3-way and 3-block PLS Regression in Consumer Preference Analysis

Valérie Lengard and Martin Kermit *

CAMO Process AS, The Unscrambler®, Nedre Vollgate 8, 0158 Oslo, Norway,
www.camo.com

Abstract

PLS regression modelling using 3-way or 3-block data is sometimes misleadingly regarded as a single method. This paper illuminates the structural differences of 3-way data compared to 3-block data applied to sensory analysis. The data used to illustrate these two powerful methods is a set of 17 tomato varieties with information acquired from chemical measurements, sensory panel evaluations and consumer likings, attitudes and demographics. A 3-way PLS regression model is used to identify positive and negative drivers of consumer likings and simultaneously evaluate the performance of individual sensory panelists. From the 3-block PLS regression model, also called L-PLS model, the tomato varieties are described using chemical information and sensory attributes. Also extracted from the L-PLS model is background information on the consumers who expressed their preferences for the different tomato types.

Key words: Consumer Preference, 3-way PLS, L-PLS, Sensory drivers of liking

1 Introduction

Sensory analysis usually faces the problem of relating data acquired from a number of sources describing a product. Identification of relationship between sources such as product information, sensory evaluation and consumer preference and attitude is key to a successful analysis. Partial Least Squares regression (PLS) is a well-known and useful tool in consumer preference analysis. Building PLS models linking two sets of descriptors for a number of samples

* Corresponding author
Email addresses: vl@camo.no, mk@camo.no (Valérie Lengard and Martin Kermit).

is relatively simple using modern multivariate software tools. However, when the number of sets increases, the complexity of the analysis grows, and model building requires extensions to classical PLS regression.

This article describes the use of two PLS extensions, 3-way PLS and L-PLS applied to build relationships within a data set where multiple sources of variations are available. The intention is not to outline the computational details of these methods. Instead, focus will be given to practical use and interpretation of 3-way PLS and L-PLS models. The overall objective of the study is to illustrate these two PLS extensions to identify segments of consumers with differing product likings and explain the differences in terms of characteristics of the consumers and descriptors of the product. A brief description of the theory behind these PLS methods can be found in Section 2. The data set chosen for the analysis is the workshop data set used in the 7th Sensometrics Meeting 2004 and is described in Section 3. The results from the 3-way analysis is presented in Section 4, while Section 5 gives the results from L-PLS. Finally, Section 6 concludes this paper.

2 Theory

Within chemometrics, two-block or two-way PLS is the standard multivariate tool for obtaining multivariate models relating an independent set of variables $\mathbf{X}$ to a dependent data set $\mathbf{Y}$. PLS decomposes the independent variables in a set of scores and the dependent variables is then regressed on these scores. Details on theory and algorithms can be found in most textbooks covering multivariate statistics, like Martens and Næs (1989); Næs *et al.* (2002). While PLS is limited to handling only two sets of data at a time, extensions using either multiple data sets or data with multiple variation modes exist. This section gives a brief description of two such PLS extensions and the application of these methods to sensory analysis.

2.1 3-way PLS regression

Multi-way PLS regression is similar to standard PLS regression by building a model incorporating a relationship between a set of independent variables $\mathbf{X}$ and a dependent variable set $\mathbf{Y}$. The extension provided in the multi-way version of PLS regression is that independent data is not limited to having only one mode of variation. The general N-way PLS was introduced by Bro (1996). An illustration of the structure of a data set with two modes of variations (3-way) is shown in Fig. 1.

