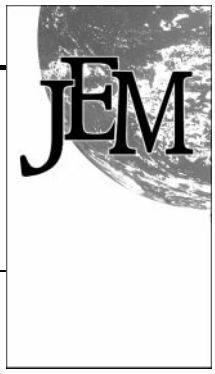

Evaluation of the effect of data pre-treatment procedures on classical pattern recognition and principal components analysis: a case study for the geographical classification of tea

Antonio Moreda-Piñeiro,[a] Ana Marcos,[b] Andrew Fisher[b] and Steve J. Hill*[b]

[a]*Department of Analytical Chemistry, Nutrition and Bromatology, Faculty of Chemistry, University of Santiago de Compostela, Avenida das Ciencias s/n, 15706 Santiago de Compostela, Spain*
[b]*Department of Environmental Sciences, Plymouth Environmental Research Centre, University of Plymouth, Drake Circus, Plymouth, Devon, UK PL4 8AA*

Received 24th April 2001, Accepted 29th June 2001
First published as an Advance Article on the web 20th July 2001

A simple transformation that uses the half-range and central value has been used as a data pre-treatment procedure for principal component analysis (PCA) and pattern recognition techniques. The results obtained have been compared with the results from classical normalisation of data (mean normalisation, maximum normalisation and range normalisation), autoscaling and the minimum–maximum transformation. Three data sets were used in the study. The first was formed by determining 17 elements in 53 tea samples (901 pieces of data). The second and third data sets arose from two long-term drift studies performed to examine instrumental stability at standard and robust conditions. The instruments used were an inductively coupled plasma atomic emission spectrometer and an inductively coupled plasma mass spectrometer. Each drift diagnosis experiment consisted of replicate determinations of a test solution containing 15 analytes at 10 mg l^{-1} over 8 h without recalibration. Twenty-nine emission lines were determined 99 times, thus, each data set was formed by 2881 pieces of data. Data pre-treatment was applied to the three data sets prior to the use of principal component analysis, cluster analysis, linear discrimination analysis and soft independent modelling of class analogy. The study revealed that the half-range and central value transformation resulted in a better classification of the tea samples than that achieved using the classical normalisation. The loadings in the PCA for the long-term stability study, under both standard and robust conditions, were found to be similar to the drift trends only when the minimum–maximum transformation and the mean or maximum normalisations were used as data pre-treatments.

Introduction

The use of chemometric tools such as principal component analysis (PCA) and pattern recognition techniques has grown rapidly in terms of their application to the classification of product brands and produce origin.[1] This growth has been aided by the development of new and powerful chemometric algorithms for data processing. However, the successful application of many chemometric techniques to a data set depends heavily on the data pre-treatment employed. It is common to encounter experimental data sets formed by variables of different natures (for instance, temperature, pressure and concentration), which require some homogenisation before applying a multivariate method. Moreover, even when the data set is homogenous (*i.e.*, same units for all variables) severe differences can be found in the magnitudes of the variables. Since chemometric tools often work by explaining variation, *e.g.*, PCA, the variables with maximum variation will be employed in the construction of the model. This is important in many applications in analytical chemistry, where the magnitudes of the different variables, *e.g.*, the concentration of elements within a sample, are often very different, resulting in the higher concentrations being more important in the model. This may have the effect of masking the information from the other variables that would have less weight in the model. Thus, before using pattern recognition techniques, it is important to prepare the data in an appropriate manner. This preliminary stage is often called pre-processing or pre-treatment.[2]

When considering pre-treatment, the direction (along columns or rows) in which the pre-treatment will be performed must be decided first. This will depend on the type of problem under study. Then, the type of pre-treatment needs to be selected. Most commercial software packages offer autoscaling or mean centring as data pre-treatments.[3] However, there are few comparative studies of different data pre-treatments for different data sets, and in fact, many published papers about classification or authentication of product brands do not specify the data pre-treatment (if any) employed. Of the comparative studies that have been reported, de Braekeleer *et al.*[4] have compared different data pre-treatments prior to the application of the orthogonal projection approach (OPA) and soft independent modelling of class and analogy (SIMCA) to study the end point of a polymorph conversion reaction by near infrared spectroscopy (NIR), and Wu *et al.*[5] have compared different data pre-treatments in pattern recognition of multivariate data.

In contrast, the situation is different for partial least-squares (PLS), principal component regression (PCR), or multiple linear regression (MLR) models where several comparative studies on the data pre-treatment have been reported.[6–15]

In this work, transformations based on the half-range (hr)

 DOI: 10.1039/b103658k

and central value (cv) have been applied for data pre-treatment to three different data sets prior to PCA and pattern recognition techniques, cluster analysis (CA), linear discriminant analysis (LDA) and SIMCA. One of the data sets is formed from 17 elemental concentrations in 53 tea samples of different geographical origin ($17 \times 53 = 901$ pieces of data). The other two data sets represent intensity against time (8 h) for 29 emission lines using a simultaneous ICP-AES instrument. Here, each variable was recorded at approximately 5 min intervals (*i.e.*, $29 \times 99 = 2881$ pieces of data). The two last data sets correspond to an ICP-AES instrument operating under standard and robust conditions. The achieved results, following half-range central value transformation, were compared with those produced after application of a classical data pre-treatment such as normalisation or autoscaling.

Theoretical background

As described previously, data pre-treatment may be performed either along the rows or along the columns. The data pre-treatment methods used in this work were the mean normalisation, maximum normalisation, range normalisation, autoscaling, minimum–maximum transformation and Half-range central value transformation. A brief description of all of these data pre-treatments is summarised in Table 1.

