CRITICAL COMPARISON OF VARIOUS FACTOR ANALYSIS METHODS IN SPECTRA ANALYSIS DETERMINING THE NUMBER OF LIGHT-ABSORBING SPECIES

Milan Meloun, Petr Mikšík a Karel Kupka*
Katedra analytické chemie,
Univerzita Pardubice
532 10 Pardubice,
*TriloByte Statistical Software Ltd., 530 01 Pardubice
Česká republika

Keywords: *principal component analysis (PCA), decomposition of absorbance matrix, factor analysis, spectra analysis (FA), rank of matrix, the number of species, Kankare method, methods of real error, extracted error, mean error and method of logarithm of eigenvalues, instrumental error of spectrophotometer*

Introduction

Principal component analysis (PCA) performs a decomposition of absorbance matrix, i.e. from the many methods of multivariate data display only those that are considered which are based on the extraction of factors. Factor data analysis (FA) also focuses on the reduction of dimensionality of data. Consider, the absorbance matrix $\boldsymbol{A}$ is of dimension $m \times n$, where m is the number of rows, i. e. solutions for which all spectra have been measured and n is the number of wavelengths at which each spectrum have been monitored.

Application of PCA means a decomposition of *the source absorbance matrix* $\boldsymbol{A}$ into a product of two matrices $\boldsymbol{T}$ and $\boldsymbol{P}^T$ and *the matrix of undescribed variability* $\boldsymbol{E}$ according to

$$\boldsymbol{A} = \boldsymbol{T}\,\boldsymbol{P}^T + \boldsymbol{E} \ .$$

The matrix $\boldsymbol{T}$ of dimension $M \times k$ is called *the matrix of latent variables* and contains k columns vectors i. e. the main components. The matrix $\boldsymbol{P}$ of dimension $N \times k$ is called *the matrix of loadings* and individual columns vectors represent a measure of contribution of a particular latent variable for description of variability of columns of source absorbance matrix. One way of calculation of matrices $\boldsymbol{T}$ and $\boldsymbol{P}$ is based on decomposition of covariance matrix $\boldsymbol{Z}$ being defined by $\boldsymbol{Z} = \boldsymbol{A}^T\boldsymbol{A}$. Decomposition is performed by diagonalization of the matrix $\boldsymbol{Z}$. Eigenvalues g_a express a variance of corresponding latent variable and are a measure of the isolated information (variability) and the matrix $\boldsymbol{Q}$ and their corresponding *eigenvectors* $\boldsymbol{q}_a$ being equal to the matrix $\boldsymbol{P}$. The matrix $\boldsymbol{T}$ is possible to calculate using relation

$$\boldsymbol{T}^T = (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{A}^T$$

Data reduction can therefore be achieved by reducing the dimensionality of the data space by removing the error associated with the absorbance data. This error is a mixture of experimental error and random deviation from the fitted model. Various techniques have been developed to identify the true dimensionality of the data. These techniques can be classified into two categories as follows:

(a) Methods based upon a knowledge of the experimental error of the data.

(b) Approximate methods requiring no knowledge of the experimental error of the data.

Theoretical

Tested example: Simulated absorbance matrix for 10 solution and 31 wavelengths concerns acid-base equilibria of Bromocresol Green (c_{BKZ} = 5.44 – 5.38×10^{-4} mol.l^{-1}) with two species L$^-$ a HL in range of pH pH = 6.49 – 3.49. The sample of random errors with the standard deviation $s_{inst}(A)$ = 0.0006 is added to absorbance data. Values of x-axis correspond to the number of light-absorbing species while the values of y-axis to the residual standard deviation $s_k(A)$. Horizontal line denotes supposed the instrumental standard deviation of absorbance of spectrophotometer used, $s_{inst}(A)$. The value $s_k(A)$ approximates best $s_{inst}(A)$ for k = 2. It is obvious that greater value k does not bring any significant decreasing a value $s_k(A)$). Analogically, the absorbance data set of this example for all following methods are used.

A. Methods based on a knowledge of the instrumental error of the data

1. Kankare method

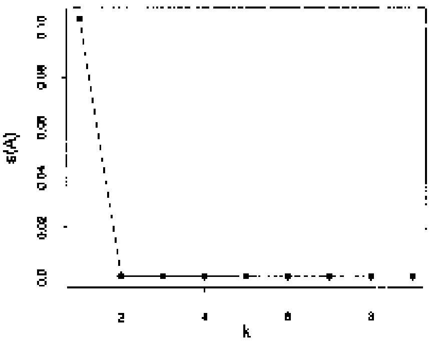

Fig. 1 ***Kankare method***. Horizontal line denotes value of the instrumental error. $s_{inst}(A)$. Best approximation of $s_{inst}(A)$ for k = 2. Higher value of k does not bring any significant decreasing of $s_k(A)$

Kankare method[2] uses the second moment M of an absorbance matrix A, i. e. $M = \frac{1}{M} A^T A$. Applying eigenvalues r_a of matrix M *the residual standard deviation of absorbance* $s_k(A)$ is estimated

$$s_k(A) = \sqrt{\frac{tr(M) - \sum_{a=1}^{k} r_a}{N-k}}$$

where $tr(M)$ is a trace of the matrix M and k is the estimated number of species in solution. The values $s_k(A)$ for different number of supposed species k are plotted against an integer k, $s_k(A) = f(k)$ and a searched number of light-absorbing species is such value k for which $s_k(A)$ is close to the instrumental standard deviation of absorbance $s_{inst}(A)$ for spectrophotometer used.

