Talanta 86 (2011) 316–323

Contents lists available at SciVerse ScienceDirect

Talanta

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

^{1}H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez*

Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007, Tarragona, Spain

ARTICLE INFO

Article history:
Received 16 June 2011
Received in revised form 5 September 2011
Accepted 12 September 2011
Available online 16 September 2011

Keywords:
Variable selection
Nuclear magnetic resonance
Food adulteration
Sudan dyes
PLS-DA

ABSTRACT

Whenever dealing with large amount of data as is the case of a NMR spectrum, carrying out a variable selection before applying a multivariate technique is necessary. This work applies various variable selection techniques to extract relevant information from ^{1}H NMR spectral data. Three approaches have been chosen, because each is based on very different foundations. The first method, called Xdiff, is based on calculating the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The second approach is the interval Partial Least Squares method (iPLS), which investigates the influential zones of the spectra that contains the most discriminating predictors calculating local PLS-DA models on narrow intervals. The last one is Genetic Algorithms (GAs) which finds the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined by the classification results obtained when multiclass Partial Least Squares-Discriminant Analysis is applied. This study has been applied to NMR spectra of culinary spices that might be adulterated with banned dyes such as Sudan dyes (I–IV). The three techniques give neither the same number nor the same selected variables, but they do select a common zone from the spectra containing the most discriminating variables. All three techniques give satisfactory classification and prediction results, being higher than 95% with iPLS and GA and around 89% with Xdiff, therefore the three variable selection techniques are suitable to be used with NMR data in the determination of food adulteration with Sudan dyes as well as the specific type of adulterant used (I–IV).

1. Introduction

In food industry, the addition of some colorants to some products is a common practice, as colorants enhance its visual aesthetics and promote sales. Up to now four Sudan (I–IV) dyes have been detected in certain food products, as culinary spices destined to human consumption, although they are normally used for coloring plastics and other synthetic materials. The current European legal framework on colors in food establishes that Sudan dyes are not included in the list of authorized colorants, as these dyes have potential carcinogenic effects [1] and even more, Sudan I may also have genotoxic effects. Therefore Sudan dyes are banned to be used as additive in food matrices for human consumption.

Previous studies have demonstrated that Proton Nuclear Magnetic Resonance Spectroscopy (^{1}H NMR) is a well suited analytical technique capable of detecting these four Sudan dyes when they are as adulterants in culinary spices [2]. NMR is a technique that generates a specific profile of the sample studied. Recent breakthroughs in NMR technology have led to measurements with increased sensitivity, resolution and reproducibility, thereby contributing to the production of high quality data [3]. These improvements in data quality, coupled with multivariate techniques, have given rise to a well-known rapid screening method [4] which has been demonstrated to be an efficient method for food screening, discrimination and characterization [2,5–7].

Because of the large amount of data obtained from a ^{1}H NMR spectrum, carrying out a variable selection before applying a multivariate technique is common practice. There are many potential benefits of variable selection: facilitating data visualization and data understanding, reducing the variables/samples ratio, eliminating noisy variables as well as redundant information, among others. All these advantages are important when chemometrics methods want to be applied. Many methods are found in the bibliography for variable selection in NMR in classification problems. Some methods include reducing noisy variables or variables of low intensity or even bucketing, with the consequent reduction of data. Other methods are supervised with the aim to find which variables are the most discriminatory in order to achieve the best discrimination when working with different groups of samples or classes.

* Corresponding author. Tel.: +34 977558490; fax: +34 977558446.
E-mail address: itziar.ruisanchez@urv.cat (I. Ruisánchez).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.09.019

Some examples of both mentioned types found in literature include stepwise discriminant analysis [8,9], supervised variable selection methods [10], self organizing maps (SOMs) [11], PLS weight coefficients [12], wavelet transform [13], univariate selection based on the maximum intensity differences [14], interval Partial Least Squares (iPLS) [15,16] and Genetic Algorithms (GAs) [17,18]. This wide variety of variable selection techniques implies that choosing the most appropriate one for a specific problem is not an easy task.

The aim of this work is to study the ability of three supervised techniques for selecting variables when working with NMR spectral data in the Sudan dyes classification problem, as it is well known that choosing the optimal variable selection technique is problem dependent. The three approaches studied have been chosen because each is based on a very different principle. Our work focuses on the region and number of selected variables, the number and type of misclassified samples and the overall performance of the classification process obtained with each technique.

The first technique is based on the computation of new variables called Xdiff, which are the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The hypothesis is that variables having different intensities will be reinforced in the new x_{diff} values, thus enabling differentiation between the classes. The second approach is the interval Partial Least Squares method (iPLS) [19], which investigates the influential zones of the spectra that contains the most discriminating predictors, and calculates local PLS-DA models in narrow intervals. The third variable selection technique used is the well known Genetic Algorithms (GAs) [20,21], which can find the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined from the classification results obtained when Partial Least Squares-Discriminant Analysis is applied.

