Analytica Chimica Acta xxx (2016) 1–10

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Two-dimensional linear discriminant analysis for classification of three-way chemical data

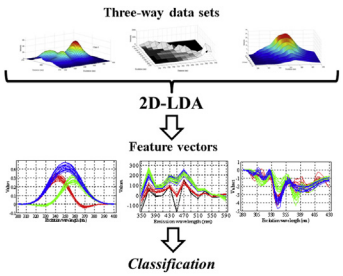
Adenilton C. da Silva [a], Sófacles F.C. Soares [a, b], Matías Insausti [c], Roberto K.H. Galvão [d], Beatriz S.F. Band [c], Mário César U. de Araújo [a, *]

[a] Universidade Federal da Paraíba, Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Caixa Postal 5093, CEP 58051-970, João Pessoa, PB, Brazil
[b] Departamento de Engenharia Química, Centro de Tecnologia (CT), Universidade Federal da Paraíba, 58051-900, João Pessoa, PB, Brazil
[c] FIA Laboratory, Analytical Chemistry Section, INQUISUR (UNS-CONICET), Av. Alem 1253, B8000CPB, Bahía Blanca, Buenos Aires, Argentina
[d] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil

HIGHLIGHTS

- Use of 2D-LDA for extraction of classification features from three-way chemical data.
- Case studies involving simulated data and real-life data sets: Parma ham and edible vegetable oils.
- Use of surface autofluorescence and total synchronous fluorescence spectrometries.
- Better results compared with the use of spectral data with no feature extraction.
- Better results compared with PLS Discriminant Analysis applied to the unfolded data, as well as PARAFAC-LDA and TUCKER3-LDA.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form
15 July 2016
Accepted 4 August 2016
Available online xxx

Keywords:
Two-dimensional linear discriminant analysis
PARAFAC-LDA
TUCKER3-LDA
Three-way fluorescence data
Dry-cured Parma ham
Edible vegetable oil

ABSTRACT

The two-dimensional linear discriminant analysis (2D-LDA) algorithm was originally proposed in the context of face image processing for the extraction of features with maximal discriminant power. However, despite its promising performance in image processing tasks, the 2D-LDA algorithm has not yet been used in applications involving chemical data. The present paper bridges this gap by investigating the use of 2D-LDA in classification problems involving three-way spectral data. The investigation was concerned with simulated data, as well as real-life data sets involving the classification of dry-cured Parma ham according to ageing by surface autofluorescence spectrometry and the classification of edible vegetable oils according to feedstock using total synchronous fluorescence spectrometry. The results were compared with those obtained by using the spectral data with no feature extraction, U-PLS-DA (Partial Least Squares Discriminant Analysis applied to the unfolded data), and LDA employing TUCKER-3 or PARAFAC scores. In the simulated data set, all methods yielded a correct classification rate of 100%. However, in the Parma ham and vegetable oil data sets, better classification rates were obtained

* Corresponding author.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araújo).

http://dx.doi.org/10.1016/j.aca.2016.08.009
0003-2670/Published by Elsevier B.V.

by using 2D-LDA (86% and 100%), compared with no feature extraction (76% and 77%), U-PLS-DA (81% and 92%), PARAFAC-LDA (76% and 86%) and TUCKER3-LDA (86% and 93%).

Published by Elsevier B.V.

1. Introduction

The use of analytical techniques that record a data matrix for each sample has become increasingly more common, as the result of advances in analytical instrumentation and methods. Examples include EEM (Excitation - Emission Matrix fluorescence spectroscopy), as well as hyphenated techniques such as HPLC-DAD (High Performance Liquid Chromatography - Diode Array Detector), GC-MS (Gas Chromatography – Mass Spectrometry), and LC-MS (Liquid Chromatography – Mass Spectrometry), among others [1–3]. Such techniques are able to provide adequate sensitivity to the analytes even in the presence of interferents, owing to the large amount of acquired information. However, the complexity of the data structure may pose difficulties for its interpretation. In this context, several chemometric tools have been employed in the literature for identification or quantification of chemical species [4–7], as well as classification of samples [8–12].

A usual approach for the classification of data matrices is based on the use of algorithms for data decomposition or unfolding, in order to obtain features that can be used with standard techniques for multivariate classification. For instance, PARAFAC (Parallel Factor analysis) [8] scores have been employed for classification of Sherry vinegar samples according to ageing using PLS-DA (Partial Least Squares - Discriminant Analysis) and SVM (Support Vector Machines) [9], characterization and classification of honey samples using PLS-DA [10], and discrimination of bacteria using LDA (Linear Discriminant Analysis) [5]. Another example consists of the use of NMF (Non-negative Matrix Factorization) [11] for compression of EEM data in order to authenticate olive oil samples employing LDA models [12].

It may be argued that the handling of two-dimensional analytical data could benefit from the use of image processing algorithms, which are also concerned with 2-D data structures. An interesting approach for image classification is the two-dimensional linear discriminant analysis (2D-LDA) algorithm, which was originally proposed by Li et al. [13] in the context of face image processing. This algorithm is based on the extraction of feature vectors from the data matrices by using projection vectors optimized with respect to the Fisher criterion [14–16]. Similarities between different images can then be evaluated in terms of the distance between the corresponding feature vectors. Similar strategies were also proposed by Liang et al. [17] and Cho et al. [18]. One of the main advantages of 2D-LDA consists of the reduction in the dimension of the data matrices while preserving relevant information for classification purposes.

Despite its promising performance in image processing tasks, the 2D-LDA algorithm has not yet been employed in applications involving chemical data. In order to bridge this gap, the present paper presents an investigation of the use of 2D-LDA in classification problems involving three-way spectral data. A brief review of 2D-LDA is initially presented, followed by a description of a distance-based procedure to classify test data on the basis of 2D-LDA features. The investigation of 2D-LDA performance in chemical classification tasks is then carried out by using three data sets, involving: (1) simulated EEM data, (2) surface autofluorescence spectra of dry-cured Parma ham samples and (3) synchronous fluorescence spectra of edible vegetable oils. The results are compared with those obtained by the distance-based procedure without feature extraction, as well as PLS-DA applied to the unfolded data (U-PLS-DA), and LDA employing either PARAFAC or TUCKER-3 [19] scores.

2. Background and theory

2.1. Notation

Matrices, vectors and scalars are represented by boldface capital, boldface lowercase and italic (either capital or lowercase) letters, respectively. The T and -1 superscripts denote the transpose and the inverse of a matrix. The dimensions of matrices and vectors are indicated within parentheses. The (i, j) element of a matrix $\mathbf{X}$ is denoted by $X_{i,j}$.

The L classes involved in the problem are indicated by C_1, C_2, …, C_L. It is assumed that a training set of N matrices $\mathbf{X}_k$ ($k = 1, 2, \ldots, N$) corresponding to samples of known classification are available for use in the 2D-LDA algorithm. The index set of the N_p training samples belonging to the pth class will be denoted by I_p, with $p = 1, 2, \ldots, L$. The notation $\sum_{k \in I_p} \mathbf{X}_k$ will be employed to indicate the sum of the matrices corresponding to the samples in the pth class.