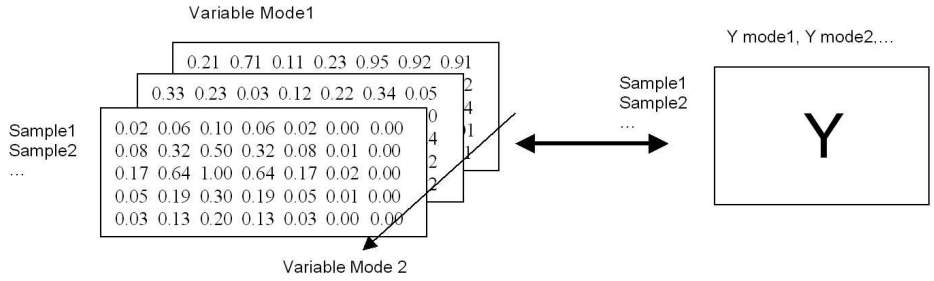

Fig. 1. A 3-way PLS regression model. For a set of samples, two variation modes are available in the independent data set **X**, thus forming a cube structure. The dependent data set **Y** is still a regular 2-D table.

A common approach to 3-way PLS models in sensory analysis is to use panelist scores and product attributes as the two variation modes for a set of samples or products. This 3-way data structure can then be related to a dependent data set consisting of either consumer evaluations or process parameters for the same set of products or samples. Several examples of this type of analysis have been described by Bro (1998).

2.2 L-PLS regression

Where classical PLS relates two blocks of data, L-PLS extends the relationship to three blocks forming an *L*-shape. The three blocks of data in L-PLS relate dependent variables **Y** $(N \times J)$ to matrices **X** $(N \times K)$ and **Z** $(J \times L)$. The model requires **Y** to be related to structures in both **X** and **Z**. The procedure can bring out structures in **Y** that can be seen both in **X** and **Y**. L-PLS is a relatively new method introduced by Martens *et al.* (2004).

In consumer studies on actual products, L-PLS can be used to combine marketing science, sensory science and product science. The 3-block structure in L-PLS regression thus integrates both sensory and physical product properties **X**, consumer liking data **Y** and the consumer background data **Z** (Martens *et al.*, 2003). An illustration of such a structure is given in Fig. 2.

2.3 Data reduction and exploratory analysis with PCA

In some cases, when the data set contains a large number of dependent variables, it may prove useful to reduce the data set into smaller segments to provide a clearer and more interpretable result. Principal component analysis (PCA) is an ideal tool for such tasks, in that a data set can be described by principal components according to the degree of variance within the data. Components with little explained variance can be left out of the analysis, thus

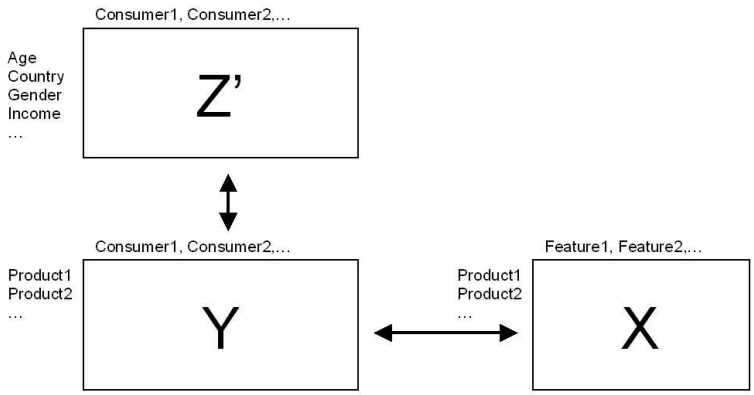

Fig. 2. An L-PLS model. The three data arrays $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}'$ form an L-shaped structure. $\mathbf{Y}$ has common rows with $\mathbf{X}$ and common columns with $\mathbf{Z}'$. $\mathbf{X}$ and $\mathbf{Z}'$ are only inter-related via $\mathbf{Y}$.

reducing the amount of data and still keeping most of the information. PCA may also be used in exploratory analysis, where plots of the principal component loadings can be used to identify variables that are similar to each other. The reader not familiar with PCA should consult Joliffe (2002).