Experimental

Instrumentation and software

A Liberty 200 ICP-AES (Varian, Cheshire, UK) and a PlasmaQuad PQ2+ ICP-MS (Thermo Elemental, Winsford,

Table 1 Summarised theoretical background of the different data pre-treatments

Data pre-treatment	Mathematical equation[a]	Description	Ref.
Mean normalisation	Along the rows $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\bar{x}_i}$ Along the columns $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\bar{x}_j}$	Each row/column of the data matrix is divided by its average. This pre-treatment will homogenise the data, by converting all the variable mean values to unity and the rest close to unity. This may be useful when data of very different magnitude are present. However there is a risk of partially erasing differences between samples.	2, 3
Maximum normalisation	Along the rows $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\text{Max}(x_i)}$ Along the columns $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\text{Max}(x_j)}$	This is an alternative to classical normalisation that divides each row/column by its maximum absolute value instead of the average. This transformation, as with the mean normalisation, homogenises the data matrix. In this case, the maximum values will become unity and the rest will range between 1 and 0.	2, 3
Range normalisation	Along the rows $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\text{Max}(x_i) - \text{Min}(x_i)}$ Along the columns $x_{ij}^{\text{normalised}} = \frac{x_{ij}}{\text{Max}(x_j) - \text{Min}(x_j)}$	Each row is divided by its range, *i.e.*, the maximum value minus the minimum value. The range normalisation will homogenise the data as in any normalisation step, but the new data array will still show differences between data points.	2, 3
Autoscaling	Along the rows $x_{ij}^{\text{autoscaled}} = \frac{x_{ij} - \bar{x}_i}{S(x_i)}$ Along the columns $x_{ij}^{\text{autoscaled}} = \frac{x_{ij} - \bar{x}_j}{S(x_j)}$	For a normal distribution of the x values, the mean and the standard deviation can be used for standardisation. This transformation converts the new data set to zero mean and unity standard deviation. The advantage of this type of pre-treatment is that every variable will be fully comparable because they show a similar variance. However, this pre-treatment also presents two disadvantages. The first concerns variables without a normal distribution (high standard deviation), and secondly, autoscaling may amplify noise for such variables whose magnitude is close to the detection limit of the technique.	2, 16
Minimum–maximum transformation	Along the rows $x_{ij}^{\text{transformed}} = \frac{x_{ij} - \text{Min}(x_i)}{\text{Max}(x_i) - \text{Min}(x_i)}$ Along the columns $x_{ij}^{\text{transformed}} = \frac{x_{ij} - \text{Min}(x_j)}{\text{Max}(x_j) - \text{Min}(x_j)}$	This is a range transformation. The advantage of a minimum–maximum transformation is that the minimum value of each variable is transformed to 0 and the maximum to unity. This facilitates the comparison between samples, because all sample data will range between 0 and 1.	16
Half-range and central value transformation	Along the rows $x_{ij}^{\text{transformed}} = \frac{x_{ij} - cv_i}{hr_i}$ Along the columns $x_{ij}^{\text{transformed}} = \frac{x_{ij} - cv_j}{hr_j}$	This transformation is used in experimental design approaches. Half-range and central value are defined as follows. This transformation converts the minimum values to -1, the maximum values to $+1$, keeping the average values to 0. This is of great use in pattern recognition because it enhances the differences in the data.	17

[a] x_{ij} corresponds to the jth element concentration in the ith tea sample, for the data set formed by element concentrations in tea samples; and x_{ij} corresponds to the intensity of the ith emission wavelength at the jth time, for the data set formed by intensities of emission lines for long-term stability in ICP-AES. $\bar{x}_i$ and $S(x_i)$ are the mean value and the standard deviation, for each element concentration, as $\bar{x}_j$ and $S(x_j)$ are for each emission line. $\text{Max}(x_i)$ and $\text{Min}(x_i)$ are the maximum and minimum value, for each element concentration, and $\text{Max}(x_j)$ and $\text{Min}(x_j)$ are the maximum and minimum value for each emission line. cv_i, cv_j and hr_i, hr_j are as defined as follows: along the rows $hr_i = \frac{\text{Max}(x_i) - \text{Min}(x_i)}{2}$ and $cv_i = \frac{\text{Max}(x_i) + \text{Min}(x_i)}{2}$, along the columns $hr_j = \frac{\text{Max}(x_j) - \text{Min}(x_j)}{2}$ and $cv_j = \frac{\text{Max}(x_j) + \text{Min}(x_j)}{2}$.

Table 2 Percentage of explained variability as defined by the first three principal components and dominating features

Data pre-treatment	Variability accounted with the first three PCs (%)	Dominating features in the first three PCs		
		1st PC	2nd PC	3rd PC
Autoscaling	56.93	Fe, V	Cr, Ba, Pb	Cr, Ni, Rb
Mean normalisation	87.26	Cr, Pb, Cu	Cr	Cs, Sr, Pb
Maximum normalisation	91.05	Ca, Mg	Pb, Fe, V, Sr	Rb, Co, Ba
Range normalisation	93.55	Ca, Mg	Sr, Pb, Fe, V	Mg, Ba, Pb
Minimum–maximum transformation	87.46	Ti, Ca, Al	Sr, Ba, Pb	Rb, Co, Zn
Half-range and central value transformation	78.95	Cr, Pb, Cs, V	Ca, Ba	Sr, Cr, Al

Cheshire, UK) were used for the determination of metals in tea samples. Optimised operating conditions as well as the emission wavelengths or isotope masses for ICP-AES and ICP-MS, respectively, have been given elsewhere.[18]

An Optima 3000 ICP-AES (PerkinElmer Corporation, Norwalk, USA) was used to acquire data to characterise the drift phenomena. The instrumental parameters employed for both standard and robust conditions were taken from refs. 19 and 20.

The total digestion of the tea samples was achieved using a hot plate (SH3, Stuart Scientific, UK) and employing the method described previously.[21]

Two commercial chemometrics packages (UNSCRAMBLER, 1998, CAMO ASA, Trondheim, Norway) and Statgraphics Plus V 4.0, 1994–1999, Manugistics Inc., Rockville M.D) were used in this study.

