Assessing the use of different chemometric techniques to discriminate low-fat and full-fat yogurts

A.G. Cruz a,*, R.S. Cadena a, M.B.V.B. Alvaro a, A.S. Sant’Ana b, C.A.F. Oliveira c, J.A.F. Faria a, H.M.A. Bolini a, M.M.C. Ferreira d

a Department of Food Technology, Faculty of Food Engineering, University of Campinas (UNICAMP), Cidade Universitária “Zeferino Vaz”, CEP 13083-862 Campinas, SP, Brazil
b Department of Food and Experimental Nutrition, Faculty of Pharmaceutical Sciences, University of Sao Paulo (USP), Av. Prof. Lineu Prestes, 580, CEP 05508-900 Sao Paulo, SP, Brazil
c Department of Food Engineering, Faculty of Zootecnics and Food Engineering, University of Sao Paulo (USP), Pirassununga, SP, Brazil
d Institute of Chemistry, University of Campinas (UNICAMP), Cidade Universitária “Zeferino Vaz”, CEP 13083-970 Campinas, SP, Brazil

Abstract

This study proposes a new approach to discriminate low and full-fat yogurts using instrumental analysis and chemometric techniques. One hundred twenty six strawberry flavored yogurts were subjected to instrumental analysis of pH, color and firmness. Exploratory methods, such as Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA), and supervised classification methods, such as K-nearest neighbors (KNN), soft independent modeling of class analogy (SIMCA), and Partial Least Square Discriminant Analysis (PLSDA) were used for assessing the data. The results showed that low- and full-fat yogurts presented different with regard to all the variables analyzed. It was not possible to obtain total separation between the samples using PCA and HCA. KNN and PLSDA presented excellent performance toward the full-fat category, with 100% correct prediction which suggests only low-fat yogurts to be subjected to the traditional fat content determination methods. This approach can be incentivized by the health agencies aimed to optimize materials and financial resources.

1. Introduction

Chemometrics is a science of multidisciplinary nature which involves multivariate statistics, mathematical modeling and information technology, specifically applied to chemical data. Actually, these methods are useful tool in the quality control of dairy products (Gaspardo, Lavrenvic, Levert, Del Zotto, & Stefanon, 2010; Hammami et al., 2010; Karoui, Mounem, Rouissi, & Blecker, 2011; Lerma-Garcia, Gori, Cerretani, Simò-Alfonso, & Caboni, 2010; Ochi et al., 2010; Sacco et al., 2009; Sola-Larrañaga & Navarro-Blasco, 2009; Souza et al., 2011).

Quality control of low/reduced fat food products has become a common procedure of increasing importance in food industry, since the information displayed on the label often does not correspond the real values present in the food product composition. This is emphasized in several Brazilian surveys (Esper, Bonets, & Kuaye, 2007; Silva, Batista, Cruz, Moura, & Carvalho, 2008), which introduces unreliability to both health authorities and consumers. Despite official methods for determination of fat level in dairy products – Bocok, Gerber and Mojonier – are simple, they require preparation of samples and chemical reagents. In addition, they need glassware and specific pieces of equipment and instruments, such as butyrometers and centrifuges, all of which require minimally trained personnel. Finally, these methods generate chemical waste which requires adequate disposal (Brasil, 2006).

This study proposes a new approach to distinguish yogurts toward their fat content using instrumental analysis such as: pH, color and firmness, using chemometric methods (Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), K-nearest neighbors (KNN), Soft independent modeling of class analogy (SIMCA), and Partial Least Square Discriminant Analysis (PLSDA)).

2. Materials and methods

2.1. Sampling

One hundred and twenty six strawberry-flavored set yogurts (83 full-fat and 43 low-fat yogurts), from 12 commercial brands and different batches were used in this work. The composition of the products was obtained from their labels, with exception of the fat levels that were determined analytically. In the class of full fat...
yogurts, the protein, fat and carbohydrates contents were 1.8—3.5, 2.0—5.4 and 17—27, respectively. In all the yogurts collected, the stabilizers are gelatin and xanthan gum. In the class of low fat yogurts, the protein, fat and carbohydrates contents were 4.3—7.2, 0.9—2.2 and 6.9—7.4, respectively. Five samples from each commercial brand were purchased randomly at different supermarkets in the city of Campinas, São Paulo, Brazil and kept under refrigeration (3—5 °C). All the commercial brands analyzed correspond to 100% of the Brazilian yogurt market.

