Journal of Chromatography, 301 (1984) 311-324
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 16,908

THREE-MODE FACTOR ANALYSIS OF DATA ON RETENTION IN NORMAL-PHASE HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY

C. L. DE LIGNY* and M. C. SPANJER

Laboratory for Analytical Chemistry, University of Utrecht, Croesestraat 77A, 3522 AD Utrecht (The Netherlands)

and

J. C. VAN HOUWELINGEN and H. M. WEESIE

Institute for Mathematical Statistics, University of Utrecht, Budapestlaan 6, 3584 CD Utrecht (The Netherlands)

(First received December 23rd, 1983; revised manuscript received May 15th, 1984)

SUMMARY

It is shown that the Snyder equation is not quite satisfactory for fitting retention data in normal-phase high-performance liquid chromatography (HPLC) on chemically bonded phases. This equation is a special case of the mathematical-statistical three-mode factor analysis model. This model, in its general form, has been used to fit two sets of literature data on the retention in normal-phase HPLC for 19 solutes on six adsorbents with two eluents, and for 39 solutes on three adsorbents with two eluents, respectively. This study represents the first application of three-mode factor analysis with missing data, and also the first application of three-mode factor analysis in the field of the natural sciences. The accuracy of the fit of the observations and of the prediction of the missing data, for various numbers of extracted factors, is discussed.

INTRODUCTION

For the correlation of data that can be classified in two modes, factor analysis[1] is often a good mathematical model. In factor analysis, the data y are correlated by the equation

$$y_{a,s} = \sum_{j=1}^{n} c_j \, A_{a,j} \, S_{s,j} \tag{1}$$

where a and s are the two modes in which the data can be classified, the parameters (or "factors") A depend only on a, the factors S depend only on s, and c_j is a scaling constant. The objective is to describe the data with a small number, n, of factors.

An example of data that can be classified in two modes is gas chromatographic

0021-9673/84/$03.00 © 1984 Elsevier Science Publishers B.V.

(GC) data on the retention of a number of solutes s on a number of absorbents a. Several authors, *e.g.* refs. 2–6, have applied eqn. 1 to data of this kind. In the case of missing data the straightforward solution of eqn. 1 is precluded, but we have developed an iterative procedure[7] for the estimation of $A_{a,j}$ and $S_{s,j}$. Once this has been performed, eqn. 1 can be used to estimate the missing data.

Of course, many data sets must be classified in three or even more modes. An example is a data set on the liquid chromatographic retention of a number of solutes s on a number of adsorbents a, obtained with a number of eluents e. The extension of eqn. 1 to three modes is[8]

$$y_{a,e,s} = \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{l=1}^{r} c_{j,k,l}\, A_{a,j}\, E_{e,k}\, S_{s,l} \tag{2}$$

where c denotes the three-mode core matrix of scaling constants. For the cases where eqn. 2 is applied to j, k and l are usually larger than one. As far as we know, eqn. 2 has only been applied in the field of the social sciences[9], and only for cases where data exist for each combination of a, e and s. Two of the present authors have recently devised a method to estimate $A_{a,j}$, $E_{e,k}$ and $S_{s,l}$ for the case of missing data[10], and we will apply this method here to a case from the field of the natural sciences, *viz.*, data on the retention in normal phase high-performance liquid chromatography (HPLC).

It is quite reasonable to investigate the merits of eqn. 2 for the correlation of data of this kind. For normal phase liquid chromatography on bare silica and alumina a good physical model exists, *viz.*, that of Snyder[11], and his equation is a special case of eqn. 2. For monosubstituted benzene solutes it can be written as

$$y_{a,e,s} = A_{a,1}\, E_{e,1}\, S_{s,1}\, c_{1,1,1} + A_{a,2}\, E_{e,1}\, S_{s,2}\, c_{2,1,2} + A_{a,2}\, E_{e,2}\, S_{s,3}\, c_{2,2,3} + A_{a,3}\, E_{e,2}\, S_{s,4}\, c_{3,2,4} \tag{3}$$

TABLE I

SYMBOLS IN THE SNYDER EQUATION THAT ARE EQUIVALENT WITH y AND THE PARAMETERS A, E AND S IN EQN. 3

V_N = Chromatographic net retention volume (cm^3); W = weight (g) of the adsorbent; V_a = measure of the specific surface area of the adsorbent; α = measure of the strength of the adsorbent; γ, ζ = measures of the heterogeneity of the adsorbent; $\varepsilon°$ = measure of the strength of the eluent; S_0 = measure of the Lewis acid or base strength of the solute; Σa_i (calc.) = surface area of the solute molecule; $\Sigma\Delta a_i$ (SiO$_2$) empirical increment of the surface area of the solute molecule for silica adsorbents; n = number of aromatic carbon atoms in the solute molecule.

