Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Phytochemistry 62 (2003) 1009–1017
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Factors affecting the robustness of metabolite fingerprinting using ^{1}H NMR spectra

Marianne Defernez*, Ian J. Colquhoun

Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

Received 11 October 2002; received in revised form 25 November 2002

Abstract

^{1}H NMR spectroscopy is one of the techniques whose potential is currently being explored in the emerging field of metabolomics. It is a non-targeted method, producing signals for all proton-containing chemical species. For crude plant materials the spectra are always complex, with many signals overlapping. Hence a most suitable approach for analysing them is 'metabolite fingerprinting', which is aimed at highlighting compositional similarities and exploring the overall natural variability in a population of samples. The most commonly used method for this is principal component analysis (PCA), as it allows the whole spectral trace to be analysed and the vast quantity of information to be simplified. In this paper we investigate whether there are factors which may affect the NMR spectra in a way that subsequently decreases the robustness of the metabolite fingerprinting by PCA. Imperfections in the signal registration (i.e. inconsistency of the peak position) are generally detrimental to analysing whole traces by multivariate methods. The sources of such problems are illustrated through specially designed repeatability studies using potato and tomato samples, and the analysis of a tea dataset containing many samples. Careful sample preparation can help to limit peak shifts; for instance here by attempting to control the pH of the extracts. In addition, some compounds are susceptible to interactions affecting their chemical shifts and mathematical alignment of peaks may be necessary. Lastly factors such as resolution can also affect analyses and must be carefully adjusted. Our choice of examples aims to raise awareness of potential problems. We do not question the validity of the NMR approach, but point out those areas where special care may need to be taken.
© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Chemometrics; Data registration; Exploratory data analysis; Instrumental factors; Metabolite profiling; Metabolomics; Multivariate; NMR; Non-targeted analysis; Principal component analysis

1. Introduction

A great deal of effort is currently being devoted to characterising and quantifying, in a systematic manner, the metabolites in plants. ^{1}H NMR is potentially a very useful technique for this, since in principle any chemical species that contains protons gives rise to signals. Hence it may be used as a rapid metabolite profiling technique, able to detect a broad range of metabolites in a non-targeted way. The 'profile' is here an NMR spectrum and comprises peaks which, for a given solvent, appear at characteristic frequencies. In crude biological extracts however, there are many peaks, leading to overlapped signals and making the profiles very complex: these are sometimes referred to as 'fingerprints'. Two-dimensional NMR is often necessary to assign the signals, but one-dimensional ^{1}H NMR spectra have been of great value as shown in many medical applications by Lindon et al. (2001). Fiehn (2002) has also mentioned NMR as a metabolite fingerprinting method for plants, where the aim is to look for compositional similarities and explore the overall natural variability.

In general, visual examination is insufficient to fully assess whole series of such profiles. One possible approach to carry out this analysis is to gather information (e.g. integrate individual peak areas) for as many constituents as possible and study the occurrence of each of these compounds separately, for example by calculating coefficients of variation or carrying out an analysis of variance (ANOVA). This is one of the tools used by Le Gall et al. (submitted for publication). A second possible approach is to look at the metabolite

* Corresponding author. Tel.: +44-1603-255316; fax: +44-1603-507723.
E-mail address: marianne.defernez@bbsrc.ac.uk (M. Defernez).

0031-9422/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0031-9422(02)00704-5

profile as a whole by using multivariate methods, in which an entire set of variables is examined: for example principal component analysis (PCA) (Krzanowski, 1988) is well suited for such exploratory analyses of the metabolite fingerprints. Early examples where this approach was used in medicine for ^{1}H NMR spectral analysis include the work on tumor classification by Howells et al. (1992). A review of the use of pattern recognition methods in this field was published recently by Lindon et al. (2001). In food research such techniques have also been found suitable to analyse NMR spectra, for instance to classify wines (Vogels et al., 1993) and to characterize orange juice (Le Gall et al., 2001). PCA provides a way to summarize the information contained in large sets of spectra. It transforms the initial variables (or 'measurements') into a much smaller set of variables or PC scores. These new variables are combinations of the initial variables but highlight the variance within the dataset and remove redundancies. Successive PCs account for decreasing amounts of variance and most of the information is contained in the first few PCs, hence the ability to reduce the dimensionality of the data. The explanation of what each PC represents in relation to the original measurements lies in the loadings, a set of weights given to each of the original variables. With ^{1}H NMR spectra, PCA could conceivably be applied to a collection of integrated values (each corresponding to a constituent). However because there is considerable signal overlap, it is of great interest to apply data exploration techniques such as PCA to the whole trace. Each variable then corresponds to the signal intensity at one of the discrete chemical shifts; hence there are initially thousands of 'measurements'. Such an approach has several advantages, not least the ability to analyze signals whose integration may present difficulties, and to potentially reveal hidden compounds. For instance Le Gall et al. (2001) found dimethylproline to be a marker compound for pulpwash in orange juice.

As the aim of metabolite fingerprinting is to globally analyse a population of samples and assess between-sample differences, it is important to gauge the precision (i.e. the repeatability) of the method. Measurement errors arise from several contributing factors, in particular instrumental and operator-related. Whilst it is common practice to measure and express the repeatability of measurements in the case of univariate data (i.e. one single characteristic that defines the sample), this is not as straightforward for multivariate measurements. Another important point is that 'lack of repeatability' only has a meaning when the magnitude of the discrepancies that are observed are compared to the effect that we are interested in studying. Therefore it is crucial to study the repeatability in terms of the output produced by the whole procedure, which includes not only the recording of the spectra but also the data analysis. The ability of chemometric approaches to highlight small differences between spectra (Lai et al., 1994) can be to our advantage if the difference originates from sample characteristics of interest, but can be detrimental if due to other, uncontrollable effects (Defernez and Wilson, 1997).