### 2.2. Analytical procedures

The fat in the yogurts were determined by the Blen-Dyer Method (Brasil, 2006). The pH, instrumental firmness and instrumental color ($L^*$, $a^*$, $b^*$) (Cielab, Hunterlab, Virginia, USA) of yogurts were performed using techniques commonly described for fermented milk products. Details about the methods are published elsewhere (Cruz, Walter, Cadena, Assis, Bolini & Fratini, 2009).

### 2.3. Chemometric techniques

The data analyses were performed using the software PIROUETTE 2.2.2 (Infometrix, Seattle, WA). The yogurt data set consisted of a 126 × 5 matrix, in which rows represented the yogurt samples, and columns the instrumental analysis values: pH, instrumental firmness and instrumental color parameters ($L^*$, $a^*$, $b^*$). Each sample was represented in the multidimensional space by a data vector, which was an assembly of the 5 features in yogurt samples. Data vectors belonging to the same category (full-fat, F and low-fat, L) were analyzed using chemometric procedures: PCA, HCA, KNN, SIMCA, PLSDA (Abdi, & Williamms, 2010; Alonso-Slaces and low-fat, L) were analyzed using chemometric procedures: PCA, HCA, KNN, SIMCA, PLSDA (Abdi, & Williamms, 2010; Alonso-Slaces & Lucey, 2009; Soukoullis, Panagiotidis, Krouelli, & Tzia, 2007). In the case of probiotic yogurts, there are additional parameters to be considered, such as the compounds used to supplement the milk (Lucas, Sodini, Monnet, Jolivet, & Corrieu, 2004), interaction between probiotic and starter cultures (Vinderola, Mocchiutti, & Reinheimer, 2002), the inoculation rate of the probiotic culture and its moment of addition during the yogurt processing (Kaur, Mishra, & Kumar, 2009), the quantity of water available during processing (Oliveira & Damín, 2003), oxygen level (Cruz et al., 2010; Cruz et al., 2012a,b) and the supplementation with a prebiotic ingredient (Oliveira, Perego, Oliveira & Converti, 2009, 2011; Debon et al., 2012).

### 3. Results and discussion

#### 3.1. Analytical results

Table 1 shows the average values obtained from the instrumental analysis of the yogurts. Significant differences were observed for all the parameters analyzed ($p < 0.05$). These findings can be due to different starter and probiotic cultures used by the manufacturers (each one with its own metabolic profiles) and the control of the operational parameters used in the yogurt processing, such as homogenization milk pressure, heat-related variables, type and amount of stabilizer used in the product formulation as well as final pH of fermentation.

The manufacture of yogurt is relatively simple, being produced in several small and medium-sized processing facilities. In many of them, the control of the inherent processing parameters is not performed. These parameters include the fat level obtained by skimming the milk, the amount of milk powder added to standardize the total solids, the heat treatment of the milk, the inoculum level of starter culture and the stabilizer used (Mortazavian, Ehsani, Mousavi, Sohrabandi, & Reinheimer, 2006; Mortazavian, Khosrokharvar, Rastegar, & Mortazaei, 2010; Peng, Horne, & Lucey, 2009; Soukoullis, Panagiotidis, Krouelli, & Tzia, 2007). In the case of probiotic yogurts, there are additional parameters to be considered, such as the compounds used to supplement the milk (Lucas, Sodini, Monnet, Jolivet, & Corrieu, 2004), interaction between probiotic and starter cultures (Vinderola, Mocchiutti, & Reinheimer, 2002), the inoculation rate of the probiotic culture and its moment of addition during the yogurt processing (Kaur, Mishra, & Kumar, 2009), the quantity of water available during processing (Oliveira & Damín, 2003), oxygen level (Cruz et al., 2010; Cruz et al., 2012a,b) and the supplementation with a prebiotic ingredient (Oliveira, Perego, Oliveira & Converti, 2009, 2011; Debon et al., 2012).

