Journal of Chromatography A, 662 (1994) 113–122
Elsevier Science B.V., Amsterdam

CHROM. 25 690

Assessment of peak origin and purity in one-dimensional chromatography by experimental design and heuristic evolving latent projections

Yi-zeng Liang

Department of Chemistry and Chemical Engineering, Hunan University, Changsha (China)

Markku D. Hämäläinen*

Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, S-750 07 Uppsala (Sweden)

Olav M. Kvalheim

Department of Chemistry, University of Bergen, N-5007 Bergen (Norway)

Roger Andersson

Department of Food Science, Swedish University of Agricultural Sciences, Box 7051, S-750 07 Uppsala (Sweden)

(First received June 25th, 1993; revised manuscript received October 20th, 1993)

ABSTRACT

A new procedure for assessing peak origin and purity in chromatographic calibration is presented. Chemical analytes were mixed according to an experimental design in order to achieve independent concentration patterns. One-dimensional chromatograms were analysed as digital profiles with the heuristic evolving latent projections (HELP) method after minimization of the retention time shifts between target peaks by a simplex technique. The origin of peaks was assessed by calculating the correlation between concentration patterns, obtained as the first loadings in HELP from principal component analysis (PCA) on selective chromatographic regions, and the patterns in the designed mixtures. Co-eluting impurities and overlapping peaks could be detected, resolved and quantified. Only a few non-overlapping data points were needed to assess the origin of peaks. Latent-variable correlation chromatograms are introduced as a powerful tool for the assignment of chromatographic areas with similar concentration patterns.

INTRODUCTION

Calibration in chromatographic analysis is commonly carried out by mixing all the chemical compounds in a stock standard solution, followed by sequential dilution to give working standard solutions of different concentrations. This creates a situation where all the chromatographic peaks from the standards are correlated. Chromatograms often contain different types of "ghost" peaks, for example from derivatization reagents, column bleed and carryover from previous samples or from partial decomposition of analytes. These peaks can overlap with the true peaks from the standards. The conventional method of calibration prevents the identification of peaks by correlation analysis as all the standards have the same concentration pattern. Cali-

* Corresponding author.

SSDI 0021-9673(93)E1143-N

bration is also usually based on area quantification. This is appropriate if the peaks are pure or overlapping impurities do not vary. However, real samples may not fulfil the above criteria, and this calls for a new strategy.

Retention time shifts and different background from one chromatogram to another make it difficult to utilize several chromatographic profiles jointly. Fortunately, in recent work [1], it was shown that by simplex optimization of the cross-correlation between selected target peaks, such retention time shifts could be minimized and the chromatographic profiles analysed by means of latent-variable projection methods [2–5]. Thus, the one-dimensional chromatograms from different runs can be made comparable by using this technique first. Fig. 1a and b show the gas chromatographic profiles for the nine calibration samples before and after baseline and retention time correction, respectively. The chromatograms from different runs can now be collected in a data matrix, in which each column represents the digital chromatographic profile of one sample and each row the chromatographic concentration of the different samples at a specific retention time point. Every region in chromatographic profiles can now be subjected to a local factor analysis as developed for coupled chromatography in heuristic evolving latent projections (HELP) methods [3–5].

In this study, we investigated a model system consisting of peracetylated aldoses. By mixing the standards according to a factorial design [6], the different chemical components are forced to

