DETERMINATION OF NUMBER OF LINEAR LATENT VARIABLES IN A SET OF EXPERIMENTAL CHEMICAL DATA

Oldřich PYTELA

Department of Organic Chemistry,
Institute of Chemical Technology, 532 10 Pardubice

Received January 26, 1989
Accepted May 31, 1989

Dedicated to Professor Otto Exner on the occasion of his 65th birthday.

A criterion has been suggested for determination of number of linear latent variables in experimental chemical sets which is based on the principle of scatter analysis applied to the method of conjugated deviations. The criterion suggested has been tested on data with preliminarily given error, and correct numbers of latent variables were obtained until the error load of 70%. The phenomenon of dominance of latent variables with the maximum variability has been analyzed and tested. The criterion suggested has been applied to three classes of data sets — homogeneous complete, homogeneous incomplete, and nonhomogeneous incomplete. A good accordance with other criteria has been obtained first of all for the first data class, in most cases the found numbers of latent variables agreed with the experiment interpreted.

Construction and analysis of latent variables from experimental results represents one of the most efficient tools for finding more general regularities. This approach is adopted in a number of scientific disciplines, but chemistry offers greatest possibilities for applications of these methods denoted as methods of analysis of latent variables. In chemistry there exist relatively extensive sets of precise experimental results which can be interpreted by means of linear models. Most frequently applied in chemistry are the factor analysis (FA)[1–4] and some of its variants, first of all the principal component analysis (PCA)[5–8], and — for relation between two or more data sets — the method of partial least squares (PLS)[9–12]. The principal problem of all these methods consists in determination of the minimum necessary number of latent variables which express the nature of the observed (manifested) features of the system measured without involving the experimental error. A number of criteria have been suggested for this purpose which are based on estimates of experimental error[6,13] or on some other basis, e.g. Malinowski's indicator function[6,14], cross-validation technique[15] or, as the case may be, on statistical tests and criteria constructed in various ways[3,5,6,16–20]. Usually it is recommended to combine several criteria because not always and not all of them lead to unambiguous results.

Another way of approach to solution of the problem of a valid number of latent

variables is offered by the method of conjugated deviations which has already been used earlier for successful parametrization of some dependences[21,22]. This method is based on the multiple linear regression, and the statistical tests currently used in this field can be adopted to decide the number of latent variables. The aim of this present communication is to verify the possibility of application of statistical tests of deciding about the number of significant latent variables determined by the method of conjugated deviations, and – in the subsequent phase – to utilize these tests for analysis of experimental chemical sets taken from literature.

THEORETICAL

The method of conjugated deviations[21,22] works with the model known from the factor analysis

$$\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{A}' + \boldsymbol{E}\,, \tag{1}$$

where $\boldsymbol{Y}$ represents the matrix of standardized experimental data of manifested variables with n rows and m columns, $\boldsymbol{P}$ means the matrix of latent variables with n rows and $p + 1$ columns, $\boldsymbol{A}$ is the matrix of regression coefficients (it corresponds to the matrix of factor loadings), and $\boldsymbol{E}$ is the matrix of errors. The first column of the matrix $\boldsymbol{P}$ is always one, the first row in the matrix $\boldsymbol{A}$ expresses the absolute term in the regression. For each element of the $\boldsymbol{Y}$ matrix it is possible to define its estimate in the form

$$\hat{y}_{ij} = a_{j0} + \sum_{k=1}^{p} a_{jk} P_{ik}\,, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, m, k = 1, 2, \ldots, p\,. \tag{2}$$

The basic idea of the method is that the fit of correlation between manifest and latent variables according to Eq. (*1*) and/or (*2*) is a function of not only the experimental error but also the value explaining the latent variables. For a given number p of the latent variables the decomposition according to Eq. (*1*) is carried out in such way that the quantity s^2 defined by Eq. (*3*) might be minimized.

$$s^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} e_{ij}^2 / \left(\sum_{j=1}^{m} N_j - np - \sum_{j=1}^{m} \bar{P}_j\right) = S_R / \nu_R \tag{3}$$

