ELSEVIER

Analytica Chimica Acta 423 (2000) 51–68

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

Critical comparison of methods predicting the number of components in spectroscopic data

Milan Meloun[a,*], Jindřich Čapek[a], Petr Mikšík[a], Richard G. Brereton[b]

[a] *Department of Analytical Chemistry, Faculty of Chemical Technology, University Pardubice, CZ-532 10 Pardubice, Czech Republic*
[b] *School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK*

Received 26 April 2000; accepted 27 July 2000

Abstract

Determining the number of chemical components in a mixture is a first important step to further analysis in spectroscopy. The accuracy of 13 statistical indices for estimation of the number of components that contribute to spectra was critically tested on simulated and on experimental data sets using algorithm INDICES in S-Plus. All methods are classified into two categories, precise methods based upon a knowledge of the instrumental error of the absorbance data, $s_{\mathrm{inst}}(A)$, and approximate methods requiring no such knowledge. Most indices always predict the correct number of components even a presence of the minor one when the signal-to-error ratio (SER) is higher than 10 but in case of RESO and IND higher than 6. On base of SER the detection limit of every index method is estimated. Two indices, RESO and IND, correctly predict a minor component in a mixture even if its relative concentration is 0.5–1% and solve an ill-defined problem with severe collinearity. For more than four components in a mixture the modifications of Elbergali et al. represent a useful resolution tool of a correct number of components in spectra for all indices. The Wernimont–Kankare procedure performs reliable determination of the instrumental standard deviation of spectrophotometer used. In case of real experimental data the RESO, IND and indices methods based on knowledge of instrumental error should be preferred. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Chemometrics; Principal component analysis; PCA; Rank of matrix; Number of components; Real error; Extracted error; Method of logarithm of eigenvalues; Instrumental error of spectrophotometer; Number of components in a mixture

1. Introduction

Determining the number of chemical components in a mixture is the first step for further qualitative and quantitative analysis in all forms of spectral data treatment. Procedures for determining the chemical rank of a matrix using a variety of empirical and statistical methods based on principal component analysis (PCA) have been reported [1]. Much work has been put into developing methods for resolution of multi-component spectra but less work has been carried out to reveal the limitations of the methods and in the estimation of the minor component of resolved spectra.

Throughout this work, it is assumed that the $n \times m$ absorbance data matrix $\boldsymbol{A} = \boldsymbol{\varepsilon C}$ containing the n recorded spectra as rows can be written as the product of the $m \times r$ matrix of molar absorptivities $\boldsymbol{\varepsilon}$ and the $r \times n$ concentration matrix $\boldsymbol{C}$. Here m denotes the number of wavelengths for which each spectrum was recorded being equal to the number of columns of $\boldsymbol{A}$ matrix, n is the number of solutions for which spectra have been recorded being equal to the number of rows

* Corresponding author. Tel.: +42-40-603-7026;
fax: +42-40-603-7068
E-mail address: milan.meloun@upce.cz (M. Meloun).

PII: S0003-2670(00)01100-4

of $\boldsymbol{A}$ matrix, r is the number of components that absorb in the chosen spectral range. The rank of the matrix $\boldsymbol{A}$ is obtained from the equation

$$\mathrm{rank}(\boldsymbol{A}) = \min[\mathrm{rank}(\boldsymbol{\varepsilon}),\ \mathrm{rank}(\boldsymbol{C}) \leq \min(m, r, n) \quad (1)$$

Since the rank of $\boldsymbol{A}$ is equal to the rank of $\boldsymbol{\varepsilon}$ or $\boldsymbol{C}$, whichever is the smaller, and since $\mathrm{rank}(\boldsymbol{\varepsilon}) \leq r$ and $\mathrm{rank}(\boldsymbol{C}) \leq r$, then provided $m \geq r$ and $n \geq r$, it will only be necessary to determine the rank of matrix $\boldsymbol{A}$ which is equivalent to the number of significant components [2,18].

Approaches for the determination of the rank of $\boldsymbol{A}$ are based on two different chemometric methods, either using pure PCA or by using PCA combined with cross-validation [3–12]. Generally, PCA will extract some of noise, i.e. experimental and/or random error which will usually be represented by the principal components with smallest size or variance. When no noise in spectra exists, the number of eigenvalues of the covariance matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ larger than zero is equivalent to the number of component r, providing that the spectra of components in mixture are linearly independent.

In a recent tutorial, Malinowski [3] concluded that spectra gleaned from a spectrophotometer often contains instrumental as well as experimental uncertainties that arise from several different sources: (a) spectrophotometer switches a filter and new uncertainty and also uncertainty of distribution are introduced; (b) changing sample cells or stock solvents will produce uncertainty distribution in the data; (c) data pretreatment such as smoothing, normalizing or standardizing the data columns or data rows can seriously effect the uncertainty distribution; (d) data can be distorted by a combination of these factors.

As all real data contain experimental noise, the number of eigenvalues different from zero is usually larger than the number of components r. Experimental and/or random error can mask the identification of the true dimensionality of a data set. Malinowski [1,4] split this error into two sources — imbedded error and extracted error. Extracted error (XE) is the error which is contained within the minor PC dimensions ($(r+1)$th, $(r+2)$th, $\ldots$, mth) and therefore be removed, or extracted, from the data by retaining only the first r dimensions. Imbedded error (IE) is the error which mixes into the factor scheme and is contained within the first r dimensions: this error can never be completely removed from a data set but may be scaled to a minimum [5].

