Chemometric analysis of high performance liquid chromatography-diode array detection-electrospray mass spectrometry of 2- and 3-hydroxypyridine

Samantha Dunkerley,[a] John Crosby,[a] Richard G. Brereton,*[a] Konstantinos D. Zissis[a] and Richard E. A. Escott[b]

[a] *School of Chemistry, University of Bristol, Cantock's Close, Bristol, UK BS8 1TS*
[b] *SmithKline Beecham Pharmaceuticals, Old Powder Mills, Near Leigh, Tonbridge, Kent, UK TN11 9AN*

Received 8th June 1998, Accepted 14th August 1998

Triply coupled high performance liquid chromatography using diode array detection and positive ion electrospray mass spectrometry of 2- and 3-hydroxypyridine is presented. Considerations of the physical method for coupling the two detectors, the influence of pH on retention times, the cone voltage of the mass spectrometer and the linear concentration ranges are described. Data from both detectors are aligned and interpolated. The analyte mass spectra are reduced to 20 significant masses. Principal components plots on the raw, normalised and standardised data, derivatives to determine composition 1 regions, deconvolution and procrustes analysis to compare data from both detectors are discussed. Common trends in both mass spectral and diode array chromatograms are interpreted. This paper represents a new approach to common processing of chromatographic data from two detectors.

1 Introduction

It is becoming increasingly common to link two or more detectors to HPLC.[1–5] A popular configuration involves a diode array detector (DAD) and MS detector as both provide complementary information.[6,7] MS can provide insights into the structures of closely eluting compounds, whereas the DAD is often more sensitive and can provide information about co-elution and purity of chromatographic mixtures. In most cases the data from these two detectors are processed independently.[8,9] There can be associated problems such as different peak shapes, data acquisition rates and start of acquisition in each independent technique.[4,10] However, there is much information common to both methods, and in this paper we explore common data analysis methods using chemometrics, to see common trends in the data.

Many experimental and data processing problems must be overcome in order to obtain a meaningful consensus in the combined information from both detectors. In this paper we select a case study of 2- and 3-hydroxypyridine. These compounds have characteristic and different electronic absorption spectra (EAS), and being isomeric, MS characterisation is primarily by fragment ions, which elute closely in HPLC. Such compounds, therefore, provide a challenge for which triply coupled HPLC-DAD-MS is suited and were chosen as a case study to validate the methods described in this paper, which can be extended to more complex situations.

Typically, a user of such techniques performs independent processing of both the HPLC-DAD and HPLC-MS systems, often using different software. Where peaks are well separated, fairly crude approaches can be employed to obtain spectra of each chromatographic peak. However, if the compounds are closely eluting, it is often not at all clear which regions of a chromatogram represent composition 1, how many peaks are in a cluster (for more than two components) or the spectral characteristics for each compound in a mixture. There are certain features in common from both detectors and data analysis can take advantage of these aspects as described below.

2 Experimental

2.1 Chemicals and solvents

All chemicals and solvents used in the analyses described in this paper were of analytical reagent grade and HPLC grade, respectively, unless stated otherwise. The two compounds of particular interest are 2-hydroxypyridine, **I**, and 3-hydroxypyridine, **II** (Acros Organics, Springfield, NJ, USA), which were 97% and 98% pure, respectively. These compounds are isomers, both consisting of a pyridine ring with a single hydroxy substituent group, and have a molecular mass of 95.10. Ammonium acetate and acetic acid were purchased from Sigma (Poole, Dorset, UK) and HPLC grade methanol and ammonia from Rathburn Chemicals (Walkerburn, UK) and deionised water was prepared using a Milli-Q filtration unit (Millipore, Bedford, MA, USA).

2.2 Reagents and standard solutions

A 0.05 M CH_3COONH_4 solution was prepared in deionised water (3.854 g L^{-1}) and adjusted to the desired pH by adding 10% CH_3COOH and 10% NH_3 dropwise. From this, the mobile phase was prepared, containing 98% 0.05 M ammonium acetate and 2% methanol. All standard solutions of compounds **I** and **II** were prepared in the pH adjusted mobile phase from stock standard solutions of 10 mg mL^{-1}.

2.3 Apparatus and instrumentation

All HPLC was carried out using a Waters (Milford, MA, USA) system comprised of a Model 616 LC pump, a Model 717 Plus autosampler and heater/cooler and a Model 600S controller. Diode array detection was performed using a Model 996 PDA photodiode array detector optics unit (Waters). The HPLC-DAD system uses Millennium Session Manager Software (Version 2.15.01, Waters) with the Millenium 2010 Chromatography Manager Add-On (Version 2.10, Waters), that runs

under Windows (Version 3.1, Standard Mode, Microsoft, Seattle, WA, USA) on a 586 PC. The stationary phase was a 100 mm × 4.6 mm id C_{18} reversed-phase symmetry column packed with 3.5 μm particles (Waters, Watford, UK). The mobile phase was isocratic and consisted of the 0.05 M ammonium acetate–methanol buffer (98 + 2) described in Section 2.2. The flow rate through the column was 0.8 ml min^{-1} and the sample injection volume was 5 μL. The analyses were carried out at ambient temperature for a run time of 10 min. The EAS were recorded using the DAD at 1 s intervals, between 200 and 400 nm (1.2 nm bandwidth resolution), with a flow cell of pathlength 10 mm.

All MS was performed using a VG Quattro Mass Spectrometer (Fisons Instruments, Altrincham, UK) controlled using MassLynx software (Version 2.1, Micromass, Altrincham, UK), which runs under Windows (Version 3.1, Microsoft) on a 486 PC. The analyses were carried out using electrospray ionisation (ESI) in the positive ion mode for these studies. ESI produces a plume of charged droplets, the result of Coulombic repulsion, which enter into the mass analyser.[11] The mass spectra were recorded every 1 s, between 40 and 200 mass units, with a cone voltage of 50 V (the effect of which is described in Section 2.6), and a source temperature of 80 °C.

2.4 HPLC-DAD-MS coupling

Two possible arrangements for coupling the HPLC, the DAD and the MS systems are shown in Fig. 1. The first of these, illustrated in Fig. 1(a), shows the analyte stream being split after the EAS have been recorded, whereas in the second arrangement, Fig. 1(b), the stream is split immediately after column separation. There are two major considerations that need to be taken into account when deciding which is the best configuration to adopt, the flow rates and the tubing lengths, and these are discussed below.

The flow rate through the column to the DAD in conventional HPLC-DAD analysis is usually around 1 mL min^{-1}. Altering this flow rate can have a marked effect on the elution times, peak shapes and resolution. The optimum flow rate through the MS is recommended to be 0.1–0.4 mL min^{-1}. Within this range, higher flow rates can cause a reduction in the signal-to-noise ratio, whereas lower flow rates can give rise to noisy signals. Problems associated with flow rates above this range can include blocking of the probe, and overloading of the detector. It is recommended that the detector loading should not exceed 30 pmol μL^{-1} as this can influence the reproducibility, hence making quantification more difficult. In order to reduce the flow rate through the HPLC to the MS, a splitting device is used. This splitter consists of a Peek T-piece with finger-tight fittings, 0.50 mm ('20 thou') id × 1/16 in od (HICHROM, Reading, UK), fitted with Peek tubing of 0.25 mm ('10 thou') id × 1/16 in od (HICHROM). Splitting the stream as shown in Fig. 1(a) allows greater flexibility in the amount of sample split between the MS and the waste.

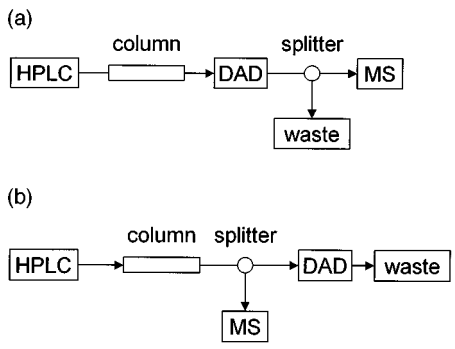

Fig. 1 Coupling arrangements for the HPLC-DAD and the MS systems with the splitting device between (a) the DAD and the waste and (b) the column and the DAD.

The second consideration that needs to be taken into account concerns the overall tube lengths. Adopting the coupling arrangement shown in Fig. 1(a) means that the analyte will have travelled a greater distance before reaching the MS. Test runs showed that the greater the distance between the DAD and the MS, the poorer is the MS resolution in comparison with the EAS elution profiles due to the extra mixing in the tubing. This mixing is similar to that found in flow injection analysis (FIA) and a consequence of laminar flow, which causes different concentrations across a peak commonly referred to as a concentration gradient.[12,13] The effect of mixing is illustrated in Fig. 2, which shows the profiles obtained in a two component system. The MS profile [Fig. 2(b)] is less well resolved than the EAS elution profile [Fig. 2(a)], mainly owing to the mixing and dead volume in the tubing and associated fittings. Furthermore, the apparent signal-to-noise ratio of the MS profile appears substantially worse than that obtained using DAD, which is due to the variability in ESI as opposed to other MS ionisation techniques. The poorer resolution observed in Fig. 2(b) underlines the necessity to keep the tubing volume to a minimum. Fig. 1(a) again illustrates that the extra tubing between the column and the MS in comparison with between the column and DAD is unavoidable in this particular arrangement, because the analyte stream has to pass through the DAD before being split to the MS. However, in the arrangement shown in Fig. 1(b) this is not the case, and even if it were difficult to keep the tube lengths to a minimum, it would at least be possible to have tubing of comparable lengths to each detector. Hence the profiles would be equally distorted and similar resolutions obtained, as illustrated in Fig. 2(c) and (d), which show the EAS and MS profiles, respectively.

An observation that is apparent from the profiles shown in Fig. 2 is that those obtained from the MS are much noisier than the corresponding EAS profiles. This is typical of ESI and is due to the variations in ionization efficiency, the detector and high chemical background noise, which can be partly overcome at higher voltages applied to the source (called the cone voltage) and a higher source temperature.[11] The effect of the cone voltage applied to the source on the fragmentation patterns of compounds **I** and **II** is discussed in Section 2.6. A final observation was that the ion count is significantly larger at the MS source using the coupling arrangement in Fig. 1(b) as opposed to that in Fig. 1(a). The actual split used in these studies involved the arrangement in Fig. 1(b) with a 50 : 50 split (both 250 mm), *i.e.*, a flow rate of 0.4 mL min^{-1} to both the EAS and MS detectors.

2.5 Effect of pH on retention time and EAS

The pH of the buffer has a significant effect on the elution times and EAS of compounds **I** and **II**. In order to demonstrate this, five mobile phases of pH 4.8, 4.9, 5.0, 5.1 and 5.2 were prepared as described in Section 2.2. Five standard solutions containing 1 mg mL^{-1} of compound **I** and 1 mg mL^{-1} of compound **II** were prepared for each of the pH adjusted buffers. These were analysed using the HPLC-DAD conditions described in Section 2.3.