In this equation e_{ij} ($e_{ij} = y_{ij} - \hat{y}_{ij}$) are elements of the $\boldsymbol{E}$ matrix, N_j represents the number of values in the j-th column of the $\boldsymbol{Y}$ matrix (which need not be complete), and $\bar{P}_j$ represents the number of statistically significant regression parameters (according to t-test) in the regression with the data of the j-th column, S_R then expresses the residual sum of squares, and ν_R symbolizes the number of degrees of freedom. The difference e_{ij} is only defined for known elements of the $\boldsymbol{Y}$ matrix, otherwise it is zero. The minimization of criterion (*3*) is iterative with a correction

according to Eq. (*4*) in each step.

$$P_{ik} = P_{ik}^{0} + \alpha \sum_{j=1}^{m} \Delta P_{ijk} \tag{4}$$

In this equation α means the damping factor with the recommended value of 0·001. In practice, the expression (*5*) proved useful for the correction ΔP_{ijk}.

$$\Delta P_{ijk} = a_{jk}(y_{ij} - \hat{y}_{ij}) / \sum_{k=1}^{p} |a_{jk}| \tag{5}$$

The quantity s is naturaly dependent on the number p of latent variables. A significant value of p can be determined on the basis of scatter analysis applied in the regression analysis[2,23]. A change in number of latent variables from p to $p + 1$ undoubtedly results in a decrease of S_R value in Eq. (*3*) (but, generally, not in a decrease of s value). If this change is statistically insignificant as compared with the S_R value found for a model with $p + 1$ latent variables, it is meaningless to adopt more than p latent variables in the model (*1*). The procedure outlined leads to a test of the hypothesis with the criterion

$$F(v_R(p) - v_R(p+1), v_R(p+1)) =$$

$$= \frac{S_R(p) - S_R(p+1)}{S_R(p+1)} \frac{v_R(p+1)}{v_R(p) - v_R(p+1)}, \quad p = 1, 2, 3, \ldots, \tag{6}$$

the values of S_R and v_R being given by Eq. (*3*). If the value of criterion (*6*) is higher than that of the corresponding quantile of Fisher–Snedecor distribution, then an addition of another latent variable is meaningful, otherwise the testing is finished.

When constructing latent variables one usually states how many % of variability of the original source matrix matrix **Y** is given a true picture by the model with p latent variables. In our case this quantity can be calculated from Eq. (*7*)

$$V = 100\,(S_T - S_R)/S_T\,(\%)\,, \tag{7}$$

where S_T means the overall variability given by Eq. (*8*)

$$S_T = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_j)^2\,, \tag{8}$$

where $\bar{y}_j$ denotes the arithmetic mean of the j-th column of the **Y** matrix.

Per analogiam, the quantity (*9*) has the meaning of a selection coefficient of multiple correlation. The relation between the expressions (*7*) and (*9*) is obvious.

$$R = (1 - S_R/S_T)^{1/2} \tag{9}$$

It is also possible to construct a criterion to test the hypothesis of $R = 0$ (insignificance of the model (*1*)) as

$$F(v_T, v_R) = (S_T/v_T)/(S_R/v_R)\,, \qquad (10)$$

where v_T symbolizes the number of columns of **Y** matrix for which at least one regression coefficient was statistically significant in the regression carried out, hence usually $v_T = m$. The model (*1*) is significant, if the value of the criterion (*10*) is greater than the corresponding quantile of Fisher–Snedecor distribution.

The methods of analysis of latent variables are usually applied to data sets of different structure and quality. Such sets can roughly be divided into two groups – homogeneous (H) and nonhomogeneous (NH). The homogeneity denotes here the same physico-chemical meaning of elements of the **Y** matrix with regard to both their form and content. Examples of homogeneous sets, e.g., involve the set of chemical shifts measured on several nuclei of substituted derivatives of some compound, the set of pK values measured in various solvents for derivatives of a compound, the set of absorbances of an indicator as a function of pH and wavelength, etc. Nonhomogeneous sets include results of measurements by different techniques, under different conditions, etc., only the property studied (e.g. solvent effects, nucleophilicity, substituent effects) being common for them. Typical nonhomogeneous sets were treated in our earlier studies[21,22]. Both homogeneous and nonhomogeneous sets can be either complete (C) or incomplete (NC). Usually homogeneous sets are complete or almost complete because they are a result of a single study. On the other hand, nonhomogeneous sets are practically always incomplete. The construction and analysis of latent variables in homogeneous and especially complete homogeneous sets usually presents no difficulties. A somewhat different situation is encountered with incomplete nonhomogeneous sets. The results presented can be distorted, and it only depends on robustness of the statistical criteria and method used, whether or not one obtains conclusions adequate to the phenomenon studied. In this respect, the advantages of the method of conjugated deviations could make themselves felt, in particular, its less sensitivity to missing data and, moreover, the phenomenon of dominance. This follows from the fact that in constructing latent variables the principle of maximum explained variability is always important. The recessive factors not related to the dominant factors selected are thus eliminated, as a consequence of application of the significance test for the regression coefficients in the method of conjugated deviations, and do not affect the result.