2. Data analysis methods

2.1. Variable selection

2.1.1. Xdiff method

This variable selection method [2,14] is applied to a multiclass problem and it is based on calculating the x_{diff} values in accordance with Eq. (1)

$$x_{diff,ij} = \frac{\left|x_{ij} - \overline{x_i}\right|}{\sigma_i} \quad (1)$$

where x_{ij} is the ith variable for the jth sample and $\overline{x_i}$ and σ_i are the mean and standard deviation, respectively, calculated from each ith variable obtained from a reference class. As we are dealing with an adulteration problem, among our predefined five classes, we have set the unadulterated class as the reference one.

The **Xdiff** matrix is calculated for all five classes. A threshold value was defined from the x_{diff} values of the reference class through a visual inspection in a way that most of the x_{diff} values are kept below. Therefore, only those original variables that correspond to x_{diff} values higher than the prefixed threshold are selected. Several threshold values around the first prefixed one are checked, retaining the one which gives the best PLS-DA classification results.

2.1.2. Interval Partial Least Squares method

Interval PLS (iPLS) develops local PLS-DA models on equidistant subintervals of the full-spectrum region and the prediction performance of these local models and the global (full-spectrum) model is compared, mainly by means of the validation parameter RMSECV (root mean squared error of cross-validation, Eq. (2)):

$$\text{RMSECV} = \frac{\sqrt{(\sum \hat{y}_i - y_i)^2}}{n} \quad i = 1, \ldots, n \quad (2)$$

with y_i as the true class assignation value for sample i, $\hat{y}_i$ as the predicted class assignation value from cross-validation and n as the number of samples.iPLS provides an overall picture of the relevant information in different spectral subdivisions, thereby removing non-relevant information from other regions.

2.1.3. Genetic Algorithms (GAs) method

The GA theory is explained in detail elsewhere [20,21], so we will limit ourselves to show the procedure followed in the present study depicted in Fig. 1 which involves an iterative process. As for running the GA algorithm, the adequate number of input variables has to be not far away from 200, the original NMR variables have to be reduced, so the average of "n" consecutive variables is obtained. To decide the optimal window size n, the Principal Components Analysis (PCA) score plots of both, the original and the mean of n variables, are compared until a similar distribution of samples is kept. The next step is to apply the GA algorithm to the mean variables, with a previous step in which a randomization test is applied to check whether the dataset is adequate to run the algorithm, in order to avoid the overfitting problem commonly found in a GA-based feature selection [22]. Therefore, the mean variables selected by GA are expanded to the n consecutive original ones used to obtain the mean variables. Finally, PLS-DA is applied to those original ones.

All this procedure is done iteratively until a subset of variables giving the optimal classification results comparing the previous and last iteration is found. It has to be remarked that the second (and so on) iterations start with the previous subset of "expanded original variables", so a lower n value is looked for.

2.2. Classification method: Partial Least Squares-Discriminant Analysis

PLS-DA is a regression technique adapted to a supervised classification task [23]. A PLS regression model is calculated, which relates the independent variables (e.g. spectra) to a binary "y" vector which has as many values as classes in order to designate the class of the sample. For example, a vector [1,0,0,0,0] means that of five possible classes, the sample belongs to class 1, and so on. Classification of an unknown sample is derived from the value predicted by the PLS model, $\hat{y}$. Ideally, this value should be close to the values used to codify the class (here either 0 or 1). A threshold value for each pre-defined class is defined between 0 and 1 so that a sample is assigned to the class for which its prediction is larger than the threshold value. Typically, a normal distribution fits $\hat{y}$ values and the threshold value is estimated using Bayes' rule [24]. The optimal number of latent variables (LVs) was chosen to minimize the root mean square-cross validation prediction error (RMSECV) for all the classes. This number is selected through a compromise between the optimal values for each class.

Both, class assignment and the percentage of predicted probability are considered for the evaluation of the multiclass PLS-DA classification results. In this study, the selected variables are autoscaled before running the multiclass PLS-DA algorithm.

2.3. Training and test set

In order to avoid overoptimistic results, the dataset is divided into training and test set, using the training set to select the most

