FULL PAPER
THE ANALYST
www.rsc.org/analyst

Classification of commercial apple beverages using a minimum set of mid-IR wavenumbers selected by Procrustes rotation

J. M. Andrade,*[a] M. P. Gómez-Carracedo,[a] E. Fernández,[a] A. Elbergali,[b] M. Kubista[b] and D. Prada[a]

[a] *Dept. Analytical Chemistry, University of A Coruña, Campus da Zapateira, s/n, E-15071 A Coruña, Galicia, Spain*

[b] *Dept. Chemistry and Biosciences, Chalmers University of Technology, Medicinaregatan 7B, S-40530 Göteborg, Sweden*

Received 8th May 2003, Accepted 24th July 2003
First published as an Advance Article on the web 5th August 2003

When infrared spectral data are used in classification and/or multivariate regression methods there can be problems related to both chemical understanding and computation speed due to the large number of wavenumbers in each spectrum. Here, it is shown that the Procrustes rotation technique can be used to select a minimum set of spectral variables (wavenumbers) to perform classification and regression. Procrustes rotation was coupled to several multivariate methods as PLS, SIMCA and potential curves (a maximum likelihood classification method). The practical problem of implementing a screening methodology for classifying apple juice-based beverages according to their contents of "pure" apple juice was addressed using attenuated total reflectance, mid-IR spectroscopy. It is found that two of the original wavenumbers are almost as good predictors as all the 176 initial ones.

Introduction

Mid infrared spectroscopy (mid-IR or FT-MIR) is an analytical technique employed worldwide in industrial quality control. Besides its "classical" qualitative purposes, its application to quantitative studies rose during the last decade, propelled by its straightforward and powerful linkage to chemometric (multivariate) methods which opened new quantitative alternatives. Here, spectra need to be digitized and typical digitalizations consider 1, 0.5, 0.25, 0.125 or 0.06925 data points per cm, resulting in large data matrices, with drawbacks such as slow computation speed, data storage capabilities and data update. Despite the good results obtained in most cases when considering complete spectra, a relevant question is: are all spectral variables (wavenumbers) really required to make a satisfactory classification or some quantitation? Or, are the spectral variables correlated and can their number be reduced? Would a reduction of the number of spectral variables perhaps simplify the chemical understanding of the problem? Moreover, variable selection in spectral analysis constitutes an important issue to develop simpler, portable equipment for routine/field analysis in quality control applications (environment, food, beverages, biosensors, *etc.*).[1] Variable selection can also be useful to find those variables in a large data set that have best discrimination power.

Several multivariate techniques have been applied for variable reduction. These include factor analysis[2] and discrimination methods,[3] which have the disadvantage that although the information is condensed into a few abstract factors, all the initial variables are still implied. Another approach for variable selection (in the sense of eliminating useless variables, not to extract an small subset) was proposed, consisting of making a preliminary regression with all variables, adding artificial random noise and, then, selecting the experimental variables that show more importance than the artificial ones (according to a criterion based on the regression coefficients).[4] Garrido-Frenich *et al.*[5] performed PLS (partial least squares) studies where a threshold level was calculated for the loadings and the variables with highest coefficients were selected. For comparison they also considered the variables that gave highest correlation with the concentrations of interest. Goicoechea and Olivieri[6] employed the error indicator derived from hybrid linear analysis regression (HLA, described by Berger *et al.*[7]), to search for the best linear fit among different spectral ranges (*i.e.*, ranges of successive variables).

Elbergali *et al.*[8] defined "resolvability indices" for 2-dimensional matrices, using a variant of the evolving factor analysis to select individual variables to resolve overlapping chromatographic peaks. Ferré and Rius[9] developed a graphical approach to select the optimum set of variables for calibration. It is based on considering the confidence region of the estimated concentrations and on optimizing five different criteria. The optimum set is found by testing all possible combinations of variables. Although quite successful, the approach may be time-consuming when analyzing spectral data sets. Heise and Bittner[1] proposed a pair-wise selection procedure of spectral variables starting from the weights of the optimum PLS-regression vector. Pairs of variables related to neighbouring minimum and maximum regression vector coefficients were selected. They studied the PLS performance with two, four, six, *etc.* wavenumbers as variables. Each additional pair introduced into the model was given lower weights than the previous pair.

Today, efforts are being made on variable selection by intensive computation and non-parametric methods using artificial neural networks and genetic algorithms. Todeschini *et al.*[10] employed Kohonen artificial neural networks to select sets of wavelengths for PLS calibration of mixtures of cresol isomers by fluorescence spectrometry. The use of genetic algorithms for variable selection prior to multivariate regression was reviewed by Leardi.[11] An interesting conclusion reached was that many workers used different genetic structures, suggesting it must be necessary to modify the algorithm for each particular problem. As an example, Smith and Gemperline[12] employed a "modified" genetic algorithm, where some genes were preserved from crossover and/or mutations as they were used to assess information either from the data themselves or from the classification model developed. Recently, Guo *et al.*[13] compared the performance of several methods, including a

DOI: 10.1039/b305185d

variant of Procrustes rotation coupled to genetic algorithms, to select variables. They adapted the consensus concept from the Generalized Procrustes Analyses to match the subspaces so that the consensus configuration was searched by the genetic algorithms.

A novel nonparametric approach for variable selection was recently developed by Heberger and Rajko[14,15] using a pairwise correlation method based on four criteria: both variables should enhance correlation, one should enhance and the other reduce correlation and *vice versa*, and finally both should reduce the correlation between a dependent variable and two independent parameters. Arranging the frequencies of the four basic events in a contingency table, significant differences can be determined by several nonparametric tests. This idea was generalized to several variables and yielded promising results.