j,k,l	*1*	*2*	*3*	*4*	*5*
$y = \log V_N/W$					
A_j	$\log V_a$	α	$\alpha\gamma$	$\alpha\zeta$	
E_k	1	$\varepsilon°$			
S_l	1	S_0	Σa_i(calc.)	$\Sigma\Delta a_i$(SiO$_2$)	$n-6$

TABLE II

INVESTIGATED ADSORBENTS

Code	*Adsorbent*	*Reference*
1	Octadecyl-silica	12
2	N-Cyanoethyl-N-methylamino-silica	13
3	Aminobutyl-silica	14
4	2,4-Dinitroanilino-silica	15
5	Bis(3-nitrophenyl)sulphone-silica	15
6	2,4,7-Trinitrofluorenimine-silica	15

whereas for polycyclic aromatic hydrocarbon solutes an additional term

$$A_{a,4}\ E_{e,1}\ S_{s,5}\ c_{4,1,5} \tag{4}$$

is required. The symbols in the Snyder equation that are equivalent to y and the parameters A, E and S in eqns. 3 and 4 are given in Table I.

However, while the Snyder equation holds good for bare adsorbents, it is less suitable for the correlation of data obtained on chemically bonded phases. Hammers and co-workers[12–15] have demonstrated this repeatedly. So, there has grown a need for a model that can correlate observations and predict missing data more accurately. It will be shown that three-mode factor analysis fulfills these requirements.

DATA

The data that we shall analyse are taken from recent investigations in our laboratory[12–15]. The first set contains data on simple solutes, *viz.*, monosubstituted benzenes and polycyclic aromatic hydrocarbons. The adsorbents are described in Table II and the solutes in Table III. The eluents were *n*-hexane (1) and 35% (v/v)

TABLE III

INVESTIGATED SOLUTES IN TABLE IV

Code	*Monosubstituted benzenes*	*Code*	*Polycyclic aromatic hydrocarbons*
1	Anisole	7	Naphthalene
2	Thioanisole	8	Acenaphthene
3	Nitrobenzene	9	Fluorene
4	Benzonitrile	10	Bibenzyl
5	Acetophenone	11	Anthracene
6	Methyl benzoate	12	Phenanthrene
		13	Pyrene
		14	Fluoranthene
		15	Chrysene
		16	3,4-Benzopyrene
		17	Perylene
		18	Triphenylene
		19	Coronene

methylene chloride in *n*-hexane (2), and the temperature was 25°C. The data are given in Table IV.

A rough estimate of the precision of the fit of these data by the Snyder equation can be obtained as follows. For adsorbents 4–6, the eluent *n*-hexane and monosubstituted benzene solutes the standard deviation of the fit was found to be 0.14, 0.12 and 0.10, respectively, when three strongly deviating data were excluded from the regression analyses by the Snyder equation[15].

For polycyclic aromatic hydrocarbon solutes the standard deviation of the fit was found to be 0.08, 0.08 and 0.12, respectively, when two strongly deviating data were excluded from the regression analyses[15]. For adsorbents 1–3 the fits of the regressions are better, but here too several data that deviate rather strongly (0.2–0.3) from the values predicted by the Snyder equation were noted[12–14]. From these considerations we estimate that the overall precision of the fit of the data in Table IV by the Snyder equation is not better than 0.15.

The second set contains data on more complicated solutes, *viz.*, monosubstituted phenols, anilines and pyridines. Retention data for these solutes have been measured on adsorbents 1–3 with eluents 2 and 3 (methylene chloride). They are given in Table V. The standard deviations of the fit of these data by the Snyder equation range from 0.10 to 0.28[12–14].