2. The method of real error

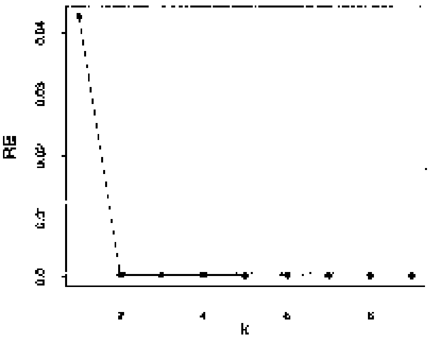

Fig. 2 ***The method of real error.*** Horizontal line denotes value of the instrumental error, $s_{inst}(A)$. Best approximation of $s_{inst}(A)$ for k = 2

The residual standard deviation RE of the second moment of an absorbance matrix or the real error can be used to identify the true dimensionality of a spectra data set. The RE is a measure of the lack of fit of a PC model to a data set and is calculated by

$$RE = \sqrt{\frac{\sum_{a=k+1}^{M} g_a}{N(M-k)}}$$

if the PCA is performed *via* the covariance matrix; where g_a is the eigenvalue associated with the kth PC dimension, N is the number of samples, M is the number of variables and k is the PC dimension (i. e. the number of latent variables) being scrutinized. The true dimensionality of a data set k is the number of dimensions required to reduce the RE to be approximately equal to the estimated experimental error of the absorbance data. The *RE* may be plotted against *k*, $RE = f(k)$, and when the *RE* reaches the value of the instrumental error of spectrophotometer used, $s_{inst}(A)$, the corresponding *k* represents the number of light-absorbing species in a mixture.

3. THE METHOD EXTRACTED ERROR

The root mean square error *XE* of an absorbance data matrix is a measure of the difference between the raw data and the data after reconstruction in the short cycle using the first *k* principal components. The *XE* is defined by

$$XE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{ij}-\hat{A}_{ij})^2}{NM}}, \text{ where } \hat{A}_{ij} = \sum_{a=1}^{k} t_{ia}\, p_{aj}, \text{ and}$$

scores are denoted by t_{ia} and loadings by p_{aj}. The alternative way of expressing *XE* is as follows

$$XE = \sqrt{\frac{\sum_{a=k+1}^{M} g_a}{NM}},$$

where g_a are eigenvalues of a covariance matrix ***Z*** and *k* is estimated number of significant latent variables. Analogically as in previous method the estimates *XE* may be plotted as a function of latent variables, $XE = f(k)$, and on base of a comparison with the magnitude of an experimental error of the spectrophotometer used the number of the light-absorbing species may be estimated. Comparing relations for *RE* and *XE*, and simplifying yields we get $XE = \sqrt{\frac{M-N}{M}}\,RE$. Although related, the *XE* and *RE* of a data set measure different sources of error. The *XE* measures the difference between raw data and reproduced data using *k* PC dimensions. The *RE* however measures the difference between the raw and the pure data containing no experimental error.

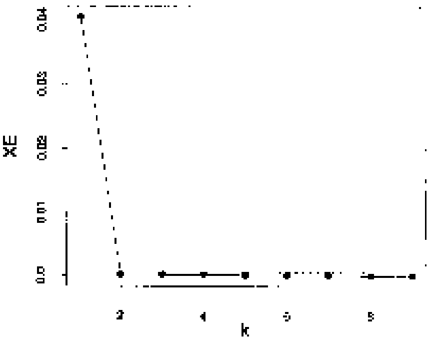

Fig. 3 ***Method of the extracted error.*** Horizontal line denotes value of the instrumental error, $s_{inst}(A)$. Best approximation of $s_{inst}(A)$ for $k = 2$

4. THE AVERAGE ERROR CRITERION

The average error of absorbance $\bar{e}$ is simply the average of the absolute values of the differences between the raw and reproduced data,

$$\bar{e} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left|A_{ij}-\hat{A}_{ij}\right|}{NM},$$

where A_{ij} a $\hat{A}_{ij}$ were described previously. The true dimensionality of absorbance data matrix is the number of dimensions required to reduce the average error to be approximately equal to the estimated average error of the data.

Values of the average error are plotted against the number of latent variables *k* and compared with the instrumental error of spectrophotometer used, $s_{inst}(A)$. When $\bar{e}$ reaches $s_{inst}(A)$, corresponding *k* estimates the number of light-absorbing species in solution.

