from 1=not important to 9=very important)
juice26	a juice have to be healthy (from 1=not important to 9=very important)
juice27	a juice have to be organic (from 1=not important to 9=very important)
juice28	a juice must have an high % of fruit (from 1=not important to 9=very important)
juice29	liking of Ace juice (from 1= "I do not like it at all" to 9="I like it very much")
juice30	liking of apricot juice (from 1= "I do not like it at all" to 9="I like it very much")
juice31	liking of pineapple juice (from 1= "I do not like it at all" to 9="I like it very much")
juice32	liking of orange juice (from 1= "I do not like it at all" to 9="I like it very much")
juice33	liking of tropical juice (from 1= "I do not like it at all" to 9="I like it very much")
juice34	liking of apple juice (from 1= "I do not like it at all" to 9="I like it very much")
juice35	liking of pear juice (from 1= "I do not like it at all" to 9="I like it very much")
juice36	liking of peach juice (from 1= "I do not like it at all" to 9="I like it very much")
juice37	liking of grapefruit juice (from 1= "I do not like it at all" to 9="I like it very much")

* Ultra High Temperature

Table A.4 – Variables describing berry fruit consumption and purchase habits (interberry data)

Variable name	Variable description
Bf1	strawberries eaten in summertime per month (1=<1;2=1-2; 3=3-4; 4=5-6; 5=7or more)
Bf2	strawberries eaten in the rest of the year per month (1=<1;2=1-2; 3=3-4; 4=5-6; 5=7or more)
Bf3	berry fruits eaten in summertime per month (1=<1;2=1-2; 3=3-4; 4=5-6; 5=7or more)
Bf4	berry fruits eaten in the rest of the year per month (1=<1;2=1-2; 3=3-4; 4=5-6; 5=7or more)
Bf5	berry fruits bought at the supermarket (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf6	berry fruits bought at the F&V shop (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf7	berry fruits bought at the market (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf8	berry fruits bought at the discount (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf9	berry fruits bought from the producers (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf10	shop where you buy strawberries sales them 1=just in summertime or 2=all the year
Bf11	shop where you buy berry fruits sales them 1=just in summertime or 2=all the year
Bf12	eaten without anything else (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf13	eaten as fruit salad (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf14	eaten as shake (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf15	eaten as pudding decoration (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf16	eaten as pudding (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf17	eaten as jam (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf18	frozen strawberries bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf19	frozen berry fruits bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf20	strawberry jam bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf21	berry fruit jam bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf22	strawberry yogurt bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf23	berry fruit yogurt bought(1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf24	strawberry ice cream bought(1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf25	berry fruit ice cream bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf26	strawberry drink bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf27	berry fruit drink bought (1=never; 2=occasionally; 3=very often; 4=often; 5=always)
Bf28	buying strawberries/berry fruits they have to be nice (from 1=not important to 9=very important)
Bf29	buying strawberries/berry fruits you look at the pick up date (from 1=not important to 9=very important)
Bf30	buying strawberries/berry fruits they have to be local (from 1=not important to 9=very important)
Bf31	buying strawberries/berry fruits you look at the price (from 1=not important to 9=very important)
Bf32	Strawberry have to be red (from 1=not important to 9=very important)
Bf33	Strawberry have to be big (from 1=not important to 9=very important)
Bf34	Strawberry have to be perfumed (from 1=not important to 9=very important)
Bf35	Strawberry have to be sweet (from 1=not important to 9=very important)
Bf36	Strawberry have to be acid (from 1=not important to 9=very important)
Bf37	Berry fruits have to have an intense colour (from 1=not important to 9=very important)
Bf38	Berry fruits have to be big (from 1=not important to 9=very important)
Bf39	Berry fruits have to be perfumed (from 1=not important to 9=very important)
Bf40	Berry fruits have to be sweet (from 1=not important to 9=very important)
Bf41	Berry fruits have to be acid (from 1=not important to 9=very important)

Table A.5 – Food neophobia scale items (interberry data)

variable name	variable description
Fns1	I'm constantly sampling new and different foods (R) (from 1=extremely disagree to 9=extremely agree)
Fns2	I don't trust new foods (from 1=extremely disagree to 9=extremely agree)
Fns3	If I don't know what is in a food, I won't try it. (from 1=extremely disagree to 9=extremely agree)
Fns4	I like foods from different countries (R) (from 1=extremely disagree to 9=extremely agree)
Fns5	Ethnic foods look too weird to eat. (from 1=extremely disagree to 9=extremely agree)
Fns6	At dinner parties I will try a new food (from 1=extremely disagree to 9=extremely agree)
Fns7	I'm afraid to eat things I have never had before (from 1=extremely disagree to 9=extremely agree)
Fns8	I am very particular about the foods I will eat (from 1=extremely disagree to 9=extremely agree)
Fns9	I will eat almost anything (R) (from 1=extremely disagree to 9=extremely agree)
Fns10	I like to try new ethnic restaurants (R) (from 1=extremely disagree to 9=extremely agree)

