ELSEVIER

PII: S0308-8146(97)00103-9

Food Chemistry, Vol. 61, No. 1/2, pp. 207–213, 1998

0308-8146/98 $19.00 + 0.00

Application of chemometrics to the ^{1}H NMR spectra of apple juices: discrimination between apple varieties

Peter S. Belton,[a] Ian J. Colquhoun,[a]* E. Katherine Kemsley,[a] Ivonne Delgadillo,[b] Paula Roma,[b] M. John Dennis,[c] Matthew Sharman,[c] Elaine Holmes,[d] Jeremy K. Nicholson[d] & Manfred Spraul[e]

[a]*Institute of Food Research, Norwich Laboratory, Norwich Research Park, Colney, Norwich NR4 7UA, UK*
[b]*Department of Chemistry, University of Aveiro, Aveiro 3800, Portugal*
[c]*CSL Food Science Laboratory, Norwich Research Park, Colney, Norwich NR4 7UQ, UK*
[d]*Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London WC1H 0PP, UK*
[e]*Bruker Analytische Messtechnik GmbH, Silberstreifen, D76287-Rheinstetten, Germany*

(Received 23 February 1997; accepted 26 March 1997)

Discrimination between apple juices produced from different varieties (Spartan, Bramley, Russet) has been achieved by applying principal components analysis (PCA) and linear discriminant analysis to ^{1}H NMR spectra of the juices. The use of covariance and correlation matrix PCA methods was investigated and different regions of the spectrum were analysed in view of the large range of signal intensities. All the methods gave a high success rate of classification, with at least 24 out of 26 samples being correctly assigned when five principal components were used. Under optimum conditions a 100% success rate was achieved. Examination of the principal component loadings showed that the levels of malic acid and sucrose were two important chemical variables, but variations in the composition of the minor constituents were also found to make a significant contribution to the discrimination. © 1998 Elsevier Science Ltd. All rights reserved

INTRODUCTION

For the analysis of complex mixtures, high resolution techniques such as nuclear magnetic resonance (NMR) and Fourier transform infrared (FTIR) spectroscopy have the attractive feature that qualitative and quantitative information may be obtained on a wide range of chemical species in a single experiment. The speed with which spectra can be obtained, often with minimal sample preparation, makes examination of many samples possible, as required for most food composition, authenticity and quality control applications. Without appropriate methods of data analysis, however, the spectroscopic detail which potentially makes these techniques so powerful would become overwhelming. To overcome this problem methods of multivariate statistics are beginning to be applied directly to spectral data in order to address questions of classification and discrimination. These methods are particularly appropriate to areas such as food analysis (Vogels *et al.*, 1993; Lai *et al.*, 1994; Defernez *et al.*, 1995) and clinical chemistry (Howells *et al.*, 1992; Holmes *et al.*, 1994), where successful classification or diagnosis depends on simultaneous consideration of many variables. There is increasing interest in applying this approach to fruits and their products. It has been shown (Belton *et al.*, 1996) that high-field ^{1}H NMR spectra of fruit juices can be obtained rapidly and directly, that suppression of the water signal is not a problem, and that numerous components can be identified and potentially quantified. Multivariate analysis techniques have been applied to ^{1}H NMR spectra of orange juices in order to detect adulteration (Vogels *et al.*, 1996) and have been applied to 2D NMR spectra of seed and leaf extracts for the differentiation of grape cultivars (Forveille *et al.*, 1996).

There have been many studies on the composition of apple juice from simple measurements of total sugars and acids in relation to sensory qualities, to quite sophisticated determinations of amounts of individual sugars (Fuleki *et al.*, 1994), non-volatile acids (Fuleki *et al.*, 1995) and phenolics (Lee and Wrolstad, 1988). The aim of most of this more detailed work has been to establish data bases for the detection of adulteration and fraud (Brause and Raterman, 1982; Mattick and Moyer, 1983). It has been well established that there are

*To whom correspondence should be addressed.

wide differences of composition between cultivars, although most surveys have only included one or two examples of each type. The cultivar appears to be a more important source of compositional variations than either the growing region, the year of production, or the length of storage, although storage does lead to some changes of sugar composition.

In this communication, we explore the potential of ^{1}H high-resolution NMR spectroscopy combined with multivariate data analysis to distinguish apple juices produced from three different cultivars. Most of the samples were prepared in the laboratory using apples obtained directly from English orchards. One or two additional samples were commercially produced juices.