Chen et al. [6] and Elbergali et al. [7] concluded that although there are many multivariate statistical methods for determination of the number of significant components that have successfully solved certain problems encountered in spectroscopy, if the spectra of components are very similar, or there are minor components, or the signal-to-noise ratio (SNR) is low, the methods may not perform well. Some methods may still fail to give satisfactory results due to the existence of some components and noise eigenvalues may be in the same magnitude. All these methods to identify the true dimensionality of a data set are classified into two categories: (a) precise methods based upon a knowledge of the instrumental error of the absorbance data, $s_{\mathrm{inst}}(A)$ before statistical examination; (b) approximate methods requiring no knowledge of the instrumental error of the absorbance data, $s_{\mathrm{inst}}(A)$. Many of these methods are empirical functions.

The purpose of our study was to make a critical comparison of various PCA methods on both simulated and experimental data and first results and algorithm were already presented [26]. In this paper, statistical properties of the instrumental random error, i.e. its homoscedastic magnitude but also heteroscedastic influence as well as a resolution under a presence of minor components in mixture with a detection limit in case of 13 various indices methods are discussed.

2. Theoretical

2.1. PCA in spectral data analysis

Principal component analysis (PCA) has been widely used for spectral data analysis since it was introduced into chemistry by Kankare [18]. PCA performs a decomposition of an absorbance matrix into a product of two matrices $\boldsymbol{T}$ and $\boldsymbol{P}^{\mathrm{T}}$ and the residual matrix $\boldsymbol{E}$ according to

$$\boldsymbol{A} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \quad (2)$$

The $n \times q$ score matrix $\boldsymbol{T}$ also called a matrix of latent variables contains q column vectors or main components. The $m \times q$ loading matrix $\boldsymbol{P}$ contains q column vectors which represents a measure of contribution of a particular latent variable. The index q is the least of n and m which in spectroscopy usually is n [9,13–17]. The second moment of an absorbance matrix is defined

$$\boldsymbol{Z} = \frac{1}{n-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \tag{3}$$

where $\boldsymbol{A}$ is usually the mean-centered absorbance matrix. The matrix $\boldsymbol{Z}$ is often called the variance–covariance matrix and contains information about the scatter of points in multidimensional space. The latent root and vector decomposition is defined by two equations

$$|\boldsymbol{Z} - g_a\boldsymbol{I}| = \boldsymbol{0} \tag{4a}$$

$$\boldsymbol{Z}\boldsymbol{p_a} = l_a\boldsymbol{p_a} \tag{4b}$$

where $\boldsymbol{Z}$ is a $m \times m$ variance–covariance matrix (sometimes data can be scaled so that each variable is standardized to equal variance down the columns: in which case the matrix becomes the correlation matrix); the matrix $\boldsymbol{I}$ is the unit matrix, and $\boldsymbol{0}$ is a matrix of zeroes. Eq. (4a) is a constrained maximization in which g is called the Lagrange multiplier; the g_a are the r latent roots and are obtained as the roots of the polynomial equation of order m defined by the determinant. The g_a denote sum of squares of scores divided by the number of elements. Eq. (2) defines the corresponding latent vectors $\boldsymbol{p_a}$ of dimension n. Two constraints are placed on the loadings vectors. Only the loadings have unit length and are mutually orthogonal, the scores do not.

The following notations are used: I is the sample number, j is the wavelength number, and a is the eigenvalue number. Then n is the number of samples, k is the current number of components being testing, r is the true number of components and q is the total possible number of components.

2.2. Precise methods

Precise methods concern such indices which are based upon a knowledge of the instrumental error of the absorbance data, $s_{\mathrm{inst}}(A)$. Determination of a number of significant components in mixture is based on a comparison of an actual index of method used with the experimental error of instrument used, $s_{\mathrm{inst}}(A)$.

2.2.1. Residual standard deviation, $s_k(A)$

Kankare in algorithm FA608 [2,18] uses the second moment $\boldsymbol{Z}$ of an absorbance matrix $\boldsymbol{A}$ (Eq. (2)). Applying eigenvalues g_a of matrix $\boldsymbol{Z}$ the residual standard deviation of absorbance $s_k(A)$ is estimated

$$s_k(A) = \sqrt{\frac{\mathrm{tr}(\boldsymbol{Z}) - \sum_{a=1}^{k} g_a}{n-k}} \tag{5}$$

where tr($\boldsymbol{Z}$) is a trace of the matrix $\boldsymbol{Z}$ and r is the estimated number of components in a mixture.

Testing criterion: the values $s_k(A)$ for different number of components k are plotted against an integer k, $s_k(A) = f(k)$, and number of significant components is such integer $r = k$ for which $s_k(A)$ is close to the instrumental standard deviation of absorbance, $s_{\mathrm{inst}}(A)$.

2.2.2. Residual standard deviation, RSD

The RSD [1] is a measure of the lack of fit of a PC model to a data set being calculated by

$$\mathrm{RSD}(k) = \sqrt{\frac{\sum_{a=k+1}^{q} g_a}{n(q-k)}} \tag{6}$$

if the PCA is performed via the covariance matrix; where g_a is the eigenvalue associated with the kth PC dimension.