#### 3.2. Chemometric techniques

**3.2.1. Cluster analysis and principal component analysis**

The Euclidean distance and incremental linkage methods were used in the pre-processing to elaborate the HCA. The presence of three distinct segments is observed as shown in the dendogram (Fig. 1). The upper segment (Cluster 1), correspond exclusively to low-fat yogurts, comprising 32 samples, while in the lower segment are found 24 samples corresponding to full-fat yogurts. Finally, in the mid-section of the figure it is noted 70 samples of both low-fat and full-fat yogurt (59 samples of full-fat yogurt, which corresponds to 84.3% of the total, and 11 samples of low-fat yogurt, corresponding to the remaining 15.7%). These results suggest that yogurt manufacturers still need to make significant investments in research and development to produce low-fat yogurts that are similar to their full fat counterparts.

Five principal components (PC) were necessary to explain the variation of the data with principal components 1 and 2 explaining...
approximately 83% of the differences between the samples. PC1 explained 50.2% of the differences and was responsible for separating the samples of low from full-fat samples (Fig. 2). The greater intensity of the red color \((a')\) characterized the full-fat samples, while the low-fat yogurts were characterized by lower firmness and pH values. These results suggest that different amounts of strawberry pulp are being added to low-fat yogurts, whereas in the case of the full-fat yogurts there are significant differences between the amounts of stabilizers used and in the conditions of the fermentation process during the manufacturing of yogurts.

### 3.2.2. Supervised chemometric techniques

KNN is based on the determination of the distances between an unknown object and each of the objects of the training set. In KNN, the k-nearest objects of the unknown sample are selected and a majority rule is applied: the unknown is classified in the group to which the majority of the \(k\) objects belong. The choice of \(k\) is optimized by calculating the prediction ability with different \(k\) values (Granato, Branco, Cruz, & Faria, 2011). Small \(k\) values between 3 and 5 are often preferred. The method presents several advantages: (i) its mathematical simplicity, which does not prevent it from achieving classification results as good as (or even better than) other more complex pattern recognition techniques, (ii) it is free from statistical assumptions, such as the normal distribution of the variables, and (iii) its effectiveness does not depend on the space distribution of the classes (Berrueta, Alonso-Salces, & Heberger, 2007). In this work, the Euclidean distance was used as the criterion for calculating the distance between samples, and the number of neighbors (\(K\)) was selected after studying the success of the classification with different \(K\) values applied to a training set with all the samples.

The PLSDA aims to find the variables and directions in the multivariate space which discriminate the established classes in the calibration set. Using this algorithm, one can build a dummy binary-coded response vector, so that, if a sample belongs to class one, it will have a “1” as the first component and the remaining components will be zero. If it belongs to class two, it will have a “1” as second component, while the remaining will be zero, and so on. Until these assumptions, compute a classification model corresponds to calculating the regression vector between the data matrix and the dummy vector of responses. The result is a linear model that has proven to be statistically equivalent to the solution obtained to Linear Discriminant Analysis (Marini, 2010).

The main idea in SIMCA is to build a confidence limit for each class with the help of PCA, project the unclassified samples into each principal components space and assign them to the class in which they fit best. Selection of the optimal number of PCs is an important step in SIMCA and is usually determined using a leave-one object out cross-validation (Stanimirova et al., 2010).

Table 2 shows recognition and prediction abilities afforded with each chemometric technique. Overall, excellent results were found for the full fat class (F), independent of the chemometric technique used while for the low-fat class (L), these findings are dependent on the method chosen. KNN achieved recognition and prediction percentages of 100% for both categories, while for PLSDA the results are 100% for the F class and 86% and 75% for L class, respectively. Following this, SIMCA attained excellent results for the F class, with 100 and 100% of correct classification and prediction, respectively. However poor results were found for the L class, in which 80 and 50.0% at the recognition and prediction step, respectively. These results are in perfect agreement with HCA and PCA results, which shows a superposition of full and low-fat samples.