In our opinion, these methods, with the exception of Guo's, have serious drawbacks. They do not identify a *minimum* set of original variables to analyze the problem. Some use trial and error strategies that are computationally inefficient and do not guarantee the best solutions. Some are also problem-specific.

In this work, the Procrustes rotation algorithm is applied to mid-IR spectroscopic data to select a minimum set of wavenumbers that account for the main information in the total data set. To the best of our knowledge, this is the first application of Procrustes rotation to such a problem. The specific problem addressed was to develop a fast procedure that is suitable for screening to estimate the amount of "real" apple juice in commercial apple juice-based beverages (soft drinks, "energetic"-soft drinks, pure apple juices, *etc.*). The overall screening was divided into a qualitative assessment of the mid-IR attenuated total reflectance (ATR) spectral profile of a commercial product and its classification into a set of predefined classes with different percentages of pure apple juice. Here, only the second step is considered. The results obtained using variables selected by Procrustes rotation are compared to those obtained with multivariate regression approaches using either complete spectra or a reduced number of wavenumbers selected by conventional approaches.

Variable selection by Procrustes rotation

Procrustes rotation is a general term for the mathematical technique to match two data sets, each one considering different variables for the same objects (samples).[16] This objective is achieved by minimizing the sum of squared differences between the two spaces after rotation, translation and stretching of one of the sets relative to the other. The approach has been applied successfully to different problems, including environmental,[17–19] quality control of aviation fuels[20,21] and spectral analysis and analyte identification.[22,23] Procrustes rotation-based techniques can also be combined with thermodynamic constraints to predict spectral profiles of imbedded species in chemical equilibrium studies, such as monomer–dimer equilibria.[24] The technique is unique in the sense that relative quantitation is possible even without the use of "standards" in a classical regression sense.[25,26] Procrustes algorithms have also been employed by González-Arjona *et al.*,[27] as a computing base for target factor analysis.

There are two main applications of Procrustes rotation: comparing subspaces to correlate 2-dimensional data sets[22–26,28] and selecting representative variables. The goal of the latter is to find a subset of original variables that conveys the main structure of the data. The strategy is to identify the variable that contributes with least information to the system and delete it. The process can then be repeated to delete additional variables until a small number remains, that containing all essential information. This approach is related to other multivariate techniques, such as rank annihilation analysis.[29] The original Procrustes rotation is based on singular value decomposition, svd, of data sets with more samples than variables although the case of rank-deficient analysis has already been studied.[28,30,31] A brief summary of the approach is presented below.

Let $\boldsymbol{X}$ be the $(n \times p)$ original data matrix; if $n \geq p$, then the data scores are $\boldsymbol{US}$ and the variable loadings are $\boldsymbol{W}^{\mathrm{T}}$ ($^{\mathrm{T}}$ denotes transposed matrix). By svd the equation $\boldsymbol{X}_{(n \times p)} = \boldsymbol{U}_{(n \times q)}\boldsymbol{S}_{(q \times q)}\boldsymbol{W}^{\mathrm{T}}{}_{(q \times p)}$ is solved for $\boldsymbol{US}$ and $\boldsymbol{W}^{\mathrm{T}}$. If $p > n$, a new matrix $\boldsymbol{B}$ can be such that $\boldsymbol{B} = \boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{B}_{(p \times n)} = \hat{\boldsymbol{U}}_{(p \times q)}\hat{\boldsymbol{S}}_{(\mathrm{q} \times \mathrm{q})}\hat{\boldsymbol{W}}^{\mathrm{T}}{}_{(q \times n)}$ is solved by svd. If the original data matrix is needed, for example for comparison, we only need to compute $\boldsymbol{X} = \boldsymbol{B}^{\mathrm{T}} = \hat{\boldsymbol{W}}\hat{\boldsymbol{S}}\hat{\boldsymbol{U}}^{\mathrm{T}}$, where $\hat{\boldsymbol{W}}_{(n \times q)}\hat{\boldsymbol{S}}_{(q \times q)}$ are the scores and $\hat{\mathbf{U}}^{\mathrm{T}}{}_{(q \times p)}$ are the loadings.

Procrustes rotation determines the sum of squared differences between corresponding points of two (or more) spatial configurations after they have been aligned through translation, rotation and reflection.[16] The subspace with all variables is fixed and the other subspaces with one or more variables deleted are "Procrustes-matched". It is more convenient to compare score subspaces instead of whole data subspaces. For this we must determine the significant dimensionality of the data set. This can be known beforehand or it can be estimated by statistical tests, the Wm one being strongly recommended.[16] A problem is an objective definition for "significant", and there will always be a degree of subjectivity in selecting the number of PCs that are sufficient to account for all important features in the data. Hereinafter, let q denote the significant dimensionality of the data set.

Let $\boldsymbol{T}_{(n \times q)}$ be the "true" target score-subspace, and let $\boldsymbol{R}_{(n \times q)}$ be the score-subspace calculated using the reduced $\boldsymbol{\chi}_{(n \times q)}$ data set. Translation matching is achieved by mean-centering the data. The sum of squared differences, M^2, is a measure of the distance between the two spaces. It is obtained as $M^2 = \mathrm{Trace}\{\boldsymbol{TT}^{\mathrm{T}} + \boldsymbol{RR}^{\mathrm{T}} - 2\boldsymbol{RA}^{\mathrm{T}}\boldsymbol{T}^{\mathrm{T}}\}$ where $\boldsymbol{A}$ is the orthogonal matrix which defines the rotation to match $\boldsymbol{R}$ with $\boldsymbol{T}$, and it is obtained from the svd decomposition of $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{T} = \boldsymbol{CLZ}^{\mathrm{T}}$ as $\boldsymbol{A} = \boldsymbol{ZC}^{\mathrm{T}}$. It is not practical to test all combinations of variables and, therefore, the calculations are performed sequentially. In each cycle M^2 is calculated deleting one variable at a time and the one that gives the lowest M^2 is permanently deleted.

