Analytica Chimica Acta, 235 (1990) 41–51
Elsevier Science Publishers B V., Amsterdam – Printed in The Netherlands

Multivariate calibration of reversed-phase chromatographic systems. Some designs based on three-way data analysis

AGE K. SMILDE *, PIET HEIN VAN DER GRAAF and DURK A. DOORNBOS

Research Group Chemometrics, University Centre for Pharmacy, A. Deusinglaan 2, NL-9713 AW Groningen (The Netherlands)

TON STEERNEMAN and ANITA SLEURINK

Department of Econometrics, Faculty of Economics, University of Groningen, PO Box 800, NL-9700 AV Groningen (The Netherlands)

(Received 4th July 1989)

ABSTRACT

When retention measurements are available for a set of solutes on different stationary phases, with varying mobile phase compositions, the resulting data set can be represented by a three-way array. Models that describe the systematic variation in this training set sufficiently, can be used to calibrate a new stationary phase. Two models are tested for this purpose: three-way partial least squares and parallel factor analysis.

One of the problems with the application of reversed-phase liquid chromatographic methods is the unsatisfactory reproducibility of materials for stationary phases. The capacity factor for a specified solute, measured for a specified mobile phase composition, varies between different stationary phase materials of the same type [1,2]. Even stationary phases of the same brand differ between the batches of the same material [3]. Especially in the area of optimization of mobile phases, this phenomenon is important; an optimized separation cannot be reproduced on a new stationary phase [4].

Calibration of a new stationary phase is therefore needed. In this context, calibration is defined as the transfer of the retention value of a solute from one system to another, in particular from one stationary phase to another. Examples of a calibration procedure based on the measurements of specially chosen solutes (markers) on the new stationary phase, were given by Smilde et al. [5,6].

Latent-variable modelling has made a promising contribution to multivariate calibration [5–10]. Recently, three-way methods of data analysis, like multi-way principal components (PC) and partial least-squares (PLS) analyses [11,12], tensorial calibration [13,14], and three-mode factor analysis [15], have proved to be useful in the modelling of chemical data. A convenient survey of methods for multi-way data analysis has been given by Law et al. [16].

THEORY

Models for two-way arrays

Here, bold lower-case characters are used for column vectors (one-way arrays), bold capitals for two-way matrices (two-way arrays) and barred bold capitals for three-way matrices (three-way arrays). The capitals I, J and K are reserved to indicate the number of levels in the different modes

(directions) in the models and D indicates the number of components in the models.

It is assumed that the data set under investigation can be represented by the matrix $\mathbf{X}$ ($I \times J$). A bilinear model of the data in this matrix $\mathbf{X}$ is

$$x_{ij} = \sum_{d=1}^{D} t_{id} p_{dj} + e_{ij} \quad (i = 1, \ldots, I;\ j = 1, \ldots, J) \tag{1}$$

or

$$\mathbf{X} = t_1 p_1' + \ldots + t_D p_D' + \mathbf{E} = \mathbf{TP}' + \mathbf{E} \tag{1a}$$

where D is the number of components in the model, $\mathbf{E}$ has typical element e_{ij}, $t_d = (t_{1d}, \ldots, t_{Id})'$, $p_d = (p_{d1}, \ldots, p_{dJ})'$, $\mathbf{T} = (t_1, \ldots, t_D)$ is an ($I \times D$) matrix (usually called the score matrix), and $\mathbf{P} = (p_1, \ldots, p_D)$ is a ($J \times D$) matrix (usually called the loading matrix). For simplicity, it is assumed that x_{ij} is a realization from an unspecified distribution and t_{id}, p_{dj} are estimates of the corresponding population parameters. Depending on the assumptions made with regard to the error terms ϵ_{ij}, with residuals e_{ij}, Eqn. 1 describes a principal components or factor analysis model. In applications, $\mathbf{T}$ and $\mathbf{P}$ are usually chosen to minimize the Frobenius norm of $\mathbf{E}$ and such that $\mathbf{P}'\mathbf{P} = I$ and $\mathbf{T}'\mathbf{T}$ is diagonal. Because each term $t_d p_d'$ has rank one, Model 1a is seen to be a successive rank-one approximation of $\mathbf{X}$.

Models for three-way arrays with unfolding

It is assumed that the data can be arranged in a three-way array $\underline{\mathbf{X}}$. An example is given in Fig. 1, where the data set used here is depicted. One of the possible ways to generalize Model 1 in the case of a three-way array $\underline{\mathbf{X}}$ is given by

$$x_{ijk} = \sum_{d=1}^{D} t_{id} p_{djk} + e_{ijk} \quad (i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K) \tag{2}$$

or

$$\underline{\mathbf{X}} = \mathbf{t}_1 \otimes \mathbf{P}_1 + \ldots + \mathbf{t}_D \otimes \mathbf{P}_D = \mathbf{T} \bullet \otimes \underline{\mathbf{P}} + \underline{\mathbf{E}} \tag{2a}$$