Testing criterion: the true dimensionality of a data set r is the number of dimensions required to reduce the RSD(k) to be approximately equal to the estimated experimental error of the absorbance data. The RSD(k) may be plotted against k, RSD(k) $= f(k)$, and when the RSD(k) reaches the value of the instrumental error of spectrophotometer used, $s_{\mathrm{inst}}(A)$, the corresponding k represents the number of significant components in a mixture, $r = k$.

2.2.3. Root mean square error, RMS

The root mean square error (RMS) [1] 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. It is also known as the extracted error XE(k) [1]. The RMS(k) is defined by

$$\mathrm{RMS}(k) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij} - \hat{A}_{ij})^2}{nm}} = \mathrm{XE}(k) \tag{7}$$

where $\hat{A}_{ij} = \bar{A} + \sum_{a=1}^{k} t_{ia} p_{aj}$ and scores are denoted by t_{ia} while loadings by p_{aj}. The alternative way of expressing RMS(k) is as follows

$$\mathrm{RMS}(k) = \sqrt{\frac{\sum_{a=k+1}^{m} g_a}{nm}} \tag{8}$$

where g_a are eigenvalues of a covariance matrix $\boldsymbol{Z}$ and k is a guess which can vary from 1 to q and we are testing to see whether $k = r$ or not.

Testing criterion: analogously as in previous method the estimates RMS(k) may be plotted as a function of latent variables, $\mathrm{RMS}(k) = f(k)$ and on base of a comparison with the magnitude of an instrumental error of the spectrophotometer used, $s_{\mathrm{inst}}(A)$, the number of the significant components may be estimated. Comparing relations for RSD(k) and RMS(k), and simplifying yields we get $\mathrm{RMS}(k) = \sqrt{m - n/m}(\mathrm{RSD}(k))$. Although related, RMS($k$) and RSD($k$) measure different sources of error. RMS(k) measures the difference between raw data and reproduced data using k PC dimensions. RSD(k), however, measures the difference between the raw and the pure data containing no experimental error.

2.2.4. Average error criterion, AE

The average error of absorbance AE (or $\bar{e}$) [1,22] is simply the average of the absolute values of the differences between the raw and reproduced data,

$$\bar{e}(k) = \mathrm{AE}(k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} |A_{ij} - \hat{A}_{ij}|}{nm} \tag{9}$$

where A_{ij} and $\hat{A}_{ij}$ were described previously in Section 2.2.3.

Testing criterion: the true dimensionality of absorbance data matrix r 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 AE(k) are plotted against the number of latent variables k and compared with the instrumental error of spectrophotometer used, $s_{\mathrm{inst}}(A)$. When AE(k) reaches $s_{\mathrm{inst}}(A)$ corresponding k is equal to the number of significant components in mixture, $r = k$.

2.2.5. χ^2 criterion

For absorbance data sets where the standard deviation varies from one absorbance point to another and is not constant Bartlett [19] proposed a 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 χ^2-criterion is defined by

$$\chi^2(k) = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\frac{A_{ij} - \hat{A}_{ij}}{\sigma_{ij}}\right)^2 \tag{10}$$

where σ_{ij} is the standard deviation associated with the measurable A_{ij} and $\hat{A}_{ij}$ is the reproduced data using k PC dimensions.

Testing criterion: the criterion is applied in an iterative manner ($k = 1, 2, \ldots, m$) and the true dimensionality of the data is the first value of k at which $\chi^2(k) < (n-k)\,(m-k)$ as $\chi^2_{\mathrm{expected}} = (n-k)(m-k)$. The number of significant components corresponds the first k value for which $\chi^2(k)$ is less than critical value $\chi^2{}_{\mathrm{expected}}$.

2.2.6. Standard deviation of eigenvalues, s(g_k)

Hugus and El-Alwady [20] related for the standard deviation of an eigenvalue of the covariance matrix $\boldsymbol{Z}$ the equation,

$$s(g_k) = \sqrt{\sum_{k=1}^{m}\sum_{j=1}^{m} q_{la}^2 q_{ja}^2 \sigma_{lj}^2} \tag{11}$$

where q_{la} and q_{ja} are elements of a matrix of eigenvalues $\boldsymbol{Q}$ and σ_{lj} are the standard deviations of elements of a matrix $\boldsymbol{Z}$ given with the relation $\sigma_{lj}^2 = \sum_{i=1}^{n}(A_{il}^2\sigma_{ij}^2 + A_{ij}^2\sigma_{il}^2)$ 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_{ij} are the estimates of standard deviations corresponding elements A_{il} and A_{ij} of an absorbance matrix $\boldsymbol{A}$.

Testing criterion: the number of significant components in mixture is equal to the number of eigenvalues which are greater than $s(g_k)$.

2.3. Approximate methods

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 [1]. Most of the techniques presented here are empirical functions.

2.3.1. Eigenvalues, g_a

Eigenvalues EV(a) or g_a are conventionally used as a measure of the size of a principal component [21]. Eigenvalues are calculated as the sum of squares of the score vectors

$$\mathrm{EV}(a) = g_a = \sum_{i=a}^{n} t_{ia,}^2 \quad a = 1, 2, \dots, r, \dots, q \qquad (12)$$

Testing criterion: the first r eigenvalues, called a set of primary eigenvalues, contain contribution from the real components and should be considerably larger than those containing only noise. The second set called the secondary eigenvalues contains $(q-r)$ eigenvalues and are referred to as non-significant eigenvalues. The secondary eigenvalues should be considerably larger, but this is not sensitive enough.

