Food Quality and Preference 47 (2016) 54–63

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers

E. Vigneau [a,b,*], E.M. Qannari [a,b], B. Navez [c], V. Cottet [d]

[a] LUNAM University, ONIRIS, Sensometrics and Chemometrics Laboratory, Nantes, France
[b] INRA, Nantes, France
[c] CTIFL, Centre de Saint Rémy de Provence, France
[d] CTIFL, Centre de Balandran, France

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form 9 January 2015
Accepted 5 February 2015
Available online 14 February 2015

Keywords:
Segmentation of consumers
Cluster analysis
CLV method
Noise cluster

ABSTRACT

Cluster analysis is often used to segment a panel of consumers according to their overall liking. In general, all the consumers are assigned to one of the segments even though they do not fit to the pattern of any cluster. Within the clustering of variables around latent variables (CLV) framework, we propose two new approaches to handle this problem. The first approach ("$K+1$" strategy) consists in explicitly identifying an additional cluster which we refer to as "noise cluster". The second approach ("Sparse LV" strategy) computes the groups' latent variables of the CLV method with a sparsity constraint. Both strategies were tested on the basis of two real hedonic case studies and compared to the *k*-means cluster analysis. They made it possible to improve the discrimination between the products within each cluster and yield homogeneous clusters of consumers for a better understanding of the main tendencies of liking.

1. Introduction

Segmenting a panel of consumers consists in identifying groups of consumers that are as homogeneous as possible according to their responses. In food marketing research, segmentation is based on consumers' attitude and consumption habits collected by means of questionnaires including qualitative or quantitative items. In preference studies, the most common setup consists in asking each consumer of a panel to assess his or her overall liking score for a set of selected products, using a hedonic scale. In both contexts, cluster analysis is commonly used for segmentation (MacFie, 2007; Næs, Brockhoff, & Tomic, 2010). However, there is a difference in the structure of the data being handled which deserves to be emphasised: in the first type of studies, consumers are the objects (or individuals) of the analysis, whereas items are the variables; in preference studies, samples are the objects and the consumers are treated as variables. In this paper, we will only consider this second context and we will make a distinction between the clustering of objects and the clustering of variables. As a matter of fact, this distinction mainly relates to the type of dissimilarity or similarity used to assess the "distance" or proximity between the statistical entities to be clustered, whereas the global structure of the algorithms that could be used is the same. This will be detailed further.

Notwithstanding, by using statistical cluster analysis, all the consumers will be assigned to the clusters, either in a crisp/hard or in a fuzzy manner. Crisp algorithms (e.g. hierarchical clustering or *k*-means algorithms) are well-known and widely available in statistical packages or software. They provide non-overlapping clusters and each consumer will be assigned to one, and only one, cluster. Fuzzy algorithms, as *c*-means algorithms (Berget, Mevik, & Næs, 2008), yields fuzzy membership values using a fuzzifier parameter, m, to be determined by the user. Usually m is fixed to 2. The crisp clustering is a special case of fuzzy clustering when $m = 1$. In Johansen, Hersleth, and Naes (2010), in the context of external preference mapping, various values of m, ranging from 1.1 to 2.2 by a step of 0.1 were tested. These authors found that the lowest value ($m = 1.1$) was the best choice. Herein, we consider only the special case of $m = 1$ (*i.e.* crisp clustering) mainly for the sake of simplicity. However, the strategy of analysis could be extended to the fuzzy clustering setup.

Very often, the partition obtained with any statistical clustering method may be unsatisfactory and not relevant (Yenket, Chambers, & Johnson, 2011). It is especially the case when there is no clear evidence of the existence of subgroups in the population under study. Moreover, atypical consumers (*i.e.* consumers whose directions of preference poorly fit the main tendencies expressed by the other consumers in the panel) may blur the "true" structure.

* Corresponding author at: ONIRIS, Site de la Géraudière, Sensometrics and Chemometrics Laboratory, Nantes, France. Tel.: +33 2 51 78 54 40; fax: +33 2 51 78 54 38.
E-mail address: evelyne.vigneau@oniris-nantes.fr (E. Vigneau).

http://dx.doi.org/10.1016/j.foodqual.2015.02.008
0950-3293/

The implicit hypothesis that there is a "true" underlying structure may appear to be somewhat strong in the context of hedonic measurements. However, this theoretical statement is useful to contrast what is considered as outlying or "background" contamination information. The objective in this paper, is to set aside the consumers who indicated a direction of preference very different from the other consumers (isolated outliers) or who do not express clear differences in their liking (background noise). Both these situations will be considered as irrelevant with regard to the main underlying structure. Westad, Hersleth, and Lea (2004) have also addressed this issue by identifying "no strong preference" and "undecided" consumers.

In the domain of the clustering of objects (observations), this issue may be addressed by means of robust clustering methods (Dave & Krishnapuram, 1997; García-Escudero, Gordaliza, Matrán, and Mayo-Iscar, 2010). The trimmed *k*-means method, for which a special attention is given in García-Escudero et al. (2010), allows us to have a proportion of unallocated observations. Hopefully, these leftover observations are the most outlying points, relating to "bridge" points or "background" points. It will be emphasized further that one of the proposed approaches in this work shares similarities with the trimmed *k*-means principle. One of our strategies of analysis draws from the "noise cluster" concept introduced by Davé (1991) which was primarily introduced to set up fuzzy clustering algorithms that are less sensitive to noise and outliers. Dave and Krishnapuram (1997) give a very interesting discussion to establish a connection between the noise clustering method and several robust statistical methods.