MATERIALS AND METHODS

The apples (cultivars Bramley, Russet and Spartan) were collected from orchards in various English counties in October 1994. Juices (about 1 litre from 5–10 kg of whole apples) were prepared at that time using an electrical mincer (Butcherboy) and were filtered through a muslin bag fitted to the output nozzle of the mincer. The juices were stored at -20°C for 6 months before samples were taken for the NMR experiments. A small quantity of each sample was centrifuged and 0.1 ml D_2O (for NMR field/frequency lock) was added to 1 ml of each supernatant. A 0.1% solution of TSP (sodium salt of 3-(trimethylsilyl)propionic-*2,2,3,3-d*$_4$ acid) in D_2O was added to 0.6 ml of each supernatant in a 5 mm o.d. NMR tube, to give a final TSP concentration of 0.01%.

500 MHz ^{1}H NMR spectra were recorded at 27°C using a Bruker DRX-500 spectrometer equipped with an automatic sample changer. One hundred and sixty scans of 64 K data points were acquired with a spectral width of 8333 Hz, acquisition time of 3.93 s, and recycle delay of 2 s. Water suppression was achieved using the NOESY-presaturation pulse sequence (Hull, 1994) with irradiation at the water frequency during the recycle and mixing time (0.08 s) delays. The data were acquired under an automation procedure, requiring about 16 min per sample.

Preliminary data processing was carried out with FELIX software, version 2.30 (Molecular Simulations) running on a Silicon Graphics Indigo workstation. The FIDs were Fourier transformed (1 Hz line broadening) and the spectra were phased, baseline corrected, reduced to the real part and compressed to 4 K data points by averaging successive pairs of points (three steps were required to reduce the data size from 32 K to 4 K real points). The resulting spectra were then aligned by right or left shifting as necessary (using the TSP signal as reference), saved as ASCII files and transferred to a PC for data analysis.

Three data matrices were constructed, each containing spectra of the 26 apple juices but for three separate regions: the 'full' spectrum (4.55–0.8 ppm, 860 points), region 1 (4.55–2.5 ppm, 482 points) and region 2 (2.5–0.8 ppm, 378 points). Thus, the dimensions of the three matrices were 860×26, 482×26 and 378×26. The upper limit of 4.55 ppm avoids any influence from the saturating irradiation applied at the water frequency. In all three cases small segments containing the methyl and methylene signals of ethanol were excluded (see below). Principal components analysis (PCA) (Jolliffe, 1986), linear discriminant analysis (LDA) (Massart *et al.*, 1988), and canonical variates analysis (CVA) (Krzanowski, 1988) were carried out using the program Win-Discrim (K.Kemsley, Institute of Food Research). PCA entails the computation of eigenvectors of a matrix $(X^TX)/(n-1)$, where X contains n observations entered into the matrix row-wise. Different variations of PCA can be performed, by varying the nature of the data in X. If X is mean-centred (column means subtracted), then $(X^TX)/(n-1)$ is a covariance matrix: we will call this form of PCA the 'covariance method'. If the data in X are standardised (mean-centred and columns scaled to unit variance), then $(X^TX)/(n-1)$ is a correlation matrix, and this form of PCA will be termed the 'correlation method'. An advantage of the covariance method is that the eigenvectors (or 'loadings') retain the scale of the original data, and will often resemble spectra which can be interpreted by the spectroscopist. In contrast, the loadings obtained by the correlation method are usually very unfamiliar in appearance. However, an advantage of this approach is that the PCA is influenced by all spectral features equally, whereas in the covariance method, larger bands tend to dominate. Consequently, the correlation method can be useful when minor constituents, with small spectral contributions, are of primary interest. In the present work, both PCA methods have been performed. The Win-Discrim software uses the NIPALS algorithm (Martens and Naes, 1989) to calculate loadings and scores for the first 15 principal components.

RESULTS AND DISCUSSION

Spectra

Table 1 shows the number of juices examined and the average pH for each of the cultivars. The pH of the samples was measured, but not adjusted, prior to recording the NMR spectra. ^{1}H NMR spectra of three of the samples (one for each cultivar) are shown in Fig. 1. The figure shows the spectral region between 0.8 and 5.5 ppm, with the vertical gain increased by a factor of 100 to display signals from species of low concentration in Fig. 1(b). The region between 6.0 and 8.0 ppm contains a number of signals from aromatic protons of phenolic constituents. The region was not, however, included in the data analysis, since no precautions were taken to prevent enzymic oxidation of the phenolics during preparation of the juices.

Table 1. Number of samples and average composition of juice from three English apple varieties

Variety	*n*	pH	Integral ratios[a,b]			
			MA/T	Suc/T	Fru/T	Glc/T
Spartan	9	4.34 ± 0.17	0.09 ± 0.02	0.11 ± 0.07	0.65 ± 0.03	0.24 ± 0.04
Bramley	10	3.51 ± 0.22	0.28 ± 0.07	0.18 ± 0.08	0.57 ± 0.04	0.25 ± 0.04
Russet	7	4.05 ± 0.14	0.12 ± 0.03	0.27 ± 0.08	0.51 ± 0.03	0.22 ± 0.05

[a]Not molar ratios, see text.
[b]MA, malic acid; Suc, sucrose; Fru, fructose; Glc, glucose; T, (Suc + Fru + Glc).