Reagents

All chemicals used were of ultrapure grade, diluted using ultrapure water of resistivity 18 MΩ cm obtained from a Milli-Q purification device (Millipore Co., Bedford, MA, USA). AnalaR nitric acid 70.0% was obtained from Merck (Poole, Dorset, UK). Stock standard solutions (1.000 or 10.000 g l^{-1}) were supplied by Merck. A cobalt stock standard solution was purchased from Aldrich (Gillingham, Dorset, UK).

Multielement determination in tea samples

Tea samples were digested four times using a method published previously.[21] Indium (as an internal standard for ICP-MS measurements) was added to each digest to give a concentration of 100 μg l^{-1} after dilution to 50 ml. The tea acid digests were kept in polyethylene vials at room temperature prior to analysis by ICP-MS and ICP-AES. The elements Co, Cr, Cs, Cu, Ni, Pb, Rb, Ti and V were determined by ICP-MS without dilution according to ref. 18. In addition Al, Ba, Ca, Fe, Mg, Mn, Sr and Zn were determined by ICP-AES, also without dilution.[18]

Multielement test solutions

A test solution containing fifteen analytes at 10 mg l^{-1} was repeatedly analysed by ICP-AES over a period of 8 h without

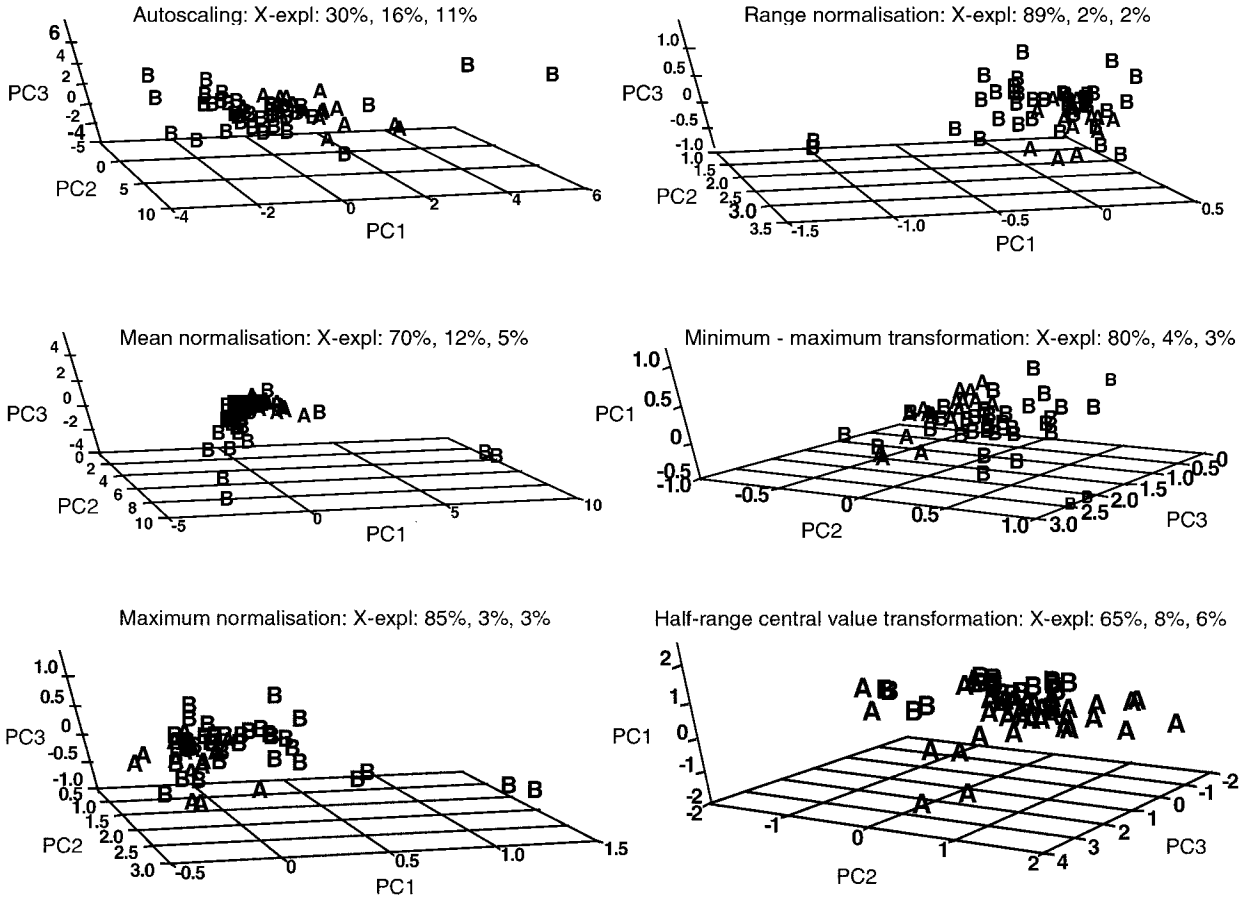

Fig. 1 Scores of the tea samples in the three-space formed by the three first principal components: A, African tea samples; B, Asian tea samples.