where $\mathbf{t}_d = (t_{1d}, \ldots, t_{Id})'$, $\mathbf{T} = (t_1, \ldots, t_D)$, $\underline{\mathbf{X}}$ has typical element x_{ijk}, $\underline{\mathbf{E}}$ has typical element e_{ijk}, $\mathbf{P}_d$ has typical element p_{djk} and $\underline{\mathbf{P}}$ is a three-way matrix with dimensions ($D \times J \times K$), with $\mathbf{P}_1$ the first horizontal slice in $\underline{\mathbf{P}}$ and $\mathbf{P}_D$ the final one. For notational details, Wold et al. [11] should be consulted. The typical element of $\mathbf{t}_d \otimes \mathbf{P}_d$ is $t_{id} p_{djk}$. Because $\mathbf{t}_d \otimes \mathbf{P}_d$ is a three-way array, Model 2a is again a successive approximation of $\underline{\mathbf{X}}$. In order to make the generalization meaningful, $\mathbf{t}_d$ and $\mathbf{P}_d$ are calculated such that $\mathbf{T}'\mathbf{T}$ is diagonal, $\mathbf{P}_d'\mathbf{P}_d = I$ for every $d = 1, \ldots, D$ and the Frobenius norm of $\underline{\mathbf{E}}$ is minimized. With this generalization, the idea of bilinearity is not extended; Model 2 is not a trilinear model [16].

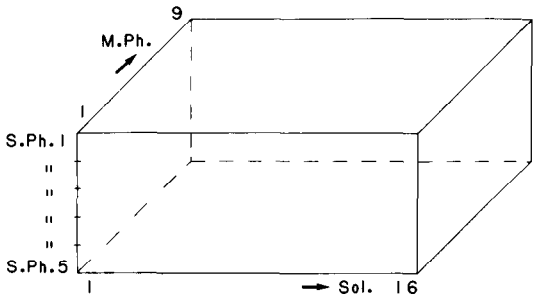

Fig 1 The arrangement of the measurements Sixteen solutes (Sol.) are measured for nine mobile phase compositions (M Ph) on five stationary phases (S.Ph.1 to S.Ph 5).

Wold et al. [11] showed that the values of t_d and P_d can be estimated by unfolding the data cube $\underline{\mathbf{X}}$. The unfolding can, of course, be done in different directions. That direction which is related to the objects should remain intact and projected onto the $\mathbf{t}_d$ vectors. In the present example, the stationary phases are regarded as objects. The unfolding is done by placing the front slice of $\underline{\mathbf{X}}$ (a 5×16 matrix) in the farthest left part of the unfolded $5 \times (16 \times 9) = 5 \times 144$ $\mathbf{X}$ matrix. The second slice of $\underline{\mathbf{X}}$ is placed to the right of the first slice in $\mathbf{X}$ and so on. For illustrative details, Wold et al. [11] should be consulted. For obvious reasons, the generalization (Model 2) is here called three-way principal component analysis with unfolding or, briefly, unfolded PCA.

A special case of the above model is appropriate when a distinction between dependent and independent variables can be made in those modes of the three-way array other than the object mode. This leads to a generalization of PLS.

It is assumed that four solutes measured at four mobile phase compositions are sufficient to pre-

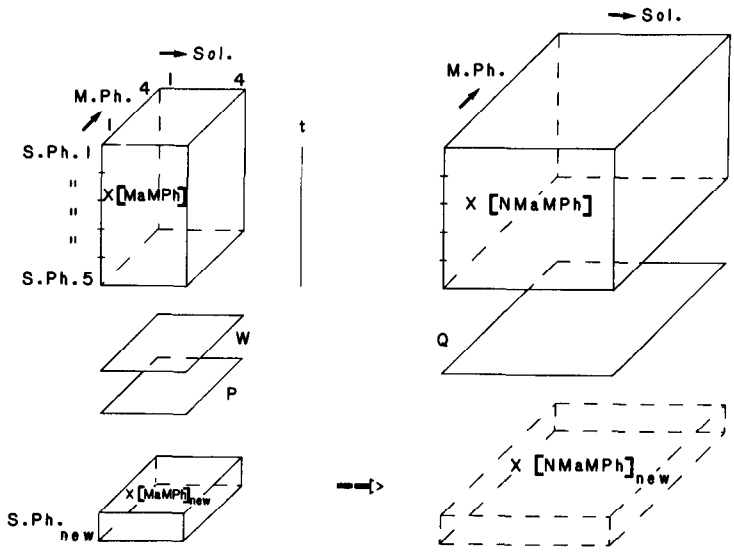

Fig. 2 The three-way PLS model visualized. The retention values of the four markers (Sol. 1 to 4), measured at four mobile phase compositions (M Ph.1 to 4) on stationary phases 1 to 5 constitute the **X**(MaMPh) block. All other retention values measured on the five stationary phase are gathered in $\bar{\mathbf{X}}$(NMaMPh) The matrices **P**, **W**, and **Q** are explained in the text, as well as t $\bar{\mathbf{X}}(\text{MaMPh})_{\text{new}}$ contains the measured retention values of the four markers for the selected four mobile-phase compositions on the new stationary phase (S Ph.$_{\text{new}}$) $\bar{\mathbf{X}}(\text{NMaMPh})_{\text{new}}$ contains the predicted values of all other solute/mobile-phase combinations on the new stationary phase

dict the retention for all other solute/mobile phase combinations. The retention values of the selected solutes [the markers (Ma)] at the selected mobile phase (MPh) compositions are gathered in $\bar{\mathbf{X}}$ (MaMPh) and, without loss of generality, the four markers are indexed with $j = 1, 2, 3, 4$ and the selected mobile phase compositions with $k = 1, 2, 3, 4$. The non-selected markers (NMa)/mobile phase combinations are gathered in $\bar{\mathbf{X}}$(NMaMPh) with the appropriate j, k index combinations. Figure 2 shows both data cubes, together with the model parameters for one PLS dimension. The following model is estimated:

$$x(\text{MaMPh})_{ijk} = \sum_{d=1}^{D} t_{id} p_{djk} + e_{ijk}$$

$$(i = 1, \ldots, I;\ j,\ k = 1, \ldots, 4) \qquad (3)$$

$$x(\text{NMaMPh})_{ijk} = \sum_{d=1}^{D} t_{id} q_{djk} + f_{ijk}$$

$$(i = 1, \ldots, I;\ j,\ k \neq 1, \ldots, 4)$$

or

$$\bar{\mathbf{X}}(\text{MaMPh}) = \mathbf{T} \cdot \otimes \bar{\mathbf{P}} + \bar{\mathbf{E}}$$
$$\bar{\mathbf{X}}(\text{NMaMPh}) = \mathbf{T} \cdot \otimes \bar{\mathbf{Q}} + \bar{\mathbf{F}} \qquad (3a)$$

where **T** is a $(I \times D)$ matrix with typical element t_{id}. The product $\otimes$ and the three-way matrices $\bar{\mathbf{P}}$, $\bar{\mathbf{Q}}$, $\bar{\mathbf{E}}$, and $\bar{\mathbf{F}}$ are defined analogously to Eqn. 2a. When both the three-way arrays, $\bar{\mathbf{X}}$(MaMPh) and $\bar{\mathbf{X}}$(NMaMPh), are unfolded in the direction which leaves the first mode intact, the results are as shown in Fig. 3. The process of unfolding was explained above. By means of ordinary PLS calculations [11], the parameters t_{id}, p_{djk} and q_{djk} of Model 3 can be estimated. For the estimation of the t_{id} parameters, use is made of weighting parameters w_{djk} for one PLS dimension gathered in **W**, which has the same dimensions as **P**. This three-way PLS is here called unfolded PLS, for obvious reasons.

When a new stationary phase becomes available, the retention values of the markers, for the previously selected mobile phase compositions,

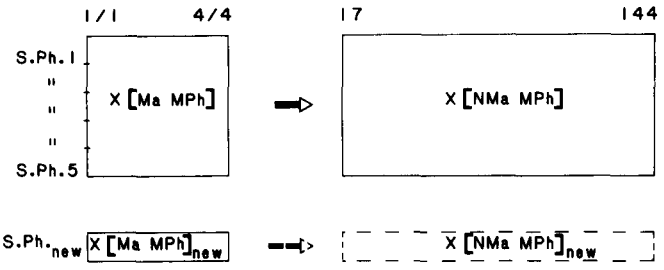

Fig. 3 The unfolding of both data cubes in Fig. 2. The numbers 1/1 indicate the retention values of the first marker for the first selected mobile phase composition, likewise 4/4 is the index for the retention value of the fourth marker with the fourth selected mobile phase composition. The numbers 17 to 144 indicate all other solute/mobile phase combinations. See Fig. 2 and the text for more explanation

must be measured on the new stationary phase to calibrate that new stationary phase. These values are gathered in $\bar{\mathbf{X}}(\text{MaMPh})_{\text{new}}$ (Fig. 2) or in $\mathbf{X}(\text{MaMPh})_{\text{new}}$ (Fig. 3). On the assumption that the previously calculated model and weighting parameters p_{djk} and w_{djk} are valid, the $t_{\text{new},d}$ values can be calculated. Predictions of the retention values of the non-selected solute/mobile phase combinations can be obtained with the use of the $t_{\text{new},d}$ and q_{djk} values. For details, Wold et al. [11] should again be consulted.

Models for three-way arrays with the PARAFAC solution

As the starting point, $\bar{\mathbf{X}}$ is assumed to be the same as above. The parallel factor analysis (PARAFAC) model has the form [16]

$$x_{ijk} = \sum_{d=1}^{D} a_{id} b_{jd} c_{kd} + e_{ijk}$$

$$(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K) \qquad (4)$$

If $\mathbf{X}_i$ is used to represent that $J \times K$ matrix which is the ith slice of the $I \times J \times K$ three-way array $\bar{\mathbf{X}}$ (in the present example, this slice contains all retention values measured on the ith stationary phase), then Model 4 can be written as

$$\mathbf{X}_i = \mathbf{B}\mathbf{A}_i\mathbf{C}' + \mathbf{E}_i \quad (i = 1, \ldots, I) \qquad (4a)$$

where $\mathbf{B}$ is a $J \times D$ matrix with typical element b_{jd}, $\mathbf{C}$ is a $K \times D$ matrix with typical element c_{kd}, and $\mathbf{E}_i$ is a $J \times K$ matrix with typical element e_{ijk}. The $D \times D$ matrix $\mathbf{A}_i$ is diagonal, with diagonal elements taken from the ith row of $\mathbf{A}$, the $I \times D$ score matrix of the first mode with typical element a_{id}. The D diagonal elements of $\mathbf{A}_i$ thus represent the effect of changes in the relative importance of the D factors on influencing retention on stationary phase i. For a given number of components in the PARAFAC model, the coefficients a_{id}, b_{jd}, and c_{kd} are calculated such that $\Sigma\Sigma\Sigma e_{ijk}^2$ is minimal, where the summations run over i, j, and k. Note that Model 4 is a trilinear model; the idea of bilinearity is extended in Model 4, in contrast to Model 2.