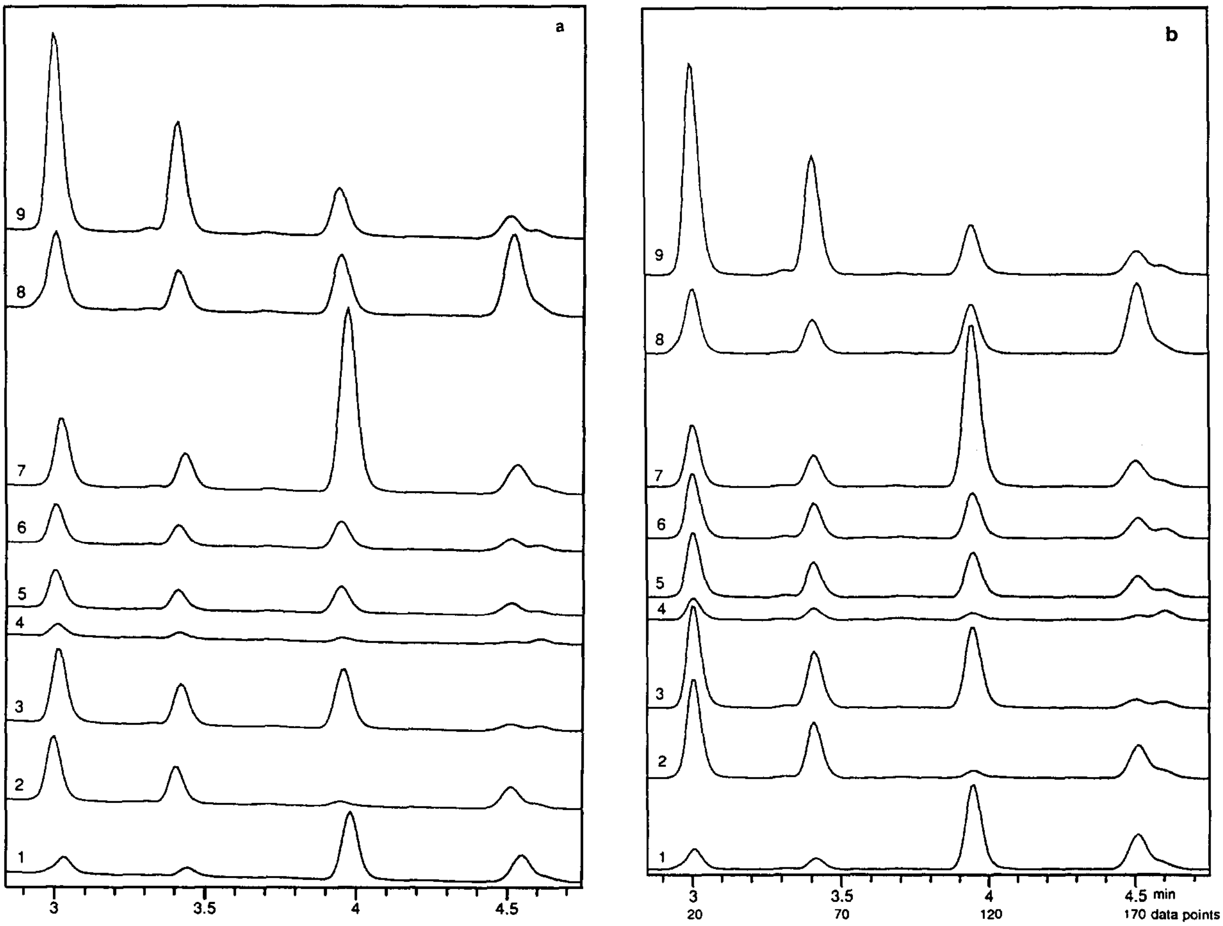

Fig. 1. Gas chromatographic profiles from peracetylated aldoses. (a) Raw data; (b) retention time, baseline and internal standard adjusted data. Time scale in min.

vary independently. The aldose standards equilibrate into anomeric isomers in acidic water, introducing several correlated peaks into the chromatograms. These correlations depict the decomposition of a compound and/or the existence of impurities in the standards, which give rise to correlated peaks in the chromatograms. There are also peaks from the derivatization reagents. Hence it should be possible to find three different kind of chromatographic areas: (i) non-correlated areas from the different experimentally designed aldoses, (ii) highly correlated areas derived from the isomerization of the aldoses and (iii) non-correlated areas, *e.g.*, from the derivatization reagents, which do not correlate with the designed concentration patterns. The first two types of areas should also correlate with the designed concentration patterns. In this paper, we present the advantages of using preprocessed chromatograms which are analysed as digital profiles by the HELP method. This provides a systematic way to find and identify the origin of chromatographic peaks in one-dimensional chromatography and to check the purity of each analyte.

THEORY

Rank analysis of a chromatographic profile

As discussed in the Introduction, a data matrix can be constructed by including different chromatographic profiles of calibration samples. For ease of comparison with the HELP resolution procedure developed for multi-detection chromatography, we let each column represent the digital chromatographic profile of one sample, and, consequently, each row the detector response of the different samples at a specific retention time point. Eigenstructure-tracking analysis (ETA) utilizing local principal component analysis (PCA) [7,8] can now be used for the analysis of the resulting matrix. This procedure performs PCA on local regions of the chromatographic profiles by moving a window of specified size from the first until the last retention time point and plotting the evolving eigenvalues in the retention time direction. The procedure starts with a window size of two and is repeated with a window of three, four, etc., until the window size exceeds by one the maximum number of co-eluting chemical components. With this window size, the last-evolving eigenvalue corresponds to the noise level over the entire elution region. The number of evolving eigenvalues above the noise level corresponds to the number of co-eluting chemical compounds in a local retention time region. For pure peaks only one eigenvalue is above the noise level. The plot of evolving eigenvalues can thus be used to assess the homogeneity of a peak [2].