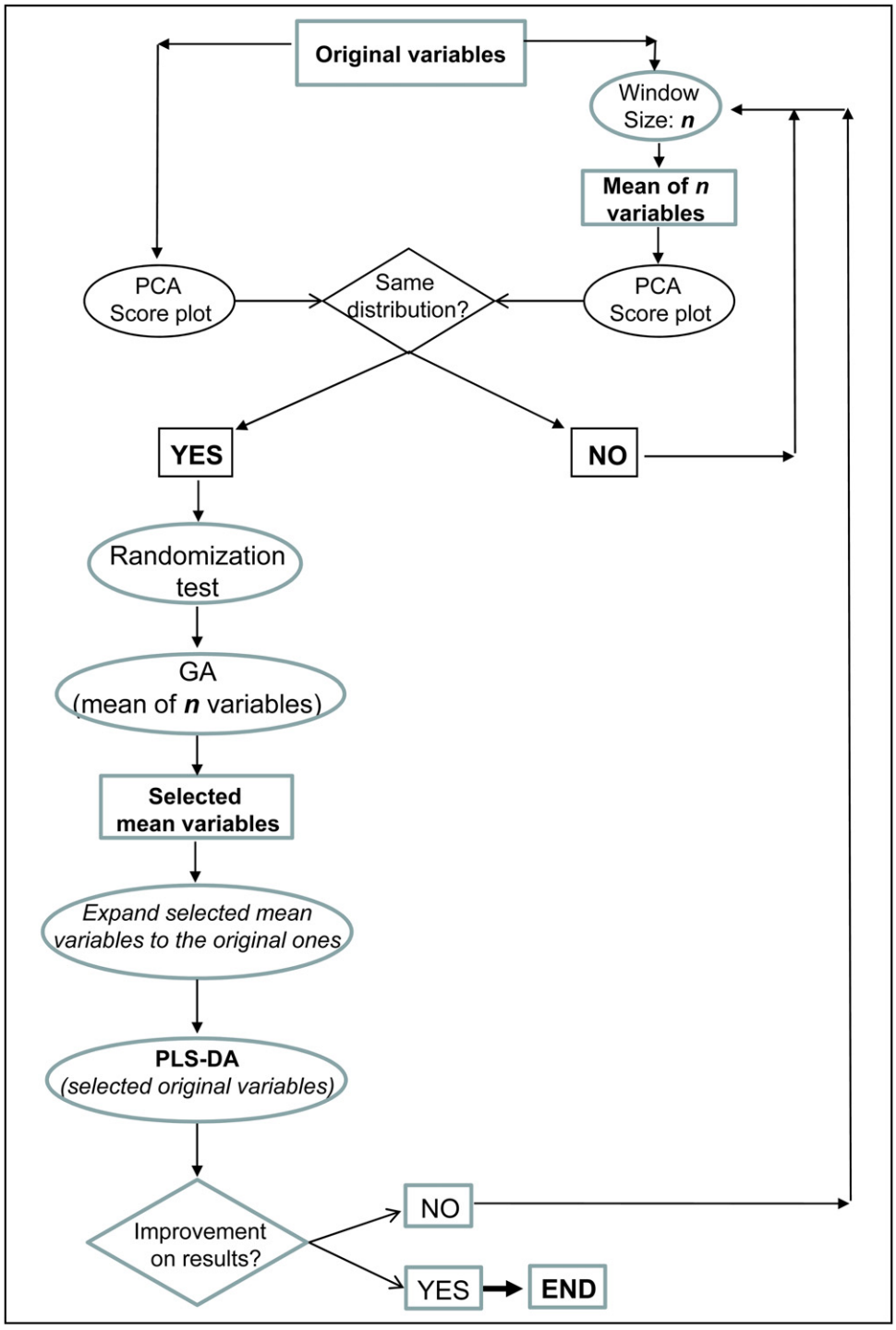

Fig. 1. Genetic Algorithm scheme for the iterative variable selection process.

relevant variables [25]. In our case, approximately 15% of samples of each class are left out to form the test set using two different strategies: a selection based on a randomly choice and on Principal Component Analysis (PCA) scores plots. In the last case, the representative samples are selected in a way to cover each class spatial distribution. Both strategies give similar global classification results, so we decided to use PCA because it selects the samples in a more homogeneous manner. Finally, the training set and test set are used to validate the PLS-DA models.

3. Experimental and data

3.1. Samples and NMR data

The studied samples correspond to twenty seven unadulterated commercial spices previously checked by HPLC-DAD they are free of any Sudan dye [26]. The samples are prepared weighing 0.1 g of each spice, dissolving it in 5 mL of deuterated chloroform and filtered. From this solution, 700 μL were taken and placed in 2 mL

flasks. Spiked samples were prepared by adding the stock Sudan (I–IV) solutions to each commercial sample to obtain a concentration within a range in which adulterated spices are commonly found: 7.1 g kg^{-1} [27]. So, five groups of twenty seven samples each give a total of 135 samples, where class 1 corresponds to the unadulterated samples and classes 2–5 correspond to the adulterated samples with Sudan I–IV, respectively.