The PARAFAC model with only one factor is depicted in Fig. 4. Predictions of retention values on a new stationary phase can be obtained as follows: suppose measurements are available for the retention of the four markers with the previously selected four mobile phase compositions on the new stationary phase. As in case of the unfolded PLS solution, the new stationary phase has to be calibrated with these measurements. These measurements are gathered in $\bar{\mathbf{X}}(\text{MaMPh})_{\text{new}}$. Then $a_{\text{new},1}$ to $a_{\text{new},D}$ can be estimated by least squares from

$$x(\text{MaMPh})_{\text{new},jk} = \sum_{d=1}^{D} b_{jd} c_{kd} a_{\text{new},d} + e_{\text{new},jk}$$

$$(j,\ k = 1, \ldots, 4) \qquad (5)$$

where the model parameters b_{jd} and c_{kd} are known from the training set (Model 4) and are assumed to be fixed. A thorough statistical treatment of this approach should incorporate investigations into the statistical properties of the estimates b_{jd}, c_{kd}, $a_{\text{new},d}$, etc. This will be the subject of future research. With the use of the $a_{\text{new},d}$ values, it is

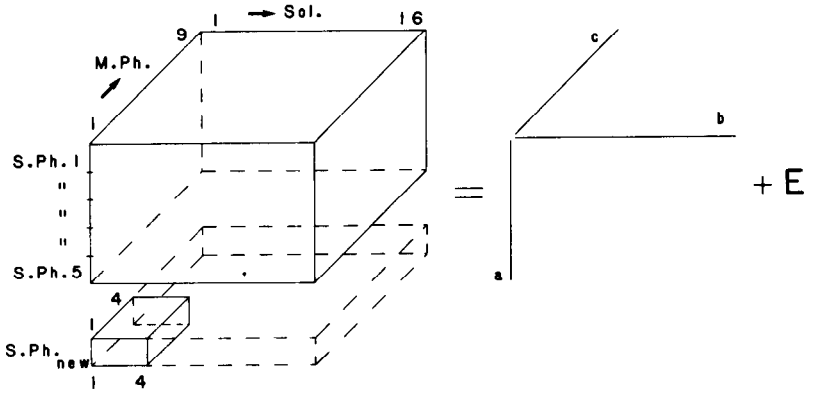

Fig. 4. The PARAFAC model. The block of data labeled as S.Ph $_{new}$ contains the measured retention values of the four markers for the selected mobile-phase compositions on the new stationary phase; a, b, and c are explained in the text.

easy to obtain predictions of the non-selected solute/mobile phase combinations:

$$x(\text{NMaMPh})_{\text{new},jk} = \sum_{d=1}^{D} a_{\text{new},d} b_{jd} c_{kd}$$

$(j,\ k \neq 1,\ldots,4)$

Conceptual differences between the PARAFAC and the unfolded PLS models

There is a conceptual difference between Models 2 and 4. Model 2 is an unconstrained model; the values p_{djk} are the factor loadings of the dth factor (component) across modes B and C of the data and no constraint is placed on these values. Such a constraint is present in Model 4: $p_{djk} = b_{jd}c_{kd}$. The meaning of this constraint can be explained as follows. By comparing $b_{1d}c_{1d}$ and $b_{1d}c_{2d}$ with $b_{2d}c_{1d}$ and $b_{2d}c_{2d}$, it is clear that the expression of the influence of factor d across mode C (as measured by c_{1d} and c_{2d}) does not depend on the level of mode B. The reverse is also true: the expression of the influence of factor d across mode B does not depend on the level of mode C.

In the present example, the second mode consists of the solutes and the third mode consists of the mobile phase compositions. Studies on the PCA of reversed-phase chromatographic data [17,18] indicate that the loadings of the solutes on the first principal component are related to the hydrophobic character of the solutes whereas the loadings of the mobile phase compositions can be related to the polarity of the eluent. With respect to the first factor, the above-mentioned constraint means that the way in which hydrophobic differences between solutes affect the differences in retention values of those solutes, does not depend on the mobile phase composition.

EXPERIMENTAL

Experimental details are already available [6]. Three stationary phases (octadecyl-modified reversed-phase material) were studied by Smilde et al. [6]. These stationary phases are called A, B, and C. On each stationary phase, the retention values of sixteen solutes were measured for nine mobile phase compositions. The data set was augmented with three stationary phases of the same brand as those used by Smilde et al. [6], but from different batches. On the three new stationary phases (labelled D, E, and F), the same test solutes were measured for the same mobile phase compositions as previously [6]. All retention values on the stationary phases, D, E, and F were measured by the same analyst/apparatus combination as for stationary phase C and are reported in Table 1.

The PLS calculations were made with the SIMCA-3B program (Sepanova, Enskede, Sweden) run on a Myami Compact AT-286 (an IBM-compatible personal computer). The PARAFAC calculations were done on the Cyber 170/760