In PCA the data matrix ($\boldsymbol{X}$) is decomposed into scores ($\boldsymbol{T}$) and loadings ($\boldsymbol{P}$) according to

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T \tag{1}$$

For selective chromatographic regions, the concentrations for one chemical component from sample to sample are proportional to the elements in the first loading vector ($\boldsymbol{p}$). This may appear strange, since the score vector maps the chromatographic concentration profile for selective chromatographic regions [3]. However, this follows from the design and the organization of the data described above. Further details can be found in ref. 2.

Note that the HELP method works on uncentred data so that correlation is defined around the origin, not around the mean as is commonly done in factor analysis (see ref. 7, p. 40). The information on peak origin can therefore be obtained by comparing the first loadings from selective regions detected by the HELP method with the uncentred orthogonal concentration patterns from the experimental design. The procedure is illustrated in Fig. 2.

One-component and zero-component regions of chromatographic profiles

Chromatographic profiles possess an attractive feature: chemical compounds appear and disappear in a continuous manner during elution. Thus, signals at neighbouring retention time points tend to correlate. This gives an opportunity to perform correlation analysis by applying PCA on interesting chromatographic regions. Let $a_{i,j}$ be an element in a data matrix (A) at the ith retention time point and from the jth calibration sample. The sub matrix A_{sub} including a

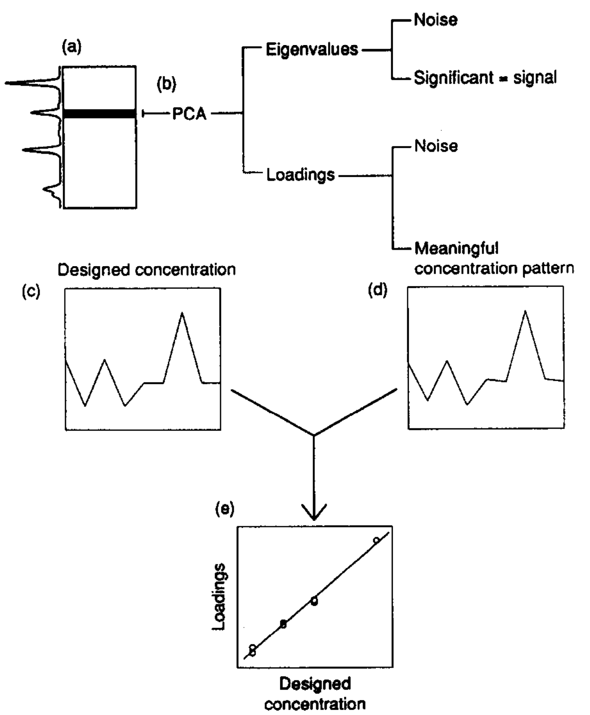

Fig. 2. Assessment of peak origin by the heuristic evolving latent projections method for one-dimensional chromatograms on several samples. (a) Pre-processed data matrix; (b) local principal component decomposition of the submatrix into loadings (eigenvectors) and eigenvalues; (c) designed concentration pattern; (d) first loading vector from a selective region; (e) assessment of peak origin by analysis of the correlation between the loadings from a selective region and the designed concentrations.

particular retention time interval (the shaded region in Fig. 2) can be expressed by

$$A_{\text{sub}} = \begin{matrix} a_{i+1,1}a_{i+1,2}a_{i+1,3}\cdots a_{i+1,N} \\ a_{i+2,1}a_{i+2,2}a_{i+2,3}\cdots a_{i+2,N} \\ \vdots \qquad\qquad \vdots \\ a_{i+k,1}a_{i+k,2}a_{i+k,3}\cdots a_{i+k,N} \end{matrix} \qquad (2)$$

Let us take a simple situation as example where there is only one chemical species, say component β, eluting from retention time point $i+1$ to $i+k$. In this case the chromatographic values for sample g, a_i+1, g, a_i+2, g, ..., a_i+k, g, should all be proportional to the designed concentration of component β in sample g. That is,

$$a_{i+j,g} \propto c_{\beta,g} \text{ or } a_{i+j,g} = r_j c_{\beta,g}$$

$$(j = 1, 2, \ldots, k, g = 1, 2, \ldots, N) \qquad (3)$$

where $c_{\beta,g}$ indicates the designed concentration of component β in sample g and r_j is a proportionality constant.