Table A.6 – Impressions on new food, exotic food, ready to eat food and familiar food coded from 1="I don't know it"; 2="I know it but I've never tried it"; 3="I have tried it but I don't use it"; 4="I eat it occasionally"; to 5="I eat it regularly" (interberry data)

variable name	variable description
NF_1	New Food: probiotic milk
NF_2	New Food: butter without cholesterol
NF_3	New Food: rice-pasta
NF_4	New Food: high digestive tolerance milk
EF_1	Exotic Food: sushi
EF_2	Exotic Food: soja sauce
EF_3	Exotic Food: couscous
EF_4	Exotic Food: avocado
RE_1	Ready to Eat: salad
RE_2	Ready to Eat: ready to cook format pie
RE_3	Ready to Eat: frozen ready meals
RE_4	Ready to Eat: lyophilized food
FF_1	Familiar Food: olive oil
FF_2	Familiar Food: parmesan cheese
FF_3	Familiar Food: tomatoes
FF_4	Familiar Food: pasta

Table A.7–Variables collecting knowledge about diet and antioxidants (interberry data)

variable name	variable description
H1	Indicate the berry that is more healthy (1=straw;2=rasp;3=black;4=currant;5=blue; 6=morethan1answer)
H2	Indicate the berry that is more artificial l(1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H3	Indicate the berry that is more natural (1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H4	Indicate the berry that is more chemical (1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H5	which berry is similar to its wild version in terms of visual aspects (1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H6	which berry is similar to its wild version in terms of odour (1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H7	which berry is similar to its wild version in terms of flavour (1=straw;2=rasp;3=black;4=currant;5=blue;6=morethan1answer)
H8	Santal juice with 10 vitamins and red fruits (1="I don't know it"; 2="I know it but I've never tried it"; 3="I have tried it but I don't use it"; 4="I eat it occasionally";5="I eat it regularly")
H9	Special K Kellogg's cereals with red fruits (1="I don't know it"; 2="I know it but I've never tried it"; 3="I have tried it but I don't use it"; 4="I eat it occasionally";5="I eat it regularly")
H10	Yougurt Yomo "Fruit and Veg" (1="I don't know it"; 2="I know it but I've never tried it"; 3="I have tried it but I don't use it"; 4="I eat it occasionally";5="I eat it regularly")
H11	choose a colour (1=maroon/violet;2=green;3=white;4yellow/orange;5=red)
H12	because I like food with that colour (from 1=extremely disagree to 9=extremely agree)
H13	because food with that colour is rich in vitamins (from 1=extremely disagree to 9=extremely agree)
H14	because food with that colour is healthy (from 1=extremely disagree to 9=extremely agree)
H15	because food with that colour is rich in antioxidants (from 1=extremely disagree to 9=extremely agree)
H16	because that colour is my favourite (from 1=extremely disagree to 9=extremely agree)
H17	do you know free radical(1=y,2=n)
H18	do you know antioxidants(1=y,2=n)
H19	do you know polyphenols(1=y,2=n)
H20	do you know antocyanins(1=y,2=n)
H21	antioxidants preserve food(1=y,2=n)
H22	antioxidants protect against cardiovascular disease(1=y,2=n)
H23	antioxidants protect against cancer(1=y,2=n)
H24	antioxidants produce free radical(1=y,2=n)
H25	antioxidants slow down brain ageing(1=y,2=n)
H26	Event1 (1=I don't know it;2=I've heard about it;3=I know it; 4=I follow it)
H27	Event2 (1=I don't know it;2=I've heard about it;3=I know it; 4=I follow it)
H28	Event3 (1=I don't know it;2=I've heard about it;3=I know it; 4=I follow it)
H29	Event4 (1=I don't know it;2=I've heard about it;3=I know it; 4=I follow it)
H30	Event5 (1=I don't know it;2=I've heard about it;3=I know it; 4=I follow it)