TABLE 1

Capacity factors of the test solutes

Mobile phase [a]	Solute [b]															
	ACP	ACT	ANS	CRE	DMP	EAB	EE	EHB	MHB	NBZ	PBL	PE	PHB	PRE	PRS	TOL
Stationary phase D																
wm1	4 32	1.63	9.13	4.36	6 27	4.71	105 02	8 43	3 47	5.35	3.64	3.55	22.19	14 02	20 17	24.11
wm2	2.23	0.93	5.08	2.33	2.61	2.06	26.66	3.71	1.70	3.04	1.58	1.92	8.74	4.28	6.47	12.88
wm3	1.22	0 55	2.84	1.27	1.22	0.98	7.98	1 67	0.84	1.72	0.79	1.08	3.53	1.59	2.40	6.75
am1	5.88	2.09	12.75	6.09	10.21	7.20	145.98	11.65	4 62	8.39	5.02	4.12	31 88	17.87	22.34	32.54
am2	3 14	1.16	7.02	3.21	4.38	3.29	33.46	4.79	2 14	4.86	2.23	2 18	11.54	5.47	6.43	16.55
am3	1 77	0.69	3 87	1.77	2.18	1.66	10.33	2.29	1.15	2.71	1.13	1.24	4.89	2.01	2.43	8.63
wa1	6 42	1.84	15.83	5.95	10.18	7.82	84.07	9.10	3.75	11.76	3.76	3.33	24.28	7.48	7.26	39.98
wa2	3.39	1.02	8.17	3.02	4 45	3.42	17.05	3.58	1.75	6.16	1 50	1.68	7.97	1.81	1.65	19 36
wa3	1.93	0.64	4.33	1.66	2.29	1.78	5.13	1 73	0.97	3 32	0.79	0.96	3.30	0.73	0.66	9.34
Stationary phase E																
wm1	4.32	1.59	9.23	4.29	6.02	4 63	96.27	8.39	3 45	5.37	3.50	3.53	22.06	13.82	19.36	23.95
wm2	2.28	0.97	5.20	2.44	2.64	2.17	26.83	3.80	1.75	3.10	1.64	1.98	8.91	4.40	6.92	13 10
wm3	1.25	0.56	2.90	1.29	1.25	1.03	8.24	1.69	0.86	1.77	0.81	1.11	3.55	1.65	2.42	6.87
am1	6.19	2.21	13.25	6.41	10.44	8.19	146.95	11.76	4.71	8.84	5.59	4.40	31.87	19.50	23.85	33 08
am2	3.26	1.21	7 23	3 32	4 58	3.46	34.87	4.94	2.22	4.98	2.48	2.31	11.78	5.54	6.72	17.25
am3	1.77	0.71	3.86	1.75	2.15	1.66	10.07	2.26	1.13	2 72	1.16	1.27	4.80	2.05	2.48	8.58
wa1	6.64	1.92	16.28	6.14	10.74	8 16	88.61	9.53	3.96	11.92	4.40	3.49	25.34	7.71	7.43	41.86
wa2	3.52	1 05	8.48	3.04	4.56	3.50	17.15	3.69	1.83	6.42	1.74	1.74	8.10	1.90	1.69	19.75
wa3	1.98	0.64	4.40	1.65	2.32	1.80	5 14	1.76	0.99	3.37	0.88	1.00	3.32	0.77	0.67	9.42
Stationary phase F																
wm1	4.39	1.58	9 53	4.33	5.80	4.70	90.43	8.39	3.44	5 32	3.37	3.48	22.18	13.42	19.52	24.48
wm2	2.23	0.92	5.23	2.39	2.58	2.10	26.57	3.71	1.69	3.07	1.62	1.95	8.73	4.27	6.58	13.62
wm3	1.27	0.58	2.99	1.36	1.25	1 03	8.20	1.76	0 89	1.81	0.83	1.14	3.71	1.68	2.54	7.13
am1	6.09	2.10	13.51	6.23	9.89	7.45	132 19	11.56	4.61	8.75	5.34	4.29	31.45	18.38	21.69	33.84
am2	3.26	1.18	7.32	3.33	4.47	3.56	32.73	4.93	2.22	5.00	2.42	2 32	11.82	5.45	6.47	17.34
am3	1.82	0.71	3 97	1.79	2.20	1.68	10.11	2.32	1.16	2 78	1 19	1.30	4.92	2.06	2.46	8.90
wa1	6 79	1.92	16.91	6.38	10.82	8.63	88.74	9.98	4.10	12.27	4.44	3.55	26.71	7.80	7.54	43.22
wa2	3 53	1.03	8.64	3.12	4.62	3 66	17.19	3.77	1.84	6.47	1.75	1.71	8.32	1.86	1.68	20.38
wa3	2.03	0.63	4.64	1.71	2.39	1.88	5.50	1.85	1.03	3.53	0.91	1.00	3.57	0.77	0.71	9.98

[a] The mobile phase compositions, given as volume fractions (w = water, m = methanol, a = acetonitrile): wm1 = 0.63/0.00/0.37; wm2 = 0.55/0.00/0.45; wm3 = 0.47/0.00/0.53, am1 = 0.71/0.11/0.18; am2 = 0.62/0 15/0 23, am3 = 0.54/0.19/0.27; wa1 = 0.78/0.22/0.00, wa2 = 0.70/0.30/0.00; wa3 = 0.62/0 38/0.00. [b] Abbreviations for the test solutes: ACP, acetophenone; ACT, acetanilide; ANS, anisole, CRE, *p*-cresol; DMP, dimethyl phthalate; EAB, ethyl aminobenzoate; EE, ethynylestradiol; EHB, ethyl hydroxybenzoate; MHB, methyl hydroxybenzoate, NBZ, nitrobenzene; PBL, phenobarbital; PE, 2-phenylethanol; PHB, propyl hydroxybenzoate; PRE, prednisone; PRS, prednisolone; TOL, toluene

computer of CDC at the Groningen University Computing Centre, with standard programs and home-made software written in PASCAL. The markers were selected with home-made software written in FORTRAN on the Cyber and the Myami Compact AT-286.