3 Description of data and software

A data set from 2001 provided by the Centre Technique des Fruits et Légumes (CTIFL) and the Institut National de la Recherche Agronomique (INRA) in France sponsored by the Ministère de la Recherche and organized by the Association des Centres Technique Agricoles (ACTA) constitutes the basis for the analysis. The data consists of information gathered from 17 tomato varieties. For every variety, descriptive evaluation of 11 sensory attributes from 14 trained panelists has been provided with one repetition on a 0-10 scale. Also available is a collection of 15 measures with two repetitions describing physical and chemical parameters of the 17 tomato varieties. The 15 measures were recorded either on entire tomato (6 measures), tomato meat (4 measures) or tomato juice and seeds (5 measures). In addition to the panel data and the physical and chemical measurements, a collection of 379 consumer evaluations exists including:

- Liking ratings of the tomatoes served one at a time on a 0-10 scale.
- 17 questions regarding usage and attitude.
- Preference test were the consumer must state a preference between two tomatoes served as a pair.
- An appearance liking test were the consumers rank their preference from a subset of the tomato varieties.

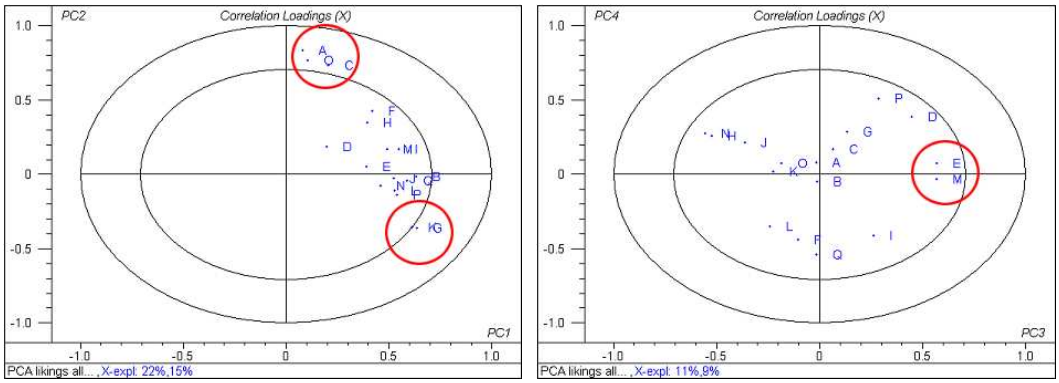

Fig. 3. Segment selection in PCA. *Left*: First and second principal component loadings of consumer likings indicating two clusters of tomatoes. *Right*: Third and fourth principal component loadings forming a third cluster of tomatoes.

In the study reported in this paper, data from the preference test and appearance test have not been used.

For model building, analysis and calculations, The Unscrambler 9.1 (2004) multivariate software tool was used. PCA and 3-way PLS regression are built-in functionalities in the program, whereas L-PLS was performed using the following three steps: Step one was the calculation of the correlation matrix of $\mathbf{Z}$ and $\mathbf{Y}'$. Step two was the organization of a data table containing $\mathbf{X}$, the correlation matrix of $\mathbf{Z}$ and $\mathbf{Y}'$, and $\mathbf{Y}$. In order for the final model to include the tomatoes on the map of loadings, dummy variables were created for the 17 tomatoes and introduced in the table. In step three a PLS regression model was built utilizing the $\mathbf{X}$ matrix and the passified (down-weighted) dummy variables as independent variables. $\mathbf{Y}$ and the correlations of $\mathbf{Z}$ and $\mathbf{Y}'$ were used as dependent variables. Leave-one-out cross-validation was performed and uncertainty testing was used to identify significant predictors in the model.

4 Results from 3-way analysis

4.1 PCA segment selection and definition

One objective was to find segments of consumers with different preferences for the 17 tomato varieties in the data set. To identify some basic structure in the data, an exploratory analysis using principal component analysis (PCA) on the consumer likings was performed.