Fig. 3 shows the Coomans diagram corresponding to the models with the complete data set. In this diagram, the axes represent categories and the ‘co-ordinates’ in these axes measure the distance of each sample from both categories. Thus, each model is drawn as a rectangle that is defined by the 95% critical distance. The horizontal rectangle corresponds to low fat yogurts and the vertical one to the full fat yogurts. Any sample whose distance to the corresponding centroid is greater than the critical distance is outside the model and is, therefore, rejected (graphically, it is plotted outside the corresponding rectangle). On the other hand, the samples located in the lower left square of the diagram belong to both models. So, if there are a large number of samples that do not belong to the category and they are plotted in this zone, it is said that the model has low specificity (Melendez, Sanchez, Iniguez, Sarabia, & Ortiz, 2001). Confirming the HCA results, it is shown that there are some samples belonging to both classes located at the lower right part of the plot, which represents an overlap area among the yogurt classes.

Recently, the use of artificial neural networks was found to be effective in classifying low-fat yogurts using instrumental analysis as variables, with a 100% performance in the previous stage (Cruz et al., 2009). However, the use of artificial neural networks do not require to know a causal knowledge of the relationship between the input and the output variables. This is because artificial neural networks are considered a class of learning methods that learn the relationships among the variables through successive training (Debska & Guzowska-Swider, 2011).

---

**Fig. 2.** Principal component analysis similarity map determined for the first two principal components, PC1 and PC2.
Table 2
Classification results for the supervised patterns.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Class</th>
<th>Recognition ability (%)</th>
<th>Prediction ability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-DA</td>
<td>F</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>KNN (K = 2); inverse squared Euclidean distance</td>
<td>F</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>SIMCA; normal range, ( \alpha = 0.05 )</td>
<td>F</td>
<td>85.7</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>80.0</td>
<td>50.0</td>
</tr>
</tbody>
</table>

![Fig. 3. Coomans plot for the squared SIMCA distances obtained from the complete data set. L = low fat yogurt, F = full fat yogurt.](image)

The results of this work suggest that it is also possible to use chemometric techniques, which presents a well-established mathematical theory, to classify yogurts according to their fat content. The chemometric classification methods – in particular KNN and PLSDA – were totally successful in the stage of recognizing and predicting full-fat yogurts category, and they can be used for discriminating this class, leaving only the low-fat yogurts class to be subjected to the traditional analytical method.

It is interesting to note, in a first understanding, the refereed parameters are not related to the fat content. However the use of chemometric methods allows finding mathematical relations which get possible to establish a link. The new methodology presents interesting with potential application, since it represents saving great amounts of chemical reagents, as it is now requires only a small number of physical–chemical analyses to be performed. In this context, the instrumental analyses of color, firmness and pH, together with chemometric methods, provide sufficient enough information to be used to discriminate yogurts with low-fat contents. These results can be interesting to the Health Agencies and dairy processors, which needs to control the quality of low fat yogurts to be in accordance with the current legislation.

However, the approach presented is limited to yogurts having the same composition as shown in Table 1, in either a qualitative or a quantitative way. Besides several possibilities exist for the use of different ingredients and technological parameters in the yogurt manufacture, any product that has different ingredients in its composition can be included in the model if an adequate number of samples are analyzed. Therefore, a new model should be elaborated and validated.

4. Conclusion

Low and full-fat yogurts presented differences with regard to all the variables analyzed. However, using PCA and HCA, it was not possible to obtain total separation between the samples, which was only achieved using supervised classification techniques – that is, KMN and PLSDA – particularly as a result of their excellent performance relative to the full-fat category.

The results of this research show that the newly proposed method, in combination with chemometric techniques, has a significant potential to be used for discriminating yogurts with regards to their fat content. This, in turns, will contribute with a low generation of chemical residues and a reduction in the amount of chemical analysis performed.

References


PIROUETTE. InfoMetrix, Inc., versão 2.02, Seattle, Washington.