COMPUTATIONAL

The calculations were carried out according to known algorithm[21,22] in the FORTRAN language using an EC 1033 computer. For testing of correctness and robustness of the criterion (*6*) we used a model matrix **Y** constructed according to

Eq. (*1*). The **P** matrix was given by 3 orthogonal (verified) vectors with 20 elements taken from tables of random numbers[24] standardized into the interval $\langle 0, 1 \rangle$. The structure of this matrix is given in Table I. For construction of the **E** matrix we created a vector with normal distribution (0, 0·985) whose 20 elements were from the interval $\langle -2{\cdot}097, 2{\cdot}097 \rangle$ symmetrically around μ. The **E** matrix was then constructed from 10 columns given by a $\varkappa$ multiple of the described vector with permuted elements. The maximum correlation coefficient found between two columns of the **E** matrix was 0·075, the variability of **E** matrix explained by 9 principal components in PCA was 90·9%. The test of dominance was carried out with the matrix **Y** derived from the matrix **P** whose first column was repeated eight times.

RESULTS AND DISCUSSION

The F Criterion Test (Eq. (6))

The calculations were carried out for $p = 1, 2, 3$, and 4 with the testing matrix Y (see Computational); the error load in the **E** matrix was varied within the limits from 0·01 to 1. The results are given in Table II from which it follows that the testing criterion (*6*) provides correct results up to the error load $\varkappa = 0{\cdot}7$. Hence the criterion given is sufficiently robust. Also correctly determined is the residual standard deviation s (calculated from residual variance (*3*)) which must correspond to the value of $\varkappa$ for the error distribution $\mathbf{N}(0, \sigma^2)$ in the **E** matrix and for a correct number of latent variables (here $p = 3$). The smaller s value as compared with $\varkappa$ for $\varkappa > 0{\cdot}5$ results from non-orthogonality of columns of the **Y** matrix at the error loads given. This phenomenon cannot practically be avoided (undoubtedly it is present in real data, too), but it is not prejudicial since the error load in question is a limiting one.

Dominance Test

In the data tested so far, the individual latent variables were present to roughly the same extent (Table I). In practice, however, such cases can be expected where one

TABLE I

The loadings in matrix **A** used for constructing the model data

Column	1	2	3	4	5	6	7	8	9	10
a_0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
a_1	0·0	0·0	0·1	0·0	0·1	0·1	0·1	0·1	1·0	10·0
a_2	0·0	1·0	0·0	1·0	0·0	1·0	1·0	0·1	1·0	10·0
a_3	10·0	0·0	0·0	10·0	10·0	0·0	10·0	0·1	1·0	10·0

or a few latent variable(s) is (are) present in all columns of the source matrix **Y**, whereas the other existing latent variables are present only in one or a few column(s) of the source matrix. With regard to the extent of explained variability, the latent variables of the first group represent the dominant component, whereas the latent variables of the second group represent the minor component. It can even happen that the variability explained by the latent variables of the second group is comparable

TABLE II

Analysis of correctness and robustness of the criterion (6) on the model data with error load within the limits from $\varkappa = 0{\cdot}01$ to $\varkappa = 1$; N means number of the data taken into the calculation, for other symbols see the text