In this paper we investigated the very basis on which metabolite fingerprinting by ^{1}H NMR rests. We examined whether and how the whole procedure from sample preparation to analysis by a data exploration technique (PCA), can be affected by factors unrelated to the sample characteristics of interest. The approach chosen is to carry out the whole procedure for a number of data sets and to examine whether the final outputs—i.e. the PCA scores—are reproducible, and to investigate the reasons behind any deficiency of the repeatability.

2. Results and discussion

2.1. Controlled repeatability study

The first set of experiments was specifically designed to assess the repeatability of an NMR/chemometric approach, and to investigate the influence and relative importance of the factors that may affect the NMR spectra. The experiments included the whole procedure: the sample extraction and preparation, the recording of ^{1}H NMR spectra, and the data exploration by principal component analysis. Three series of NMR spectra (series 1–3), which encompassed different sample types and/or experimental conditions, were recorded. Each series comprised six different samples on which the sample extraction procedure was repeated three times, giving 18 extracts. The recording of the NMR spectra was carried out under automation 'in block' (i.e. all 18 extracts from one series were measured consecutively in an automation 'run'), and this was repeated on separate occasions with these same 18 tubes (4 'runs' for series 1 and 3 'runs' for series 2 and 3). The recording of several such series allowed the study to cover more than one type of sample and more than one operator.

Fig. 1 shows the scores obtained on PC1 and 2 for the PCA of the first series of samples (potato). The major difference that is highlighted by the PCA is the difference between samples, labelled A–F, confirming that even though the samples are all potato tubers cv. Desiree, grown, harvested, and analyzed together, the biggest source of variance arose from sample compositional differences. Repeats of a given sample form well-defined clusters that are mainly separated on PC1, which accounts for 45.8% of the variance in the dataset. The scores suggest that sample B is particularly different from the others, and the spectra reveal that one of sample B' s characteristics is that it has much lower glucose levels than all other samples. On PC2, which

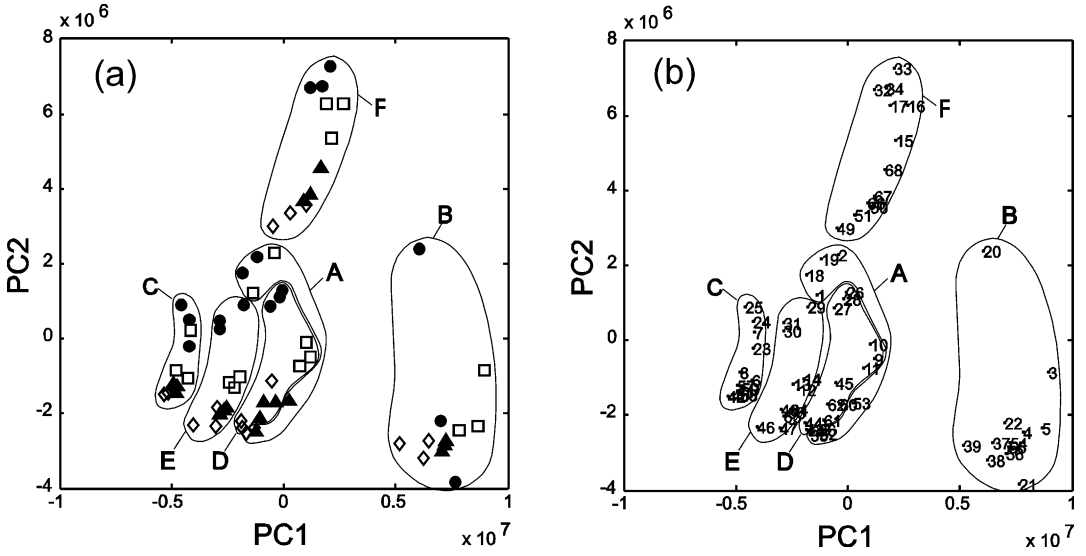

Fig. 1. Scores obtained on PC1 and 2 for the PCA of the first series of samples (Desirée potato), coded A–F, (a) highlighting NMR-run number (□ run 1, ● run 2, ◇ run 3, ▲ run 4) and samples (delimited by outline), (b) highlighting NMR-code number.

accounts for 23.4% of the variance, sample F is the most different. However the NMR 'run' also has a strong influence on this PC. The corresponding loadings reveal that the bandshape of virtually all peaks is related to PC2, because when plotted against variable number (which corresponds to the chemical shift), the peaks had an apparent differential character (in NMR terminology the loadings resemble dispersion rather than absorption curves). For instance for sample B, the spectrum numbered 20 has a much higher score than all others. Examination of the spectra confirms that the peaks of this spectrum have a marked shoulder on the low chemical shift side (Fig. 2). For sample A, the spectra numbered 1, 2, 18, and 19 have higher scores than those numbered 35, 36, 52, and 53. Examination of the spectra showed that the lines of those with high scores also have a marked shoulder on the low chemical shift side. For sample E, the spectra numbered 29, 30, and 31 have higher scores than those numbered 12, 13, and 14, which in turn have higher scores than those numbered 46, 47, 48, 63, 64, and 65. The lines of the spectra numbered 29, 30, and 31 have a slight shoulder on the low chemical shift side. A similar situation also exists for sample D. As a summary, the first run yields higher PC2 scores than the second run, which has higher scores than runs 3 and 4. This is explained by the resolution of runs 1 and 2 being worse than that of runs 3 and 4. In a simple attempt to remedy this deficiency the spectra were deresolved (Lindon et al., 2001) by factors of 2, 6 and 25, but in all cases the influence of the run number still had a marked influence on the PCA.