Table A.8–Variables description (cheese data)

variable name	variable description
Sex	gender (1=male; 2=female)
Age	Age classes (1= until 19 years; 2= from 20 to 29; 3= from 30 to 39; 4= from 40 to 49; 5= from 50 to 59; 6= from 60 to 69; 7= from 70 to 79; 8= from 80 to 89)
Members	Number of family members
Prov	Consumer's province of origin (1= Trento; 2= Bolzano; 3= other Italian provinces; 4=abroad)
Edu	school level (from 1=lowest to 5=highest)
Weekfreq	Weekly cheese consumption, excluding grated Parmesan (1=<1; 2=1 or 2 times; 3=3 or 4; 4=5 or 6; 5=7 or more)
Pref1	First preferred cheese (1=like Parmesan cheese; 2=fresh cheese; 3=bloomy rind/blue-veined cheeses; 4=semi-hard cheese; 5=hard cheese; 6= no vaccine cheese and 7=no preference)
Pref2	Second preferred cheese (1=like Parmesan cheese; 2=fresh cheese; 3=bloomy rind/blue-veined cheeses; 4=semi-hard cheese; 5=hard cheese; 6= no vaccine cheese and 7=no preference)
Pref3	Third preferred cheese (1=like Parmesan cheese; 2=fresh cheese; 3=bloomy rind/blue-veined cheeses; 4=semi-hard cheese; 5=hard cheese; 6= no vaccine cheese and 7=no preference)
Where	Where do you generally buy cheeses? (1= supermarket; 2=cheese shop; 3=dairy)
Prod	Do you generally buy (1=local; 2=national; 3=other; 4=1+2) cheese?
Tre	Monthly "Trentingrana" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Puz	Monthly "Puzzone di Moena" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Spr	Monthly "Spressa delle Giudicarie" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Pri	Monthly "Nostrano di Primiero" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Vez	Monthly "Vezzena" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Cam	Monthly "Campitello di Fassa" cheese consumption (1=never; 2 =<1; 3=1 or 2 times; 4= 3 or 4 and 5=more than 4)
Mount	Mountain cheese is better than that made in plain?(1=Yes; 2=No)

Appendix B

Correlation coefficients between Z characteristics and each latent component for INTERBERRY and CHEESE datasets.

Table B.1–Correlations between Z variables and group latent components (interberry data)

Z modalities	c1	c2	c3	Z modalities	c1	c2	c3
RE_1_5	0.518	--	--	fruit26_1	--	--	0.719
child_2	0.448	--	--	FF_2_4	--	--	0.717
status_1	0.429	--	--	staff_2	--	--	0.716
fruit3_1	0.406	--	--	H20_1	--	--	0.716
Fns10_9	0.323	--	--	juice2_2	--	--	0.710
age_cod_7	0.291	--	--	NF_4_2	--	--	0.708
child_4	0.276	--	--	school_1	--	--	0.705
Bf5_5	0.274	--	--	Bf21_1	--	--	0.705
school_3	0.264	--	--	H17_1	--	--	0.700
staff_1	0.234	--	--	fruit27_9	--	--	0.698
fruit11_2	0.208	--	--	Fns8_7	--	--	0.697
fruit27_5	0.208	--	--	child_1	--	--	0.694
Fns10_1	0.191	--	--	status_2	--	--	0.694
fruit38_5	0.185	--	--	Bf3_5	--	--	0.682
Bf31_8	0.128	--	--	sex_1	--	--	0.680
EF_3_5	0.099	--	--	Fns1_6	--	--	0.678
Bf26_4	--	0.723	--	juice36_9	--	--	0.675
juice32_7	--	0.690	--	Fns3_6	--	--	0.672
age_cod_5	--	0.686	--	juice4_2	--	--	0.669
juice17_1	--	0.676	--	juice35_9	--	--	0.669
juice11_2	--	0.663	--	H28_1	--	--	0.664
juice15_1	--	0.639	--	fruit8_1	--	--	0.662
juice8_1	--	0.639	--	Bf8_1	--	--	0.655
H18_2	--	0.629	--	Bf38_6	--	--	0.650
fruit3_2	--	0.620	--	H22_1	--	--	0.649
juice11_1	--	0.599	--	H26_1	--	--	0.645
FF_1_5	--	0.595	--	juice14_1	--	--	0.639
H4_1	--	0.593	--	EF_3_1	--	--	0.639
fruit8_4	--	0.582	--	H18_1	--	--	0.638
school_2	--	0.581	--	juice15_4	--	--	0.631
Bf40_9	--	0.579	--	EF_2_1	--	--	0.629
juice23_1	--	0.577	--	juice27_3	--	--	0.628