$\varkappa$	p	s	v_R	V	N	$F(6)$	$F_{0.95}$
0·01	1	0·576	142	72·4	180	—	—
	2	0·234	113	96·4	180	25·76	1·57
	3	0·011	100	99·994	200	4 075·46	1·58
	4	0·012	77	99·995	200	0·47	1·67
0·05	1	0·574	142	72·7	180	—	—
	2	0·242	113	96·1	180	23·65	1·57
	3	0·054	100	99·8	200	166·89	1·58
	4	0·056	77	99·87	200	0·71	1·67
0·1	1	0·573	142	72·7	180	—	—
	2	0·260	113	95·5	180	19·91	1·57
	3	0·108	101	99·4	200	46·55	1·85
	4	0·114	78	99·5	200	0·50	1·67
0·5	1	0·658	142	64·0	180	—	—
	2	0·606	132	74·5	200	3·55	1·90
	3	0·480	108	86·9	200	4·29	1·62
	4	0·492	85	89·2	200	0·76	1·66
0·7	1	0·722	142	56·7	180	—	—
	2	0·702	135	65·0	200	2·18	2·08
	3	0·624	110	77·0	200	2·43	1·61
	4	0·645	88	80·8	200	0·68	1·66
0·8	1	0·752	142	53·1	180	—	—
	2	0·735	136	61·3	200	2·09	2·17
	3	0·688	111	72·3	200	1·77	1·61
	4	0·708	88	76·8	200	0·73	1·65
1·0	1	0·804	142	46·4	180	—	—
	2	0·794	136	54·9	200	1·60	2·17
	3	0·772	112	64·8	200	1·32	1·61
	4	0·795	89	70·4	200	0·73	1·65

with the unexplained variability, and these latent variables thus appear to be statistically insignificant. Therefore, the source matrices **Y** were constructed with different ratios of the individual latent variables present, using the error load $\varkappa = 0{\cdot}05$ (Table III). As shown in Table III, even at the ratio of 2 : 1 : 1 one can observe a transient increase of statistics (*3*) with the number of assessed latent variables. As this statistics forms a basis for the test (*6*), this phenomenon could represent a source of incorrect decision about the number of latent variables in cases of analyses based on an unsufficient extent of the p quantity. The results of test (*10*) (Table III) confirm the correctness of decision about the number of latent variables based on the criterion (*6*). In addition it can be expected that the dominance will make itself felt to a smaller extent with a greater error load, as it is indicated by the data of Table II.

Analysis of H–C *Type Sets*

This type of matrices has been investigated in detail recently by Malinowski[18] with regard to the number of latent variables. Therefore, this present communication also

TABLE III

Analysis of dominance at various ratios of proportions of latent variables for $\varkappa = 0{\cdot}05$; for the symbols see Table II and the text

Ratio	p	s	v_R	V	N	$F(6)$	$F_{0.95}$
8 : 1 : 1	1	0·052	124	99·8	160	—	—
	2	0·344	139	91·3	200	—	—
	3	0·052	114	99·8	200	237·37	1·60
	4	0·054	92	99·86	200	0·60	1·66
6 : 1 : 1	1	0·052	88	99·8	120	—	—
	2	0·397	104	89·2	160	—	—
	3	0·054	79	99·8	160	222·16	1·65
	4	0·058	56	99·9	160	0·56	1·72
4 : 1 : 1	1	0·052	52	99·8	80	—	—
	2	0·489	68	85·7	120	—	—
	3	0·056	44	99·88	120	211·62	1·77
	4	0·065	20	99·93	120	0·56	2·08
2 : 1 : 1[a]	1	0·055	16	99·87	40	—	—
	2	0·734	30	99·89	80	—	—
	3	0·107	4	99·94	80	54·14	5·84

[a] At this ratio it is impossible to carry out the calculation for $p = 4$ because of the insufficient number of degrees of freedom.

