Journal of Pharmaceutical & Biomedical Analysis
Vol. 10, Nos 10–12, pp. 837–844, 1992
Printed in Great Britain

0731–7085/92 $5.00 + 0.00
Pergamon Press Ltd

Assessment of chromatographic peak purity of drugs by multivariate analysis of diode-array and mass spectrometric data*

DAEMON LINCOLN,‡ ANTHONY F. FELL,†‡ NICHOLAS H. ANDERSON§ and DAVID ENGLAND§

‡ *Pharmaceutical Chemistry, School of Pharmacy, University of Bradford, Bradford BD7 1DP, UK*
§ *Sterling Winthrop Research Centre, Willowburn Avenue, Alnwick NE66 2JH, UK*

Abstract: Numerous multivariate chemometric approaches have been developed for LC–UV data acquired using a diode-array detector (DAD), but these methods have not been widely exploited for LC–MS data. Principal component analysis (PCA) and subsequent axis rotation within the reduced factor space are assessed for LC–DAD and LC–MS data as approaches for estimating the number of components (i.e. the rank of the data) under a single chromatographic peak for compounds whose UV-spectra are very similar. Multivariate techniques for LC–DAD data are shown to suffer from inherent limitations of sensitivity for the minor components. The novel technique in LC–MS of plotting the rotated PCA data in two-dimensional factor space generates characteristic ion clusters, giving a visual criterion of peak purity. Single ion chromatograms produced subsequently confirm the profile of each coeluting component and give evidence of the degree of peak overlap. The application of this new chemometric technique to the detection of low levels of coeluting impurities by LC–MS is discussed as a novel approach for the validation of LC separations in pharmaceutical research and development.

Keywords: *Liquid chromatography; thermospray mass spectrometry; peak purity; multivariate analysis; principal component analysis; diode-array detection.*

Introduction

One of the primary aims of LC method validation in the pharmaceutical industry today is to assess the purity of the analyte peak. In fact, homogeneity rather than purity is the property that is usually measured, since very few techniques currently available are able to detect the case of exact coelution. An example where peak purity is an essential requirement for validating an analytical method would be in the long-term stability study of a finished product. For this, stressed samples of the product are analysed by the proposed method and the resulting analyte peak is subjected to one or more of the so-called peak purity tests. These include graphical methods such as spectral overlays [1], iso-absorbance plots [2], three-dimensional projections of wavelength–time–absorbance data [3], higher derivative transformation of spectra [4], as well as numerical techniques such as absorbance ratio [5], spectral suppression [6], purity parameter [7], absorbance index [8], multiple absorbance ratio correlation [9] and peak area correlation [10].

The major limitations of these univariate approaches can be summarized as follows:

(i) With graphical methods, close overlap of a related compound always produces a composite spectrum which shows little significant difference from that of the pure compound. Such small differences are thus difficult to detect, so that only relatively high levels of impurity can be observed.

(ii) With numerical methods, where there is no information on the spectrum of the overlapping impurity or its retention time, many of these techniques suffer from inappropriate wavelength and/or timepoint selection. Some recent techniques such as the purity parameter [7] and peak area correlation [10] overcome this limitation very elegantly by sampling ranges of data from both the spectral and the time domains. In spite of this all these univariate techniques suffer from an under-utilization of the available information, resulting in a lack of generality of application.

The main thrust of developments in multivariate analysis originated from the early work of Malinowski [11] and Kowalski *et al.* [12], in which they propounded the use of principal

*Presented at the "Fourth International Symposium on Drug Analysis", May 1992, Liège, Belgium.
†Author to whom correspondence should be addressed.

component analysis and factor analysis for the examination of chemical data. These mathematical procedures have since been applied to the examination of multivariate data sets generated by various hyphenated analytical techniques.

PCA is often applied as the first stage of data analysis, in order to determine the rank of the data, i.e. the most probable number of overlapping components in the data set, excluding noise. The final solution is generated after rotation of the principal axes within the reduced factor space, in order to simplify the data structure by maximizing the association of the high loading variables with the relevant factor, as discussed below.

Developments of this form of multivariate analysis include rank annihilation factor analysis (RAFA) [13], evolving factor analysis (EFA) [14], iterative target transformation factor analysis (ITTFA) [15, 16] and heuristic evolving latent projections (HELP) [17].

