CROATICA CHEMICA ACTA CCACAA **69** (3) 1175–1188 (1996)

ISSN-0011-1643
CCA–2367
Original Scientific Paper

A Correlation Rank Determinator for Principal Components

Damir Pavković, Vladislav Tomišić, Revik Nuss, and Vladimir Simeon

Laboratory of Physical Chemistry, Faculty of Science, University of Zagreb,
10000 Zagreb, P.O. Box 163, Croatia

Received Juli 14, 1995; revised November 14, 1995; accepted November 27, 1995

A new statistic, defined as $R_0(k) = s(k)/\sqrt{s_1^2 + \ldots + s_b^2}$ {*s(k)* denoting the k-th singular value of data matrix}, is proposed for assessing the significance of individual components in principal-components analysis (PCA). R_0 was shown to be the correlation coefficient of the prediction from the k-th principal component to the original data. The common significance test on R_0 was applied as a semi-empirical determinator of the effective rank (»pseudorank«) of data matrices. By examining the performance of this simple test on R_0 on a number of data matrices of known effective ranks (UV/Vis and Raman spectra of aqueous solutions of various inorganic salts), it was shown to be a serious competitor to the rank determinators commonly used in PCA.

INTRODUCTION

Many empirical[1–4] and statistical[5–9] criteria have been used in principal components analysis (PCA) for assessing the number of significant components, the so-called *effective rank* (called also *pseudorank* or *chemical rank*), *i.e.* the rank the data matrix would have in the absence of any random measurement errors. Although the problem of effective rank assessment, the central task of principal components analysis, is very unlikely to have a general solution, a straightforward and robust criterion can still be useful, especially in those cases where PCA is the last step in the data analysis so that PCA results cannot be checked in subsequent target or evolving factor analyses. In the present paper, we describe a new semi-empirical determinator of the effective rank, based upon the correlation of predictions from individual principal components to the original data. This rank determinator is applicable in the frequently encountered case when the data variability is not known. In the present, very initial stage, it was natural to focus the attention mainly (though not exclusively) on the so-called low-rank problems.

THEORETICAL

Model

Let $\boldsymbol{X}$ denote an $l \times b$ $(l > b)$ data matrix whose elements are sampled from $l \times b$ independent, homoscedastic unimodal distributions. In other words, it will be assumed that every necessary data pretreatment (such as centering, standardization, weighting[10,11] or alike) has already been done. A statistical model, appropriate in many real situations (UV/Vis, Raman and NMR spectra, GC, MS, HPLC, ...) can now be formulated in the following way:

$$\boldsymbol{X} = \boldsymbol{X}^* + \boldsymbol{E} = \boldsymbol{P}^* \cdot \boldsymbol{C}^* + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{E}$ is the matrix of random errors with zero expectance and constant variance; $\boldsymbol{X}^*$ denotes the expectance of $\boldsymbol{X}$, which is thought of as being the product of two factors, *spectral profile matrix*, $\boldsymbol{P}^*$, and *concentration matrix*, $\boldsymbol{C}^*$ (the terminology has been adapted to the spectrometric context of this paper). The dimensions of these two matrices are $(l \times r)$ and $(r \times b)$, respectively, and their rank is $r \leq b$, in contrast to $\boldsymbol{X}$ which is usually full-rank (generally, $r \leq \text{rank } \boldsymbol{X} \leq b$). In most real situations, the *effective rank* (*pseudorank*) of $\boldsymbol{X}$ matrix, r, is unknown and has to be inferred from the data; this is one of main objectives of the *analysis into prinicipal components* (PCA).

Principal components

Singular value decomposition (SVD)[12] of $\boldsymbol{X}$ yields

$$\boldsymbol{X} = \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}^{\mathrm{T}} = \tilde{\boldsymbol{P}} \cdot \boldsymbol{V}^{\mathrm{T}} = \tilde{\boldsymbol{P}} \cdot \tilde{\boldsymbol{C}} \ . \tag{2}$$