The region between 2.5 and 5.5 ppm (Fig. 1(a)) contains signals from the components of highest concentration: sucrose, fructose, glucose and malic acid. Signals from minor components such as sorbitol, quinic acid, citric and aspartic acids are also expected in this region, but have not yet been specifically identified. Two features which are clear from Fig. 1(a) are the dependence of the malic acid chemical shifts on pH (the sugar resonances are not very sensitive) and the inter-cultivar differences in the relative concentrations of malic acid and the individual sugars. As a rough guide to these variations, selected signals (indicated in Fig. 1(a)) associated with major components were integrated and their ratios compared. For malic acid the multiplet at ~2.8 ppm was used. Average values of the ratios are given in Table 1. It should be noted that these are not molar ratios, since for glucose and fructose the signals selected arose from only one of the anomeric forms. There was considerable variation in the ratios between samples of a given cultivar, but the spectra displayed in Fig. 1 were chosen to best represent the average values. The high-field region of the spectrum (0.8–2.5 ppm) contains signals from many species including quinic acid, ethanol, lactic and acetic acids. The last three are normally present at very low levels in freshly squeezed apple juice. Ethanol levels in some of the samples were unusually high, probably as a consequence of microbial degradation during the production/storage period. The small segments of the spectrum containing the methyl and methylene signals of ethanol were therefore excluded