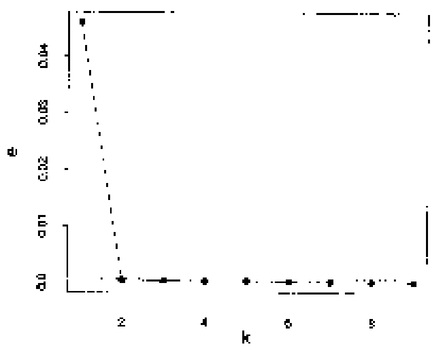

Fig. 4 ***Method of the average error.*** Horizontal line denotes value of the instrumental error, $s_{inst}(A)$. Best approximation of $s_{inst}(A)$ for $k = 2$

5. The Method of χ^2 criterion

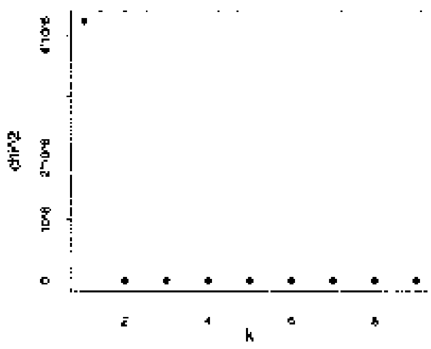

Fig. 5 ***Method of the χ^2 kritéria.*** Horizontal line denotes magnitude of χ^2_{krit} and a vertical line separates values of k, for which H_0 was accepted

Ability of absorbance matrix A^{pred} approximating the real matrix $\boldsymbol{A}$ depends on the number of latent variables k. With increasing k the degree of approximation also increases. The first matrix A^{pred} which reproduces a matrix A in range of experimental errors is considered as the best approximation and corresponding k is taken as the number of light-absorbing species in solution.

For absorbance data sets where the standard deviation varies from one absorbance point to another and is not constant Bartlett proposed a chi-squared (χ^2) criterion. This method takes into account the variability of the error from one data point to the next, but has the major disadvantage that one must have a reasonably accurate error estimate for each data point. The chi-squared (χ^2) criterion is defined

$$\chi^2 = \sum_{i-1}^{N}\sum_{j-1}^{M}\left(\frac{A_{ij}-\hat{A}_{ij}}{\sigma_{ij}}\right)^2,$$

where σ_{ij} is the standard devia-tion associated with the measurable A_{ij} and $\hat{A}_{ij}$ is the reproduced data using k PC dimensions. The criterion is applied in an iterative manner (k = 1, 2, ..., M) and the true dimensionality of the data is the first value of k at which $\chi^2_k < (N - k)(M - k)$ as $\chi^2_{(expected)} = (N - k)(M - k)$. The number of light-absorbing species corresponds the first k value for which χ^2_k is less than critical value $\chi^2_{(expected)}$.

6. The method of standard deviations of eigenvalues

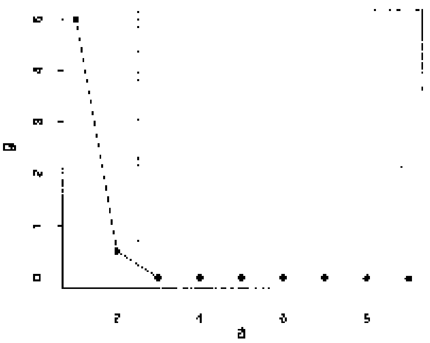

Fig. 6 ***Method od standard deviations of eigenvalue.*** Horizontal line denotes the magnitude of standard deviations of eigenvalues s_g while vertical line separates those g_n which are greater than corresponding standard deviation. First two eigenvalues are greater than their standard deviations

Hugus a El-Awady[5] related for the standard deviation of an eigenvalue of the covariance matrix $\boldsymbol{Z}$ the equation,

$$\sigma_{g_a} = \sqrt{\sum_{l=1}^{M}\sum_{m=1}^{M} q_{la}^2\, q_{ma}^2\, \sigma_{lm}^2}$$

where q_{la} a q_{ma} are elements of a matrix of eigenvalues $\boldsymbol{Q}$ and σ_{lm} are the standard deviations of elements of a matrix $\boldsymbol{Z}$ given with the relation $\sigma_{lm}^2 = \sum_{i-1}^{N}(A_{il}^2\,\sigma_{im}^2 + A_{im}^2\,\sigma_{il}^2)$ for $l \neq m$ and $\sigma_{ll}^2 = 4\sum_{i-1}^{N}(A_{il}^2\,\sigma_{il}^2)$ for $l = m$, where σ_{il} and σ_{im} are the estimates of standard deviations corresponding elements A_{il} a A_{il} of an absorbance matrix $\boldsymbol{A}$. The number of light-absorbing species in solution[4] is equal to the number of eigenvalues which are greater than σ_{g_a}.

B. Methods based on no knowledge of the experimental error of the data

If no knowledge of the experimental error associated with the data is available then one of the following techniques has to be applied to approximate the true dimensionality of the data, although the results obtained from these could be used to approximate the size of the error contained in the data. Most of the techniques presented here are empirical functions.

7. The average eigenvalue criterion

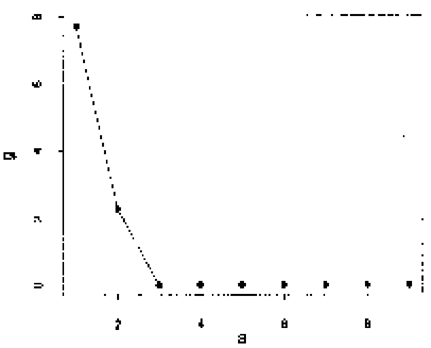

Fig. 7 ***Average value criterion.*** First two eigenvalues are greater than an arithmetic mean of eigenvalues

The average eigenvalue criterion is based upon retaining only those k latent variables or k PC dimensions whose eigenvalues g_a are above the average eigenvalue.