The ^{1}H NMR spectra was acquired in Varian NMR System 400 at 400.13 MHz using a 4 μs pulse (45°), an acquisition time of 2.2 s (32,768 complex points) with a 15 s delay time to allow full relaxation and a spectral width of 7217 Hz (18 ppm). Sixteen scans were recorded per sample. Spectra were processed using Mestrec-C version 4.7.0 software. All the free induction decays (FID) were Fourier transformed (FT) by applying, first, a 32 k zero filling and, then, an exponential filter function with line broadening (LB) of 0.5 Hz. The spectra were automatically phase corrected and the baseline correction was made by manual multipoint with 14 points interpolated by a cubic spline function. All spectra were calibrated by setting the $CDCl_3$ peak at 7.26 ppm. The spectral region between 0.5 and 8.9 ppm is selected, because it is the zone where most of relevant signals are located, although the entire spectrum was used to make baseline corrections. The most intense range signal coming from the solvent was eliminated, leading to 8391 final variables. These variables constitute the raw spectra used for the different variable selection processes.

3.2. Software

All algorithms were run in the Matlab 6.5 (The MathWorks, Natick, MA) computing environment and PLS Toolbox 3.5 (Eigenvector Research Incorporated). The Matlab packages for the iPLS Toolbox as well as the PLS-GA Toolbox are freely available [28].

4. Results and discussion

As an initial trial, PLS-DA was applied without selecting any variables, being the recognition and prediction ability around 50%. These results suggest that a variable selection is quite necessary. Therefore as a preliminary study, the aromatic zone (from 6.6 to 8.9 ppm) was selected just based on a visual inspection of the spectra, as Sudan dyes have most of NMR signals there, while samples without Sudan dyes do not have any relevant signals. The percentage of correct classification was 87.8% and 85% for the training and test set, respectively. These preliminary results, although they are quite satisfactory, might be improved if a systematic variable selection approach is implemented.

Regarding the Xdiff approach, as a result of applying Eq. (1), the x_{diff} maximum values are around 40 while the maximum x_{diff} value obtained for the non-contaminated samples is 7 times lower. To select the optimal threshold value, several values were checked on the basis of the minimum classification error obtained. The optimal value of 5.5 is resulted in 1059 selected variables. These selected variables are in accordance with previous studies [2].

For the iPLS variable selection, the raw spectra are first divided into 20 equally sized subintervals containing 420 variables each, and the PLS-DA models were applied to each one. Fig. 2a shows the RMSECV results for each NMR interval, and for the global model (shown by the dotted line). At that point, the strategy is to select the intervals with an RMSECV value below the dotted line (intervals 2–4, 7 and 15) and to make a second iPLS selection. A further division into 10 equally sized subintervals is made (Fig. 2b) and intervals 4 and 5 are those that are located under the RMSECV global value. The classification results considering the 420 variables included in those intervals improve the previous classification results so they are kept as the final selected variables.

Table 1
Proton chemical shifts for the first variables selected with Xdiff, iPLS and GA. The correspondence with the Sudan dyes signal is also indicated.

Chemical shift (ppm)	Sudan dyes presence	Xdiff	iPLS	GA
8.086	III			x
7.970	I–II–III–IV	x		
7.992	I–II–III–IV	x		
7.755	I–II–III–IV		x	
7.738	I–II–III–IV		x	x
7.715	I–II–III–IV		x	
7.596	I–II–III–IV		x	
7.563	I–II–III–IV	x		
7.519	I–II–III–IV		x	x
7.505	I–II–III–IV		x	
7.484	I–II–III–IV		x	
7.466	I–II–III–IV		x	
7.209	I–II–III–IV			x
6.995	I–II–III–IV			x
6.938	I–II–III–IV	x		
6.915	I–II–III–IV	x		
6.885	I–II–III–IV			x
5.989	–			x
4.270	–			x
2.534	II–IV	x		
2.385	II–IV	x		
2.380	II–IV	x		

Among the different GA structures that can be used, we selected the GA parameters that have been successfully applied to spectral datasets [29]. The optimized GA parameters are those as follows: population size, 30 chromosomes; probability of mutation, 1%; probability of cross-over, 50%; cross-validation, 5 deletion groups; number of runs, 100; maximum number of features selected in the same chromosome, 30; average number of features per chromosome in the original population, 5.

As mentioned above, Fig. 1 shows the scheme for the GA iterative feature selection process used in this study. The first window size "n" is fixed to a value of 25. This value is chosen looking at the PCA scores plots (not shown) of the original variables and the mean centered variables in order to assure that a similar distribution of samples is maintained. With $n = 25$, 336 mean variables are obtained which corresponds to 8391 divided by 25. Then GA selects 96 mean variables that when there are expanded to the original ones, correspond to 2400 (96×25) original variables. In the second iteration, the maximum value for n which gives comparable PCA scores distribution was 5. Then GA is applied over 480 mean variables selecting 141, which when are expanded correspond a final number of 705 original variables (141×5). Further iteration trials do not improve the PLS-DA classification results.