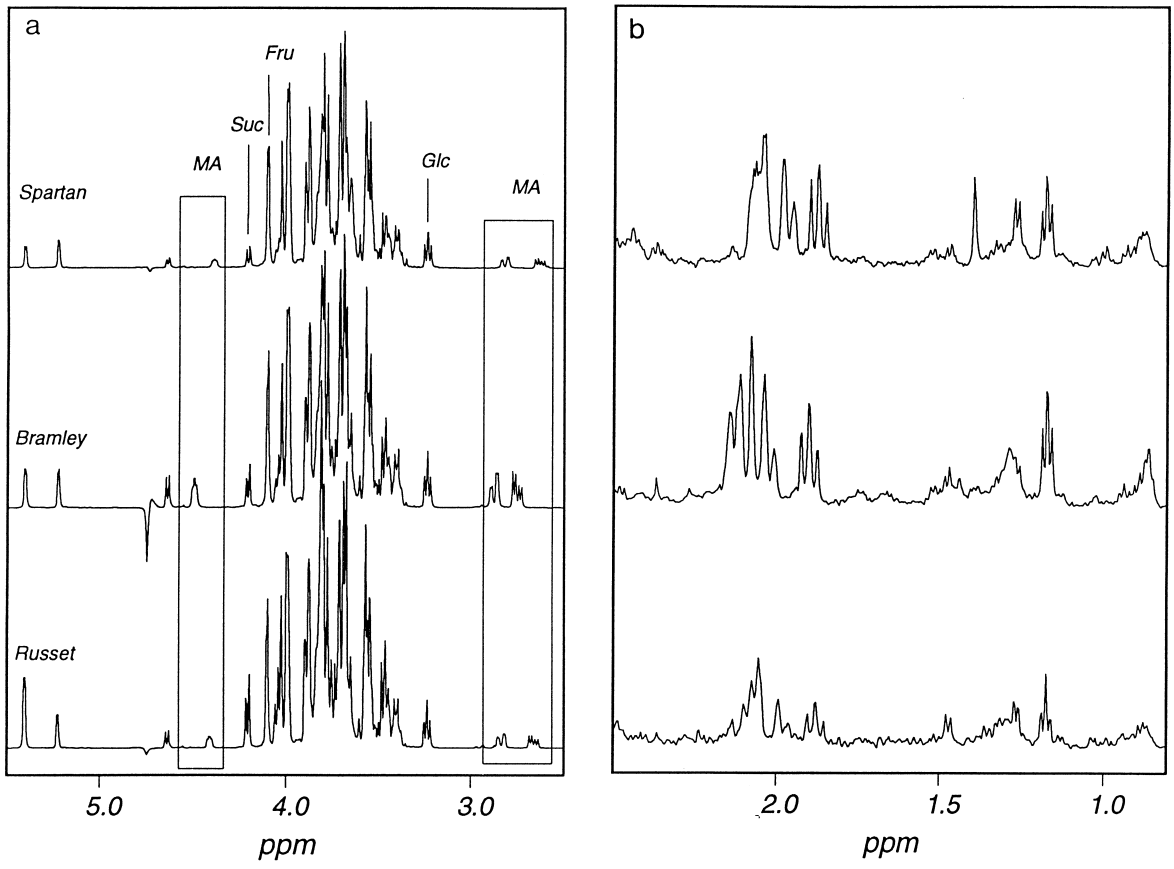

Fig. 1. Five hundred megaHertz ^{1}H NMR spectra of apple juices: (a) 5.5–2.5 ppm; (b) 2.5–0.8 ppm. In (b) the vertical gain has been increased by a factor of 100. Data matrices were constructed for the 'full' spectrum (4.55–0.8 ppm), region 1 (4.55–2.5 ppm) and region 2 (2.5–0.8 ppm). MA, malic acid; Suc, sucrose; Fru, fructose; Glc, glucose signals.

from the data analysis, although the regions containing the much weaker acetate and lactate resonances were included.

Data analysis

In view of the large differences in signal intensity between the regions of the spectrum shown in Figs 1(a) and (b) it was decided to compare the correlation and covariance PCA methods. The covariance method is influenced most by stronger signals so that PCA of the full spectrum is minimally affected by variations in the 0.8–2.5 ppm region, even though this region was found to possess considerable discriminatory power (see below).

Figure 2 shows plots of the PC scores for the most important PCs (PC1 vs PC2 and PC1 vs PC3) and compares results of covariance and correlation method PCAs on the full spectra. Clustering according to variety is evident in all the plots but somewhat better separation between groups is evident for the correlation method. Scores on PC1 lead to separation between Spartan apples and the other two types (Fig. 2(a)), whilst PC3 (Fig. 2(b)) shows clearer distinction between Bramley and Russet cultivars. The scatter within the individual groups is not surprising, given the different locations from which the apples were obtained, possible variations in ripeness and the changes which may have taken place between juice extraction and sample measurement.

Although Fig. 2 provides clear indications of sample clustering, the plots are based only on that part of the information contained in the first three PCs. Figure 3 illustrates how the cumulative percentage of variance explained depends on the number of PCs, as this number increases up to 15 (the maximum calculated). There is an obvious difference between the two curves obtained from use of the covariance and correlation methods. In the case of the covariance method, the first three or four PCs account for a high proportion of the total variance, whereas the correlation method gives a curve showing a more gradual approach to 100% variance explained. For data of the type shown in Fig. 1, with a very wide range of signal intensities, it is expected that more chemical variables will become significant when the correlation method is used. In particular, variations in the concentrations of the minor components (high field part of the spectrum) will be relatively more important.

Whilst the plots of Fig. 2 provide a useful visual impression, the likely success rate of the PC-based model in correctly assigning future unknown samples is