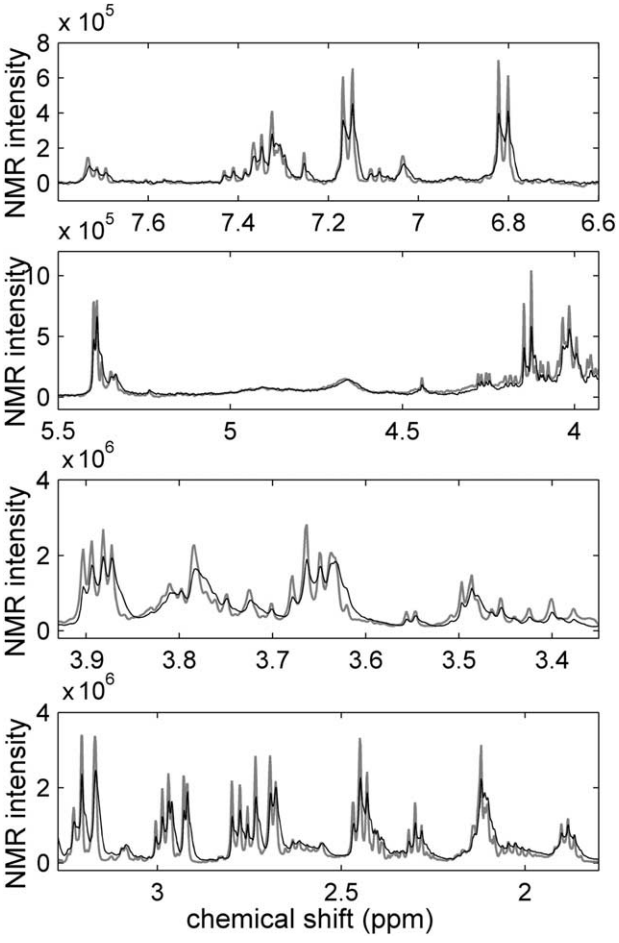

Fig. 2. Selected regions of Desirée potato spectra used in the analysis of the first series (— spectrum numbered 20, — spectrum numbered 21).

With the second series of spectra (tomato), the major difference highlighted by the PCA was again the difference between samples: they were mainly separated on PC1 and PC2 (37.2 and 24.9% variance, respectively). In this case the between-run and between-extraction variances were much lower than the between-sample variance: there was no apparent splitting of the clusters highlighted by PCA based on the two former criteria (Fig. 3). Although these were very satisfactory results, it was noted that the size of the clusters corresponding to each sample differed. Therefore a scaling of the spectra was attempted to alleviate any instrumental factor that could be the source of an amplitude difference. This consisted of normalising each spectrum so that the area under the curve corresponding to the TSP reference peak was the same for all spectra (Lommen et al., 1998). Whilst this improved very slightly the clustering of four of the samples and left one sample unchanged, it also split the repeats of one sample into two, probably because one lot had a slightly different reference amount (sample preparation error). We did not attempt to normalise the total spectral intensity, a procedure that is sometimes used (Lindon et al., 2001).

On the whole the third series of spectra (again potato but variety, sample preparation, and operator are different from series 1) again highlighted the similarity between spectra of the same sample, prepared several times (different extractions), compared to other samples; each sample formed a cluster on PC1 and 2 (respectively 81.0 and 6.3% variance). The loadings suggest this is due to 'natural variability'. However a slight split was observed on PC2 for several samples; for one of them (sample B) the split separated spectra of one of the extracts (three runs) from their repeats, but for others (sample D and F) it did not correspond to any obvious criterion. In both cases it was difficult to pinpoint the reason for this split by observing the spectra. More importantly however, on PC3 (3.74% variance) a clear difference appeared between runs; spectra from run 3 had consistently higher scores than those from the two other runs, and run 2 also tended to have higher scores than those from run 1 (Fig. 4). The difference between run 3 and the others could be traced back to a different lineshape; none of the runs had completely symmetrical peaks, with run 3 having slightly skewed shape on the high chemical shift side and runs 1 and 2 having a very slight shoulder on the low chemical shift side (Fig. 5).

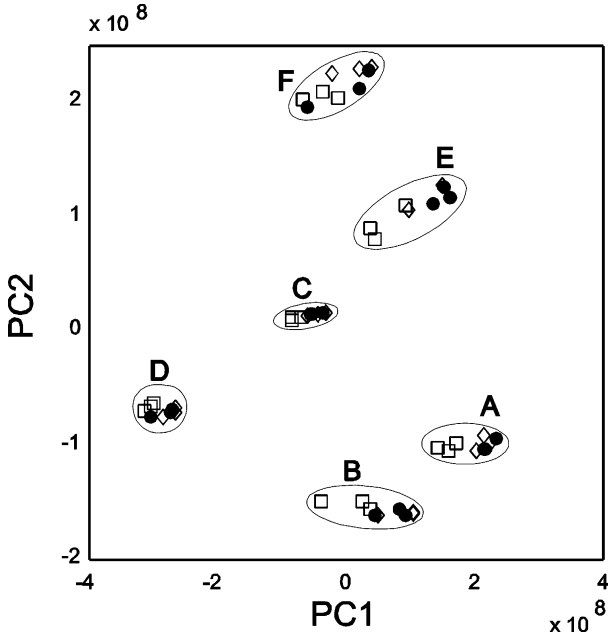

Fig. 3. Scores obtained on PC1 and 2 for the PCA of the second series of samples (tomato), coded A–F, highlighting NMR-run number (□ run 1, ● run 2, ◇ run 3) and samples (delimited by outline).

2.2. Tea data set

The fourth set of data was valuable in allowing us to explore some of the potential effects of the variability in the exact chemical shifts of chemical species, sometimes observed even within datasets of similar samples such as this one. For this study, ^{1}H NMR spectra of a set of green tea samples (see experimental section), on which an exploratory data analysis was to be carried out, were used.