As the final number of selected variables is not fixed in advance, each technique selects a different number: 1059, 705 and 420 for Xdiff, GA and iPLS, respectively. In order to compare the three variable selection methods, Fig. 3 shows the original and selected variables obtained with the three selection methods studied for a random adulterated sample. Also, both the spectrum of the pure dye contained in that sample and the unadulterated original sample are shown. Table 1 presents the variables which are most discriminant in each PLS-DA model by evaluating the loading weights of the two first PLS components. The first column presents the NMR chemical shift for the mentioned variables, the second column shows the correspondence with the Sudan dye signals and the last three columns contain the first eight selected variables by each method.

As it can be seen from Fig. 3, the spectrum of the pure Sudan II dye has most of its signals in the aromatic region between 6.8 and 9 ppm (Fig. 3b) and some in the aliphatic zone between 1 and 3 ppm. The spectra of both the unadulterated and adulterated

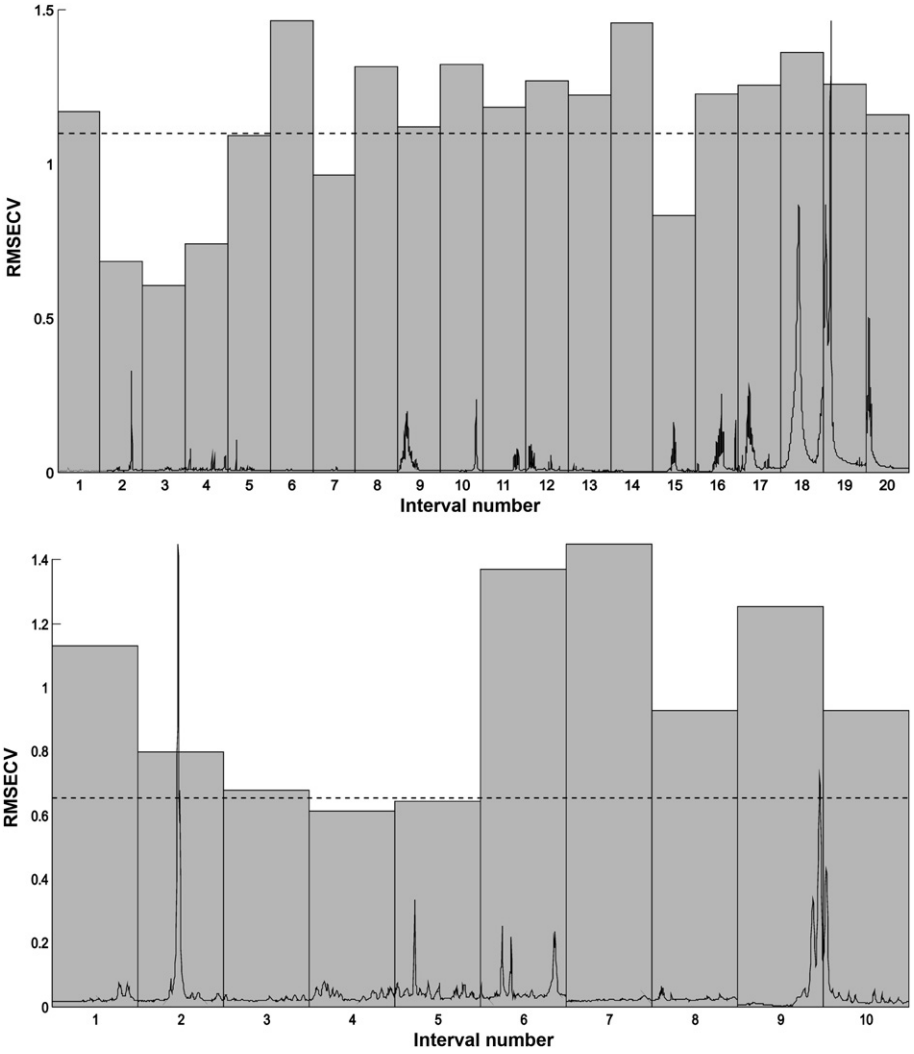

Fig. 2. iPLS plots of the cross-validated classification performance (RMSECV) (a) with the first 20 intervals and (b) with the last ten intervals obtained from the previous selected ones. In both plots, the dash line corresponds to the RMSECV value for the global model.

sample (Fig. 3a and c, respectively) have in addition, some signals between 4 and 5.5 ppm which are not present in the spectrum of the pure dye. The common selected variables obtained with the three different methods (Fig. 3d–f) are placed in the aromatic zone between 6.8 and 9 ppm, which is the spectral zone where the Sudan II dye as well as Sudan I, III and IV dyes (not shown) present most of their signals. It must be emphasized that although the spectral region is the same, the variables are not exactly the same. Xdiff (Fig. 3d) looks for individual variables that may or may not be consecutive; iPLS (Fig. 3e) looks for intervals of variables, such that there are consecutives within each interval, and finally GA (Fig. 3f) looks for consecutive variables within a short interval of variables, those resulting from expanding the selected variables in the last iteration. Evidence of this is a signal located at 8.086 ppm (behind the most intense peak found in this zone) which is only selected by GA. Moreover, GA and Xdiff also select variables located at the aliphatic zone.