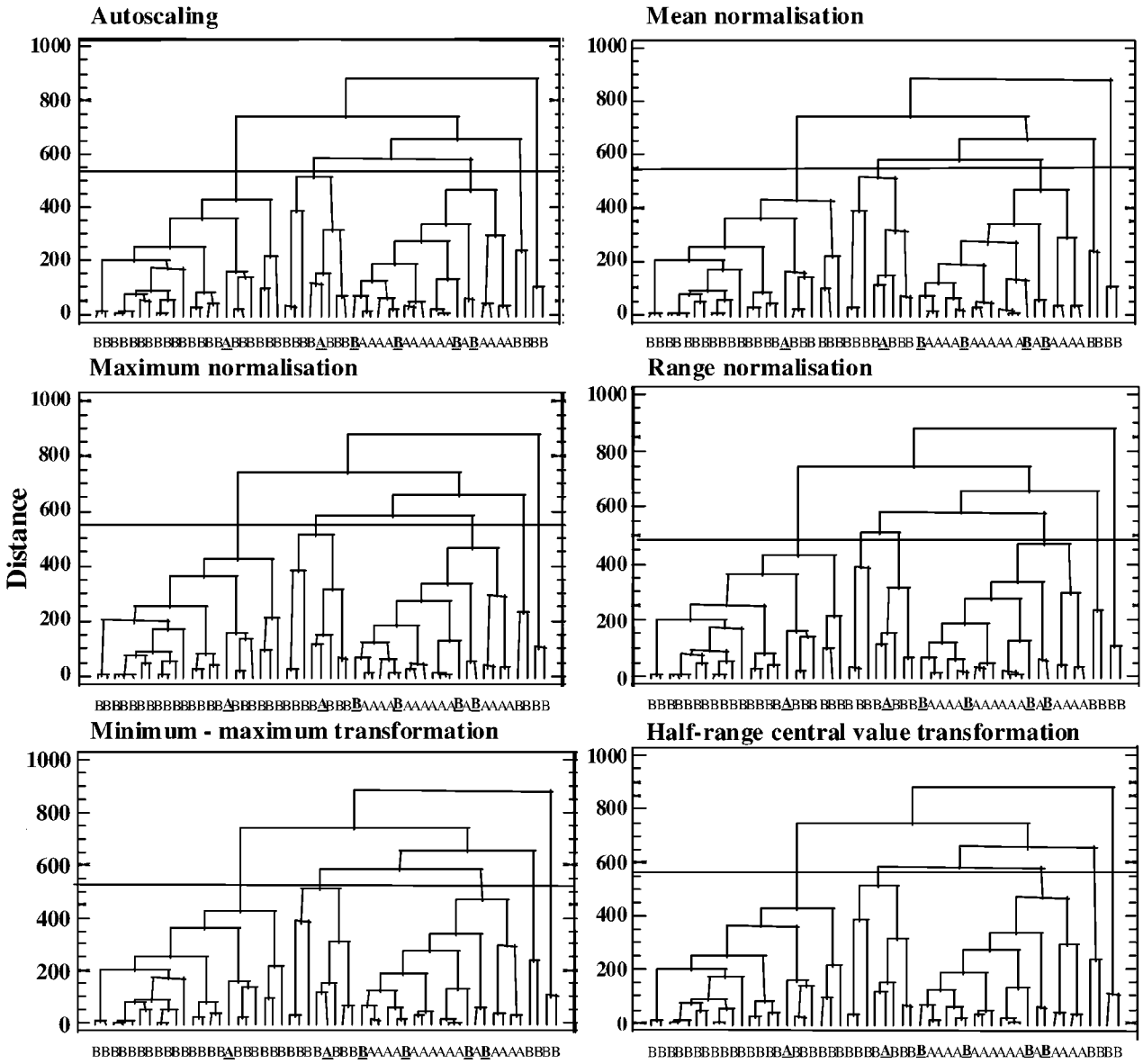

Fig. 2 Dendrogram of cluster analysis: A, African tea samples; B, Asian tea samples.

re-calibration. The test solution was matched with 2% nitric acid. The instrumental parameters for both standard and robust conditions and the emission wavelengths are given in refs. 19 and 20.

Data sets

A data set comprising 901 pieces of data (17 variables measured over 53 tea samples) was used in the study. Thirty-six of the tea samples were from different Asian countries and 17 samples from African countries. The tea samples were obtained from commercial retailers.

The second and third data sets represent the emission intensity against time (8 h) for the 29 emission lines (15 analytes) using ICP-AES under both standard and robust conditions. By taking measurement at approximately 5 min intervals, this represents $29 \times 99 = 2881$ pieces of data.

Results and discussion

The data under study were arranged in such a way that data pre-treatment was performed along the columns in the case of the tea samples data set (element concentrations for the classification study of tea samples) and along the rows in the other two cases concerning the study of drift in ICP-AES (emission lines for the study of drift diagnosis).

Geographical classification of tea samples

Principal component analysis and cluster analysis. A PCA, using cross validation as validation method, was performed on the data set (17 variables and 53 tea samples) for the raw data and after the application of each data pre-treatment. Seventeen PCs were calculated and the cumulative variance for the first three PCs listed in Table 2. It can be seen that most of the variation in the data is explained by the first three PCs, except when using autoscaling for data pre-treatment (53.9% of the explained variance). That indicates that the data structure after autoscaling is too complex to be accounted for by only three PCs, probably as a result of the pre-treatment itself, since division by the standard deviation may enhance the noise. Thus, all data pre-treatment methods except autoscaling appear to be adequate as data pre-treatment methods when considering only three PCs. It is very important to note that, in the absence of any data pre-treatment, only the more concentrated elements are considered by the PCA model, although the variance explained by PC1 is close to 100%. In addition, the loadings are different after each data pre-treatment.

Table 3 Classification of tea samples with linear discriminant analysis[a,b]

Actual group	Size	Predicted group	
		Africa (A)	Asia (B)
Africa (A)	17	17 (100.0%)	0 (0.0%)
Asia (B)	36	1 (2.8%)	35 (97.2%)

[a]Data pre-treatment: Autoscaling; mean normalisation; maximum normalisation; range normalisation; minimum–maximum transformation and half-range central value transformation. [b]Discriminant function: $F = 0.47Al - 0.60Ba - 0.54Ca + 0.80Co - 0.34Cr - 0.70Cs - 0.32Cu - 0.68Fe - 0.33Mg + 0.43Mn + 0.16Ni - 0.54Pb + 0.61Rb + 1.21Sr + 0.49Ti - 0.16\,V + 0.43Zn$.

The scores for the tea samples in the three-dimensional space formed by the first three principal components after each data pre-treatment are given in Fig. 1. It can be seen that the separation between tea samples labelled as A (Africans) and B (Asians) is clearer following the use of the half-range central value transformation than when using the other data pre-treatments.