Note that eqn. 3 implies that all the rows, $\boldsymbol{a}_j = (a_{i+j,1}a_{i+j,2}a_{i+j,3}\cdots a_{i+j,N})$, in the submatrix $\boldsymbol{A}_{\text{sub}}$ can be linearly expressed by one vector, $\boldsymbol{c}_\beta = (c_{\beta,1}\, c_{\beta,2}c_{\beta,3}\cdots c_{\beta,N})$. In mathematical terms, the rank of the matrix $\boldsymbol{A}_{\text{sub}}$ is one. Such a one component region is called a selective region in the HELP method [3,4].

Because of the measurement error from the instrument, eqn. 3 should include an error term:

$$a_{i+j,g} = r_j c_{\beta,g} + e_{i+j,g} \qquad (4)$$

For this reason, ETA is used for resolving significant signals from noise, *i.e.*, for finding the "chemical rank" (number of components under a studied peak). The task is completed by comparing the eigenvalues obtained from the matrix $\boldsymbol{A}_{\text{sub}}$ with the first eigenvalue from the regions with no chemical signals above the baseline, *i.e.*, the so-called zero-component regions [2–5]. The rationale for this comparison has been published recently [9].

Loading pattern and designed concentration pattern

As discussed in the last section, a submatrix $\boldsymbol{A}_{\text{sub}}$ containing only one chemical component can be decomposed to provide a loading vector, $\boldsymbol{p}_\beta = (p_{\beta,1}p_{\beta,2}p_{\beta,3}\cdots p_{\beta,N})$, with the concentrations of the chemical components as elements. This vector can be used to find the origin of the peak. It is the experimental design with independent concentration patterns for different chemical analytes, which permits such identification. If the studied peak originates from the calibration set, the first loading vector from a selective region is proportional to one of the designed concentration vectors, *i.e.*, $\boldsymbol{p}_\beta \propto \boldsymbol{c}_\beta$. On the other hand, if the first loading vector obtained from PCA for a one-component region does not correlate with any designed concentration pattern, the peak is a so-called "ghost" peak.

Latent-variable correlation chromatograms

A noise-reduced evolving correlation pattern, latent-variable correlation chromatogram (LVCC), can be constructed in the following way:

(1) Selective regions are identified by ETA.

(2) Target concentration profiles are calculated with a local PCA, one for each selective region. The whole selective regions are used to provide maximum noise reduction. The first loadings vector in a selective region defines the target concentration profile, $\boldsymbol{p}_t$, for a chemical component.

(3) A local PCA is performed, starting from the three first retention time points, moving in step by one until the end of elution. The first evolving loading vector, $\boldsymbol{p}_e$, is stored. Step 3 provides a further noise reduction.

(4) The latent-variable correlation chromatograms are finally constructed by calculating and displaying the evolving correlation coefficients between the target loading vectors in step 2 and the evolving loading vectors in step 3, *i.e.*, $\boldsymbol{p}_t^T \boldsymbol{p}_e$.

EXPERIMENTAL

Sample description

The sample set used in this work was mixed from individual standards of xylose, arabinose and rhamnose. A four-level factorial design (Table I) was used in the mixing of the standards in order to define their concentration patterns. The first four experiments are half of a two-level factorial design for three factors followed by two centre points. The last three experiments are the high star points of a star design. This design is not orthogonal by the common definition used in experimental design if the columns are mean-centred. However, as discussed under Theory, the PCA modelling in the HELP method is based on uncentred data and therefore the concentration patterns between the aldoses are totally independent (orthogonal; $x_i^T x_j = 0$). Table II shows the amounts of the three aldoses in the mixture samples. The aldoses were peracetylated with acetic anhydride using 1-methylimidazole as catalyst [10]. All samples were analysed on a Packard Model 427 gas chromatography equipped with a flame ionization detector and a CP-SIL 88 (9 m × 0.22 mm I.D.) capillary column with helium as carrier gas at 150 cm/min (splitting ratio 1:20). Detailed conditions are given in ref. 7. Nelson 2600 chromatography software was used for collecting digital chromatographic elution profiles. The first 200 data points (retention time range 2.85–4.75 min, sampling interval 0.6 s) from the profiles were used in this paper.