Table 1 gives additional information which resumes the above discussion considering the first eight selected variables. It can be seen that the aromatic zone is where most variables are selected by the three methods. Continuing with Table 1, the last three variables can be related to the NMR signals of the methyl groups which are only present in the chemical structure of the Sudan I and IV dyes [26]. This information related to the methyl groups is only considered by the first eight relevant variables selected by Xdiff although, as it can be seen in Fig. 3f, GA also selects some variables in the same zone that have lower relevance in the PLS loading weights.

Concerning the classification results, Table 2 shows the misclassified samples obtained by the three PLS-DA models built using the variables selected from each technique (Xdiff, iPLS and GA). The errors are discussed considering samples wrongly assigned (depicted in bold), samples not assigned to any class (called none) and samples assigned to more than one class (showing the two classes where these samples are assigned to). For the sake of clarity, when the above mentioned samples are correctly assigned with some of the models, the results are not shown (empty spaces).

Focusing at the training set, the first point concerns sample 47 which is the only wrong assigned sample using the three variable selection approaches. This sample belongs to class 3 and the percentage of predicted probability is 100% to class 2. So, in our opinion, this misclassification might be due to an experimental error and it can be considered as an outlier. A particular case is sample 98 that belongs to class 5 and it is wrongly assigned to class 1 with almost 99% of probability when using either GA or iPLS. However,