The cluster analysis results appear to be independent of the data pre-treatment method used. It can be seen in Fig. 2 that the dendrograms obtained using the square Euclidean distance and the Ward's method are very similar for all data pre-treatments. At a distance of 550, a first cluster comprising 21 tea samples from Asia and 1 tea sample from Africa (bold and underlined) was obtained, followed by a second cluster of 4 Asian teas and 1 African tea sample (bold and underlined). A third cluster formed mainly of African teas (15 tea samples) included 4 Asian teas (bold and underlined). Finally, a fourth and a fifth cluster composed of 2 Asian tea samples in each was found.

These results suggest that PCA is more dependent on the data pre-treatment carried out than cluster analysis. PCA works in terms of variance and, thus, "only" those variables with large variance will be taken into account to form the PCA model. When applying a data pre-treatment, the variance of each original variable is (modified or) scaled to a certain magnitude of variance (*i.e.*, variance 1 for auto-scaling, or range 2 for minimum–maximum transformation). Using such transformation, all variables will present the same magnitude of variance (or range) for all original variables. Therefore, the variability explained by the first PC would contain information of all original variables, and a better separation of objects will be obtained. If a very heterogeneous data set is not pre-treated, variables will have very different variances and the first PCs will be formed by using only those original variables that offer high variance. In such a case, the latent information from variables with low variance will not be taken into account. Thus, if one of these last variables is important to classify the different objects, it is evident that a worse classification will be obtained because it does not contribute to the PCA model. However, clustering techniques are methods that use all the information in the data matrix working in terms of distance. The data pre-treatment will change the absolute distances between objects but not their relative positions (*i.e.*, like a map on a different scale) and thus, the similarity between objects will remain after the pre-treatment.

Linear discriminant analysis and soft independent modelling of class and analogy. No effects following the data pre-treatment procedures on the tea classification by LDA were found since the percentages of correctly classified cases of 100.0 and 97.2% were reached for the African (A) and Asian (B) classes, respectively, using all data pre-processing methods (Table 3). Similarly, the standardised discriminant function was the same following all data pre-treatments (also shown in Table 3). It can also be seen that the variables Sr, Co, Cs and Fe offer the highest weightings. This suggests that all of the data pre-treatments are adequate prior to performing an LDA approach and the results obtained are independent of the data pre-processing. This is because PCA produces new functions (principal components) maximising the sum of squares instead of maximising the ratio of the between groups to the within groups square error as is the case with LDA.

However, a very different situation was found when using SIMCA because this method produces a PCA model for each category. Thus, as PCA results are dependent on the data pre-treatment, SIMCA results will also be dependent on the function of the data pre-processing used. Results after SIMCA (with a significance level of 5%) are presented in Table 4 and it can be seen that many cases are only correctly assigned to the category Asian teas (B) after the application of the autoscaling and the half-range central value transformation as data pre-treatments (97.2 and 88.9%, respectively). Percentages lower than 60.0% were obtained for the alternative data pre-processing methods. For African teas (samples labelled as A)

Table 4 Classification of tea samples using SIMCA

Actual group	Size	Predicted group			
		Africa (A)	Asia (B)	Both (A or B)	Neither
Autoscaling					
Africa (A)	17	**4 (19.0%)**	0 (0.0%)	13 (81.0%)	0 (0.0%)
Asia (B)	36	0 (0.0%)	**35 (97.2%)**	0 (0.0%)	1 (2.8%)
Mean normalisation					
Africa (A)	17	**6 (35.3%)**	0 (0.0%)	10 (58.8%)	1 (5.9%)
Asia (B)	36	0 (0.0%)	**17 (47.2%)**	12 (33.3%)	7 (19.5%)
Maximum normalisation					
Africa (A)	17	**6 (35.3%)**	0 (0.0%)	10 (58.8%)	1 (5.9%)
Asia (B)	36	0 (0.0%)	**19 (52.8%)**	11 (30.6%)	6 (16.6%)
Range normalisation					
Africa (A)	17	**7 (41.2%)**	0 (0.0%)	9 (52.9%)	1 (5.9%)
Asia (B)	36	0 (0.0%)	**20 (55.5%)**	10 (27.8%)	6 (16.7%)
Maximum–minimum transformation					
Africa (A)	17	**6 (35.3%)**	0 (0.0%)	10 (58.8%)	1 (5.9%)
Asia (B)	36	0 (0.0%)	**18 (50.0%)**	13 (36.1%)	5 (13.9%)
Half-range central value transformation					
Africa (A)	17	**15 (88.2%)**	0 (0.0%)	2 (11.8%)	0 (0.0%)
Asia (B)	36	0 (0.0%)	**32 (88.9%)**	4 (11.1%)	0 (0.0%)

