Journal of Chromatography, 539 (1991) 123–132
Elsevier Science Publishers B.V., Amsterdam

CHROM. 22 866

Multi-component principal component regression and partial least-squares analyses of overlapped chromatographic peaks

J. FERNANDO FAIGLE, RONEI J. POPPI, IEDA S. SCARMINIO and ROY E. BRUNS*
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13081 Campinas, SP (Brazil)
(First received May 18th, 1990; revised manuscript received September 25th, 1990)

ABSTRACT

Principal component and partial least-squares in latent variable regression methods were applied to the multivariate calibration of overlapping chromatographic peaks for toluene, isooctane and ethanol mixtures. The degree of peak overlap was varied using column temperatures of 105, 120 and 130°C. Even using the most severely overlapped peaks (130°C), the analysis errors obtained for validation set samples using both regression techniques were of the same size as those encountered using simple linear regression for individual determination of the three constituents. Truncation of the overlapped peak chromatograms appeared to lower the noise level without a significant loss of statistical information about the constituent concentrations.

INTRODUCTION

Quantitative chromatographic analysis of complex samples is often complicated by the occurrence of overlapping peaks of the mixture constituents. Expensive investment in sophisticated chromatographic equipment capable of separating overlapped peaks can be made, although more modest outlays involved in computer software and hardware can lead to accurate quantitative determintations of the constituents of mixtures. In some instances optimization techniques can be used to eliminate overlap, but longer elution times often result [1]. In others, direct numerical treatment of the overlapped band system is applied. Chemometric techniques, such as the partial least-squares in latent variables (PLS) [2,3] and the principal component regression [4] (PCR) methods, have been applied to complex multi-component analysis using a variety of chemical instrumentation [5,6], including high-performance liquid chromatography [7]. The theoretical basis and mathematical formulations of these multivariate methods have also been described [8–10].

In this work, the quantitative gas chromatographic analysis of toluene, isooctane and ethanol mixtures using the multivariate PLS and PCR methods is reported. Although diode-array detectors permit more sophisticated multivariate applications, the chromatographic system employed here has a simple thermal conductivity detector. Different degrees of peak overlap were investigated using chromatograms obtained at different column temperatures. The results of the multi-component analysis compare favourably with those obtained for the individual analyses of the three constituents. Preprocessing of the raw chromatographic data, eliminating analytical signal for the wing portions of the band system, reduces the number of principal components or latent variables needed to perform the analysis, indicating that truncation reduces the noise level without a significant loss of statistical information about the concentrations of the constituents of the mixture.

EXPERIMENTAL

Fourteen calibration samples of toluene, isooctane and ethanol mixtures with mass ranges of 0.300–1.014, 0.201–0.787 and 0.308–0.964 mg, respectively, were prepared by direct weighing so that the concentration interval between these limits was more or less evenly covered. Six validation set mixtures were prepared in the same manner as the calibration samples. Analytical-reagent grade reagents from Merck (ethanol) and Carlo Erba (toluene and isooctane) were used. Ethanol was first treated with a molecular sieve to eliminate water.

All chromatographic work was done using a Varian Model 920 chromatograph with a thermal conductivity detector and a 2 m × 1/8 in. I.D. stainless-steel column packed with 5% SE-30 on Chromossorb W (80–100 mesh). The measurements were performed at an injector temperature of 160°C and a detector temperature of 180°C with hydrogen as the carrier gas at a flow-rate of 30 ml min^{-1}. In Fig. 1 representative chromatograms obtained with column temperatures of 150, 120 and 130°C for a toluene–isooctane–ethanol mixture (0.613, 0.608 and 0.311 mg, respectively) are shown.

PLS calculations were performed on an 8-bit CPM DICON microcomputer using the SIMCA-3B program acquired from Principal Data Components. [11]. PCR