The notation trace($\mathbf{M}$) indicates the trace of a square matrix $\mathbf{M}$, i.e. the sum of its diagonal elements. By using this notation, the sum of squares of the elements of a matrix $\mathbf{X}$ ($m \times n$) can be expressed as $trace(\mathbf{X}^T\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{i,j})^2$.

2.2. Two-dimensional linear discriminant analysis (2D-LDA)

2.2.1. Determination of the projection vectors

Let $\mathbf{X}$ be an ($m \times n$) matrix of data recorded for a given sample. In the case of EEM data, for example, the number of rows (m) and columns (n) correspond to the number of emission and excitation wavelengths, respectively. A feature vector $\mathbf{y}$ ($m \times 1$) is obtained by multiplying $\mathbf{X}$ by a projection vector $\mathbf{b}$ ($n \times 1$), as

$$\mathbf{y} = \mathbf{Xb} \tag{1}$$

The ith component of vector $\mathbf{y}$ is given by the scalar product between the ith row of matrix $\mathbf{X}$ and the projection vector $\mathbf{b}$, i.e.

$$y_i = \sum_{j=1}^{n} X_{i,j} b_j \tag{2}$$

In the case of EEM data, for example, the ith row of $\mathbf{X}$ corresponds to the excitation spectrum for the ith emission wavelength, as illustrated in Fig. 1.

An optimal projection vector $\mathbf{b}_{opt}$ can be obtained by maximizing Fisher's linear projection criterion [14–16] as

$$\mathbf{b}_{opt} = \arg\max_{\mathbf{b}} \frac{\mathbf{b}^T\mathbf{S_B}\mathbf{b}}{\mathbf{b}^T\mathbf{S}_W\mathbf{b}} \tag{3}$$

where $\mathbf{S_B}$ ($n \times n$) and $\mathbf{S_W}$ ($n \times n$) denote the between-class and within-class scatter matrices, which are calculated from the set of training data as

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

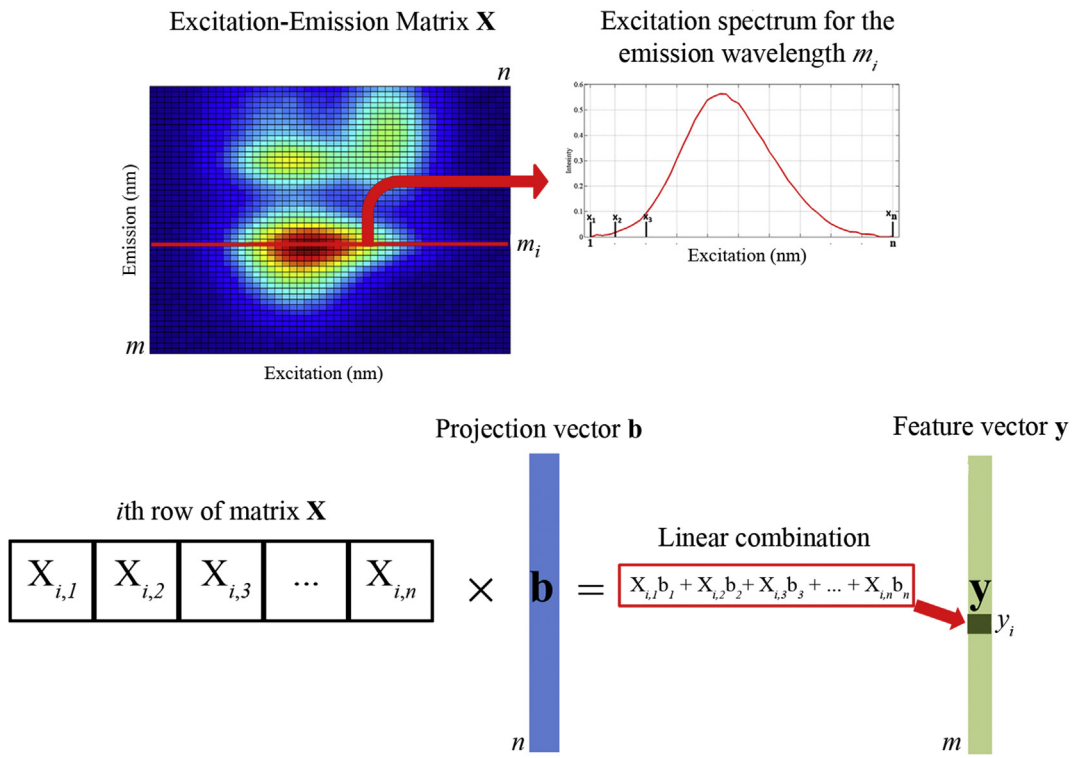


Fig. 1. Calculation of each element of a 2D-LDA feature vector from EEM data.

$$\mathbf{S_B} = \sum_{p=1}^{L} N_p (\overline{\mathbf{X}}_p - \overline{\mathbf{X}})^T (\overline{\mathbf{X}}_p - \overline{\mathbf{X}}) \tag{4}$$

$$\mathbf{S_W} = \sum_{p=1}^{L} \sum_{k \in I_p} (\mathbf{X}_k - \overline{\mathbf{X}}_p)^T (\mathbf{X}_k - \overline{\mathbf{X}}_p) \tag{5}$$

where $\overline{\mathbf{X}}_p$ denotes the mean of the $\mathbf{X}$ matrices for samples belonging to class C_p and $\overline{\mathbf{X}}$ denotes the mean over the entire data set, i.e.

$$\overline{\mathbf{X}}_p = \frac{1}{N_p} \sum_{k \in I_p} \mathbf{X}_k \tag{6}$$

$$\overline{\mathbf{X}} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{X}_k \tag{7}$$

If $\mathbf{S_W}$ is nonsingular, $\mathbf{b}_{\text{opt}}$ can be obtained as an eigenvector resulting from a generalized eigenvalue problem, i.e. $\mathbf{b}_{\text{opt}}$ must satisfy the following equation:

$$\mathbf{S_W^{-1}} \mathbf{S_B} \mathbf{b}_{\text{opt}} = \lambda \mathbf{b}_{\text{opt}} \tag{8}$$

where λ is the largest eigenvalue of matrix $\mathbf{S_W^{-1}S_B}$ ($n \times n$). This procedure can be extended in order to obtain up to M projection vectors, in decreasing order of relevance for the classification problem, where $M \leq n$ is the rank of $\mathbf{S_W^{-1}S_B}$. The qth projection vector $\mathbf{b}_q$ is obtained as the solution of

$$\mathbf{S_W^{-1}} \mathbf{S_B} \mathbf{b}_q = \lambda_q \mathbf{b}_q \tag{9}$$

where λ_q is the qth largest eigenvalue of $\mathbf{S_W^{-1}S_B}$, with $q = 1, 2, \ldots, M$.

Let $r \leq M$ be the number of projection vectors obtained in this manner. By arranging $\mathbf{b}_1$, $\mathbf{b}_2$, …, $\mathbf{b}_r$ as the columns of a projection matrix $\mathbf{B}$ ($n \times r$), a feature matrix $\mathbf{Y}$ ($m \times r$) can be calculated from a given data matrix $\mathbf{X}$ as

$$\mathbf{Y} = \mathbf{XB} \tag{10}$$

The columns of $\mathbf{Y}$ correspond to r feature vectors in decreasing order of relevance for the classification problem.