2.3.2. Logarithms of eigenvalues, log g_a

The method of logarithms of eigenvalues [9] 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.

Testing criterion: 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 significant components in relatively small quantities. Therefore some information about instrumental noise 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 significant components in a mixture.

2.3.3. Exner function, ψ

The Exner ψ-function [23] is another approach for identifying the true dimensionality of a data. This function is defined as

$$\psi = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij} - \hat{A}_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij} - \bar{A})^2} \times \frac{nm}{(nm) - k}} \qquad (14)$$

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.

Testing criterion: the $\psi(k) = (k)$ function can vary from zero to infinity, with the best fit approaching zero. A $\psi(k)$ 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 $\psi(k)$ value, because this means the fit is twice as good as guessing the overall mean for each data point. Using this reasoning $\psi(k) = 0.3$ can be considered a fair correlation, $\psi(k) = 0.2$ can be considered a good correlation and $\psi(k) = 0.1$ an excellent correlation. It means that for $\psi(k) < 0.1$ the corresponding k can be taken as the number of significant components in solution.

2.3.4. Scree test, RPV

The scree test [1,24] 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

$$\mathrm{RV}(k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij} - \hat{A}_{ij})^2}{nm} \qquad (15)$$

which is equal to the square of the RMS(k) error. The residual percent variance can be expressed as a percentage

$$\mathrm{RPV}(k) = 100\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij} - \hat{A}_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}A_{ij}^2}\right) \qquad (16)$$

In terms of the eigenvalues of the data matrix, this expression can be converted to

$$\mathrm{RPV}(k) = 100\left(\frac{\sum_{a=k+1}^{m} g_a}{\sum_{a=1}^{m} g_a}\right) \qquad (17)$$

Testing criterion: when the residual percent variance is plotted against the number of k PC dimensions used in the data reproduction, $\mathrm{RPV}(k) = f(k)$, 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.

2.3.5. Imbedded error, IE

The imbedded error function (IE) [1,4] 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

$$\mathrm{IE}(k) = \sqrt{\frac{k\sum_{j=k+1}^{m} g_a}{nm(q-k)}} \tag{18}$$

which is equivalent to $\sqrt{k/m}$ RSD(k) and represents a measure of the difference between reconstructed and pure data and describes this part of errors which remains in reconstructed data.

Testing criterion: the behavior of the IE(k) function, as k varies from 1 to q, can be used to deduce the true dimensionality of the data. The IE(k) 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, IE(k) 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. Imbedded error can also be related to RMS(k) and RPV(k).

2.3.6. Factor indicator, IND

The factor indicator function IND(k) [1,4] is an empirical function which appears more sensitive than the IE(k) function to identify the true dimensionality of an absorbance data matrix. The function is composed of the same components as the IE(k) function, and is defined by

$$\mathrm{IND}(k)=\frac{\sqrt{\sum_{j=k+1}^{q} g_j/(n(m-k))}}{(q-k)^2}=\frac{\mathrm{RSD}(k)}{(q-k)^2} \tag{19}$$

where RSD(k) is the residual standard deviation of absorbance.

Testing criterion: this function, like the IE(k) function, reaches a minimum when the correct number of latent variables or r 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(k) function exhibits no minimum.

2.3.7. F-test

Malinowski [1,4] developed a test for determining the true dimensionality of a data set based on the Fisher variance ratio test, F-test. The F 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 [25]. 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: $g_a^{\mathrm{red}} = \bar{g}_a^{\mathrm{red}}$ versus alternative H_A: $g_a^{\mathrm{red}} > \bar{g}_a^{\mathrm{red}}$ is formulated. In case of validity of null hypothesis the test criterion

$$F(1,q-k) = \frac{\sum_{a=k+1}^{m}(n-a+1)(m-a+1)}{(n-k+1)(m-k+1)} \times \frac{g_a}{\sum_{a=k+1}^{q} g_a} \tag{20}$$

with 1 and $q-k$ degrees of freedom is applied.

Testing criterion: when testing the k is varied from the smallest eigenvalues in range $m-1, m-2, \ldots, 1$. The first kth reduced eigenvalue for which it is valid that $F(1, q-k)$ is greater than critical value for given significance level is taken as the smallest and corresponding k represents the number of significant components in a mixture.

2.3.8. Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA, RESO

A recommended procedure for determining the number of components in mixtures using RESO [6] contains the principal components analysis for the measured spectra set using the SVD algorithm to find the eigenvalues g_i^0 corresponding to ordinary PCA. Then the smoothed principal component analysis (SPCA) is applied to the same data set, doing the generalized eigenvalues problems $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{s}} = g_{a,i}^s(\boldsymbol{I} + a\boldsymbol{G})\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{s}}$ with different a. Details may be found in original paper describing RESO [6].

Testing criterion: calculate index RESO_i^a or the ratios between $g_{a,i}^s$ and g_i^0 for different a and plot $\log(\mathrm{RESO}_i^a)$ versus component number. Estimate the number of components by examining the $\log(\mathrm{RESO}_i^a)$ versus component number plots. Locate the number of $\log(\mathrm{RESO}_i^a)$s which are very close to each other and do not change substantially with the variation of k in comparison with the remaining $\log(\mathrm{RESO}_i^a)$s. This is the number of components existing in the mixture examined [6].