The elution profiles for the mixtures at each pH are shown in Fig. 3. To summarise, at pH 4.8 [Fig. 3(a)] the peaks are partially overlapping, with compound **II** eluting first. Increasing the pH to 4.9 [Fig. 3(b)] increases the extent of this overlap, with **II** still eluting first. At pH 5.0 [Fig. 3(c)] the compounds are almost co-eluting, but **II** still elutes before **I**. At pH 5.1 [Fig. 3(d)] the compounds are completely co-eluting. A final increase to pH 5.2 [Fig. 3(e)] changes the order of elution, with **I** now eluting first. Further studies revealed that the separation with **I**

eluting prior to **II** is improved further still by increasing the pH above 5.2. The results show that the retention times of both compounds decrease with increase in pH, although that of **I** decreases more significantly. The corresponding EAS for **I** and **II** also change with pH; however, these are not illustrated in this paper as the pH was used purely as a means of changing the resolvability.

2.6 Effect of cone voltage on MS fragmentation

To determine the effect of the cone voltage on the fragmentation pattern, two 1 mg mL^{-1} standards of compounds **I** and **II** were prepared as described in Section 2.2 in a buffer adjusted to pH 4.8. The mass spectra were recorded (5 μL injections) using the conditions described in Section 2.3 with the cone voltage set at 30, 50 and 70 V. The results of the analysis are shown in Fig. 4 (it should be noted that the mass spectra illustrated in Fig. 4 were baseline corrected as described in Section 3.1.3.2).

The mass spectra for compound **I** are shown in Fig. 4(a)–(c). The mass spectrum in Fig. 4(a) shows that there is virtually no fragmentation of the molecular ion peak, *m/z* 96, at a cone voltage of 30 V, with the only other mass to have an intensity greater than 10% that of the parent peak being at *m/z* 155 (an adduct). Increasing the cone voltage to 50 V [Fig. 4(b)] causes a significant increase in the *m/z* 78 peak that corresponds to the loss of H_2O from the protonated molecule and is now greater in intensity than the parent peak. This increase also gives rise to a second fragment at *m/z* 51 with other masses greater than 10% being at *m/z* 155 and 191 (a dimer). Setting the cone voltage to 70 V [Fig. 4(c)] increases the proportion of the *m/z* 51 peak to the *m/z* 78 peak, with the molecular ion peak again being the most significant. As before, the *m/z* 155 and 191 peaks are of greater than 10% relative intensity.

The mass spectra for compound **II** are shown in Fig. 4(d)–(f), and again illustrate that there is no fragmentation of the parent peak up to 30 V [Fig. 4(d)] with the only other mass greater than 10% being at *m/z* 155. At a cone voltage of 50 V [Fig. 4(e)] there are two main fragments, one at *m/z* 41 and the other at *m/z* 68. Finally, at a cone voltage of 70 V [Fig. 4(f)], the fragment at *m/z* 41 has increased slightly and that at *m/z* 68 has decreased to below 10% relative intensity. As before, the only other significant peak is at *m/z* 155.

In this example, most structurally discriminant information is in the fragment ions, so a cone voltage of 50 V was chosen.

2.7 Linearity ranges

Chemometric factor analysis using multiple linear regression requires a linear concentration range. In order to determine the linear range for the DAD, the stock standard solution (Section 2.2) was used to prepare 0.25, 0.50, 0.75, 1.00, 1.25 and 1.50 mg mL^{-1} solutions at pH 4.8 for both compounds **I** and **II**. Injections of 5, 10 and 15 μL were made for each of the standards for each compound using the conditions described in Section 2.3. The λ_{max} wavelengths for compounds **I** and **II** are 292 and 281 nm, respectively. Graphs of the absorbance at λ_{max} against injection volume and concentration, for each compound, indicate the linear absorbance regions across the injection volume and concentration. The linear range for compound **I** was found to be up to approximately 1.5 absorbance and corresponds to a 5 μL injection of a 1.50 mg mL^{-1} standard solution. The linear range for compound **II** was found to be up to approximately 1.5 absorbance and corresponds to a 5 μL injection of a 1.25 mg mL^{-1} standard solution. It is important to note that the value quoted for compound **II** is extremely sensitive to changes in pH.

For determining the linear range of the MS, standard solutions of 0.50, 1.00, 1.50, 2.00, 2.50, 3.00, 4.00 and 5.0 mg mL^{-1} were prepared for both compounds from the stock standard solution (Section 2.2). These were analysed using the conditions described in Section 2.3 (10 μL injections). The linear ranges were determined at two ions for each compound: the common parent molecular ion, at *m/z* 96, and an independent daughter ion, at *m/z* 51 for **I** and *m/z* 41 for **II**. For compound **I** the ion count was linear up to 2.00 mg mL^{-1} and for compound **II** up to 3.00 mg mL^{-1}.

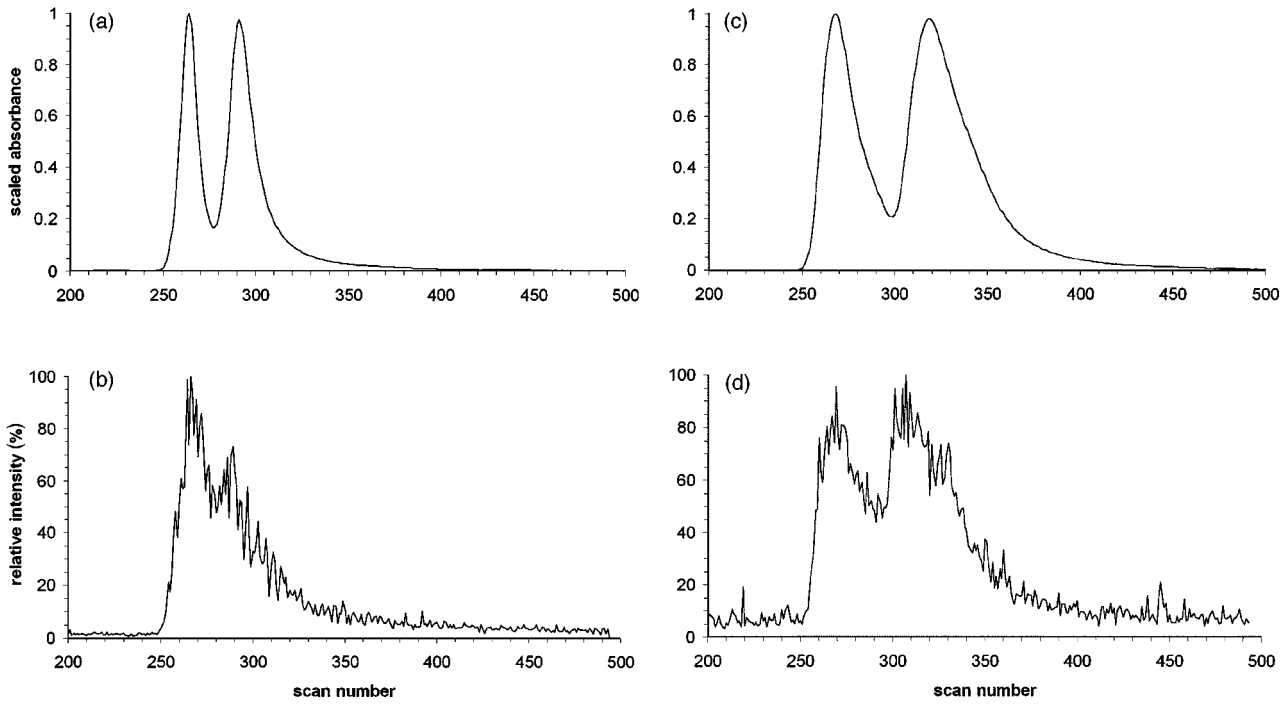

Fig. 2 Comparison of DAD and MS profiles with different coupling arrangements: (a) DAD and (b) MS profile with the coupling as in Fig. 1(a), and (c) DAD and (d) MS profile with the coupling as in Fig. 1(b).

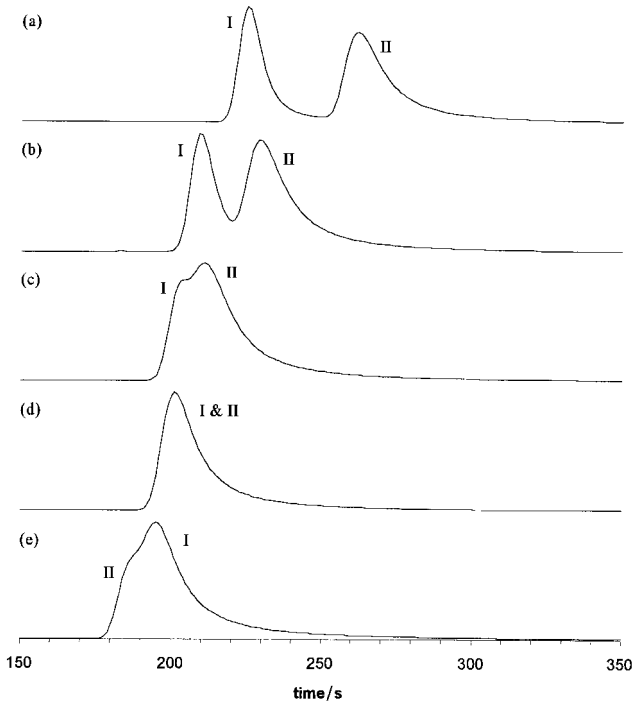

Fig. 3 Scaled HPLC-DAD elution profiles of a 1 + 1 mixture of compounds **I** and **II** at pH (a) 4.8, (b) 4.9, (c) 5.0, (d) 5.1 and (e) 5.2.

2.8 Software

2.8.1 Decoding programs. Two programs are required to decode the data acquired from the HPLC-DAD analysis. The first of these is the 2010 DDE Assistant for Raw Data macro (Version 2.10, Waters) that runs in Excel (Version 5.0a, Microsoft). This is used in conjunction with the Millennium software in order to extract the appropriate data from the Oracle database. The resulting file is then converted to a matrix using a second VBA macro that was written in-house. The MS data are decoded using a C^{++} program written by Dr. R. L. Erskine.

2.8.2 Data analysis. The preprocessing methods, chemometric algorithms and other techniques were written in-house in MatLab (Version 4.2c.1, Math Works, South Natick, MA, USA) and as VBA macros for Excel.