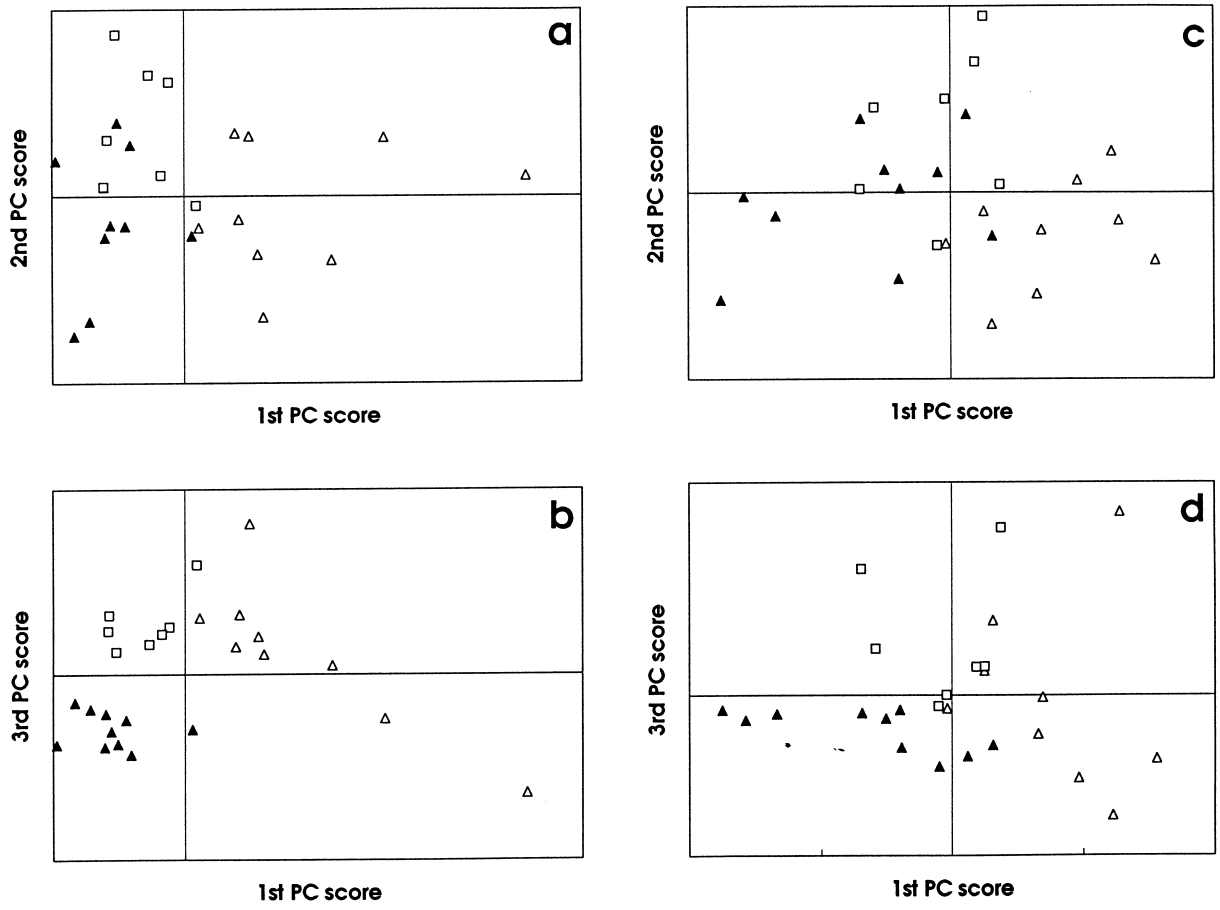

Fig. 2. PC scores from principal component analysis of the full spectrum: (a),(b) correlation method; (c),(d) covariance method. Apple varieties: △, Spartan; ▲, Bramley; □, Russet.

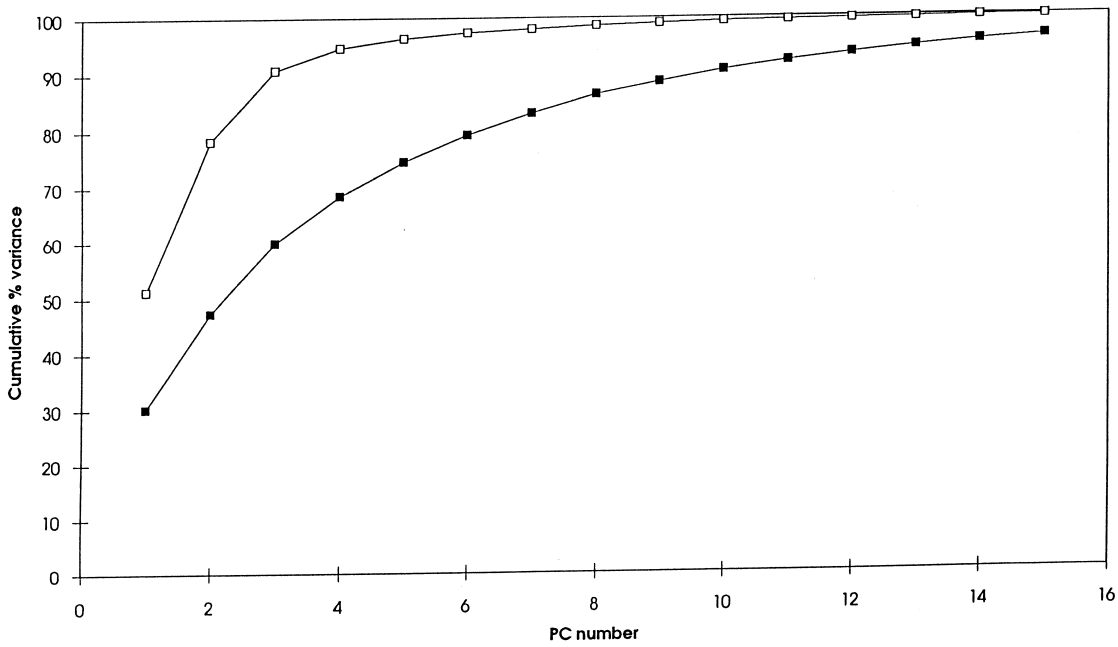

Fig. 3. Cumulative percentage of variance explained in relation to the number of PCs used: □, covariance; ■, correlation.

more effectively gauged by a linear discriminant analysis procedure. In LDA, the samples of known origin are first assigned to groups (here, according to cultivar). The positions of the group means, in the multi-dimensional PC space are then calculated from the scores using a chosen number of PCs. To avoid over-fitting, the number of PCs should be substantially fewer than the total number of samples and preferably smaller than the size of the smallest group. Using the squared Mahalonobis distance (Jolliffe, 1986; Massart *et al.*, 1988), the nearest group mean to each sample is then identified, the samples are re-classified into groups according to this criterion and the number of correct assignments is determined. This procedure allows an objective comparison of the success rate achieved in classification, not only between covariance and correlation methods, but also between data matrices based on the full spectrum and on regions 1 and 2 taken separately.