Operations in multivariate analysis

Multichannel chromatographic data can be represented in the form of an s by t matrix, **D**, where s is the number of spectral channels (wavelengths or m/z values) and t is the number of timepoints across the chromatographic peak at which the spectra were collected. Therefore, the signal measured at a single spectral channel i and timepoint t for a system consisting of n components can be expressed algebraically as:

$$D_{ij} = \sum_{k=1}^{k=n} A_{ik} \cdot C_{kj}, \qquad (1)$$

where A_{ik} is the normalized spectral amplitude of compound k at spectral channel i, and c_{kj} is the chromatographic elution profile of compound k at time j [11]. The above equation assumes linear additivity of the individual component signals and an absence of systematic error in the measurement process.

Equation (1) can be more conveniently expressed as the product of two matrices:

$$[\mathbf{D}] = [\mathbf{A}]\,[\mathbf{C}], \qquad (2)$$

where [**A**] is an s by n matrix containing the normalized spectra of the n components, and [**C**] is an n by t matrix containing chromatographic concentration profiles of the corresponding components.

Of course, real data contains noise (indeterminate error) and it is a primary goal of principal component analysis to extract and partition this uncorrelated variance from the raw data matrix, thus allowing estimation of the true dimensionality or rank of the data. It is assumed that the true rank of the data will be equal to the number of components within the chromatographic peak, provided the following requirements are met: (i) the chromatographic profiles (i.e. shape and retention times) of the individual components do not exactly overlap; (ii) the spectral patterns of each component are measurably different (i.e. the correlation coefficient between the spectra is less than a value determined by the spectral resolution of the detector); (iii) systematic errors, such as those produced by the process of spectral scanning as the peak elutes, are either absent or corrected for.

The first stage of a factor analysis usually involves the decomposition of either the covariance matrix or the correlation matrix [18] to obtain the associated eigenvectors according to the relation:

$$\mathbf{CX} = \lambda\mathbf{x} \qquad (3)$$

the column vectors, λ, being subsequently ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$.

The next stage, and this is a far from trivial procedure, is to determine how many of these eigenvectors are required to adequately describe the rank or dimensionality of the system. Although this is an area of continuing research, as evidenced by some recent papers published on this topic [17, 19, 20], at the time of writing no generally applicable approach had been universally accepted.

Finally, the principal components (normally referred to as *factors* after this rotation step) can be rotated in the direction of the constituent vectors within the pre-determined subspace. The aim of this rotation is to produce a simplification of the data structure by minimizing the number of variables that have high loadings on a particular factor. Thus if the partial similarity of a factor with many variables is eliminated by emphasising the strong similarity (or dissimilarity) of that factor with a few variables, the data are structured into a more easily interpretable form. This can be readily achieved using the Varimax technique [11], whereby the mutually orthogonal principal components are rotated so that the total

sum of squares of the loadings along each new axis (or factor) is maximized.

Application of multivariate techniques to peak purity analysis

The data matrices generated by LC–DAD and by LC–thermospray-MS systems are characterized by intrinsically different information content. Thus MS data are much more likely to display clusters of ions representing individual components coeluting under the chromatographic peak envelope, than are UV data to yield specific spectral regions which characterize each component in the overlapping peak. In both cases the utility of multivariate analysis is expected to depend on the relative peak concentrations, the extent of peak overlap, the degree of similarity of the spectra and the noise characteristics of each system.

However the higher information content of mass spectra, which consequently require larger data matrices than UV spectra acquired during chromatography, implies that LC–MS should be at least as effective as LC–DAD in detecting the presence of overlapping peaks by multivariate analysis, except in the case of structural isomers.

In the present work the use of factor analysis is discussed for estimating the rank of a multivariate data set and for the detection of minor coeluting impurities using LC–DAD and LC–thermospray-MS in the separation of drugs and related substances. Some novel graphical approaches for visualising the results of PCA are presented; these have been specifically designed for LC–DAD and LC–MS data.

Experimental

DAD instrumentation

A Varian LC Star System Workstation (Varian Associates, Walnut Creek, CA, USA) operating on a Compaq Deskpro 286 computer was used to manage a Varian 9010 Solvent Delivery System fitted with a Rheodyne 7125 manual loop (10-μl) injection valve and a Varian Polychrom 9065 detector.