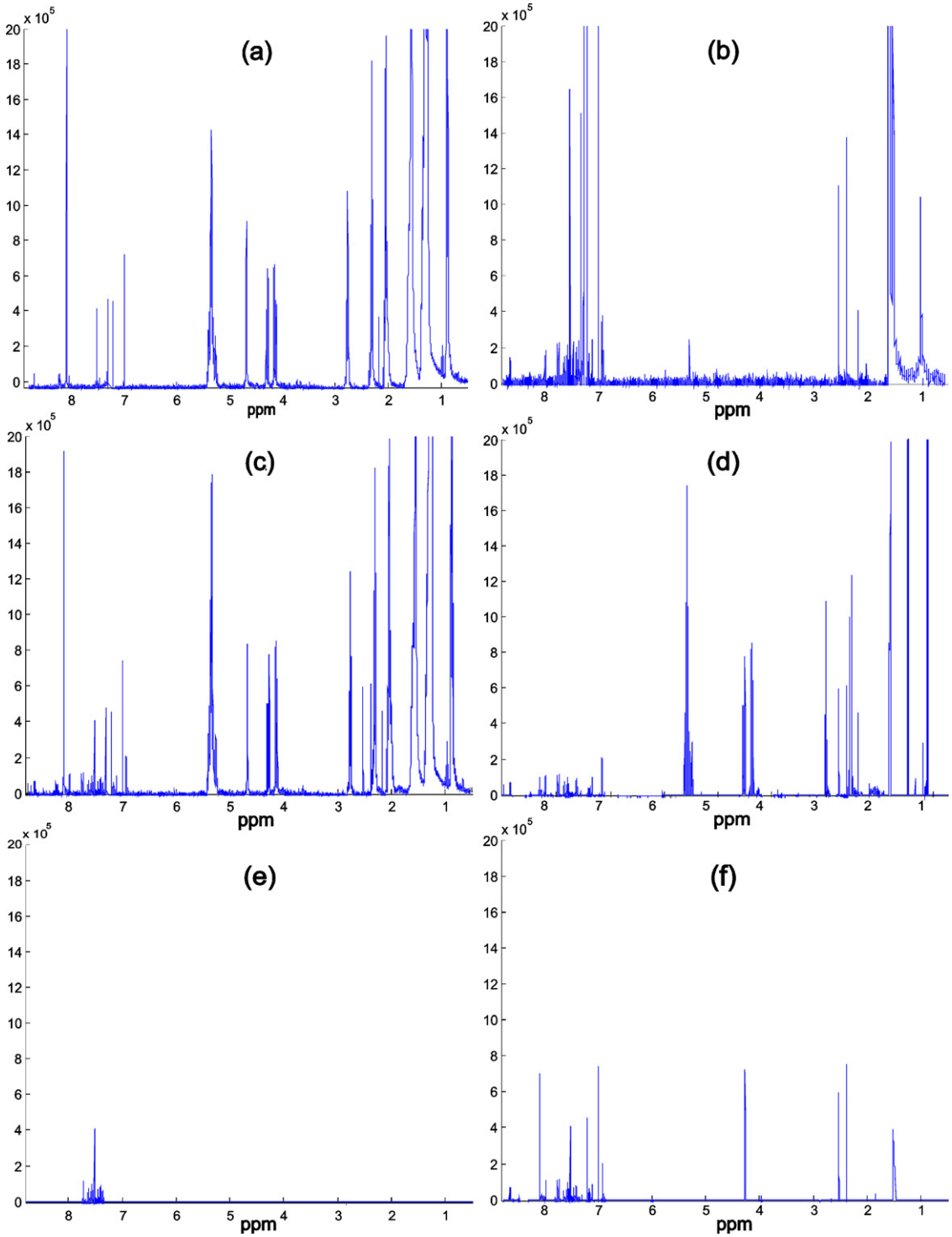

Fig. 3. Spectrum of the (a) original variables of an unadulterated random sample, (b) pure Sudan II dye, (c) original variables of the sample adulterated with Sudan II and (d) variables selected with Xdiff, (e) iPLS and (f) GA methods.

it is correctly assigned only when using Xdiff. In our opinion this fact might be an indication that, in this case, some of the variables only selected by the Xdiff technique are crucial in obtaining the right classification.

There are some samples which are not assigned to any class (called none) either with Xdiff and/or GA. In some ways we think that it is better not to assign a sample to any class than to make a wrong assignation, particularly when dealing with foodstuff adulteration problems. It has to be stated that these samples are correctly assigned when using the iPLS technique.

The rest of samples shown in Table 2 correspond to samples assigned to more than one class, which have high predicted probability to belong to both of them and being one of the two classes its true class. Most of these samples are given when Xdiff technique

Table 2
PLS-DA class assignation errors considering the variables selected by three techniques (Xdiff, iPLS and GA) for the training and test set. Wrongly assigned samples are depicted in bold, samples not assigned to any class are called as none and samples assigned to more than one class are presented (showing the two classes where these samples are assigned to). The sample number and its true class are indicated in the first two columns.

Sample	True class	Class assignment selecting variables with		
		Xdiff	iPLS	GA
Training set				
1	1	1,5		
2	1			1,5
4	1			1,5
6	1			1,5
36	2	2,5		
46	2		2,5	
47	**3**	**2**	**2**	**2**
80	4	4,5		
81	4	4,5		
82	4	4,5		
96	5	None		None
98	**5**		**1**	**1**
100	5	None		
101	5	5,4		
108	5	5,4		
113	5	5,4		
Test set				
128	4		4,5	
132	5	5,1		
133	5	None		

is used, mainly for samples between class 4 (spices spiked with Sudan III) and class 5 (spices spiked with Sudan IV). On the other hand, the assignation results for the test set (Table 2) are similar to those presented above regarding the training set.

If the type of error and its implications are considered, some comments can be made: there are some unadulterated samples with also high probability to be assigned as contaminated with Sudan dyes, which implies that these samples must be discarded until their real status is confirmed. But the most important fact is when the consumer health is under risk, i.e., samples contaminated with Sudan dyes which are classified as unadulterated ones. In our particular case, only one sample was obtained under this condition (sample 98). Finally, the general trend regarding the samples assigned to more than one class, is that there are adulterated samples with also high probability to belong to another adulterated class (classes 4 and 5), which has neither economical nor healthy implications as in any case they will be withdrawn from the markets.

The final overall recognition abilities regarding the training set are 87.8%, 95.7% and 99.1% for Xdiff, GA and iPLS, respectively. Considering the prediction abilities in the test set, a 90%, 100% and 95% is achieved by Xdiff, GA and iPLS, respectively. These results show that the use of an appropriate variable selection technique really improves the performance classification when dealing with NMR data.

5. Conclusions

We demonstrate that in a classification problem, when dealing with hundreds or thousands of variables as is the case of NMR data, the application of a variable selection approach is almost mandatory. The selection of variables, not only makes a considerable data reduction, but also eliminates noisy areas from the spectra or areas that do not contain relevant information in a way to achieve an optimal classification performance.

Although a visual inspection of the spectra in some cases, might allow selecting characteristics regions, it has been demonstrated that the application of variable selection techniques improves the classification results. This study focuses on three different approaches and the following conclusions can be mentioned each one:

Xdiff is easy to implement, and has a clear and simple structure which allows individual variables to be selected. In this particular case, it is the only selection technique that selects variables across the entire spectra. An obvious example of this can be seen in the classification of sample 98, as it is only assigned correctly by this technique.

On the other hand, iPLS is conceptually and mathematically easy to implement and it is very effective at selecting the interesting parts of the spectrum. However, as it selects zones from across the entire spectra, it might also incorporate noisy variables, in such a way it cannot select punctual variables.

Finally, GA may also be considered a good variable selection technique as this has been demonstrated. Several GA-based features selection methods using different GA structures have been developed, making this technique a flexible way to be applied in a variety of analytical problems. Nevertheless, it is conceptually more complex than the other two techniques studied and a larger number of parameters have to be considered as inputs to the algorithm.