If the PCA is performed *via* its correlation matrix $\boldsymbol{Z}$ the average eigenvalue will be unity as the variance of each variable is unity. Therefore only those dimensions whose eigenvalue are greater than 1 should be retained. For this reason this method is also known as the *eigenvalue-one* criterion.

The number of latent variables k is then equal to the number of light-absorbing species in solution.

8. The method of Exner function

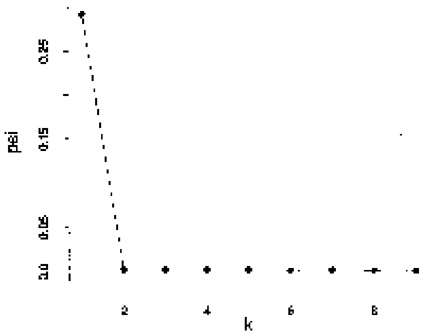

Fig. 8 ***Method of Exner function.*** Value $\psi \leq 0.1$ is achieved for $k = 2$. Higher value k does not bring significant decreasing value ψ

The Exner psi (ψ) function may be used for identifying the true dimensionality of a data. This function is defined as

$$\psi = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{ij}-\hat{A}_{ij})^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{ij}-\bar{A})^2}\cdot\frac{NM}{(NM)-k}},$$ where $\bar{A}$ represents the overall mean of the absorbance matrix $\boldsymbol{A}$ and $\hat{A}_{ij}$ is the reproduced data using the first k latent variables. The ψ function can vary from zero to infinity, with the best fit approaching zero. A ψ equal to 1.0 is the upper limit for significance as this means the data reproduction using k dimensions is no better than saying each point is equal to the overall data mean. Exner proposed that 0.5 be considered the largest acceptable ψ value, because this means the fit is twice as good as guessing the overall mean for each data point. Using this reasoning $\psi = 0.3$ can be considered a fair correlation, $\psi = 0.2$ can be considered a good correlation and $\psi = 0.1$ an excellent correlation. It means that for $\psi < 0.1$ the corresponding k can be taken as the number of light-absorbing species in solution.

9. The method of residual variance (The scree test)

The scree test for identifying the true dimensionality of a data set is based on the observation that the residual variance should level off before those dimensions containing random error are included in the data reproduction. The residual variance associated with a reproduced data set, is defined as $RV = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{ij}-\hat{A}_{ij})^2}{NM}$ which is equal to the square of the XE error. The residual variance can be expressed as a percentage as

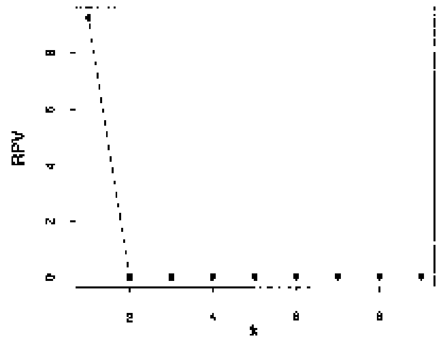

Fig. 9 *The scree test*

$RPV = 100\left(\sum_{i=1}^{N}\sum_{j=1}^{M}(A_{ij}-\hat{A}_{ij})^2 \,/\, \sum_{i=1}^{N}\sum_{j=1}^{M}A_{ij}^2\right)$. In terms of the eigenvalues of the data matrix, this expression can be converted to $RPV = 100\left(\sum_{a=k+1}^{M} g_a \,/\, \sum_{a=1}^{M} g_a\right)$. When the residual percent variance is plotted against the number of k PC dimensions used in the data reproduction, $RPV = f(k)$, the curve should drop rapidly and level off at some point. The point where the curve begins to level off, or where a discontinuity appears, is taken to be the dimensionality of the data space. This is explained by the fact that successive eigenvalues (k PC dimensions) explain less variance in the data and hence this explains the continual drop in the residual percent variance. However the error eigenvalues will be equal, if the experimental error associated with the data is truly random, and hence the residual percent variance will be equal. Discontinuity appears in situations where the errors are not random, in such situations PCA exaggerates the non-uniformity in the data as it aims to explain the variation in the data.

10. The method of logarithms of eigenvalues

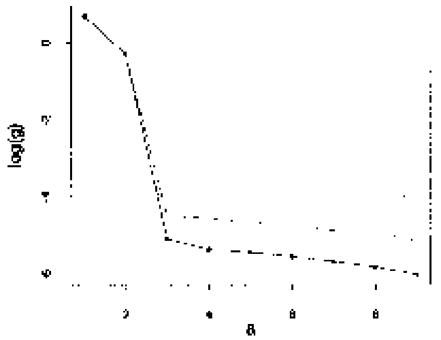

Fig. 10 *Method of logarithms of eigenvalues*. Dashed curve denote eigenvalues for absorbance matrix with superimposed random noise with the standard deviations $2\times s_{inst}(A)$ and $4\times s_{inst}(A)$. Only for a = 1 and 2 the value log (g_a) does not depend on generated noise

The method of logarithms of eigenvalues[3] comes from an assumption that primary eigenvalues of the covariance matrix $\boldsymbol{Z}$ significantly differ in a magnitude from secondary eigenvalues as their magnitude is approximately same. Therefore the primary and secondary eigenvalues can be separated graphically in a plot $\log(g_a) = f(a)$, where a is the order of given eigenvalue in descending order.

