Comparison of different ways of handling consumer segments using L-shape data

Article

Accepted Version

Nguyen, Q. C., Asioli, D. ORCID: https://orcid.org/0000-0003-2274-8450, Varela, P. and Næs, T. (2023) Comparison of different ways of handling consumer segments using L-shape data. Journal of Sensory Studies, 38 (4). e12836. ISSN 1745-459X doi: 10.1111/joss.12836 Available at https://centaur.reading.ac.uk/111900/

Comparison of Different Ways of Handling Consumer Segments using L-shape Data

Quoc Cuong Nguyen[1,2*], Daniele Asioli[3], Paula Varela[4], Tormod Næs[4]

[1]Department of Food Technology, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam.

[2]Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam.

[3]Department of Agri-Food Economics and Marketing, School of Agriculture, Policy, and Development,

University of Reading, Reading, United Kingdom.

[4]Nofima AS, Ås, Norway.

* Corresponding author: nqcuong@hcmut.edu.vn

ABSTRACT

Different approaches for handling consumer segments in L-shape data are compared in a study conducted in Norway. Consumers evaluated eight different yoghurt samples with profiles varying in three intrinsic attributes following a full factorial design. Three blocks of data were collected including sensory properties, liking ratings, and consumer attributes. Data were analysed using two different approaches. In approach one, the one-step simultaneous L-Partial Least Square (L-PLS) Regression with average consumer liking to represent the segments was used, while approach two was based on a two-step procedure (TSP) based on Partial Least Square (PLS) Regression using dummy variables to represent the segments. The methods were compared in terms of interpretations, flexibility, and outcomes. Methodological implications, recommendations, and future research avenues are discussed.

PRACTICAL APPLICATIONS

This manuscript has been devoted to two different ways of handling segmentation in L-shape data of consumer liking, sensory properties, and consumer attributes. Overall, both L-PLS and TSP approaches provide similar interpretation of results. The TSP approach, however, has the advantage of interpretating the horizontal and vertical direction in the L separately using standard regression methods. It is of interest of product development and marketing activities to identify which food product characteristics are important for consumer preferences and to better understand the characteristics of the consumers (e.g., socio-demographics) that drive the consumer acceptance of the different products.

Keywords: Individual differences; L-shape data; Method comparison; One-step L-PLS; Segmentation; Two-step TSP; Yoghurt.

1. INTRODUCTION

Often, in the analysis of consumer liking data, one is interested not only in the liking data themselves, but also in how liking ratings relate to sensory properties of the food products and consumer attributes, such as socio-demographics, attitudes, and habits. The data sets for such situations can be formulated within a so-called L-shape as depicted in Figure 1. In these types of datasets, the consumer liking data (**Y**) are linked to sensory properties (**X**) along the horizontal dimension, and to the consumer attributes (**Z**) along the vertical dimension. It is common that a set of I products have been assessed by a set of J consumers, e.g. with respect to degree of liking. In addition, each of the I products have been measured by K product properties, reflecting chemical or physical measurements, sensory properties, etc. Moreover, each of the J consumers have been characterised by L consumer attributes, comprising individual characteristics like socio-demographics variables like gender, age, income, etc., as well as the individual's general attitudes, consumption habits (Lengard & Kermit, 2006; Martens et al., 2005). The information obtained from investigating all three data sets and their links is important for product developers and marketers to improve product properties, product communication, and marketing strategies of new food products (Asioli, Nguyen, Varela, & Næs, 2022).

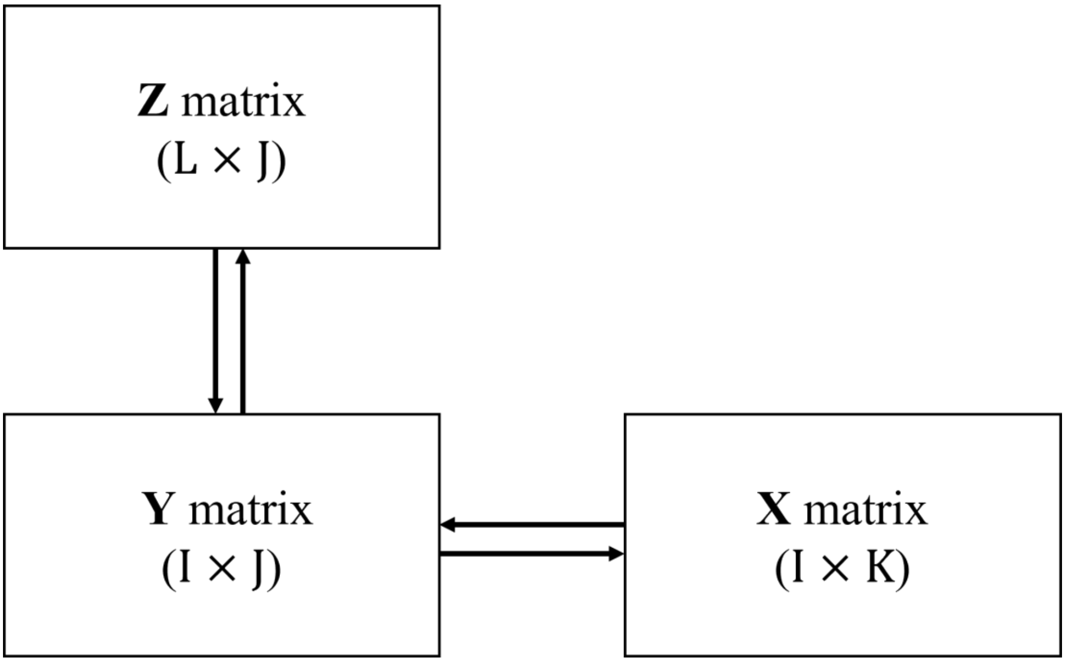

Figure 1. L-shape data: sensory properties – *X* matrix (*I* products × *K* sensory properties), consumer liking ratings – *Y* (*I* products × *J* consumers), and consumer attributes – *Z* matrix (*L* consumer attributes × *J* consumers).

Several studies have investigated L-shape data. For example, Martens et al. (2005) investigated sensory properties of different apple cultivars, consumer degree of liking of such products, and consumer attributes (e.g. food choice, consumption frequency, age, gender, etc.) in Denmark. Frandsen, Dijksterhuis, Martens, and Martens (2007) conducted a sensory and consumer study (authenticity test and questionnaire comprised of demographic, willingness to buy) by investigating different types of milk in Denmark using L-shape data. Pohjanheimo and Sandell (2009) investigated sensory properties of different yoghurts, consumer degree of liking of such products, and consumer attributes (e.g., food choice questionnaire, consumers' concerns about food and health) with Finnish consumers. Asioli et al. (2022) performed a sensory and consumer study (liking ratings, consumers' attitudes toward the health and taste characteristics of foods) by investigating different types of yoghurt in Norway using the same L-shape data.

L-shape data can be analysed in different ways, see for example Smilde, Næs, and Liland (2022); Vinzi, Guinot, and Squillacciotti (2007). Here we will focus on a one-step approach called L-Partial Least Squares (L-PLS) regression, and a two-step procedure (TSP) using standard Partial Least Squares (PLS) regression methods along the horizontal and vertical direction in the L-shape separately. The two approaches have been compared and found to give similar results in Asioli et al. (2022), but more studies are needed to better understand the differences and similarities of these methods, aiming at generalising recommendations.

A closely related aspect which also needs more research is how to analyse L-shape data in a context of segmentation. The TSP approach has previously been used for this purpose using a dummy variable coding of the segments in the second step (Asioli, Næs, Granli, & Lengard Almli, 2014; Næs, Varela, & Berget, 2018). In other words, the segments are represented in a matrix coded with 1's and 0's to represent the segments for all consumers and regressed onto the consumer characteristics. To the best knowledge of the authors there is no research of this type based on the one-step L-PLS approach. In this manuscript, we propose an alternative approach for this methodology based on using the average degree of liking for consumers in the same segment as the new matrix **Y**. This approach can be used both for a priori and a posteriori segmentation. (see e.g., Næs, Varela, & Berget, 2018).

The main aim of this manuscript is to investigate the one-step L-PLS approach where the consumer liking data are replaced by average degree of liking for each segment separately, as compared to the two-step procedure (TSP), which is an already established benchmark, based on dummy coding for the segments (Asioli et al., 2014). The main contribution of this manuscript lies in the methodology for analysis of segmented data using the L-PLS approach.

The segmentation used is interpretation-based clustering, which we find appealing, but any other ways of clustering, for example, automatic segmentation of L-shape data(Endrizzi, Gasperi, Calò, & Vigneau, 2010; Vigneau, Endrizzi, & Qannari, 2011) could also have been used for this illustrative purpose. The method will be tested on data from an experiment investigating consumers' preferences for yoghurts in Norway (same data set as used in Asioli et al., 2022). Issues related to interpretations, flexibility, and outcomes of the two approaches will be compared and discussed. Some discussion will also be given on the comparison of the conventional L-PLS approach and the new procedure for the same data.

The manuscript is structured as follows: first the statistical methods used are described, second, the methodological approach is illustrated, including the experimental design, and data analysis. Then, we will present and discuss the results and provide methodological implications and recommendations as well as outline some future research avenues.

2. THEORY: STATISTICAL METHODS

In this section we will briefly describe the basic theories of the statistical methods used in this manuscript, such as the one-step L-PLS, and the two-step procedure (TSP) approaches.