On the second principal component loadings, a group consisting of tomatoes A, C and O was clearly separated from the rest. Varieties K and G indicated

a clustering behavior uncorrelated to the first group. These two groups are indicated in the PCA loadings plot to the left of Fig. 3. A third group appearing in the higher principal components consisted of tomato varieties labeled E and M, shown in the PCA loading plot to the right of Fig. 3. These three groups of tomato types were used as a basis to define three consumer segments with corresponding likings. The overall consumer liking has been provided on a scale of 0-10, and a threshold was used to identify consumers with a preference for each of the groups. For the ACO-group, a segment was formed from consumers with a liking above 7 for at least one of the tomato varieties A, C or O and no liking below 5 for all three. The two other groups, KG and EM were created in a similar fashion but with only one threshold set to 5. Since tomato variety A was liked by all groups, the segments were renamed to ACO, GAK and EMA where the names indicate the liking for each group in the preferred order. It should be emphasized that the three segments had some overlapping members in that 27% of the consumers belonged to more than one group and 31% did not fall into any of the three groups.

Taking into account background knowledge on the tomatoes, we can proceed to an external validation of these results. It is observed that the three identified segments very much correspond to the shapes of the tomatoes: Tomato types A and O are cocktail tomatoes, tomato types E and M are ovoid tomatoes, while tomatoes G and K are regular-shaped tomatoes.

4.2 3-way data structure

To extract positive and negative drivers of liking for the defined consumer segments, a 3-way PLS regression model was constructed. In the model, the array of dependent variables $\mathbf{X}$ is a 3-way array where the 17 tomato varieties constitute the objects mode, the sensory attributes constitute the first variable mode and the panelists constitute the second variable mode according to Fig. 1. The average likings for each defined segment were used as dependent variables $\mathbf{Y}$. The $\mathbf{X}$-data of sensory properties is presented in a cube (17 tomatoes $\times$ 11 attributes $\times$ 14 panelists). The $\mathbf{Y}$-data of preference evaluations is presented in a 2-D table (17 tomatoes $\times$ 3 consumer segments).

4.3 Extraction of positive and negative drivers

On the first variable mode (sensory attributes) the model indicates that tomato flavor is the only positive driver common to all three segments, while mealiness is the only shared negative driver. The EMA and GAK segments differentiate most from each other and three negative drivers of the EMA segment (firm, firm inside and acidity) are positive drivers of the GAK segment (Fig. 4). It

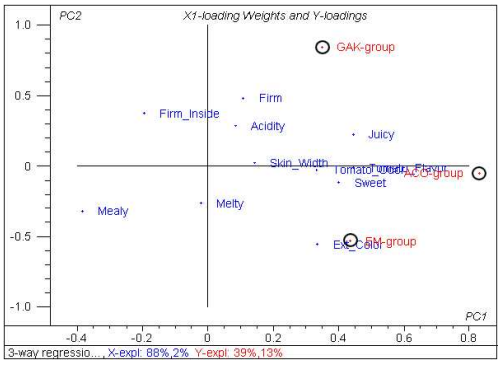

Fig. 4. 3-way PLS result: Loading weights from sensory attributes and loadings of 3 consumer segments indicating the drivers of liking for the different segments.