Although the svd approach works well, the algorithm can be slow, which may be a drawback when large data matrices are compared. An alternative (not applied here) to speed up the calculation is to consider the NIPALS algorithms. Consider the previous $\boldsymbol{X}$ and $\boldsymbol{B}$ matrices. NIPALS decomposition gives $\boldsymbol{X}_{(n \times p)} = \boldsymbol{T}_{(n \times q)}\boldsymbol{P}_{(q \times p)}$ and $\boldsymbol{B}_{(p \times n)} = \tilde{\boldsymbol{T}}_{(p \times q)}\tilde{\boldsymbol{P}}_{(q \times n)}$ from which the score matrices are obtained directly. The two approaches are equivalent since $\tilde{\boldsymbol{T}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{S}}$ and $\tilde{\boldsymbol{P}} = \hat{\boldsymbol{W}}^{\mathrm{T}}$.

Experimental

Samples

Samples with known contents of pure apple juice were prepared from Gloster, Golden, Granny Smith, Reineta, Royal Gala and Starking varieties of apples in the laboratory. Juice was squeezed out, centrifuged, filtered and immediately characterized by FTMIR-ATR. This showed that the juice spectrum did not depend on apple variety. Aliquots of pure apple juice were then diluted with Milli-Q water (18.1 MΩ.cm resistivity, Millipore, Barcelona, Spain) to produce 20 samples with 2% apple juice, 20 samples with 4%, 31 samples with 6%, 27 samples with 8%, 27 samples with 10%, 27 samples with 16%, 27 samples with 20%, 39 samples with 25%, 32 samples with 50%, 19 samples with 70% and 26 samples of pure apple juice. For practical purposes they can be termed "laboratory standards".

The European legislation distinguishes between soft drinks, that contain at least 10% juice, nectars with at least 50% juice and pure juices that should be 100% juice with no sugars added. Comparison of commercial beverages with the above laboratory standards is not possible because most commercial soft drinks and "energetic" beverages have added sugar. In Europe it is not necessary to declare the amount of sugar(s) added, which complicates the analysis. But this will change soon owing to new EU Directives.[32] To account for the presence of the three most important sugars of apple juice (glucose, fructose, sucrose)[33,34] in commercial beverages, an indirect strategy was followed. The total amounts of sugars can be quantified either in the beverages or in the laboratory standards by a matching stage where their spectra are compared against those of standard solutions of mixtures of the three sugars.

The laboratory standards were split into two groups with different amount of juice. Further, each group was divided into a training and a testing set. The first set contained 2–20% juice. It had a total of 173 standard samples, of which 134 were dedicated to train the model and 39 were used for validation. The second set contained 25%–100% juice. It had 130 standard samples of which 86 were used for training and 44 for validation. The model was then used to classify 23 commercial juice beverages available in Spain.

FTMIR-ATR spectra (1250–900 cm^{-1}, 50 scans per spectrum, background substracted, Beer–Norton strong apodization, 4 cm^{-1} nominal resolution) were measured using a 16PC-FTMIR instrument (PerkinElmer, Überlingen, Germany) and an ATR device (PerkinElmer, ZnSe crystal, 45° incidence angle, 12 nominal reflections). The equipment was daily as well as weekly assessed by routine quality control tests.[35] The spectra were baseline corrected, digitized (despite the resolution being 4 cm^{-1}, PerkinElmer's software gives one datum every 2 cm^{-1}) and exported to ASCII files which are input into the statistical software (Matlab, The Mathworks Inc, Natick, MA, v. 4.2c.1). All studies were made with mean centred data (using the means of the calibration sets).

Classification models

Three multivariate classification techniques (PLS, SIMCA and potential curves) were applied to the entire set of 176 wavenumbers and to the minimum subset selected by Procrustes rotation. If the same results were obtained with the entire set of variables and the reduced set it was concluded that the reduction did not lose essential information.

PLS has become an standard methodology[36,37] to regress an *X*-block (spectra) on an *Y* block (type of juice). The *Y* data were binary codes for each group of laboratory standards (class), *e.g.* 1000000 for class one (2% apple juice), 0100000 for class two (4% apple juice), *etc.* Accordingly, after developing a model, a new sample will be classified by its highest score in the PLS output. As real outputs are not pure 0s or 1s (*e.g.* 0 0 0.998 1.014 0.054 0 0), all the figures for each sample should be closely scrutinized in order to assess the final assignment (the sample might even not be included in any class).