3 Chemometrics and data analysis

3.1 Data preprocessing

3.1.1 Interpolation. The DAD and MS detectors were set-up to acquire spectra every 1 s. In practice, it was found that the scan intervals were not even throughout a run, and were different for each instrument. In addition, there is a small delay before the first scan is recorded. Factors that may influence the

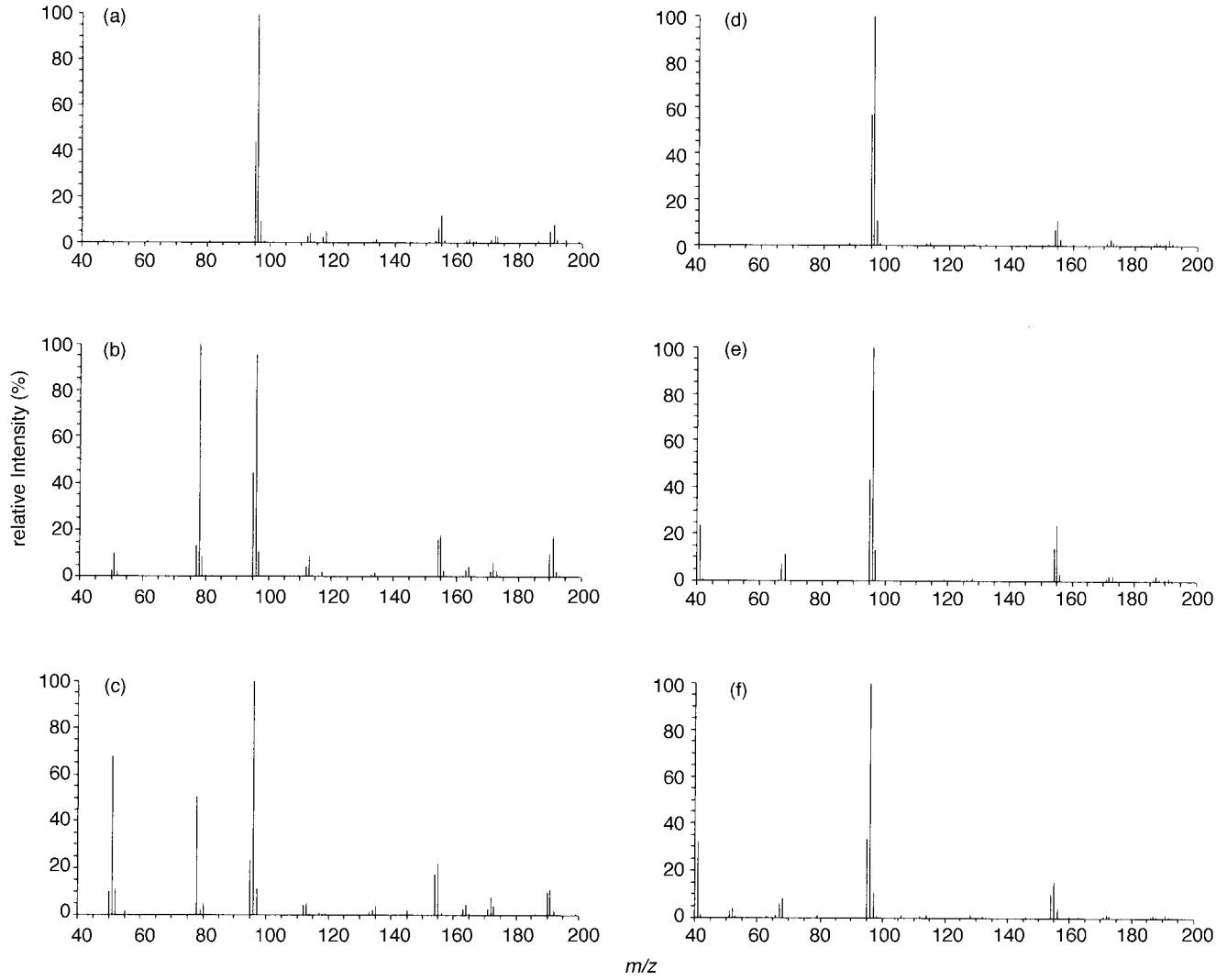

Fig. 4 ESI-MS of compounds **I** and **II** at increasing cone voltage. (a)–(c) Compound **I** and (d)–(f) compound **II** at 30, 50 and 70 V, respectively.

scan interval and offset include the time required to acquire each set of data (itself dependent on the scan range and resolution) and the speed of data transfer. As a result of these discrepancies, it is necessary to interpolate the data in order to put them on a comparable time-scale of 1 s.

3.1.2 Alignment. Although the HPLC-DAD and the MS systems are started simultaneously, the DAD does not begin to acquire data until sample injection is complete. It is therefore necessary to align the two data sets so that they can be compared directly. This requires shifting the HPLC-DAD data forward by a predetermined number of interpolated scans, a shift that varies slightly for each run. At this stage *I* common interpolated points in time are selected for both the HPLC-DAD and HPLC-MS data.

3.1.3 Baseline correction and background suppression. *3.1.3.1 HPLC-DAD.* It is found that small negative absorbance values are exhibited at some of the wavelengths obtained in HPLC-DAD analysis. To overcome this problem it is necessary to correct for all wavelengths separately as each one responds differently. This procedure first involves selecting two baseline regions, one after and one before compound elution. Linear regression is performed across this baseline region to obtain the coefficients of the line using the least squares equation,

$$\boldsymbol{b} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y} \quad (1)$$

where $\boldsymbol{X}$ is a matrix whose first column are 1 s and whose second column are the scan times in the baseline regions, $\boldsymbol{y}$ is a vector of the absorbances for the baseline regions and $\boldsymbol{b}$ is a row vector of two coefficients. These coefficients are used to construct a profile for the entire scan range for each wavelength, which is then subtracted from the original data matrix to obtain ${}^{\mathrm{b,DAD}}\boldsymbol{X}$.

3.1.3.2 HPLC-MS. Baseline correction is essential for the MS data because of the cumulative ion effect between analyses, ion suppression and the high solvent baseline. With regard to the high solvent baseline, it is found that masses that do not vary with time are often more intense than those which change during compound elution. Another problem is that of ion count suppression, in which the relative intensity of some masses decreases during the compounds' elution. In this particular case, the total ion current (TIC) could not be used to determine these baseline regions owing to the high solvent contribution and suppression that meant that the elution region was not actually visible in the TIC. As an alternative, the parent peak common to both compounds, *m/z* 96, was used to determine the baseline region. Owing to problems with tailing, a baseline region after elution could not be used for correction. Having identified a suitable baseline region, the mean intensity for each mass, b_j, is determined in this region and subtracted from the corresponding masses for the whole run, *i.e.*,

$${}^{\mathrm{b,MS}}x_{i,j} = {}^{\mathrm{MS}}x_{i,j} - \sum_{i=1}^{I} b_j \quad (2)$$

3.1.4 Significant mass selection. It is common to use a low cut-off mass owing to the large number of volatiles detected, which, owing to the low molecular mass of the compounds, was taken as *m/z* 40. A second way in which the size of the data is reduced is by removing those masses which appear to have no relevance during the compounds' elution. The data are further reduced, in order to alleviate sparse matrices, or those in which a large number of the masses remain unchanged with time. This is achieved by determining the most significant masses for the MS data, by dividing the ion count for the individual masses by the total ion count:

$$s_j = \frac{\sum_{i=1}^{I} {}^{\mathrm{b,D}}x_{i,j}}{\sum_{i=1}^{I}\sum_{j=1}^{{}^{\mathrm{raw,MS}}J} {}^{\mathrm{b,D}}x_{i,j}} \quad (3)$$

These ratios s_j, are then sorted into descending order and the significance against rank is plotted. The top ${}^{\mathrm{MS}}J$ masses are selected according to graphical criteria as discussed in Section 4.1.4. Note that the masses that are intense in the original HPLC-MS dataset, but do not vary substantially, have been reduced in intensity, thereby eliminating any non-significant masses.

3.1.5 Normalisation. Normalisation is a common method of preprocessing in which each value in a scan (row) is divided by its sum, so that the total for each scan becomes equal to one, *i.e.*, the scans are transformed so that effects of concentration are removed. For example, a baseline corrected, data matrix, ${}^{\mathrm{b,D}}\boldsymbol{X}$, would be normalised using the equation:

$${}^{\mathrm{n,D}}x_{i,j} = \frac{{}^{\mathrm{b,D}}x_{i,j}}{\sum_{j=1}^{J} {}^{\mathrm{b,D}}x_{i,j}} \quad (4)$$

It is essential that ions with negative intensity are removed for this procedure to be meaningful, hence for HPLC-MS this could not be performed without prior baseline correction.

3.1.6 Standardisation. Another preprocessing technique is standardisation in which all the variables (columns) are transformed on to the same scale by subtracting the variable mean from the values in that variable and dividing this by the standard deviation for that variable. This is particularly important when comparing results from different sources, and in MS where intensities can differ by an order of magnitude. For example, the baseline corrected data matrix, ${}^{\mathrm{b,D}}\boldsymbol{X}$, would be standardised using

$${}^{\mathrm{s,D}}x_{i,j} = \frac{{}^{\mathrm{b,D}}x_{i,j} - {}^{\mathrm{b,D}}\bar{x}_j}{\sqrt{\sum_{i=1}^{I} \frac{\left({}^{\mathrm{b,D}}x_{i,j} - {}^{\mathrm{b,D}}\bar{x}_j\right)^2}{(I-1)}}} \quad (5)$$

3.2 Principal components analysis

Principal components analysis (PCA) is a method commonly used for reducing the dimensionality of measured data.[14–16] For example, an $I \times J$ data matrix, $\boldsymbol{X}$, can be decomposed into its $I \times K$ scores matrix, $\boldsymbol{T}$, its $K \times J$ loadings matrix, $\boldsymbol{P}$, and the residual errors $I \times J$ matrix, $\boldsymbol{E}$, *i.e.*,

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \quad (6)$$

where I is the number of scans, J is the number of variables and K is the number of principal components (PCs). The scores relate to the compounds' concentrations, the loadings to their spectra and the number of significant PCs should equal the number of compounds. Non-iterative partial least squares (NIPALS), which was originally developed by Wold and Lyttkens,[17] extracts the PCs from the data matrix one at a time. The first PC is the most significant, or descriptive, with each successive PC describing less and less of the information contained within the original data matrix.

3.3 Derivatives

Derivative plots are one of a number of methods that can be used to identify the different composition of regions in a chromatogram of mixtures. The implementation here consists of a number of steps illustrated by a seven point simulation in Fig. 5 and described as follows.

1. The first step is to normalise the data using eqn. (4). The key to the success of this method is to scan peaks to see if the spectra differ significantly; normalisation provides a suitable method for putting all the scans on to a common scale, hence giving them equal weighting, as illustrated in Fig. 5(b) and (c).
2. The five point Savitsky–Golay smoothed first derivative[18–20] is calculated at each wavelength for the HPLC-DAD data using

$$^{\mathrm{DAD}}d_{i,j} = (-2^{\mathrm{n,DAD}}x_{i-2,j} - {}^{\mathrm{n,DAD}}x_{i-1,j} + {}^{\mathrm{n,DAD}}x_{i+1,j} + 2^{\mathrm{n,DAD}}x_{i+2,j})/10 \quad (7)$$

and the nine point Savitsky–Golay smoothed first derivative is calculated at each mass number for the HPLC-MS data using

$$^{\mathrm{MS}}d_{i,j} = (-4^{\mathrm{n,MS}}x_{i-4,j} - 3^{\mathrm{n,MS}}x_{i-3,j} - 2^{\mathrm{n,MS}}x_{i-2,j} - {}^{\mathrm{n,MS}}x_{i-1,j} + {}^{\mathrm{n,MS}}x_{i+1,j} + 2^{\mathrm{n,MS}}x_{i+2j} + 3^{\mathrm{n,MS}}x_{i+3j} + 4^{\mathrm{n,MS}}x_{i+4,j})/60 \quad (8)$$

The effect is shown in Fig. 5(d). If a peak is pure over five or nine points in time, the normalised spectral intensity should not vary significantly resulting in a derivative close to zero.