TABLE I

EXPERIMENTAL DESIGN MATRIX USED IN THE MIXING OF STANDARDS

The true amounts can be found in Table II.

Standard No.	Xylose	Arabinose	Rhamnose
1	+1	+1	−1
2	−1	+1	+1
3	+1	−1	+1
4	−1	−1	−1
5	0	0	0
6	0	0	0
7	+3	0	0
8	0	+3	0
9	0	0	+3

Data pretreatment and analysis

Retention time adjustment, baseline correction and intensity normalization were first performed on a IBM PC-486 compatible computer

TABLE II

AMOUNTS OF SUGARS (IN mg) AT THE DIFFERENT DESIGN LEVELS

The coded design levels correspond to the amounts in Table I.

Sugar	Design level			
	−1	0	+1	+3
Xylose	1	4.75	8.5	16
Arabinose	1	4.75	8.5	16
Rhamnose	1	3.25	5.5	10

by means of ChromPro software (BioTriMark, Björkkulla, Funbo, Uppsala, Sweden). The chromatograms were normalized in regions with high intensities in order to correct for the effect of increased noise with increased signal, *i.e.*, heteroscedasticity [4,11]. The data were then analysed by means of the HELP method. The software used for data analysis was written in VAX FORTRAN and implemented on a VAX-station 2000 [3].

RESULTS AND DISCUSSION

Defining peak purity by local PCA

PCA can be used for resolving signals from noise. If a region of a chromatographic area contains only one pure chemical component, the chemical rank of the submatrix A_{sub} (representing the chromatographic signal intensity for the region of elution of that peak) is one and the data contain only one principal component above the noise level. Such a region is referred to as a selective region [3–5]. The eigenvalues from PCA obtained from the matrix A_{sub} are compared with the first eigenvalue from the zero-component regions [3,4]. If the second eigenvalue of the studied region is significantly smaller than the first eigenvalue obtained from the zero-component region (noise level), the studied region can be considered to have chemical rank one. Fig. 3 shows the results from an ETA obtained with window sizes three and two. This will give a rank map for every local retention time region [5]. For instance, Fig. 3 shows that the second eigenvalue is larger than the noise level around data points 18, 57, 110 and 182. Hence the number of co-eluting chemical components is two in these regions.

A comparison of eigenvalues between the zero-component regions and the selective regions is shown in Table III. The first eigenvalues from the selective regions are all significantly larger than those from the zero-component regions (noise level or detection limit), showing the presence of chemical substances. In selective regions the second eigenvalues are all smaller than the first eigenvalues from the zero-component regions, indicating the presence of only *one* compound in the regions. If the second eigenvalues from some regions are larger than the zero-component region and the third eigenvalue smaller, these regions are two-component regions.

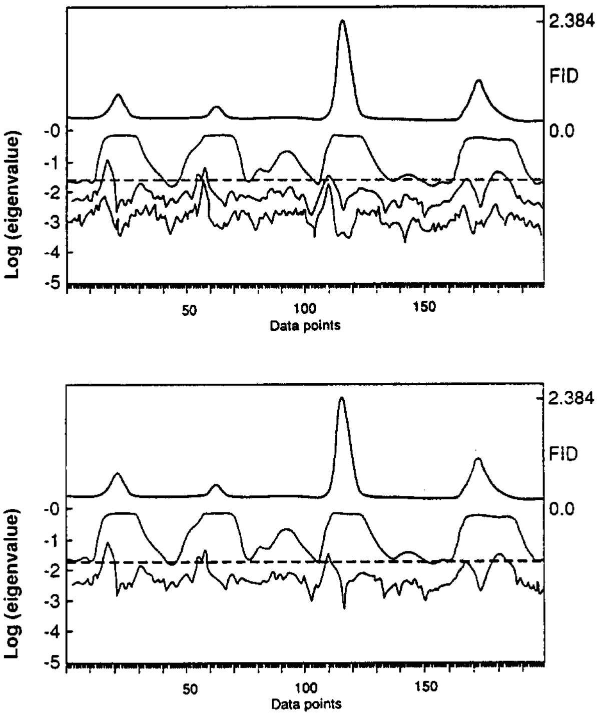

Fig. 3. Eigenstructure-tracking analysis of the data using window sizes three (top) and two (bottom). The upper trace in both parts is one of the chromatograms, followed by the first, second and (in the top part) third eigenvalues. The dashed horizontal lines indicate the noise level.