The exact chemical shifts of NMR signals depend on several factors, and in particular on the solvent used. This is why reference standards for signal assignment are best prepared in exactly the same way as the samples of interest. In addition to the solvent, the effective pH of the extract may also have an influence, in particular on the chemical shifts of compounds containing acidic and basic groups. Interactions with metal ions, hydrogen bonding and other intermolecular interactions can also cause chemical shift displacements. In the case of teas, the variable position of some signals may be explained by the hydrophobic nature of caffeine and its tendency to self-associate and form complexes with polyphenols in water and other polar solvents such as the mixture used here (Seshadri and Dhanraj, 1988; Lakenbrink et al., 2000; Le Gall, 2002). For these reasons it is important to record series of spectra intended for metabolite fingerprinting under carefully considered conditions and to keep these conditions as constant as possible throughout. In our work we are aware of these issues and endeavour to apply these principles. Nevertheless, variations in sample composition lead to inconsistencies in signal position even when every care is taken to ensure uniform sample extraction. The effect of chemical shifts not being constant throughout a dataset can vary. In our work, we have sometimes encountered relatively subtle effects of such variability: the analysis by PCA may not be obviously perturbed but the loadings, when plotted as values against variable number,

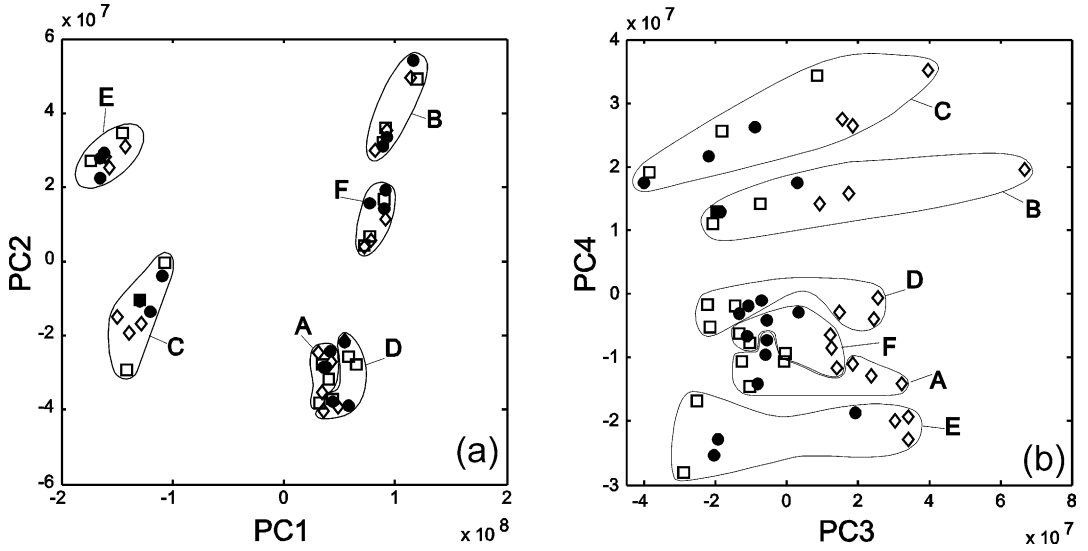

Fig. 4. Scores obtained on (a) PC1 and 2 and (b) PC3 and 4 for the PCA of the third series of samples (Bintje potato), coded A–F, highlighting NMR-run number (□ run 1, ● run 2, ◇ run 3) and samples (delimited by outline).

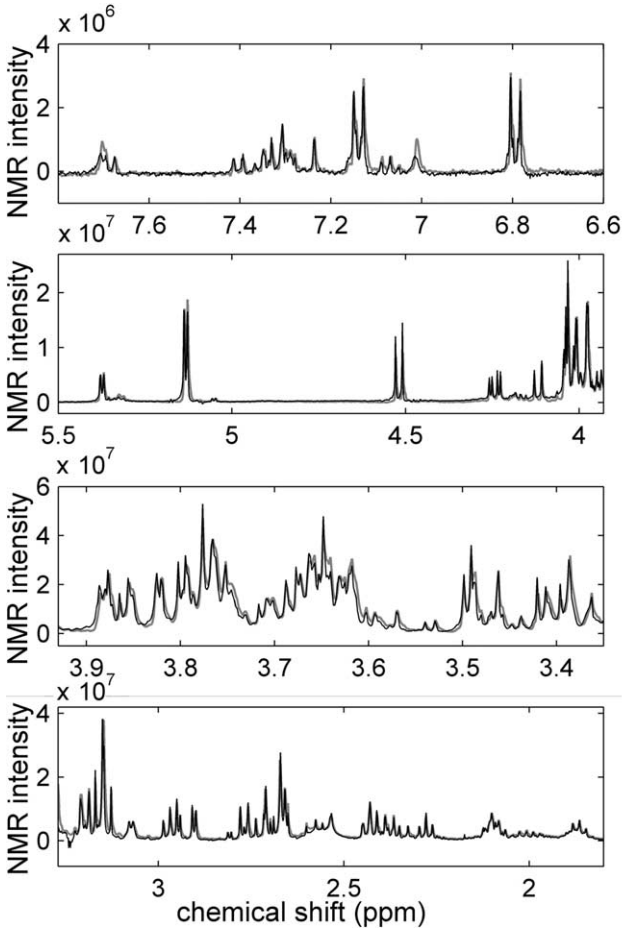

Fig. 5. Selected regions of potato spectra used in the analysis of the third series, sample C (—— spectrum from run 3, —— spectrum from run 2).

may show an apparent differential character to the peaks.