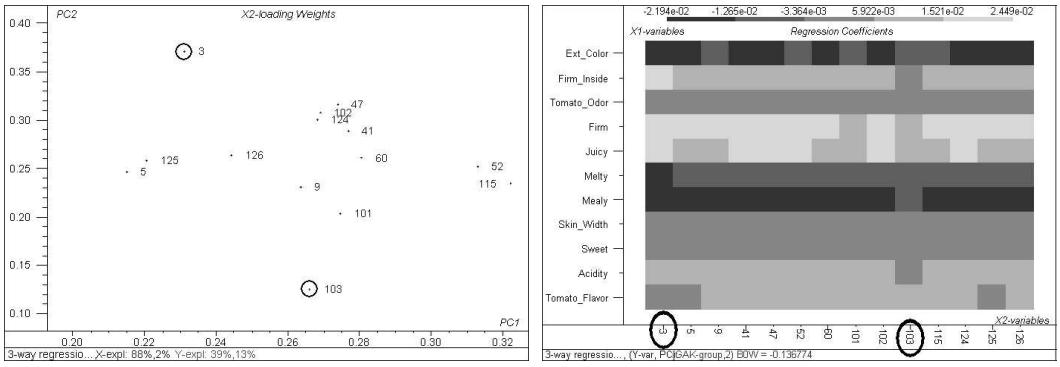

Fig. 5. 3-way PLS result. *Left:* Loading weights from the second variable mode, panelists, indicating two outliers. *Right:* Regression coefficient map showing drivers of liking for the GAK-group. The two outlying panelists in disagreement with the rest of the panel are indicated.

is to be noted that sensory attribute skin width is the only attribute with neither a positive nor a negative effect on the consumer preferences. Detailed findings are presented in Table 1.

On the second variable mode, the model also indicates a strong agreement between sensory panelists, with panelists 3 and 103 showing most deviation from the rest of the panel as shown to the left in Fig. 5. The deviation was not found important enough to discard these assessors' evaluations. However, in the perspective of conducting appropriate further training of the panelists it is of interest to identify where their individual weaknesses lie. Fig. 5 shows to the right a map of regression coefficients for the modeling of likings of the GAK consumer group. The assessment from panelist 3 of attribute firm inside shows this property as a strong positive driver of liking (pale grey on the plot). A similar assessment from panelist 103 does not indicate firm inside as a relevant driver of liking (intermediate grey). The rest of the panel's assessments show firm inside as a positive driver of liking, although not to the same extent as panelist 3 (light grey). Panelist 103 also slightly disagrees with the rest of the panel for attributes mealy and acidity.

The model's main difference between panelist 3 and panelist 103 is that while assessments from panelist 3 indicate most attributes to be strongly positive (pale grey) or strongly negative (black) drivers, assessments from panelist 103 do not indicate any strong drivers. Conclusively, assessor 103 shows signs of magnitude error in his evaluations.

5 Results from L-PLS

5.1 Data reduction with PCA and PLS

Due to the many sources of information gathered into a single model, L-PLS results can prove to be overwhelming with information. It is therefore of good advice to simplify the data by applying data reduction techniques before launching into the L-PLS model calibration.

A first data treatment was performed on the sensory data, where the assessments were averaged over panelists. As the prior 3-way PLS model did not indicate any serious defects in sensory evaluations, all assessors were included in the average. Further, the physical and chemical measurements were averaged over repetitions. A PCA calibration performed on these parameters showed very strong correlations between similar chemical measurements performed on the three locations in the tomato: the entire tomato, the tomato meat and the tomato juice and seeds. Therefore, only five physical and chemical properties measured on the entire tomato were selected for further modeling. Finally, a PLS regression was run on the answers from the consumer attitudinal questionnaire (independent variables) vs. consumer likings (dependent variables). Uncertainty testing was used in the software and **X**-variables detected as significant at a 95% level were kept for further analysis. The selected variables include information on the consumer's origin (Lyon, Dijon or Nantes), age (four age groups), gender, reasons for satisfaction and dissatisfaction, tomato usage, consumed tomato types and preferred tomato types.