The results, presented in Table 2, show that a high success rate is achieved with three or four PCs, whichever method or spectral region is used. It is interesting to note that with the same number of PCs, the success rates are identical for the full spectrum and for region 1 when covariance is used, whereas use of correlation leads to some differences. This implies that, for covariance, only the strong signals (sugars, malic acid) contribute to the discrimination when the full spectrum is treated, and the inclusion of the region containing weaker signals is irrelevant. For region 2 taken on its own, however, the results are comparable with those obtained from region 1 or from the full spectrum. This shows that the composition of minor components is potentially as good an indicator of cultivar type as the composition of major components. For adulteration purposes it would obviously be much harder to tamper with the composition of a large number of minor species than with that of three or four readily available major components.

Table 2. Success rate of discriminant analysis procedure employing different spectral regions and PCA methods

Region (ppm)	Method	Samples correctly assigned (out of 26) No of PCs					
		1	2	3	4	5	6
4.55–0.8	Covariance	17	19	24	23	25	26
	Correlation	20	20	24	24	26	26
4.55–2.5	Covariance	17	19	24	23	25	26
	Correlation	15	21	25	25	24	25
2.5–0.8	Covariance	20	22	23	24	24	25
	Correlation	16	19	23	26	26	26

Canonical variates analysis is a further discriminatory technique which can be applied to the PC scores. It is a data rotation method which involves finding a linear combination of a subset of the PC scores which maximises the ratio of between-groups to within-groups variance, that is it simultaneously maximises the distances between the groups and minimises the scatter within each group. Since CVA is a non-rigid rotation method, care must be taken to avoid over-fitting. In view of the relatively small size of our data set, we have restricted the subset of PCs to a maximum of 5. Plots of the canonical variate scores give an optimal impression of the capacity of the data to discriminate between groups (Fig. 4). The plot in Fig. 4, which uses the first five PCs, may be compared with those for pairs of PCs shown in Fig. 2. This technique should be valuable in future work, where the number of cultivars examined will be increased and some of the inter-group differences will

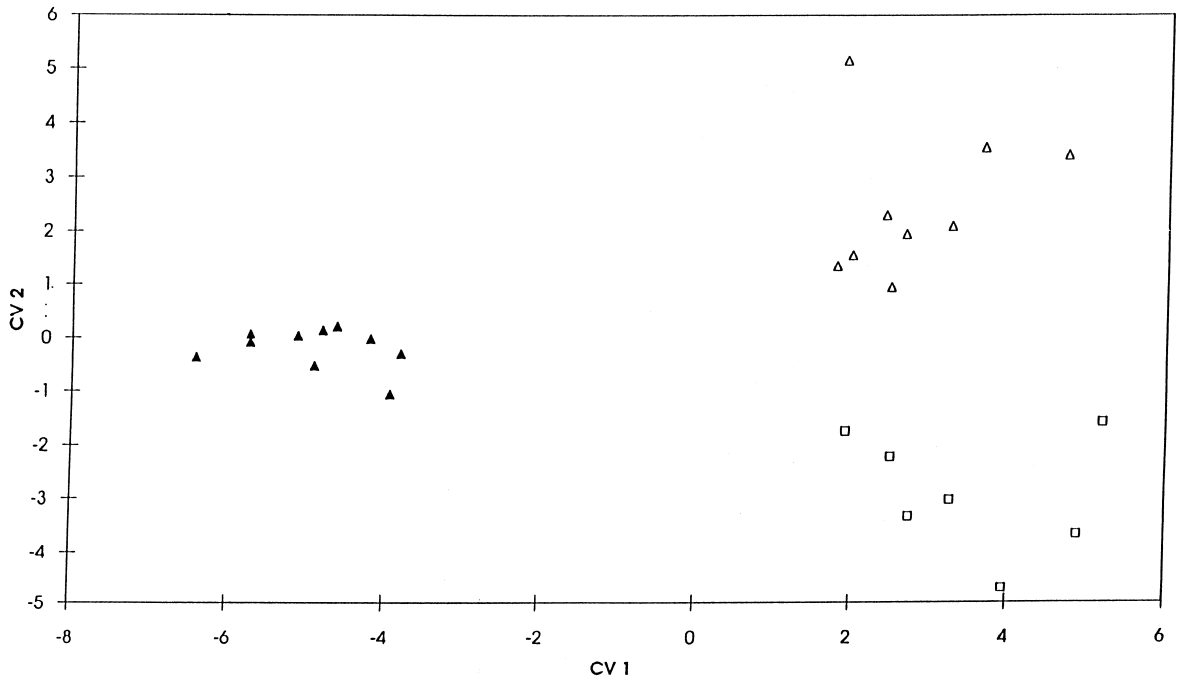

Fig. 4. CV scores from canonical variates analysis (full spectrum, correlation method, first five PCs used). Apple varieties: △, Spartan; ▲, Bramley; □, Russet.

inevitably be smaller than for the three types selected here.