Matrices $\boldsymbol{U}$ and $\boldsymbol{V}^{\mathrm{T}}$ are orthonormal. Matrices $\tilde{\boldsymbol{P}}$ and $\tilde{\boldsymbol{C}}$ will be called *abstract spectral profiles* and *abstract concentrations*, respectively. The diagonal $b \times b$ matrix $\boldsymbol{S}$ contains *singular values* in a descending order:

$$\boldsymbol{S} = \mathrm{diag}(s_1 \geq s_2 \geq \ldots \geq s_b) \ . \tag{3}$$

Singular values are positive square roots of the eigenvalues,

$$s_k = \sqrt{\lambda_k} \ , \tag{4}$$

of the matrix of crude second moments (*dispersion matrix*),

$$\boldsymbol{D} = \boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{X} \, . \tag{5}$$

Matrix $\boldsymbol{V}$ contains eigenvectors of $\boldsymbol{D}$ (known also as right singular vectors of $\boldsymbol{X}$). From the dyadic decomposition theorem,[12]

$$\boldsymbol{X} = \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{b} \boldsymbol{u}_i \cdot \boldsymbol{s}_i \cdot \boldsymbol{v}_i^{\mathrm{T}} \ , \tag{6}$$

it can be concluded that each eigenvector is the predictor of an additive part of $\boldsymbol{X}$, such that, for k = 1, ..., b:

$$\tilde{\boldsymbol{X}}_k = \boldsymbol{u}_k \cdot s_k \cdot \boldsymbol{v}_k^{\mathrm{T}} = \boldsymbol{X} \cdot (\boldsymbol{v}_k \cdot \boldsymbol{v}_k^{\mathrm{T}}) \ . \tag{7}$$

Partial predictions, $\tilde{\boldsymbol{X}}_k$, defined in Eq. (7) are mutually orthogonal, *i.e.* statistically independent.

First r eigenvectors, termed here *structural components*, contain all information on the data structure whereas the remaining $b - r$ ones (*residual components*) contain nothing but the experimental error. Therefore, structural components can be used to smooth the original data:

$$\tilde{\boldsymbol{X}}_{1\ldots r} = \sum_{k=1}^{r} \tilde{\boldsymbol{X}}_k \ \ . \tag{8}$$

The same result is obtained by deleting the residual components from $\boldsymbol{V}$ matrix, as well as the corresponding columns from $\boldsymbol{U}$ matrix and the corresponding singular values from $\boldsymbol{S}$ matrix:

$$\tilde{\boldsymbol{X}}_{1\ldots r} = \tilde{\boldsymbol{U}}_{1\ldots r} \cdot \tilde{\boldsymbol{S}}_{1\ldots r} \cdot \tilde{\boldsymbol{V}}_{1\ldots r}^{\mathrm{T}} = \tilde{\boldsymbol{P}}_{1\ldots r} \cdot \tilde{\boldsymbol{V}}_{1\ldots r}^{\mathrm{T}} = \tilde{\boldsymbol{P}}_{1\ldots r} \cdot \tilde{\boldsymbol{C}}_{1\ldots r} \tag{9}$$

where dim $\tilde{\boldsymbol{U}} = l \times r$, dim $\tilde{\boldsymbol{S}} = r \times r$, dim $\tilde{\boldsymbol{V}} = b \times r$. Thus, Eq. (9) defines an estimator of $\boldsymbol{X}^*$ as well as estimators of $\boldsymbol{P}^*$ and $\boldsymbol{C}^*$, apart from the (unknown) $r \times r$ matrix defining the rotation of the SVD coordinate frame back to the »natural« coordinate frame defined by the (unknown) factors of $\boldsymbol{X}^*$, *i. e.* $\boldsymbol{P}^*$ and $\boldsymbol{C}^*$ (*cf.* Eq. 1).

Effective rank determinators

Eigenvalues. The effective rank determinators based upon the eigenvalues of the dispersion matrix,[1–5,8] besides being very economical from the computational viewpoint, have an important theoretical property that each eigenvalue is an estimator of the sum of squares (of the deviations from the origin) generated by the respective eigenvector. Nevertheless, it proved to be rather difficult to devise theoretically sound criteria for discriminating the structural eigenvalues from those generated by the residual components. Quite generally, these determinators are greatly influenced by the shape of the »data ellipsoid«, *i.e.* by relative magnitudes of the eigenvalues, that is the data structure.