In the present case however, the effects seem much more severe. The scores on PC1 and 2 are shown in

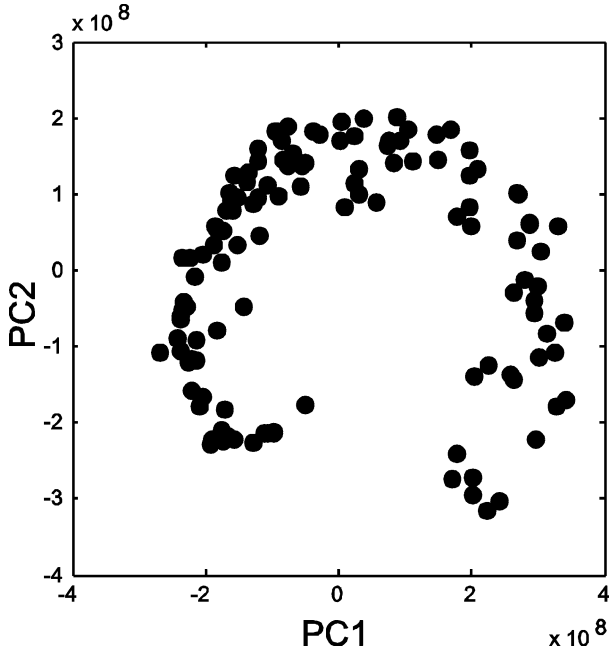

Fig. 6. Scores obtained on PC1 and 2 of PCA of tea data set.

Fig. 6, and the corresponding loadings in Fig. 7. The unusual distribution of the scores on the PCs can be explained by examining the loadings. These are greatest (in absolute value) for the large peaks in the spectra, whose position varies and whose maxima span a certain range of chemical shifts throughout the dataset. PC2 has maximal, positive, weights for the average chemical shift positions and less high, slightly negative, weights on either side, whilst PC1 has high negative weights to the high chemical shift side of the mean chemical shift position and high positive values to the other side. As a consequence, spectra whose peaks appear at the average

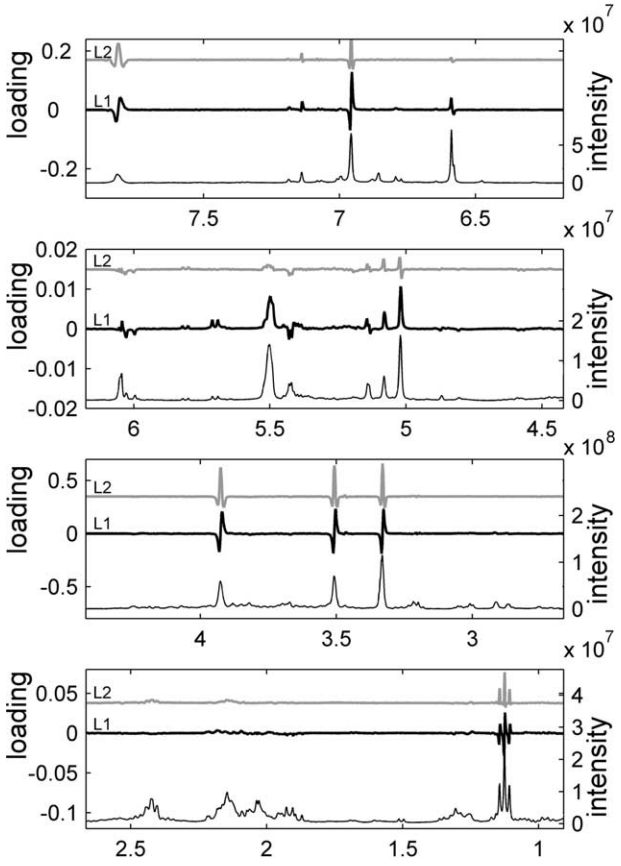

Fig. 7. Loadings of PC1 (——) and PC2 (——) of PCA of the tea dataset, shown above the average spectrum of the dataset (——). Both loadings are on the same scale but loading 2 is offset from loading 1 for clarity. The *y*-axis on the left is for the loadings, and that on the right for the spectrum.

chemical shift have average (i.e. near null) PC1 scores and high positive PC2 scores, spectra whose peaks are displaced slightly to lower chemical shifts have high positive PC1 scores and average (i.e. near null) PC2 scores, and spectra whose peaks are shifted to yet lower chemical shifts have somewhat less high positive PC1 scores and low, negative, PC2 scores. For spectra with signals that are displaced to higher chemical shifts the above observations are reversed as regards the PC1 scores, but the effect on the PC2 scores is unchanged (because of the symmetry of the loadings about the mean chemical shifts). It is the combination of the loadings of PC1 and 2 that explains the peculiar distribution of the scores. Thus when repeating the analysis following the application of a peak alignment procedure based on that described by Vogels et al. (1993, 1996a, b), the distribution of the scores is much more homogeneous (data not shown).

To examine further the circumstances leading to such behaviour, the data constituting individual peaks whose position was not constant was examined. Firstly it can be noted that the positions of the peak maxima at 7.85, 3.94, 3.52, and 3.34 ppm are highly correlated (correlation coefficient 0.9 to 0.98): these peaks all arise from caffeine. However, the distributions of the peak maxima (see selection on Fig. 8, and Table 1) show that the spread of the possible values taken for these maxima varies from one peak to another. The consequence of this is that the loadings, and hence the scores' distribution, also differ between them, as seen in examples in Fig. 9. The analysis by individual PCAs of the peaks centered at 7.85, 3.94, 3.52, and 3.34 ppm led to similar loadings to those observed for these same peaks in the whole spectrum analysis. Those four caffeine peaks were ones for which the standard deviation of the distribution of the peak position exceeded 1.89 datapoints. It was also found that when the spread of the peak maxima gave these peculiar loadings, loadings 1 and 2 then

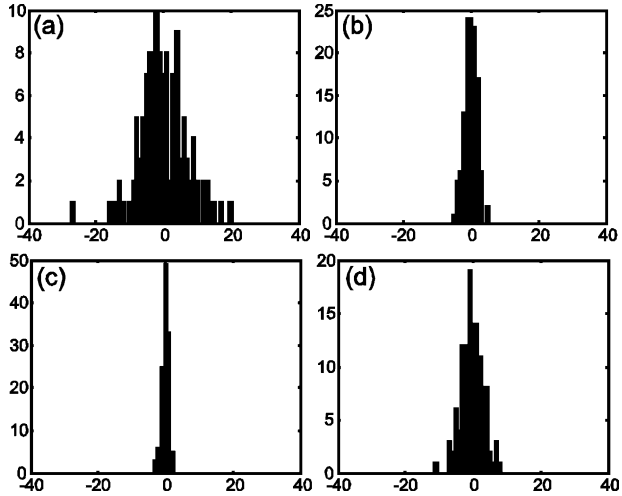

Fig. 8. Tea dataset, distribution of difference (in number of datapoints) between position of peak maxima and average position of peak maxima for peaks centred at (a) 7.85 ppm, (b) 6.95 ppm, (c) 5.02 ppm, (d) 3.52 ppm.