However, this test is not sufficiently sensitive on a presence of light-absorbing species in relatively small quantities. Therefore some information about a noise in absorbance should also be supplied. When in one graph various levels of experimental error in absorbance are plotted then the primary and secondary eigenvalues may be easily recognized. The number of primary eigenvalues is then equal to the number of light-absorbing species in solution.

11. The method of Imbedded Error Function

The imbedded error function IE is an empirical function developed to identify those k latent variables or PC dimensions containing error without relying upon an estimate of the error associated with the absorbance data matrix. The imbedded error is a function of the error eigenvalues and takes the following form $IE = \sqrt{\dfrac{k\sum_{a=k+1}^{M} g_a}{NM\,(M-k)}}$ and represents a measure of the difference between reconstructed and pure data and describes this part of errors which

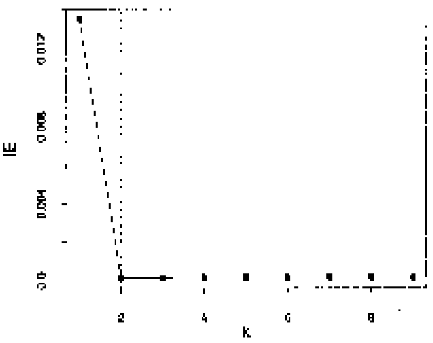

Fig. 11 ***Method of imbedded error function.*** Function *IE* = f(*k*) reaches a minimum for *k* = 2

remains in reconstructed data. The behavior of the *IE* function, as *k* varies from 1 to *M*, can be used to deduce the true dimensionality of the data. The *IE* function should decrease as the true dimensions are used in the data reproduction. However when the true dimensions are exhausted, and the error dimensions are included in the reproduction, the *IE* should increase. This should occur because the error dimensions are the sum of the squares of the projections of the error points on the error axis. If the errors are uniformly distributed, then their projections onto the error dimensions should be approximately equal.

12. The method of Factor Indicator Function

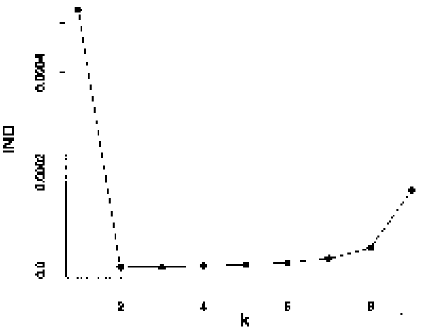

Fig. 12 *Method of factor indicator function*

The factor indicator function *IND* is an empirical function which appears more sensitive than the IE function to identify the true dimensionality of an absorbance data matrix. The function is composed of the same components as the *IE* function, and is defined as

$$IND = \frac{\sqrt{\dfrac{\sum_{a=k+1}^{M} g_a}{N(M-k)}}}{(M-k)^2} = \frac{RE}{(M-k)^2}$$

, where *RE* is the residual standard deviation of absorbance. This function, like the *IE* function, reaches a minimum when the correct number of latent variables or *k* PC dimensions have been employed in the data reproduction. However, it has been seen that the minimum is more pronounced and can often occur in situations where the *IE* function exhibits no minimum.

13. Malinowski F–test

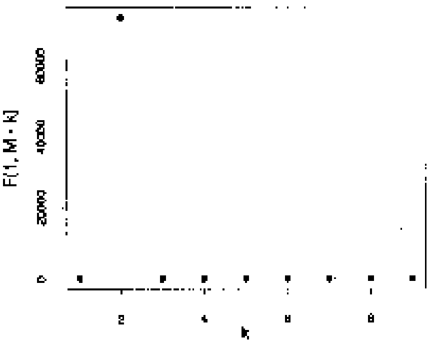

Fig. 13 *Malinowski F–test.* Values on x-axis correspond to the number of light absorbing species in mixture while values on y-axis to testing criterion *F(1, M - k)*. Highest *k* fulfilling H_0 is for *k* = 2

Malinowski[6, 7] developed a test for determining the true dimensionality of a data set based on the Fisher variance ratio test (F-test). The F-test is a quotient of two variances obtained from two independent pools of samples that have normal distributions. As the eigenvalues obtained from a PCA are orthogonal, the condition of independence is satisfied. It is common to assume that the residual errors in the data have a normal distribution; if this is true, then the variance expressed by the error eigenvalues should also follow a normal distribution. This will not be the case if the errors in the data are not uniform or if systematic errors exist. The pooled variance of the error eigenvectors is obtained by dividing the sum of error eigenvalues by the number of pooled vectors *M* - *k*. For distinguishing primary and secondary eigenvalues the null hypothesis H_0:

$g_a^{red} = \bar{g}_a^{red}$ vs. alternative H$_A$: $g_a^{red} > \bar{g}_a^{red}$ is formulated. In case of validity of null hypothesis the test criterion

$$F(1, M-k) = \frac{\sum_{a=k+1}^{M} (N-a+1)(M-a+1)}{(N-k+1)(M-k+1)} \cdot \frac{g_a}{\sum_{a=k+1}^{M} g_a}$$

with 1 and M - k degrees of freedom is applied. When testing the k is varied from the smallest eigenvalues in range M - 1, M - 2, ..., 1. The first k-th reduced eigenvalue for which it is valid that $F(1,\ M - k)$ is greater than critical value for given significance level is taken as the smallest and corresponding k represents the number of light-absorbing species in solution.