5.2 L-PLS data structure

The second phase of the study aims at relating three blocks of data. An L-PLS model was constructed to include the following three sources of information into a single, overview model:

(1) The sensory profile, physical and chemical information for each tomato, contained in a matrix **X** (17 tomatoes $\times$ 16 sensory and chemistry),

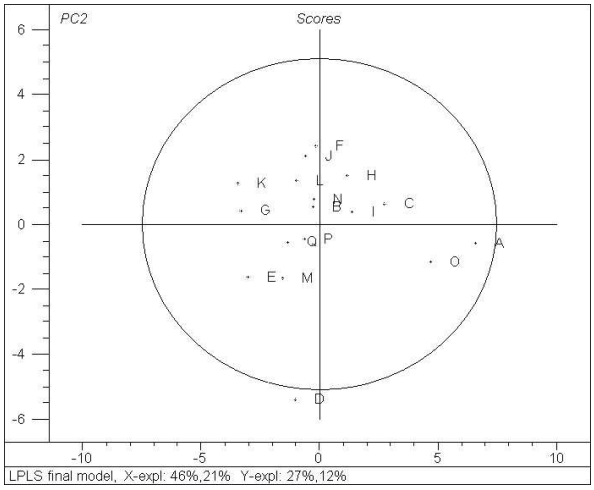

Fig. 6. 95% confidence ellipse showing tomato variety D at the lower end of the second principal component as a strong outlier.

(2) The consumer likings per tomato, contained in a matrix **Y** (17 tomatoes × 379 consumers) and
(3) The attitudinal characteristics of the consumers, contained in a matrix **Z** (379 consumers × 34 attitudes).

5.3 Relating preference ratings, consumer attitudinal information, sensory evaluations and chemical measurements

A first L-PLS model built on all 17 tomato types reveals tomato variety D to be a strong outlier. This is highlighted by display of a 95% confidence Hotelling T^2 ellipse on the score plot (Fig. 6). By studying the loadings plot we identify that tomato D is characterized by very high melty and mealy properties. Very few consumers are projected in the direction of tomato D which indicates that it is a much disliked tomato.

Results from a second L-PLS calibration without tomato D are presented hereafter. The result of most interest in the L-PLS model is the correlations loadings plot shown in Fig. 7 as it displays all available sources of information into one map. In Fig. 7 objects A to Q are the tomatoes, the black objects are the independent variables (sensory, physical and chemical information) and the grey objects are the dependent variables (consumer likings, represented by a simple dot, and consumer attitudinal background). The plot shows the first two PLS-components, which represent 69% of the variance in the independent variables and 39% of the variance in the dependent variables. The significance test indicates that most sensory and chemical predictors are significant in the model, validating the variable selection performed at data reduction stage in Section 5.1.