Remark: The decomposition procedure described above could be alternatively applied to the transpose matrix $\mathbf{X}^T$, in order to extract features from the columns of $\mathbf{X}$, rather than the rows. In the case of EEM data, for example, the transposition of $\mathbf{X}$ amounts to switching the roles of the excitation and emission modes in the decomposition procedure. It is worth noting that $\mathbf{S_W}$ will have dimensions ($n \times n$) if the decomposition is applied to $\mathbf{X}$ and dimensions ($m \times m$) if the decomposition is applied to $\mathbf{X}^T$. Therefore, a possible guideline consists of choosing the option that results in the smallest dimensions for $\mathbf{S_W}$, in order to avoid ill-conditioning problems in the calculation of $\mathbf{S_W^{-1}}$. This is the criterion that will be adopted in the present work.

2.2.2. Classification procedure

A test sample $\mathbf{X}_{\text{test}}$ can be classified on the basis of the similarity of its feature matrix $\mathbf{Y}_{test} = \mathbf{X}_{test}\mathbf{B}$ with respect to the feature matrices $\mathbf{Y}_k = \mathbf{X}_k\mathbf{B}$, $k = 1, 2, \ldots, N$, of the samples in the training set. In the present work, the similarity is evaluated in terms of an Euclidian distance $d(\mathbf{Y}_{test}, \mathbf{Y}_k)$ calculated as

$$d(\mathbf{Y}_{test}, \mathbf{Y}_k) = \sqrt{trace\left[(\mathbf{Y}_{test} - \mathbf{Y}_k)^T (\mathbf{Y}_{test} - \mathbf{Y}_k)\right]} \tag{11}$$

for $k = 1, 2, \ldots, N$. The average distance between the test sample and the N_p training samples belonging to class C_p is then obtained as

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

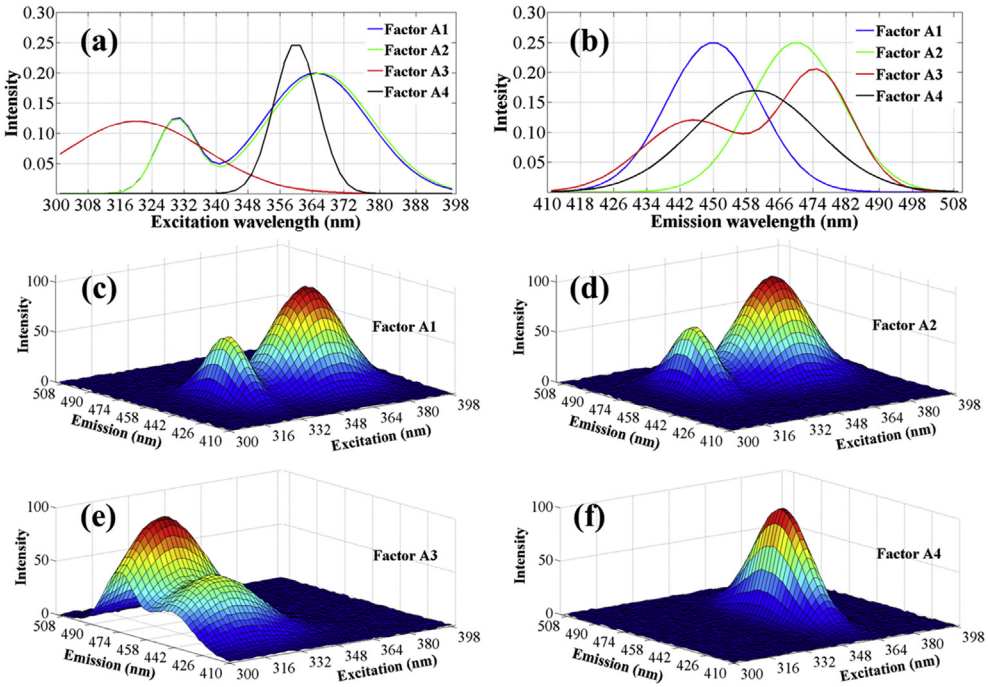


Fig. 2. Factors employed in the generation of the simulated EEM data set. (**a**) Excitation profiles, (**b**) emission profiles, (**c**) A1 factor, (**d**) A2 factor, (**e**) A3 factor, (**f**) A4 factor.

$$\overline{d}(\mathbf{Y}_{test}, C_p) = \frac{1}{N_p} \sum_{k \in I_p} d(\mathbf{Y}_{\text{test}}, \mathbf{Y}_k) \tag{12}$$

Finally, the test sample is assigned to the class p^* corresponding to the smallest average distance, i.e.

$$\overline{d}(\mathbf{Y}_{test}, C_{p^*}) = \min_{p=1,2,\ldots,L} \overline{d}(\mathbf{Y}_{test}, C_p) \tag{13}$$

3. Experimental

3.1. Simulated EEM data set

The simulated EEM data set was generated according to the following procedure:

a) The wavelengths were varied from 300 to 398 nm for excitation (Fig. 2a) and from 410 to 508 nm for emission (Fig. 2b), with a 2 nm interval, which resulted in an EEM data matrix with $m = 50$ rows and $n = 50$ columns for each sample.
b) The sample matrices were generated as the linear combination of up to four simulated factors (A1, A2, A3, A4, which are depicted in Fig. 2c–f, respectively);
c) Three classes were defined by varying the factors employed in the generation of the samples: Class 1 (factors A1, A2, A3), Class 2 (factors A2, A3, A4) and Class 3 (factors A1, A2, A3, A4). Examples of samples in each of these classes are presented in Fig. 3.
d) Within-class variability was simulated by randomly varying the linear combination coefficients in the generation of the samples. For this purpose, the coefficient values were drawn from Gaussian distributions with the means and standard deviations indicated in Table 1.

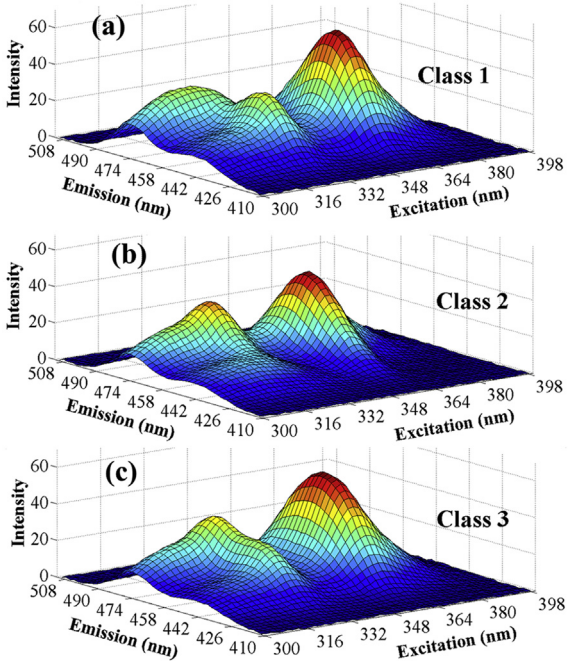


Fig. 3. Examples of simulated EEM data for (**a**) Class 1, (**b**) Class 2 and (**c**) Class 3.

e) Gaussian noise was added to resulting EEM matrices, with an intensity of 0.5% of the maximum peak value in each sample.
f) A total of 90 samples were generated (60 for training and 30 for test) as indicated in Table 2.