Table 1
Summary of peak characteristics and behaviour in PCA for peaks analyzed individually

δ (ppm)	Standard deviation of peak maximum (datapoints)	Exhibiting overall set behaviour in PCA	% Variance PC1	% Variance PC2
7.85[a]	7.2	Yes	32	22.7
7.14[b]	1.24	No	60.1	24.7
6.95[c]	1.89	Somewhat	54.7	25.4
6.59[c]	0.59	No	46.1	30.7
5.02[c]	1.04	No	65.3	31.4
3.94[d]	4.3	Yes	39.7	25.4
3.52[d]	3.26	Yes	42.2	25.2
3.34[d]	2.77	Yes	41.8	25.1

[a] Caffeine (NH).
[b] Gallic acid.
[c] Epigallocatechin-3-gallate.
[d] Caffeine (CH_3).

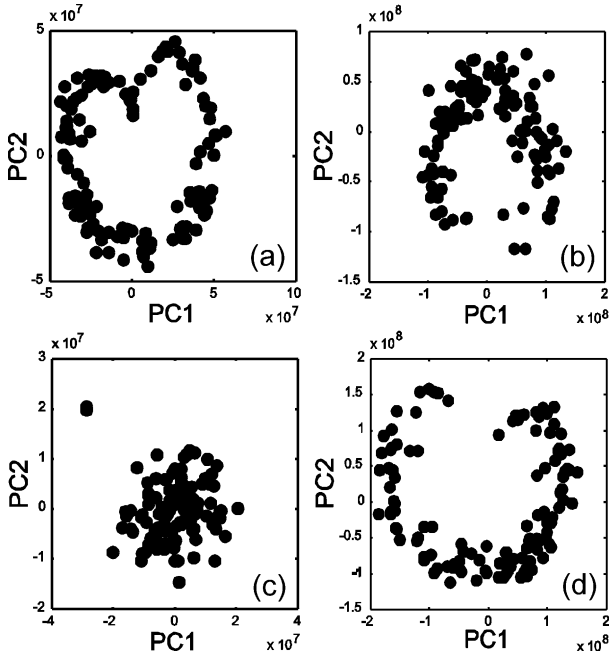

Fig. 9. Tea dataset, scores obtained on PC1 and 2 of separate PCAs carried out with peaks centred at (a) 7.85 ppm, (b) 6.95 ppm, (c) 5.02 ppm, (d) 3.52 ppm.

explained about 32–42 and 22–25% (respectively) of the variance, as opposed to about 45–65 and 25–31% respectively, when the spread was smaller (Table 1).

The controlled repeatability study and analysis of a somewhat peculiar real dataset presented above allowed us to highlight effects that may be observed in PCA scores plots of NMR data, and which are not related (or related indirectly) to the sample composition. The magnitude of these effects was variable, but we were able to trace them back to features in the spectra. The two main 'spectral defects' that introduced unwanted information in the PCA scores plot concerned the lineshape (resolution) and unstable frequency of given peaks throughout the dataset. The variations of lineshape seen here were somewhat exaggerated (especially in series 1) and the experiments described helped to diagnose a fault in the NMR spectrometer hardware (subsequently corrected) that affected the auto-shimming procedure. Reference deconvolution (Morris et al., 1997) might be used to correct such faults post-acquisition but it is obviously preferable to avoid the need for such corrections. When signal displacement is a problem modifications to the extraction procedure may be worth considering: if this is not successful segment-wise signal alignment procedures can be helpful although no completely general method has yet been demonstrated.

Lineshape changes and inconsistencies in chemical shift (Vogels et al., 1996b) are issues that have been described by other authors, albeit not necessarily for ^{1}H NMR spectra. A number of authors have been working with ^{31}P NMR data and observed such effects, for example Brown and Stoyonova (1996) and Witjes et al. (2000a, 2001). Kuesel et al. (1996) comment on the fact that peak shifts in ^{31}P NMR spectra may be due to either instrumental factors (field differences) or sample characteristics (pH, ionic strength, solvent, and temperature), and that their influence may be reduced but not always eliminated. Frequency shifts and lineshape changes may also occur in Raman spectroscopy, as observed by Witjes et al. (2000b, 2001); Bowie et al. (2000) explain that a change or move of limiting aperture and changes in temperature are two potential sources of shifts in Raman. Infrared spectroscopy, despite not always yielding well-resolved peaks, may nevertheless also be susceptible to similar effects, as described for the near-infrared region by Wülfert et al. (1998, 2000) and in the mid-infrared region by Defernez and Wilson (1997). Beyond spectroscopy, instability in the referencing of data may be observed for electrophoretic data (Rønn, 2000, 2001); here the peak corresponding to a given compound may appear at somewhat different positions on the time axis, in a set of measurements. The same thing is true of chromatographic data (Nielsen et al., 1998).