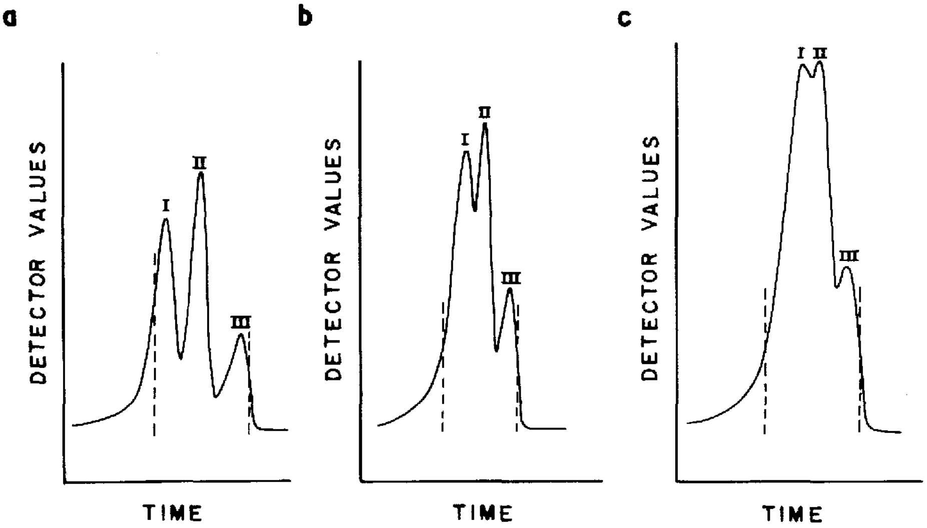

Fig. 1. Representative chromatogram obtained with column temperatures of (a) 105, (b) 120 and (c) 130°C for (I) Toluene (0.613 mg), (II) isooctane (0.608 mg) and (III) ethanol (0.311 mg).

was carried out using the KARLOV subroutine of the ARTHUR/75 computer program [12] adapted for the PC-XT microcomputer [13] and a standard multiple regression program.

PREPROCESSING

The calibration concentration matrix is 14 × 3, each row containing the masses of the three constituents of the mixtures. Each row of the 14 × 41 chromatographic matrix contains the chromatographic responses of 41 evenly spaced intervals chosen to cover completely the overlapping peak system. Each row of both the concentration and the chromatographic matrix were normalized so that the sum of their elements was equal to one. The validation set chromatographic matrix (6 × 41) was formed in the same way as that used in the calibration step. Visual inspection of the overlapped peak chromatograms showed that the wings of the chromatogram superimpose for the calibration and validation set samples. As these portions of the peak system probably contribute little information about the constituent concentrations and also introduce noise into the PCR and PLS calculations the effect of peak truncation was tested using only the 26 central detector responses falling between the limits illustrated in Fig. 1. Normalization of the rows in the (14 × 26) calibration and (6 × 26) validation matrices of the truncated data set was also carried out.

RESULTS AND DISCUSSION

Individual analysis

Separate chromatographic quantitative analyses of each of the three mixture constituents were performed to estimate the errors in the measurement procedures. The chromatographic conditions were identical with those employed for the mixture analyses, except that the column temperature was held constant at 120°C and benzyl alcohol was used as a mixture component. Eight calibration set samples with masses in approximately the same ranges as those used in the mixture analyses were used. Validation set standard prediction errors for three samples were calculated using

$$SEP = \sqrt{\sum (y_{\mathrm{calc}} - y_{\mathrm{exp}})^2/n} \tag{1}$$

where y_{calc} and y_{exp} are the percentage constituent masses calculated from the calibration graph and by direct weighing, respectively. The number of validation set samples, n, was three for the individual analyses. Standard prediction errors (*SEP*), expressed as percentages, for toluene, isooctane and ethanol of 0.95, 0.82 and 1.22%, respectively, were estimated by this procedure.

Principal component regression

The number of principal components appropriate for use in a multi-component analysis is not known *a priori*. Table I contains information relevant for the determination of these numbers for analyses performed using the three column temperatures. *F*-values calculated using standard statistical equations for regressions, fluctuate over wide ranges for all the constituents at all column temperatures as the number of principal components included in the regression is increased. However,

TABLE I

CALIBRATION SET *F*-TEST AND CORRELATION COEFFICIENT VALUES AND VALIDATION SET STANDARD ERROR VALUES FOR TOLUENE, ISOOCTANE AND ETHANOL CONCENTRATIONS FOR 41 VARIABLE EXPERIMENTS WITH COLUMN TEMPERATURES OF 105, 120 AND 130°C