3.2. Dry-cured Parma ham data set

Parma ham is a traditional product from the city of Parma (Italy), with protected designation of origin, which develops a distinctive

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

Table 1
Mean and standard deviation (Std) values employed in the random generation of the coefficients for the construction of the simulated data sets.

Factors	Class 1		Class 2		Class 3	
	Mean	Std	Mean	Std	Mean	Std
Training set						
A1	0.8675	0.0954	–	–	0.8674	0.0765
A2	0.8194	0.0948	0.8352	0.0975	0.8563	0.0893
A3	0.8546	0.0836	0.8664	0.0949	0.8427	0.0889
A4	–	–	0.8833	0.0892	0.8501	0.0658
Test set						
A1	0.8893	0.0661	–	–	0.8744	0.0771
A2	0.9342	0.05396	0.8870	0.0636	0.8782	0.0481
A3	0.9147	0.0650	0.8914	0.0536	0.9001	0.0555
A4	–	–	0.9134	0.0446	0.9066	0.0576

Table 2
Division of the samples into training and test sets.

Data set	Class	Training set	Test set	Total
Simulated EEM data	Class 1	20	10	30
	Class 2	20	10	30
	Class 3	20	10	30
	Total	60	30	90
Dry-cured Parma ham	Raw	4	2	6
	Salted	9	5	14
	Matured	17	7	24
	Aged	16	7	23
	Total	46	21	67
Edible vegetable oil	Soybean	10	3	13
	Corn	14	6	20
	Sunflower	12	4	16
	Total	36	13	49

flavour and aroma after 12 months of maturation [20,21]. The data set employed herein, which was made publicly available at www.models.life.ku.dk/datasets by Moller et al. [22], concerns the use of surface autofluorescence spectroscopy as an alternative to traditional methods for the evaluation of ageing-related quality parameters. This data set was also employed by Durante et al. [8] for classification of samples according to ageing state using the N-SIMCA chemometrics tool.

The data set comprises 67 samples, with EEM spectra recorded at a BioView instrument (Delta Light and Optics, Lyngby, Denmark) fitted with an optical fibre probe. In the present work, the number of variables and the division of the samples into classes followed the study presented by Durante et al. [8]. Thus, the data matrix for each sample consisted of $m = 13$ emission wavelengths in the range 350–590 nm, and $n = 11$ excitation wavelengths in the range 270–470 nm. Four classes were defined on the basis of the ageing period of the Parma ham sample: raw meat, salted (3 months), matured (11–12 months) and aged (15–18 months). Fig. 4 presents EEM spectra for representative samples of each class. The division of the samples into training and test sets is presented in Table 2.

3.3. Edible vegetable oil data set

Edible vegetable oils have nutritional and health properties that vary according to the feedstock, in addition to other factors such as processing and storage [23,24]. The identification of feedstock can be a challenging task, in view of the similarity among some types of oil. In the present work, total synchronous fluorescence was employed, as an alternative to other analytical techniques reported in the literature [25–28].

Total synchronous fluorescence is a highly sensitive technique, which can be used as an alternative to standard molecular fluorescence to avoid the superposition of excitation and emission bands [29]. For this purpose, the emission and excitation monochromators are scanned simultaneously, with a constant wavelength difference ($\Delta\lambda = \lambda_{emission} - \lambda_{excitation}$) [30]. Applications reported in the literature include the assessment of adulterations in virgin olive oil [29], the discrimination between edible and lampante virgin olive oil [30], the classification of edible oils in n-hexane solutions [31] and the classification of biodiesel samples with respect to the type of oil feedstock [32].

The data set comprises 49 samples of edible oil from three types of feedstock: soybean (13 samples), corn (20 samples) and sunflower (16 samples). The spectral data were acquired with a computer-controlled Aminco Bowman Series 2 spectrofluorometer fitted with a xenon discharge light source (150 W). The measurements were carried out at a scan rate of 5 nm s^{-1}, with precision and repetitivity of ±0.5 nm and ±0.25 nm, respectively. For each sample, a volume of 600 μL was employed and eight synchronous spectra were obtained by moving the emission and excitation monochromators with constant wavelength differences ($\Delta\lambda$) of 10, 15, 20, 25, 30, 35, 40, and 45 nm. The excitation range was the same for all spectra (280–430 nm), whereas the emission range varied from 290-440 nm to 325–475 nm, according to the wavelength difference ($\Delta\lambda$) employed. Fig. 5 presents spectra for representative samples of each class.

In order to use the 2D-LDA algorithm, the spectral data for each sample were arranged in a matrix with $m = 150$ rows

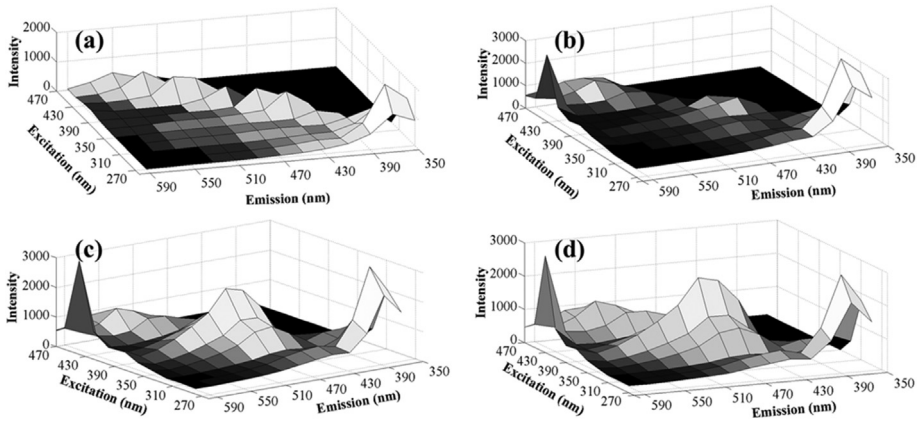


Fig. 4. Examples of surface autofluorescence spectra of Parma ham samples: (**a**) raw, (**b**) salted, (**c**) matured, (**d**) aged.