MS instrumentation

A Finnigan MAT TSQ 700 triple stage quadrupole mass spectrometer operating on a DEC Station 2100 was interfaced through a Finnigan TSPII thermospray device to a HP 1050 Quaternary LC system.

Column and eluent

LC–UV. A 5-μm Spherisorb S5 ODS-2 (25 cm × 4.6 mm i.d.) reversed-phase column was used in conjunction with methanol–water (30:70, v/v) as mobile phase.

LC–MS. A μBondapak (Waters) cyano column (5-μm material, 12.5 cm × 4.6 mm i.d.) was employed as stationary phase. The mobile phase composition was: ammonium acetate (0.5% w/w; pH 6.0)–methanol (25:75, v/v). The buffer was adjusted to pH 6.0 with glacial acetic acid prior to mixing.

Reagents and materials

Methanol (HPLC grade, Rathburn Chemicals, Walkerburn, UK), ammonium acetate ('Analar', BDH Lab supplies, Poole, UK) and glacial acetic acid ('Analar', BDH Lab supplies, Poole, UK) were used as received. The buffer salt was dissolved in distilled water and filtered using Millipore 0.45-Im filters.

Theophylline (Batch number: T-1633) and paraxanthine (Batch number: 38f40581) from Sigma Chemical Co. (St Louis, MO, USA) were used as received (Fig. 1). The two drug substances, COMPD1 (MW 390) and COMPD2 (MW 362) (Fig. 2) employed in the LC–MS studies were kindly provided by Sterling Research Group (Alnwick, UK).

Standard solutions of each of the analytes alone, and in binary mixtures, were prepared in the appropriate HPLC mobile phase. A solution of COMPD1 spiked with 10% w/v of COMPD2 solution, was aged for a period of 10 weeks, by storage under ambient light at room temperature.

Selection of data submatrix for SPSS/PC+ software package

A submatrix of each data set was selected for manual input into the SPSS/PC+ Version 3.0 statistical software package (SPSS Inc., 444 N. Michigan Avenue, IL 606611, USA).

LC–UV. Spectra (220–287 nm) were selected at regular 0.05 min intervals throughout the chromatographic peak, producing 15 × 15 submatrices of spectrochromatographic data.

LC–MS. Spectra (composed of the 10 most

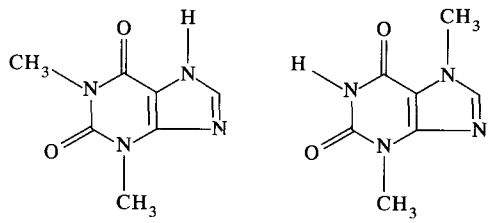

Figure 1
Structures of theophylline and paraxanthine.

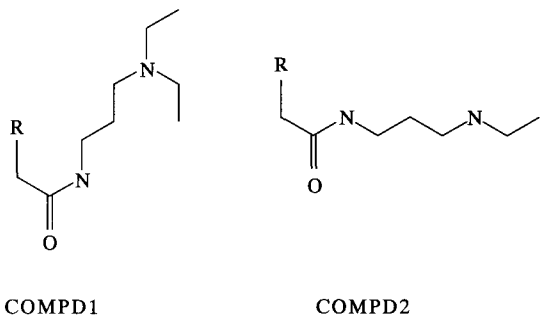

Figure 2
Structures of the parent compound (COMPD1) and a potential process impurity (COMPD2).

abundant ions, as determined at the chromatographic peak apex) were taken at 1.0 s intervals throughout the chromatographic peak, producing 20 × 10 submatrices of spectrochromatographic data.

Results and Discussion

LC–DAD experiments

The data summarized for the sample solutions analysed by the photodiode array detector are presented in Tables 1–4. Table 1 shows quite clearly that even for a concentration range of standard solutions between 0.02 and 0.50 mg ml^{-1}, representing an absorbance range from 25 to 600 milliabsorbance units (mAU), the eigenvalue for the first principal component (or factor) remains very stable. There is in fact only a very small, but nonetheless predictable, increase in the eigenvalue for the second PC or factor as the concentration level is decreased, emphasizing the excellent signal-to-noise ratio of the detector.