In the L-shape data set, the matrix $\mathbf{Y}(I \times J)$, represents the degree of liking ratings given by J consumers for I products, the descriptive sensory data $\mathbf{X}(I \times K)$, contains intensities for K sensory properties of the same I products. The data set $\mathbf{Z}(L \times J)$ represents the L attributes for the J consumers (i.e., consumer attributes).

2.1 A priori vs a posteriori segmentation

A priori segmentation means that the consumer segments are determined before data analysis starts (Næs et al., 2018). One may for instance be interested in comparing results for women and men or old, and young consumers. A posteriori segmentation means that segments are determined based on the data, either by e.g., cluster analysis (CA) or visual interpretation of PCA plots (Principal Component Analysis) based on the consumer degree of liking data (Næs et al., 2018). In this manuscript the focus will be on the latter method. For the visual interpretation based on PCA, the data are either organised with consumers as rows and products as columns or vice versa. Then, the segmentation is conducted based on interpreting the scores and loadings and focusing on the pattern one is most interested in. When a priori segmentation is used, the methods below will have to eliminate the segmentation variable in the analyses (second step of TSP) in order to avoid the double use of a variable.

It is important to emphasise that this clustering method chosen is different from what is used in many other application areas. For instance, Chang (1983) is generally sceptical about using principal components for clustering. Similar viewpoints can be found in Witten and Tibshirani (2010); Green and Krieger (1995); Yeung and Ruzzo (2001). Our situation is, however, different from those considered in these cases: consumer liking data are always very noisy, as it is subjective data where each consumer has his/her own opinion and uses the scale in different ways. In such cases one will seldom find any information of interest, except noise, in components beyond for instance 3.

As was emphasised in for instance Næs et al. (2018) and also shown in the example below there is often no clear cluster tendency in liking data, only a continuum of individual liking differences. This means that the outcome of an automatic clustering procedure may be uncertain and unstable due to the lack of a clear cluster structure. The result will depend on criterion/distance (for instance Euclidean, Mahalanobis or others) and procedure (hierarchical or criterion based) used. This has been demonstrated in for instance Endrizzi, Gasperi, Rødbotten, and Næs (2014); Castura, Meyners, Varela, and Næs (2022) (see also Yenket and

Chambers IV (2017); Yenket, Chambers IV, and Johnson (2011)). The results may therefore depend heavily on sometimes arbitrary decisions (criterion and procedure) made prior to analysis. It is therefore often safer and closer to a user's need to use interpretation-based segmentation based on what is seen in PCA plots and what is meaningful to consider. This strategy can be seen as more transparent and more directed towards an interpretable perspective of interest. We refer to Endrizzi et al. (2014); Almli et al. (2011); Rødbotten et al. (2009) for other applications based on PCA and visual interpretation for clustering.

Since the main purpose of the manuscript is to analyze clustered L-shape data, any other clustering could have been used. The clustering will here be validated by checking the interpretation of the clusters using simple columns plots.

2.2 ANOVA for investigating product average liking

The methods discussed in this research focus on consumer degree of liking for individual consumers or segments of consumers. However, in most cases one will also be interested in analysing the average degree liking of products. This can be done using the Analysis Of Variance (ANOVA) model:

$$y_{ij} = \mu + \alpha_i + C_j + \varepsilon_{ij} \qquad (1)$$

where i refers to product, j refers to consumer, y_{ij} is the $(ij)^{\text{th}}$ observation, μ is the general mean and the α_i's are the fixed main effects of the product factor. The C_j's represent the random main effects of the consumers, and e_{ij} is the independent random noise. One is interested in both the product differences themselves and how significantly different these differences are.

As an alternative to visual segmentation based on PCA of raw data it was advocated in Endrizzi, Menichelli, Johansen, Olsen, and Næs (2011) that the residuals from model (1) above may sometimes be easier to use for highlighting differences in preference pattern among the consumers (Almli et al., 2011; Endrizzi et al., 2011; Hersleth, Lengard, Verbeke, Guerrero, & Næs, 2011). We therefore chose this approach. The residuals are double centered.

2.3 Data analysis of L-shape data: standard methods in situations without segmentation

2.3.1 One-step L-PLS regression

The L-PLS regression approach introduced by Martens et al. (2005) is based on one single analysis combining all the three blocks of data together (i.e., sensory properties, consumers' degree of liking ratings, and consumers' attributes) (Vinzi et al., 2007). The matrices **X** and **Z** are centred for properties and attributes respectively, while matrix **Y** is supposed to be centered with respect to both its rows and its columns (double centered). The L-PLS regression method used here is based on components calculated from the first singular vectors of the *Singular Value Decomposition* (SVD) of $\mathbf{X}'\mathbf{Y}\mathbf{Z}'$ with deflation, i.e., only the first singular vector is then used in each SVD computation of residual (Martens, 2005). L-PLS regression can be arranged as *endo*-L-PLS or *exo*-L-PLS depending on how the deflation is done (see Martens et al., 2005 and Sæbø, Martens, and Martens, 2010 for more details). For a recent application of the L-PLS approach, we refer to Asioli et al. (2022).

The relations between three blocks of data **X** (sensory properties), **Y** (consumers' degree of liking ratings), and **Z** (consumers' attributes) can be shown in the correlation loadings plot (Martens et al., 2005). In case of the *endo*-L-PLS, **X** (or **Z**) correlation loadings are calculated by correlating the **X** (or **Z**) variables onto **X** (or **Z**) scores. Both columns and rows of **Y** are regressed onto the two sets of scores to obtain correlation loadings (Sæbø et al., 2010).

Since Y is double centred, information about the actual liking of the different products is less visible in the plot as compared to in standard preference mapping. Therefore, it is good practice to add the results from the ANOVA described above to obtain a more comprehensive interpretation.

2.3.2 Two-step Procedure (TSP)

The TSP approach is based on the PLS regression performed according to the following procedure. In *step 1,* PLS regression is used for linking sensory properties (**X**), and consumer degree of liking (**Y**) using either **Y** or **X** as response corresponding to external and internal preference mapping, respectively. Internal preference mapping is based on first using PCA for the centered consumer liking data, and then regressing centered sensory data onto the principal components. External preference mapping is based on first using PCA of the sensory data before the liking values for the individual consumers are regressed onto the principal components of the sensory profiles. Detailed explanation of preference mapping is described in the literature (McEwan, 1996; Næs, Brockhoff, & Tomic, 2010; Næs et al., 2018). In *step 2*, a PLS regression model is used for relating the consumer loadings from the *step 1* to the consumer attributes in **Z**. For a detailed description of the TSP approach we refer to Næs et al. (2018).

2.3.3 Comparison of the one-step L-PLS regression and the two-step Procedure (TSP)

The two approaches presented here of analysing L-shape data have both similarities and differences. In Asioli et al. (2022) it was found that the two methods provide very similar results for interpretation of a data set based on yogurt samples. Regarding the differences, the two approaches differ in the way interpretation is done. Indeed, in the one-step L-PLS approach the results are visible all in one single plot while for the TSP approach the interpretation should be based on multiple plots which can be more cumbersome. On the other hand, the TSP approach is based on more well-known methods and the interpretation can be done sequentially for the horizontal and vertical direction in the L-shape. The L-PLS is based on double centred Y-data which may make it less intuitive to interpret (see Asioli et al., 2022). Adding results from the ANOVA above is therefore useful for a more comprehensive interpretation. The TSP can be used both for raw consumer liking data and for double centred data.

2.4 Incorporation of consumer segments in the analysis

In this section, we will propose a new way of using average degree of liking in segments for the L-PLS and discuss an established way of using TSP for incorporating segments in L-shape data (Asioli et al., 2014; Smilde et al., 2022) using a dummy matrix (a 0/1 matrix) to represent segments. We focus on the visual segmentation, but some automatic segmentation in the context of L-shape data can be found, for example, in Endrizzi et al. (2010).

2.4.1 Y-average approach for L-PLS

The **Y**-average matrix will have as many rows as there are products and as many columns as there are consumers (see Table 1 taken from the empirical study below). With this approach the matrix **Y** has a similar structure as for the original liking data (products $\times$ consumers), but the average likings of consumer segments are used instead of original liking values. For example in Table 1, the first, fourth and fifth column are given the same values since consumer C1001, C1004, and C1006 belong to the same segment. Note that the **Y**-average approach can be used in both the TSP, and the L-PLS regression approaches, but here it will only be used for the latter.

Table 1. An illustration of matrix Y-average with products in rows, and consumers in columns. Consumers belonging to the same segment have the same liking values.

PRODUCT	C1001	C1002	C1003	C1004	C1006	C1007	C1008	C1009
thin_fla_low	-4.53	-18.16	-18.16	-4.53	-4.53	-18.16	-21.51	-21.51
thick_fla_low	0.42	3.17	3.17	0.42	0.42	3.17	15.58	15.58
thin_flo_low	-11.13	-1.21	-1.21	-11.13	-11.13	-1.21	-16.47	-16.47
thick_flo_low	-2.05	9.18	9.18	-2.05	-2.05	9.18	8.07	8.07
thin_fla_opt	7.15	-15.95	-15.95	7.15	7.15	-15.95	-1.29	-1.29
thick_fla_opt	4.65	10.02	10.02	4.65	4.65	10.02	19.54	19.54
thin_flo_opt	1.09	-2.63	-2.63	1.09	1.09	-2.63	-13.57	-13.57
thick_flo_opt	4.40	15.58	15.58	4.40	4.40	15.58	9.65	9.65

2.4.2 The Y-dummy approach for TSP

A simple method which has been used for TSP is to relate the *consumer segments* represented as dummy variables (**Y**-dummy) to the consumer attributes (**Z**) in the second step using some type of discriminant analysis, for example PLS discriminant analysis (Asioli et al., 2014; Asioli et al., 2022; Endrizzi et al., 2011). In the dummy **Y**-approach, a matrix of 0/1 response values are generated based on cluster membership. The matrix has as many rows as there are consumers, and as many columns as there are segments. Then, for each consumer a 1 is placed in the column corresponding to the segment that the consumer belongs to.