Concentration patterns, peak origin and resolution

The first loading vector in the selective regions obtained by local PCA displays the chromatographic concentration pattern (Fig. 4). These loadings should be proportional to some of the designed concentration vectors if the chemical components in these regions are derived from the standards or their decomposition products (illustrated by anomeric isomers). By comparing the loading vectors (Fig. 4) with the designed concentration profiles for the three standards (Fig. 5), the origin of the peaks can be established. The region denoted *C* in Fig. 4 is noisy as this chromatographic peak is very small, but one can still easily identify its origin (rhamnose). It is

TABLE III

EIGENVALUE COMPARISON BETWEEN THE ZERO-COMPONENT REGIONS AND SELECTIVE REGIONS

Retention time points	First eigenvalue	Second eigenvalue	Peak in Fig. 4
Zero-component regions			
4–6	0.0254	0.0082	
44–47	0.0250	0.0064	
156–159	0.0269	0.0076	
Selective regions			
12–15	0.2138	0.0151	A
23–26	0.8037	0.0058	B
48–51	0.1351	0.0136	C
62–65	0.8154	0.0101	D
117–120	0.8414	0.0108	E
171–173	0.6931	0.0053	F

interesting to look at the peaks at retention time 3 min. The first peak can, with some difficulty, be visually detected in sample 1 and sample 8 (Fig. 1b). However, the loading pattern from local PCA based on four retention time points (Table III) in the beginning of the peak cluster, denoted A in Fig. 4, correlates well with the concentration profile of arabinose (Fig. 5). This

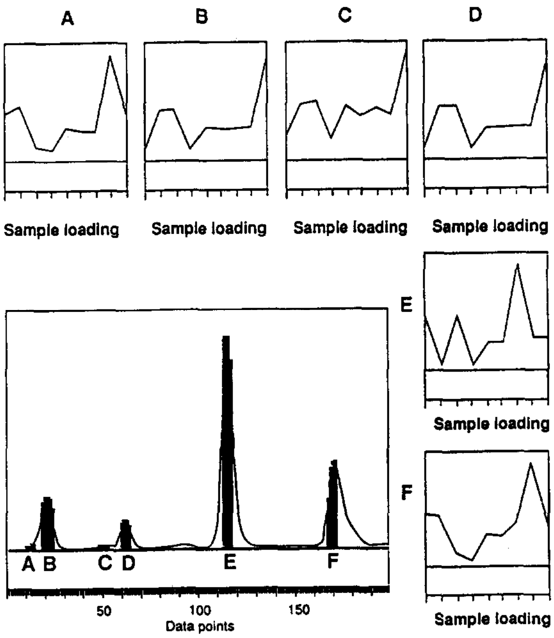

Fig. 4. Sample loading pattern (concentration) in selective areas (A–F) which correlates with the designed concentrations (Fig. 5).

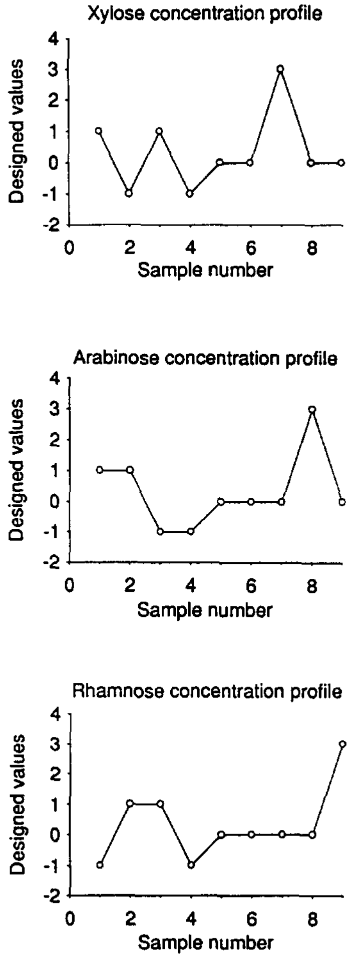

Fig. 5. Designed concentration pattern in the nine calibration mixtures.

is a clear advantage in comparison with the older integration approaches where the correlation can only be analysed between integrated peak areas (one needs at least ten data points for good integration results). Fig. 6 shows the first loading vectors for three other selective regions where the concentration patterns are far away from the designed ones. These peaks are impurities from the derivatization step.