3. It is desirable to compute derivatives at each of the wavelengths or masses to provide a consensus over all wavelengths (or mass number), because of the difference in the magnitude of intensity (especially apparent in the significant masses). Therefore, the individual derivatives are renormalised to a common scale to allow approximately equal significance at each mass or wavelength. The absolute values of the derivatives [Fig. 5(e)] are normalised for each variable as shown in Fig. 5 as opposed to each time as in step 1, using the equation

$$^{\mathrm{n}}\left|{}^{\mathrm{D}}d_{i,j}\right| = \frac{\left|{}^{\mathrm{D}}d_{i,j}\right|}{\sum_{i=1}^{I}\left|{}^{\mathrm{D}}d_{i,j}\right|} \quad (9)$$

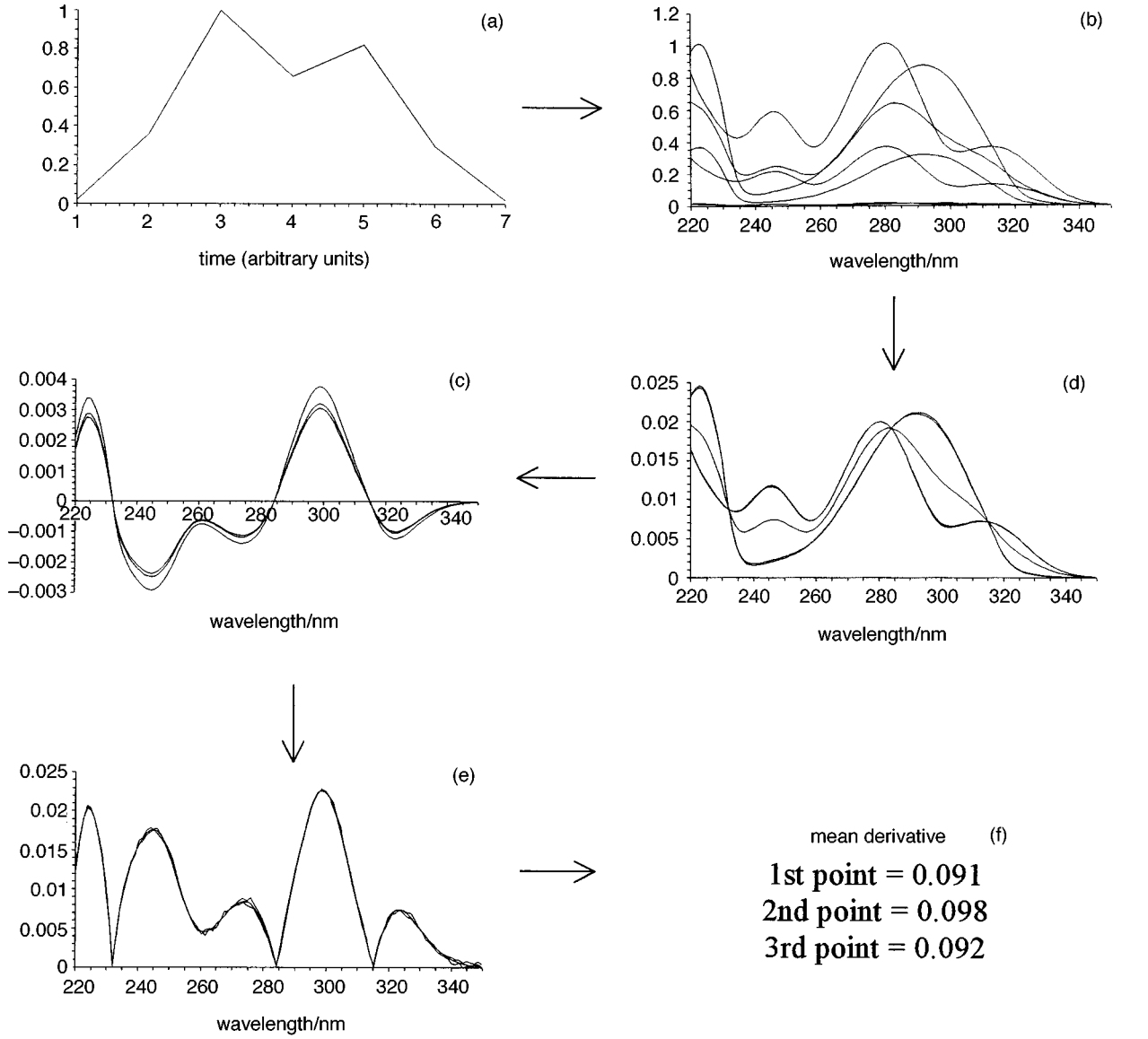

Fig. 5 Worked example of derivative process starting with (a) the elution profile, (b) the individual spectra, (c) the normalised spectra, (d) the derivative plots, (e) the renormalised absolute derivative plots and (f) the mean derivative values.

4. Finally, the weighted average of the derivative (Fig. 5) at each point in time is calculated using

$$ {}^{\mathrm{D}}\bar{d}_j = \frac{\sum_{i=1}^{I} {}^{\mathrm{n}}\left|{}^{\mathrm{D}}d_{i,j}\right|}{J} \quad (10) $$

The mean derivative values for the simulation are shown in Fig. 5(f). Composition 1 and 2 regions are always clear of the noise regions. Pure peaks are indicated by low absorbance values; note that the size of the derivatives increases both in the noise regions and the composition 2 region.

3.4 Deconvolution

The purpose of deconvolution is to determine the estimated concentration profiles, $\hat{\boldsymbol{C}}$, and spectra, $\hat{\boldsymbol{S}}$, of the components in the data matrix, $\boldsymbol{X}$.[14,21,22] The concentration profiles (related to the scores) and the spectral profiles (related to the loadings) have a similar relationship to that shown in eqn. (6):

$$ \boldsymbol{X} = \hat{\boldsymbol{C}} \cdot \hat{\boldsymbol{S}} + \boldsymbol{E} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \quad (11) $$

where $\boldsymbol{X}$ is an $I \times J$ data matrix, $\hat{\boldsymbol{C}}$ is an $I \times K$ matrix of concentrations, $\hat{\boldsymbol{S}}$ is a $K \times J$ matrix of spectra and $\boldsymbol{E}$ is an $I \times J$ matrix of residuals. The deconvolution steps are different for HPLC-DAD and HPLC-MS, with the former starting with spectral profiles as obtained from the composition 1 regions of the derivative plots and the latter starting with the concentration profiles of the two most diagnostic masses. The steps are described below.

3.4.1 HPLC-DAD. Multiple linear regression (MLR) is commonly used in HPLC-DAD for factor analysis. As the two compounds absorb at all wavelengths in the region of interest, it is not possible to start with the elution profiles. Instead, as the spectra are very diagnostic and can be fully characterised at all wavelengths it is necessary to start with the pure spectra. To deconvolute the DAD data matrix initial estimates of the EAS, ${}^{1,\mathrm{DAD}}\hat{\boldsymbol{S}}$, are required. These are obtained from the derivative plots by calculating the mean spectrum in the composition 1 region determined for each component which are then entered into each row of the matrix, ${}^{1,\mathrm{DAD}}\hat{\boldsymbol{S}}$. The pseudo-inverse is used to obtain a first estimate of the concentration profiles, ${}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}}$, for the baseline corrected data matrix, ${}^{\mathrm{b,DAD}}\boldsymbol{X}$:

$$ {}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}} = {}^{\mathrm{b,DAD}}\boldsymbol{X} \cdot {}^{1,\mathrm{DAD}}\hat{\boldsymbol{S}}'({}^{1,\mathrm{DAD}}\hat{\boldsymbol{S}} \cdot {}^{1,\mathrm{DAD}}\hat{\boldsymbol{S}}')^{-1} \quad (12) $$

The first estimate of the concentration profiles, ${}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}}$, is then used to determine the second estimate of the EAS, ${}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}}$, using

$$ {}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}} = ({}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}}' \cdot {}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}})^{-1} \cdot {}^{1,\mathrm{DAD}}\hat{\boldsymbol{C}}' \cdot {}^{\mathrm{b,DAD}}\boldsymbol{X} \quad (13) $$

This improved estimate of the EAS spectra, ${}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}}$, is then regressed back on to the baseline corrected data matrix, ${}^{\mathrm{b,DAD}}\boldsymbol{X}$, to obtain a second estimate of the concentration profiles, ${}^{2,\mathrm{DAD}}\hat{\boldsymbol{C}}$, using the equation

$$ {}^{2,\mathrm{DAD}}\hat{\boldsymbol{C}} = {}^{\mathrm{b,DAD}}\boldsymbol{X} \cdot {}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}}' \cdot ({}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}} \cdot {}^{2,\mathrm{DAD}}\hat{\boldsymbol{S}}')^{-1} \quad (14) $$

No significant improvement is obtained by further regression.

3.4.2 HPLC-MS. The situation with HPLC-MS contrasts with that with HPLC-DAD. Very diagnostic masses can be found for each component, allowing a good first estimate of elution profiles, which is not possible for HPLC-DAD. However, because the spectra obtained by HPLC-MS are considerably noisier and are unstable, it is not easy to obtain a good first guess of the mass spectra. To deconvolute the MS data an initial estimate of the concentration profiles, ${}^{1,\mathrm{MS}}\hat{\boldsymbol{C}}$, is required. For this the most diagnostic mass of each compound, determined using the PCA loadings plot,[21,23] is entered into each column of the matrix, ${}^{1,\mathrm{MS}}\hat{\boldsymbol{C}}$. Again, the pseudo-inverse is used to obtain a first estimate of the spectra, ${}^{1,\mathrm{MS}}\hat{\boldsymbol{S}}$, for the reduced mass, baseline corrected data matrix, ${}^{\mathrm{b,MS}}\boldsymbol{X}$:

$$ {}^{1,\mathrm{MS}}\hat{\boldsymbol{S}} = ({}^{1,\mathrm{MS}}\hat{\boldsymbol{C}}' \cdot {}^{1,\mathrm{MS}}\hat{\boldsymbol{C}})^{-1} \cdot {}^{1,\mathrm{MS}}\hat{\boldsymbol{C}}' \cdot {}^{\mathrm{b,MS}}\boldsymbol{X} \quad (15) $$