2.4. Signal-to-noise ratio SNR (or SER) and detection limit

In any simulation study of this type, the level of noise employed will be critical. Therefore, it is necessary to have a consistent definition of the SNR so that the impact of this parameter can be critically assessed. Traditional approaches to SNR are based on the ratio of the maximum signal to maximum noise value. As an alternative, the concept of instrumental error was again employed and the SER is defined where for an error the instrumental standard deviation of absorbance, $s_{\mathrm{inst}}(A)$ is used.

Attention should be paid to the methods' ability to detect a minor component in the presence of major ones. The detection limit is equivalent to the amount of 'detectable impurity' or the smallest relative concentration of the minor component. Approaching the detection limit, no methods can accurately determine the minor component in mixture. The detection limit depends on several factors, such as (i) spectral similarity of the minor component with other ones; (ii) instrumental resolution; (iii) noise level and noise type, and (iv) SNR with respect to the minor component.

3. Experimental

The comparison of indices methods tested has been demonstrated by real data as well as simulated ones which were designed to cover some typical situations to access in experimental practice.

3.1. Procedure

All spectra evaluation and data simulation were performed in the S-Plus programming environment and the algorithm INDICES is available on internet, http://meloun.upce.cz/indices [27].

Most indices methods (Fig. 1) are functions of the number of PC(k)'s into which the spectral data usually are plotted against k, and when the PC(k) reaches the value of the instrumental error of spectrophotometer used, $s_{\mathrm{inst}}(A)$, the corresponding k represents the number of significant components in a mixture, $r = k$. The dependence $f(k)$ decreases steeply with increasing number PCs as long as the PCs are significant. When k is exhausted the indices fall off, some of indices even display a minimum. At this point $r = k$ for all indices except g for which $r = k + 1$ is valid. The indices values at this point can be predicted from the properties of the noise, which may be used as a criterion to determine r [3,8].

3.2. Simulated data sets

To investigate all statistical properties of absorbance data matrix which were designed to be quite similar to real experimental data and cover some typical situations of analytical practice, several data sets of absorption spectra were simulated for a three-components system in mixture: potassium bichromate, cobalt(II) sulphate and copper(II) sulphate, a mixture abbreviated {Cr–Co–Cu}. An absorbance matrix was created by multiplying absorptivity spectra of three components (Fig. 2a) by their simulated concentration profiles (Fig. 2b) to reach resulting absorbance. Each matrix data set contains n digitized spectra consisted of m digitized wavelengths. Random noise was added to the spectra by generating random numbers with a Gaussian distribution with mean 0 and standard deviation equal to the pre-selected noise level, $s_{\mathrm{inst}}(A)$, to reach an optioned SER value. Most of simulated spectra sets for examination of five factors (i.e. concentra-