TABLE IV

DATA ON LOG V_N/W IN NORMAL-PHASE HPLC, FOR THE ADSORBENTS AND THE SOLUTES LISTED IN TABLES II AND III, RESPECTIVELY

Solute	*Adsorbent*											
	1		*2*		*3*		*4*		*5*		*6*	
	Eluent											
	1	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
1	0.69	−0.28	0.54	−0.33	0.45		0.19		0.26		0.35	
2	0.35	−0.76	0.24	−0.55	0.32		0.14		0.17		0.27	
3	1.05	−0.26	0.85	−0.17	0.82	−0.33	0.88	−0.07	0.92	−0.04	0.86	−0.20
4	1.45	0.16	1.34	0.14	1.12	−0.17	1.28	0.15	1.35	0.21	1.15	−0.01
5		0.72	1.79	0.55	1.38	0.07	1.49	0.35	1.64	0.46	1.46	0.27
6	1.66	0.33	1.25	0.19	0.97	−0.20	1.05	−0.03	1.18	0.10	1.00	−0.08
7	0.09		−0.06	−0.60	0.21		0.15		0.13		0.23	
8	0.18	−0.95	0.00		0.31		0.29		0.28		0.34	
9	0.37		0.18	−0.50	0.52	−0.73	0.45		0.46		0.56	
10	0.47		0.23	−0.66	0.38		−0.01		0.00		0.23	
11	0.37	−0.91	0.22	−0.58	0.66	−0.68	0.78	−0.19	0.79	−0.16	0.85	−0.14
12	0.39	−0.91	0.23	−0.48	0.73	−0.59	0.79	−0.16	0.80	−0.16	0.86	−0.12
13	0.45	−0.92	0.29	−0.50			1.15	0.21	1.14	0.19	1.20	0.20
14	0.51	−0.91	0.32		0.94	−0.43	1.13	0.13	1.14	0.11	1.21	0.14
15	0.70	−0.76	0.51	−0.47	1.17	−0.29	1.46	0.30	1.47	0.28	1.56	0.42
16	0.78	−0.68	0.59	−0.45	1.31	−0.24	1.85	0.68	1.85	0.63	1.98	0.86
17	0.82	−0.61	0.64	−0.35	1.41	−0.16	1.92	0.78	1.92	0.73	2.02	0.93
18	0.70	−0.67			1.18	−0.31						
19	0.96	−0.46	0.77	−0.31	1.63	0.01						

TABLE V

DATA ON LOG V_N/W IN NORMAL-PHASE HPLC, FOR MONOSUBSTITUTED PHENOLS, ANILINES AND PYRIDINES

Solute			*Adsorbent*					
Code	*Series*	*Substituent*	*1*		*2*		*3*	
			Eluent					
			2	*3*	*2*	*3*	*2*	*3*
20	Phenols	*m*-F		0.79	1.00	0.53	1.95	1.18
21		*p*-F		0.42	1.01	0.67	1.81	1.03
22		*m*-Cl	0.78	0.48	0.98	0.71	2.01	1.22
23		*p*-Cl	0.83	0.46	1.00	0.68	1.93	1.14
24		*m*-Br	0.78	0.49	1.00	0.76	2.00	1.17
25		*p*-Br	0.84	0.48	1.02	0.71	1.98	1.09
26		*m*-CH_3	0.75	0.37	0.96	0.55	1.57	0.73
27		*p*-CH_3	0.78	0.41	0.98	0.57	1.57	0.74
28		*m*-OCH_3	1.22	0.72	1.34	0.87	1.92	0.94
29		*p*-OCH_3	1.32	0.43	1.41	0.84	1.84	0.87
30		*m*-NO_2	1.45	1.08	1.64	1.27		1.57
31		*p*-NO_2	1.63	1.38	1.88	1.36		2.09
32		*m*-CN		1.12	1.94	1.23		1.51
33		*p*-CN	1.77	1.30	2.04	1.46		1.82
34		*m*-$COOCH_3$		1.22	1.84	1.23		1.26
35		*p*-$COOCH_3$		1.37	1.94	1.37		1.54
36		*m*-$COCH_3$		1.61	2.16	1.48		1.52
37		*p*-$COCH_3$		1.77	2.37	1.78		1.82
38	Anilines	*m*-F	0.80	0.09	0.72	0.12	0.63	−0.25
39		*p*-F	1.22	0.55	1.03	0.47	0.85	0.00
40		*m*-Cl	0.78	0.06	0.68	0.10	0.63	−0.24
41		*p*-Cl	1.01	0.27	0.85	0.24	0.75	−0.12
42		*m*-Br	0.78	0.05	0.67	0.08	0.66	−0.24
43		*p*-Br	0.97	0.21	0.83	0.21	0.75	−0.14
44		*m*-CH_3	1.25	0.61	0.94	0.46	0.73	−0.09
45		*p*-CH_3	1.29	0.80	1.13	0.45	0.83	0.03
46		*m*-OCH_3	1.66	0.85	1.33	0.64	1.01	0.03
47		*p*-OCH_3		1.28	1.67	0.99	1.24	0.29
48		*m*-NO_2	1.15	0.09	1.01	0.19	1.00	−0.13
49		*p*-NO_2	1.20	−0.02	1.19	0.22	1.26	0.08
50		*m*-CN	1.56	0.49	1.34	0.47	1.14	−0.03
51		*p*-CN	1.48	0.28	1.33	0.38	1.21	0.02
52		*m*-$COCH_3$		1.33	1.93	1.02	1.45	0.26
53		*p*-$COCH_3$		1.19	2.00	1.04	1.58	0.33
54	Pyridines	3-Cl		0.96	1.10	0.84	0.57	−0.07
55		3-Br		0.89	1.17	0.81	0.57	−0.08
56		4-CH_3	2.07	2.01	2.13	1.70	1.30	0.74
57		3-CN		1.18	1.61	1.04	0.94	−0.06
58		4-CN	1.74	1.15	1.65	1.11	1.01	0.06

RESULTS

The results of three-mode factor analysis with missing data[10], applied to the data in Tables IV and V, are presented in Tables VI and XI, respectively. The results of an analysis of variance in which only the main effects of the adsorbents, eluents and solutes (but not their interactions) were taken into account are also presented in these tables. The computer program, called GEPCAM (generalized principal components analysis with missing values), as well as a mathematical treatment of its underlying theory[16], are available on request from the second pair of authors.