Predicted values. An alternative approach to the problem of assessing the effective rank is the analysis of partial Eq. (7) or cumulative Eq. (8) predictions. Such an approach is computationally somewhat more expensive than the eigenvalue-based rank determinators but still acceptable, even with a modest personal computer. Several methods based upon the 'prediction sum of squares' (PRESS) have been proposed,[13] all of them requiring rather complex and lengthy computations. Recently, Tomišić and Simeon[9] used, with some success, the Kolmogorov-Smirnov (KS) distribution test to compare successive cumulative predictions to the original data matrix, although the important condition of statistical independence of distributions to be compared was not met.

Correlation and regression coefficients. Let us consider a standardized measure of the proximity of data predictions, $\tilde{x}_{ij}(k)$, to the original data, x_{ij}, *viz.* the raw product-moment correlation coefficient, $R_0(k)$:

$$R_0(k) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{b} x_{ij}\,\tilde{x}_{ij}\,(k)}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{b} x_{ij}^2}\,\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{b} [\tilde{x}_{ij}(k)]^2}}. \qquad (10)$$

Provided the measurement errors are not too large, one can expect (in view of Eq. (8)) each element data of the data matrix to contain r significant contributions generated by the structural components, $\tilde{x}_{ij}(k)$ $(1 \le k \le r)$, *plus* $b - r$ distinctly smaller additive contributions from residual components $(r + 1 \le k \le b)$. Therefore, the first r partial predictions can be expected to be better correlated to the data than the residual partial predictions $(r + 1 \le k \le b)$, which should be nearly orthogonal to the original data.

Since each eigenvalue equals the sum of squared deviations from zero, predicted by the respective eigenvector the fraction of the over-all data variation due to the k-th principal component is given by the $R_0^2(k)$ statistic which is completely analogous to the coefficient of determination in regression analysis:

$$R_0^2(k) = \frac{\lambda_k}{\sum\limits_{j=1}^{b} \lambda_j} = \frac{s_k^2}{\sum\limits_{j=1}^{b} s_j^2} \ . \tag{11}$$

The algebraical equivalence of (squared) Eq. (10) and Eq. (11) can easily be proved by inserting the $\tilde{x}_{ij}(k)$ values, as given by Eq. (8), into squared Eq. (10) and making use of the mutual orthogonality of individual predictions to get Eq. (11). It can also be shown, in a similar way, that $R_0^2(k)$ is equal to the slope of the k-th prediction regression *vs.* original data:

$$b_0(k) = R_0^2(k) \ . \tag{12}$$

Correlation tests. As already said, a good correlation of the predictions from a component to the data indicates an appreciable presence of that component (*cf.* Eq. (11)) while a prediction that is almost orthogonal to the data can hardly be considered as anything else but noise. Therefore, the significance test of the $R_0(k)$ coefficient could be a potential determinator of the effective rank, having a practical advantage of being computationally economical (it requires very little additional computation since there is no need to compute the predictions, unless they are required for some other purpose).

Unfortunately, it is not easy to devise a theoretically sound test on R_0, for the following reasons: (*i*) Error distribution of the data is not known in the general case (although it is frequently not far from normality); (*ii*) the expectance of R_0 is known to be greater than zero, for every component, but in view of (*i*) and (*ii*) may be difficult to assess.

In default of a straightforward theoretical solution, an empirical alternative has to be adopted. For example, R_0 can be subjected to one of the coventional correlation tests[14] (single-sided normal, Student's t or Fisher's z test) although the assumptions for any of these procedures are certainly not fulfilled (neither the partial predictions are sampled from a bivariate normal population nor the null hypothesis, $\mathbf{H}_0$: $\rho = 0$, is true, *etc.*). The results obtained by applying such a semi-empirical rank determinator to several sets of spectrometric data will be described in the following parts of this paper.