Table 2. An illustration of matrix Y-dummy with consumers in rows, and segments in columns. Consumers belonging to a cluster have the 1's, otherwise 0's.

CONSUMER	CLUSTER 1	CLUSTER 2	CLUSTER 3
C1001	1	0	0
C1002	0	1	0
C1003	0	1	0
C1004	1	0	0
C1006	1	0	0

C1007	0	1	0
C1008	0	0	1
C1009	0	0	1

3. MATERIALS & METHODS

This section briefly describes the methodology applied in this manuscript, including the description of participants, products, consumer tests, sensory description, consumer attributes, and statistical data analysis. More detailed information can be found in Asioli et al. (2022).

3.1 Participants

3.1.1 Participants

One hundred and one Norwegian consumers participated in a study in October 2017 at Nofima AS (Ås, Norway). Only consumers who regularly consume yoghurt at least once a month were included in the study. All data were collected with EyeQuestion (Logic8 BV, The Netherlands).

The manuscript is written according to Nofima's ethical standards and code of conduct as set down by the Ethical Board of Nofima As. The manuscript is designed and written in accordance with the guidelines laid out in the Declaration of Helsinki (revised 2008). All participants signed an informed consent and were free to withdraw from the studies at any time without providing a reason for withdrawal and without penalty.

3.2 Products

Several yoghurt samples were prepared following an experimental design based on the same ingredients, but with varying in texture, including three intrinsic attributes with two levels each: viscosity (thin/thick), particle size (flake/flour), and flavour intensity (low/optimal). The flavour is added as follows: optimal samples (1000 grams yoghurt with 0.5 grams vanilla and 0.25 grams acesulfame potassium), low samples (1000 grams yoghurt with 0.25 grams vanilla and 0.125 grams acesulfame potassium). The samples had the same calories and composition, and were originally formulated with the purpose of studying satiety expectations driven by food texture, for more details see Nguyen, Næs, and Varela (2018). Table 3 shows the samples with different levels of viscosity, particle size, and flavour intensity.

Table 3. Formulation of yoghurts and the symbols used in plots.

SAMPLE	VISCOSITY	PARTICLE SIZE	FLAVOUR INTENSITY
thin_fla_low	Thin	Flakes	Low
thick_fla_low	Thick	Flakes	Low
thin_flo_low	Thin	Flour	Low
thick_flo_low	Thick	Flour	Low
thin_fla_opt	Thin	Flakes	Optimal
thick_fla_opt	Thick	Flakes	Optimal
thin_flo_opt	Thin	Flour	Optimal
thick_flo_opt	Thick	Flour	Optimal

3.3 Consumer test

The consumer test was held in the sensory laboratory of Nofima AS. Consumers were asked to taste each of the eight samples, and rate their degree of liking using a Labeled Affective Magnitude (LAM) scale (Schutz & Cardello, 2001). Consumer attributes data (i.e., health and taste attitudes and socio-demographics) were also collected.

All the sensory evaluations were conducted in standardized individual booths according to ISO 8589:2007. See Nguyen, Næs, Almøy, and Varela (2020) for more details.

3.4 Sensory description: Quantitative descriptive analysis

Sensory profiling of the eight samples was performed via quantitative descriptive analysis following a generic descriptive analysis procedure (based on QDA), as described by (Lawless & Heymann, 2010; Stone, Bleibaum, & Thomas, 2012). The final list of sensory properties used in the experiment included six odours (*total intensity of all odours, acidic, vanilla, stale, sickening/cloying,* and *oxidized*), three tastes (*sweet, acidic,* and *bitter*), six flavours (*total intensity of all flavours, sour, vanilla, stale, sickening,* and *oxidized*), and six textures (*thick, full, gritty, sandy, dry,* and *astringent*). The sensory properties with definition are referred from Asioli et al. (2022), and Nguyen et al. (2018).

3.5 Consumers' attributes

Several consumers' attributes were also collected, such as consumers' attitudes toward the health and hedonic characteristics of foods (Roininen, Lahteenmaki, & Tuorila, 1999) by including the three health-related factors (*general health interest*, *light product interest*, and *natural product interest*), and the three taste-related factors (*craving for sweet foods, using food as a reward*, and *pleasure*). In addition, consumers' socio-demographics were collected. For more details, see Nguyen et al. (2020) and Asioli et al. (2022).

3.6 Statistical data analysis

3.6.1 Approach one: The L-PLS approach with average liking for each segment

We use *endo*-L-PLS, reflecting the *inward-pointed regression* of a single response **Y** from two outer predictors (**X** and **Z**) as illustrated in Martens et al. (2005) and Mejlholm and Martens (2006). The matrices **X** and **Z** are centered and standardized, **X** for each sensory properties, and **Z** for each consumer attribute. The matrix **Y** of averages is then subjected to a double centering across both rows and columns. For details of *endo*-L-PLS and centering, interested readers are referred to Sæbø et al. (2010).

3.6.2 Approach two: TSP – with dummy Y-matrix

In this approach, to compare with the approach one (the L-PLS approach), the matrices **X** and **Z** are also centered and standardized. The matrix **Y** is centered across rows (i.e., column-centered). In the *first step*, a standard PLS regression is run with consumers degree of liking as **Y** and sensory properties as **X**. In the *second step*, a matrix **Y**-dummy (1 in the position corresponding to the segment that a consumer belongs to, and 0 elsewhere) is regressed, using PLS, onto the matrix **Z** (consumer attributes). The regression and corresponding scatter plots illustrate the relation between consumer attributes, and consumer segments.

The computations are done in R version 4.2.2 (R Core Team, 2022) using the package *multiblock* (Liland, 2022) and in-house codes.

4. RESULTS

4.1 ANOVA of the liking data

The overall liking of products is shown in Figure 2. The thick products are the most liked, the products with optimal flavour intensity are most liked within both thick and thin products.

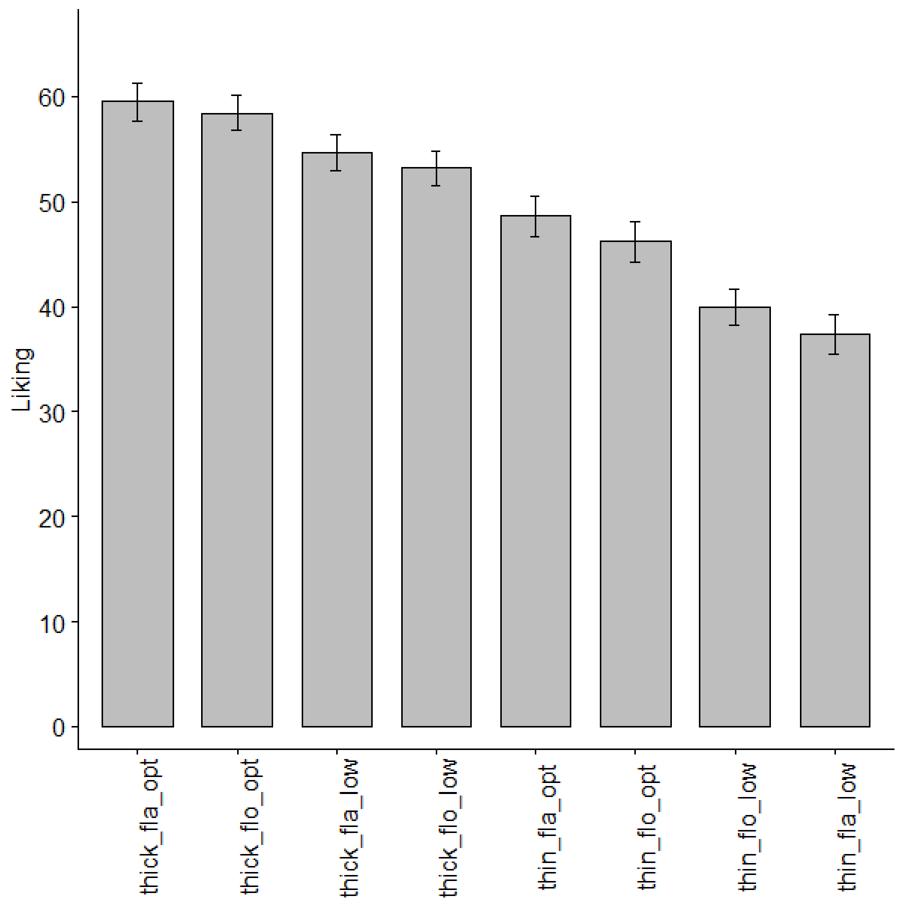

Figure 2. Overall liking of products.

Table 4 shows the results of the ANOVA model (1) with product and consumer factors. We can see that both factors product and consumer were significant for the degree of liking (i.e., p-values < 0.001).

Table 4. Results from ANOVA model.