includes the data used by the author[18] (as far as the source matrices were available) and the results are compared (Example 2–5). Example *1* can serve as a test of the criterion (*6*) on a real matrix. This is formed by the relative content of 17 amino acids in a mixture of ribonuclease and chymotrypsinogen[25,26]. The **Y** matrix should contain 2 latent variables, which also was unambiguously found (Table IV). Example *2* is of similar nature. The data matrix is formed by relative intensities of MS peaks of a mixture hexane–cyclohexane[19,27]. Surprisingly (but in accordance with ref.[18]), the number of latent variables found was $p = 3$. The m/e 28 signal manifested univocally as one of the latent variables. Probably this signal is caused by the background (N_2, CO_2) and was not excluded (for unknown reasons) from the mass spectrum. If this is done, one obtains (again in accordance with ref.[18]) the expected number $p = 2$ (Table IV, Example *3*). Example *4* is based on the (at present already classic) matrix describing the ^1H chemical shifts of methane derivatives in 14 solvents[28]. In this case, too, our results are in accordance with literature, as it can be seen in Table IV. However, different results were obtained in the next Example 5. The data matrix was obtained from the ^{19}F chemical shifts of 19 fluorinated hydrocarbons in 8 solvents[29]. Whereas the analysis by Malinowski[18] indicates 3 latent variables, the result of the criterion (*6*) indicates 4. The difference is probably due to high precision of the data and excellent explanation of experimental variability by the model used (see Table IV), which makes it possible to prove a greater number of existing latent variables. Example *6* represents an analysis of another (already classic) matrix whose elements are formed by logarithm of ratio of dissociation constants of substituted and unsubstituted benzoic acids in 8 solvents[30]. As the *ortho* derivatives are included too, one can expect the number of latent variables to be equal to at least two. The author[30] estimates this number as 3–4, the criterion (*6*) gives an univocal result of $p = 3$. In the analysis the dominance was obvious for $p = 1$ as indicated by the values of the criterion (*6*) for $p = 1$ (5·86) and $p = 2$ (15·27). The next two examples show situations where the testing with this criterion failed. Example *7* is based on a matrix representing pK values of 13 derivatives of N-(subst.phenyl)benzenesulfonamide in 5 solvents[31]. In the paper cited an analysis was carried out regarding the number of factors which affect the variability of the experimental matrix obtained, and it was stated that it is greater than 2. The application of the criterion (*6*) leads to a value $p \geqq 3$, and the testing cannot be continued because of the insufficient number of degrees of freedom. This problem is not encountered in treating the matrix in the last example of this series (Example 8), but another problem consists in a slow convergence of the criterion (*6*) which, after a relatively significant decrease for $p = 1$ (7·44) begins to stagnate (for $p = 2$ it is $F = 3{\cdot}96$), and for $p = 6$ it still reaches the value of 3·35. This is probably due to the way of treatment of the matrix whose columns represent relatively different compounds (e.g. benzene, butanol, pyridine) for which the retention indexes (rows) were measured in a large number (226) of liquid phases. Wold[15] using PCA with

TABLE IV

Number of latent variables p determined according to criterion (*6*) at significance level $\alpha = 0{\cdot}05$, number of latent variables p_{lit} given in literature, residual standard deviation s, number of degrees of freedom v_R, variability explained by the method of conjugated deviations V, and by principal component analysis V_{PCA}, values of the F criterion according to Eq. (*10*) and dimensions of source matrix Y

Example No.	p	p_{lit}	s	v_R	V	V_{PCA}	$F(10)$	$n \times m$	Ref.
1	2	2	0·075	50	99·7	99·7	$2{\cdot}88 \cdot 10^3$	17 × 6	25, 26
2	3	3	0·048	44	99·91	99·91	$7{\cdot}33 \cdot 10^3$	18 × 7	27, 18
3	2	2	0·044	64	99·89	99·89	$8{\cdot}26 \cdot 10^3$	17 × 7	27, 18
4	3	3	0·005	51	99·999	99·999	$4{\cdot}42 \cdot 10^5$	14 × 9	28, 18
5	4	3	0·001	19	100·000	100·000	$2{\cdot}35 \cdot 10^7$	19 × 8	29, 18
6	3	4(3)	0·038	64	99·94	99·94	$1{\cdot}27 \cdot 10^4$	19 × 8	30
7[a]	≧3	2	0·093	9	99·87	99·86	$1{\cdot}40 \cdot 10^3$	13 × 5	31
8[a]	≧7	—[b]	0·049	600	99·94	99·93	$9{\cdot}57 \cdot 10^4$	226 × 10	32, 15, 16
9	2(1)	1	0·141	186	98·4	98·4	$1{\cdot}62 \cdot 10^3$	35 × 7	33, 34
10	2	2	0·270	77	96·0	95·6	$3{\cdot}23 \cdot 10^2$	26 × 6	35
11	4	2	0·150	152	98·7	95·7	$9{\cdot}99 \cdot 10^2$	26 × 12	36
12[a]	≧4	2	0·052	53	99·9	93·1	$7{\cdot}98 \cdot 10^3$	49 × 6	37, 38
13	2	2	0·544	361	77·0	74·9	$3{\cdot}83 \cdot 10^1$	28 × 41	21
14[a]	≧6	—[b]	0·323	4 534	92·0	—[c]	$9{\cdot}59 \cdot 10^0$	51 × 367	22, 40