SIMCA (soft independent modelling of class analogy) and simplified potential curves are pattern recognition techniques to classify new samples according to the probability that the sample belongs to each of a predefined set of classes. Although SIMCA is an established method[38,39] potential curves are still not and is therefore briefly outlined here. "Potential curves" is a maximum likelihood classification method and can be exemplified as follows.[40,41] Suppose that standard juices are distributed along the PC1–PC2 subspace (or PC1–PC3, *etc.*). Assume each group of juices (2%, 4%, *etc.*) is homogeneously distributed and does not overlap with any other. An iso-probability bivariate Gaussian region can be defined for each group considering its scores (see eqn. (1) and (2)).

$$f(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[\frac{-A}{2(1-\rho^2)}\right] \quad (1)$$

$$\boldsymbol{A} = \left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right) \quad (2)$$

Where *X* and *Y* are the PC1 and PC2 sample scores; μ_X and μ_Y, the average scores for each group and σ_X and σ_Y their respective standard deviations; ρ is the calculated correlation coefficient between the *X*s and the *Y*s. Intercepting this curve with horizontal parallel planes, elliptic sections are obtained which represent the iso-probability sections. The equation defining the iso-probability ellipses is: $a = A(1/(1-\rho^2))$, where *a* is a positive constant. The area of each ellipse can be related to the probability of a sample belonging to this group by means of the expression: Area = $\exp(-a/2)$ = Prob (sample $\in$ group). If new samples are to be classified, they must be projected on the PC1–PC2 subspace, the constants recalculated and, then, the samples will be classified according to the probability of belonging to each class. It is worth note that (i) PCs are used instead of original variables to avoid useless information and take advantage of the sample grouping in the reduced PC-subspace and (ii) eqn. (1) and (2) are a natural simplification of the equivalent determinant method from Forina *et al.*[40]

Results and discussion

Classification using all the spectral variables

Table 1 summarizes the main characteristics and classification results of the multivariate models for the 23 commercial samples. Fig. 1 shows the iso-probability potential region obtained for each class using the training samples. The PC1–PC2 subspace accounted for slightly more than 95% of the information. The contour of each isoprobability ellipse (on the PC1–PC2 plane) represents a probability level (*e.g.*, 85%, 90%, 95%, *etc.*). Despite the moderate overlap among several classes in the 2–20% range the final results are quite satisfactory considering the intrinsic simplicity of the algorithm. Moreover, when an erroneous classification was made either in the calibration or the validation step, the sample was almost always assigned to an adjacent group. This was in fact the case for all classification models.

The first conclusion that can be derived from Table 1 is that all commercial samples but one of the pure (100%) juices were classified according to manufacturers' declaration. The outlier sample had a quite different IR spectral profile, suggesting it contained inappropriate amount of sugar(s) (Fig. 2). By PLS and SIMCA it was classified as belonging to the 70% group, while potential curves put it in the 100% group. Nevertheless, this different classification is not crucial as the beverage was identified in the first step of the screening methodology as a suspicious beverage.

A second conclusion is that the 2%–20% range is more difficult to model than the 25%–100%, most likely because the former has more groups of juices, some of which partially overlap. Accordingly, the models lead to some wrong classifications. Of course, overlap could be reduced by defining broader groups (for example, by adding 2% + 4%, 8% + 10%, and 16% + 20% groups), but it was decided not to in order to better reflect real situations where it might be interesting to differentiate between similar juice contents.

SIMCA behaved very well on both the training and the commercial samples, making only a single error. However it performed worse on the validation samples. This was probably due to the overlap among the classes, which were not satisfactorily resolved by the local PC models of this technique.

The PLS model yielded more errors during calibration and validation than the others. But it performed well on the commercial samples. It required only two latent variables (LV). Noteworthy all errors corresponded to samples placed between two classes which were assigned to an adjacent group. As the assignments were not always definite, unclear classifications should be further assessed before decision making (*e.g.*, studying the spectrum of the sample or applying other classification methods).

Three quality parameters[40] were calculated for all the classification methods: (i) efficiency: the percentage of objects correctly classified, (ii) sensitivity: the percentage of objects belonging to the class in which they are classified, and (iii) specificity: the percentage of objects not belonging to a class which are correctly classified outside such class. Table 2 shows that PLS and potential curves performed almost the same whereas SIMCA produced less good results. Overlap reveals important among several groups of standards, namely, the 8% and the 10%, the 10% and the 16% and (to some degree), the 16% and the 20%.

Classification using a minimum set of spectral variables

Two PCs were considered to retain most of the information of the mean-centred spectra, as suggested by the eigenvalues and Wm statistic. Therefore, the minimum number of variables (*i.e.*, original wavenumbers) that can be selected by Procrustes

Table 1 Multivariate models and results using all the variables[a]

Model	Range	Calibration	Validation	Commercial
PLS	2%–20%	*n* = 134 # LV = 2; # errors = 29	*n* = 39 # errors = 11	*n* = 2 # errors = 0
	25%–100%	*n* = 86 # LV = 2; # errors = 11	*n* = 44 # errors = 5	*n* = 21 # errors = 1
SIMCA	2%–20%	*n* = 134 # factors for each class: 2%: 1; 4%: 2; 6%: 3; 8%: 4; 10%: 2; 16%: 3; 20%: 4 # errors = 19	*n* = 39 # errors = 14	*n* = 2 # errors = 0
	25%–100%	*n* = 86 # factors for each class: 20%: 2; 25%: 2; 50%: 2; 70%: 3; 100%: 4 # errors = 15	*n* = 44 # errors = 12	*n* = 21 # errors = 1
Potential curves	2%–20%	*n* = 134 factor subspace: PC1–PC2 # errors = 4	*n* = 39 # errors = 9	*n* = 2 # errors = 0
	25%–100%	*n* = 86 factor subspace: PC1–PC2 # errors = 4	*n* = 44 # errors = 6	*n* = 21 # errors = 0

[a] *n* = number of samples; # LV = number of latent variables.