Another important feature of these data, as highlighted by Table 2, is the apparent lack of any other significant sources of systematic variability. The orthogonal rotation of the principal axes within the pre-determined subspace was attempted, but succeeded only in reproducing the original eigenvalues, as shown in Table 1, because of the lack of significant latent variability. This gives a further indication of the presence of only one component in this dataset.

The data in Table 3 give unequivocal evidence that a second source of systematic variation is present for the solutions spiked with theophylline. However, this is only true for concentrations at 10% w/v and above. The case for detection of impurity in the solution spiked at 2% w/v is doubtful. Although the second eigenvector certainly exceeds that of the 'noise predictor' for the corresponding standard solution, the Varimax routine is unable to rotate the principal axes to emphasize any latent variability in the secondary eigenvalues, which would permit them to be more readily detected (Table 4).

It could be argued that the possibility of an empirically based method for rank determination is suggested by the results in Tables 1–4. If one assumes that the second eigenvalue is indeed due to indeterminate error from the PCA–factor analysis of a standard solution, then the use of this value for the determination of the noise threshold level may be possible. Repeated injections of the standard could ultimately facilitate the incorporation of a confidence interval for the expected noise threshold at the desired probability level.

LC–MS experiments

As indicated in the experimental section, the LC was linked to the mass spectrometer via a thermospray interface. This produced the somewhat noisy total ion chromatograms as typified by Fig. 3. Even the most optimistic of analysts would not normally attempt to investigate the composition of such a chromatogram if it had been produced by any other detector. However, it was hoped that the excellent selectivity and sensitivity of the mass spectrometric detector would, in combination with the factor analysis approach, yield some meaningful answers as to the underlying structure of the data.

Presented graphically in Fig. 4 are the eigenvalues produced from the decomposition of the transpose of the raw data matrices of the solutions: (A) freshly prepared; and (B) after storage under the prescribed conditions. A

Table 1
Comparison of the principal component eigenvalues for a series of pure paraxanthine standard solutions of decreasing concentration

		Principal components			
Solution	Conc. (mg ml^{-1})	PC1	PC2	PC3	PC4
1	0.50	13.9999	0.00004	0.00003	0.00000
2	0.10	13.9999	0.00007	0.00000	0.00000
3	0.02	13.9997	0.00016	0.00005	0.00003

Table 2
Comparison of the rotated principal component eigenvalues for a series of pure paraxanthine standard solutions

		Factors			
Solution	Conc. (mg ml^{-1})	F1	F2	F3	F4
1	0.50	13.9999	0.00004	0.00003	0.00000
2	0.10	13.9999	0.00007	0.00000	0.00000
3	0.02	13.9997	0.00016	0.00005	0.00003

Table 3
Comparison of the principal component eigenvalues for a series of paraxanthine standard solutions (0.1 mg ml^{-1}) spiked with decreasing levels of theophylline

		Principal components			
Solution	% w/v TH level	PC1	PC2	PC3	PC4
4	50	13.7054	0.29440	0.00013	0.00001
5	10	13.9973	0.00269	0.00004	0.00000
6	2	13.9996	0.00020	0.00011	0.00003
7	1	13.9999	0.00009	0.00003	0.00000
2	0	13.9999	0.00007	0.00000	0.00000

Table 4
Comparison of the rotated principal component eigenvalues for a series of paraxanthine standard solutions (0.1 mg ml^{-1}) spiked with decreasing levels of theophylline

		Factors			
Solution	% w/v TH level	F1	F2	F3	F4
4	50	7.92090	6.07900	0.00016	0.00001
5	10	7.13210	6.86825	0.00005	0.00000
6	2	13.9997	0.00020	0.00005	0.00000
7	1	13.9999	0.00009	0.00003	0.00000
2	0	13.9999	0.00007	0.00000	0.00000

novel method for presenting the transposed data from LC–MS is illustrated in Figs 5 and 6. These represent the eigenvalues of the transpose, **D′**, of the original raw data matrix, **D**. By this method it is clear that the inherent structure within the individual eigenvectors is better represented when plotting these latent ion vectors in retention time space.

In essence, this new technique is analogous to that developed concurrently with the present work by Kvalheim *et al.* [17], who analysed selected regions in spectrochromatograms by LC–DAD. However, for thermospray mass spectrometry, which produces little if any fragmentation of the quasi-molecular ion and hence few or no common ions, the effect is to produce characteristic groupings of ions for each component.