3. Conclusions

It is clear from this work and the discussion above that imperfections in the data registration, confusing the analysis of metabolite fingerprints, are far from being confined to a particular technique, and that some of the effects on PCA described here may be observed for other types of data. It is certainly the case, however, that the magnitude of these imperfections differs; for instance, a change of column in chromatography may have as consequence a considerable change in retention times, whilst in mid-infrared spectroscopy changes will be much more subtle, even when changing a major component like the infrared source. The discussion here concerns cases where the peak shift is such that there remains an overlap between peak traces in the whole data set, even though the position of the peak maxima may vary. The effect on a PCA that we have been able to highlight with real data has been described in several pieces of work with simulated spectra constituted by a single line. Brown and Stoyonova (1996), and Witjes et al. (2000b) show the loadings of a PCA carried out with sets of simulated spectra constituted by a single Lorentzian peak, where the 2nd PC loadings are very similar to our 1st loadings and their 3rd loading to our 2nd loading. Siuda et al. (1998) also show derivative-type loadings in the analysis of a set of simulated spectra constituted by a single Gaussian-shaped curve. For the tea samples included in this study the spectra are dominated by a few large peaks whose positions are variable within the dataset, and this is why the result so closely

matches the simulated case of single lines. However in other sets of complex spectra sources of variance other than peak shifts may have a considerable contribution, spreading the effect of peak shifts onto more PCs and making the phenomenon less easily recognisable.

This work has focused primarily on potential difficulties that may arise from analyzing ^{1}H NMR spectra by chemometric methods. However it should be stressed that the study examples were chosen to illustrate what can go wrong. All of them had occurred in the normal course of our work: they were not set up to prove a point. We have shown the importance of carefully adjusted experimental conditions, and hinted at the possibility to carry out post-acquisition corrections (e.g. segment-wise alignment). This strongly suggests that care has to be taken when using the NMR/chemometrics approach, but a growing literature testifies to its advantages and overall validity across many applications.

4. Experimental

4.1. Sample preparation

4.1.1. Controlled repeatability study

Three series (each consisting of 6 different samples×3 extractions) of extracts were prepared for NMR spectroscopy.

4.1.1.1. Series (1). Each extract was prepared by addition of 1 ml 70% methanol-d_4/30% buffer (100 mM K_2HPO_4/KH_2PO_4, 1 mM TSP in D_2O) to 0.04 g of freeze-dried potato powder (cv. Desirée, five genetically modified lines (three different modifications) and one control line; samples from whole tubers, combined from single plants). The mixture was stirred at room temperature for 30 min and centrifuged at 10,000 rpm for 10 min (Jouan A14 centrifuge). Each NMR sample consisted of 700 μl of the supernatant, which was stored in the NMR tube at −18 °C until required for analysis (and between automation runs). The spectra were collected in four different runs.

4.1.1.2. Series (2). Tomato extracts of six lines of tomato plants were prepared in the same way (Le Gall et al., submitted for publication) except that the buffer solution was also 1 mM in EDTA (0.048 g freeze dried powder extracted in 1.2 ml solvent, 750 μl of supernatant transferred to the NMR tube). Spectra were collected in three different runs.

4.1.1.3. Series (3). Potato samples (cv. Bintje; samples from flesh only, individual tubers from a single soil-grown batch) were prepared as for *Series 1*. The variation present was expected to be less than in *Series 1*. The spectra were collected on three different runs.

4.1.2. Tea data set

The second set of recordings was of a group of 121 green tea samples of different type, grade and geographical origin. Tea leaves were ground to a fine powder in a coffee grinder. The powder (0.096 g) was extracted into 1.2 ml 70% methanol-d_4/30% buffer (10 mM Na_2HPO_4 with 1 mM TSP in D_2O) with stirring for 10 min at 70 °C. After cooling 20 μl of a solution containing EDTA and ascorbic acid (both 100 μg/ml) was added, the solution was centrifuged as above and 700 μl of the supernatant were transferred to an NMR tube.

4.2. NMR spectroscopy

^{1}H NMR spectra were recorded at 27 °C on a 400 MHz JEOL GX spectrometer. For the repeatability studies the recording of the NMR spectra was carried out 'in block' (i.e. all 18 tubes consecutively), and this was repeated on separate occasions 3–4 times. Each block of 18 samples was run under full automation using a sample changer, with automatic adjustment of Z1 and Z2 shims prior to data acquisition for each sample. For the tea data set the extracts were run once each in automation blocks of 30–40 samples at a time. Tuning of the spectrometer and manual shimming were carried out on the first sample before the start of each automation run. Methanol-d_4 was used as the internal lock. Each spectrum consisted of 104 scans (304 scans for *Series 2*) of 8192 complex data points with a spectral width of 5000 Hz, an acquisition time of 1.64 s and a recycle delay of 2 s per scan. The pulse angle was 50°. The receiver gain was set at the same value for all samples within a series. A presaturation sequence was used to suppress the residual water signal with low power selective irradiation at the water frequency during the recycle delay. Spectra were Fourier transformed with 1 Hz line broadening, phased and baseline corrected using the JEOL (Delta) software. Spectra were converted to Felix 2000 software format and saved as ASCII files. Spectra were further transferred to a personal computer for data analysis.

4.3. Data analysis

The data analysis was carried out using Matlab 6.1 (The Mathworks, Inc) running on a desktop computer. For the repeatability study PCA was carried out with each of the sets of spectra described above, using whole spectra except for the baseline at high and low fields and the suppressed water signal. Typically this left about 5000 data points. The data was mean-centred for each variable prior to analysis. For the first series the analysis was repeated after de-resolving the spectra by factors of 2–25 with an in house-written routine. Each point in the new spectra was defined as a weighted sum of neigh-

bouring points in the original spectra. The tea data set was also initially subjected to a PCA involving the whole spectra with the exception of the extreme baseline regions and suppressed water signal. The analysis appeared to be strongly influenced by a few strong peaks whose chemical shift varied from one sample to the other. Therefore separate PCAs were also carried out with a few selected peaks (7.85, 7.14, 6.95, 6.59, 5.02, 3.94, 3.52, and 3.34 ppm), typically constituting 20–80 datapoints, to understand the phenomenon. In addition, a procedure designed to align peaks was applied to segments of the spectra prior to a re-analysis by PCA of the whole spectra.