[a] The values are given for the declared number of latent variables; [b] see the text; [c] for technical reasons the determination was impossible.

cross-validation estimate, arrived at the value of $p = 2$, other authors[16] in similar way arrived at the value of $p = 3$. The discussion can be concluded by the statement that the dominant variability is explained by 2 latent variables which, however, do not statistically significantly explain all the experimental variability.

On the whole it can be stated that the sets of H–C type in chemistry can be interpreted by the model (*1*) very well, as it can be seen from the residual standard deviations s as well as from the explained variability V and values of the criterion (*10*) (Table IV). It is noteworthy that the s values are better than or comparable with the estimated or declared experimental error in all the cases. The criterion (*6*) provides numbers of latent variables which agree well with those determined according to other criteria or by analysis of the problem.

Analysis of H–NC *Type Sets*

The first illustration is represented by Example *9* in which the source matrix is formed by pK values of 35 *meta* and *para* substituted benzoic acids in 7 solvents[33]. The matrix is filled up to 95·1%. The value of the criterion (*6*) can be denoted as a limiting one, since for $p = 1$ it is $F(35, 151) = 1{\cdot}49$ at the critical values of $F_{0.95} = 1{\cdot}50$ and $F_{0.90} = 1{\cdot}37$. With regard to the closeness of interpretation by the given residual standard deviation ($s = 0{\cdot}141$, Table IV) it is more correct to reject the hypothesis for $p = 1$. The value of criterion (*6*) for $p = 2$ is $F = 1{\cdot}09$, whereas the critical values are $F_{0.95}(36, 115) = 1{\cdot}52$ and $F_{0.90}(36, 115) = 1{\cdot}39$. Hence the set analyzed contains 2 latent variables, although only one was used for the sugestion of the scale of substituent constants based on this matrix[34]. Example *10* represents an analysis of matrix of logarithms of distribution coefficients of 26 solutes in 6 solvents[35]. The matrix was filled up to 94·2%. The criterion (*6*) leads to a quite univocal decision for $p = 2$. This result agrees with that found by the PCA method[35]. Another situation is encountered in Example *11*. The source matrix is formed by logarithms of rate constants of solvolyses of 26 tosyl and brosyl esters in 6 individual or mixed solvents (the mixture acetone–water was excluded because of its nonconstant composition) at 25°C and 50°C (ref.[36]). The matrix was filled up to 90·1%. The criterion (*6*) provided univocal results, because for $p = 3$ the value F is equal to 3·60 ($F_{0.95}$. .$(29, 152) = 1{\cdot}55$) and for $p = 4$ it is $F = 0{\cdot}50$ ($F_{0.95}(28, 120) = 1{\cdot}57$). Such a result agrees with the expectation, since one factor is introduced by temperature, one to two factors are due to the solvent change, and one to two are ascribable to the change of substrate structure. From this standpoint, the value $p = 2$ given in the original paper[36] seems to be too low. In the last illustration of data analysis of this type of sets (Example *12*) no decision was arrived at before exhausting the degrees of freedom. The source matrix[37,38] contained the retention data of 49 samples in 6 eluents and was filled up to 46·4%. Although the proportion of explained variability was rapidly increasing, the value of criterion (*6*) did not much change after

$p = 2$. The number of latent variables determined in another way[38] was $p = 2$, i.e. again in the break of the trend of criterion (*6*). Obviously, the situation is analogous to that of Example *8*.