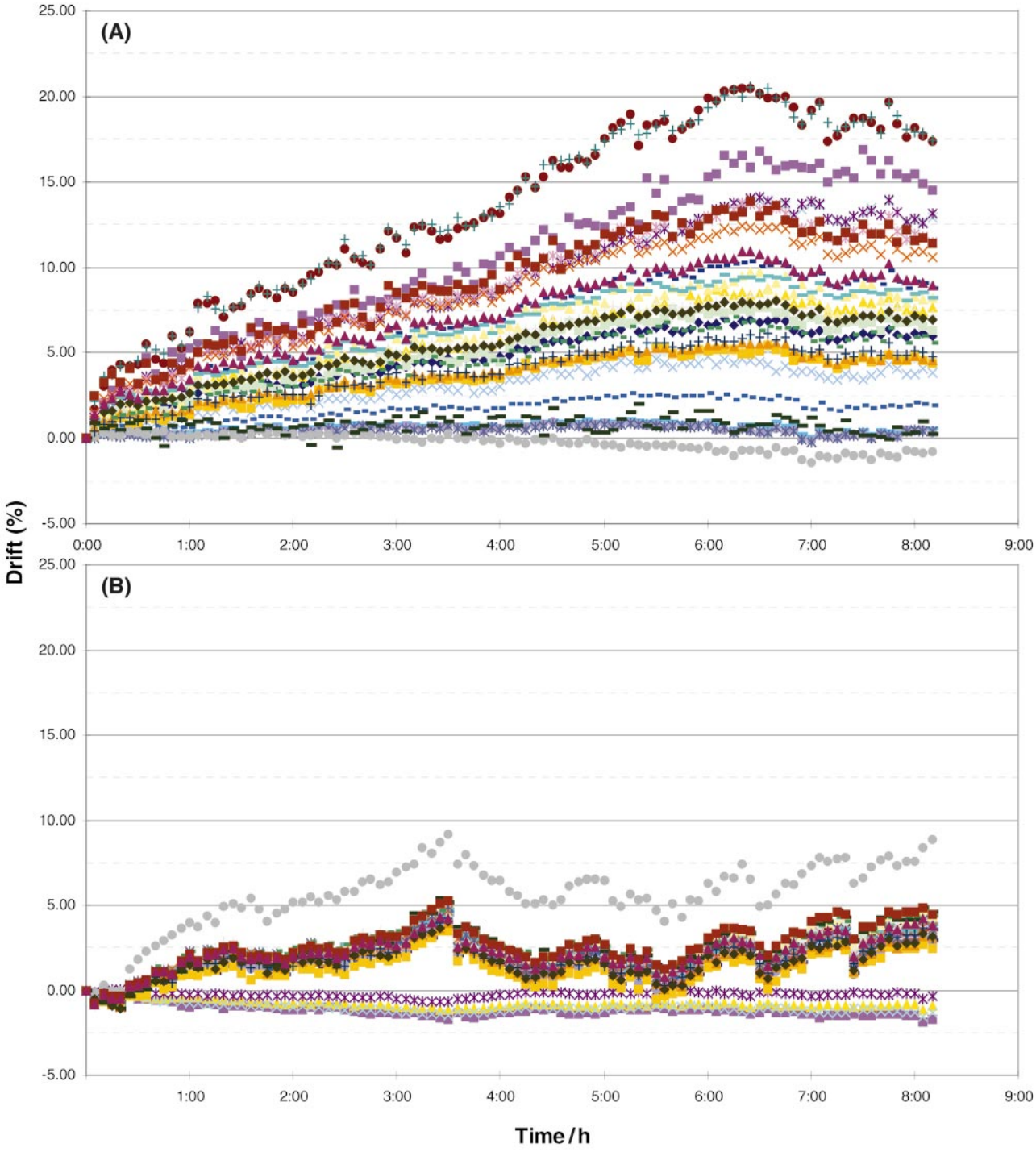

Fig. 3 Drift pattern under different experimental conditions: A, standard conditions; B, robust conditions.

88.2% of tea samples were correctly classified using the half-range central value transformation. Percentages lower than 50.0% were obtained after the use of the other data pre-treatments. Therefore, only the data pre-treatment using the half-range central value transformation gave acceptable SIMCA results (classification percentages higher than 90%) for the two classes.

The poor SIMCA results achieved following the application of most of the other data pre-treatment methods are because SIMCA is a more restrictive pattern recognition technique than LDA. For this reason, the use of appropriate data pre-processing is essential in order to obtain acceptable chemical data classification by using SIMCA.

The results obtained in this study indicate that the half-range central value transformation appears to be an excellent pre-treatment approach to avoid data noise. HR-CV rearranges the data set in such a way that the average values become zero, meanwhile extreme values became absolute maximums. This means that the procedure enhances the extreme cases, *i.e.*, those that offer the maximum variability (differences) among samples. As the mean values tend towards zero, the information for samples/variables that lies in the average is neglected. This is an advantage over the minimum–maximum transformation, a technique that also enhances the differences but does not cancel the average values.

Drift diagnosis in ICP-AES: long-term stability

Fig. 3 shows the intensities of a range of emission lines studied over a period of 8 h using both standard and robust conditions. It can be seen that the trends are slightly different for both operating conditions.[19,20]

Table 5 Percentage of explained variability as accounted for by the first principal component

Data pre-treatment	Variability accounted for by the first PC (%)
Standard conditions	
Autoscaling	85.2
Mean normalisation	100.0
Maximum normalisation	100.0
Range normalisation	100.0
Minimum–maximum transformation	97.0
Half-range central value transformation	89.3
Robust conditions	
Autoscaling	86.1
Mean normalisation	100.0
Maximum normalisation	100.0
Range normalisation	100.0
Minimum–maximum transformation	77.8
Half-range central value transformation	87.2

A PCA, using cross validation for validation, was performed on the two data sets (intensities of 29 emission lines each measured every 4–5 min over 8 h) representing both sets of conditions. Ten PCs were calculated following each of the data pre-treatments. It was found that the first PC explained a variance percentage higher than 95% when using normalisation (mean, maximum, or range normalisation) as data pre-treatments for the two data sets (Table 5). However, percentages lower than 90% are explained when using other forms of data pre-processing. This was expected due to the high degree of homogenisation associated with normalisation pre-treatments. It also appears that the variability in the data for the robust conditions set is less well explained than for the standard conditions using the minimum–maximum and half-range central value transformations (Table 5). This is clearly due to the presence of two different trends in the robust conditions data set (Fig. 3). The choice of data pre-treatment in these data sets will condition the type of loadings obtained after PCA. Figs. 4 and 5 show the loadings over the first principal component when using each type of data pre-treatment for both operating conditions. It can be seen that the loadings obtained after the application of the mean and maximum normalisation and the minimum–maximum transformation data pre-treatments are similar to trends shown in the drift phenomena (Fig. 3). Therefore, these data pre-treatments must be used before the application of PCA to these data sets in order to study drift phenomena. However, the three other data pre-treatments partially modified the long-term patterns. In fact, data pre-treatments involving mean subtraction, as is the case in half-range central value and autoscaling, are not adequate for these data sets. This is because mean subtraction