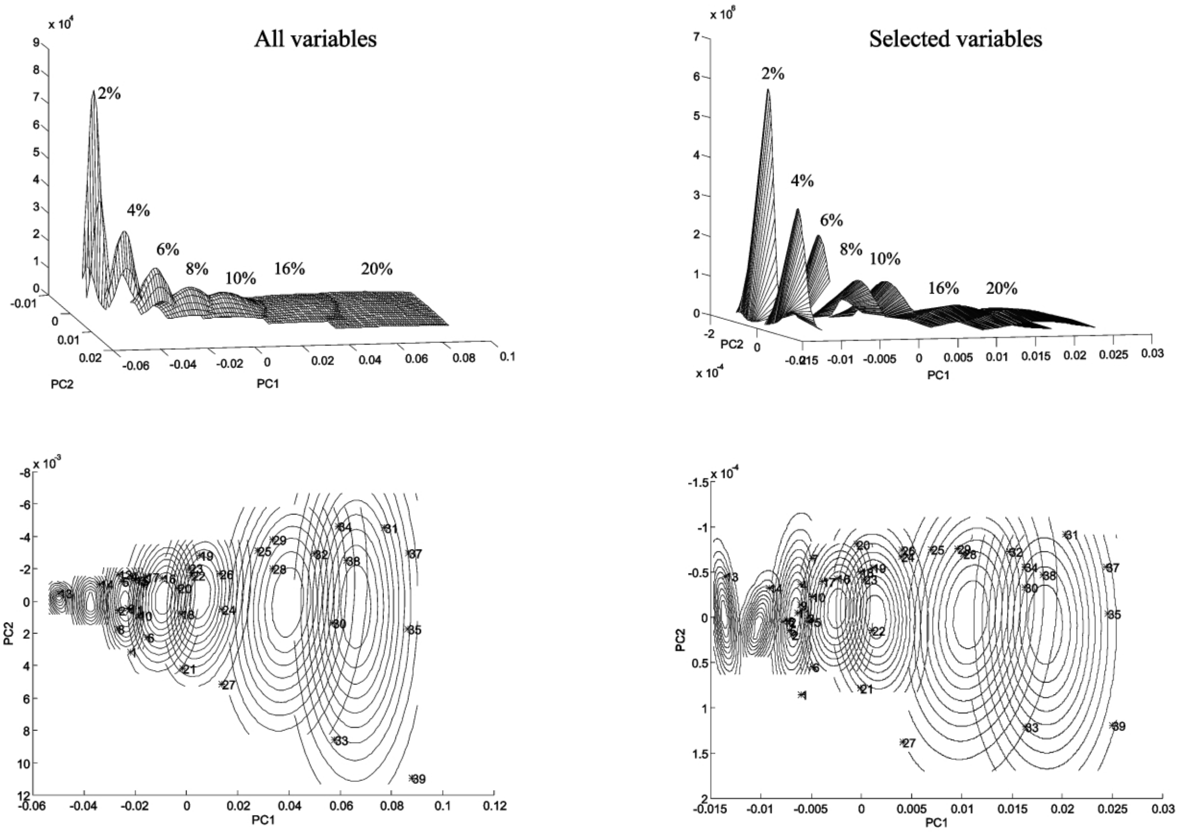

Fig. 1 Iso-probability potential regions. Upper figures, bivariate Gaussian distributions (potential curves). Lower figures, iso-probability ellipses in the PC1–PC2 subspace (the asterisks correspond to validation samples).

rotation are two. These were 1064 cm^{-1} and 1062 cm^{-1}, which correspond to the highest spectral band in the 1290–900 cm^{-1} region. The 1062 cm^{-1} band has been assigned to fructose.[42] Close examination of the spectra showed that a slight peak shift occurred for some samples, and maybe this would be accounted for by the two selected variables. Unfortunately, this issue could not be ascertained fully because whenever data scaling is changed the retained variables change, as expected. For instance, when data were autoscaled 1074 cm^{-1} and 1148 cm^{-1} (2%–20% range) and 996 cm^{-1} and 1158 cm^{-1} (25–100% range) were retained and, so, the eventual “peak-shift effect” was not obvious. More research has to be done here.

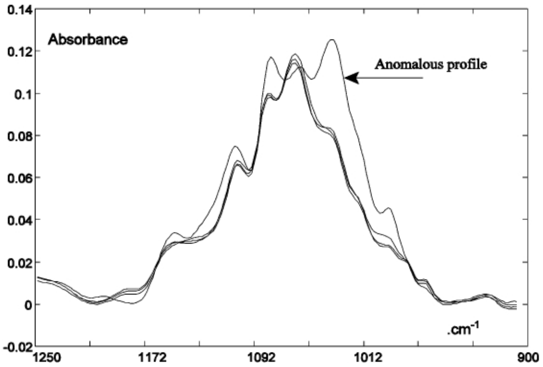

Fig. 2 Anomalous IR spectral profile of a commercial 100% juice, suggesting inappropriate sugar addition.

New classification models were developed based on the two selected variables (see Table 3). Again PLS and potential curves performed better than SIMCA, which did not give a useful model. This seems reasonable considering that three to four PCs were required to model each class when all 176 variables were employed. Since, only two PCs can be obtained from two variables SIMCA fails. PLS and potential curves maintained their good properties. The PLS model reduced the total number of errors during calibration and validation, although their behaviour remained almost unchanged. The potential curves model increased the number of errors during training and validation. Both methods classified commercial samples satisfactory.