EXPERIMENTAL

Experimental design

Spectral experiments were designed in such a way that, almost in all instances, the number of solutions was considerably smaller than the number of wavelengths,

TABLE I

Composition and spectral range of the examined solution sets

System	c/(mol dm^{-3})	λ/nm
Methyl Orange[9] [(a)]	$2.55 \cdot 10^{-5}$	350...600
Methyl Red[9] [(b)]	$3.09 \cdot 10^{-5}$	350...600
Methyl Red[15] [(c)]	unspecified	unspecified
$Mg(NO_3)_2$	0.02...2	250...350
$Co(NO_3)_2$	0.02...4.25	252...600
$K_2Cr_2O_7$	$10^{-5}...10^{-3}$	249...650
$KMnO_4$	$10^{-5}...10^{-3}$	249...650
KNO_3	$2 \cdot (10^{-5}...10^{-3})$	249...650
$MgSO_4$ [(d)]	0.02...2.6	100...1800
$K_2Cr_2O_7$ + KNO_3		249...650
$K_2Cr_2O_7$ + KNO_3 + $KMnO_4$		249...650

[(a)] $2.96 \le pH \le 9.50$.

[(b)] $4.08 \le pH \le 7.72$.

[(c)] pH range unspecified.

[(d)] Raman spectra; wavenumber/cm^{-1} is quoted.

a frequent case in spectrometric work. The examined aqueous solutions differed in kind and/or concentration(s) of the solute(s), their compositions and, consequently, the effective ranks of data matrices (*i.e.* the number of chemical components in each set) being precisely known. Concentration ranges of the solutions and the wavelength (wavenumber) ranges of their spectra are shown in Table I. Since many rank determinators are known to fail just in those cases where the effective rank is expected te be 1, many of the examined data sets (33 out of 49) contained only one solute ($K_2Cr_2O_7$ or KNO_3 or $KMnO_4$ or $Mg(NO_3)_2$ or $Co(NO_3)_2$); 8 of these 33 unisolute data sets contained repeated measurements of the same ($K_2Cr_2O_7$) spectrum (number of repeats ranging from 3 to 30). The remaining 16 data sets were Raman spectra of $MgSO_4$ solutions and UV/Vis spectra of solutions containing varying concentrations of either two ($K_2Cr_2O_7$ + KNO_3) or three ($K_2Cr_2O_7$ + KNO_3 + $KMnO_4$) solutes. Some data sets were »trimmed« by omitting wavelength ranges with near-zero absorbances and/or some solutions.

Materials and methods

All chemicals used were of analytical reagent purity grade and were not further purified. Water was distilled twice, in an all-glass still. Absorbances in the UV/Vis spectral range were measured by means of a Varian Cary 5 spectrometer to ±0.001 or better (including base-line instability), at constant temperature of (25.00 ± 0.05) °C, sampling interval 1 nm. Raman spectra were recorded by means of a Spex 1401 spectrometer at room temperature (approx. 22 °C). Most of the computations were done using own software (numerical precision 16 decimal digits or better); Statgraphics® (version 4.0) was used for the rest.

Along with the single-sided normal test on the $R_0\sqrt{lb-1}$ variate, the effective ranks were assessed by means of two additional criteria frequently used in chemometric applications of PCA, *viz.* Malinowski's IND function[2] and F test,[8] as well as the Kolmogorov-Smirnov test on cumulative predictions.[9]

RESULTS

In order to give an idea of the shapes of the examined spectra, UV/Vis absorption spectra of three unisolute solutions are displayed in Figure 1 and the Raman spectra of a series of aqueous magnesium sulphate solutions are shown in Figure 2.

The results obtained with 33 unisolute data sets are summarized in Figure 3 where the frequency distribution of error in the effective rank assessment is shown for all four rank determinators examined. The analogous plot for 16 multisolute data sets is given in Figure 4. Both error distributions indicate that the proposed R_0 rank determinator – though not infallible – might be a serious competitor to the other three examined criteria.

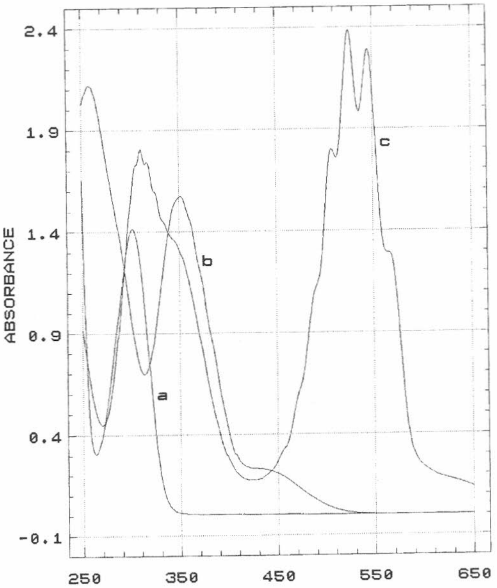

Figure 1. UV/Vis absorption spectra of a series of aqueous solutions of (**a**) KNO_3, (**b**) $K_2Cr_2O_7$ and (**c**) $KMnO_4$.