Response: Degree of liking	Mean Sq	Sum Sq	Df	F value	Pr(>F)
Product	67779	47456	7	30.84	0.0000
Consumer	1047	104735	100	4.76	0.0000
Residuals	220	153902	700		

4.2 Visual segmentation based on residual data

The residuals from the model (1) were computed, then they were put into a matrix with the rows corresponding to products and the columns corresponding to the consumers (8×101).

Then, PCA was run on the matrix of residuals to obtain the score (Figure 3) and loading (Figure 4) plots. The explained variances for the first two components were 44.3% of the total variance. The third component explained 16.5% of the variance giving about 60.8% explained variance after 3 components. This indicates that one should not put too much emphasis on components beyond 3.

The first component is strongly related to the viscosity of the yoghurts tested in which, on the right side, there were consumers who tended towards yoghurts with a thin viscosity (*thin_fla_low*, *thin_flo_opt*, *thin_fla_opt*) whereas those who appreciated a thick viscosity (*thick_flo_opt*, *thick_flo_low*, *thick_fla_low*, *thick_fla_opt*) were positioned on the left side except for product *thin_flo_low*. The second component is related to particle size of oat flakes added (flakes vs flour): negative values of this component were related to small particle size (yoghurts coded with *flo* as the second position of text) except for product *thin_fla_low*, and positive values were with large particle size (yoghurts coded with *fla*). The third component spreads differences in flavour perception i.e., low vs optimal flavour intensity (data not shown). This corresponds reasonably well to the experimental design of yoghurts in this study.

It is important to emphasize that because of double centering for instance the consumer to the right in the plot do not necessarily prefer the thin yoghurts, they lie simply more in this direction than the average consumer. The same holds for the other interpretations above.

The segmentation chosen for visualisation in this research (based on the two dominant components) is shown in Figure 3 and Figure 4. In particular, the segment G1 (1) consisted of consumers with a higher liking for *thin* yoghurts as compared to the average consumer. Consumers in segment G2 (2) had a higher liking for *thick* yoghurts with *oat flour* added than the average consumer, and consumers in segment G3 (3) went more in the direction of *thick* yoghurts with *oat flakes* added. The differences can be also seen by plotting average likings of segments (Figure 5). As can be seen, this broadly confirms the above interpretation.

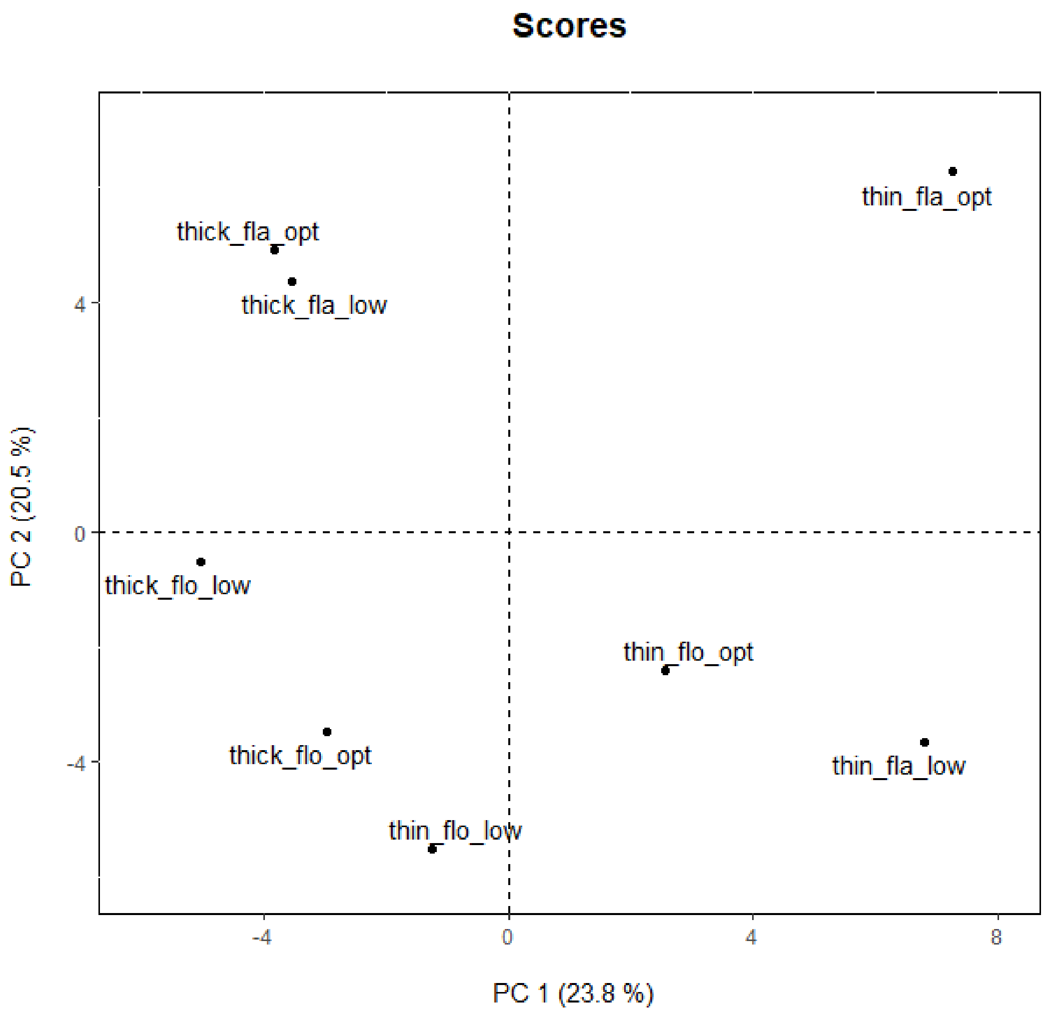

Figure 3. Scores plot with segment numbers from the PCA of the double centered residuals.

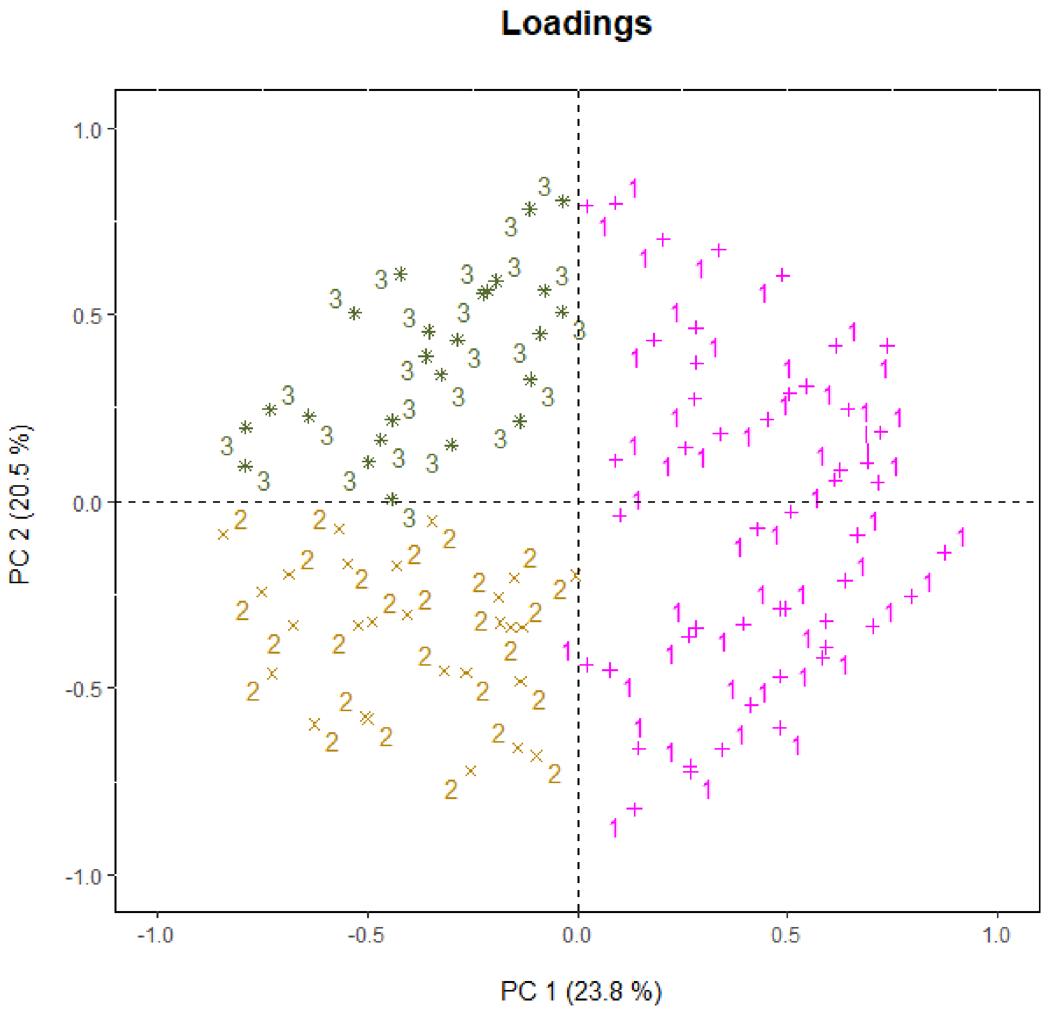

Figure 4. Loadings plot with segment numbers from the PCA of the double centered residuals.

It is important to note that this segmentation approach is only one of several methods that can be used. The actual segmentation chosen is also one of several that can be chosen depending on the focus of the study. The simple one used here must be considered merely as an illustration of the methodology. As could be seen, however, it is also meaningful for distinguishing between important preference differences.