This is illustrated in Fig. 6, where the unique ions for both COMPD1 and COMPD2 are closely grouped together. Of course PCA is not able to extract all the error from the data; that which is imbedded or non-orthogonal to the principal component axes will remain [13]. This imbedded error is responsibile for the fact that ions *m/z* 413 and *m/z* 345 are separated

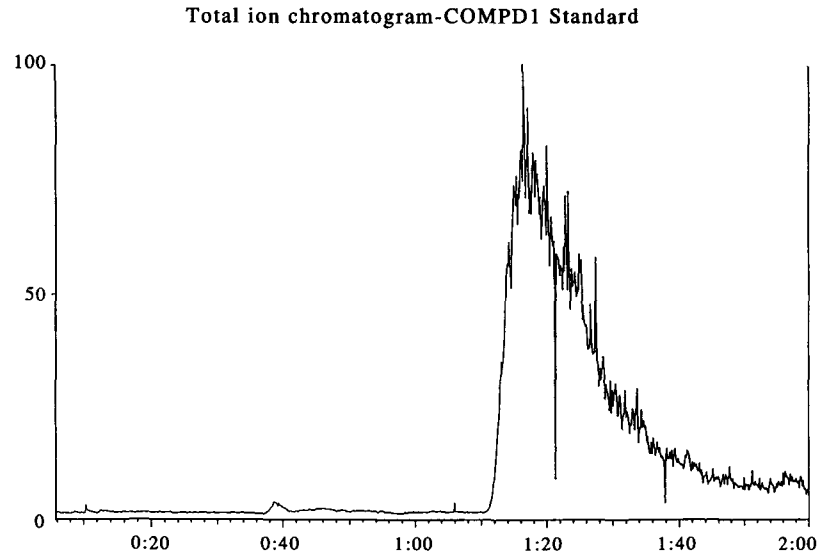

Figure 3
Total ion chromatogram of a 0.2 mg ml^{-1} COMPD1 standard solution obtained by thermospray LC–MS.

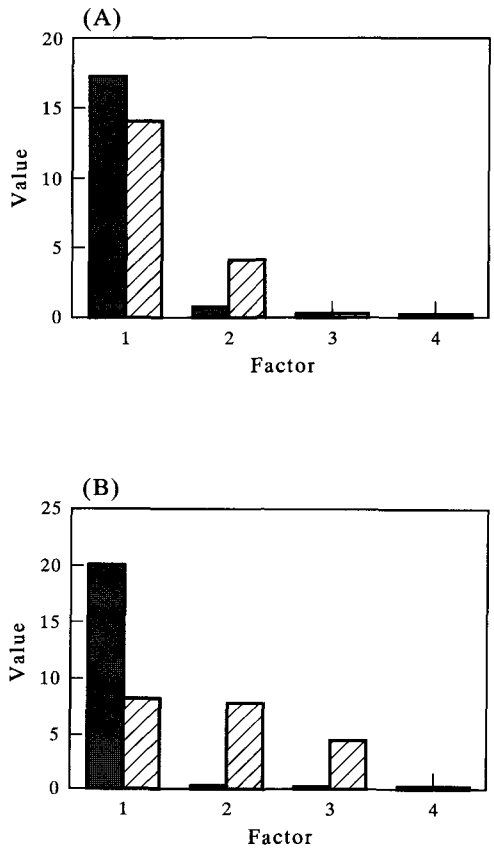

Figure 4
Graphical representation of the principal components produced by LC–MS of a 0.2 mg ml^{-1} standard solution of COMPD1, spiked with 0.02 mg ml^{-1} of COMPD2. (A), fresh solution; (B), solution stored at room temperature in ambient light for 10 weeks. Solid block, unrotated data; hatched block, data after Varimax rotation.

from the main component cluster at *m/z* 391. The main outlier from these two clusters of data is, however, the *m/z* 278 ion in the bottom left corner of the plot. This unexpected ion was further investigated by plotting the normalized single-ion chromatogram (after smoothing) to ascertain whether it represented a partially resolved coeluting impurity (Fig. 7). The resultant plot is inconclusive, but it must be borne in mind that (assuming that the response factor is similar to that of COMPD1), this ion probably represents a concentration of the minor component of less than 0.2% w/w in the fresh solution of COMPD1. However, by plotting the single ion intensity of the *m/z* 278 and the *m/z* 391 ions for the stored solution (Fig. 8), it is possible to see that the level is now sufficient (*ca* 0.5% w/w) to produce an identifiable chromatogram for a third solute, well resolved from that of COMPD1. This solute was subsequently identified as the hydrolysis product of COMPD1, whose concentration was, as expected, found to increase with time on storage in aqueous solution.