RESULTS AND DISCUSSION

Preliminary calculations on the whole data cube

For the reasons previously mentioned [6], the logarithms of the capacity factors were used as the retention values. The data cube was arranged such that the first direction (mode A) consisted of the stationary phases (index $i = 1,\ldots,6$), the second direction (mode B) of the solutes (index $j = 1,\ldots,16$), and the third direction (mode C) of the mobile phase compositions (index $k = 1,\ldots,9$). In order to obtain some insight into the data at hand, the data cube was modelled according to Eqns. 2 and 4. Before application of the model, the data were centred such that for each j, k the sum $\Sigma x_{ijk} = 0$, where the summation covered $i = 1,\ldots,6$. Because differences between stationary phases were the primary interest, this centring operation seems reasonable.

After the centring operation, the mean sum of squares (MS) for each stationary phase i, before application of the model, was calculated as $\Sigma x_{ijk}^2/(16 \times 9)$, where the summation ran over all combinations of j and k. These MS_{bef} values can be regarded as the mean sum of squares to be explained by the model. After application of the model, analogous values were calculated for the residuals (MS_{res} values). When a data cube is modelled, it is important to know the appropriate number of components. Because the experiment was designed such that estimates of the variance of the measurement error were available [6], these estimates served as a yardstick to choose the number of components by comparing the MS_{res} values with these measurement error variances.

Table 2 summarizes the results of Models 2 and 4 applied to the data cube. The measurement error differs considerably between the stationary phases and it is therefore hard to establish the number of components, but one or perhaps two components seems to be a reasonable choice. The signal-to-noise ratio (MS_{bef}/s^2_{repro}) is not very high for stationary phases D, E, and F. A calibration training set with low s^2_{repro} values would be very advantageous.

The main differences between the unfolded solution (Model 2) and the PARAFAC solution (Model 4) can be summarized as follows. The first component in the unfolded solution explains almost completely the variation in stationary phase B, whereas the second one does the same for stationary phase A. It may be noted that sta-

TABLE 2

Results for unfolded PCA and PARAFAC on the whole data cube (MS_{bef} is the mean sum of squares before application of the models. $M_{res}(k)$ is the mean sum of squared residuals after application of k components in the model. The cumulative percentages of the sums of squares explained by the first two components of the unfolded PCA model are 65 0% and 86.1%. For the PARAFAC model these numbers are, respectively, 56.5% and 72.0% The values of s^2_{repro}, MS_{bef} and MS_{res} must be multiplied by 10^{-4})

Stationary phase	s^2_{repro}	MS_{bef}	Overall results of the models				Scores of stat phases on components			
			Unfolded PCA		PARAFAC		Unfolded PCA		PARAFAC	
			$MS_{res}(1)$	$MS_{res}(2)$	$MS_{res}(1)$	$MS_{res}(2)$	1	2	1	2
A	2.25	30.56	26 69	0.28	22.11	8 25	0.24	−0.62	−0.52	1.56
B	31 36	82.66	0.50	0.41	14.76	10.78	−1 09	−0.04	−3 34	3 51
C	2.89	22.60	10.86	8.45	13 07	10 11	0 41	0 19	1 70	−2 07
D	11 56	8 90	8 30	3 26	7.09	3 96	−0.09	0 27	0 25	−0.74
E	6.25	7 35	4.24	3.62	5.36	4 87	0.21	0.10	0.71	−0 83
F	38.44	15.14	7.93	7.23	10 39	8 80	0 32	0 10	1.19	−1 43
Av	15.46	27 87	9 76	3 87	12.13	7 79				

tionary phase B scores high on the first unfolded component, whereas stationary phase A scores high on the second one. The PARAFAC solution explains the variation in the stationary phases more regularly, although the overall performance is a bit poorer. The PARAFAC model is more restricted than Model 2, and this has several consequences. First, the number of parameters to estimate in Model 2, assuming two components, is 444 ($16 \times 9 + 16 \times 9 + 16 \times 9 + 2 \times 6$), whereas this number is 206 ($16 \times 9 + 2 \times 16 + 2 \times 9 + 2 \times 6$) in the case of Model 4. So, when the PARAFAC model is applied a considerable gain in degrees of freedom is obtained. Secondly, in the case of the unfolded solution, stationary phase B is permitted to have a considerable influence on the first component. Whether or not this is justified depends on the information contained in the variation of the retention values on stationary phase B. When this information is relevant, the strong dependence of the first unfolded component on this stationary phase is justified. When stationary phase B is an outlier, this outlying behaviour is modelled. The PARAFAC solution is less pronounced in this respect.

The solutes EE and PBL remain largely unexplained by the PARAFAC model after application of two components. The percentages of the explained variation for EE and PBL by the PARAFAC model are, respectively, 28% and 61%. These percentages are 77% and 83%, respectively, with the unfolded PCA model. This also illustrates a difference between Models 2 and 4: the more flexible Model 2 allows the solutes EE and PBL to describe much of the variation in the components, whereas PARAFAC is more restricted in this sense. This was confirmed by the high loadings of EE and PBL of the unfolded PCA components (results not shown), whereas the loadings of these solutes on the PARAFAC components were not extremely high.