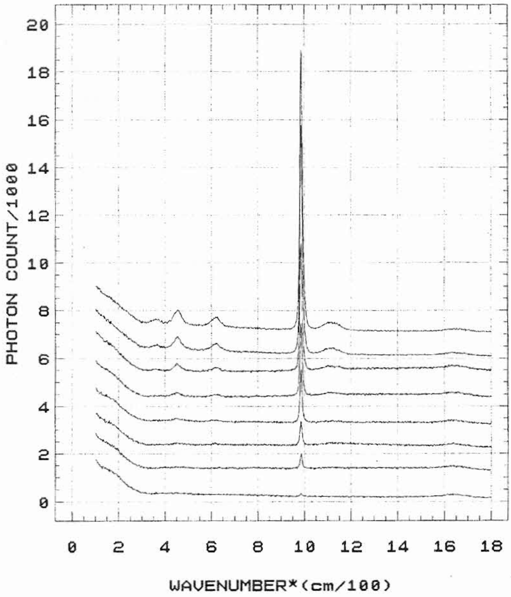

Figure 2. Raman spectra of a series of aqueous $MgSO_4$ solutions.

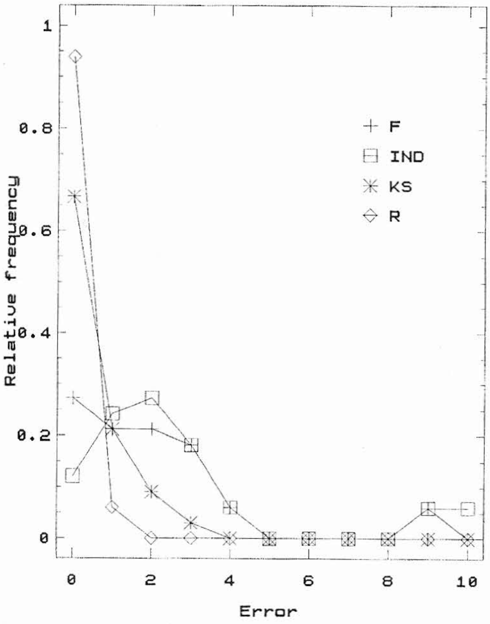

Figure 3. Distribution of error in effective rank assessment (true rank = 1).

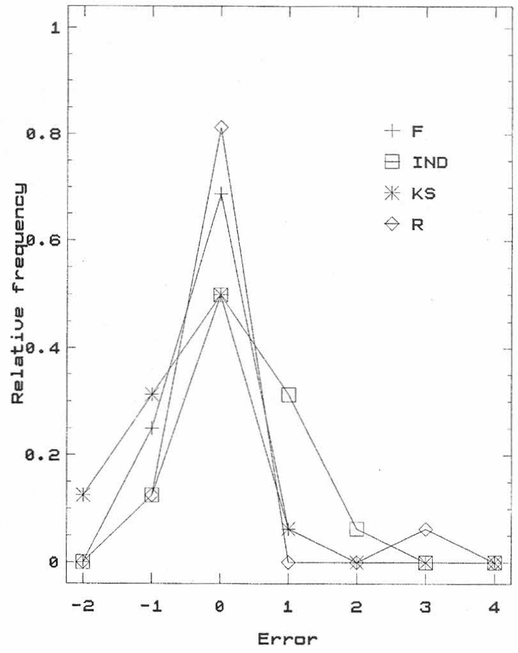

Figure 4. Distribution of error in effective rank assessment (true rank > 1).

DISCUSSION

Unisolute data sets

The performance of rank determinators on solution sets containing only one solute is seldom scrutinized. Most probably, this case has been considered trivial although the data displayed in Figure 3 convincingly demonstrate that it is all but trivial:

(*a*) IND and *F* criteria, when applied to repeated measurements on one single $K_2Cr_2O_7$ solution, tended to overestimate the effective rank, especially for larger data matrices. For the largest data matrix (dimension 402 × 30), the rank estimate amounted to 10 (*F*) or 11 (IND)! In contrast, KS and R_0 invariably indicated the correct rank of 1, the difference in significance levels of the first and second correlations being at least 0.9.