Table 4 presents the three quality parameters after variable selection. The SIMCA models were excluded since they failed for the reduced set. PLS maintained its properties and there are even improvements in specificity for the 8%,10% and 16% classes. Potential curves show somewhat less good results; efficiency, sensitivity and specificity decreased for the 8%, 10% and 16% classes. But the overall performance is good. The isoprobability ellipses as well as the projection of the validation samples can be seen in Fig. 1. The patterns are almost the same as when using all variables. All models performed less well for the 8%–10%, 10%–16% and to a less degree 16%–20% classes with the reduced set. This is a consequence of the complexity of the original problem where not all the different classes of laboratory standard juices can be clearly discriminated because of their similar IR spectra. Although the huge reduction in the number of variables loses some information, the key question is if the remaining information is sufficient for useful classifications and decision making. As the number of errors when

Table 2 Efficiency (Eff), sensitivity (Sen) and specificity (Spe), expressed as % of total samples, for each of the multivariate models

		Class: % of apple juice										
		2%	4%	6%	8%	10%	16%	20%	25%	50%	70%	100%
Eff	PLS	100	100	100	50	55.6	60	62.5	80.9	92.3	100	100
	SIMCA	100	50	100	37.5	40	83.3	83.3	76.2	83.3	25	57.1
	Pot C[a]	100	100	100	60	57.1	60	66.7	80.9	100	100	100
Sen	PLS	100	100	84.6	33.3	83.3	50	83.3	94.4	100	100	100
	SIMCA	100	0	46.2	50	66.7	83.3	83.3	88.9	76.9	100	83.3
	Pot C[a]	100	100	84.6	50	66.7	50	100	94.4	100	100	100
Spe	PLS	100	100	100	93.9	87.9	93.9	90.9	84.6	96.8	100	100
	SIMCA	100	97.4	100	84.9	81.8	97	97	80.8	93.6	93	92.1
	Pot C[a]	100	100	100	93.9	90.9	93.9	91	84.6	100	100	100

[a] Pot C = potential curves.

Table 3 Multivariate models and results using the minimum number of selected variables[a]

Model	Range	Calibration	Validation	Commercial
PLS	2%–20%	n = 134	n = 39	n = 2
		# LV = 2 ; # errors = 29	# errors = 9	# errors = 0
	25%–100%	n = 86	n = 44	n = 21
		# LV = 2 ; # errors = 6	# errors = 6	# errors = 1
SIMCA	2%–20%	n = 134	n = 39	n = 2
		# factors for each class: 2%: 2; 4%: 2; 6%: 2; 8%: 2; 10%: 2; 16%: 2; 20%: 2	# errors = 17	# errors = 0
		# errors = 36		
	25%–100%	n = 86	n = 44	n = 21
		# factors for each class: 20%: 2; 25%: 2; 50%: 2; 70%: 2; 100%: 2	# errors = 27	# errors = 8
		# errors = 44		
Potential curves	2%–20%	n = 134	n = 39	n = 2
		factor subspace: PC1–PC2	# errors = 13	# errors = 0
		# errors = 29		
	25%–100%	n = 86	n = 44	n = 21
		factor subspace: PC1–PC2	# errors = 6	# errors = 0
		# errors = 5		

[a] n = number of samples; # LV = number of latent variables.

validating and, more importantly when classifying commercial samples, does not increase (except for the SIMCA case) it suggests that Procrustes rotation is a good procedure to reduce large data sets to a minimum number of significant variables.

PLS regression

The problem we are considering can also be approached by multivariate regression. Using PLS ("1-block") with percentage of apple juice as the predictand variable and either the 176 or the 2 selected wavenumbers as predictors gave the results summarized in Table 5. In order to select the best model not only the root mean squared error of calibration, RMSEC; root mean squared error of prediction by cross-validation leave-one-out, RMSEP-CV-LOO; and root mean squared error of prediction, RMSEP were studied but the relative errors associated to each group of samples (*i.e.*, error × 100/nominal percentage). Using the two variables selected by Procrustes rotation the PLS model performed as good as when using all 176 variables.

For comparison, a reduced set of variables was selected by considering the wavenumbers related to the most important regression coefficients of the full-spectra PLS regression. This gave 76 wavenumbers for the 2%–20% range and 67 wavenumbers for the 25%–100% range. The PLS predictions using these sets of variables are presented in Table 5. Although the sets were fairly large they did not predict better than the two variables selected by the Procrustes technique, in fact in some cases they even predicted slightly worse.

Conclusions

It was shown that Procrustes rotation can be used to select an small number of mid-IR wavenumbers (not spectral ranges, but individual variables) intended to develop multivariate models which can be used to deploy a screening methodology to evaluate the amount of apple juice in commercial apple-based beverages. Several classification models were implemented considering the original (176 wavenumbers) and the Procrustes-reduced (2 wavenumbers) data sets; namely, a simplified mode of potential functions (termed potential curves), SIMCA and discriminant PLS. Another study compared regression models developed using the above two data sets and another reduced data set obtained by a classical approach (selection of variables with highest loadings/regression coefficients). Except for SIMCA models, the small subset of variables selected by

Table 4 Minimum number of selected variables. Efficiency (Eff), Sensitivity (Sen) and Specificity (Spe), expressed as % of total samples, for each of the multivariate models

		Class: % of apple juice										
		2%	4%	6%	8%	10%	16%	20%	25%	50%	70%	100%
Eff	PLS	100	100	100	66.7	55.6	75.5	66.7	80	92.9	100	100
	SIMCA	50	100	100	40	57.1	50	50	85.7	100	52.6	57.1
	Pot C[a]	100	100	100	40	50	50	60	80	92.7	100	100
Sen	PLS	100	100	92.3	33.3	83.3	50	100	88.9	100	100	100
	SIMCA	100	0	53.9	66.7	66.7	33.3	66.7	33.3	15.4	100	66.7
	Pot C[a]	100	100	76.9	33.3	66.7	33.3	100	88.9	100	100	100
Spe	PLS	100	100	100	97	87.9	97	90.9	84.6	96.8	100	100
	SIMCA	50	100	100	84.9	90.9	93.9	87.9	96.2	100	60.5	92.1
	Pot C[a]	100	100	100	90.9	87.9	93.9	87.9	84.6	96.8	100	100

[a] Pot C = potential curves.