If the results of testing of number of latent variables in the sets of H–NC type are summarized, it can be stated that – in comparison with the sets of H–C type – the accordance with literature data is less, but the results are reasonably adequate to composition of experimental data. As the missing data can be considered to constitute an additional error load, correct results can be expected with respect to robustness of the criterion (*6*) used. On the other hand, a discrepance with literature data can indicate – in some cases – incorrect results given in the literature, because other criteria, too, will necessarily be affected by an incomplete filling of the source matrix. As compared with the sets type H–C, the sets of H–NC type exhibit higher residual standard deviation (Table IV) due predominantly to the incompleteness of source matrix mentioned.

Analysis of NH–NC *Type Sets*

Analysis of these sets is of the highest importance from practical standpoint because of the generality of the results obtained. Among such sets, e.g., belongs the set used for the suggestion of a new nucleophilicity scale[21] (Example *13*). In contrast to the original paper, the source matrix was reduced by rejecting the columns Nos 10, 19, 39–42 which contain less than 7 data each, and the source matrix was then filled up to 45·2%. The criterion (*6*) leads to two latent variables (for $p = 2$ it is $F(31, 330) = 1{\cdot}37$) if the testing is carried out at the significance level $\alpha = 0{\cdot}05$ ($F_{0.95} = 1{\cdot}49$). At the significance level $\alpha = 0{\cdot}1$ ($F_{0.90} = 1{\cdot}36$) the hypothesis of insignificance of decrease of residual variability is closely rejected. The value of criterion (*6*) for higher p values somewhat increase, the number of latent variables exceeding the value of $p = 6$ for which the testing was finished. The heterogeneity of set and dominance make themselves felt in a typical way in the method of conjugated deviations. The nucleophilicity as a phenomenon is obviously described by two basic factors (as indicated in earlier studies[21,39]), addition of further latent variables can then lead only to interpretation of individual columns or small groups of columns of the source matrix with a small communality. Example *14* represents an illustration of another intensively studied phenomenon – the solvent effects. For the analysis we used a data matrix analyzed earlier[22,40] and rejected the columns containing small numbers of data (Nos 10, 49, 219–225, 266, 286); the matrix was filled up to 33·4%. When treating this matrix by the method of conjugated deviations we observed a new phenomenon – the number of degrees of freedom increases with increasing number of the latent variables involved until $p = 3$. The reason consists in the increasing number of the columns of source matrix for which at least one regression coefficient is significant in the regression with the latent variables, and the

respective column is then included into the statistics. All columns of the source matrix are included into the calculation only above $p = 5$. The values of criterion (*6*) monotonously decrease: $F(p = 3) = 10{\cdot}44$, $F(p = 4) = 5{\cdot}30$, $F(p = 5) = 5{\cdot}12$. The number of significant latent variables is thus higher than or equal to six, at which value the testing was stopped for practical reasons. Obviously, the solvent effects represent a phenomenon exhibiting many aspects, and determination of number of significant factors in the model (*1*) is difficult if not impossible[40]. In this situation, probably the best way is to use arbitrarily a defined number of parameters – latent variables which interpret the experimental data with sufficient accuracy. From this point of view it is interesting to mention the proportions of explained variability V which increase from $p = 1$ to $p = 6$ in the following order (in %): 67·7, 78·6, 85·4, 88·2, 90·4, 92·0. A similar trend of decrease within the same range of p is observed also for the residual standard deviations: 0·590, 0·490, 0·412, 0·376, 0·347, 0·323. With regard to the fact that the method of conjugated deviations does not afford orthogonal latent variables, the selection of 3 latent variables for description of solvent effect (suggested earlier[22]) seems to be sufficient from the practical point of view.