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

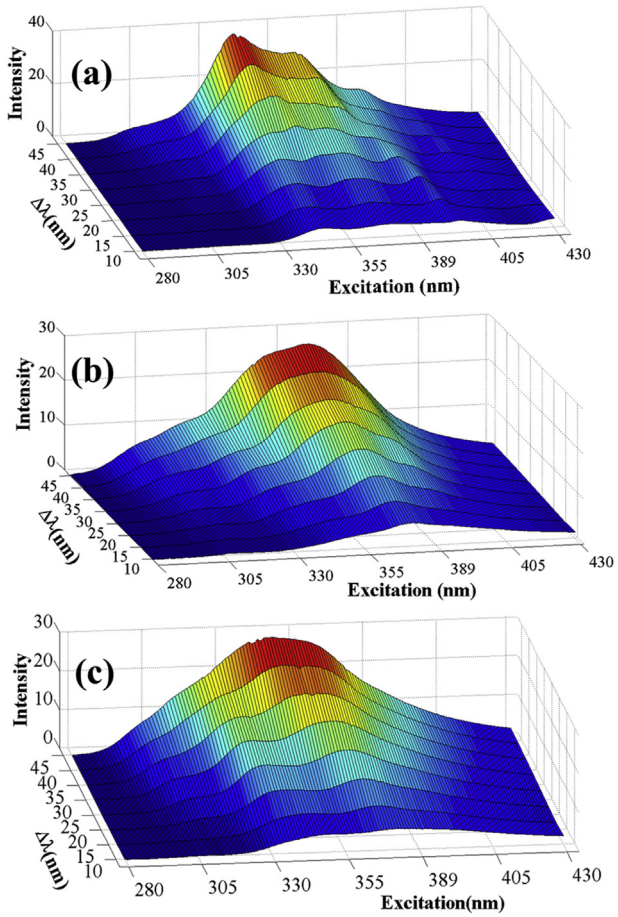


Fig. 5. Examples of total synchronous fluorescence spectra of edible vegetable oil samples: (**a**) soybean, (**b**) corn, (**c**) sunflower.

corresponding to the excitation wavelengths ($\lambda_{excitation}$), and $n = 8$ columns corresponding to the wavelength differences ($\Delta\lambda$).

The division of the samples into training and test sets is presented in Table 2.

3.4. Software

The computation of 2D-LDA projection vectors and the calculation of Euclidian distances described in Section 2.2.1 and 2.2.2 were implemented in Matlab®2010b (Mathworks) by using a lab-made code. The best number r of projection vectors was determined on the basis of the correct classification rate (CCR) obtained by leave-one-out cross-validation in the training set. If the same CCR was obtained with different values of r, the smallest value was selected, for parsimony reasons. Finally, the selected projection vectors were employed in the classification of the test samples.

For comparison, the classification procedure described in Section 2.2.2 was also employed by using the original data, instead of the feature matrices. For this purpose, the calculations in Equations (11)–(13) were carried out by using $\mathbf{X}_{test}$ and $\mathbf{X}_k$ in place of $\mathbf{Y}_{test}$ and $\mathbf{Y}_k$, respectively.

In addition, the results were compared with those obtained by using U-PLS-DA, PARAFAC-LDA and TUCKER3-LDA. The unfolding operation employed in U-PLS-DA consists of concatenating the rows of each data matrix $\mathbf{X}$ ($m \times n$) into a row vector ($1 \times nm$). As a result, the N training samples are arranged in the form of a single matrix of dimensions ($N \times nm$) [33], as in the standard PLS-DA classification algorithm [34]. More details concerning this toolbox can be found elsewhere [35]. In the PARAFAC-LDA and TUCKER3-LDA algorithms, the score values obtained from the decomposition of the three-way data were employed as inputs to a standard LDA classifier. Non-negativity and orthogonality constraints were adopted in the PARAFAC and TUCKER-3 decompositions, respectively. The number of latent variables in U-PLS-DA and decomposition factors in PARAFAC-LDA and TUCKER3-LDA were chosen in order to maximize the CCR value obtained by cross-validation. If the maximal CCR value was obtained with different numbers of latent variables/factors, the smallest number was selected, for parsimony reasons, as in the 2D-LDA case.

The PLS-DA calculations on the unfolded data were carried out in Matlab®2010b (Mathworks) by using the Classification Toolbox 3.1 available at http://michem.disat.unimib.it/chm/download/softwares.htm. The TUCKER-3-LDA and PARAFAC-LDA calculations were carried out in Matlab®2010b (Mathworks) by using the N-way toolbox v. 3.30 available at http://www.models.life.ku.dk/algorithms and a lab-made LDA code.

4. Results and discussion

4.1. Simulated EEM data set

In order to use the 2D-LDA algorithm, the training data set was initially employed to calculate the $\mathbf{S_B}$ and $\mathbf{S_W}$ matrices, as in Equations (4) and (5). As a result, $\mathbf{S_W^{-1}S_B}$ was a full-rank matrix of dimensions (50×50). Equation (9) was then solved with $q = 1, 2, \ldots, 50$ to obtain the projection vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{50}$. Fig. 6a presents the correct classification rate obtained by cross-validation in the training set, as a function of the number r of projection vectors included in the projection matrix $\mathbf{B}$. As can be seen, a correct classification rate of 100% was obtained by using $r = 1$ up to $r = 5$ projection vectors. Therefore, in view of the parsimony criterion, the final classification model included only the first projection vector.

Fig. 7 presents the feature vectors obtained for the training samples. As can be seen, there are clear differences among the three classes. It is worth noting that the feature vectors for the samples in classes 1, 2 and 3 are similar to the emission factors A1, A2, and A4, respectively, which are shown in Fig. 2b.

Finally, by using the 2D-LDA projection vector obtained in the training set, all test samples were correctly classified.

In this case, a perfect classification of the test samples was also achieved by using the PARAFAC-LDA, TUCKER3-LDA and U-PLS-DA classifiers (with factors/latent variables chosen on the basis of the cross-validation results in Fig. 6b, c, d), as well the distance-based method with no feature extraction. It is worth noting that this simulated example does not pose significant difficulties for classification and was mainly aimed at illustrating the steps involved in the 2D-LDA classification procedure. The advantages of 2D-LDA will be more apparent in the two next case studies, which involve real-life data sets of a more complex nature.

4.2. Dry-cured Parma ham data set

In this case, a total of $n = 11$ projection vectors were obtained in the 2D-LDA algorithm. Fig. 8 presents the correct classification rate obtained by cross-validation in the training set, as a function of the number of projection vectors employed. On the basis of the parsimony criterion, the best choice consisted of using 5 projection vectors.

Fig. 9 presents the feature vectors obtained for the samples in the training set. As can be seen in Fig. 9a, the first projection vector provides a clear discrimination between the raw and salted classes, which are also separated from the remaining two classes (matured

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

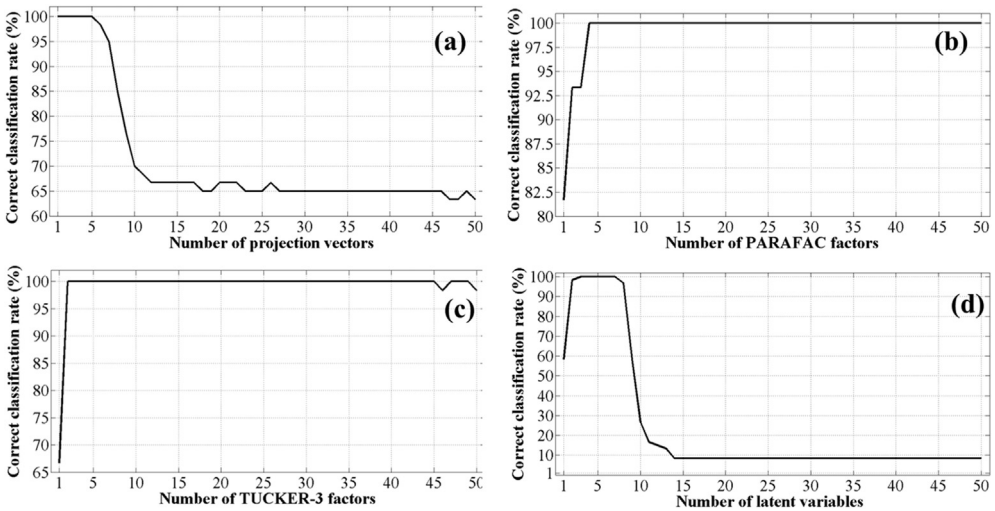


Fig. 6. Simulated EEM data set: Correct classification rate obtained by cross-validation *versus* number of (a) 2D-LDA projection vectors, (b) PARAFAC factors, (c) TUCKER-3 factors and (d) latent variables in U-PLS-DA.