Table 5 Main characteristics of the PLS regression models developed considering all the spectral wavenumbers, the two Procrustes-selected and a reduced set of variables considering the correlation coefficients (see text)

		All variables			Two selected variables			Reduced range		
		Relative error			Relative error			Relative error		
Range		Calibration	Validation	Commercial	Calibration	Validation	Commercial	Calibration	Validation	Commercial
2%–20%	2%	18.7	10.0[a]		18.6	16.3[a]		27.3	10.4[a]	
	4%	14.0	10.0[a]		13.7	17.0[a]		14.8	13.6[a]	
	6%	12.3	11.4		10.8	10.6		12.0	11.6	
	8%	12.7	16.3		14.6	14.7		14.1	17.0	
	10%	11.0	18.2		13.5	13.8		13.4	18.8	
	16%	12.4	11.9	2.0[a]	13.3	11.9	1.0[a]	13.1	16.1	2.5[a]
	20%	10.4	14.9	3.8[a]	11.2	11.7	11.0[a]	10.4	14.4	6.9[a]
	General features	LV = 2; % of EI in X = 99.7; % of EI in Y = 94.8; RMSEC = 1.4; RMSEPCV = 1.4; RMSEP = 1.9; F-test = 3.6[b]			LV = 2; % of EI in X = 100; % of EI in Y = 94.3; RMSEC = 1.4; RMSEPCV = 1.5; RMSEP = 1.8; F-test = 3.9[b]			LV = 2; % of EI in X = 99.9; % of EI in Y = 94.6; RMSEC = 1.4; RMSEPCV = 1.4; RMSEP = 1.9; F-test = 3.7[b]		
25%–100%	25%	10.8	11.3	15.3	13.8	12.9	15.1	16.4	24.3	15.5
	50%	9.0	8.5		8.7	8.5		9.1	8.6	
	70%	8.6	1.1[a]		8.8	2.2[a]		8.6	1.7[a]	
	100%	10.3	16.6	8.3	11.0	12.1	8.5	10.2	16.7	9.1
	General features	LV = 2; % of EI in X = 99.9; % of EI in Y = 96.0; RMSEC = 6.0; RMSEPCV = 6.3; RMSEP = 7.4; F-test = 1.7[b]			LV = 2; % of EI in X = 100; % of EI in Y = 95.6; RMSEC = 6.3; RMSEPCV = 6.6; RMSEP = 7.3; F-test = 1.9[b]			LV = 2; % of EI in X = 99.9; % of EI in Y = 96.0; RMSEC = 6.0; RMSEPCV = 6.3; RMSEP = 7.5; F-test = 1.7[b]		

[a] Only one sample available. [b] EI = explained information; RMSEC = root mean squared error of calibration; RMSEPCV = root mean squared error of prediction by cross-validation leave-one-out calculated from the calibration set; RMSEP = root mean squared error of prediction calculated from the validation set; F-test is the joint test for the slope and the intercept (real *vs.* predicted values); F(tabulated, 95%) = 3.1, F(tabulated, 99%) = 4.8.

Procrustes rotation yielded similar results than the other options assayed in this work and, so, it can be concluded that Procrustes rotation is an efficient procedure for variable selection for classification and regression methods.

Appendix, notation

In this paper bold italic lower case denote vectors whereas bold italic upper case denote matrices, (^{T}) means a transposed matrix.

n = number of samples.

p = number of variables in the original data set ($\boldsymbol{X}$).

q = number of the optimum principal components (latent variable) which is also the number of variables that will be retained by Procrustes rotation.

Wm = statistic developed to evaluate the q optimum number of principal components which take account of the main information in the data set (see ref. 16 for details).

$\boldsymbol{T}$ = subspace of the first q ("optimum") principal component scores, this is the true (or target) matrix to which any other principal component subspace will be compared by Procrustes rotation.

$\boldsymbol{R}$ = subspace of the first q principal component scores obtained after deleting a given variable.

M^2 = statistic representing the sum of the squared differences amongst $\boldsymbol{T}$ and $\boldsymbol{R}$, it corresponds to the "Procrustes match" after translation, rotation and stretching (data was mean centred before starting calculations; if not, the equations are a bit more complex).

$\boldsymbol{A}$ = matrix defining the orthogonal rotation to match $\boldsymbol{R}$ with $\boldsymbol{T}$.

Matrix decomposition is typically made by two well-known methods:

A.—Singular Value Decomposition (svd) decomposes a data matrix ($\boldsymbol{X}$) into three matrices, svd($\boldsymbol{X}$) = $\boldsymbol{USW}^T$. The decomposition is generally done using the Householder diagonalization. The columns of the matrix $\boldsymbol{W}$ correspond to the basis set vectors (loadings), the columns of the matrix $\boldsymbol{US}$ correspond to the sample scores.

B.—NIPALS decompositon (= Nonlinear Iterative Partial Least Squares) is performed by calculating two new matrices such as $\boldsymbol{X} = \boldsymbol{TP}^T$ (the loadings being the columns of $\boldsymbol{P}$ and the scores correspond to the columns of $\boldsymbol{T}$). $\boldsymbol{T}$ and $\boldsymbol{P}$ are calculated from iterative algorithms which have been described elsewhere (*e.g.* refs. 36,37,43).