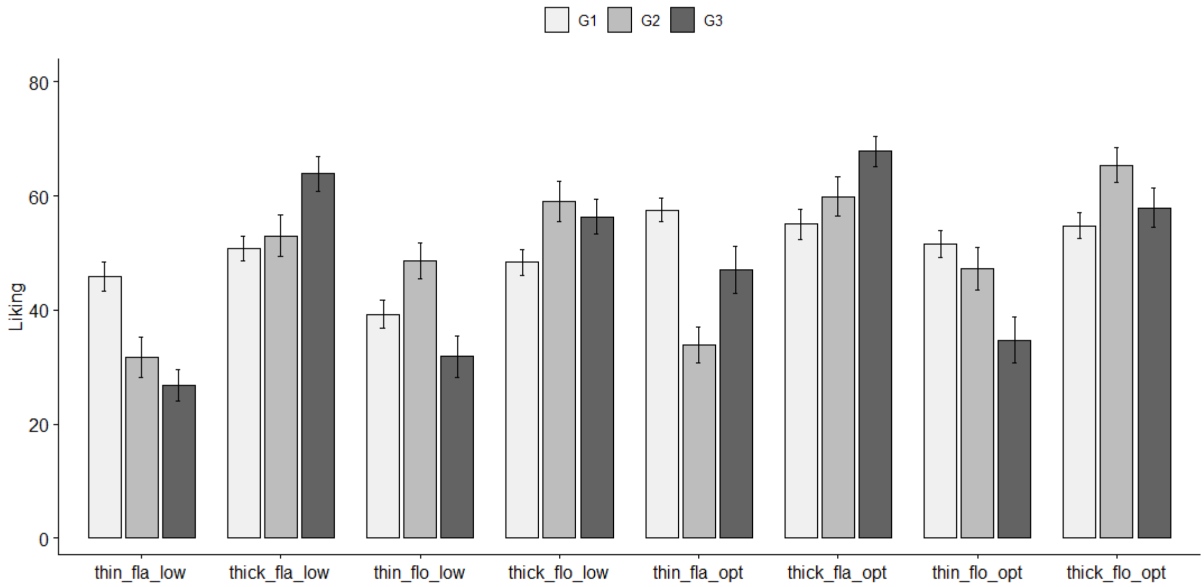

Figure 5. Overall consumer degree of likings of yoghurts in segments G1, G2, G3.

4.3 L-PLS approach based on average liking in segments.

The sensory description in Figure 6 shows that the first component (comp.1) is interpreted by both textures (*Sandy, Dry* on the left vs *Gritty* on the right), and flavours (*Oxidized*, *Bitter* on the left vs *Sour*, *Acidic* on the right). Note that *Vanilla*, and *Sweet* are located on the right of the component 1 and, to some extent, related to *Sour*, and *Acidic*. The second component (comp.2) is described by textures *Full* and *Thick* vs the property *Sickening* flavour. *Sickening* (cloying) flavour was more intense in the yoghurts with flour (small particles), and it may have been more distinguishable in the thin viscosity samples (*thin_flo_low* and *thin_flo_opt*).

Consumers in segment G1 liked thin yoghurts (*thin_fla_opt, thin_fla_low, thin_flo_opt*) described by *Sweet*, and *Vanilla* flavours better than the average consumer. These consumers were characterised by taste-related factors *craving for sweet foods* (e.g., cra_2, cra_4, cra_5) and *using food as a reward* (e.g., rew_1, rew_4, rew_5). Not surprisingly, this highlights those consumers driven by taste and food as reward preferring the sweeter yoghurts, and more intense in vanilla flavour. Consumers in segment G2 liked flour-added yoghurts (*thick_flo_low*, *thick_flo_opt*, *thin_flo_low*) with sensory perceptions *Bitter*, *Dry*, *Sandy*, *Astringent, Oxidized* better than the average consumer. Those consumers tended towards products with the attributes *Astringent* and *Oxidized* rather than more indulgent sensory properties of yoghurts, such as *Sweet*, *Vanilla,* and *Sour*. Possible explanation is that the G2 consumers pay more attention to textures than flavours. Furthermore, consumers in G2 did not have high values of taste-related factors as these were in the opposite direction of *craving for sweet foods* (e.g., cra_4, cra_6), and *using food as a reward* (e.g., rew_1, rew_2, rew_3). Consumers in segment G3 liked thick-flakes-yoghurts (*thick_fla_low, thick_fla_opt*) described by *Thick, Full* and, to some extent, *Gritty* better than the average consumer. The G3 consumers were characterised by health-related factors *general health interest* (e.g., gen_3, gen_4, gen_7, gen_8), *light product interest* (e.g., lig_3, lig_4). In addition, the G3 consumers lie close to the consumer attribute *age* (i.e., older consumers). Possibly, consumers in G3 pay more attention to health aspects of food consumption, and preferred products that could be perceived as healthier, i.e., big flakes may signal higher fibre content, while thicker yoghurts are perceived as more satiating.

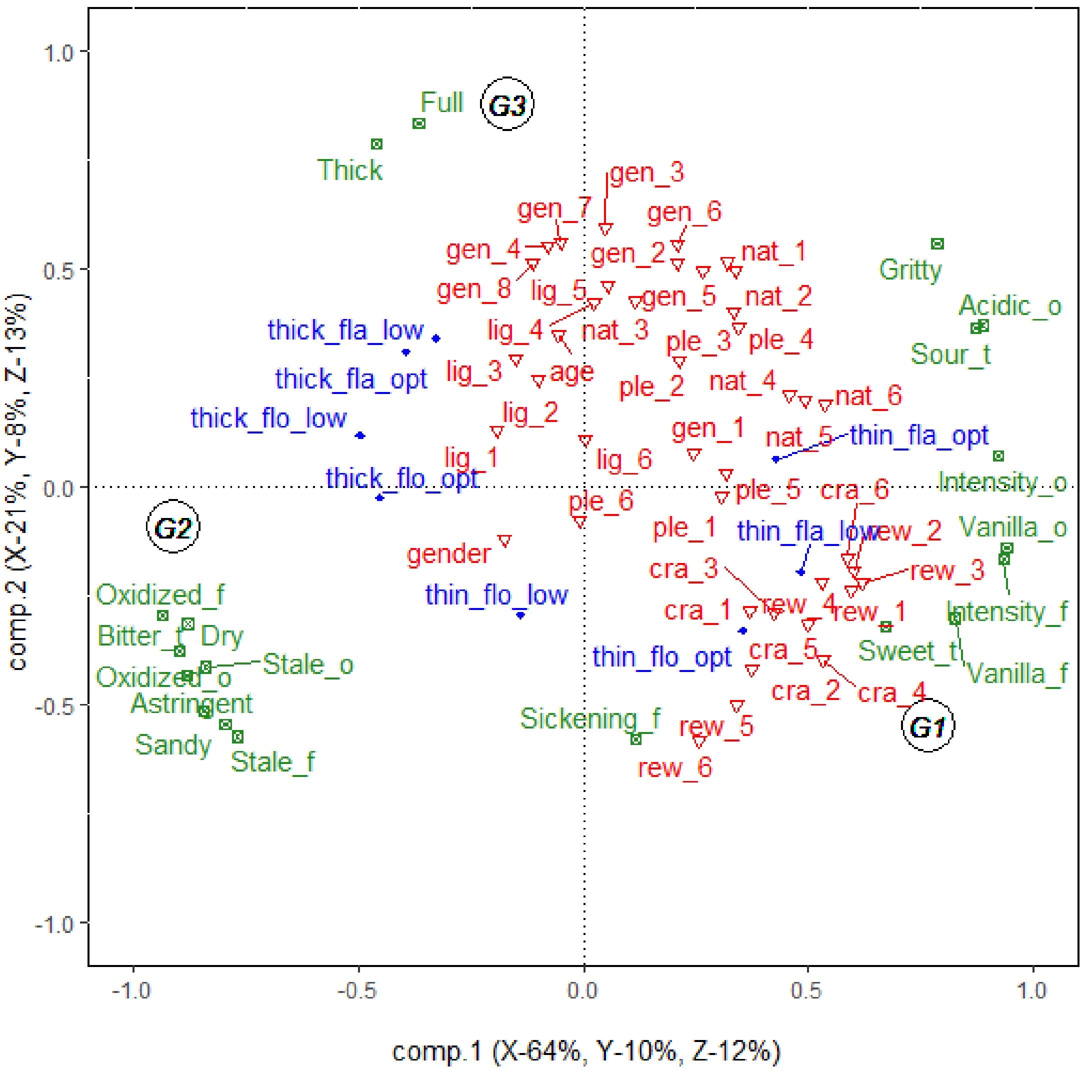

Figure 6. Endo-L-PLS. Sensory properties (X): centered and standardized for each property; Consumer degree of liking (Y): double-centered; Consumer attributes (Z): centered and standardized for each attribute. Endo-PLS plot shows consumer preferences for the eight yoghurts (in blue) in relation to both sensory properties (in green) and consumer attributes (in red); three consumer segments are shown as G1, G2, G3.

4.4 TSP based on dummy coding of the segments

Figure 7 and Figure 8 exhibit the relation between sensory properties and consumer degree of liking (i.e., consumers in different segments noted by different symbols). Consumers in segment G1 (on the right of component 1) preferred yoghurts described by *Gritty* and some flavours such as *Acidic_o*, *Sour_t*, *Sweet_t*, and *Vanilla_o*. These consumers did not prefer thick-viscosity yoghurts as the textures *Thick*, *Full* were located on the left side of the component 1. Consumers in segment G2 (on the left of component 1) preferred flour-yoghurts characterised by *Dry*, and *Sandy*. Consumers in segment G3 mostly preferred thick-flakes-yoghurts as they were close to *Thick*, and *Full*.