Column temperature (°C)	Compound	Parameter	Number of PC[a]				
			1	2	3	4	5
105	Toluene	*F*-value[b]	186	131	217	492	352
		Corr. coef.[c]	0.969	0.980	0.992	0.998	0.998
		SEP[d] (%)	3.35	3.28	1.84	1.45	1.41
	Isooctane	*F*-value	1.47	153	201	646	663
		Corr. coef.	0.331	0.983	0.992	0.998	0.999
		SEP (%)	12.7	2.55	2.25	2.24	2.02
	Ethanol	*F*-value	9.99	1510	1270	862	823
		Corr. coef.	0.674	0.998	0.999	0.999	0.999
		SEP (%)	11.3	1.50	1.07	1.09	1.01
120	Toluene	*F*-value	107	492	202	143	195
		Corr. coef.	0.948	0.948	0.992	0.992	0.996
		SEP (%)	6.21	6.23	2.63	2.43	2.77
	Isooctane	*F*-value	2.21	20.7	171	141	156
		Corr. coef.	0.394	0.889	0.990	0.992	0.995
		SEP (%)	13.8	11.5	3.17	2.73	1.10
	Ethanol	*F*-value	6.47	156	167	177	128
		Corr. coef.	0.592	0.983	0.990	0.994	0.994
		SEP (%)	9.17	5.85	3.47	2.30	2.64
130	Toluene	*F*-value	67.1	31.7	93.1	84.3	392
		Corr. coef.	0.921	0.923	0.983	0.987	0.998
		SEP (%)	2.45	1.85	1.58	1.39	0.78
	Isooctane	*F*-value	1.58	46.5	106	73.1	183
		Corr. coef.	0.341	0.946	0.985	0.985	0.996
		SEP (%)	12.7	2.22	1.61	1.60	2.27
	Ethanol	*F*-value	7.26	342	357	580	428
		Corr. coef.	0.614	0.992	0.995	0.998	0.998
		SEP (%)	10.8	2.02	2.14	2.05	1.97

[a] Number of principal components included in regression calculations.

[b] Standard *F*-value calculated for the sums of squares due to regression and to residuals.

[c] Correlation coefficient, *r*, for experimental and calculated concentrations.

[d] Calculated using eqn. 1.

once the *F*-values remain essentially constant or decrease, correlation coefficients for the experimental and analysed values of the constituent concentrations of the calibration set samples are larger than 0.99. For the analyses at column temperatures of 105 and 120°C, three principal components included in the regression result in correlation coefficients of 0.99 or higher for all constituents. However, for the more severely overlapped groups of peaks, obtained with a column temperature of 130°C, four or five principal components are necessary to obtain correlation coefficients this high.

In principle, the standard errors for the six validation set samples could also be used to estimate the appropriate number of principal components by comparing their values for the overlapped band systems with the standard errors obtained for the individual analyses of toluene, isooctane and ethanol. The standard error values in Table I are almost always much larger than those obtained for the individual analyses. Even when five principal components are included in the regression, errors in the PCR results of about twice the size of the errors for the individual analyses can be encountered. The larger errors in the mixture analyses can be the result of several error contributions that do not exist in the individual analyses. As shown later, some validation set samples have constituent concentrations which are not in the concentration domain of the calibration set samples. This is not a problem for the individual analyses as the validation set sample concentrations were always obtain by interpolation procedures applied to the calibration graphs. Also, it is reasonable to expect that errors propagated in the mixture results may be larger than those for the individual results owing to more severe detector non-linearity problems and possibly matrix effects.

For some constituents, analysis with certain column temperatures can be carried out using less than three principal components. For example, the determination of ethanol in the extremely overlapped peak system obtained with a column temperature of 130°C provides acceptable results using only two principal components. In Fig. 2 the ethanol concentrations of both the calibration and validation set samples as a function of the scores of the first two principal components are shown. These points form a well defined plane in this space, illustrating the usefulness of a two-variable regression in the analysis of the overlapped ethanol peak.

Partial least-squares regression

The PLS technique is especially convenient for estimating the number of components or latent variables to be included in the regression and for studying the effects of the peak truncation suggested above. Total standard prediction errors obtained from results of two block PLS treatments, for the validation set samples were calculated using eqn. 1. The y_{calc} and y_{exp} quantities are the weight percentages (normalized constituent masses) determined by the analyses and those obtained from the masses of the mixture preparations. The sum is taken over all the validation set

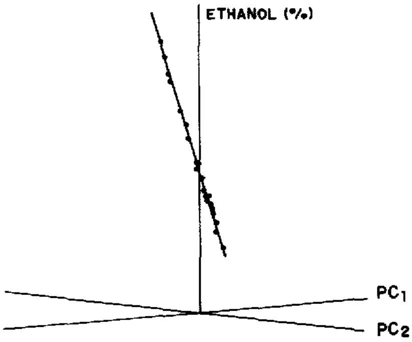

Fig. 2. Ethanol concentrations of both the calibration and validation set samples as a function of the scores of the first two principal components, obtained with a column temperature of 130°C.