14. Bartlett isotropy test

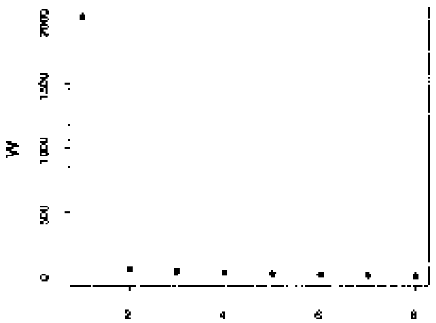

Fig. 14 ***Bartlett test.*** Vertical line separates values k for which H$_0$ is not accepted

Bartlett isotropy test uses the null hypothesis about an equality of M - k lowest eigenvalues of a covariance matrix Z in form: $H_0 : g_{k-1} = g_{k-2} = \ldots = g_M$.

The test criterion W is defined

$$W = \left(N - \frac{2M+11}{6}\right)(M-k) \ln\left(\frac{\bar{g}_k^{a}}{\bar{g}_k^{g}}\right)$$

where $\bar{g}_k^{a}$ is the arithmetic mean and $\bar{g}_k^{g}$ the geometric mean of M - k lowest eigenvalues of matrix **Z**. Supposing a validity of H$_0$ the criterion W has approximately χ^2 distribution with 0.5 (M – k +2) (M – k – 1) degrees of freedom. The integer k is changed k = 0, 1, 2, ... up to the first value of k for which the W is lower than χ^2. The number of light-absorbing species is then equal to a number k-1 for which H$_0$ was not accepted.

Discussion

1. Estimation of the instrumental error of spectrophotometer used, $s_{inst}(A)$

For a determination of the instrumental error of spectrophotometer used, $s_{inst}(A)$, Wernimont-Kankare method[9] based on the concentration dependence of a spectral standard was applied. One spectral standard in solution means one light-absorbing species and therefore a rank of absorbance matrix is equal to 1 and corresponding residual standard deviation of absorbance $s_k(A)$ may be estimated from a graph $s_k(A) = f(k)$. Analogically, $s_{inst}(A)$ may be estimated with the use of *the real error RE, the extracted error XE* and *the average error* $\bar{e}$ for k = *1*.

Table 1 Estimation of the instrumental error of spectrophotometer used, $s_{inst}(A)$,

	K_2CrO_4		$CoSO_4 \cdot 7H_2O$		$CuSO_4 \cdot 5H_2O$	
Repeatability =	1.	2.	1.	2.	1.	2.
Kank, $k = 1$	0.00070	0.00069	0.00038	0.00031	0.00036	0.00024
RE, $k = 1$	0.00048	0.00040	0.00018	0.00015	0.00022	0.00018
XE, $k = 1$	0.00047	0.00038	0.00018	0.00015	0.00021	0.00017
$\bar{e}$, $k = 1$	0.00024	0.00021	0.00019	0.00016	0.00018	0.00015

In determination of $s_{inst}(A)$ the most pessimistic estimates of $s_k(A)$, *RE*, *XE*, and $\bar{e}$ were taken for potassium bichromate i. e. $s_l(A) = 0.00070$, $RE = 0.00048$, $XE = 0.0047$ a $\bar{e} = 0.0024$.

2. Criticism of the algorithm used

An algorithm for estimation of the number of light-absorbing species in solution was validated using four various sets of absorbance data: 16 sets of spectra of dissociation equilibria of sulphonephtaleins Bromocresol Green, Phenol Red, Thymol Blue and 4 sets of spectral standards potassium bichromate, cobalt(II) sulphate and copper(II) sulphate.

2.1 Acid-base equilibria of some sulphonephtaleins (Table 2)

For simple acid-base equilibrium of sulphonephtaleins HL ⇄ $H^+ + L^-$ and with the use of known matrix of molar absorptivities the absorbance matrix A^{pure} with precision of 9 valid digits is calculated. Generated random errors with the instrumental standard deviation $s_{inst}(A)$ (=0.0003, 0.0006, 0.0009, 0.0020 a 0.0040) have been superimposed to all absorbance values. Absorbance data describing two various light-absorbing species L^- and HL were calculated for Bromocresol Green (sets A1 through A5), Phenol Red (sets A6 through A10) and Thymol Blue (sets A11 through A15). Last set A16 concerns six light-absorbing species from all three sulphonephtaleins used was of experimental nature. Table 2 shows that found $s_k(A)$ is always slightly lower than an optioned value $s_{inst}(A)$ used for a simulation.

Except the Bartlett isotropy test the all other methods reached true estimation $k = 2$. When a number of light-absorbing species was $k = 6$, only seven methods i. e. $s_k(A)$, *RE*, *XE*, $\bar{e}$, *chi^2*, σ_g and *log(g)* found true estimates *k*. It is interesting that successful methods are mostly based on preliminary knowledge of the instrumental error $s_{inst}(A)$.

2.2 Spectra of three standards

For a comparison of effectivity of factor analysis for a search of the number of light-absorbing species a set of spectra of a Beer concentration dependence of 3 spectral standards was applied. Four experimental sets of spectra (B1 through B4) were measured in which spectral standards $K_2Cr_2O_7$, $CoSO_4 \cdot 7H_2O$ a $CuSO_4 \cdot 5H_2O$ were combined (Table 3). Very low values of $s_{inst}(A)$ prove quite reliable spectrophotometer and experimental technique used.