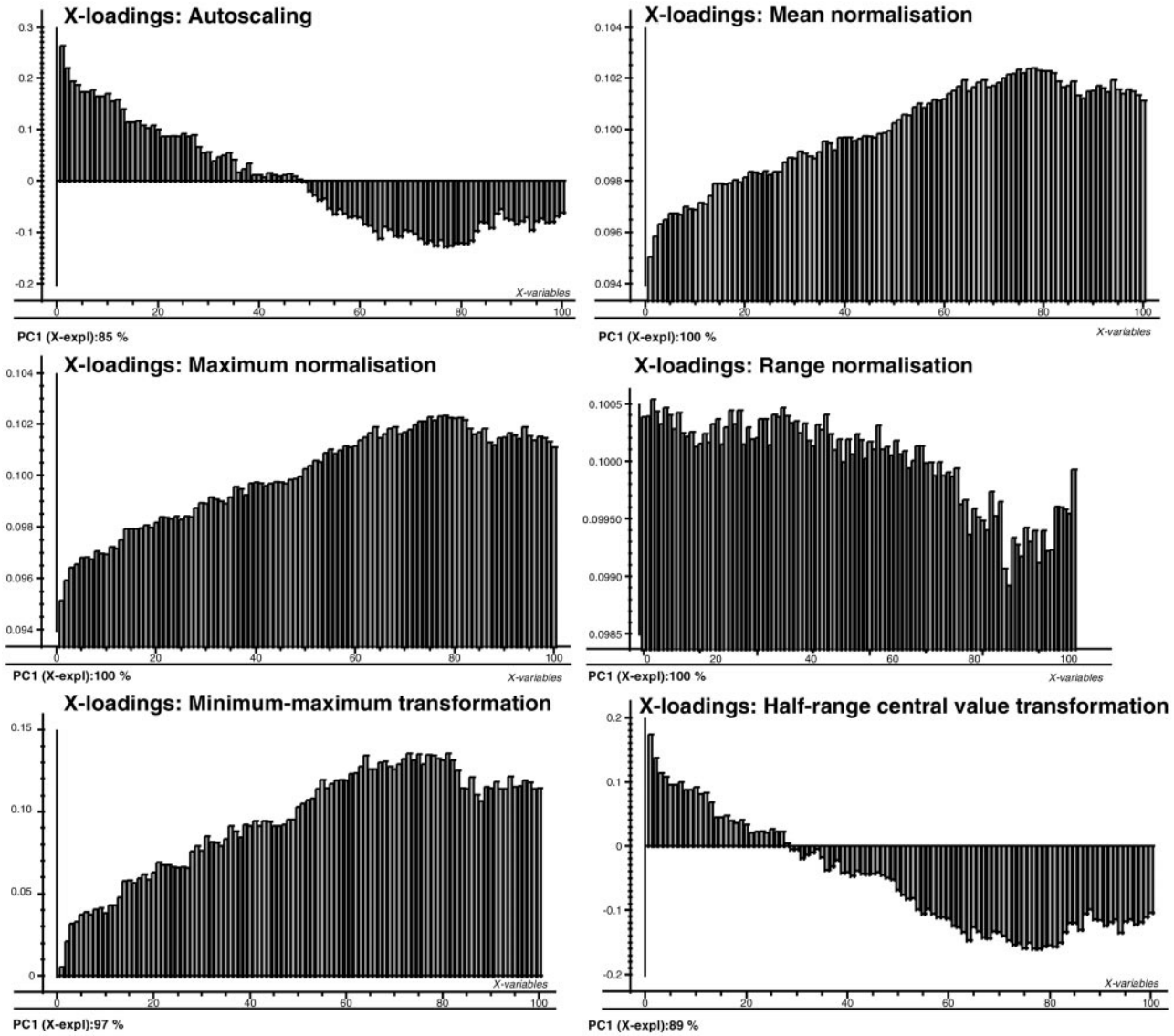

Fig. 4 Loadings of the emission intensities under standard conditions along the first principal component after the different data pre-treatments.