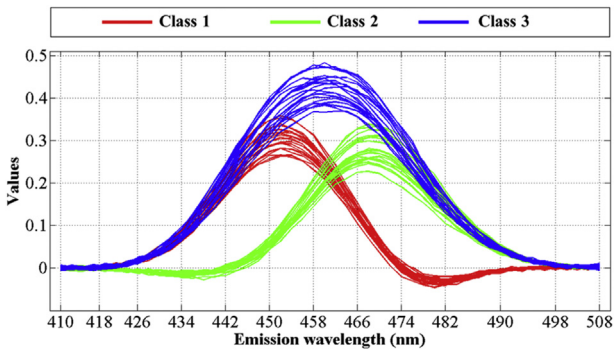


Fig. 7. Simulated EEM data set: 2D-LDA features of the training samples.

The test samples were classified by using the 2D-LDA procedure with the five projection vectors obtained in the training set. The results are presented in the form of a confusion matrix in Table 3. All the raw and salted samples were correctly assigned to their true classes, which is consistent with the clear class separation observed in the training set. Some classification errors involving the matured and aged samples were obtained, which is also consistent with the training set results. These findings are in agreement with previous investigations [8] [22], which also revealed an overlapping between the matured and aged classes in terms of PARAFAC scores. As discussed in Ref. [22], this overlapping may be ascribed to the fluorescence profile of tertiary lipid oxidation products, which emerge late in the maturation of dried hams.

For comparison, Table 4 presents the CCR values obtained by

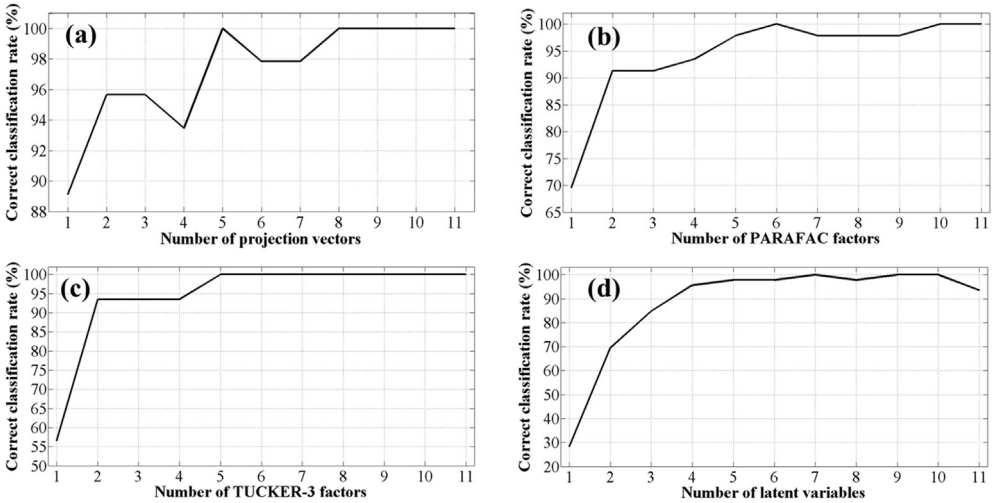


Fig. 8. Parma ham data set: Correct classification rate obtained by cross-validation *versus* number of (a) 2D-LDA projection vectors, (b) PARAFAC factors, (c) TUCKER-3 factors and (d) latent variables in U-PLS-DA.

and aged). The distinction between the matured and aged classes is less evident, but some separation can be observed in the features corresponding to the third (Fig. 9c) and fifth (Fig. 9e) projection vectors.

using the PARAFAC-LDA, TUCKER3-LDA and U-PLS-DA classifiers (with factors/latent variables chosen on the basis of the cross-validation results in Fig. 8b, c, d), as well the distance-based method with no feature extraction. As can be seen, 2D-LDA

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

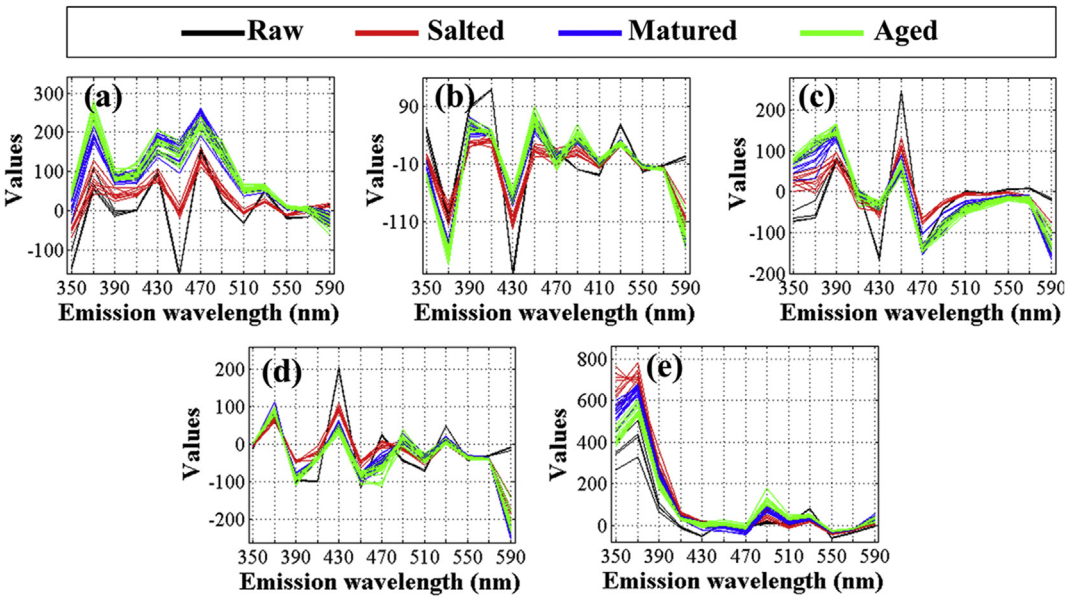


Fig. 9. Parma ham data set: 2D-LDA features of the training samples obtained with (**a**) first, (**b**) second, (**c**) third, (**d**) fourth, and (**e**) fifth projection vectors.

Table 3
Parma ham data set: Classification results in the test set.