The first estimate of the mass spectra, ${}^{1,\mathrm{MS}}\hat{\boldsymbol{S}}$, is then used to determine the second estimate of the concentration profiles, ${}^{2,\mathrm{MS}}\hat{\boldsymbol{C}}$, using

$$ {}^{2,\mathrm{MS}}\hat{\boldsymbol{C}} = {}^{\mathrm{b,MS}}\boldsymbol{X} \cdot {}^{1,\mathrm{MS}}\hat{\boldsymbol{S}}' \cdot ({}^{1,\mathrm{MS}}\hat{\boldsymbol{S}} \cdot {}^{1,\mathrm{MS}}\hat{\boldsymbol{S}}')^{-1} \quad (16) $$

This improved estimate of the concentration profiles, ${}^{2,\mathrm{MS}}\hat{\boldsymbol{C}}$, is then regressed onto the baseline corrected data matrix containing all the masses, ${}^{\mathrm{all,b,MS}}\boldsymbol{X}$, to obtain a second estimate of the mass spectra, ${}^{2,\mathrm{MS}}\hat{\boldsymbol{S}}$, using the equation

$$ {}^{2,\mathrm{MS}}\hat{\boldsymbol{S}} = ({}^{2,\mathrm{MS}}\hat{\boldsymbol{C}}' \cdot {}^{2,\mathrm{MS}}\hat{\boldsymbol{C}})^{-1} \cdot {}^{2,\mathrm{MS}}\hat{\boldsymbol{C}}' \cdot {}^{\mathrm{all,b,MS}}\boldsymbol{X} \quad (17) $$

3.5 Procrustes analysis

Procrustes analysis is a statistical method used to compare two dimensional, or two component, datasets.[24,25] The scores matrices obtained from the PCA of the HPLC-DAD and HPLC-MS data matrices are equivalent in dimensionality; however, they are different in size and in orientation. Despite this, it is possible to overlay them onto one another by rotating, scaling and reflecting them in such a way that there is maximum overlap.

In some cases it may be necessary to reflect the data, which simply involves multiplying one or both of either the DAD or MS score vectors by -1. This is because it is not possible to control the sign of the PCs. The second step involves rotating and scaling the scores. To rotate and scale the DAD scores on to the MS scores the following equations are used:

$$ {}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,1} = \mu[(\cos\theta) \cdot {}^{\mathrm{DAD}}\boldsymbol{t}_{i,1} - (\sin\theta) \cdot {}^{\mathrm{DAD}}\boldsymbol{t}_{i,2}] \quad (18) $$

$$ {}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,2} = \mu[(\sin\theta) \cdot {}^{\mathrm{DAD}}\boldsymbol{t}_{i,1} + (\cos\theta) \cdot {}^{\mathrm{DAD}}\boldsymbol{t}_{i,2}] \quad (19) $$

where μ is the scaling factor, θ is the angle of rotation, ${}^{\mathrm{DAD}}\boldsymbol{t}_{i,1}$ and ${}^{\mathrm{DAD}}\boldsymbol{t}_{i,2}$ are the original first and second DAD scores for the ith scan and ${}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,1}$ and ${}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,2}$ are the first and second DAD scores for the ith scan after procrustes analysis.

To measure the effectiveness of the procrustes analysis in this particular example, the error between the transformed DAD scores and the non-transformed MS scores is measured using

$$ {}^{\mathrm{p}}\boldsymbol{E} = \sqrt{\frac{\sum_{k=1}^{2}\sum_{i=1}^{I}\left({}^{\mathrm{MS}}\boldsymbol{t}_{i,k} - {}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,k}\right)^2}{I}} \quad (20) $$

where ${}^{\mathrm{p}}\boldsymbol{E}$ is the error, ${}^{\mathrm{MS}}\boldsymbol{t}_{i,k}$ is the score of the ith scan and kth component for the original MS scores, ${}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,k}$ is the score of the ith scan and kth component for the transformed DAD scores and I is the number of scans.

4 Results

For the chemometric and other data analysis described in this section, the standard solution containing 1 mg ml^{-1} each of compounds **I** and **II** analysed at pH 4.9 was used.

4.1 Preprocessing

4.1.1 Interpolation. Interpolation was performed as described in Section 3.1.1 to correct for the scan intervals, which were approximately 1.00 s for the DAD and 1.24 s for the MS.

4.1.2 Alignment. The HPLC-DAD and HPLC-MS raw data matrices were aligned as described in Section 3.1.2, in which it was necessary to shift the HPLC-DAD data forward by 47 interpolated scans, *i.e.*, so that scan 1 becomes scan 48.

4.1.3 Baseline correction. Baseline correction was performed on the HPLC-DAD data as described in Section 3.1.3.1 using the first and last 20 scans covering 40 s. Baseline correction was then performed as on the HPLC-MS described in Section 3.1.3.2 using the baseline region of 1–240 s. A much longer baseline region was required for MS data because of a considerably worse and more variable baseline. The resultant matrices are denoted by $^{b,D}\boldsymbol{X}$.

4.1.4 Significant masses and wavelength reduction. The significance of each mass was determined using the criterion described in Section 3.1.4. The significance of each mass was sorted and a graph of significance ratio against mass number for the first 30 masses is shown in Fig. 6. Overall there is in the region of 30 masses with significant positive intensities and about 50 masses that are negative or close to zero after baseline correction.

The first 20 positively significant masses were selected for further analysis and are listed in Table 1 with those unique to compounds **I** and **II** indicated. The baseline corrected data were reduced to the interpolated scan range of 200–600 s for both the DAD and MS data. The actual wavelength range used for chemometric analysis was 220–350 nm as prior to this region there are interferences from the mobile phase, and after this range the compounds do not absorb. A plot comparing the summed DAD and TIC of the significant masses is shown in Fig. 7.

4.2 Principal components analysis

The scores and loadings were calculated for the baseline corrected DAD and MS data matrices, $^{b,D}\boldsymbol{X}$, and the corresponding normalised and standardised versions using the NIPALS algorithm described in Section 3.2. The results for the DAD data are shown in Fig. 8 and those for the MS data in Fig. 9. The scores plots show the relationship between time and the loadings plots show the relationship between the variables.

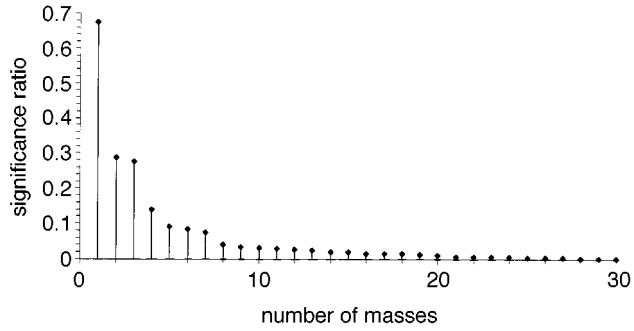

Fig. 6 Plot of the first 30 sorted significance ratios *versus* the number of masses.

Table 1 Twenty most significant masses with those unique to 2-hydroxypyridine (**I**) or 3-hydroxypyridine (**II**) indicated in parentheses

Rank	*m/z*	Rank	*m/z*
1	96	*11*	113 (**I**)
2	95	*12*	172
3	78 (**I**)	*13*	51 (**I**)
4	155	*14*	67 (**II**)
5	97	*15*	156
6	41 (**II**)	*16*	190 (**I**)
7	154	*17*	173
8	68 (**II**)	*18*	164 (**I**)
9	191 (**I**)	*19*	112 (**I**)
10	79 (**I**)	*20*	171

For the baseline corrected data each linear segment of the scores plot indicates a pure compound in the mixture with the extreme points of each segment corresponding to the peak maxima [Fig. 8(a) and 9(a)]. The DAD scores plot is much better resolved than the MS scores plot, which is as expected when comparing the elution profiles (Fig. 7). In all cases the top right hand limb corresponds to the earlier eluting compound **II** and the bottom right hand limb to **I**.

The MS loadings plot in Fig. 9(b) is fairly hard to interpret. The most intense peaks are further from the origin, and there are approximately three limbs. A lower limb containing *m/z* 78 contains mass diagnostic to **I**, whereas *m/z* 41 corresponds to **II**. Masses in the central limb (*m/z* 95 and 96) are common to both compounds. Unfortunately, many diagnostic masses are weak in intensity and not so easy to distinguish in the graphs. However, *m/z* 51 and 113, for example, appear to lie on the same line as **I**, so some further information can be gleaned. The DAD loadings plot shown in Fig. 8(b) is more interesting, as both the compounds absorb at all the wavelengths chosen for this analysis. These plots can be interpreted with reference to the EAS [the deconvoluted spectra are given in Fig. 13(b)]. For example, compound **II** shows a pure but weakly absorbing band centred at 246 nm, which is clearly indicated in Fig. 8(b). Another maximum is exhibited at 281 nm, which again forms a clear turning point in the graph. However, because **I** also absorbs at this maximum, the direction of this turning point is half way between the direction of pure **II** and pure **I**. Above 320 nm, compound **II** dominates so the graph turns back in on itself. Compound **I** exhibits an absorbance maximum at 223 nm, as shown by the turning point in Fig. 8(b), but this is not as pure as 246 nm, despite the lower intensity, because the spectrum of **I** decreases in intensity more rapidly than **II** after 223 nm.

The normalised plots for HPLC-DAD are illustrated in Fig. 8(c) and (d). It is important to pick regions dominated by peaks for successful application of this approach, and in this paper we use the region in time of 250–300 s. The loadings plots do not change significantly, because the scaling is in the direction of elution time. However, the scores plots show a dramatic change, and now all the points lie on a straight line. The top of the line corresponds to **II** and the bottom to **I**. Regions of co-elution (265–273 s) lie in the centre of the line, with the pure points at the end.

The normalised scores plot for HPLC-MS is given in Fig. 9(c) and, again, two groups of elution times can be distinguished. However, the MS data are substantially more noisy, so the scores do not lie so clearly on a straight line. It is, though, precisely this sort of data where chemometric techniques can help, and it is still fairly obvious in which regions co-elution takes place (268–273 s), which is extremely close to the region identified in HPLC-DAD. This also corresponds to the region

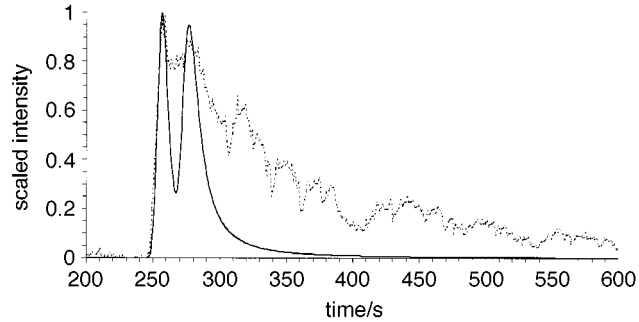

Fig. 7 Comparison of the summed DAD profile (solid line) and the TIC for the 20 most significant masses (dotted line).

demonstrated in the derivative plots as discussed below. Simple inspection of the chromatograms would not obviously show co-elution, unlike for HPLC-DAD, because of the inherently much more noisy peak shapes, but these plots are excellent indications that the regions of co-elution and so chromatography are, in fact, fairly similar for both detectors.