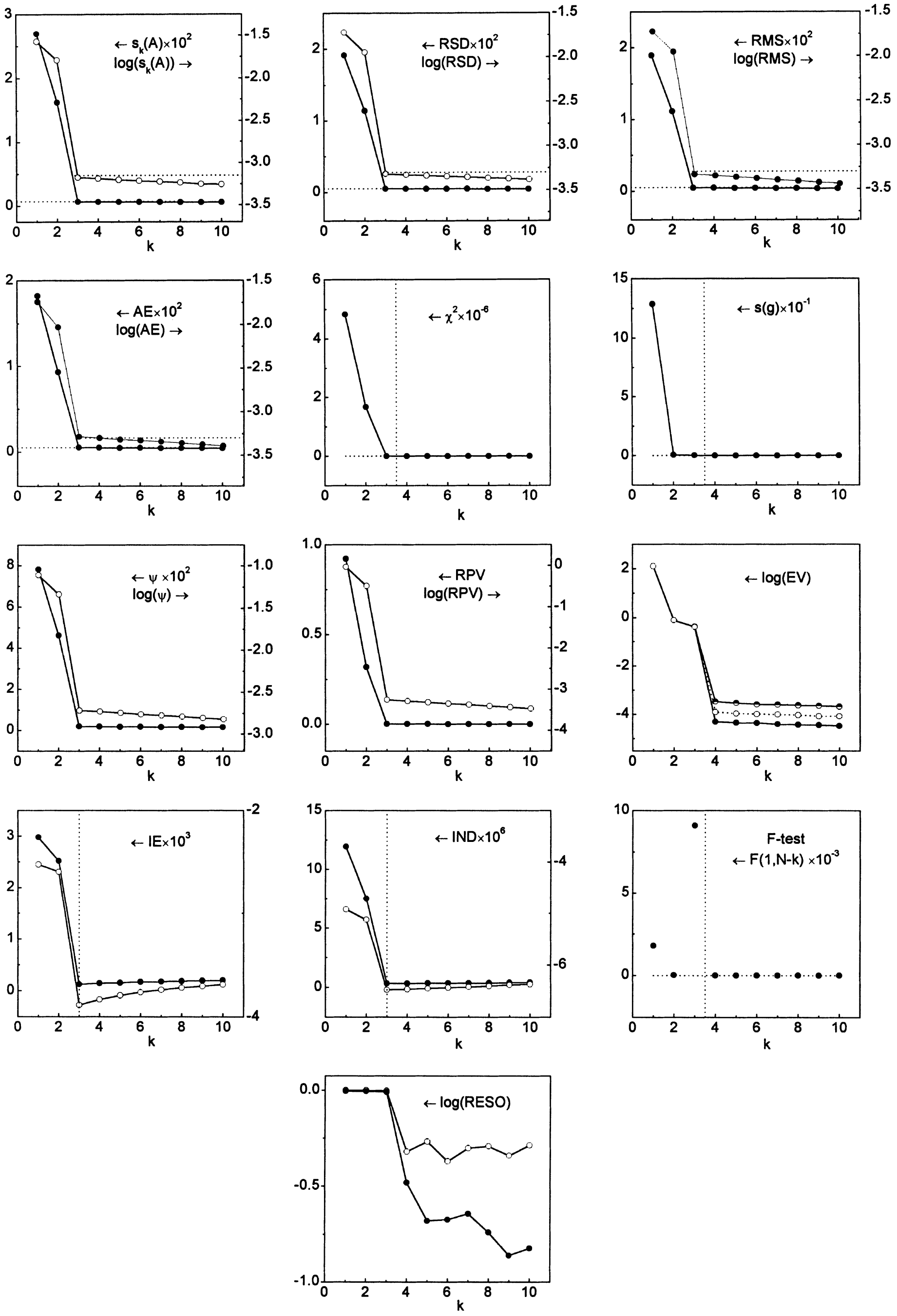

Fig. 1. The indices (full circles) and logarithm of the indices (empty circles) of 13 methods as a function of the number of principal components k for a simulated three-components system in mixture, potassium bichromate–cobalt(II) sulphate–copper(II) sulphate, with $r = 3, n = 82, m = 41$ and SER = 1570, S-Plus.

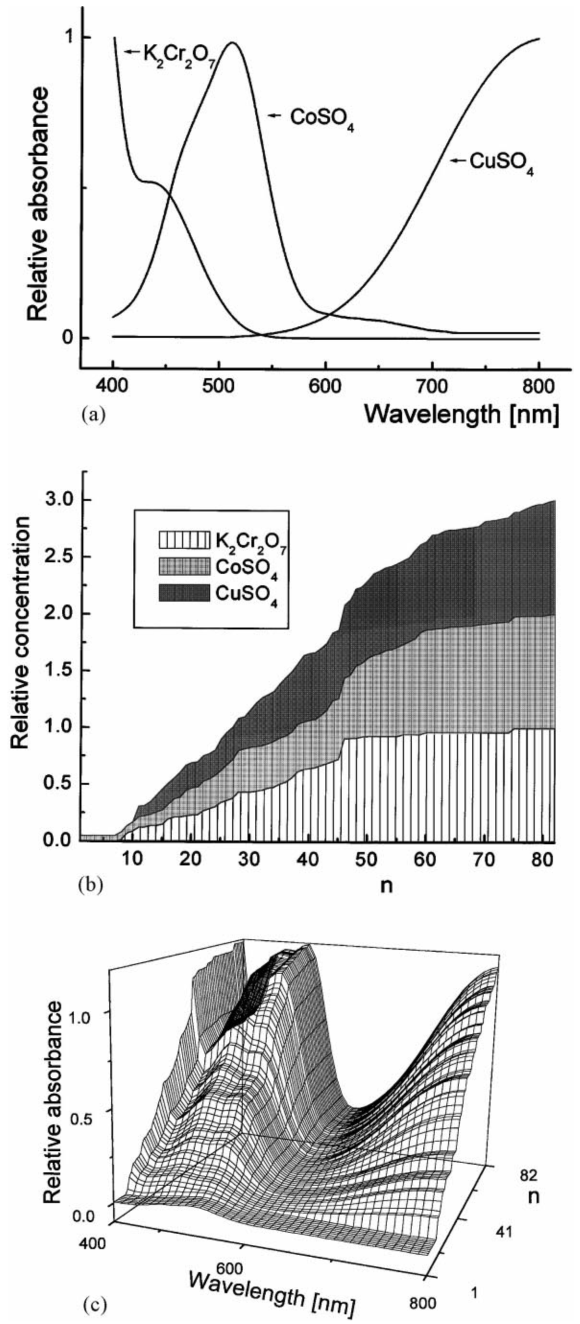

Fig. 2. (a) Spectra of relative absorbance for three components described as in Fig. 1. (b) Diagram of a relative concentration of three components in mixture for a simulated data set of three components described as in Fig. 1. (c) 3D-relative absorption spectra for a simulated data set of three components described as in Fig. 1.

tion, homoscedasticity noise, collinearity in spectra, heteroscedasticity noise and sample size), are of sample size $n = 82$ spectra and $m = 41$ wavelengths (Fig. 2c).

3.2.1. Concentration variation

One of the three components in system was arranged at different minor concentration levels, actually 0.25, 0.50, 1.0, 1.5 and 2.5% with respect to a sum of other two components which each had the same highest absorbance (Tables 1 and 2). For noise level $s_{\mathrm{inst}}(A) = 0.7\,\mathrm{mAU}$ the minor concentration levels cause SER values 4.0, 8.0, 15.9, 23.9 and 39.9.