DISCUSSION

It follows from Table VI that the present data can be fitted better by the three-mode factor analysis model (even with only one factor for each mode) than by the (additive) analysis of variance model. An analysis of variance with first order interactions is hardly feasible because it involves too many free parameters, namely 138, and it is less suitable for making predictions.

Table VI shows further that the factor analysis model, eqn. 2, with only three factors for the solutes and two for the adsorbents and for the eluents already gives a better fit of the data in Table IV than does the Snyder equation. Introduction of a third factor for the adsorbents reduces the standard deviation of the fit even further, to a value that is only 1/3 of our estimate for the overall precision of the Snyder equation. With three factors for the solutes and the adsorbents and two for the eluents, eqn. 2 explains 99.7% of the variance of the data in Table IV. It can also be concluded that observations that strongly deviate from predictions by the Snyder equation can be fitted well by the three-mode factor analysis model.

Detailed information on this model is given in Table VII, which shows that the fit for each adsorbent is about equally good. The same conclusion applies to the eluents and to the solutes.

TABLE VI

SUMMARY OF THE RESULTS OF ANALYSIS OF VARIANCE AND OF THREE-MODE FACTOR ANALYSIS OF THE DATA, PRESENTED IN TABLE IV

Number of observations: 183. Number of missing values: 45. NFA = Number of factors for the adsorbents; NFE = number of factors for the eluents; NFS = number of factors for the solutes; NPAR = number of estimated parameters; DF = degrees of freedom; $\hat{\sigma}$ = standard deviation of the model; S_p = standard deviation of (new observation − prediction) when the new observations are generated by a random process; $\hat{\sigma}_{\text{pred.}}$ = average standard deviation of (new observation − prediction) for the missing values in the data set.

Model	*NFA*	*NFE*	*NFS*	*NPAR*	*DF*	$\hat{\sigma}$	$\sqrt{S_p}$	$\hat{\sigma}_{pred.}$
Analysis of variance				25	158	0.54		
Factor analysis	1	1	1	25	158	0.38	0.41	0.41
	2	2	2	50	133	0.23	0.27	0.26
	3	2	2	55	128	0.22	0.26	0.25
	2	2	3	68	115	0.11	0.14	0.20
	3	2	3	75	108	0.05	0.07	0.10
	3	2	4	93	90	0.05	0.07	32.15

TABLE VII

FIT OF THE THREE-MODE FACTOR ANALYSIS MODEL (3,2,3) FOR THE INDIVIDUAL ADSORBENTS, ELUENTS AND SOLUTES IN TABLE IV

NO = Number of observations; RMRSS = root of the mean residual sum of squares.

Adsorbent			*Eluent*			*Solutes*		
Code	*NO*	*RMRSS*	*Code*	*NO*	*RMRSS*	*Code*	*NO*	*RMRSS*
1	34	0.04	1	105	0.03	1	8	0.07
2	34	0.04	2	78	0.04	2	8	0.03
3	31	0.03				3	12	0.03
4	28	0.03				4	12	0.06
5	28	0.03				5	11	0.03
6	28	0.05				6	12	0.04
						7	7	0.03
						8	7	0.03
						9	8	0.03
						10	7	0.08
						11	12	0.03
						12	12	0.03
						13	10	0.04
						14	11	0.03
						15	12	0.03
						16	12	0.03
						17	12	0.03
						18	12	0.00
						19	6	0.03

It must be realized that a comparison of the fit of the data by the Snyder model and the factor analysis model alone does not do justice to the merits of the former. The Snyder model contains far less parameters than the factor analysis model (3,2,3), *i.e.*, 24 *vs.* 75, as the values for the constants, characterizing the eluents and the solutes, can be taken from literature. Moreover, the Snyder model gives a good deal of physical insight, whereas the factor analysis model is not a physical, but a mathematical–statistical model.