Examination of the eigenvector loadings (Fig. 5), together with the original spectra (Fig. 1) can be useful in understanding the basis of the clustering behaviour. Figure 5 shows the contrast between loadings obtained from PCA of the full spectra when covariance and correlation methods are used. Only in the latter case do the weaker signals in region 2 make an appreciable contribution. Loadings 1, 2 and 3 (Figs 5(a)–(c)) should be considered in relation to PC scores plots in Figs 2(c) and (d) since these plots result from PCA of the covariance matrix. For loading 1, associated with PC1, the only negative values appear at positions corresponding to the malic acid signals in the low pH samples (essentially the Bramley apples). Positive values are strongest at positions corresponding to fructose and glucose (but not sucrose), as well as the malic acid positions in juices of higher pH (Spartan, Russet). The difference in malic acid chemical shifts is therefore one parameter responsible for the separation along PC1, with the Bramley apple scores lying towards the negative side of this axis. Although the spread of pH values makes this particularly evident, one would also expect the difference in malic acid levels (Table 1) to be important even if the pH of all the samples (and hence the malic acid chemical shifts) had been made the same. Loading 2 is not so readily interpretable, but in loading 3 the positions associated with sucrose are positive and, compared with loading 1, are stronger relative to fructose and glucose (which are both negative). In accord with this, the scores on PC3 tend to be more positive for the Russet juices, i.e. those with the highest proportion of sucrose

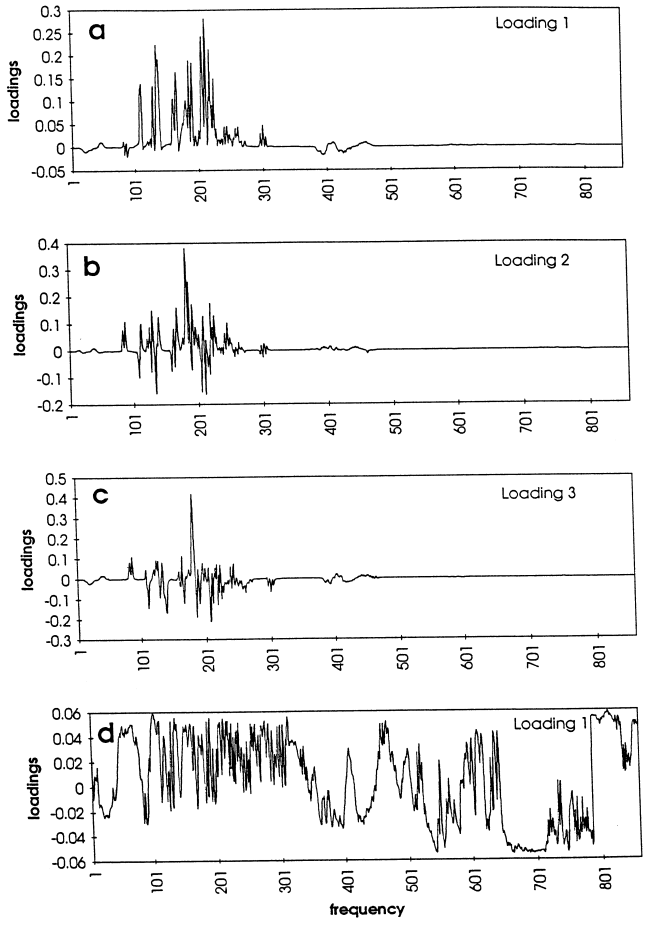

Fig. 5. Eigenvector loadings (full spectrum): (a)–(c) covariance; (d) correlation. Full spectrum is 860 data points, region 1 corresponds to points 1–482, region 2 to points 483–860.

(Table 1). Each of the eigenvector loadings in fact represents a combination of chemical variables, and interpretations such as that given above are greatly oversimplified. Nevertheless, it is useful to have some indication of the chemical basis for the discrimination. Such information should be particularly interesting with regard to the loadings for region 2, when assignment of the many signals in this part of the spectrum has been accomplished by 2D NMR methods.