For the standardised data, there is little difference made to the scores plots [Fig. 8(e) and 9(e)]. However, the loadings plots exhibit very different characteristics. Those for HPLC-DAD fall mainly on an arc of a circle, suggesting that the data are well represented by two components within the region of the chromatogram analysed. As a result of standardisation, each wavelength has equal significance and the most important factor influencing the position in the loadings plot is the ratio of the absorbances for each compound at each wavelength. For example, at 334 nm, the ratio of absorbances between **II** and **I** is a maximum, so this wavelength is on the top of the arc. In contrast, the ratio of absorbances between **I** and **II** is a maximum at 301 nm, at the other extreme end of the arc. The distance from the end of the arc is plotted against the ratio of absorbance of compound **II** to **I** in Fig. 10 and it can be seen that these are directly related, as expected.

The standardised HPLC-MS loadings plots do not fit exactly on an arc of a circle because of the greater level of noise, meaning that the data are not exactly modelled by two PCs. However, the significance of the different masses is easier to see, as they are no longer crowded close to the origin. The reconstructed mass spectra are discussed below and illustrated in Fig. 11(b) and (c). A small ion at around *m/z* 68 is diagnostic of compound **II** and can be seen in the top left hand corner of Fig. 9(f). However, the ion cluster around *m/z* 79 is diagnostic of **I** in addition to the small ion at *m/z* 51 in the lower half of the graph. The common ions are much more clearly in the centre of the loadings plot. Note that less intense ions may, nevertheless, be highly diagnostic, hence *m/z* 51, for example, is clearly indicated in Fig. 9(f) in contrast to Fig. 9(b). Standardised loadings plots are more useful than raw loadings plots, especially for isomers where there will be dominant and common fragment ions, such as the molecular ion. Note that there are a number of high molecular mass ions possibly due to the background, but the only ones that appear to be significantly correlated to either structure at *m/z* 112 and 113, which are characteristic of compound **I**, and probably due to addition of OH (= $M^{\cdot+}$ + 16 and MH^+ + 16) and loss of H, respectively.

4.3 Derivative plots

The results of the derivative analysis for the DAD and MS data are shown in Fig. 12 and 13, respectively. The purpose of the

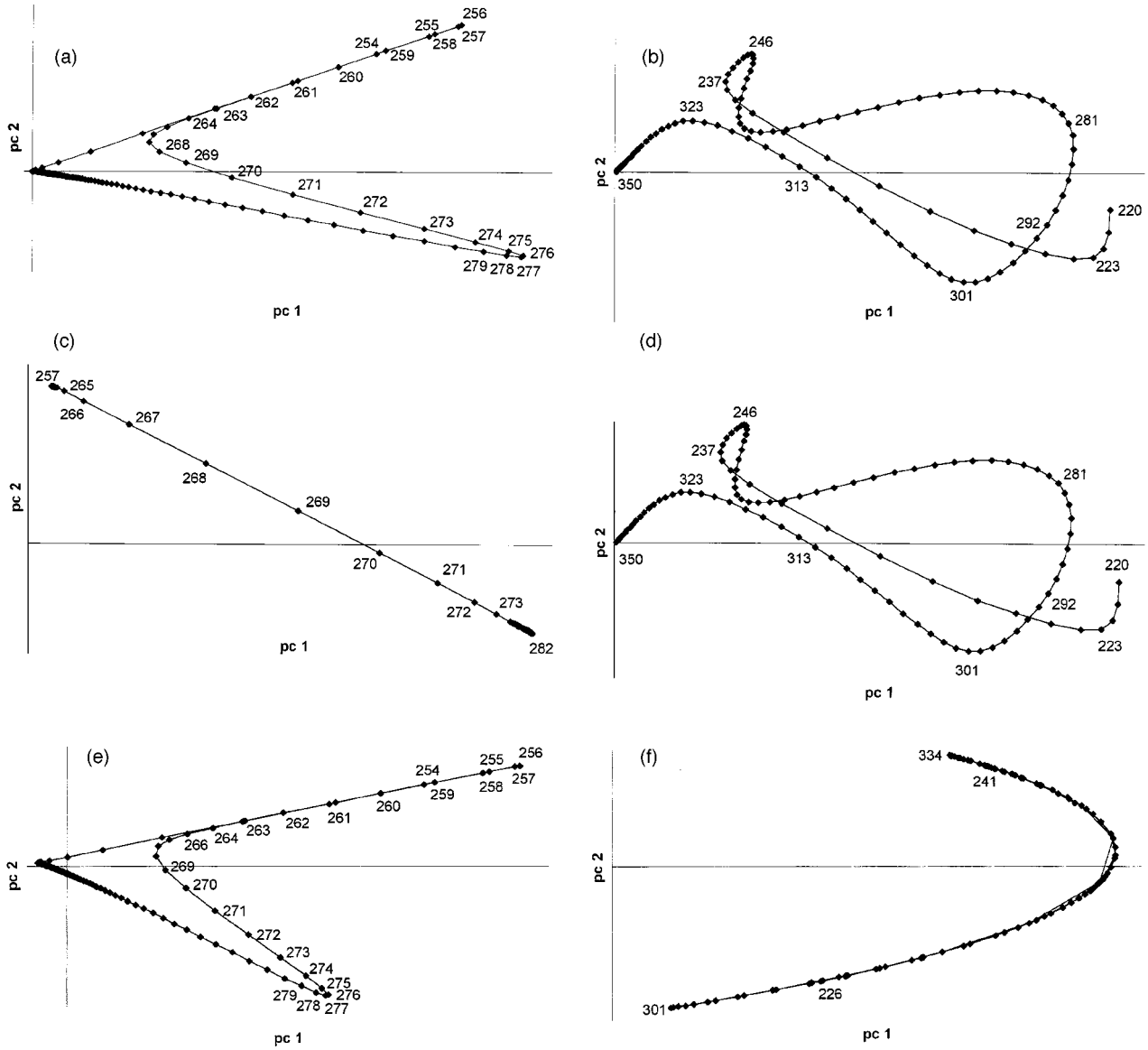

Fig. 8 Principal components analysis of the HPLC-DAD data: (a) non-transformed scores plot; (b) non-transformed loadings plot; (c) normalised scores plot; (d) normalised loadings plot; (e) standardised scores plot; and (f) standardised loadings plot.

derivative plots is to determine which regions in time are compositions 1 and 2 in the profile. In addition, in this paper the logarithm of the DAD derivative plot [Fig. 12(b)] is used to find the purest spectral profiles so that they can be used as first estimates for deconvolution.

For the DAD data there is a clear composition 2 region between the peaks, as expected. The logarithmic plot is even clearer, and suggests that scans at 257 and 281 s represent the purest points for the two components. At first the slight cusp in the derivative at 281 s is unexpected, but we have shown in a previous paper that unusual peak shapes often represent unexpected changes in purity. Owing to tailing of both peaks the slower eluting peak (compound **I**) is never completely pure. However, these points in time provide excellent first guesses of the pure spectra, which are refined by deconvolution as discussed below.

For MS, there is a clear composition 2 region centred around 270 s. This is compatible with the PC plots in Fig. 9. Note that, because of the very noisy baseline and the way in which the derivatives are scaled, these plots are much noisier than for HPLC-DAD. Hence it is less easy to perform deconvolution starting with pure mass spectra, and the alternative approach of using pure masses as a first guess of elution profiles is more appropriate.

4.4 Deconvolution

The logarithm of the derivative plot was used to determine the purest EAS for each compound as shown in Fig. 12(b). The purest spectra are shown to be at 257 and 281 s and are used to create the first estimate of the spectral matrix, ${}^{1,\mathrm{DAD}}\hat{\mathbf{S}}$, with each row of the matrix corresponding to each spectrum. This spectral matrix is used in eqn. (12) to start the deconvolution algorithm (Section 3.5.1) and obtain a first estimate of the concentration profiles. A second estimate of the spectral and then the

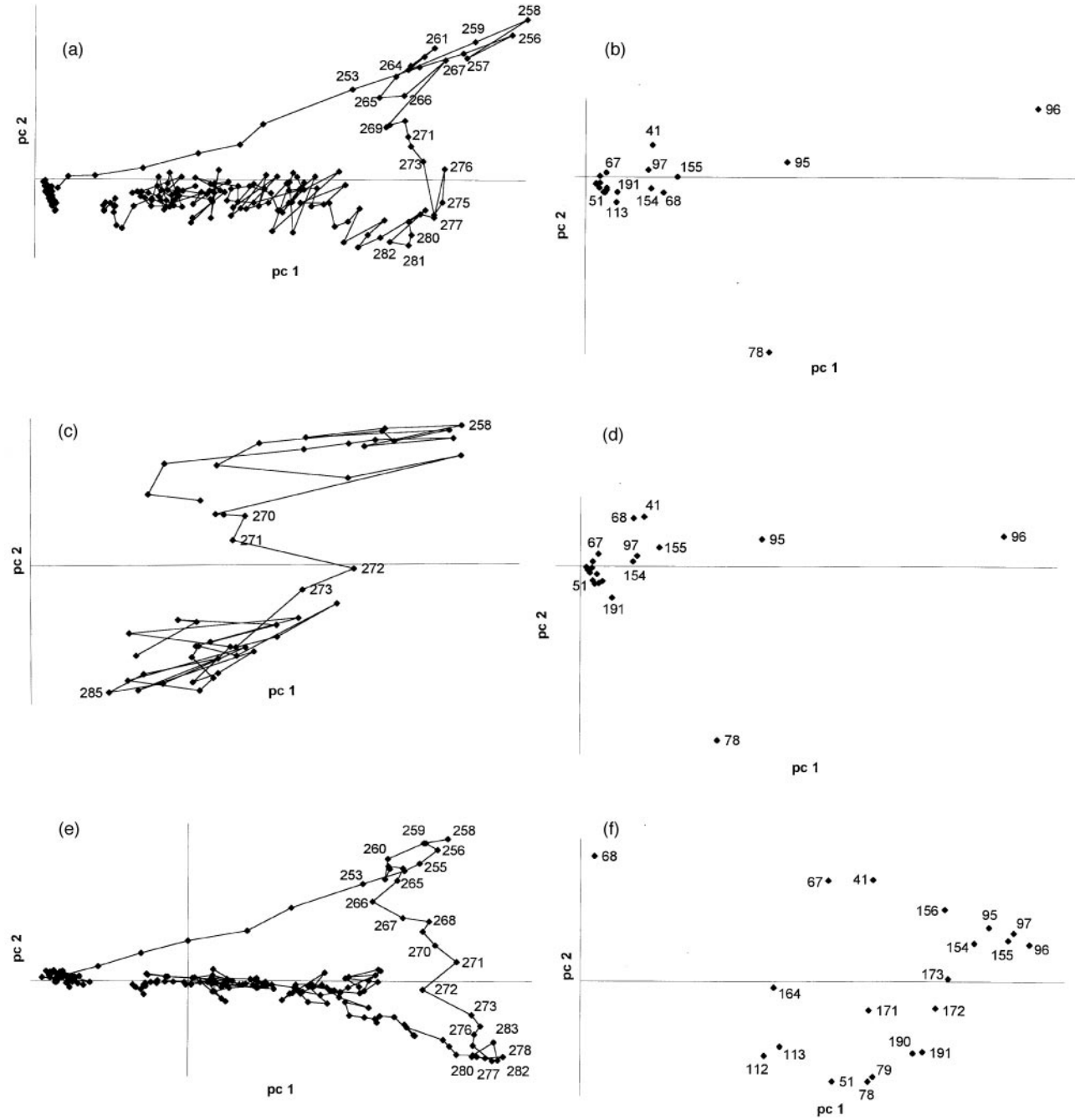

Fig. 9 Principal components analysis of the HPLC-MS data: (a) non-transformed scores plot; (b) non-transformed loadings plot; (c) normalised scores plot; (d) normalised loadings plot; (e) standardised scores plot; and (f) standardised loadings plot.

concentration profiles are then performed as shown in eqns. (13) and (14), respectively. The second estimates of the spectral, $^{2,\mathrm{DAD}}\hat{\boldsymbol{S}}$, and concentration profiles, $^{2,\mathrm{DAD}}\hat{\boldsymbol{C}}$, are shown in Fig. 13. The tailing peak shapes are obvious from the deconvoluted spectra. The EAS are virtually identical with the known EAS of the individual components at pH 4.9, which are not illustrated for brevity.