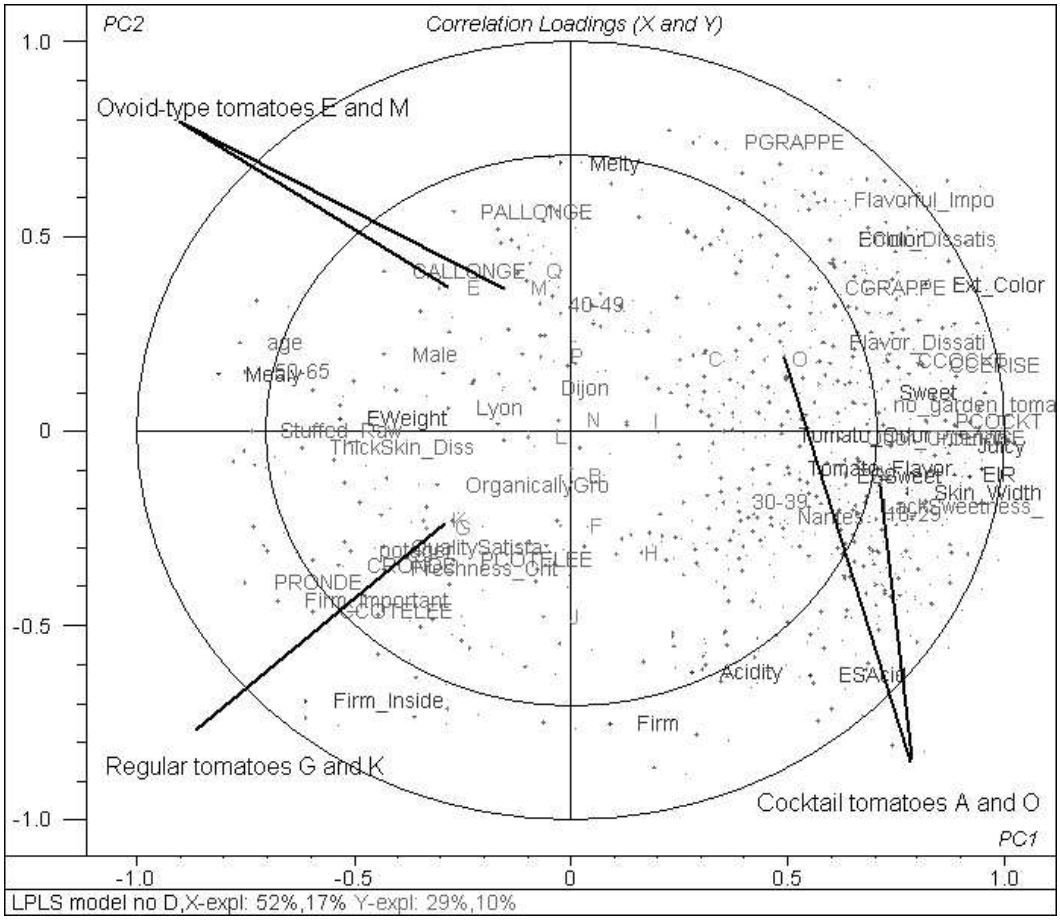

Fig. 7. Correlation loadings plot of the L-PLS model for the first two components. Objects A to Q are the 17 tomato varieties. Sensory attributes, chemical and physical parameters are in black. Consumers are represented by dots; consumer background information issued from the questionnaire is shown in grey. A high density of consumers lie in the direction of cocktail-tomatoes A and O, characterised by a strong tomato flavor, tomato odor and sweetness, and preferred by younger consumers. Ovoid-type tomatoes E and M are characterised by a low acidity; regular-sized tomatoes K and G present inside firmness. These four tomato varieties are preferred by middle-aged to elder consumers.

The best described products along component 1 are tomatoes A, O and C (Fig. 7). A large density of consumers is lying in the direction of cocktail-tomato A, thus indicating a preference for this specific tomato. These conclusions are in agreement with the segmentation results found in Section 4.1, which showed tomato A to be the overall most preferred tomato, and where consumer group AOC was the largest segment with 44% of the consumers. Tomato A is characterized by sensory properties sweet, tomato flavor, tomato odor, juicy, external color, skin width and low mealiness, all described along component 1. These results are in accordance with the drivers of likings detected in the 3-way PLS regression model (Section 4.3). Regarding chemical properties, tomato-type A is high in sugar content (ESSweet and EIR on

Size of Segment (%)	Segment 1: ACO (44%)	Segment 2: EMA (29%)	Segment 3: GAK (27%)
Tomato varieties liked most	A, C and O	E, M and A	G, A and K
Tomato varieties liked least	E, P and G	P, D and K	E, D and P
Positive drivers of liking	Tomato flavor and odor, sweet, juicy, external color	External color, tomato flavor and sweet	Firm, firm inside, juicy, acidity, tomato flavor
Negative drivers of liking	Mealy and firm inside	Mealy, acidity, firm and firm inside	External color, melty and mealy
Chemical and physical properties	High sugar content	Low acidity	Weak color
Key demographics and attitudinal parameters	Young adults, no access to garden tomatoes, flavor and lack of sweetness as potential dissatisfaction criteria, consume and prefer cherry and cocktail tomatoes	Middle-aged to elder consumers, often males	Middle-aged to elder consumers, indicate firmness as an important attribute in tomatoes

Table 1
Summary findings from the 3-way PLS regression and L-PLS models.