(*b*) As seen from Figure 3, each of the examined rank determinators occasionally failed with data sets containing spectra of unisolute solutions of varied concentration; R_0 failed only with two sets of $KMnO_4$ solutions. Some additional information (*viz.* eigenvalues, correlations and significance levels) for one of these data sets (all 7 solutions, full wavelength range) can be

TABLE II

Statistical parameters for selected data matrices (dimension: $l \times b$): eigenvalues $\{\lambda(k)\}$, correlation coefficients $\{R_0(k)\}$, and significance levels; last significant ($P \leq 0.075$) values are underlined (twice if chemically correct); results of IND, F, KS and R_0 procedures are summarized at the bottom.

k	$KMnO_4$	$KMnO_4$	$K_2Cr_2O_7$ + KNO_3	$K_2Cr_3O_7$ + KNO_3	$K_2Cr_2O_7$ + KNO_3 + $KMnO_4$	$MgSO_4$ (Raman)
	(402 × 7)	(402 × 6)	(402 × 5)	(272 × 4)	(402 × 6)	(851 × 16)
1	4.167E+2 0.999270 0.000000	4.166E+2 0.999588 $\underline{\underline{0.000000}}$	1.147E+2 0.998843 0.000000	5.967E+1 0.999103 0.000000	8.680E+1 0.987391 0.000000	5.448E+9 0.937049 0.000000
2	5.978E–1 0.037849 $\underline{0.0224}$	3.361E–1 0.028392 0.0816	2.424E–1 0.045907 $\underline{\underline{0.0198}}$	8.870E–2 0.038519 $\underline{\underline{0.0102}}$	1.884 0.145452 0.000000	6.565E+8 0.325279 1.7E–315
3	6.179E–3 3.848E–3 0.419	6.583E–3 3.973E–3 0.423	2.161E–2 0.013706 0.269	1.779E–2 0.017249 0.285	3.171E–1 0.059678 $\underline{\underline{0.0017}}$	9.114E+7 0.121203 $\underline{\underline{1.04E–45}}$
4	4.340E–3 3.225E–3 0.432	4.195E–4 1.003E–3 0.480	1.293E–3 0.003352 0.441	6.795E–4 0.003371 0.456	2.701E–2 0.017417 0.196	2.260E+6 0.019087 0.0130
5	2.269E–4 7.37E–4 0.484	1.056E–4 5.03E–5 0.490	5.854E–4 0.002256 0.460		2.650E–3 0.005456 0.394	1.606E+6 0.018088 0.0302
6	9.897E–5 4.87E–4 0.490	2.924E–5 2.65E–4 0.495			7.884E–4 0.002976 0.442	1.038E+6 0.012936 $\underline{0.0656}$
7	2.848E–5 2.44E–4 0.495					6.421E+5 0.010173 0.118
...						...
16						2.643E+5 0.006527 0.223
IND	4	3	3	1	4	$\underline{\underline{3}}$
F	4	3	$\underline{\underline{2}}$	1	$\underline{\underline{3}}$	$\underline{\underline{3}}$
KS	2	2	$\underline{\underline{2}}$	1	$\underline{\underline{3}}$	$\underline{\underline{3}}$
R_0	2	$\underline{\underline{1}}$	$\underline{\underline{2}}$	$\underline{\underline{2}}$	$\underline{\underline{3}}$	$\underline{\underline{3}}$ or 6

found in the first column of Table II. R_0 indicated the effective rank of 2 whereas IND and F were in error by as much as 4. Upon omitting one of the solutions, the significance level of $R_0(2)$ changed from 0.02 to 0.08 (see the second column in Table II). The other set (not reproduced in Table II), where all of the examined rank determinators failed, contained the data for 5 solutions, including the »critical« one which was seen to be responsible for blowing up the significance of $R_0(2)$. It is possible that a chemical change of permanganate solution (*e.g.* formation of solid MnO_2) in the »critical« solution was responsible for a too high effective rank estimate. The question of the decision probability level will be discussed later in this section.