The ability of a statistical model to fit observations is not its most useful property. Far more important is its ability to predict accurate values for missing data. (This statement holds for a science like chemistry, where data are usually very precise, but their measurement is often costly or time-consuming. In such a situation there is little need for data smoothing but a clear need for the prediction of missing data. So, the choice for a particular statistical model should be made on the basis of its ability to predict missing data. In the social sciences the opposite situation exists, and here the choice of a statistical model should be based on its fit to the observations.) The ability of a model to predict missing data can be measured by the socalled S_p criterion, defined by $S_p = \hat{\sigma}^2 \left(1 + \frac{p}{n - p - 1}\right)$, where p = number of parameters (noted as NPAR in Table VI), n = number of observations and $\hat{\sigma}^2$ = variance of the model. The quantity S_p is an estimate of the variance of ($y_{\text{new}} - y_{\text{pred.}}$), where y_{new} is a new observation and $y_{\text{pred.}}$ its prediction, under the assumption that the

TABLE VIII

PREDICTIONS OF THE MISSING DATA IN TABLE IV BY THE THREE-MODE FACTOR ANALYSIS MODEL (3,2,3) AND THEIR 0.95 INTERVALS

Also given, for the sake of comparison, are predictions by Snyder's model.

Missing data			*Predictions according to*	
j	*k*	*l*	*Model (3,2,3)*	*Snyder*
3	2	1	−0.58 ± 0.16	−0.69
4	2	1	−0.52 ± 0.17	−0.42
5	2	1	−0.50 ± 0.17	−0.39
6	2	1	−0.58 ± 0.18	−0.36*
3	2	2	−0.78 ± 0.17	−0.74
4	2	2	−0.64 ± 0.18	−0.65
5	2	2	−0.65 ± 0.18	−0.63
6	2	2	−0.67 ± 0.19	−0.51
1	1	5	2.22 ± 0.18	2.07
1	2	7	−0.83 ± 0.18	−0.86
3	2	7	−0.69 ± 0.20	−0.67
4	2	7	−0.49 ± 0.19	−0.48
5	2	7	−0.52 ± 0.19	−0.48
6	2	7	−0.49 ± 0.20	−0.23*
2	2	8	−0.66 ± 0.13	−1.09*
3	2	8	−0.78 ± 0.17	−0.86
4	2	8	−0.51 ± 0.18	−0.64
5	2	8	−0.54 ± 0.18	−0.65
6	2	8	−0.51 ± 0.19	−0.37
1	2	9	−0.81 ± 0.15	−0.90
4	2	9	−0.39 ± 0.15	−0.28
5	2	9	−0.41 ± 0.15	−0.29
6	2	9	−0.40 ± 0.16	−0.02*
1	2	10	−0.96 ± 0.19	−1.18*
3	2	10	−1.05 ± 0.23	−0.94
4	2	10	−0.94 ± 0.23	−0.54*
5	2	10	−0.96 ± 0.23	−0.56*
6	2	10	−1.00 ± 0.24	−0.24*
3	1	13	0.91 ± 0.12	0.53*
3	2	13	−0.45 ± 0.12	−0.30*
2	2	14	−0.56 ± 0.12	−0.81*
2	1	18	0.54 ± 0.13	0.31*
2	2	18	−0.42 ± 0.13	−0.86*
4	1	18	1.53 ± 0.21	1.92*
4	2	18	0.46 ± 0.20	0.62
5	1	18	1.53 ± 0.20	1.94*
5	2	18	0.44 ± 0.20	0.61
6	1	18	1.64 ± 0.22	2.00*
6	2	18	0.50 ± 0.22	0.89*
4	1	19	2.25 ± 0.23	3.07*
4	2	19	1.08 ± 0.23	1.62*
5	1	19	2.26 ± 0.23	3.09*
5	2	19	1.07 ± 0.23	1.60*
6	1	19	2.41 ± 0.25	3.12*
6	2	19	1.63 ± 0.25	1.88

observations are generated by a kind of random mechanism[17]. Using this criterion we learn from the last but one column of Table VI that (3,2,3) is the best model and that it is no use introducing more factors.

The assumption about the way observations are generated is not very plausible in our situation. A better idea about the appropriateness of a model can be obtained by computing $\sigma^2_{\text{pred.}}$ = variance of $(y_{\text{new}} - y_{\text{pred.}})$ averaged over the missing data. The value of $\sigma^2_{\text{pred.}}$ can be computed by extension of the procedure, developed for the two-mode factor analysis case[10], to the three-mode situation. Using a linearization of the model around the true parameters, the variance of $(y_{\text{new}} - y_{\text{pred.}})$ can be computed for all missing data. From the values of $\hat{\sigma}_{\text{pred.}}$, presented in the last column of Table VI, we learn again that (3,2,3) is the model to be preferred and that (4,2,3) is of no use, because it gives nonsensical predictions. Moreover, this technique enables us to give 0.95-prediction intervals for future observations of a missing datum. These prediction intervals, calculated with the model (3,2,3), are given in Table VIII together with the calculated values for the missing data.