Conclusions

A certain amount of the data in a spectrochromatogram is low in information content or even redundant. This is particularly true for mass spectrometric data. Principal component analysis and factor analysis offer a means of extracting the pertinent information from the

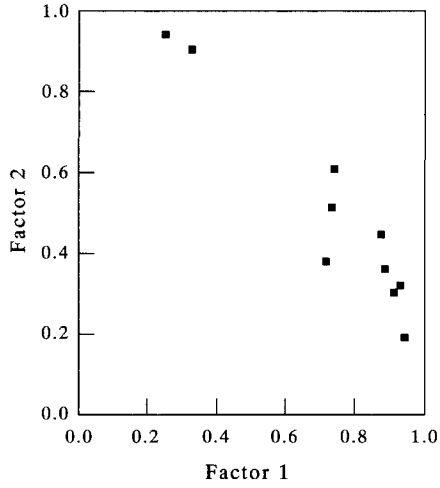

Figure 5
Plot of the eigenvector loadings in reduced factor space for the thermospray LC–MS of standard solution of COMPD1, after Varimax rotation.

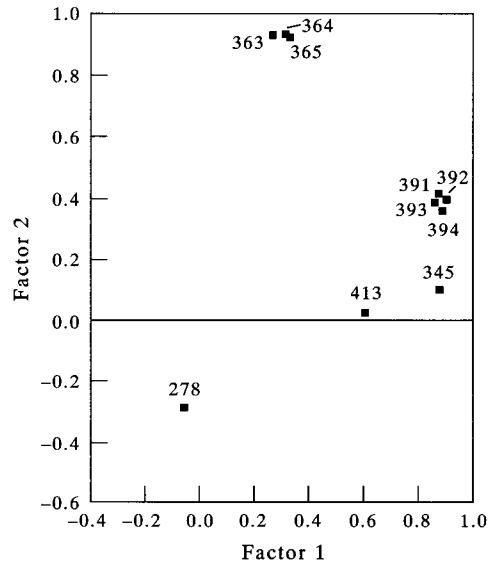

Figure 6
Plot of the eigenvector loadings in reduced factor space for the thermospray LC–MS of a 0.2 mg ml^{-1} standard solution of COMPD1, spiked with 0.02 mg ml^{-1} of COMPD2, after Varimax rotation.

data set by what is essentially a form of data compression. The technique can improve the effective signal-to-noise ratio and can ultimately lead to increased interpretability of the data.

The technique does not require *a priori* knowledge of the potential coeluting impurities either in the spectral or time domains. A major advantage of this approach over its univariate counterparts for peak purity assessment is

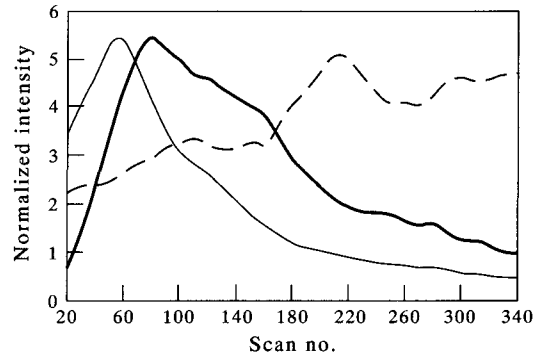

Figure 7
Selected ion plots (at 300 ms per scan) for a fresh solution of COMPD1 spiked at 10% w/v with COMPD2, revealing the coelution of an ion at *m/z* 278. Data in these plots were smoothed using a three-point moving average algorithm, prior to normalization. — *m/z* 391; —— *m/z* 363; - - - - *m/z* 278.