Actual class	Predicted class			
	Raw	Salted	Matured	Aged
Raw	100%	–	–	–
Salted	–	100%	–	–
Matured	–	–	86%	14%
Aged	–	–	29%	71%

outperformed all the other methods, with the exception of TUCKER3-LDA, which yielded the same CCR value (86%).

4.3. Edible vegetable oil data set

In this case, a total of $n = 8$ projection vectors were obtained in the 2D-LDA algorithm. As shown in Fig. 10, the best cross-validation result was obtained by using 4 projection vectors, which resulted in

Table 4
Comparative results: Correct classification rates obtained in the test sets. The number of projection vectors, factors or latent variables employed in each model is indicated in parenthesis.

Data set	2D-LDA	PARAFAC-LDA	TUCKER3-LDA	U-PLS-DA	No feature extraction
Simulated EEM data	100% (1)	100% (4)	100% (2)	100% (3)	100%
Dry-cured Parma ham	86% (5)	76% (6)	86% (5)	81% (7)	76%
Edible vegetable oils	100% (4)	86% (5)	93% (5)	92% (5)	77%

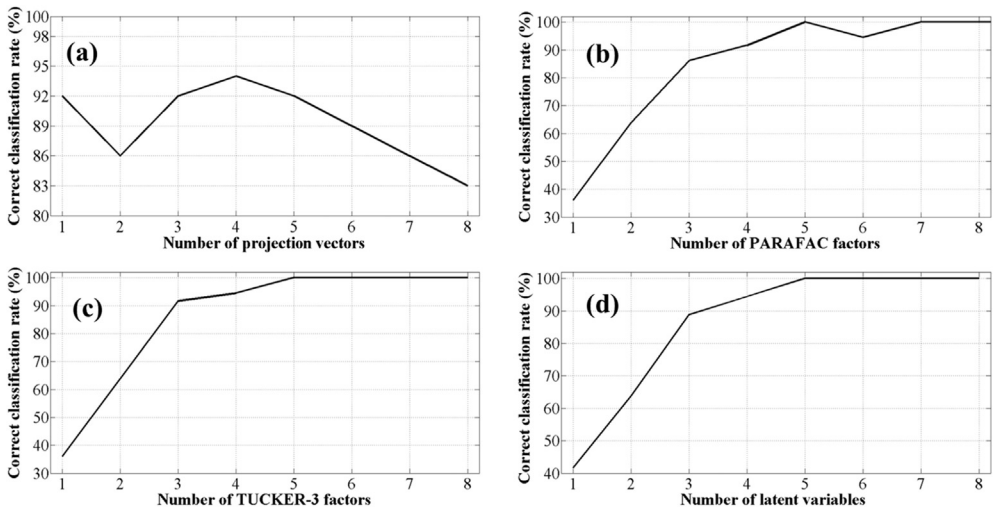


Fig. 10. Edible vegetable oil data set: Correct classification rate obtained by cross-validation *versus* number of (a) 2D-LDA projection vectors, (b) PARAFAC factors, (c) TUCKER-3 factors and (d) latent variables in U-PLS-DA.

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009

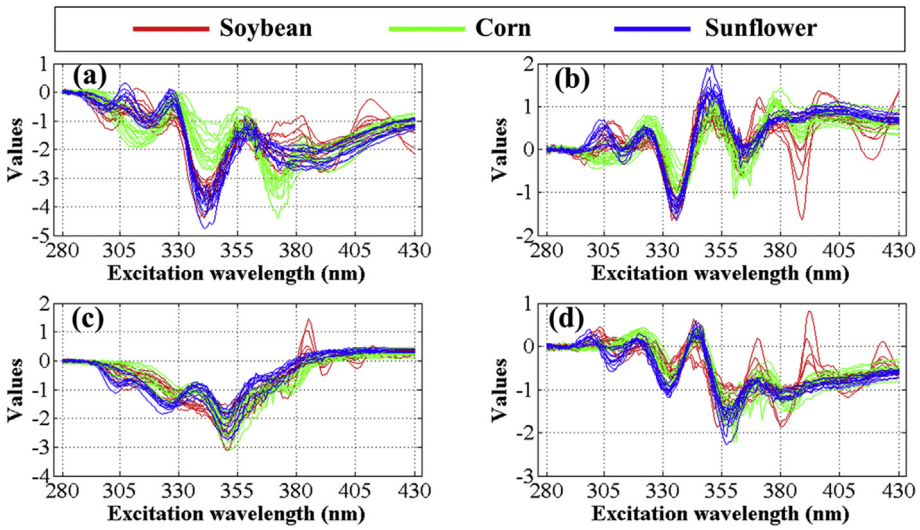


Fig. 11. Edible vegetable oil data set: 2D-LDA features of the training samples obtained with (**a**) first, (**b**) second, (**c**) third, and (**d**) fourth projection vectors.

a correct classification rate of 94%.

Fig. 11 presents the feature vectors obtained in the training set, which in this case are associated to excitation wavelengths in the range 280–430 nm. A separation between soybean and the other two classes can be observed between 330 and 340 nm (third projection, Fig. 11c) and also around 390 nm (second and fourth projections, Fig. 11b and d). The corn samples have a distinctive feature profile between 305 and 380 nm (first projection, Fig. 11a). In the case of sunflower, the best discrimination from the other classes is observed in the third and fourth projections (Fig. 11c and d), between 305 and 330 nm. These wavelength ranges are similar to those employed in a previous study [29] concerning the use of synchronous fluorescence for quantification of olive oil adulterations by soybean (315–365 nm), corn (315–392 nm), sunflower (315–365 nm) and other oils.

By using the four 2D-LDA projection vectors, all the test samples were correctly classified. In contrast, the other methods employed for comparison yielded CCR values ranging from 77% to 92%, as shown in Table 4. Interestingly, the PARAFAC-LDA, TUCKER3-LDA and U-PLS-DA classifiers achieved a CCR of 100% in cross-validation, as can be seen in Fig. 10. Such a finding suggests that 2D-LDA may have better generalization capabilities, i.e. better ability to classify samples that were not included in the model-building procedure.

On the overall, the results in Table 4 indicate that the use of decomposition methods tends to improve the classification accuracy, as compared to the simple distance-based criterion applied to the original data (i.e. with no feature extraction). Within this scope, the better classification performance of 2D-LDA over the other decomposition-based methods may be ascribed to the use of features specifically related to the discrimination among the classes.

5. Conclusion

This paper presented an investigation of two-dimensional linear discriminant analysis (2D-LDA) as a feature extraction tool for use in classification problems involving three-way chemical data. A brief review of the 2D-LDA algorithm was given and a distance-based classification method based on Euclidean distances between the feature matrices was described.

The use of 2D-LDA for classification of three-way chemical data was illustrated by using simulated EEM data, as well as real-life data sets involving the classification of dry-cured Parma ham according to ageing by surface autofluorescence spectrometry and the classification of edible vegetable oils according to feedstock using total synchronous fluorescence spectrometry.