The above-mentioned methods provide good tools for the detection of peak impurities and for the assessment of peak origin. Calibration results can be improved by eliminating the influence of impurities. This can be done by first resolving the peaks by the HELP method [3,5] and then using the resolved peak of the analytes for calibration. For example, the peak cluster around retention time 4.4–4.7 min contains two peaks (figs. 1b and 4), where the latter is an impurity (denoted C in Fig. 6). These peaks were resolved with the HELP method (Fig. 7).

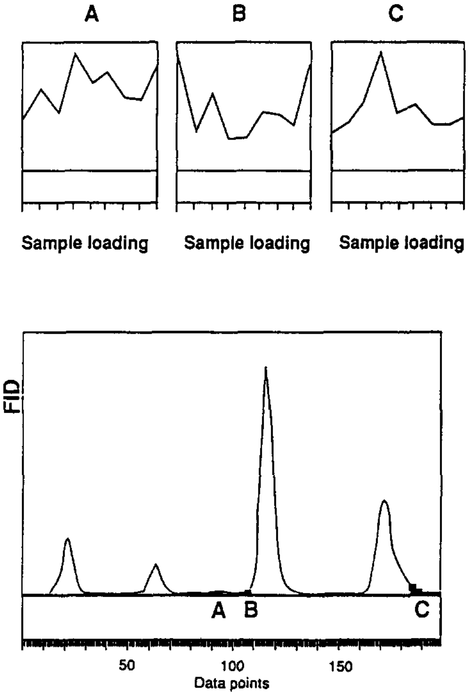

Fig. 6. Sample loading pattern (concentration) in selective areas derived from "ghost" peaks.

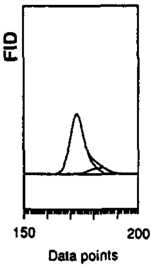

Fig. 7. Two-peak cluster at 4.5 min resolved by the HELP method.

Latent-variable correlation chromatography

In order to investigate thoroughly the correlation between one selected peak and all the other areas in the chromatogram, we introduce here the concept of latent-variable correlation chromatograms (LVCC). A chemical compound which is in equilibrium with different isomers, or a minor decomposition product, or an impurity in a standard, should have the same concentration profile with only differences in magnitude. The simplest way to assess similarities in concentration profiles is to calculate the correlation coefficient between two data points in the different chromatograms. However, a simple correlation coefficient is noise sensitive and we therefore use the loadings from principal component analysis for the calculation of the correlations (see Theory). A correlation coefficient near 1.0 indicates that the compounds have the same origin, whereas two compounds with independent concentration patterns have a correlation close to zero. It is also possible to obtain a correlation coefficient close to −1 if a single sample is run several times and a compound in the mixture decomposes giving a new, perhaps highly overlapped peak in the chromatogram. This is useful, for example, when studying the stability of compounds.

Fig. 8a shows three selected pure one-component areas (1, 2 and 3). These areas were selected by the ETA proceduce (at the minimum of the second eigenvalue at the bottom of Fig. 8a). Each of the three areas consists of three data points and are centred around retention times 3.0, 3.95 and 4.5 min, respectively (Fig.

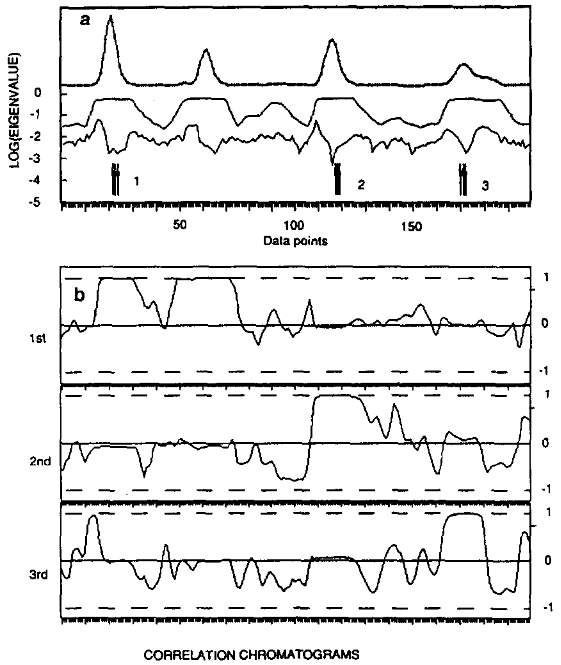

Fig. 8. (a) ETA using two retention time points as the window size. The top trace is one of the chromatograms, followed by the first and second eigenvalues. Three pure one-component areas (1–3) were selected as the target concentration patterns. (b) Three latent-variable correlation chromatograms (1st, 2nd and 3rd) calculated for each selected area.