As a final conclusion, based on the quite high correct classification results, the three variable selection techniques are very powerful tools in the determination of either food adulteration as well as the type of adulterant used when dealing with NMR data.

In our particular case study, iPLS and GA give better classification and prediction results. It has to be emphasized that most of the predictions errors are between two adulterated classes and there is only one case of an adulterated sample being assigned unadulterated by iPLS and GA but not with Xdiff, which is of great importance whenever dealing with a food safety problem.

Acknowledgements

The authors would like to thank the Management Agency for University and Investigation Support of the Catalan Government (AGAUR) for providing Carolina Di Anibal a doctoral fellowship and the Spanish Ministry of Education, Culture and Sports (Project CTQ2007-311 61474/BQU) for economic support.

References

[1] IARC: International Agency for Research on Cancer, IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man: Some Aromatic Azo Compounds, vol. 8, Lyon, 1975, pp. 224–231.
[2] C.V. Di Anibal, I. Ruisanchez, M.P. Callao, Food Chem. 124 (2011) 1139–1145.
[3] R.A. Davis, A.J. Charlton, S. Oehlschlager, J.C. Wilson, Chemom. Intell. Lab. Syst. 81 (2006) 50–59.
[4] J.W.E. Vogels, L. Terwel, A.C. Tas, F. van den Berg, F. Dukel, J. van der Greef, J. Agric. Food Chem. 44 (1996) 175–180.
[5] D.W. Lachenmeier, E. Humpfer, F. Fang, B. Schütz, P. Dvortsak, C. Sproll, M. Spraul, J. Agric. Food Chem. 57 (2009) 7194–7199.
[6] R. Consonni, L.R. Cagliani, M. Stocchero, S. Porretta, J. Agric. Food Chem. 57 (2009) 4506–4513.
[7] B. Biais, J.W. Allwood, C. Deborde, Y. Xu, M. Maucourt, B. Beauvoit, W.B. Dunn, D. Jacob, R. Goodacre, D. Rolin, A. Moing, Anal. Chem. 81 (2009) 2884–2894.
[8] S. Rezzi, I. Giani, K. Héberger, D.E. Axelson, V.M. Moretti, F. Reniero, C. Guillou, J. Agric. Food Chem. 55 (2007) 9963–9968.
[9] S. Rezzi, D.E. Axelson, K. Héberger, F. Reniero, C. Mariani, C. Guillou, Anal. Chim. Acta 552 (2005) 13–24.
[10] M. Cuny, E. Vigneau, G. Le Gall, I. Colquhoun, M. Lees, D.N. Rutledge, Anal. Bioanal. Chem. 390 (2008) 419–427.
[11] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, Chemom. Intell. Lab. Syst. 98 (2009) 149–161.
[12] A. Jankevics, E. Liepinsh, E. Liepinsh, R. Vilskersts, S. Grinberga, O. Pugovics, M. Dambrova, Chemom. Intell. Lab. Syst. 97 (2009) 11–17.
[13] S.B. Kim, Z. Wang, S. Oraintara, C. Temiyasathit, Y. Wongsawat, Chemom. Intell. Lab. Syst. 90 (2008) 161–168.
[14] A.J. Charlton, P. Robb, J.A. Donarski, J. Godward, Anal. Chim. Acta 618 (2008) 196–203.
[15] H. Winning, E. Roldán-Marín, L.O. Dragsted, N. Viereck, M. Poulsen, C. Sánchez-Moreno, M.P. Cano, S.B. Engelsen, Analyst 134 (2009) 2344–2351.

[16] H. Winning, N. Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen, Food Hydrocolloids 21 (2007) 256–266.
[17] H.W. Cho, S.B. Kim, M.K. Jeong, Y. Park, T.R. Ziegler, D.P. Jones, Expert Syst. Appl. 35 (2008) 967–975.
[18] M. Wasim, R.G. Brereton, Chemom. Intell. Lab. Syst. 81 (2006) 209–217.
[19] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[20] D.B. Hibbert, Chemom. Intell. Lab. Syst. 19 (1993) 277–293.
[21] R. Leardi, J. Chemom. 15 (2001) 559–569.
[22] R. Leardi, A. Gonzalez Lupíañez, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[23] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.
[24] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, J. Chemom. 20 (2006) 341–351.
[25] R.G. Brereton, Trends Anal. Chem. 25 (2006) 1103–1111.
[26] C.V. Di Anibal, M. Odena, I. Ruisanchez, M.P. Callao, Talanta 79 (2009) 887–892.
[27] American Spice Trade Association, http://www.astaspice.org/files/public/SudanWhitePaper.pdf, 2005 (accessed 20.09.10).
[28] http://www.models.kvl.dk/source/.
[29] R. Leardi, J. Chemom. 14 (2000) 643–655.