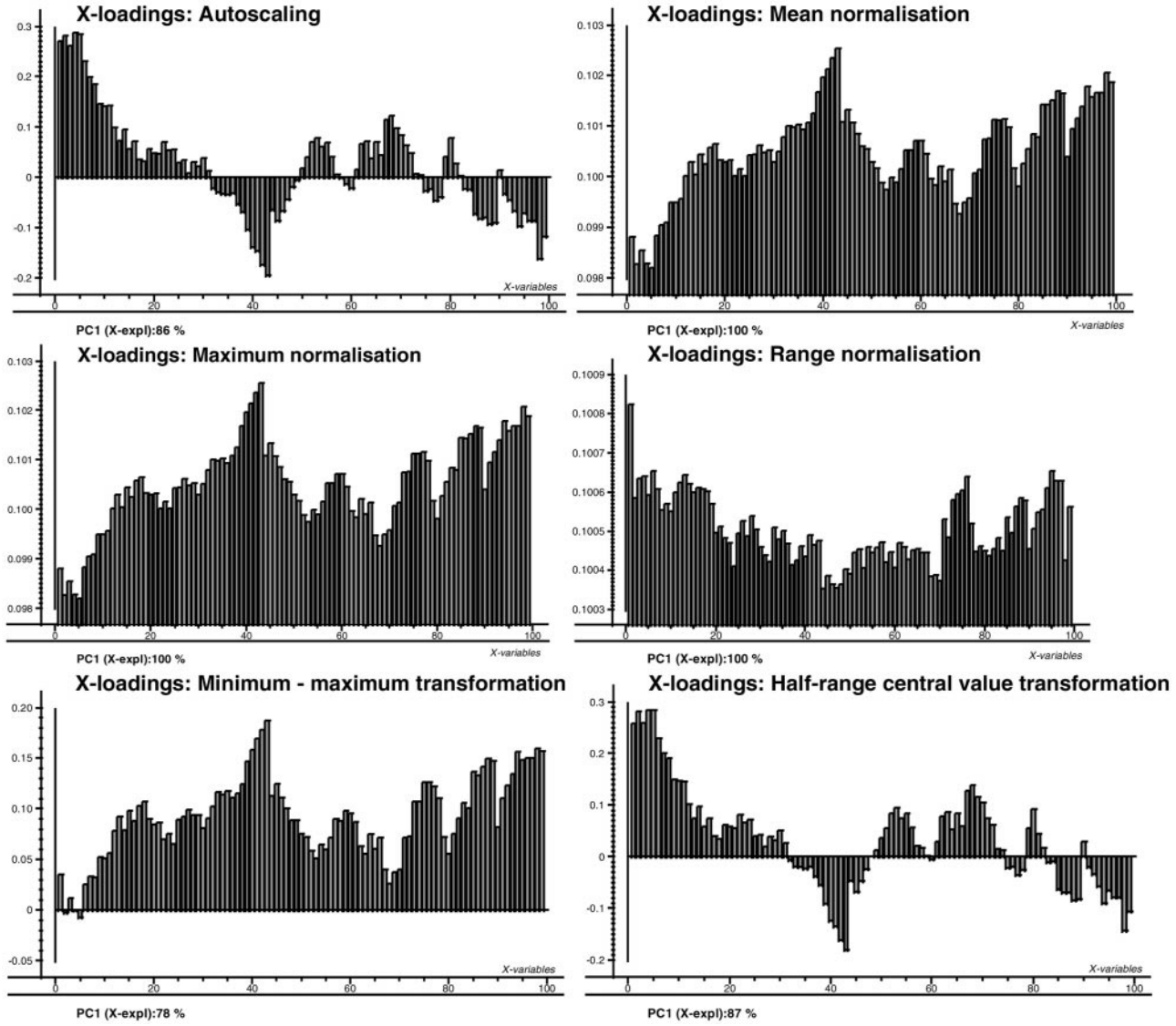

Fig. 5 Loadings of the emission intensities under robust conditions along the first principal component after the different data pre-treatments.

inverses the data trend giving high loading to the initial measurements where no drift bias is recorded and low loading where high drift error is observed (Fig. 4).

Conclusions

The results presented here emphasise the need to use an appropriate data pre-treatment to homogenise data prior to using chemometric tools. It has been shown that approaches such as PCA and related models such as SIMCA offer different results when different data pre-treatments are used prior to modelling. However, the effect of the type of data pre-processing is insignificant when using CA and LDA. Finally, it is important to choose an appropriate data pre-treatment approach relevant to the problem under study, in order to enhance the information that needs to be modelled. SIMCA is a more restrictive method than LDA and a good classification of different classes or categories is only obtained if a satisfactory SIMCA classification is reached. It is evident that the application of half-range central value transformation has been found to be a very useful data pre-treatment method, since successful classification was not possible using the alternative data pre-processing methods studied.

Acknowledgements

AMP would like to acknowledge the financial support provided by MEC "Ministerio de Educación y Cultura", Spain, for a "programa de becas de formación de personal investigador en el extranjero" post-doctoral grant.

References

1 S. D. Brown, R. S. Bear and T. B. Blank, *Anal. Chem.*, 1992, **64**, 22R.
2 R. G. Brereton, *Chemometrics - applications of mathematics and statistics to laboratory systems*, Ellis Horwood Limited, West Sussex, UK, 1990, ch. 7, pp 241–244.
3 The Unscrambler User Manual, CAMO ASA, Trondheim, Norway, 1998, ch. 2, pp 45–47.
4 K. de Braekeleer, R. de Maesschalck, P. A. Hailey, D. C. A. Sharp and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 1999, **46**, 103.
5 W. Wu, Q. Guo, D. Jouan-Rimbaud and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 1999, **45**, 39.
6 A. Garrido-Frenich, M. Martínez-Galera, J. L. Martínez-Vidal and M. D. Gil-García, *J. Chromatogr.*, 1996, **727**, 27.
7 P. R. Griffiths, *J. Near Infrared Spectrosc.*, 1995, **3**, 60.
8 A. Donachie, A. D. Walmsley and S. J. Haswell, *Anal. Commun.*, 1996, **33**, 293.
9 M. Blanco, J. Coello, H. Iturriaga, S. Maspoch and C. de la Pezuela, *Appl. Spectrosc.*, 1997, **51**, 240.

10 J. G. Sun, *J. Chemom.*, 1997, **11**, 525.
11 M. Rupprecht and T. Probst, *Anal. Chim. Acta*, 1998, **358**, 205.
12 Y. Ootake and S. Kokot, *J. Near Infrared Spectrosc.*, 1998, **6**, 251.
13 A. Donachie, A. D. Walmsley and S. J. Haswell, *Anal. Chim. Acta.*, 1999, **378**, 235.
14 N. M. Faber, *Anal. Chem.*, 1999, **71**, 557.
15 Q. Ding, G. W. Small and M. A. Arnold, *Appl. Spectrosc.*, 1999, **53**, 402.
16 J. W. Einax, H. W. Zwanziger and S. Geiβ, *Chemometrics in Environmental Analysis*, VCH, Weinheim, Germany, 1997, ch. 5, pp. 140–144.
17 J. W. Einax, H. W. Zwanziger and S. Geiβ, *Chemometrics in Environmental Analysis*, VCH, Weinheim, Germany, 1997, ch. 3, pp. 78–79.
18 A. Moreda-Piñeiro, A. Marcos, A. Fisher and S. J. Hill, *J. Anal. At. Spectrom.*, 2001, **16**, 350.
19 A. Marcos and S. J. Hill, *Analyst*, 2000, **125**, 1015.
20 A. Marcos, M. Foulkes and S. J. Hill, *J. Anal. At. Spectrom.*, 2001, **16**, 105.
21 A. Marcos, A. Fisher, G. Rea and S. J. Hill, *J. Anal. At. Spectrom.*, 1998, **13**, 521.