In case of B1 seven methods, ***RE*, *XE*, $\bar{e}$, *chi^2*, σ_g, *g*** and ***RPV***, true value $k = 2$ were found. For B2 nine methods, i. e. ***$s_k(A)$*, *RE*, *XE*, $\bar{e}$, *chi^2*, σ_g, *RPV*, *log(g)*** and ***F*** true value $k = 2$ were found. For B3 same result as for B2 was received. For three spectral standards in a mixture in case of B4 nine methods, i. e. ***$s_k(A)$*, *RE*, $\bar{e}$, *chi^2*, σ_g, *log(g)*, *IND*, *F*** and ***W***, found true values.

2.3 Reliability of validated methods

Critical comparison of all methods of determination of the number of light-absorbing species in solution was carried out (Table 4). The percentage of successful determination of *k* for the individual methods is presented here. It may be concluded that **the first group** contains ***very reliable methods*** (Kankare method $s_k(A)$, the method of real method *RE*, the method extracted error *XE*, the average error criterion $\bar{\varepsilon}$, the method of standard deviations of eigenvalues σ_g, the method of logarithms of eigenvalues log(*g*)) with a successful determination of 91 through 100%. The **second group** contains ***reliable methods*** (the method of χ^2 criterion *chi^2*, the scree test *RPV* and Malinowski *F*–test) with a successful determination of 81 through 90%. The **third group** contains ***unreliable and quite unreliable methods*** (the average eigenvalue criterion g_a, the method of Exner function *PSI*, the method of imbedded error function *IE*, the method of factor indicator function *IND*, Bartlett test *W*) with a successful determination less than 80%.

Conclusion

1. Wernimont-Kankare procedure was used as the best method for determination of the instrumental standard deviation of spectrophotometer used, $s_{inst}(A) = 0.0007$ in range 380 - 650 nm.
2. Analysing 20 sets of spectra with 14 methods of factor analysis there were found the most reliable methods: Kankare method, the method of real error, the method of extracted error, the average error criterion, the method of standard deviations of eigenvalues and the method logarithms of eigenvalues.
3. Generally, the most reliable methods seem to be methods based on knowledge of instrumental error.
6. In determination of the number light-absorbing species it is more reliable and highly recommended to make a comparison of different methods of factor analysis. The methods based on knowledge of instrumental error are preferred.

References

1. Malinowski E. R. (1977) Anal. Chem. 49: 606.
2. Kankare J. J. (1970) Anal. Chem. 42: 1322.
3. Brereton R. G., *Data Handling in Science and Technology – Volume 9: Multivariate Pattern Recognition in Chemometrics, Chapter 5*, J. M. Deane: *Data Reduction Using Principal Components Analysis*, Elsevier, Amsterodam – London – New York – Tokyo, 1992.
4. Čepan M., Pelikán P., Liška M.: *Metódy faktorovej analýzy v molekulovej spektrometrii I. Abstraktná faktorová analýza a určenie počtu aktívnych zložiek v spektrách zmesi*, (in press).
5. Hugus Z. Z. Jr., El–Awady A. A. (1971) J. Phys. Chem. 75: 2954.
6. Malinowski E. R. (1987) J. Chemometrics 1: 33.
7. Malinowski E. R. (1989) J. Chemometrics 3: 49.
8. Antoon M. K. , D'Esposito L. , Koenig J. L. (1979) Appl. Spectrosc. 33: 351.
9. Wernimont G. (1967) Anal. Chem. 39: 554.

Table 2 Spectra of the acid-base equilibria of sulphonephtaleins: BKG - Bromocresol Green, FC - Phenol Red, TM - Thymol Blue, (Algorithms: **Kank** Kankare method, **RE** the method of real method, **XE** the method extracted error, **$\bar{e}$** method of average error, **chi^2** the method of χ^2 criterion, **g.** the method of standard deviations of eigenvalues, **psi** the method of Exner function, **RPV** the scree test, **log(g)** the method of logarithms of eigenvalues, **IE** the method of imbedded error function , **IND** the method of factor indicator function, **F** Malinowski *F*–test, **Bart** Bartlett test, (–) denotes unavailable estimation)

Set	Spectra for equilibrium	Typ of data, $s_{inst}(A)$, $s_k(A)$ found (Kank)	Found number of light-absorbing species in brackets for every method
A1	BKG – H^+	Simulated data, 0.0003, 0.00027	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A2	BKG – H^+	Simulated data, 0.0006, 0.00051	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (2)
A3	BKG – H^-	Simulated data, 0.0009, 0.00078	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A4	BKG – H^-	Simulated data, 0.0020, 0.00173	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A5	BKG – H^-	Simulated data, 0.0040, 0.00376	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (3), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (–), IND (2), F (2), Bart (3)
A6	FC – H^-	Simulated data, 0.0003, 0.00024	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A7	FC – H	Simulated data, 0.0006, 0.00053	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A8	FC – H^+	Simulated data, 0.0009, 0.00081	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A9	FC – H^+	Simulated data, 0.0020, 0.00172	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)