Values for missing data can also be calculated by the Snyder equation and these results are collected in the last column of Table VIII. On comparing the last two columns of this table one notices several values, predicted with the Snyder equation, that deviate so strongly that they do not fall within the 0.95-prediction interval calculated with the (3,2,3) model. (These cases are marked with an asterisk in the last column of Table VIII.)

It would be interesting to learn with which physicochemical properties of the three "modes" the factors of the model (3,2,3) are related. In Snyder's physical model of adsorption chromatography the adsorbents are characterized by four variables, the eluents by one variable and the solutes by four variables. The physical meaning of these variables is indicated in the legend to Table I. Thus, it would be interesting to investigate the regression of the factors characterizing the six adsorbents in the factor analysis model on the four variables characterizing the adsorbents in Snyder's model. However, an attempt to estimate four regression coefficients from only six data is clearly not warranted. The same situation occurs for the eluents, where regression analysis would mean the estimation of one regression coefficient from two data. However, in the case of the solutes we have nineteen data, from which we can estimate the regression coefficients, Q, in the equation:

$$S_l = Q_0 + Q_1 S_0 + Q_2 \Sigma a_i \text{(calc.)} + Q_3 \Sigma \Delta a_i (\text{SiO}_2) + Q_4 (n-6) \qquad (5)$$

It appeared that the coefficient Q_3 is not significantly different from zero, for l = 1–3. Therefore, we investigated the regression equation:

$$S_l = Q_0 + Q_1 S_0 + Q_2 \Sigma a_i \text{(calc.)} + Q_3 (n-6) \qquad (6)$$

The values of the variables for the solutes are given in Table IX, and the results of the regression analysis are given in Table X. The values of $\hat{\sigma}$, and in particular those of f, in this table show that there is a close relationship between S_1 and Snyder's variables characterizing the solutes, but only a poor relationship between S_2 or S_3 and Snyder's variables (a value of $f < 0.1$ indicates good precision in regression analysis[18]).

The ratio of the mean values of S_0, Σa_i (calc.) and $n-6$ is approximately 2:3:5.

TABLE IX

VARIABLES, CHARACTERIZING THE SOLUTES IN THE THREE-MODE FACTOR ANALYSIS MODEL (3,2,3) AND IN SNYDER'S MODEL

Solute code	*Factor analysis model*			*Snyder's model*		
	S_1	S_2	S_3	S_0	Σa_i*(calc.)*	$n-6$
1	0.074	0.237	−0.229	1.83	1.1	0
2	0.052	0.354	−0.089	1.29	1.7	0
3	0.186	0.084	−0.207	2.77	1.3	0
4	0.259	−0.058	−0.341	3.33	0.6	0
5	0.325	−0.234	−0.556	4.69	1.5	0
6	0.230	−0.045	−0.428	3.45	2.3	0
7	0.037	0.319	0.042	1.0	2.1	4
8	0.062	0.347	0.043	1.0	3.7	4
9	0.100	0.286	−0.003	1.5	3.7	6
10	0.039	0.490	−0.155	1.5	6.4	6
11	0.153	0.234	0.089	2.0	4.2	8
12	0.156	0.214	0.080	2.0	4.2	8
13	0.215	0.113	0.169	2.5	4.5	10
14	0.216	0.133	0.143	2.5	4.5	10
15	0.276	0.042	0.136	3.0	6.3	12
16	0.342	−0.075	0.223	3.5	6.8	14
17	0.355	−0.120	0.215	3.5	6.8	14
18	0.287	−0.003	0.151	3.0	6.3	12
19	0.416	−0.231	0.246	4.5	7.8	18

Thus it can be concluded from the values of the regression coefficients Q in Table X that the factor S_1 is mainly related with Snyder's variable S_0. In the factor S_2 the contributions of Snyder's variables Σa_i (calc.) and $n-6$ are relatively more important, and the factor S_3 is mainly a measure of the number of aromatic carbon atoms, n.

The correctness of the three-mode factor analysis model for these relatively simple compounds raised our interest in further investigations on more complicated molecules. Therefore a data set for substituted phenols, anilines and pyridines, measured in eluents 2 and 3 (pure methylene chloride) on adsorbents 1, 2 and 3 was compiled from references 12–14 (Table V). The results of an analysis of variance, and of three-mode factor analysis of this data set, are summarized in Table XI. Again it appears that the data can be fitted better by the three-mode factor analysis model (even with only one factor for each mode) than by the analysis of variance model. With three factors for the solutes and two for the adsorbents and the eluents the standard deviation of the fit by the factor analysis model ($\hat{\sigma}$ = 0.08) is already better than that of the fit by regression analyses according to the Snyder equation ($\hat{\sigma}$ 0.10–0.28). With the (3,2,4) model, $\hat{\sigma}$ is as low as 0.05, an impressive figure for these complicated solutes. From the data on $\hat{\sigma}_{\mathrm{pred.}}$ in the last column of Table XI it follows that the (3,2,4) model is the best one to make predictions for the missing data in Table V*.