1b), corresponding to the data point intervals 22–24, 117–119 and 170–172 in Fig. 8. Three separate PCAs are applied to these areas and the first loading vector is used as the target concentration patterns. A concentration profile chromatogram is constructed by ETA using three retention time points as the window size. For each of the three target areas an LVCC is constructed by a sequential calculation and display of the correlation between the first loading vector from the target area and all the loadings in the concentration profile chromatogram (Fig. 8b 1st–3rd). In the first LVCC, we can see that the correlation coefficients are close to one between data points 18–30 and 50–70. The first interval (18–30) includes the first target window and the second (50–70) a peak with a highly correlated concentration pattern. In the second LVCC (Fig. 8b, 2nd), only the target area has a correlation close to 1, which shows that the second target area has a unique concentration. The last LVCC (Fig. 8b, 3rd) indicates that the minor peak, with the maximum at data points 13–15, strongly correlates with the third target concentration profile. This small peak has only a few pure data points at the beginning of the peak cluster and it is notable that it was possible to identify its origin.

CONCLUSIONS

The following stepwise procedure is proposed for improved control of the calibration process in one-dimensional chromatography:

(1) In the calibration step, the standards are mixed according to an experimental factorial design, Plackett–Burman design [12], an orthogonal array or a response surface design to create uncorrelated concentration patterns.

(2) The chromatograms are made comparable by adjustment of retention time and baseline shifts.
This allows the analysis of the chromatograms as digital profiles.

(3) Heuristic evolving latent projections is applied on the profiles using the following steps:

(a) the noise level is determined from baseline regions (zero-component regions);

(b) ETA is performed in order to distinguish between pure one-component regions (peak purity check) and areas with overlapping peaks;

(c) the peak origin is established by analysis of the congruence between the sample loadings from pure one-component areas and designed concentration patterns;

(d) if a peak from a standard overlaps with other peaks which have a different concentration pattern, it can be resolved with the HELP method and the resolved areas are used in the calibration.

(4) Latent variable correlation chromatograms provide a powerful tool for finding chromatographic regions with similar concentration patterns. This method can also be used without calibration.

ACKNOWLEDGEMENTS

Y.L. is grateful for a postdoctoral fellowship from the Royal Norwegian Council for Scientific and Industrial Research (NTNF). M.D.H. thanks the Swedish Council for Technical Research for financial support.

REFERENCES

1 R. Andersson and M.D. Hämäläinen, *Chemom. Intell. Lab. Syst.*, in press.

2 M.D. Hämäläinen, Y. Liang, O.M. Kvalheim and R. Andersson, *Anal. Chim. Acta*, 271 (1993) 101–114.

3 O.M. Kvalheim and Y. Liang, *Anal. Chem.*, 64 (1992) 936–945.

4 Y. Liang, O.M. Kvalheim. H.R. Keller, D.L. Massart, P. Kiechle and F. Erni, *Anal. Chem.*, 64 (1992) 946-953.

5 Y. Liang, O.M. Kvalheim, A. Rahmani and R.G. Brereton, *J. Chemom.*, 7 (1993) 15–43.

6 S.N. Deming and S.L. Morgan, *Experimental Design: a Chemometric Approach*, Elsevier, Amsterdam, 1987, pp. 214–215.

7 E.R. Malinowski and D.G. Howery, *Factor Analysis in Chemistry*, Wiley, New York, 2nd ed., 1991.

8 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 2 (1987) 37–52.

9 Y. Liang, O.M. Kvalheim, and A. Höskuldsson, *J. Chemom.*, 7 (1993) 277–290.

10 M.D. Hämäläinen, I.-E. Ternrud, E. Nordkvist and O. Theander, *Carbohydr. Res.*, 207 (1990) 167–175.

11 H.R. Keller, D.L. Massart, Y. Liang and O.M. Kvalheim, *Anal. Chim. Acta*, 263 (1992) 29-36.

12 R.L. Plackett and J.P. Burman, *Biometrica*, 33 (1946) 305.