We can also clearly see from the comparison of scores and loadings that the products *thin_fla_opt*, *thick_flo_low*, and *thick_fla_low* in the score plot are the ones that are best liked, which corresponds to the bar plot of the average liking (Figure 5). Note that this information is not immediately available in the L-PLS approach without the addition of the ANOVA results.

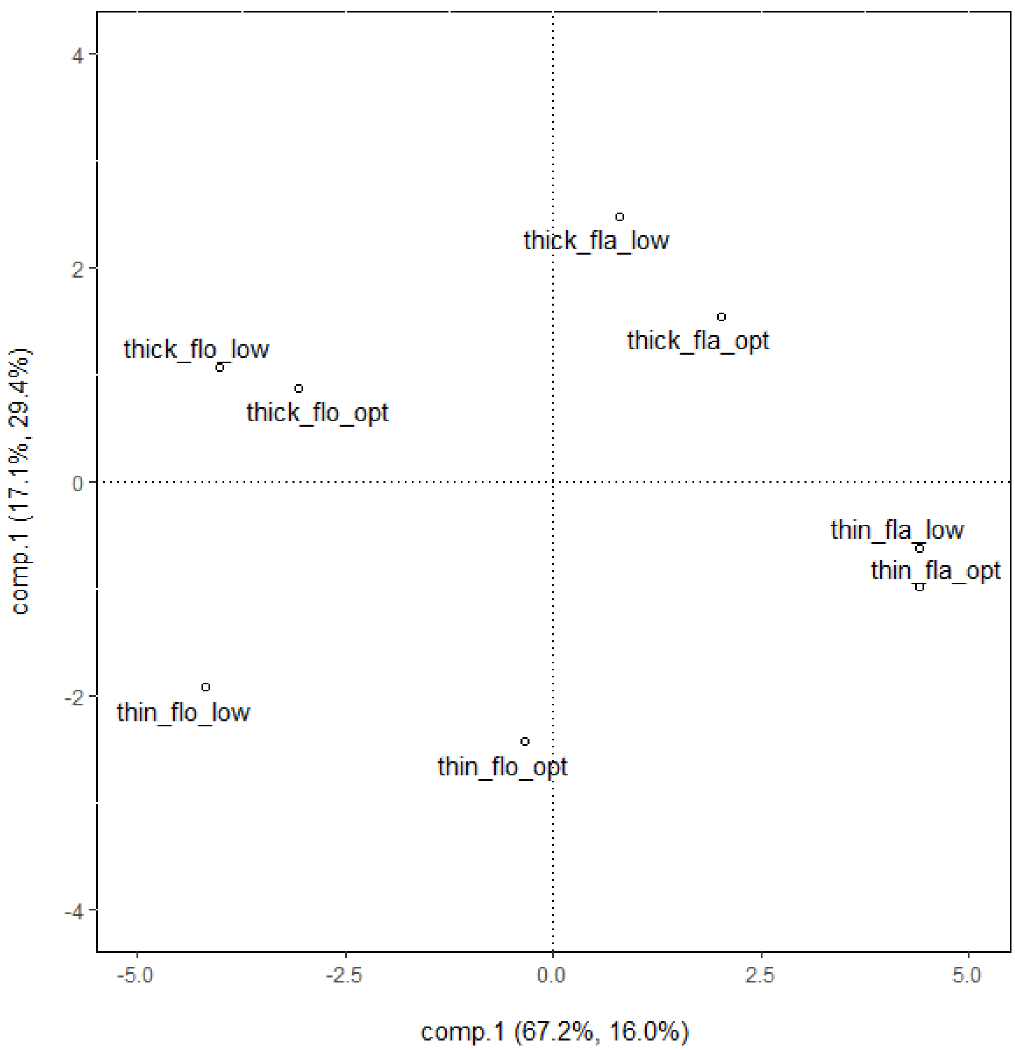

Figure 7. Score plot between Y data - consumer likings and X data - sensory properties.

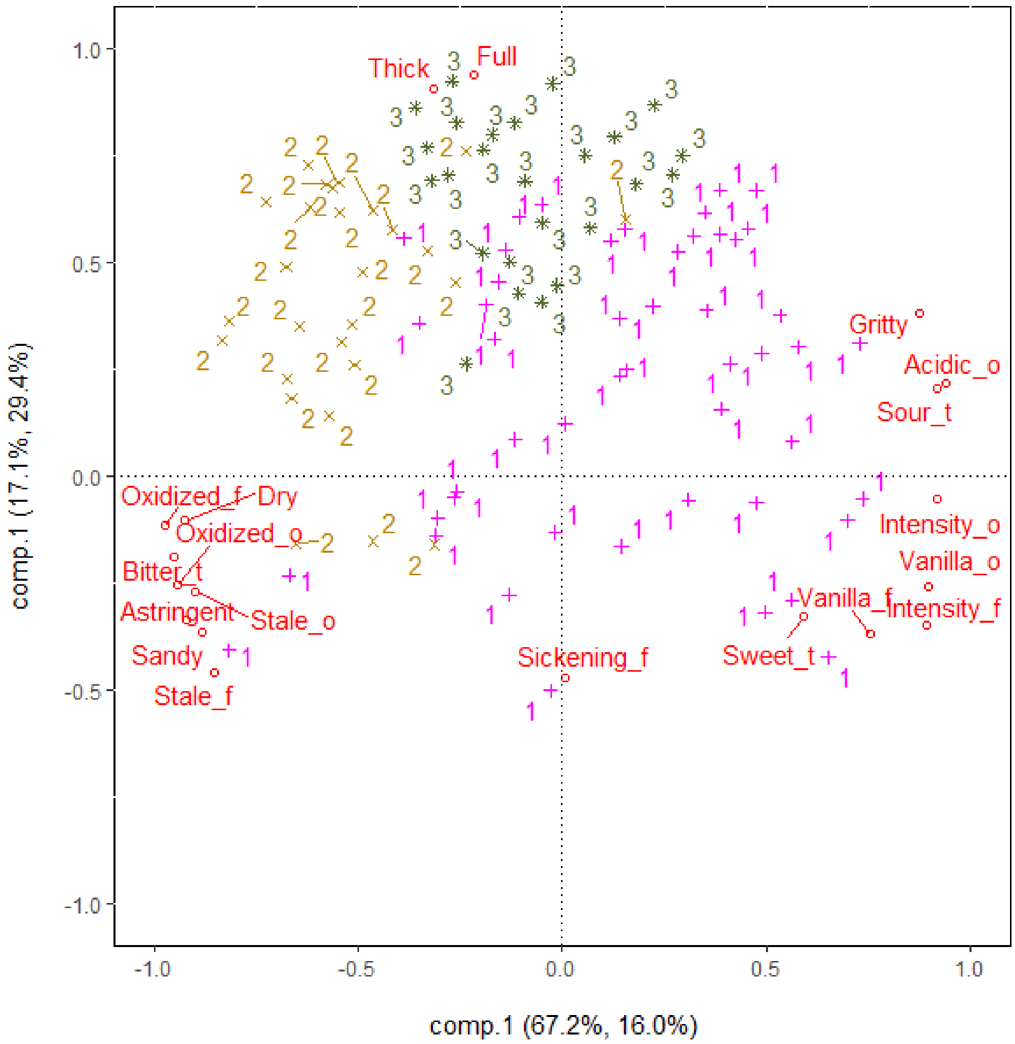

Figure 8. Correlation loading plot between Y data - consumer likings and X data - sensory properties. The correlation loading plot shows consumer preferences in relation to sensory properties (in red). Three consumer segments G1, G2.

Figure 9 highlights the map for consumer attributes linked to consumer segments (results taken from Figure 3 and Figure 4). We can see that consumers in segment G1 who preferred thin-yoghurts are characterized by consumer attributes related to taste-related factor *pleasure* (e.g., ple_1, ple_4) and health-related factor *natural product interest* (e.g., nat_4, nat_5). Consumers in segment G3 who preferred thick-flakes-yoghurts are described by consumer factors related to health-related attitudes *light product interest* (e.g., lig_1, lig_2, lig_3) and *general health interest* (e.g., gen_4). Consumers in segment G2 who preferred flour-yoghurts did not relate any specific consumer attributes as these consumers located in the opposite site of the consumer attributes (Figure 9). Although the relation is not so strong, the G2 consumers might associate with taste-related factors *craving for sweet foods* (e.g., cra_1, cra_2) and *using food as a reward* (e.g., rew_6). Furthermore, these consumers had low values of *general health interest* (e.g., gen_2, gen_3, gen_5), *natural product interest* (e.g., nat_1, nat_2, nat_5, nat_6), and *light product interest* (e.g., lig_5).