* In the (3,2,3) and the (3,2,4) model the maximum number of factors for the adsorbents (3) and the eluents (2) is extracted. The consequence is that these two modes can be combined to a single adsorbent/eluent mode, so that the data can be classified in two modes. Accordingly, they can be analysed with two-mode factor analysis with identical results.

TABLE X

RESULTS OF THE REGRESSION ANALYSIS ACCORDING TO EQN. 6

$\hat{\sigma}$ = Standard deviation of the regression; RMSS = root of the mean sum of squares of S_l.

Factor	S_1	S_2	S_3
Q_0	−0.051±0.016	0.49 ±0.03	0.06 ±0.04
Q_1	0.089±0.005	−0.175±0.009	−0.102±0.012
Q_2	−0.013±0.006	0.047±0.012	−0.059±0.015
Q_3	0.011±0.002	−0.019±0.005	0.062±0.006
$\hat{\sigma}$	0.02	0.04	0.05
$f = \hat{\sigma}$/RMSS	0.09	0.18	0.22

Detailed information on this model is given in Table XII. Again it appears that the fit for each adsorbent, each eluent and each solute is about equally good.

This model has been used to calculate values for missing data, which are presented in Table XIII, together with the 0.95-prediction intervals.

For the sake of comparison we made also predictions according to the Snyder equation (which is for these solutes more complicated than eqn. 3). The predicted values are given in the last column of Table XIII. Now we notice that in almost all cases predictions by the Snyder equation deviate significantly from the values calculated by the (3,2,4) model. These cases are marked with an asterisk in Table XIII.

TABLE XI

SUMMARY OF THE RESULTS OF ANALYSIS OF VARIANCE AND OF THREE-MODE FACTOR ANALYSIS OF THE DATA, PRESENTED IN TABLE V

Number of observations: 213. Number of missing values: 21. For other symbols see Table VI.

Model	*NFA*	*NFE*	*NFS*	*NPAR*	*DF*	$\hat{\sigma}$	$\sqrt{S_p}$	$\hat{\sigma}_{pred.}$
Analysis of variance				42	171	0.58		
Factor analysis	1	1	1	42	171	0.38	0.42	0.47
	2	2	2	84	129	0.21	0.27	0.32
	2	2	3	122	91	0.08	0.12	0.14
	3	2	3	126	87	0.07	0.11	0.13
	3	2	4	164	49	0.05	0.11	0.11

CONCLUSIONS

The Snyder equation, which has proven its value for bare silica and alumina adsorbents, is not quite satisfactory for fitting retention data in normal phase HPLC on chemically bonded phases. This equation is in fact a special case of the mathematical–statistical three-mode factor analysis model. This model, in its general form, appears to be able to fit data on the retention of nineteen simple solutes (*viz.*, monosubstituted benzenes and unsubstituted polycyclic aromatics) on six adsorbents with the eluents *n*-hexane and 35% (v/v) methylene chloride in *n*-hexane very satisfactorily. With only three parameters or "factors" for the solutes and the adsorbents, and two for the eluents, the standard deviation of the fit of log V_N/W data is only 0.05. This model is also able to fit a set of data on 39 more complicated solutes (*viz.*,

TABLE XII

FIT OF THE THREE-MODE FACTOR ANALYSIS MODEL (3,2,4) FOR THE INDIVIDUAL ADSORBENTS, ELUENTS AND SOLUTES IN TABLE V

NO = Number of observations; RMRSS = root of the mean residual sum of squares.

Adsorbent			*Eluent*			*Solute*		
Code	*NO*	*RMRSS*	*Code*	*NO*	*RMRSS*	*Code*	*NO*	*RMRSS*
1	65	0.00	2	96	0.03	20	5	0.00
2	78	0.03	3	117	0.03	21	5	0.00
3	70	0.01				22	6	0.00
						23	6	0.00
						24	6	0.03
						25	6	0.03
						26	6	0.03
						27	6	0.03
						28	6	0.00
						29	6	0.05
						30	5	0.00
						31	5	0.00
						32	4	0.00
						33	5	0.00
						34	4	0.00
						35	4	0.00
						36	4	0.00
						37	4	0.00
						38	6	0.03
						39	6	0.00
						40	6	0.03
						41	6	0.00
						42	6	0.03
						43	6	0.00
						44	6	0.04
						45	6	0.00
						46	6	0.03
						47	5	0.00
						48	6	0.00
						49	6	0.04
						50	6	0.00
						51	6	0.03
						52	5	0.00
						53	5	0.03
						54	5	0.04
						55	5	0.03
						56	6	0.04
						57	5	0.03
						58	6	0.03

substituted phenols, anilines and pyridines) on three adsorbents with the eluents 35% (v/v) methylene chloride in *n*-hexane and pure methylene chloride. With four factors for the solutes, three for the adsorbents and two for the eluents, the standard deviation of the fit of the observations is only 0.05. The most important feature of a mathematical–statistical model is its ability to predict missing data. In the analysed

TABLE XIII

PREDICTIONS OF THE MISSING DATA IN TABLE V BY THE THREE-MODE FACTOR ANALYSIS MODEL (3,2,4) AND THEIR 0.95 INTERVALS

Also given, for the sake of comparison, are predictions by Snyder's model.