Acknowledgements

The authors would like to acknowledge the financial support from the Biotechnology and Biological Sciences Research Council's competitive strategic grant. They thank Gareth S. Catchpole, Yvonne M. Gunning, and Gwénaëlle Le Gall for preparing samples and recording the spectra. Howard Davies and Louise Shepherd (Scottish Crop Research Institute, UK) are acknowledged for the provision of potato samples and the Tea Research Institute, Hangzhou, China, for the tea samples.

References

Bowie, B.T., Chase, D.B., Griffiths, P.R., 2000. Factors affecting the performance of Raman spectrometers. Part 1: instrumental effects. Appl. Spectrosc. 54, 164A–173A.

Brown, T.R., Stoyanova, R., 1996. NMR spectral quantitation by principal component analysis. II. Determination of frequency and phase shifts. J. Magn. Res., Ser. B 112, 32–43.

Defernez, M., Wilson, R.H., 1997. Infrared spectroscopy: instrumental factors affecting the long-term validity of chemometrics models. Anal. Chem. 69, 1288–1294.

Fiehn, O., 2002. Metabolomics—the link between genotypes and phenotypes. Plant Mol. Biol. 48, 155–171.

Howells, S.L., Maxwell, R.J., Peet, A.C., 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.

Krzanowski, W.J., 1988. Principles of Multivariate Analysis: A User' s Perspective. Oxford University Press, New York.

Kuesel, A.C., Stoyanova, R., Aiken, N.R., Li, C.-W., Szwergold, B.S., Shaller, C., Brown, T.R., 1996. Quantitation of resonances in biological ^{31}P NMR spectra via principal component analysis: potential and limitations. NMR in Biomedicine 9, 93–104.

Lai, Y.W., Kemsley, E.K., Wilson, R.H., 1994. Potential of Fourier transform-infrared spectroscopy for the authentication of vegetable-oils. J. Agric. Food Chem. 42, 1154–1159.

Lakenbrink, C., Lapczynski, S., Maiwald, B., Engelhardt, U.H., 2000. Flavonoids and other polyphenols in consumer brews of tea and other caffeinated beverages. J. Agric. Food Chem. 48, 2848–2852.

Le Gall, G., Puaud, M., Colquhoun, I.J., 2001. Discrimination between orange juice and pulp wash by ^{1}H nuclear magnetic resonance spectroscopy: identification of marker compounds. J. Agric. Food Chem. 49, 580–588.

Le Gall, G., 2002. Food Authenticity and Quality Assessment using ^{1}H NMR Spectroscopy and Chemometrics. PhD thesis, University of East Anglia, UK.

Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J., Verhoeyen, M.E., 2002. Metabolite profiling of tomato using ^{1}H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. J. Agric. Food Chem. (submitted for publication).

Lindon, J.C., Holmes, E., Nicholson, J.K., 2001. Pattern recognition methods and applications in biological magnetic resonance. Prog. NMR Spectrosc. 39, 1–40.

Lommen, A., Weseman, J.M., Smith, G.O., Noteborn, H.P.J.M., 1998. On the detection of environmental effects on complex matrices combining off-line liquid chromatography and ^{1}H-NMR. Biodegradation 9, 513–525.

Morris, G.A., Barjat, H., Horne, T.J., 1997. Reference deconvolution methods. Prog. NMR Spectrosc. 31, 197–257.

Nielsen, N.P.V., Carstensen, J.M., Smedsgaard, J., 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J. Chromatogr. A 805, 17–35.

Rønn, B., 2000. Alignment of curves by non-parametric maximum likelihood estimation (personal communication).

Rønn, B., 2001. Nonparametric maximum likelihood estimation for shifted curves. J. R. Statist. Soc. B 63, 243–259.

Seshadri, R., Dhanraj, N., 1988. Flavour interactions in tea. In: Proceedings of the 5th International Flavour Conference, Porto Karras, Chalkidiki, Greece, 1–7 July 1987, Elsevier, Amsterdam.

Siuda, R., Balcerowska, G., Aberdam, D., 1998. Spurious principal components in the set of spectra subjected to disturbances. I. Presentation of the problem. Chemom. Intell. Lab. Sys. 40, 193–201.

Vogels, J.T.W.E., Tas, A.C., van den Berg, F., 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. Chemom. Intell. Lab. Sys. 21, 249–258.

Vogels, J.T.W.E., Terwel, L., Tas, A.C., van den Berg, F., Dukel, F., van der Greef, J., 1996a. Detection of adulteration in orange juice by new screening method using proton NMR spectroscopy in combination with pattern recognition techniques. J. Agric. Food Chem. 44, 175–180.

Vogels, J.T.W.E., Tas, A.C., Venekamp, J., van der Greef, J., 1996b. Partial linear fit: a new NMR spectroscopy pre-processing tool for pattern recognition applications. J. Chemom. 10, 425–438.

Witjes, H., Melssen, W.J., in 't Zandt, H.J.A., van der Graaf, M., Heerschap, A., Buydens, L.M.C., 2000a. Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets. J. Magn. Res. 144, 35–44.

Witjes, H., van den Brink, M., Melssen, W.J., Buydens, L.M.C., 2000b. Automatic correction of peak shifts in Raman spectra before PLS regression. Chemom. Intell. Lab. Sys. 52, 105–116.

Witjes, H., Pepers, M., Melssen, W.J., Buydens, L.M.C., 2001. Modelling phase shifts, peak shifts and peak width variations in spectral data sets: its value in multivariate data analysis. Anal. Chim. Acta 432, 113–124.

Wülfert, F., Kok, W.T., de Noord, O.E., Smilde, A.K., 2000. Correction of temperature-induced spectral variation by continuous piecewise direct standardization. Anal. Chem. 72, 1639–1644.

Wülfert, F., Kok, W.T., Smilde, A.K., 1998. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. Anal. Chem. 70, 1761–1767.