Z modalities	c1	c2	c3	Z modalities	c1	c2	c3
Fns6_9	--	0.570	--	Bf29_7	--	--	0.623
juice9_1	--	0.554	--	juice23_2	--	--	0.621
H20_2	--	0.547	--	fruit3_3	--	--	0.619
Bf35_6	--	0.546	--	juice20_1	--	--	0.618
juice1_1	--	0.540	--	Fns8_6	--	--	0.617
juice2_1	--	0.538	--	H25_1	--	--	0.616
H2_1	--	0.538	--	EF_4_1	--	--	0.615
H17_2	--	0.532	--	fruit21_9	--	--	0.613
Bf33_3	--	0.464	--	fruit8_3	--	--	0.611
Bf3_3	--	0.464	--	Fns1_5	--	--	0.610
Bf17_1	--	0.449	--	juice29_7	--	--	0.608
juice4_1	--	0.448	--	Bf35_9	--	--	0.601
juice28_9	--	0.443	--	Bf26_2	--	--	0.591
RE_3_3	--	0.430	--	Bf38_5	--	--	0.584
H19_2	--	0.424	--	fruit1_1	--	--	0.579
Fns1_9	--	0.422	--	H25_2	--	--	0.565
fruit10_1	--	0.411	--	H28_2	--	--	0.559
Bf9_2	--	0.408	--	Bf24_2	--	--	0.552
fruit6_4	--	0.388	--	fruit38_9	--	--	0.548
Bf7_3	--	0.359	--	juice12_1	--	--	0.544
H16_5	--	0.313	--	Fns7_1	--	--	0.544
age_cod_6	--	0.290	--	H22_2	--	--	0.540
juice29_6	--	0.277	--	juice17_2	--	--	0.529
sex_2	--	0.224	--	juice24_7	--	--	0.526
H5_1	--	--	0.807	H14_7	--	--	0.525
H19_1	--	--	0.782	juice9_3	--	--	0.522
juice26_3	--	--	0.770	fruit5_1	--	--	0.512
juice1_2	--	--	0.758	Bf13_4	--	--	0.423
Bf13_5	--	--	0.743	juice29_9	--	--	0.421
age_cod_3	--	--	0.743	Fns9_9	--	--	0.387
H7_2	--	--	0.742	Fns6_6	--	--	0.366
H14_3	--	--	0.722	fruit5_4	--	--	0.365
age_cod_2	--	--	0.722	H26_2	--	--	0.352
fruit10_3	--	--	0.721	juice14_2	--	--	0.259

Table B.2–Correlations between Z variables and group latent components (cheese data)

variable	c1	c2	variable	c1	c2
where_7	1.000	--	puz_1	-0.048	--
spr_3	1.000	--	prod_4	-0.103	--
cam_4	1.000	--	where_2	--	0.999
weekfre_5	0.999	--	pref1_1	--	0.994
cam_2	0.999	--	pref1_6	--	0.990
weekfre_1	0.999	--	pref1_3	--	0.982
pri_4	0.998	--	vez_3	--	0.982
pref1_4	0.996	--	vez_5	--	0.982
where_6	0.994	--	prov_3	--	0.982
sex_2	0.993	--	pref3_5	--	0.979
pref2_2	0.988	--	cam_1	--	0.976
vez_2	0.988	--	prod_2	--	0.963
pref1_2	0.984	--	pref3_6	--	0.945
parm_3	0.982	--	age_7	--	0.941
vez_1	0.981	--	pref2_4	--	0.939
prov_1	0.968	--	pref2_7	--	0.929
edu_3	0.956	--	age_4	--	0.919
pref2_1	0.941	--	where_4	--	0.908
mount_2	0.927	--	edu_4	--	0.871
where_3	0.921	--	pref3_3	--	0.821
pref1_5	0.909	--	pref2_5	--	0.778
pref2_3	0.876	--	members_3	--	0.773
puz_2	0.876	--	age_5	--	0.767
spr_5	0.876	--	spr_2	--	0.757
pri_2	0.876	--	vez_4	--	0.757
pref3_2	0.869	--	age_6	--	0.757
age_1	0.868	--	pref2_6	--	0.742
members_4	0.866	--	sex_1	--	0.715
age_3	0.866	--	weekfre_3	--	0.702
parm_2	0.855	--	pref3_1	--	0.677
spr_1	0.850	--	pref1_7	--	0.654
prod_3	0.832	--	members_7	--	0.654
members_5	0.829	--	pri_1	--	0.638
puz_4	0.815	--	parm_1	--	0.619
pref3_4	0.749	--	cam_5	--	0.608
where_5	0.735	--	weekfre_4	--	0.590
pri_3	0.681	--	puz_5	--	0.583
members_2	0.645	--	puz_3	--	0.531
prov_2	0.613	--	parm_5	--	0.521
prod_1	0.603	--	mount_1	--	0.498
parm_4	0.591	--	members_1	--	0.418
edu_2	0.518	--	spr_4	--	0.188
age_8	0.482	--	prov_4	--	0.046
members_6	0.482	--	pri_5	--	-0.092
where_1	0.462	--	pref3_7	--	-0.328
cam_3	0.458	--	edu_1	--	-0.384
weekfre_2	0.285	--	age_2	--	-0.434
edu_5	0.023	--			