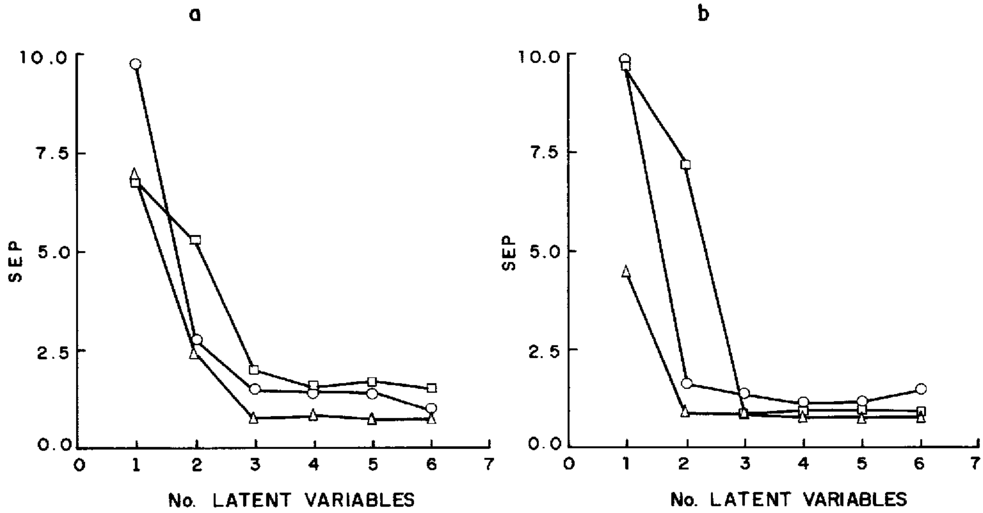

Fig. 3. *SEP* values as a function of the number of components used in the PLS analysis for (a) the complete chromatograms (41 detector responses) and (b) the truncated peak systems (26 responses). Column temperature: △ = 105; □ = 120; ○ = 130°C.

samples and all three constituents. Hence n is equal to the number of validation set samples times the number of constituents. Fig. 3 shows the graphs of the SEP values as a function of the number of components used in the PLS analysis for both the complete chromatograms (41 detector responses) and the truncated peak systems (26 responses). It can be observed that the SEP values become almost constant for three components or more for the complete 41 response data set. For the truncated 26 detector response set, values of SEP below 2.0% can be obtained using only two components for the 105 and 130°C column temperature data sets. As the results of cross-validation were not clear for determining the number of components, two components were used to describe the 105 and 130°C truncated systems and three components for the others.

Truncation of chromatograms leads to a reduction in the number of components necessary to perform the analysis for two reasons. First, the discarded data belong to regions where the chromatograms superimpose and apparently do not depend on the concentrations of the constituents. The data corrsponding to these regions of the chromatograms probably contain more noise than useful information about the constituent concentrations. Second, as the truncated regions correspond to low concentration values, the effects of non-linearity of the detector responses are also minimized.

The 120°C data set represents a special case for which the intermediate resolution of the chromatograms results in the appearance or not of a third peak depending on the constituent concentrations. A graph of the scores of the first two principal components, shown in Fig. 4, confirms the existence of sub-classes with two or three partially resolved peaks. In spite of this, a one-class model was used in all our calculations.

PLS-predicted weight percentages for the three constituents are compared with the experimental values for the six validation set samples in Table II. Results are

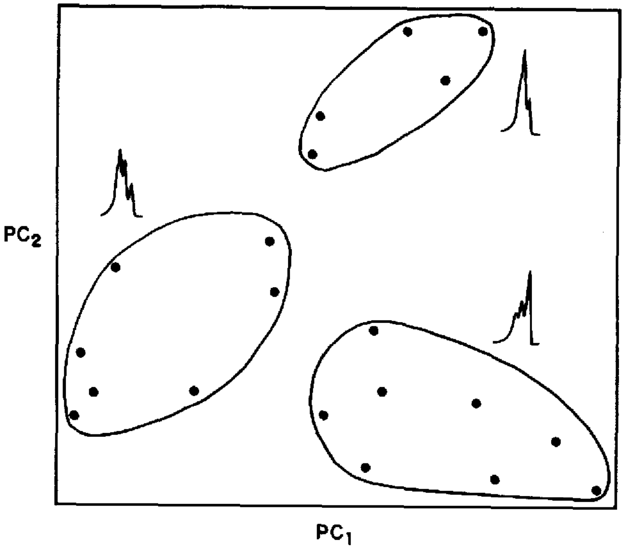

Fig. 4. Score plot of the first two principal components for the 120°C data set containing 87% of the total data variance. Clustering of samples for different types of peak profiles is clearly evident.