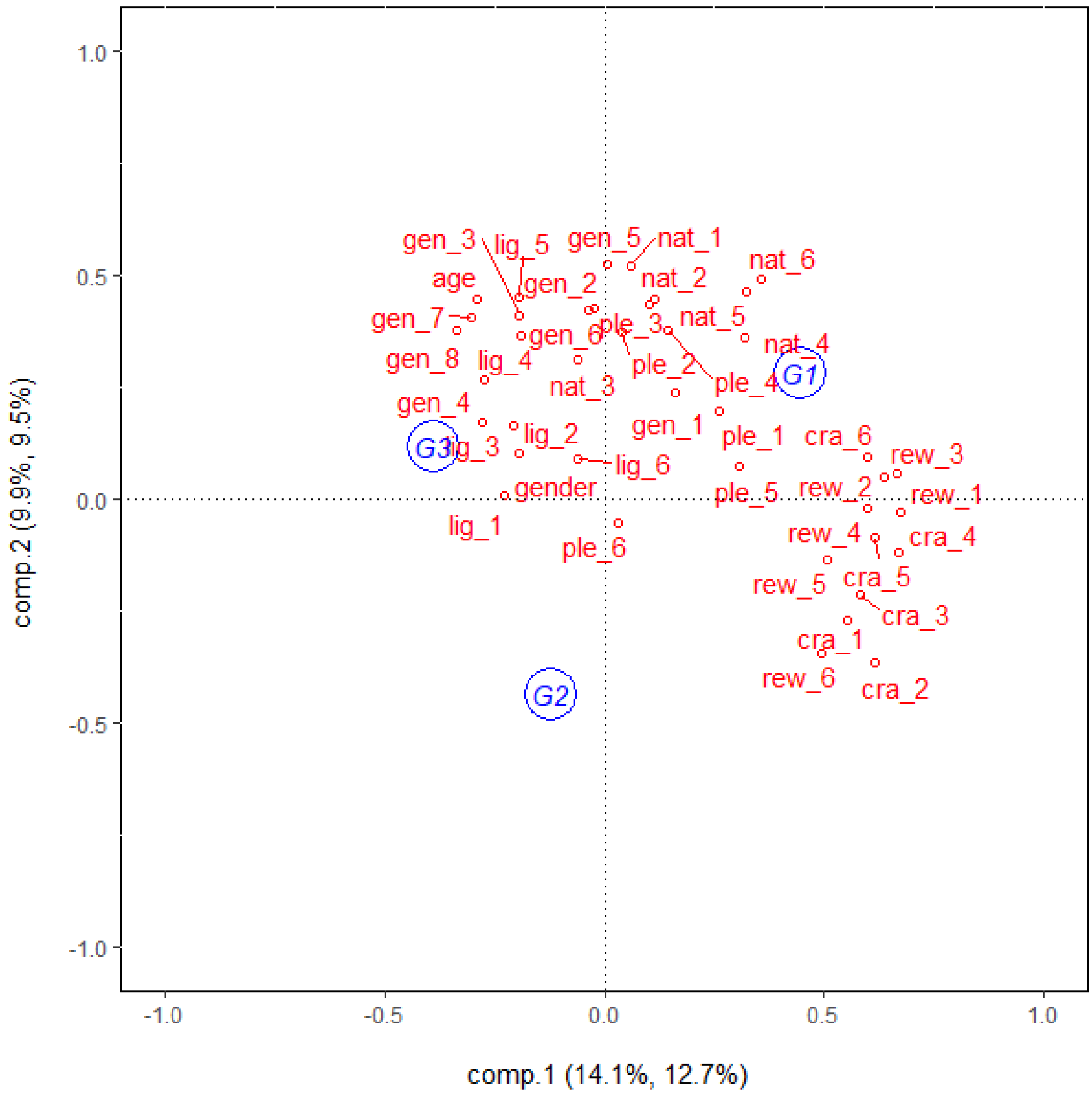

Figure 9. Correlation loading between Y data - dummy and Z data - consumer attributes.

The correlation loading plot shows consumer segments in relation to consumer attributes (in red). The three consumer segments are shown as G1, G2, G3.

4.5 Comparison of consumer segmentation in L-PLS regression and TSP approaches

By using the **Y**-average matrix, the L-PLS approach indicates how different consumer segments relate to sensory properties and consumer attributes (Figure 6). These results can also be observed in the TSP approach (Figures 6 and 7). The two approaches provide similar results in the relation between consumer segments and their interpretation based on their sensory properties (segment G1 – *Sweet,* and *Vanilla*; segment G2 – *Dry,* and *Sandy*; segment G3 – *Thick,* and *Full*). In addition, the two approaches highlight the same relation between consumer segments and consumer attributes: segment G1 – *attitudes to taste,* segment G3 – *attitudes to health,* and segment G2 – opposite side to both attitudes.

5. DISCUSSION & CONCLUSIONS

This manuscript investigates and compares the one-step L-PLS approach with a two-step PLS approach (TSP) for L-shape data when segments are the focus. For the L-PLS consumer degree of liking data are replaced by average liking for each segment separately. As a benchmark we use the established two-step procedure (TSP) based on dummy coding for the segments.

Segmentation

Using an automatic segmentation procedure can be problematic in many cases and there is no guarantee that the segments are identified according to a meaningful interpretation. It is also very seldom to find clearly separated segments in consumer science, indicating that segmentation will often have a strong subjective element in it (Endrizzi et al., 2014; Endrizzi et al., 2011) depending on which method (criterion and procedure) is used. In this manuscript, the segmentation and interpretation of segments is graphically oriented according to PCA score and loading plots of the individual differences. As the focus in this research is to investigate different consumer segments preferring different types of yoghurts, consumer segments are determined according to their preferences to *thin* yoghurts, *flour* yoghurts, and *thick-flake* yoghurts.

Procedures for incorporating segments in the analyses

Here we proposed to represent the consumers in the different segments by the average degree of liking values in the segments they belong to. This strategy worked very well for the L-PLS approach and gave reasonable interpretations, comparable to results obtained without segmentation. The 'average of consumers with segment' method can also be used for the TSP approach, but here we decided to use the more established TSP approach based on dummy response variables and discriminant PLS (i.e., PLS-DA).

Interpretation

Overall, both L-PLS and TSP approaches provide similar interpretation results. In the one-step L-LPS approach results are visible in a single plot, while the TSP approach needs plots for both steps 1 and 2. The TSP approach, however, has the advantage of interpretating the horizontal and vertical direction in the L separately using standard regression methods. In L-PLS, a double-centred matrix (of average likings of consumers) is applied which implies that it highlights differences between consumers in their relative position. In TSP, column-centred matrix (of original likings) is used that gives a more direct interpretation of the liking of the different products.

Possible extensions

It is worth noting that the comparison between L-PLS and TSP is qualitative as it is based on the interpretations. The main issue here is that, in the case of L-shape data, there are three different sources of information (i.e., consumer likings, sensory properties, and consumer attributes) and there are differences in presenting results of L-PLS (one figure) and TSP (two figures); therefore, it is not easy to establish a quantitative criterion for comparison. This issue should be addressed in the future studies.

In addition, when two of the main blocks, i.e., sensory properties and consumer attributes, consist of variables representing different aspects, the TSP approach can handle the relations between blocks in L-shape data by using multiblock regression such as Sequential and Orthogonalized - Partial Least Square (SO-PLS, Jørgensen & Næs, 2008 and Jørgensen, Segtnan, Thyholt, & Næs, 2004) in each step of the TSP approach. Future research should make some comparation to identify pros and cons of these approaches.

Conclusions

In conclusion, this manuscript has been devoted to two different ways of handling segmentation in L-shape data: average likings of consumers in each segment in L-PLS, original likings and dummy variables in TSP approach. Prior to applying either L-PLS or TSP approach, the segmentation can be done based on visual interpretations of the PCA results. Both the L-PLS and TSP approaches highlight the relation between the consumer segments to sensory properties, and consumer attributes. Although the methods have different advantages, when considering the overall interpretation, results are comparable. Therefore, it is not possible to give a strict recommendation based on this manuscript. The interpretations are also comparable to what is obtained without clustering, but the segmentation approach may be slightly preferred since it focuses more on overall pattern than on all possible individual consumers.

ACKNOWLEDGEMENTS

Author Quoc Cuong Nguyen is funded by Vietnam National University Ho Chi Minh City under grant number C2022-20-22. Author Quoc Cuong Nguyen acknowledges Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study. Authors P. Varela & T. Naes acknowledge the financial support received from the Research Council of Norway and the Norwegian Fund for Research Fees for Agricultural Products (FFL) through the project "FoodForFuture" (Project number 314318; 2021-2024). Special thanks go to Hilde Kraggerud (Tine, Norway) for the support with the sample materials and to Stefan Sahlstrøm (Nofima) for his help with the milling procedure.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHORSHIP CONTRIBUTIONS

Quoc Cuong Nguyen: Methodology, Formal analysis, Software, Validation, Writing - original draft. **Daniele Asioli:** Writing - original draft. **Paula Varela:** Funding acquisition, Project administration, Writing - review & editing. **Tormod Næs**: Conceptualization, Methodology, Supervision, Writing - review & editing.

DATA AVAILABILITY STATEMENT

Research data are not shared.