PLS studies were carried out using the PLS-Toolbox for Matlab, v. 1.5.2, Eigenvector Technologies, Manson, WA, USA.

Procrustes rotation, potential curves and SIMCA studies were performed by in-house Matlab programs.

References

1 H. M. Heise and A. Bittner, *Fresenius' J. Anal. Chem.*, 1997, **359**, 93–99.

2 J. Grimalt and J. Olive, *Anal. Chim. Acta*, 1993, **278**, 159–176.

3 W. J. Krzanowski, *J. Chemom.*, 1995, **9**, 509–520.

4 V. Centner, D. L. Massart, O. E. De Noord, S. de Jong, B. M. Vadeginste and C. Sterna, *Anal. Chem.*, 1996, **68**, 3851–3858.

5 A. Garrido-Frenich, M. D. Gil-García, J. L. Martínez-Vidal and M. Martínez-Galera, *Quím. Anal.*, 1999, **18**, 319–327.

6 H. C. Goicoechea and A. C. Olivieri, *Talanta*, 1999, **49**, 793–800.

7 A. J. Berger, T. Koo, I. Itzkan and M. S. Feld, *Anal. Chem.*, 1998, **70**, 623–627.

8 A. Elbergali, R. G. Brereton and A. Rahmani, *Analyst*, 1995, **120**, 2207–2216.

9 J. Ferré and F. X. Rius, *Trends Anal. Chem.*, 1997, **16**, 155–162.

10 R. Todeschini, D. Galvagni, J. L. Vílchez, M. del Olmo and N. Navas, *Trends Anal. Chem.*, 1999, **18**, 93–98.

11 R. Leardi, *J. Chemom.*, 2001, **15**, 559–569.

12 B. M. Smith and P. J. Gemperline, *Anal. Chim. Acta*, 2000, **423**, 167–177.

13 Q. Guo, W. Wu, D. L. Massart, C. Boucon and S. de Jong, *Chemom. Intell. Lab. Syst.*, 2002, **61**, 123–132.

14 K. Heberger and R. Rajko, *J. Chemom.*, 2002, **16**, 436–443.

15 R. Rajko and K. Heberger, *Chemom. Intell. Lab. Syst.*, 2001, **57**, 1–14.

16 W. J. Krzanowski, *Principles of Multivariate Analysis*, Clarendon Press, Oxford, UK, 1988.

17 A. Carlosena, J. M. Andrade, M. Kubista and D. Prada, *Anal. Chem.*, 1995, **67**, 2373–2378.

18 M. B. Richman and S. J. Vermette, *Atmos. Environ.*, 1993, **27A**, 475–481.

19 J. R. King and D. A. Jackson, *Environmetrics*, 1999, **10**, 67–77.

20 J. M. Deane and H. J. H. Macfie, *J. Chemom.*, 1989, **3**, 477–491.

21 J. M. Andrade, S. Muniategui, P. Lopez and D. Prada, *Fuel*, 1997, **76**, 51–59.

22 I. Scarminio and M. Kubista, *Anal. Chem.*, 1993, **65**, 409–416.

23 J. Nygren, A. Elbergali and M. Kubista, *Anal. Chem.*, 1998, **70**, 4841–4846.

24 J. Nygren, J. M. Andrade and M. Kubista, *Anal. Chem.*, 1996, **68**, 1706–1710.

25 M. Kubista, J. Nygren, A. Elbergali and R. Sjöback, *Crit. Rev. Anal. Chem.*, 1999, **29**, 1–28.

26 M. Kubista, *Chemom. Intell. Lab. Syst.*, 1990, **7**, 273–279.

27 D. González-Arjona, G. López-Pérez and G. A. González, *Talanta*, 1999, **49**, 189–197.

28 E. Vigneau, M. F. Devaux and M. Safar, *J. Chemom.*, 1995, **9**, 125–135.

29 K. S. Booksh and B. R. Kowalski, *J. Chemom.*, 1994, **8**, 287–292.

30 C. E. Anderson and J. H. Kalivas, *Appl. Spectrosc.*, 1999, **53**, 1268–1276.

31 W. Wu, D. L. Massart and S. de Jong, *Chem. Intell. Lab. Syst.*, 1997, **36**, 165–172.

32 L. Braakman, *Food Engineering and Ingredients*, 2002, **27**, 14–19.

33 B. M. Silva, R. M. Seabra, P. B. Andrade, M. B. Oliveira and M. A. Ferreira, *Cienc. Tecnol. Aliment.*, 1999, **2**, 184–191.

34 P. Stöber, G. G. Martin and T. L. Peppard, *Dtsch. Lebensm.-Rundsch.*, 1998, **94**, 309–316.

35 B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*, CRC Press, Boca Raton, USA, 1996.

36 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.

37 R. G. Brereton, *Analyst*, 2000, **125**, 2125–2154.

38 S. Wold and M. Sjöström, in *Chemometrics: Theory and Application*, ed. B. R. Kowalski, ACS Symposium Series, Washington, USA, 1977.

39 B. Mertens, M. Thompson and T. Fearn, *Analyst*, 1994, **119**, 2777–2784.

40 M. Forina, C. Armanino, R. Leardi and G. Drava, *J. Chemom.*, 1991, **5**, 435–453.

41 X. Tomas and J. M. Andrade, *Quim. Anal.*, 1999, **18**, 225–231.

42 S. Sivakesava and J. Irudayaraj, *Appl. Eng. Agric.*, 2000, **16**, 543–550.

43 S. Wold, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37–52.