Selection of the markers and the important mobile phase compositions

The selection of solute/mobile phase combinations which can be used to calibrate a new stationary phase is based on the work of McCabe [19] as adapted by Smilde et al. [6] for use in reversed-phase calibration. The purpose of this paper is not to give an extensive evaluation of methods for solute/mobile phase selection, thus only brief comments will be made on the choice in the calibration considered. The calibration procedures will be validated by successively leaving out one stationary phase, building the models with the five remaining stationary phases, and using the omitted stationary phase as an independent test sample. Hence the problem comes down to choosing solute/mobile phase combinations from five stationary phases.

The first step in the selection procedure was the unfolding of the $5 \times 16 \times 9$ data cube such that the direction of the solutes was left intact. The result was a 45×16 matrix, where the objects are stationary/mobile phase combinations and the variables are the 16 solutes. Those four solutes were selected that gave the highest sum of all the multiple correlation coefficients between the four solutes and the 12 unselected solutes. These selected solutes explained the unselected ones best. The outcome of this procedure was the following set of solutes: anisole, dimethyl phthalate, ethynylestradiol, and prednisolone. This result was found six times; for each omitted stationary phase, the procedure was repeated and gave the same results. This places some confidence on the selection procedure.

The second step in the selection procedure was the unfolding of the $5 \times 16 \times 9$ data cube such that the direction of the mobile phases was left intact. This resulted in a 80×9 data matrix, with the mobile phase compositions as variables. The same procedure as above was applied and resulted in the choice of the wm1, wm3, wa1, and wa3 mixtures. These mixtures were found six times; for each omitted stationary phase, the procedure was repeated. The conclusion is that a new stationary phase can be calibrated by measuring the retention values of the markers, the solutes ANS, DMP, EE, and PRS, at the mobile phase compositions wm1, wm3, wa1, and wa3.

Calibration of a new stationary phase with the unfolded PLS model

The calibration procedure was followed with each stationary phase left out once. This sta-

TABLE 3

Results of the unfolded PLS and PARAFAC calibrations

(The percentage X(MaMPh) and X(NMaMPh) values are the percentages of the variation in the respective matrices explained by the PLS model The R^2 values are the fraction of the variation explained by the PARAFAC components in the data cube The rmsep value is $[(1/m)\Sigma(y_i - \hat{y}_i)^2]^{1/2}$, where y_i is a measured ln k value and $\hat{y}_i$ is the value predicted by the model The summation runs from 1 to m, where m is the number of predictions made. The percentage TEST is the percentage of variation in the test set explained by the model)

Stationary phase	s_{repro}	PLS calibration				PARAFAC calibration		
		X(MaMPh) (%)	X(NMaMPh) (%)	rmsep	TEST (%)	R^2	rmsep	TEST (%)
A	0 015	92 98	89 70	0.065	0.23	0 77	0 069	– [a]
B	0 056	84 43	68 49	0 106	4.64	0 60	0 089	32 65
C	0 017	92 35	91 35	0 046	37.23	0 77	0 048	32 60
D	0 034	88 80	85 74	0 026	46.13	0 73	0 027	41 84
E	0.025	87 36	86 67	0.026	34.17	0 74	0 031	10 41
F	0.062	90 52	89.85	0 039	21 87	0 77	0 039	19 69
Av	0.039	89 41	85 30	0 058	24 05	0.73	0 055	22 87

[a] This value is not given because it was negative, i e , meaningless

tionary phase was then used as independent test set. Hence the calibration was done six times. Column-centring was always done in the way described in the Preliminary Calculations section. The results are presented in Table 3. The number of components in the model was always chosen to be two. This decision was based on cross-validation and on the comparison of the s^2_{repro} values with the residual variance after application of the successive dimensions in the model.

In judging the prediction results, several issues should be kept in mind. First, the root-mean-squared error of prediction (rmsep) values of the respective stationary phases can be compared with the s_{repro} values for the stationary phases, because they are both expressed in the same units. Secondly, the nature of the percentages of variation explained in the test set is as follows: from the test-set values, the corresponding means from the training set are subtracted; the sum of squares of these corrected values gives the sum of squares that has to be explained. Likewise, the unexplained sum of squares can be obtained by squaring (and summing) all prediction errors (observed values minus values predicted by the model). The percentage of the variation explained is then easily calculated. When the sum of squared values that have to be explained is compared with the sum of squares arising from the measurement error (which can be calculated from the s_{repro} values), signal-to-noise ratios can be calculated in the same way as in the Preliminary Calculations section. These signal-to-noise ratios were 7.2, 2.5, 5.2, 0.8, 1.1, and 0.5 for stationary phases A, B, C, D, E and F, respectively. The interpretation for stationary phases, D, E and F is that the sum of squares that has to be explained cannot be distinguished from noise. This illustrates a serious problem: the differences between stationary phases, although present, are small.

In discussing the results of the unfolded PLS calibration, it must be noted that Table 3 shows stationary phases A and B to be exceptions. Application of the unfolded PLS model is useless, because the rmsep values are high in comparison to the corresponding s_{repro} values. In the case of stationary phase B, there is some evidence in the training set that the model does not fit very well; the percentages of variation explained in the X(MaMPh) and X(NMaMPh) blocks are low. Stationary phase C has also a high rmsep compared to s_{repro}, but application of the model is not useless, although there is some lack of fit. In the case of stationary phases D and E, the model predicts with approximately the same error as the measurement error, although the low signal-to-noise ratios for these test sets (see above) cast doubt on the relevance of these predictions. For stationary phase

F, there is not only the low signal-to-noise ratio, but the rmsep value is smaller than the corresponding s_{repro} value, which indicates overfitting of the model. Probably the number of components in the unfolded PLS model is too high; the model complexity is not well established.