REFERENCES

Almli, V. L., Næs, T., Enderli, G., Sulmont-Rossé, C., Issanchou, S., & Hersleth, M. (2011). Consumers' acceptance of innovations in traditional cheese. A comparative study in France and Norway. *Appetite, 57*(1), 110-120. doi:https://doi.org/10.1016/j.appet.2011.04.009

Asioli, D., Næs, T., Granli, B. S., & Lengard Almli, V. (2014). Consumer preferences for iced coffee determined by conjoint analysis: an exploratory study with Norwegian consumers. *International Journal of Food Science & Technology, 49*(6), 1565-1571. doi:10.1111/ijfs.12485

Asioli, D., Nguyen, Q. C., Varela, P., & Næs, T. (2022). Comparison of different ways of handling L-shaped data for integrating sensory and consumer information. *Food Quality and Preference, 96*, 104426. doi:https://doi.org/10.1016/j.foodqual.2021.104426

Castura, J. C., Meyners, M., Varela, P., & Næs, T. (2022). Clustering consumers based on product discrimination in check-all-that-apply (CATA) data. *Food Quality and Preference, 99*, 104564. doi:https://doi.org/10.1016/j.foodqual.2022.104564

Chang, W.-C. (1983). On Using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 32*(3), 267-275. doi:10.2307/2347949

Endrizzi, I., Gasperi, F., Calò, D. G., & Vigneau, E. (2010). Two-step procedure for classifying consumers in a L-structured data context. *Food Quality and Preference, 21*(3), 270-277. doi:https://doi.org/10.1016/j.foodqual.2009.06.004

Endrizzi, I., Gasperi, F., Rødbotten, M., & Næs, T. (2014). Interpretation, validation and segmentation of preference mapping models. *Food Quality and Preference, 32*, 198-209. doi:https://doi.org/10.1016/j.foodqual.2013.10.002

Endrizzi, I., Menichelli, E., Johansen, S. B., Olsen, N. V., & Næs, T. (2011). Handling of individual differences in rating-based conjoint analysis. *Food Quality and Preference, 22*(3), 241-254. doi:http://dx.doi.org/10.1016/j.foodqual.2010.10.005

Frandsen, L. W., Dijksterhuis, G. B., Martens, H., & Martens, M. (2007). Consumer evaluation of milk authenticity explained both by consumer background characteristics and by product sensory descriptors. *Journal of Sensory Studies, 22*(6), 623-638. doi:10.1111/j.1745-459X.2007.00114.x

Green, P. E., & Krieger, A. M. (1995, 1995/07//). Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society, 37,* 221.

Hersleth, M., Lengard, V., Verbeke, W., Guerrero, L., & Næs, T. (2011). Consumers' acceptance of innovations in dry-cured ham: Impact of reduced salt content, prolonged aging time and new origin. *Food Quality and Preference, 22*(1), 31-41. doi:https://doi.org/10.1016/j.foodqual.2010.07.002

Jørgensen, K., & Næs, T. (2008). The use of LS–PLS for improved understanding, monitoring and prediction of cheese processing. *Chemometrics and Intelligent Laboratory Systems, 93*(1), 11-19. doi:http://dx.doi.org/10.1016/j.chemolab.2008.03.001

Jørgensen, K., Segtnan, V., Thyholt, K., & Næs, T. (2004). A comparison of methods for analysing regression models with both spectral and designed variables. *Journal of Chemometrics, 18*(10), 451-464. doi:10.1002/cem.890

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices*: Springer New York.

Lengard, V., & Kermit, M. (2006). 3-Way and 3-block PLS regressions in consumer preference analysis. *Food Quality and Preference, 17*(3–4), 234-242. doi:http://dx.doi.org/10.1016/j.foodqual.2005.05.005

Liland, K. H. (2022). multiblock: Multiblock Data Fusion in Statistics and Machine Learning. *R package version 0.8.3*.

Martens, H. (2005). *Domino PLS: a framework for multi-directional Path Modelling*. Paper presented at the Proceedings of PLS'05 International Symposium, 2005, Paris.

Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L. H., Høy, M., Westad, F., . . . Martens, M. (2005). Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. *Computational Statistics & Data Analysis, 48*(1), 103-123. doi:http://dx.doi.org/10.1016/j.csda.2003.10.004

McEwan, J. A. (1996). Preference mapping for product optimization. In *Multivariate analysis of data in sensory science* (pp. 71-102): Elsevier.

Mejlholm, O., & Martens, M. (2006). Beer identity in Denmark. *Food Quality and Preference, 17*(1), 108-115. doi:http://dx.doi.org/10.1016/j.foodqual.2005.10.001

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). Preference Mapping for Understanding Relations between Sensory Product Attributes and Consumer Acceptance. In *Statistics for Sensory and Consumer Science* (pp. 127-153): John Wiley & Sons, Ltd.

Næs, T., Varela, P., & Berget, I. (2018). Chapter 7 - Individual Differences in Consumer Liking Data (Rating Based). In T. Næs, P. Varela, & I. Berget (Eds.), *Individual Differences in Sensory and Consumer Science* (pp. 109-169): Woodhead Publishing.

Nguyen, Q. C., Næs, T., Almøy, T., & Varela, P. (2020). Portion size selection as related to product and consumer characteristics studied by PLS path modelling. *Food Quality and Preference, 79*, 103613. doi:https://doi.org/10.1016/j.foodqual.2018.11.020

Nguyen, Q. C., Næs, T., & Varela, P. (2018). When the choice of the temporal method does make a difference: TCATA, TDS and TDS by modality for characterizing semi-solid foods. *Food Quality and Preference, 66*, 95-106. doi:https://doi.org/10.1016/j.foodqual.2018.01.002

Pohjanheimo, T., & Sandell, M. (2009). Explaining the liking for drinking yoghurt: The role of sensory quality, food choice motives, health concern and product information. *International Dairy Journal, 19*(8), 459-466. doi:https://doi.org/10.1016/j.idairyj.2009.03.004

R Core Team. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rødbotten, M., Martinsen, B. K., Borge, G. I., Mortvedt, H. S., Knutsen, S. H., Lea, P., & Næs, T. (2009). A cross-cultural study of preference for apple juice with different sugar and acid contents. *Food Quality and Preference, 20*(3), 277-284. doi:https://doi.org/10.1016/j.foodqual.2008.11.002

Roininen, K., Lahteenmaki, L., & Tuorila, H. (1999). Quantification of Consumer Attitudes to Health and Hedonic Characteristics of Foods. *Appetite, 33*(1), 71-88. doi:http://dx.doi.org/10.1006/appe.1999.0232

Sæbø, S., Martens, M., & Martens, H. (2010). Three-Block Data Modeling by Endo- and Exo-LPLS Regression. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H.

Wang (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications* (pp. 359-379). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schutz, H. G., & Cardello, A. V. (2001). A labeled affective magnitude (LAM) scale for assessing food liking/disliking. *Journal of Sensory Studies, 16*(2), 117-159. doi:10.1111/j.1745-459X.2001.tb00293.x

Smilde, A. K., Næs, T., & Liland, K. H. (2022). *Multiblock Data Fusion in Statistics and Machine Learning: Applications in the Natural and Life Sciences*: Wiley.

Stone, H., Bleibaum, R., & Thomas, H. A. (2012). *Sensory Evaluation Practices*: Academic.

Vigneau, E., Endrizzi, I., & Qannari, E. M. (2011). Finding and explaining clusters of consumers using the CLV approach. *Food Quality and Preference, 22*(8), 705-713. doi:https://doi.org/10.1016/j.foodqual.2011.01.004

Vinzi, V. E., Guinot, C., & Squillacciotti, S. (2007). Two-step PLS regression for L-structured data: an application in the cosmetic industry. *Statistical Methods and Applications, 16*(2), 263-278. doi:10.1007/s10260-006-0028-2

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *J Am Stat Assoc, 105*(490), 713-726. doi:10.1198/jasa.2010.tm09415

Yenket, R., & Chambers IV, E. (2017). Influence of cluster analysis procedures on variation explained and consumer orientation in internal and external preference maps. *Journal of Sensory Studies, 32*(5), e12296. doi:https://doi.org/10.1111/joss.12296

Yenket, R., Chambers IV, E., & Johnson, D. E. (2011). Statistical package clustering may not be best for grouping consumers to understand their most liked products. *Journal of Sensory Studies, 26*(3), 209-225. doi:10.1111/j.1745-459X.2011.00337.x

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics, 17*(9), 763-774. doi:10.1093/bioinformatics/17.9.763

TABLES

Table 1. An illustration of matrix Y-average with products in rows, and consumers in columns. Consumers belonging to the same segment have the same liking values.

Table 2. An illustration of matrix Y-dummy with consumers in rows, and segments in columns. Consumers belonging to a cluster have the 1's, otherwise 0's.

Table 3. Formulation of yoghurts and the symbols used in plots.

Table 4. Results from ANOVA model.

FIGURES

Figure 1. L-shape data: sensory properties – X matrix (I products × K sensory properties), consumer liking ratings – Y (I products × J consumers), and consumer attributes – Z matrix (L consumer attributes × J consumers).

Figure 2. Overall liking of products.

Figure 3. Scores plot with segment numbers from the PCA of the double centered residuals.

Figure 4. Loadings plot with segment numbers from the PCA of the double centered residuals.

Figure 5. Overall consumer degree of likings of yoghurts in segments G1, G2, G3.

Figure 6. Endo-L-PLS. Sensory properties (X): centered and standardized for each property; Consumer degree of liking (Y): double-centered; Consumer attributes (Z): centered and standardized for each attribute. Endo-PLS plot shows consumer preferences for the eight yoghurts (in blue) in relation to both sensory properties (in green) and consumer attributes (in red); three consumer segments are shown as G1, G2, G3.

Figure 7. Score plot between Y data - consumer likings and X data - sensory properties.

Figure 8. Correlation loading plot between Y data - consumer likings and X data - sensory properties. The correlation loading plot shows consumer preferences in relation to sensory properties (in red). Three consumer segments G1, G2.

Figure 9. Correlation loading between Y data - dummy and Z data - consumer attributes.