This paper is structured as follows: In Section 2, the clustering around latent variables (CLV) approach by Vigneau and Qannari (2003), used for panel segmentation on the basis of their liking, will be recalled. Moreover, two new strategies, within the CLV framework, will be proposed for dealing with atypical or "noise" consumers. The first strategy, based on the "noise cluster" idea proposed by Davé (1991) for the clustering of observations, is adapted herein to the case of variables. It aims at obtaining a more robust partition of the consumers into a given number of groups (to be fixed). The second strategy is related to the sparse latent variables approach whereby the prototypes, or centroids, of the clusters are defined without involving those variables (*i.e.* consumers in our context) that are deemed to be atypical. To address the issue regarding the validity and the stability of the segments and the identified preference tendencies, some criteria are introduced and a cross-validation procedure is discussed in Section 2.4. Section 3 is devoted to the illustrations of the methods. Firstly, we will consider a real, but rather simple, case study to illustrate the efficiency of the proposed new CLV approaches in presence of atypical consumers. The second case study deals with a real and more complex example, with a high level of noise, for which the partition and the prototypes are clearly different according to the methodology used to perform the clustering.

2. Methods

2.1. The clustering of variables around latent variables (CLV) method

In the following, we consider data collected in preference mapping, where p consumers gave overall liking scores to n products (for instance using a 9-points scale). These scores are arranged in an $(n \times p)$ matrix, $\mathbf{X}$. The aim in segmenting the p consumers is to partition the consumers into groups (or clusters) with similar patterns of liking.

A wide range of approaches are available to address this issue. Besides non-automatic, but visual, clustering rules (Næs et al., 2010), or probabilistic approaches (De Soete & Winsberg, 1993; Séménou, Courcoux, Cardinal, Nicod, & Ouisse, 2007), *k*-means clustering or hierarchical clustering methods are the most common approaches. The *k*-means clustering, as the fuzzy clustering (*c*-means), are partitioning methods that may be considered as non-probabilistic versions of the model-based methods. In the preference mapping context, Wajrock, Antille, Rytz, Pineau, and Hager (2008) claimed that the partitioning methods outperform the hierarchical methods. However, all these procedures are based on the Euclidean distance as dissimilarity index.

Alternatively, we can be more interested in assessing the similarities between the directions of preference of the consumers by means of the angles (or the correlations) between their individual vectors of scores. The cluster analysis of variables around latent variables (CLV) method has been proposed within the framework of internal or external preference mapping (Vigneau, Charles, & Chen, 2014; Vigneau, Qannari, Punter, & Knoops, 2001). Similarly to the *k*-means algorithm, the CLV algorithm is an alternating optimization procedure. It aims at defining groups of variables (*i.e.* the consumers in our context) as homogeneous as possible, and a set of latent variables, each of them being associated with a group. These latent variables make it possible to pinpoint the main directions of preference in the data set.

The CLV method makes it possible to define clusters of variables of two different types: the directional groups which are elongated along an axis (i.e. positively and negatively correlated variables are merged together), and the local groups which include variables pointing to the same direction (i.e. only positively correlated variables are merged together). Herein, the objective is to separate segments of consumers who have distinctly different directions of preference, and, therefore, only the case of the local groups will be considered. Readers interested in the case of directional groups can refer to Chen and Vigneau (2014), Lovaglio (2011), Vigneau and Qannari (2003), Vigneau, Sahmer, Qannari, and Bertrand (2005).

The CLV method, for local groups, aims at maximizing the internal cohesion within each group of consumers and, simultaneously, determining a latent variable within each group. Formally, this is achieved by the maximization of the following criterion:

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

where $\mathbf{x_j} = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})^{\mathbf{T}}$ is the vector of the liking scores given by the consumer j ($j = 1, \ldots, p$) to the n products, $\mathbf{c_k}$ is an n-dimensional vector which represents the latent variable in the group G_k ($k = 1, \ldots, K$) and $\delta_{jk} = 1$ if consumer j belongs to group G_k, $\delta_{jk} = 0$ otherwise. In this expression, $\mathrm{cov}(\mathbf{x_j},\mathbf{c_k})$ stands for the covariance between $\mathbf{x_j}$ and $\mathbf{c_k}$, and $\mathrm{var}(\mathbf{c_k})$ stands for the variance of $\mathbf{c_k}$. It can be easily shown (Vigneau & Qannari, 2003) that, for a given partition of the variables, the optimum of criterion S is obtained when each latent variable $\mathbf{c_k}$ is proportional to the average score vector, $\overline{\mathbf{x}}_\mathbf{k}$, of the consumers belonging to G_k. More precisely, $\mathbf{c_k} = \overline{\mathbf{x}}_\mathbf{k}/\sqrt{\mathrm{var}(\overline{\mathbf{x}}_\mathbf{k})}$.

By way of comparing methods, we can note that with the *k*-means clustering, the centroid of each cluster C_k is the average vector of the (centered) scores associated to consumers in C_k. We will illustrate, in Section 3.2, the differences between the outcomes of CLV and *k*-means approaches, especially when the liking scores are not standardized.

In order to maximize the criterion S, for a fixed number of groups K, the CLV method is based on an alternating optimization algorithm (similar to the *k*-means algorithm). This algorithm starts from an initial partition of the variables (*i.e.* consumers). This initial partition can be generated at random, or, preferably, selected from the partition into K groups obtained with a hierarchical clustering based on the criterion S. Thereafter, two main steps are alternatively undergone: