Food Quality and Preference 71 (2019) 8–20

Contents lists available at ScienceDirect

Food Quality and Preference

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

Segmentation in projective mapping

Ingunn Berget[a,*], Paula Varela[a], Tormod Næs[a,b]

[a] Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway
[b] University of Copenhagen, Faculty of Life Sciences, Department Food Science, Rolighedsvej 30, 1958 Fredriksberg, Copenhagen, Denmark

ABSTRACT

Projective mapping (PM) or napping® has attained much attention in recent literature as a method for fast sensory profiling and measurement of consumer perception. However, little work has been done to understand the consumer's individual differences in these experiments. In this work, segmentation criteria based on the *Procrustes* distance are explored. The Procrustes distance can be applied with hierarchical clustering using the *Proclustrees* method, which consists of doing hierarchical clustering on the pairwise Procrustes distance between consumers. An alternative strategy called sequential *clusterwise rotations* (SCR) is proposed. SCR extracts clusters by a sequentially partitioning obtained by combining fuzzy clustering techniques and general Procrustes analysis. The methods were tested on simulated and real data and compared with clustering based on MFA results. The simulations show that the MFA approach was outperformed by the other methods when the underlying classes were of same size and there are noise configurations present in the data. For the real data, all methods provided at last one cluster similar to the consensus but differed with respect to the number of clusters identified as well as the interpretation of the clusters. Differences between the methodologies point out the need for external cluster validation in such experiments.

1. Introduction

Due to costs of maintaining a trained sensory panel for sensory profiling of food products, alternative methods, also known as rapid methods, have attained much focus in recent literature (Varela & Ares, 2014). One of these methods is projective mapping (PM) (Risvik, McEwan, Colwill, Rogers, & Lyon, 1994), also referred to as Napping® (Pages, 2003, 2005). PM is also similar to the spatial arrangement method (SpAM, (Goldstone, 1994)). In PM the assessors are asked to arrange a set of samples on a sheet of paper according to how similar or dissimilar the samples are.

Because PM can be applied both with trained panels and untrained consumers, the method is useful for studying consumer perception. However, consumers may use diverse criteria for their evaluation, particularly for complex products. As consumers are not trained, neither given instructions on what to focus on, individual differences may exist. In such cases the consensus will not always give the complete picture. Another issue is that similarity can be perceived in different ways for the different modalities. Also, for problems outside the food and drink area, some types of samples may appear similar visually, but have different functionality (Goldstone, 1994). It is therefore important to study individual variations in projective mapping experiments.

To obtain a consensus, PM data are typically analysed by multiple factor analysis (MFA, (Escofier & Pagès, 1994)), INDSCAL (Carroll & Chang, 1970) or Generalized Procrustes Analysis (GPA, (Gower, 1975)). For PM data GPA is constrained to two components, whereas both MFA and INDSCAL can provide more components. It has been shown that when focusing on the two first components, these methods provide similar consensus solutions although they are conceptually different (Næs, Berget, Liland, Ares, & Varela, 2017; Tomic, Berget, & Næs, 2015). Nevertheless, in studies with complex products, more than two components are sometimes necessary to provide a fully adequate consensus solution. More than two components may, however, be conceptually more demanding since the original data are given in two dimensions (x and y coordinates from each sheet). An alternative is to use clustering for the PM data to enhance interpretability, as discussed in (Vidal et al., 2016). In a comprehensive study on several data sets, each with 80–100 consumers they identified 2–4 segments with different descriptions of the samples in all cases. Moreover, they observed that more segments were needed when more components were needed in the MFA model. The study by Vidal et al., clearly shows that it may not be sufficient to consider the consensus only. Little attention has, however, been put on how to do the segmentation in projective mapping.

In (Vidal et al., 2016) the clustering was based on the consumers' correlations to the MFA components from a global MFA model using a hierarchical agglomerative clustering (HAC). Hence, the consumers

* Corresponding author.
E-mail address: ingunn.berget@nofima.no (I. Berget).

https://doi.org/10.1016/j.foodqual.2018.05.007
Received 21 December 2017; Received in revised form 24 April 2018; Accepted 18 May 2018
Available online 19 May 2018

were considered similar if they had high contributions to the same MFA components, this means that similarity between two assessors will depend on the other assessors present in the data.

In this work, we present two alternative methods which do not depend on a global model, both are based on Procrustes rotations (Gower, 1975; Gower & Dijksterhuis, 2007). Procrustes analysis is particularly suited for this purpose since it removes rotation and scaling effects. The first alternative, is based on a hierarchical clustering strategy, whereas the second utilises a sequential approach for identifying clusters. The hierarchical strategy was developed for outlier detection in traditional sensory panels, under the name Proclustrees (Dahl & Næs, 2004). In this method, the Procrustes distance is computed between all pairs of assessors, and then HAC is applied for the segmentation. This method is expected to be useful for projective mapping as well, although it has not been explored yet.

The second approach is based on the sequential partitioning strategy proposed in (Berget, Mevik, Vebo, & Naes, 2005; Dahl & Næs, 2009). The sequential approach first identifies the most obvious segment the "good" cluster, removes this as the first cluster and then repeats this procedure until a maximum number of clusters is identified. For consumer studies this is advantageous as the data often have high noise levels and there may be several assessors without any clear structure in their data. For each step in the sequence a GPA-like criterion is used for identification of the good cluster for each sequence. The proposed method is named *Sequentially Clusterwise Rotation (SCR)*.