The deconvolution process for the HPLC-MS data is described in Section 3.5.2. In order to deconvolute the MS data, estimates of the concentration profiles are required, and these were determined from the non-normalised, non-standardised loadings plot [Fig. 9(b)], which illustrate the most diagnostic masses for each compound (*m/z* 41 and 78). The profiles at *m/z* 41 and 78 are used to form the concentration matrix $^{1,\mathrm{MS}}\hat{\boldsymbol{C}}$, and to obtain first estimates of the spectral profiles, $^{1,\mathrm{MS}}\hat{\boldsymbol{S}}$, as shown in eqn. (15). A second estimate of the concentration profiles, $^{2,\mathrm{MS}}\hat{\boldsymbol{C}}$, is then obtained using eqn. (16). Finally, a second estimate of the spectral profiles, $^{2,\mathrm{MS}}\hat{\boldsymbol{S}}$, is obtained for all the masses (as opposed to just the significant masses in the previous iterations). Further iterations can be made to improve the estimates, but they were found to make little difference. The results after these two iterations are shown in Fig. 14.

The elution profiles in both techniques are relatively similar apart from the tailing of the peak at long elution times and the much noisy MS profile. Graphs of the reconstructed profiles for each compound between 250 and 350 s are given in Fig. 15(a) and (b), and form roughly a linear trend. Interestingly, the elution maxima are remarkably similar for both methods. For compound **I** these are 277 s (DAD) and 281 s (MS) and for compound **II** 256 s (DAD) and 258 s (MS). The difference in elution maxima between compounds **I** and **II** is 21 s for DAD and 23 s for MS. However, because the mass spectra are inherently more noisy it is harder to pinpoint the maximum easily. This suggests that the chromatography for both techniques is extremely similar given the set-up proposed in this paper.

The reconstructed spectra are also of considerable value. Above we have referenced the MS loadings plots to Fig. 11(b) and (c). Unlike DAD, the electrospray MS detector is much noisier, so taking a single point in time will not necessarily

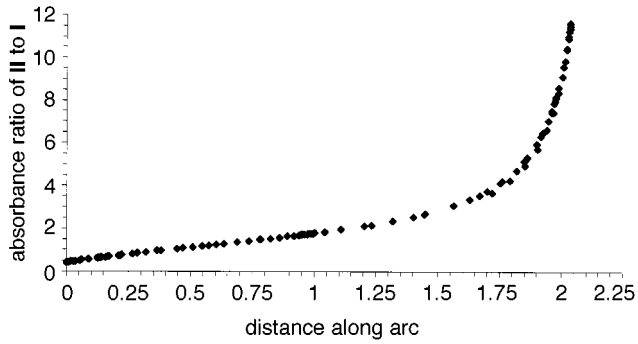

Fig. 10 Plot of absorbance ratio for compound **II** to **I** *versus* the distance along the arc (obtained from the standardised loadings plot).

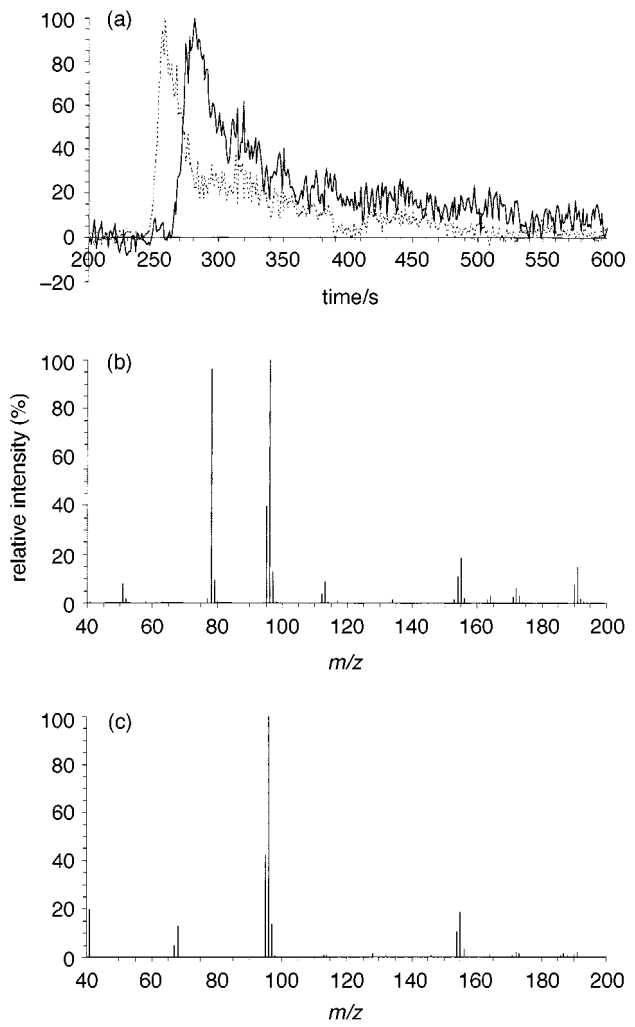

Fig. 11 Results of the HPLC-MS deconvolution; (a) the concentration profiles for compounds **I** (solid line) and **II** (dotted line) and the spectral profiles of (b) compound **I** and (c) compound **II**.

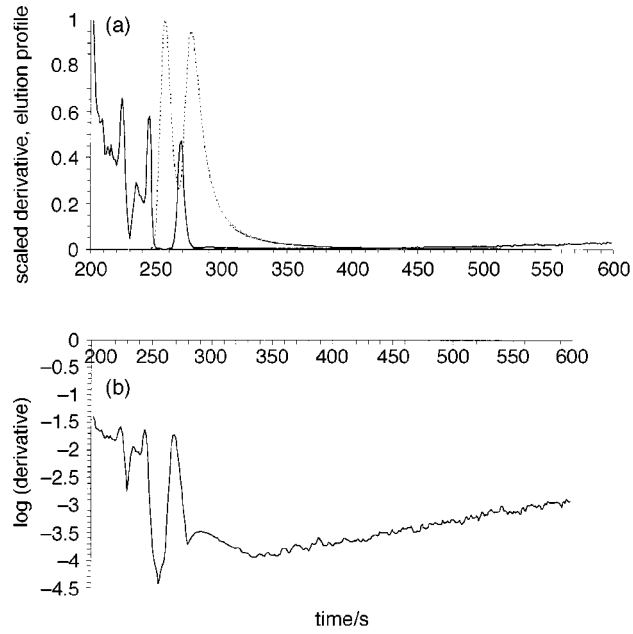

Fig. 12 Comparison of (a) the scaled HPLC-DAD derivative (solid line) and scaled elution profile (dotted line) and (b) a plot of log (HPLC-DAD derivative).

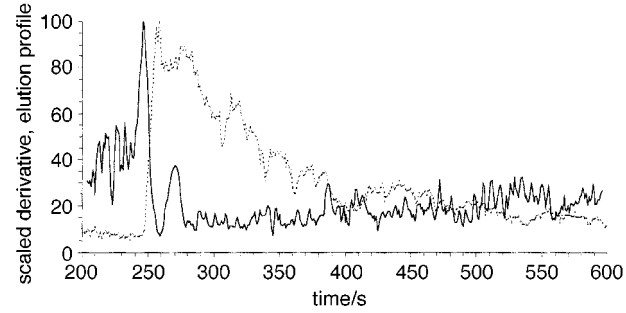

Fig. 13 Comparison of the scaled HPLC-MS derivative (solid line) and scaled elution profile (dotted line).

produce a consistent answer, with the ratio of peak intensities changing considerably. Deconvolution provides a good average spectrum. It is particularly valuable where there are common ions because it redistributes the intensities between spectra. This may have a major influence on recognition of spectra *via* library searching. A weakness of several ionisation techniques including electrospray is that, although the molecular ion is easily obtained, good fragment ion information is harder to elucidate. For noisy partially overlapping peaks, chemometric deconvolution is an important aid.

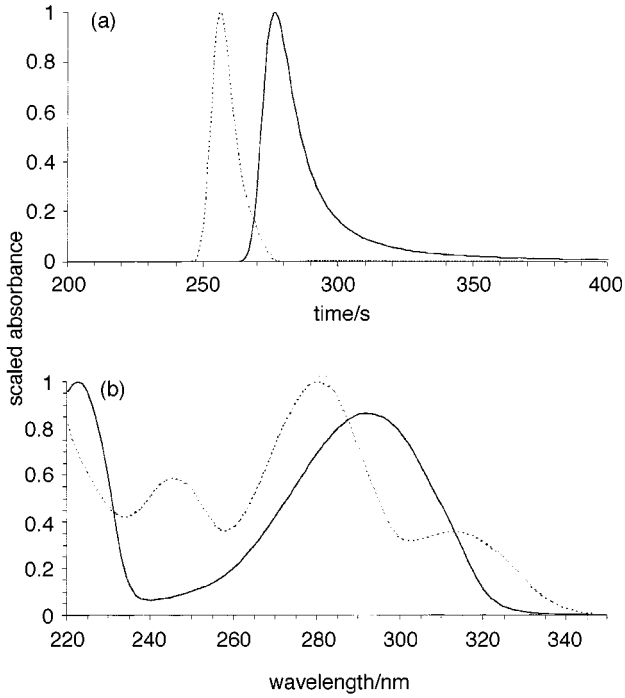

Fig. 14 Results of the HPLC-DAD deconvolution; (a) the concentration profiles and (b) the spectral profiles for compounds **I** (solid line) and **II** (dotted line).