included for the three column temperatures studied. The residual standard deviations of the calibration set samples,

$$S_0 = \sum_{i}^{n} \sum_{k}^{p} e_{ki}^2/[(p-A)\ (n-A-1)]^{1/2} \qquad (2)$$

and the individual residual standard deviations of the validation set samples,

$$S_t = \sum_{k}^{p} e_k^2/(p-A)^{1/2} \qquad (3)$$

are also presented in Table II. In these equations the e_{ki} and e_k are the differences between chromatographic detector values and those predicted by the PLS model. The sums are taken over the independent variables (p = 26 for the truncated system and 41 for the complete peak system) and the number of samples in the calibration set (n = 14). The number of components (latent variables), A, is two for the 105 and 130°C column temperature analyses and three for those at 120°C. Only those validation set samples fitting the PLS calibration model within about $2S_0$ can be expected to have reliable PLS prediction values if a 95% confidence envelope for the calibration model is used as the classification criterion. For this reason, the sample numbers labelled *b* in Table II were not included in the calculation of the standard prediction errors of the validation set samples using eqn. 1. These error values (see Table II) are 1% or less for the three constituents analysed using column temperatures of 105 and 120°C for which overlap is less severe. For a column temperature of 130°C the errors are larger, but always less than 2%.

TABLE II

SAMPLE COMPOSITION AND PLS-PREDICTED VALUES (%, w/w) FOR TWO- AND THREE-COMPONENT PLS MODELS FOR 26-VARIABLE CHROMATOGRAPHIC ANALYSIS WITH COLUMN TEMPERATURES OF 105, 120 AND 130°C

Column temperature (°C)	S_0	Sample No.	Toluene		Isooctane		Ethanol		S_T
			Exp.	PLS	Exp.	PLS	Exp.	PLS	
105	0.254	1	40.4	39.9	25.8	25.9	33.8	34.1	0.215
		2[b]	13.9	16.4	61.5	57.2	24.7	26.4	0.752
		3	26.2	26.3	40.7	38.9	33.1	34.8	0.328
		4	26.7	26.6	19.5	18.8	53.9	54.7	0.145
		5	40.0	39.8	39.7	39.3	20.3	20.9	0.179
		6	33.8	33.0	32.2	33.1	34.0	33.9	0.246
		SEP (%)		0.4		1.0		0.9	
120	0.284	1	40.4	40.9	25.8	27.0	33.8	32.1	0.151
		2[b]	13.9	16.2	61.5	65.9	24.7	17.9	1.302
		3	26.2	27.4	40.7	40.3	33.1	32.3	0.389
		4	26.7	26.5	19.5	19.5	53.9	53.9	0.202
		5	40.0	39.3	39.7	40.5	20.3	20.2	0.409
		6	33.8	33.7	32.2	32.6	34.0	33.7	0.212
		SEP (%)		0.7		0.7		0.9	
130	0.247	1[b]	40.4	45.4	25.8	25.3	33.8	29.3	0.571
		2	13.9	14.3	61.5	61.8	24.7	23.9	0.130
		3	26.2	25.6	40.7	39.9	33.1	34.5	0.121
		4	26.7	26.4	19.5	19.3	53.9	54.3	0.128
		5	40.0	37.7	39.7	43.5	20.3	18.8	0.174
		6	33.8	32.7	32.2	33.4	34.0	33.9	0.123
		SEP (%)		1.2		1.8		1.0	

[a] S_T and S_0 values calculated using eqns. 2 and 3. *SEP* values calculated using eqn. 1. PLS results are from two block calculations. For all calculations PLS_2 and PLS_1 calculations gave identical results within experimental error.

[b] Results for these samples were not included the calculation of *SEP* as they do not fit the PLS calibration models, *i.e.*, $S_T > 2.0S_0$.