In this work Proclustrees and *SCR* are presented as new methods for segmentation of projective mapping data. The proposed methodology is compared with the MFA approach previously applied in the literature, for both for real and simulated data. Results will be summarised, and finally advantages and disadvantages of the different approaches will be discussed.

2. Methodology

2.1. Clustering PM data with MFA and hierarchical cluster analysis

Multiple factor analysis (MFA) (Abdi, Williams, & Valentin, 2013; Escofier & Pagès, 1994) can be used for analysis of several blocks or tables of data and is frequently used for data from projective mapping (PM). In MFA, all the coordinates from the consumers are organized into a wide matrix with N rows and $2 \times M$ columns where M consumers test N products.

Let $\mathbf{X}_k$ denote the data table for consumer k. Then each $\mathbf{X}_k$ is centred and scaled by its first singular value obtaining $\mathbf{Z}_k$. The concatenated matrix $\mathbf{Z} = [\mathbf{Z}_1 \ldots \mathbf{Z}_M]$ is then analysed by PCA according to the model.

$$\mathbf{Z} = \mathbf{T}\mathbf{P}^{\mathbf{t}} + \mathbf{E} \tag{1}$$

The $\mathbf{T}$ represents the scores, $\mathbf{P}$ the loadings each with A components, and $\mathbf{E}$ the residuals. Note that A can be larger than two, even though each matrix $\mathbf{Z}_k$ only has two columns.

Frequency tables of words used to describe samples can be added as supplementary tables, to guide interpretation of the consensus (Abdi et al., 2013). The supplementary variables (word frequencies) can be projected on to the principal components.

There are different ways of using results from a global model such as MFA for clustering consumers. Here we will apply the same approach as in (Vidal et al., 2016) where pairwise distances between consumers were computed as the Euclidean distances between so-called consumer coordinates with A components. The consumer coordinates show how much contribution the consumers have to the components. For component a, and consumer k, these can be computed as

$$y_{ka} = \sum_{j}^{J_k} \mathbf{P}_{ja}^2 \tag{2}$$

i.e. the sum of squared loadings for the variables in table k. For general applications of MFA, the number of variables per block (J_K) may vary, but for projective mapping $J_k = 2$ for all k ($k = 1,\ldots, M$). The distance between consumer k and l, for A components are computed as the Euclidean distance between the vectors $\mathbf{y}_k = [y_{k1} \ldots y_{kA}]$ and $\mathbf{y}_l = [y_{l1} \ldots y_{lA}]$.

Clustering can then be performed by hierarchical clustering (HAC) on the distance matrix computed from the consumer coordinates $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_M$… Here we apply the ward linkage as in (Vidal et al., 2016), however, other linkage methods can be applied as well. In the following this approach will be referred to as MFA-HAC.

The MFA based approach is simple, since when the consumer coordinates (Eq. (2)) are computed, the cluster analysis can be performed in any statistical software using standard methods. Potential problems are, however, the dependency on the number of components, as well as on the global model because the consensus may be misleading if there are outliers in the data, or when groups of consumers base their mapping on very different characteristics of the products. In addition, hierarchical methods are known to be prone to noise and outliers, and some studies indicate that partitions methods are better than hierarchical methods (Wajrock, Antille, Rytz, Pineau, & Hager, 2008).

2.2. Proclustrees

A simple way of implementing the Procrustes distance for segmentation of PM data is the Proclustrees method proposed in (Dahl & Næs, 2004). This was originally proposed for outlier detection in descriptive analysis (DA), but is potentially useful for projective mapping as well. The method is based on establishing the Procrustes distance between all pairs of individuals, and then running a hierarchical clustering on the distance matrix. Hence the clustering *strategy* is the same as for MFA-HAC, whereas the *criterion* for similarity is different and only based on raw individual data.

The Procrustes distance between two assessors k and l is given by

$$\widetilde{D}_{kl} = \min\|p\widetilde{\mathbf{X}}_k\mathbf{R} - \widetilde{\mathbf{X}}_l\| \tag{3}$$

where $\widetilde{\mathbf{X}}$ denotes matrix $\mathbf{X}$ after centring and scaling so that $tr(\widetilde{\mathbf{X}}_k^t\widetilde{\mathbf{X}}) = 1$, $\|\cdot\|$is the Frobenius norm, p is an isotropic scaling factor and $\mathbf{R}$ a rotation/reflection matrix that minimises the distance between the two configurations. More details are given in (Dahl & Næs, 2004). When $\widetilde{D}$ is computed for all pairs of k and l, HAC is applied as for any other distance matrix. In the present paper we have applied the ward linkage. Again, this is only one of several options. The group structure can be studied by inspection of the dendrogram, and the consensus for each cluster can be computed after they are identified using MFA, INDSCAL or GPA. In the following, the acronym PT will be used for the Proclustrees method. MFA was applied for computing cluster consensus.

A clear advantage of the Proclustrees method is that is does not rely on any additional parameters, hence it is easy to use. Potential drawbacks related to hierarchical methods is the same as for MFA-HAC described above.

2.3. Sequential partitioning by clusterwise rotations (SCR)

2.3.1. The sequential strategy with fuzzy clustering and the noise cluster modification

The sequential approach was introduced in (Berget et al., 2005), and was also discussed in (Dahl & Næs, 2009). The basic idea, is to extract and shave off one cluster at a time, then repeat this procedure for the remaining data to find the next cluster. The division into cluster and "noise" is done until a maximum number of clusters are identified, or until there are too few observations left in the data. Typically, a part of the data will remain as a final "rest" cluster, comprising consumers that do not form a cluster.

The partition of data into "good" and "noise" cluster is achieved by