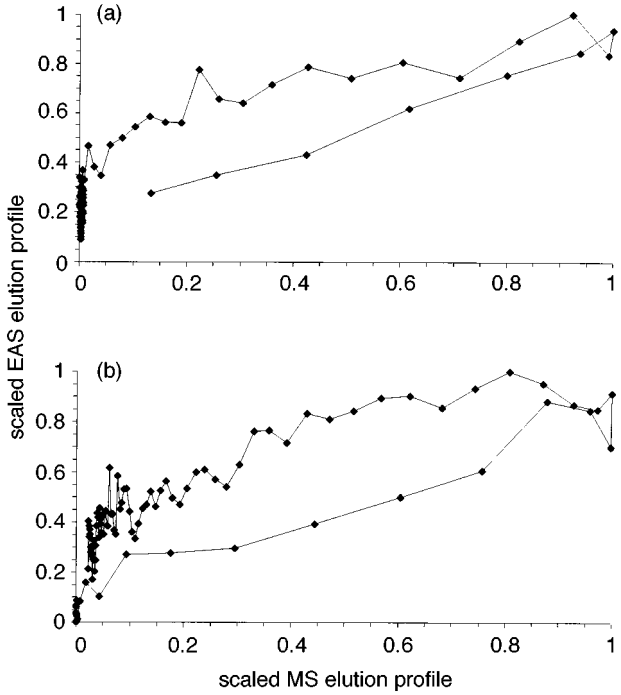

Fig. 15 Plots of the EAS deconvoluted elution profiles *versus* the MS deconvoluted elution profile for (a) compound **I** and (b) compound **II**.

4.5 Procrustes analysis

Another approach to direct comparison of DAD and MS plots is to use procrustes analysis. Procrustes analysis was described in Section 3.5, and was used to 'align' the HPLC-DAD and HPLC-MS scores, by a process of reflecting, scaling and rotating. In this particular case it was necessary to multiply the second MS score by -1 in order to reflect them on to the DAD scores. Owing to the DAD and MS data having a large difference in their order of magnitude, both sets of scores were scaled to a maximum of 1. As a result the scale factor, μ, in eqns. (18) and (19) was set to 1. It was found that a theta value of only $-3°$ was required to give the best alignment between the two sets of scores, *i.e.*, the smallest error according to eqn. (20). The results of procrustes analysis are shown in Fig. 16. This scores plot provides a useful visual comparison of the two techniques.

5 Conclusions

Despite the possibilities of coupling two or more detectors to HPLC, and the slow but steady increase in the use of HPLC-DAD-MS both as an integrated system and by coupling two separate machines, there has been limited work on data analysis. Conventionally, information from both systems is processed and stored independently. For straightforward systems this approach is adequate. If there are clearly distinguished peaks, simply selecting the chromatographic maxima on each peak and drawing out the relevant spectra are sufficient. However, in the case of closely eluting peaks, it is often desirable to process both sets of data simultaneously. Many problems have to be overcome, such as alignment of data points, and achieving comparable resolution on both systems, as reported in this paper. PC scores plots are a very flexible graphical method for exploring such datasets, and an understanding of how scaling influences the appearance and interpretation of these graphs is important. Deconvolution can be employed to obtain pure elution profiles and spectra. In the case of electrospray, where there is considerable instrumental noise, fragment ions can be identified using approaches for spectral clean-up.

This paper presents a first approach to the conjoint analysis of HPLC data from more than one detector. Most spectroscopic and chromatographic data are presented graphically, but standard software analyses each profile independently. In the case where peaks are partially overlapping it is important to be able to obtain simultaneous spectral information on individual data points, especially in the case of noisy, partially resolved, poor peak shapes. MS provides the best structural and diagnostic information whereas EAS results in the best chromatographic profiles. With the growth in technical feasibility, there must be a parallel growth in software and chemometric techniques. It is, of course, possible to extend these methods to more complex situations. However, it is good

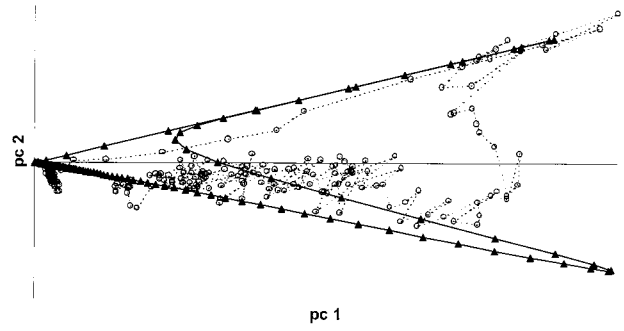

Fig. 16 Comparison of the HPLC-DAD (▲) and HPLC-MS (○) scores after procrustes analysis.

practice to report approaches when first applied to reasonably tractable problems to validate the approaches as in this paper.

We thank the EPSRC Analytical Science Programme for their financial support and P. Hindmarch and R. L. Erskine for their help in the mass spectral data decoding.

Appendix

Table of notation

I	number of scans in the selected region after interpolation
J	number of variables (wavelengths or masses)
i	individual scan number, *i.e.*, the *i*th scan
j	individual variable (wavelengths or masses) number, *i.e.*, the *j*th variable
${}^{\mathrm{DAD}}J$	number of wavelengths in HPLC-DAD
${}^{\mathrm{raw,MS}}J$	number of raw mass numbers in HPLC-MS
${}^{\mathrm{MS}}J$	number of reduced mass numbers in HPLC-MS
DAD	diode-array detector
MS	mass spectrometric detector
$\boldsymbol{X}$	data matrix of dimensions $I \times J$ (right hand side superscripts distinguish type)
$x_{i,j}$	individual intensity in a data matrix (with superscripts corresponding to the parent matrix)
${}^{\mathrm{b,D}}\boldsymbol{X}$	baseline corrected data matrix after interpolation of dimensions $I \times J$ for detector D
${}^{\mathrm{all,b,MS}}\boldsymbol{X}$	baseline corrected data matrix containing all masses of dimensions $I \times {}^{\mathrm{raw,MS}}J$
${}^{\mathrm{b,D}}x_{i,j}$	intensity at interpolated poin *i* and *j*th detector variable of ${}^{\mathrm{b,D}}\boldsymbol{X}$ for detector D
${}^{\mathrm{b,D}}\overline{x}_j$	mean, baseline corrected intensity at point *j* for detector D
s_j	significance at the *j*th mass of ${}^{\mathrm{b,MS}}\boldsymbol{X}$
${}^{\mathrm{n,D}}\boldsymbol{X}$	normalised data matrix of dimension $I \times J$ for detector D
${}^{\mathrm{s,D}}\boldsymbol{X}$	standardised data matrix of dimension $I \times J$ for detector D
K	total number of components
k	component number, *i.e.*, the *k*th component
$\boldsymbol{T}$	scores matrix of dimension $I \times K$
$\boldsymbol{P}$	loadings matrix of dimension $K \times J$
$\boldsymbol{E}$	error matrix of dimensions $I \times J$
$\hat{\boldsymbol{C}}$	concentration profile matrix of dimensions $I \times K$ for detector D
$\hat{\boldsymbol{S}}$	spectral profile matrix of dimensions $K \times J$
${}^{m,\mathrm{D}}\hat{\boldsymbol{C}}$	*m*th estimate of the concentration profile matrix using detector D from $\boldsymbol{X}$
${}^{m,\mathrm{D}}\hat{\boldsymbol{S}}$	*m*th estimate of the spectral matrix using detector D from $\boldsymbol{X}$
μ	scaling factor used in procrustes analysis
θ	rotation angle used in procrustes analysis
${}^{\mathrm{D}}\boldsymbol{t}_{i,k}$	score at point *i* for detector D for component *k*
${}^{\mathrm{p,DAD}}\boldsymbol{t}_{i,k}$	rotated DAD score at point *i* after procrustes analysis for component *k*
${}^{\mathrm{p}}\boldsymbol{E}$	error after procrustes analysis
${}^{\mathrm{D}}D_{i,j}$	derivative values for detector D
$\lvert{}^{\mathrm{D}}d_{i,j}\rvert$	absolute derivative values for detector D
${}^{\mathrm{n}}\lvert{}^{\mathrm{D}}d_{i,j}\rvert$	renormalised absolute derivative values for detector D
${}^{\mathrm{D}}\overline{d}_j$	mean derivative spectrum for detector D

6 References

1 N. Mistry, I. M. Ismail, M. S. Smith, J. K. Nicholson and J. C. Lindon, *J. Pharm. Biol. Anal.*, 1997, **16**, 697.
2 H. Iwahashi and T. Ishii, *J. Chromatogr. A*, 1997, **773**, 23.
3 J. L. Wolfender, S. Rodriguez, K. Hosettmann and W. Hiller, *Phytol. Anal.*, 1997, **8**, 97.
4 R. Andreoli, M. Careri, P. Manini, G. Mori and M. Musci, *Chromatographia*, 1997, **44**, 605.
5 I. Ogura, D. L. Duval and K. Miyajima, *J. Am. Oil Chem. Soc.*, 1995, **72**, 827.
6 A. C. Hogenboom, I. Jagt, R. J. J. Vreuls and U. A. T. Brinkman, *Analyst*, 1997, **122**, 1371.
7 N. Jitsufuchi, K. Kudo, H. Tokunaga and T. Imamura, *J. Chromatogr. B*, 1997, **690**, 153.
8 R. Draisci, L. Giannetti, L. Lucentini, L. Palleschi, G. Brambilla, L. Serpe and P. Gallo, *J. Chromatogr. A*, 1997, **777**, 201.
9 A. Lagana, C. Fago and A. Marino, *Anal. Chem.*, 1998, **70**, 121.
10 M. Careri, A. Mangia, P. Manini and N. Taboni, *Fresenius' J. Anal. Chem.*, 1996, **355**, 48.
11 F. A. Mellon, *VG Monographs*, 1991, 1.
12 M. Valcárcel and D. Luque de Castro, *Flow-Injection Analysis: Principles and Applications*, Ellis Horwood, Chichester, 1987.
13 J. Ruzicka, *Anal. Chem.*, 1983, **55**, 1040A.
14 E. Malinowski, *Factor Analysis in Chemistry*, Wiley, New York, 2nd edn., 1991.
15 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37.
16 I. T. Jolliffe, *Principal Components Analysis*, Springer, New York, 1986.
17 S. Wold and E. Lyttkens, *Bull. Int. Stat. Inst. Proc.*, 1969, **37**, 1.
18 A. Savitsky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1927.
19 M. J. Adams, *Chemometrics in Analytical Spectroscopy*, Royal Society of Chemistry, Cambridge, 1995.
20 R. G. Brereton, *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*, Ellis Horwood, Chichester, 1993.
21 P. Hindmarch, C. Demir and R. G. Brereton, *Analyst*, 1996, **121**, 993.
22 R. G. Brereton, *Analyst*, 1995, **120**, 2313.
23 O. M. Kvalheim, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 283.
24 M. Kubista, *Chemom. Intell. Lab. Syst.*, 1990, **7**, 273.
25 C. Demir, P. Hindmarch and R. G. Brereton, *Analyst*, 1996, **121**, 1443.

Paper 8/04345K