The importance of excluding samples that do not fit the calibration model should be emphasized. Their constituent concentrations were not included in the calculation of the SEP values shown in Fig. 3 or in the determination of the appropriate number of latent variables to be included in the PLS analysis. Also, the PCR validation set sample errors in Table I are grossly inflated by errors from the constituent concentrations of these same samples which are not accurately described by the PCR model. Even though this three-analyte system is not complicated and the detector might be expected to provide linear additivity, it is usually risky to extrapolate with PLS or PCR.

In Table III, standard prediction errors for the validation set samples for analyses made with the three column temperatures are presented for calculations using both the PCR and PLS methods. The number of principal components or latent variables for the calculations at each column temperature are indicated. The PLS and PCR prediction errors are very similar.

TABLE III

STANDARD PREDICTION ERRORS (%) FOR VALIDATION SET SAMPLES USING THE PLS AND PCR METHODS FOR COMPLETE AND TRUNCATED CHROMATOGRAMS

Method	Column temperature (°C)	A^a	Toluene	Isooctane	Ethanol
PLS, complete	105	3	0.4	0.9	1.0
	120	3	2.6	1.3	1.7
	130	3	0.7	1.5	1.6
PLS, truncated	105	2	0.4	1.0	0.9
	120	3	0.7	0.7	0.9
	130	2	1.2	1.8	1.0
PCR, complete	105	4	0.3	1.0	1.0
	120	4	2.1	1.8	1.9
	130	4	0.7	1.7	1.5
PCR, truncated	105	3	0.8	1.2	0.9
	120	4	0.8	1.1	0.8
	130	3	0.7	1.7	1.1
Individual analyses			1.0	0.8	1.2

[a] A = number of latent variables (PLS) or principal components (PCR) used in the regressions.

The effect of eliminating detector readings for the wings of the overlapped peak systems on the standard prediction errors can also be seen in Table III. For either the PLS or PCR regression techniques the peak truncation performed here does not appear to have decreased the accuracies of the analyses. Some standard prediction errors decreased dramatically with peak truncation (*e.g.*, PLS calculations for analysis with a column temperature of 120°C) whereas others increased slightly.

CONCLUSIONS

Both the PCR and PLS regression techniques result in similar standard prediction errors for the toluene–isooctane–ethanol mixtures studied here. The PCR method required one more regression variable (or component) than PLS. This has already been observed for multi-component analysis using fluorescence spectra [14]. Increasing overlap of the chromatographic peaks did not result in large increases in the prediction errors. This is especially interesting because the calibration and validation set samples are identical for the three column temperatures used to control the degree of peak overlap. Peak truncation effected by eliminating the detector values of the wing portions of the overlapped peak system resulted in prediction errors slightly smaller than those obtained using the entire overlapped peak system.

ACKNOWLEDGEMENTS

The authors thank Professor Antonio Pires Valente for use of his chromatographic system and his encouragement of these studies. Stimulating discussion with Professor William J. Dunn, III, were very helpful in the completion of this work.

REFERENCES

1 S. L. Morgan and S. N. Deming, *J. Chromatogr.*, 112 (1975) 267–285.

2 H. Wold, in K. G. Joreskog and H. Wold (Editors), *Systems Under Indirect Observation, Part 2,* North-Holland, Amsterdam, 1982, pp. 1–54.

3 K. R. Beebe and B. R. Kowalski, *Anal. Chem.*, 59 (1987) 1007A–1015A.

4 K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, New York, 1979, pp. 213–246.

5 W. Lindberg and B. R. Kowalski, *Anal. Chim. Acta*, 206 (1988) 125–135.

6 M. Otto and W. Wegscheider, *Anal. Chem.*, 57 (1985) 63–69.

7 W. Lindberg, J. Öhman, S. Wold and H. Martens, *Anal. Chim. Acta*, 174 (1985) 41–51.

8 S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemometr.*, 1 (1987) 41–56.

9 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 185 (1986) 1–17.

10 A. Lorber, L. E. Wargen and B. R. Kowalski, *J. Chemometr.*, 1 (1987) 19–31.

11 S. Wold, *SIMCA-3B*; available from Principal Data Components, Columbia, MO 65201, U.S.A.

12 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, *ARTHUR*; available from Infometrix, Seattle, WA 98121, U.S.A.

13 I. S. Scarminio and R. E. Bruns, *Trends Anal. Chem.*, 8 (1989) 326–327.

14 M. Sjöström, S. Wold, W. Lindberg, J. A. Persson and H. Martens, *Anal. Chim. Acta*, 150 (1983) 61–70.