3.2.2. Homoscedastic noise

To all previous absorbance matrices described in Section 3.2.1 the normally distributed homoscedastic random errors generated with zero expectation and different absorbance standard deviation $s_{\mathrm{inst}}(A) = 0.3$, 0.7, 1.4 and 2.8 mAU were added which represent about 0.03, 0.07, 0.14 and 0.28% of the maximum absorbance, respectively. These correspond to various levels of SER (Table 1), enabling to examine a detection limit of every indices method.

3.2.3. Heteroscedastic noise

The heteroscedastic noise with zero expectation was proportionally increased with increasing wavelength (a) from $s_{\mathrm{inst}}(A) = 0.01$ to 0.3 mAU which is about from 0.001 to 0.03% of the maximum absorbance, (b) from $s_{\mathrm{inst}}(A)$ 0.01 to 0.7 mAU which is about from 0.001 to 0.07% of the maximum absorbance, (c) from $s_{\mathrm{inst}}(A) = 0.01$ to 1.4 mAU which is about from 0.001 to 0.14% of the maximum absorbance, (d) from $s_{\mathrm{inst}}(A) = 0.01$ to 2.8 mAU which is about from 0.001 to 0.28% of the maximum absorbance added to precise values of a noiseless absorbance matrix. For comparison the concentrations of three components were used as described previously in Sections 3.2.1 and 3.2.2. These correspond to various levels of SER (Table 2), enabling to examine a detection limit of every indices method.

3.2.4. Collinearity in spectra

To examine collinearity in spectra, the spectrum s_r of system of three components produced according to the following equation with $p_2 = 1.1$, 2.0, 3.0, 5.0 and 8.0 and $p_3 = 1.1$, 3.0, 5.0, 8.0, i.e. $\boldsymbol{s_r} =$

Table 1
Search of a detection limit for 13 indices procedures proposing a number of components for simulated three-component system with various levels of homoscedastic noise level from 0.3 to 2.8 mAU and various concentrations of third minor component[a]

Noise (mAU)	SNR		SER		Precise methods						Approximate methods						
	All	Minor	All	Minor	$s_k(A)$	RSD	RMS	AE	χ^2	$s(g)$	ψ	RPV	g_a	IE	IND	F-test	RESO
Concentration of minor component 0.25%																	
0.3	932	2.3	3670	9.2	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
0.7	485	1.2	1594	4.0	2	2	2	2	2	2	2	2	2	2	2	2	2
1.4	226	0.6	802	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2
2.8	107	0.3	402	1.0	2	2	2	2	2	2	2	2	2	2	2	2	2
Concentration of minor component 0.5%																	
0.3	932	4.7	3670	18.3	**3**	**3**	**3**	**3**	4	2	**3**	**3**	2	**3**	**3**	**3**	**3**
0.7	485	2.4	1594	8.0	2	2	2	2	**3**	2	2	2	2	2	**3**	2	**3**
1.4	226	1.1	802	4.0	2	2	2	2	2	2	2	2	2	2	2	2	2
2.8	107	0.5	402	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2
Concentration of minor component 1%																	
0.3	932	9.3	3670	36.7	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.7	485	4.9	1594	15.1	**3**	**3**	**3**	**3**	2	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
1.4	226	2.3	802	8.0	2	2	2	2	2	2	2	2	**3**	2	**3**	2	**3**
2.8	107	1.1	402	4.0	2	2	2	2	2	2	2	2	2	2	**3**	2	2
Concentration of minor component 1.5%																	
0.3	932	13.8	3670	55.0	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.7	485	7.3	1594	23.9	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
1.4	226	3.4	802	12.0	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**
2.8	107	1.6	402	6.0	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
Concentration of minor component 2.5%																	
0.3	932	23.3	3670	91.7	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.7	485	12.1	1594	39.9	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
1.4	226	5.7	802	20.0	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
2.8	107	2.7	402	10.0	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	2	**3**	2	**3**
Estimation of detection limit for SER					10	10	10	10	24	10	10	10	8	16	4	12	6

[a] Bold digit means correct value found.

$(p_2\boldsymbol{s}_2+p_3\boldsymbol{s}_3)/(p_2+p_3)$, was added and $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$ are the vectors of spectra of component 2 and 3, respectively. For comparison, the relative concentrations of the three components in system were held constant 1, 1, 1, respectively, and the resulting spectrum was either (a) loaded with a homoscedastic noise level 0.7 mAU which is about 0.07% of the maximum absorbance, or (b) loaded with a heteroscedastic noise level varied in range from 0.01 to 0.7 mAU which is about from 0.001 to 0.07% of the maximum absorbance.

3.2.5. Sample size

Influence of a sample size or size of absorbance matrix on resulting number of estimated components r was investigated. The size of the absorbance matrix was decreased in several steps from $n \times m = 82 \times 41$ to a final size 10×10 (Table 3).

3.3. Real data sets

After determination of instrumental error of spectrophotometer used $s_{\text{inst}}(A)$ the real spectra of three components in mixture and protonation equilibria of a mixture of three sulphonephtaleins were investigated.