The 2D-LDA classification results were compared with those obtained by using the distance-based procedure with no feature extraction (i.e. with the original spectral data), as well as by using U-PLS-DA (Partial Least Squares Discriminant Analysis applied to the unfolded data), PARAFAC-LDA (LDA employing PARAFAC scores) and TUCKER3-LDA (LDA employing TUCKER-3 scores). In the simulated data set, all methods yielded a correct classification rate of 100%. However, in the Parma ham and vegetable oil data sets, better classification rates were obtained by using 2D-LDA (86% and 100%), compared with the distance-based procedure with no feature extraction (76% and 77%), U-PLS-DA (81% and 92%), PARAFAC-LDA (76% and 86%) and TUCKER3-LDA (86% and 93%). These findings indicate that 2D-LDA is indeed a promising method for the classification of samples on the basis of three-way chemical data.

Acknowledgements

The authors acknowledge the support of CAPES/SPU (Brazil/Argentina international cooperation grant PPCP013/2011), CONICET (Argentina) and CNPq (Brazil) for research fellowships (303714/2014-0 and 301569/2009-6) and scholarships (160951/2013-5).

References

[1] A.C. Olivieri, G.M. Escandar, H.C. Goicoechea, A.M. de la Peña, Fundamentals and Analytical Applications of Multiway Calibration, first ed., vol. 29, Elsevier, 2015.
[2] V. Gomez, M.P. Callao, Analytical applications of second-order calibration methods, Anal. Chim. Acta 627 (2008) 169–183.
[3] A.C. Olivieri, G. Escandar, Practical Three-way Calibration, first ed., Elsevier, 2014.
[4] G.M. Escandar, H.C. Goicoechea, A.M. de la Peñac, A.C. Olivieri, Second- and higher-order data generation and calibration, A Tutor. Anal. Chim. Acta 806 (2014) 8–26.
[5] K.S. Booksh, B. Bronk, J. Czege, Three-way calibration, Compr. Chemom. 3 (2009) 379–412.
[6] Y. Wang, O.S. Borgen, B.R. Kowalski, M. Gu, F. Turecek, Advances in second-order calibration, J. Chemom. 7 (1993) 117–130.
[7] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Sys 38 (1997) 149–171.
[8] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on

SIMCA methodology, Chemom. Intell. Lab. Sys 106 (2011) 73–85.
[9] R.M. Callejóna, J.M. Amigo, E. Pairo, S. Garmón, J.A. Ocaña, M.L. Morales, Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches, Talanta 88 (2012) 456–462.
[10] L. Lenhardt, R. Bro, I. Zeković, T. Dramićanin, M.D. Dramićanin, Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey, Food Chem. 175 (2015) 284–291.
[11] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[12] F. Guimet, R. Boqué, J. Ferré, Application of non-negative matrix factorization combined with Fisher's linear discriminant analysis for classification of olive oil excitation–emission fluorescence spectra, Chemom. Intell. Lab. Sys. 81 (2006) 94–106.
[13] M. Li, B. Yuan, 2D-LDA: a statistical linear discriminant analysis for image matrix, Pattern Recogn. Lett. 26 (2005) 527–532.
[14] A. Rozza, G. Lombardi, E. Casiraghi, P. Campadelli, Novel Fisher discriminant classifiers, Pattern Recog 45 (2012) 3725–3737.
[15] Y. Xu, J. Yang, Z. Jin, A novel method for Fisher discriminant analysis, Pattern Recog 37 (2004) 381–384.
[16] Z. Ji, P. Jing, T. Yu, Y. Su, C. Liu, Ranking Fisher discriminant analysis, Neurocomputing 120 (2013) 54–60.
[17] Z. Liang, Y. Li, P. Shi, A note on two-dimensional linear discriminant analysis, Pattern Recogn. Lett. 29 (2008) 2122–2128.
[18] D.U. Cho, U.D. Chang, B.H. Kim, S.H. Lee, Y.L.J. Bae, S.C. Ha, 2D Direct LDA algorithm for Face recognition, in: Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06), 2006.
[19] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966).
[20] L. Bolzoni, G. Barbieri, R. Virgili, Changes in volatile compounds of Parma ham during maturation, Meat Sci. 43 (1996) 301–310.
[21] A.K. Agarwal, Business and Intellectual Property: Protect Your Ideas, first ed., Random House India, London, 2010.
[22] J.K. Møller, G. Parolari, L. Gabba, J. Christensen, L.H. Skibsted, Monitoring chemical changes of dry-cured Parma ham during processing by surface autofluorescence spectroscopy, J. Agric. Food Chem. 51 (2003) 1224–1230.
[23] R.D. O'brien, Fats and Oils: Formulating and Processing for Applications, third ed., CRC Press, Florida, 2009.
[24] S.C. Savva, A. Kafatos, Vegetable oils: dietary importance, Encycl. Food Health (2016) 365–372.
[25] C.E.T. da Silva, V.L. Filardi, I.M. Pepe, M.A. Chaves, C.M.S. Santos, Classification of food vegetable oils by fluorimetry and artificial neural networks, Food Control.. 47 (2015) 86–91.
[26] Y.G. Martın, J.L.P. Pavón, B.M. Cordero, C.G. Pinto, Classification of vegetable oils by linear discriminant analysis of Electronic Nose data, Anal. Chim. Acta 384 (1999) 83–94.
[27] A.S. Luna, A.P. da Silva, J. Ferré, R. Boqué, Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy, Spectrochim. Acta A 100 (2013) 109–114.
[28] F.F. Gambarra-Neto, G. Marino, M.C.U. Araújo, R.K.H. Galvão, M.J.C. Pontes, E.P. de Medeiros, R.S. Lima, Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis, Talanta 77 (2009) 1660–1666.
[29] K. Poulli, G. Mousdis, C. Georgiou, Rapid synchronous fluorescence method for virgin olive oil adulteration assessment, Food Chem. 105 (2007) 369–375.
[30] K.I. Poulli, G.A. Mousdis, C.A. Georgiou, Classification of edible and lampante virgin olive oil based on synchronous fluorescence and total luminescence spectroscopy, Anal. Chim. Acta 542 (2005) 151–156.
[31] E. Sikorska, T. Górecki, I.V. Khmelinskii, M. Sikorski, J. Kozioł, Classification of edible oils using synchronous scanning fluorescence spectroscopy, Food Chem. 89 (2005) 217–225.
[32] M. Insausti, A.A. Gomes, F.V. Cruz, M.F. Pistonesi, M.C.U. Araujo, R.K.H. Galvão, C.F. Pereira, B.S.F. Band, Screening analysis of biodiesel feedstock using UV–vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm, Talanta 97 (2012) 579–583.
[33] K.D. Zissis, R.G. Brereton, S. Dunkerley, R.E.A. Escott, Two-way, unfolded three-way and three-mode partial least squares calibration of diode array HPLC chromatograms for the quantitation of low-level pharmaceutical impurities, Anal. Chim. Acta 384 (1999) 71–81.
[34] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173.
[35] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (2013) 3790–3798.

Please cite this article in press as: A.C.d. Silva, et al., Two-dimensional linear discriminant analysis for classification of three-way chemical data, Analytica Chimica Acta (2016), http://dx.doi.org/10.1016/j.aca.2016.08.009