(*c*) For the formally unisolute cobalt(II) nitrate solution set, an effective rank of 2 was indicated (not shown in Table II). This was explained by assuming the formation of $[Co(NO)_3]^+$ associate at higher concentrations affecting the UV absorption band of nitrate ion but having no influence on the cobalt(II) ligand-field band in the visible range.

Multisolute data sets

The performance of the R_0 rank determinator with the systems containing two or more chemical components can be seen from the error distribution depicted in Figure 4. Although generally satisfactory, it has failed in three instances. One of these was an incomplete set of $K_2Cr_2O_7$ + KNO_3 solutions where the effective rank was underestimated (indicated: 1, expected: 2), the significance level of $R_0(2)$ being borderline (0.10). The other three rank determinators also failed in this instance.

For the well-known Wallace and Katz[15] data on the absorption spectra of Methyl Red (4'-dimethylaminoazobenzene-2-carboxylic acid), R_0 indicated the effective rank of 2 although the accurate value might be 3 as well (see the »scree plot«[3] of $R_0(k)$ values in Fig. 5). IND function and KS test indicated chemically impossible effective ranks of 4 and 1, respectively. F test gave an ambiguous answer, the significance levels of the third and fourth components being 0.017 and 0.086, respectively, thus pointing to the effective rank of at least 3, possibly 4. It is questionable, however, whether Wallace and Katz data are a suitable benchmark set, having in view their limited size (8 × 8) and modest absorbance measurement accuracy (estimated[6] to be around ±0.003). Moreover, the authors did not specify the pH range of their solutions, in spite of its importance in the case of ionization of this *dibasic* acid. Our data for Methyl Red,[9] measured in an appropriate pH range, were correctly indicated to have the rank of 2.

The effective rank of Raman data ($MgSO_4$ solutions) was expected to be 3, because of the contributions of two chemical species (SO_4^{-2} ion and $[Mg^{+2}SO_4^{-2}]$ ion pair), superposed onto a clearly curved base-line giving rise to a third significant component. All examined rank determinators indicated the correct rank, except for R_0 which overestimated it by as much as 3 units (see the last column in Table II). This serious failure of the proposed rank

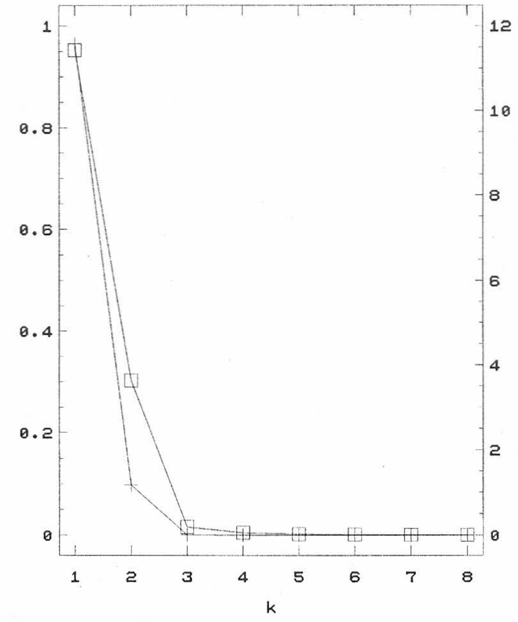

Figure 5. »Scree plots«[3] for Wallace and Katz [15] data; left ordinate axis: $R_0(k)$ values (□), right ordinate axis: eigenvalues, $\lambda(k)$ (+).

determinator can hardly be conclusively explained without more detailed study. Possible causes may have been a less favourable signal/noise ratio in the Raman data, as well as non-normal (Poisson) error distribution. However, the pattern of significance levels (Table II) deserves to be pointed out. P values for the first three R_0's were extremely small (less than 10^{-45}), in contrast to the next three (formally still significant) P values ranging from 0.013 to 0.065. By comparing successive correlation coefficients (using Fisher's z test for this purpose) the $R_0(3) - R_0(4)$ difference was found to be highly significant ($P < 0.001$), contrary to the next one $\{R_0(4) - R_0(5)\}$ where $P = 0.40$.