Missing data			*Predictions according to*	
j	*k*	*l*	*Model (3,2,4)*	*Snyder*
1	1	20	0.96 ± 0.22	1.10
1	1	21	0.81 ± 0.17	1.09*
3	1	30	2.58 ± 0.24	1.26*[a]
3	1	31	3.31 ± 0.28	1.62*[a]
1	1	32	1.78 ± 0.20	1.97
3	1	32	2.91 ± 0.27	1.11*[a]
3	1	33	3.12 ± 0.25	1.34*[a]
1	1	34	1.74 ± 0.19	1.95*
3	1	34	2.44 ± 0.26	1.16*[a]
1	1	35	1.79 ± 0.20	2.00*
3	1	35	2.73 ± 0.27	1.37*[a]
1	1	36	2.08 ± 0.21	2.44*
3	1	36	2.83 ± 0.28	1.28*[a]
1	1	37	2.17 ± 0.20	2.52*
3	1	37	3.14 ± 0.28	1.61*[a]
1	1	47	1.85 ± 0.18	2.35*
1	1	52	2.13 ± 0.18	2.59*
1	1	53	2.12 ± 0.17	2.48*
1	1	54	1.29 ± 0.18	[b]
1	1	55	1.32 ± 0.18	[b]
1	1	57	1.84 ± 0.18	[b]

[a] Calculated with unpublished β, ρ and δ values of the authors of ref. 14.

[b] Impossible to predict because of lack of A_s values for pyridine[11].

sets of data, about 20 and 10% of the data are missing, respectively. The missing data can be predicted by three-mode factor analysis with a standard deviation of 0.10 and 0.11 respectively.

REFERENCES

1 H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, Chicago, 1970.

2 P. T. Funke, E. R. Malinovski, D. E. Martire and L. Z. Pollara, *Separ. Sci.*, 1 (1966) 661.

3 P. H. Weiner and J. R. Parcher, *J. Chromatogr. Sci.*, 10 (1972) 612.

4 S. Wold and K. Andersson, *J. Chromatogr.*, 80 (1973) 43.

5 R. B. Selzer and D. G. Howery, *J. Chromatogr.*, 115 (1975) 139.

6 D. H. Mc Closkey and S. J. Hawkes, *J. Chromatogr. Sci.*, 13 (1975) 1.

7 C. L. de Ligny, G. H. E. Nieuwdorp, W. K. Brederode, W. E. Hammers and J. C. van Houwelingen, *Technometrics*, 23 (1981) 91.

8 L. R. Tucker, in C. W. Harris (Editor), *Problems in Measuring Change*, University of Wisconsin Press, Madison, 1963.

9 P. M. Kroonenberg and J. de Leeuw, *Psychometrica*, 45 (1980) 69.

10 H. M. Weesie and J. C. van Houwelingen, *GEPCAM Users' Manual*, Institute of Mathematical Statistics, Utrecht, 1983.

11 L. R. Snyder, *Principles of Adsorption Chromatography*, Marcel Dekker, New York, 1968.

12 W. E. Hammers, R. H. A. M. Janssen, A. G. Baars and C. L. de Ligny, *J. Chromatogr.*, 167 (1978) 273.

13 W. E. Hammers, C. H. Kos, W. K. Brederode and C. L. de Ligny, *J. Chromatogr.*, 168 (1979) 9.

14 W. E. Hammers, M. C. Spanjer and C. L. de Ligny, *J. Chromatogr.*, 174 (1979) 291.

15 W. E. Hammers, A. G. M. Theeuwes, W. K. Brederode and C. L. de Ligny, *J. Chromatogr.*, 234 (1982) 321.

16 J. C. van Houwelingen, *Proceedings of the 3rd Prague Symposium on Asymptotic Statistics, Prague, 1983*, North-Holland, Amsterdam 1984, in press.

17 L. Breiman and D. Freedman, *J. Amer. Stat. Assoc.*, 78 (1983) 131.

18 S. Ehrenson, R. T. C. Brownlee and R. W. Taft, *Progr. Phys. Org. Chem.*, 9 (1958) 287.