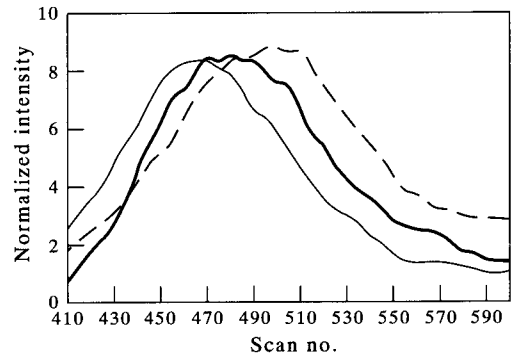

Figure 8
Selected ion plots (at 300 ms per scan) for a solution of COMPD1 spiked at 10% w/v with COMPD2, showing the increased level of the ion at *m/z* 278, after storage (cf. Fig. 4). Data treatment and key as in Fig. 7.

therefore its suitability for incorporation into an expert system.

The novel presentation of eigenvector loadings on each factor, plotted in reduced factor space, adequately demonstrates the advantage of utilizing the structural information content of the eigenvectors, rather than simply condensing down the information into a single-figure eigenvalue. This information was found to facilitate the location of characteristic groupings of ions for the thermospray data, leading to the generation of single-ion plots for the individual components eluting under the same chromatographic envelope. The success of this approach is clearly dependent on obtaining unique ions for an individual component with an adequate signal-to-noise ratio.

Further work in this area is continuing with the development of methods for analysing the complete LC–MS data set output, rather than a pre-selected sub-matrix, with and without additional information from tandem photodiode array detection.

Acknowledgements — The authors would like to thank Dr John W. Firth for collection of the LC–MS data, Dennis Lendrem for helpful discussions, and Dr John B. Castledine for his helpful comments on the manuscript. Thanks are also expressed to Varian Associates for the generous loan of equipment.

References

[1] B.J. Clark, A.F. Fell, H.P. Scott and D. Westerlund, *J. Chromatogr.* **286**, 261–273 (1984).
[2] B.F.H. Drenth, R.T. Ghijsen and R.A. de Zeeuw, *J. Chromatogr.* **238**, 113–120 (1982).
[3] A.F. Fell, H.P. Scott, R. Gill and A.C. Moffat, *Chromatographia* **16**, 69–78 (1982).
[4] A.C.J.H. Drouen, H.A. Billiet and L. de Galan, *Anal. Chem.* **56**, 971–978 (1984).
[5] P.C. White, *J. Chromatogr.* **200**, 271–276 (1980).
[6] A.F. Fell, H.P. Scott, R. Gill and A.C. Moffat, *J. Chromatogr.* **273**, 3–17 (1983).
[7] T. Alfredson and T. Sheehan, *Am. Lab.* **17**, 40–54 (1985).
[8] A.F. Poille and R.D. Conlon, *J. Chromatogr.* **204**, 149–152 (1981).
[9] J.G.D. Marr, G.G.R. Seaton, B.J. Clark and A.F. Fell, *J. Chromatogr.* **506**, 289–301 (1990).
[10] J.B. Castledine, A.F. Fell, R. Modin and B. Sellberg, *J. Chromatogr.* **592**, 27–36 (1992).
[11] E.R. Malinowski and D.G. Howery, *Factor Analysis in Chemistry*. Wiley Interscience, New York (1980).
[12] L.S. Ramos, K.R. Beebe, W.P. Carey, E. Sanchez, B.C. Erickson, B.E. Wilson, L.E. Wangen and B.R. Kowalski, *Anal. Chem.* **58**, 294R–315R (1986).
[13] M. McCue and E.R. Malinowski, *J. Chrom. Sci.* **21**, 229–234 (1983).
[14] M. Maeder, *Anal. Chem.* **59**, 527–530 (1987).
[15] G.G.R. Seaton and A.F. Fell, *Chromatographia* **247**, 208–216 (1987).
[16] B.G.M. Vandeginste, W. Derks and G. Kateman, *Anal. Chim. Acta* **173**, 253–263 (1985).
[17] O.M. Kvalheim and Y.-Z. Liang, *Anal. Chem.* **64**, 936–946 (1992).
[18] R.W. Rozett and E. Mclaughlin Petersen, *Anal. Chem.* **47**, 1301–1308 (1975).
[19] H.R. Keller and D.L. Massart, *Anal. Chim. Acta* **246**, 379–390 (1991).
[20] X.M. Tu, *J. Chemometrics* **5**, 333–343 (1991).

[Received for review 3 July 1992; revised manuscript received 21 July 1992]