In the trisolute case ($K_2Cr_2O_7$ + KNO_3 + $KMnO_4$ – see the penultimate column in Table II), three rank determinators (F, KS and R_0) were successful whereas IND overestimated the effective rank by 1.

CONCLUSIONS

Although the performance of the proposed statistic was examined on a limited amount of experimental data, the available information seems to qualify the R_0 statistic as a serious competitor to the existing effective rank determinators, despite its semi-empirical character. Even in those few instances where the effective rank was not correctly estimated, the error was within ±1, which is tolerable in most chemometric applications. The question of the appropriate decision probability level, however, can hardly be given a definite answer because of the limited experience with the new statistic. Nevertheless, it seems that, in most cases, the range $0.05 \leq P \leq 0.1$ is the borderline zone. Therefore, in analyzing the present data, the mean of this range (0.075) was taken as the critical P value. This provisional decision level should not be observed too rigidly, without recourse to the chemical common sense or intuition. Also, it will still be wise to analyze the results with other rank determinators, to examine the P *versus* k pattern (*cf.* the example of Raman data discussed above) and, last but not least, to perform target or evolving factor analyses whenever possible.

Acknowledgement. – We thank Prof. M. Schwuger and Dr. E. Koglin of the Institute for Applied Physical Chemistry in Jülich, Germany, who gave us the opportunity to use the Raman spectrometer. Financial support by the Ministry of Science and Technology of the Republic of Croatia is gratefully acknowledged.

REFERENCES

1. H. F. Kaiser, *Brit. J. Statist. Psychol.* **14** (1961) 1.
2. E. R. Malinowski, *Anal. Chem.* **49** (1977) 612.
3. R. B. Cattell, *Multivariate Behav. Res.* **1** (1966) 245.
4. J. L. Horn, *Psychometrika* **30** (1965) 179.
5. M. S. Bartlett, *Brit. J. Psychol. Statist. Sect.* **3** (1950) 77; D. N. Lawley, *Biometrika* **43** (1956) 128.
6. Z. Z. Hugus, Jr. and A. A. El-Awady, *J. Phys. Chem.* **75** (1971) 2954.
7. R. N. Carey, S. Wold, and J. O. Westgard, *Anal. Chem.* **47** (1975) 1824; S. Wold, *Technometrics* (1978) 397.
8. E. R. Malinowski, *J. Chemometrics* **3** (1988) 49; *ibid.* **4** (1990) 102.
9. V. Tomišić and Vl. Simeon, *J. Chemometrics* **7** (1993) 381.
10. R. N. Cochran and F. H. Horne, *Anal. Chem.* **49** (1977) 846.
11. Vl. Simeon and D. Pavković, *J. Chemometrics* **6** (1992) 257.
12. S. van Huffel and J. Vandewalle, *The Total Least Squares Problem*, SIAM, Philadelphia 1991, pp. 29–32.
13. E. R. Malinowski, *Factor Analysis in Chemistry*, 2d. ed., J. Wiley, New York 1991, pp. 117–119, 144–145 and references therein.
14. B. E. Cooper, *Statistics for Experimentalists*, Pergamon, Oxford 1969, pp. 206–212.
15. R. M. Wallace and S. M. Katz, *J. Phys. Chem.* **68** (1964) 3890.

SAŽETAK

Korelacijski determinator ranga u metodi glavnih komponenata

Damir Pavković, Vladislav Tomišić, Revik Nuss i Vladimir Simeon

Predložen je novi statistik, $R_0(k) = s(k) / \sqrt{s_1^2 + \ldots + s_b^2}$, za utvrđivanje statističke značajnosti pojedinih glavnih komponenata realne podatkovne matrice ($s(k)$ označuje k-tu singularnu vrijednost podatkovne matrice). Pokazano je da je $R_0(k)$ Pearsonov koeficijent korelacije između izvornih podataka i predikcijâ izračunanih iz glavne komponente. Obična parametarska kušnja značajnosti R_0 može poslužiti kao determinator efektivnog ranga (pseudoranga) podatkovne matrice. Provjera te kušnje na većem broju podatkovnih matrica (UV/Vis i Ramanski spektri vodenih otopina nekolikih anorganskih soli) pokazala je da je ona ozbiljan takmac determinatorima ranga koji se najčešće rabe u analizi glavnih komponenata (PCA).