REFERENCES

1. Überla K.: *Faktorenanalyse.* Springer, Berlin 1971.
2. Rao R. C.: *Linear Statistical Inference and Its Applications.* Wiley, New York 1973.
3. Aivazyan S. A., Bezhaeva Z. I., Staroverov O. V.: *Klassifikaciya mnogomernych nablyudenii.* Statistika, Moscow 1974.
4. Blahuš P.: *Faktorová analýza a její zobecnění.* SNTL, Prague 1985.
5. Bolch B. W., Huang C. J.: *Multivariate Statistical Methods for Business and Economies.* Prentice-Hall, New Jersey 1974.
6. Malinowski E. R., Howery D. G.: *Factor Analysis in Chemistry.* Wiley, New York 1980.
7. Sharaf M. A., Illman D. L., Kowalski B. R.: *Chemometrics.* Wiley, New York 1986.
8. Massart D. L., Vandeginste B. G. M., Deming S. N., Michotte Y., Kaufman L. in: *Chemometrics: Data Handling in Science and Technology* (B. G. M. Vandeginste and L. Kaufman, Eds), Vol. 2. Elsevier, Amsterdam 1988.
9. Geladi P.: J. Chemometrics *2*, 231 (1988).
10. Lorber A., Wangen L. E., Kowalski B. R.: J. Chemometrics *1*, 19 (1987).
11. Stähle L., Wold S.: J. Chemometrics *1*, 185 (1987).
12. Höskuldsson A.: J. Chemometrics *2*, 211 (1988).
13. Vallis L. V., MacFie H. J., Gutteridge C. S.: Anal. Chem. *57*, 704 (1985).
14. Malinowski E. R.: Anal. Chem. *49*, 612 (1977).
15. Wold S.: Technometrics *20*, 397 (1978).
16. Eastment H. T., Krzanowski W. J.: Technometrics *24*, 73 (1982).
17. Rossi T. M., Warner I. M.: Anal. Chem. *54*, 810 (1986).
18. Malinowski E. R.: J. Chemometrics *3*, 49 (1988).
19. Malinowski E. R.: J. Chemometrics *1*, 33 (1987).
20. Cartwirght H.: J. Chemometrics *1*, 111 (1987).
21. Pytela O., Zima V.: Collect. Czech. Chem. Commun. *54*, 117 (1989).

22. Pytela O.: Collect. Czech. Chem. Commun. *54*, 136 (1989).
23. Bakytová H., Hátle J., Novák I., Urgon M.: *Statistická indukce pro ekonomy*. SNTL/ALFA, Prague 1986.
24. Sadowski W.: *Statystaka matematyczna*, 2nd ed. Panstwowe wydawnictwo ekonomiczne, Warsaw 1969.
25. Gold R. J. M., Tenenenhouse H. S., Adler L. S.: Biochem. J. *159*, 157 (1976).
26. Spjøtvoll E., Martens H., Volden R.: Technometrics *24*, 173 (1982).
27. Ritter G. L., Lowry S. R., Isenhour T. L., Wilkins C. L.: Anal. Chem. *48*, 591 (1976).
28. Weiner P. H., Malinowski E. R., Levinstone A. R.: J. Phys. Chem. *74*, 4537 (1970).
29. Abraham R. J., Wileman D. F., Bedford G. R.: J. Chem. Soc., Perkin Trans. *2*, *1973*, 1027.
30. Weiner P. H.: J. Am. Chem. Soc. *95*, 5845 (1973).
31. Ludwig M., Pytela O., Javůrková H., Večeřa M.: Collect. Czech. Chem. Commun. *52*, 2900 (1987).
32. McReynolds W. O.: J. Chromatogr. Sci. *8*, 685 (1970).
33. Ludwig M., Baron V., Kalfus K., Pytela O., Večeřa M.: Collect. Czech. Chem. Commun. *51*, 2135 (1986).
34. Pytela O., Ludwig M., Večeřa M.: Collect. Czech. Chem. Commun. *51*, 2143 (1986).
35. Dunn W. J., Wold S.: Acta Chem. Scand., B *32*, 536 (1978).
36. Albano C., Wold S.: J. Chem. Soc., Perkin Trans. 2, *1980*, 1447.
37. Snyder L. R.: *Principles of Adsorption Chromatography*. Dekker, New York 1968.
38. de Ligny C. L., Nieuwdrop H. G. E., Brederode W. K., Hammers E. W., van Honwelingen J. C.: Technometrics *23*, 91 (1981).
39. Zima V., Pytela O., Večeřa M.: Collect. Czech. Chem. Commun. *53*, 814 (1988).
40. Pytela O.: Collect. Czech. Chem. Commun. *53*, 1333 (1988).

Translated by J. Panchartek.