A10	FC – H^+	Simulated data, 0.0040, 0.00347	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (–), IND (2), F (2), Bart (2)
A11	TM – H^+	Simulated data, 0.0003, 0.00027	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A12	TM – H^-	Simulated data, 0.0006, 0.00054	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (–), IND (2), F (2), Bart (1)
A13	TM – H^-	Simulated data, 0.0009, 0.00079	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IF (2), IND (2), F (2), Bart (1)
A14	TM – II^+	Simulated data, 0.0020, 0.00195	Kank (2), RF (2), XE (2), $\bar{c}$ (2), chi^2 (3), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (–), IND (2), F (2), Bart (1)
A15	TM – II^+	Simulated data, 0.0040, 0.00324	Kank (2), RF (2), XE (2), $\bar{c}$ (2), chi^2 (2), σ_g (2), g (2), psi (2), RPV (2), log(g) (2), IE (2), IND (2), F (2), Bart (1)
A16	BKG – FC – TM – H^+	Experimental data, 0.0015, 0.00008	Kank (6), RE (6), XF (6), $\bar{c}$ (6), chi^2 (6), σ_g (6), g (3), psi (2), RPV (5), log(g) (6), IE (–), IND (–), F (31), Bart (–)

Table 3 Spectra of standards $K_2Cr_2O_7$, $CoSO_4 \cdot 7H_2O$ and $CuSO_4 \cdot 5H_2O$,
(Algorithms: **Kank** Kankare method, **RE** the method of real method, **XE** the method extracted error, **$\bar{e}$** method of average error, **chi^2** the method of χ^2 criterion, **σ_g** the method of standard deviations of eigenvalues, **psi** the method of Exner function, **RPV** the scree test, **log(g)** the method of logarithms of eigenvalues, **IE** the method of imbedded error function , **IND** the method of factor indicator function, **F** Malinowski *F*–test, **Bart** Bartlett test, (–) denotes unavailable estimation)

Sct	Spectra for equilibrium	Typ of data, $s_{inst}(A)$ $s_k(A)$ found (Kank)	Found number of light-absorbing species in brackets for every method
B1	$K_2Cr_2O_7$ – $CoSO_4 \cdot 7II_2O$	Experimental data, 0.0007, 0.00089	Kank (3), RE (2), XF (2), $\bar{c}$ (2), chi^2 (2), σ_g (2), g (2), psi (1), RPV (2), log(g) (3), IE ( ), IND (5), F (3), Bart (14)
B2	$CoSO_4 \cdot 7H_2O$ – $CuSO_4 \cdot 5H_2O$	Experimental data, 0.0007, 0.00047	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (1), psi (1), RPV (2), log(g) (2), IF (–), IND (8), F (2), Bart (11)
B3	$K_2Cr_2O_7$ – $CuSO_4 \cdot 5H_2O$	Experimental data, 0.0007, 0.00047	Kank (2), RE (2), XE (2), $\bar{e}$ (2), chi^2 (2), σ_g (2), g (1), psi (1), RPV (2), log(g) (2), IE (–), IND (7), F (2), Bart (10)
B4	$K_2Cr_2O_7$ – $CoSO_4 \cdot 7H_2O$ – $CuSO_4 \cdot 5H_2O$	Experimental data, 0.0007, 0.00023	Kank (3), RF (3), XE (3), $\bar{c}$ (3), chi^2 (3), σ_g (2), g (2), psi (2), RPV (2), log(g) (3), IE (–), IND (3), F (3), Bart (3)

Table 4 Accuracy of the estimated number of light-absorbing species in a mixture: 1 denotes true result, 0 denotes false results

(Algorithms: **Kank** Kankare method, **RE** the method of real method, **XE** the method extracted error, **ē** method of average error, **chi^2** the method of χ^2 criterion, $\boldsymbol{\sigma_g}$ the method of standard deviations of eigenvalues, **psi** the method of Exner function, **RPV** the scree test, **log(g)** the method of logarithms of eigenvalues, **IE** the method of imbedded error function , **IND** the method of factor indicator function, **F** Malinowski *F*–test, **Bart** Bartlett test, (–) denotes unavailable estimation)

Soub.	Kank	RE	XE	ē	chi^2	σ_g	g	psi	RPV	log(g)	IE	IND	F	Bart
A1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A3	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A4	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A5	1	1	1	1	0	1	1	1	1	1	0	1	1	0
A6	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A7	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A8	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A9	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A10	1	1	1	1	1	1	1	1	1	1	0	1	1	1
A11	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A12	1	1	1	1	1	1	1	1	1	1	0	1	1	0
A13	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A14	1	1	1	1	0	1	1	1	1	1	0	1	1	0
A15	1	1	1	1	1	1	1	1	1	1	1	1	1	0
A16	1	1	1	1	1	1	0	0	0	1	0	0	0	0
B1	0	1	1	1	1	1	1	0	1	0	0	0	0	0
B2	1	1	1	1	1	1	0	0	1	1	0	0	1	0
B3	1	1	1	1	1	1	0	0	1	1	0	0	1	0
B4	1	1	1	1	1	0	0	0	0	1	0	1	1	1
Agree	19	20	20	20	18	19	16	15	18	19	11	16	18	3
Success [%]	**95**	**100**	**100**	**100**	**90**	**95**	**80**	**75**	**90**	**95**	**55**	**80**	**90**	**15**
Group	I	I	I	I	II	I	III	III	II	I	III	III	II	III