CONCLUSIONS

Multivariate statistical methods have been used to analyse high resolution proton NMR spectra of apple juices of known origin. Samples of the same variety showed clustering in PC space and a discriminant analysis procedure gave a high success rate in assigning samples to the correct groups. The PC loadings were examined and different regions of the spectrum were treated independently in order to learn something about the chemical basis for the clustering behaviour. Although the major components were important, it was apparent that minor constituents could also contribute significantly to the discrimination under certain conditions. This has been a preliminary investigation and the number of cultivars, as well as the number of samples, within each group will need to be increased considerably in order to make a proper evaluation of the scope of the method. The speed of the measurements and the potential for further automation, however, make the approach an attractive one for further studies of fruit juice authenticity and quality.

ACKNOWLEDGEMENTS

Funding by the BBSRC Competitive Strategic Grant (PSB, IJC, EKK) and the EU COMETT programme (PR) is gratefully acknowledged.

REFERENCES

Belton, P. S., Delgadillo, I., Holmes, E., Nicholls, A., Nicholson, J. K. and Spraul, M. (1996) Use of high -field NMR spectroscopy for the analysis of liquid foods. *Journal of Agricultural and Food Chemistry* **44**, 1483–1487.

Brause, A. R. and Raterman, J. M. (1982) Verification of authenticity of apple juice. *Journal of the Association of Official Analytical Chemists* **65**, 846–849.

Defernez, M., Kemsley, E. K. and Wilson, R. H. (1995) Use of infrared spectroscopy and chemometrics for the authentication of fruit purees. *Journal of Agricultural and Food Chemistry* **43**, 109–113.

Forveille, L., Vercauteren, J. and Rutledge, D. N. (1996) Multivariate statistical analysis of two- dimensional NMR data to differentiate grapevine cultivars and clones. *Food Chemistry* **57**, 441–450.

Fuleki, T., Pelayo, E. and Palabay, R. B. (1994) Sugar composition of varietal juices produced from fresh and stored apples. *Journal of Agricultural and Food Chemistry* **42**, 1266–1275.

Fuleki, T., Pelayo, E. and Palabay, R. B. (1995) Carboxylic acid composition of varietal juices produced from fresh and stored apples. *Journal of Agricultural and Food Chemistry* **43**, 598–607.

Holmes, E., Foxall, P. J. D., Nicholson, J. K., Neild, G. H., Brown, S. M., Beddell, C. R., Sweatman, B. C., Rahr, E., Lindon, J. C., Spraul, M. and Neidig, P. (1994) Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal Biochem* **220**, 284–296.

Howells, S. L., Maxwell, R. J., Peet, A. C. and Griffiths, J. R. (1992) An investigation of tumor ^{1}H nuclear magnetic resonance spectra by the application of chemometric techniques. *Magn Reson Med* **28**, 214–236.

Hull, W. E. (1994) In: *Two Dimensional NMR Spectroscopy. Applications for Chemists and Biochemists* 2nd edn. VCH, New York. pp. 125.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer, New York, US, Chapters 1–4, pp. 1–49.

Krzanowski, W. J. (1988) *Principles of Multivariate Analysis: A User's Perspective*. Clarendon Press, Oxford, UK, chapter 11, pp. 292–309.

Lai, Y. W., Kemsley, E. K. and Wilson, R. H. (1994) Potential of FTIR spectroscopy for the authentication of vegetable oils. *Journal of Agricultural and Food Chemistry* **42**, 1154–1159.

Lee, H. S. and Wrolstad, R. E. (1988) Apple juice composition: sugar, nonvolatile acid and phenolic profiles. *Journal of the Association of Official Analytical Chemists* **71**, 789–794.

Martens, H. and Naes, T. (1989) *Multivariate Calibration*. Wiley, Chichester, UK, pp. 111–112.

Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y. and Kaufman, L. (1988) *Data Handling in Science and Technology* (Vol 2: Chemometrics, A Textbook), ed. B. G. M. Vandeginste and I. Kaufman. Elsevier, Amsterdam, The Netherlands, pp. 386–388.

Mattick, L. R. and Moyer, J. C. (1983) Composition of apple juice. *Journal of the Association of Official Analytical Chemists* **66**, 1251–1255.

Vogels, J. T. W. E., Tas, A. C., Van den Berg, F. and Van der Greef, J. (1993) A new method for classification of wines based on proton and carbon-13 NMR spectroscopy in combination with pattern recognition techniques. *Chemometrics Intell Lab Syst : Lab Inf Manage* **21**, 249–258.

Vogels, J. T. W. E., Terwel, L., Tas, A. C., Van den Berg, F., Dukel, F. and Van der Greef, J. (1996) Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques. *Journal of Agricultural and Food Chemistry* **44**, 175–180.