Fig. 7). Consumers with the highest preferences for tomato A belong to the younger age categories (18-29 and to some extent 30-39), live mainly in Nantes, consume and prefer cherry and cocktail-type tomatoes, do not have access to garden tomatoes, indicate flavor and lack of sweetness as potential dissatisfaction criteria for tomatoes, name tomato odor as a purchase criterion, and do not prepare stuffed raw tomatoes.

Component 2 on Fig. 7 describes in particular ovoid tomatoes E and M. These are projected in the direction of parameters CALLONGE (consumes ovoid tomatoes) and PALLONGE (prefers ovoid tomatoes), which were two questions issued from the questionnaire presented to the consumers. In addition, tomatoes E and M are characterized by a very low acidity both from a sensory point of view (Acidity) and on the chemical side (ESAcid). This is consistent

with the results from the 3-way PLS regression model, which indicated acidity as a negative driver of liking for consumer group EMA (Section 4.3). E and M tomato types are preferred by middle-aged to elder male consumers.

In the fourth quadrant of the correlation loadings plot, regular, rounded tomatoes K and G are displayed. These are characterized by inside firmness and low external color (detected both by sensory attribute Ext-Color and physical measurement EColor), which were two identified drivers of likings (respectively positive and negative) for the GAK consumer group. G and K are preferred by middle-aged to elder consumers who indicate firmness as an important criteria of satisfaction. All major findings from correlation loadings plot are summarized in Table 1.

6 Conclusion

3-way PLS regression and 3-block PLS regression often lead to confusion because of their nearly identical names. Through this study, the structural differences in the shape of 3-way data compared to 3-block data were exposed and illustrated. A 3-way PLS regression model was utilized to relate a 3-way sensory data array of two variable modes: panelists and attributes, to preference ratings of three pre-defined consumer segments. This analysis enabled us to kill two birds with one stone, namely identify positive and negative sensory drivers of liking as well as study the individual assessors' performances. Further, an L-PLS regression model was built on three data arrays where one of them (the $\mathbf{Y}$ data array) plays the role of a hinge between the other two (the $\mathbf{X}$ and $\mathbf{Z}$ data arrays). The L-PLS model is a powerful tool to incorporate and relate multiple sources of variation stored in data arrays of un-matching dimensions. From the presented L-PLS model, a large amount of information regarding the tomato varieties were linked together. The correlation loadings plot described the tomato varieties using chemical information and sensory attributes. Also extracted from the L-PLS model was background information on the consumers who expressed their preferences for the different tomato types.

Although model building of 3-way and 3-block data is not extensively more challenging than handling 2-way and 2-block data, these techniques are strongly inviting to perform data reduction so as not to be overwhelmed with information. Further, appropriate visualization software play a determining role in achieving a successful interpretation.

References

Rasmus Bro. Multi-way calibration. Multi-linear PLS. *Journal of Chemometrics*, 10:47–61,1996

Rasmus Bro. *Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications.* Ph.D. thesis, University of Amsterdam (NL), 1998

Ian Joliffe *Principal Component Analysis, 2nd ed.* Springer Series in Statistics, 2002

Harald Martens and Tormod Næs. *Multivariate calibration.* J. Wiley & Sons, Chichester, 1989.

Harald Martens, E. Andersen, L. H. Gidskehaugl, M. Høy and U. Indahl. *Some Extentions of the PLS Regression.* 3rd International Symposium on PLS and Related Methods (PLS'03), Lisbon, Portugal, September 15-17, 2003.

Harald Martens, Endre Anderssen, Arnar Flatberg, Lars Halvor Gidskehaug, Martin Høy, Frank Westad, Anette Thybo and Magni Martens. Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. *Computational Statistics & Data Analysis*, In Press, 2004

Tormod Næs, Tomas Isaksson, Tom Fearn and Tony Davies. *A user-friendly guide to Multivariate Calibration and Classification.* NIR Publications, 2002

The Unscrambler 9.1 Camo (USA, Norway and India) 30-day free trial installation available at www.camo.com