Calibration of a new stationary phase with the PARAFAC model

The procedure was the same as with the unfolded PLS model. Each stationary phase was left out once and then calibrated with the use of the model. Again column-centring was done and two components were chosen in the PARAFAC model. The results are presented in Table 3. Again stationary phase B is seen to be an exception, with a low R^2 value in the training set. The predictions for stationary phase B are better than in the case of the unfolded PLS model, but there is still lack of fit, for both stationary phases B and C. Stationary phase A is very badly predicted; actually the model does not work at all. Discussion of the results for stationary phases D, E and F is essentially the same as the above discussion for the unfolded PLS model.

Comparison of the results of both calibration schemes and general conclusions

No clear preference for either of the two calibration models is apparent, although PARAFAC performs better in calibrating stationary phase B than unfolded PLS. A closer look at the prediction errors of both models shows that the solutes EE and PBL are almost always badly predicted with PARAFAC. This pattern does not emerge from the unfolded PLS prediction errors where, for example, the solutes ACP and ACT are badly predicted on stationary phase B. This is in agreement with the remarks made in the Preliminary Calculations section on the differences between PARAFAC and unfolded PCA with respect to the flexibility of the models.

In discussing the calibration results, it should be kept in mind that stationary phases A and B were tested with a different analyst/apparatus combination than that used for stationary phases C–F. Hence there are not only stationary phase differences in the data set but also differences arising from other measurement conditions. This may explain the bad predictions on stationary phases A and B; both unfolded PLS and PARAFAC are unsuitable for dealing with these differences. For example, if there is a systematic difference between the measurements on stationary phase B and the other stationary phases which results in a constant absolute difference between the measurements on stationary phase B and the corresponding average values of the other five stationary phases, then it is doubtful that unfolded PLS and PARAFAC, as applied here, would be able to handle the situation. Incorporation of variables in the Models 2 and 4 that describe and handle the above-mentioned systematic differences may be profitable. Another solution to the problem of systematic differences might perhaps be found by applying a different form of centring and/or scaling of the data cube, but the subject of centring and/or scaling is difficult [16].

The problem of low signal-to-noise ratios in the test set can be tackled as follows. The first step in the calibration of a new stationary phase is the measurement of four markers for four mobile phase compositions. These 16 values can be compared with the averages of the corresponding retention values for the stationary phases in the training set. The signal-to-noise ratio of the new stationary phase can be approximated with the use of these new retention values, by dividing the sum of squared differences between the measured values and the corresponding averages in the training set by the sum of squares arising from measurement error. When this is calculated for the six stationary phases, the outcome is 10.2, 3.6, 3.8, 1.1, 1.1, and 1.0 for stationary phases A, B, C, D, E and F, respectively. With the use of these values, inferences can be made with regard to the signal-to-noise ratio of all values in the test set. Stated otherwise, the *a priori* sense in applying a model can be judged from the signal-to-noise ratio for the 16 measured retention values on the new stationary phase.

The (statistical) merits of both kinds of models, unfolded PLS and PARAFAC, have still to be established, keeping in mind their conceptual dif-

ferences. Other three-way models should also be evaluated. Diagnostic tools to judge what kind of model is appropriate in a particular application are needed. Research on these topics is underway.

REFERENCES

1 J G Atwood and J Goldstein, J Chromatogr. Sci., 18 (1980) 650
2 L. Hansson and L Trojer, J. Chromatogr., 207 (1981) 1
3 R M Smith, T G. Hurdley, R. Gill and A C Moffat, Chromatographia, 19 (1984) 407
4 A.K Smilde, C.H P Bruins, D.A Doornbos and J Vink, J Chromatogr , 410 (1987) 1.
5 A K Smilde, C.H.P. Bruins, P M.J. Coenegracht and D A Doornbos, Anal. Chim. Acta, 212 (1988) 95.
6 A.K. Smilde, P M.J Coenegracht, C H P. Bruins and D A Doornbos, J. Chromatogr., 485 (1989) 169
7 H A. Martens, Ph.D. Thesis, Technical University of Norway, Trondheim, 1985
8 S. Wold, M. Sjöström, R Carlson, T Lundstedt, S. Hellberg, B Skagerberg, C Wikstrom and J. Öhman, Anal Chim Acta, 191 (1986) 17.
9 D.M. Haaland and E V. Thomas, Anal Chem , 60 (1988) 1193
10 P. Geladi, J Chemomet., 2 (1988) 231
11 S. Wold, P Geladi, K Esbensen and J Ohman, J Chemometr., 1 (1987) 41
12 K. Esbensen, S Wold and P Geladi, J Chemomet , 3 (1988) 33
13 E Sanchez and B R Kowalski, J Chemomet , 2 (1988) 247
14 E Sanchez and B.R. Kowalski, J Chemomet , 2 (1988) 265
15 C L. de Ligny, M C Spanjer, J C. van Houwelingen and H M. Weesie, J Chromatogr , 301 (1984) 311
16 H G Law, C.W Snyder, J A Hattie and R P McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984
17 H Wijnne, Ph D Thesis, University of Amsterdam, Amsterdam 1983
18 M F Delaney, A N Papas and M J Walters, J. Chromatogr , 410 (1987) 31
19 G P. McCabe, Technometrics, 26 (1984) 137