3.3.1. Instrumental error $s_{inst}(A)$

For determination of the instrumental error of spectrophotometer used, $s_{\text{inst}}(A)$, Wernimont–Kankare method [18] was applied. If there is one component

Table 2
Search of a detection limit for 13 indices procedures proposing a number of components for simulated three-component system with various levels of heteroscedastic noise level (a) from 0.01 to 0.3 mAU; (b) from 0.01 to 0.7 mAU; (c) from 0.01 to 1.4 mAU; (d) from 0.01 to 2.8 mAU and various concentrations of third minor component[a]

Noise [mAU]	SNR		SER		Precise methods						Approximate methods						
	All	Minor	All	Minor	$s_k(A)$	RSD	RMS	AE	χ^2	$s(g)$	ψ	RPV	g_a	IE	IND	F-test	RESO
Concentration of minor component 0.25%																	
(a)	1256	3.1	6215	15.5	**3**	**3**	**3**	**3**	2	2	**3**	**3**	2	2	**3**	**3**	**3**
(b)	433	1.1	2676	6.7	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
(c)	220	0.6	1380	3.5	2	2	2	2	2	2	2	2	2	2	2	2	2
(d)	159	0.4	684	1.7	2	2	2	2	2	2	2	2	2	2	2	2	2
Concentration of minor component 0.5%																	
(a)	1256	6.3	6215	31.1	**3**	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(b)	433	2.2	2676	13.4	**3**	2	2	**3**	2	2	**3**	**3**	2	2	**3**	**3**	**3**
(c)	220	1.1	1380	6.9	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
(d)	159	0.8	684	3.4	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
Concentration of minor component 1%																	
(a)	1256	12.6	6215	62.2	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(b)	433	4.3	2676	26.8	**3**	**3**	**3**	**3**	2	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(c)	220	2.2	1380	13.8	**3**	**3**	**3**	**3**	2	2	**3**	**3**	**3**	**2**	**3**	**3**	**3**
(d)	159	1.6	684	6.9	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
Concentration of minor component 1.5%																	
(a)	1256	18.8	6215	93.2	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(b)	433	6.5	2676	40.2	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(c)	220	3.3	1380	20.7	**3**	**3**	**3**	**3**	2	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(d)	159	2.4	684	10.3	2	2	2	2	2	2	2	2	**3**	2	**3**	**3**	**3**
Concentration of minor component 2.5%																	
(a)	1256	31.4	6215	155.4	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(b)	433	10.9	2676	67.1	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(c)	220	5.5	1380	34.5	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
(d)	159	4.0	684	17.1	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
Estimation of detection limit for SER					13.4	13.8	13.8	13.4	34.5	17.1	13.4	13.4	13.8	17.1	6.7	10.3	3.4

[a] Bold digit means correct value found.

Table 3
Number of components predicted by 13 indices for simulated three-component system for various size of absorbance matrix, homoscedastic noise level 0.7 mAU, SER = 15.7 and relative absorbance of all three components 1:1:0.02[a]

Matrix size ($n \times m$)	Precise methods						Approximate methods						
	$s_k(A)$	RSD	RMS	AE	χ^2	$s(g)$	ψ	RPV	g_a	IE	IND	F-test	RESO
10 × 10	2	2	2	2	**3**	2	2	2	2	7	5	2	**3**
12 × 10	**3**	**3**	**3**	**3**	**3**	2	2	2	**3**	7	**3**	**3**	**3**
20 × 10	**3**	**3**	**3**	**3**	4	2	2	2	**3**	7	**3**	2	**3**
20 × 20	**3**	**3**	**3**	**3**	4	2	**3**	2	**3**	**3**	**3**	–	**3**
40 × 20	**3**	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
41 × 41	**3**	**3**	**3**	**3**	4	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
82 × 41	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	3

[a] Bold digit means correct value found.

Table 4

Number of components predicted by 13 indices for simulated (first digit in each cell being valid for corresponding SER value) and experimental (second digit in each cell) spectral data of three-component system and various concentrations of third minor component c_3 (%) when for simulated data a homoscedastic noise level 0.7 mAU was used while for experimental data a value $s_{\text{inst}}(A) = 0.7$ mAU was found[a]

c_3	SER	Precise methods						Approximate methods[a]						
		$s_k(A)$	RSD	RMS	AE	χ^2	$s(g)$	ψ	RPV	g_a	IE	IND	F-test	RESO
0.5	7.9	2, 2	2, 2	2, 2	**3**, 2	2, **3**	2, 2	–, 2	–, 2	**3**, 2	2, 2	2, **3**	5, **2**	**3**, **3**
1.0	15.7	**2**, **3**	**3**, **3**	2–**3**, **3**	**3**, **3**	**3**, **2**	**3**, **2**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	5, **3**	**3**, **3**
1.5	23.6	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	5, **3**	**3**, **3**
2.5	39.3	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	**3**, **3**	5, **3**	**3**, **3**

[a] Bold digit means correct value found; –: means that a value cannot be estimated.

in the solution, this means that the true rank of the absorbance matrix is equal to one, $r = 1$, and the corresponding residual standard deviation of absorbance $s_k(A)$ being estimated from a graph $s_k(A) = f(k)$ for $k = 1$, and analogously, the real error RE(1), the extracted error XE(1) and the average error $\bar{e}$ for $k = 1$ were estimated: for potassium bichromate $s_1(A) = 0.7$ mAU, $RE(1) = 0.5$ mAU, $XE(1) = 0.5$ mAU and $\bar{e}(1) = 0.5$ mAU.

3.3.2. *Mixture of three components*

For a three-components system {Cr–Co–Cu}, the absorbance matrix of $n = 30$ spectra for various concentration combinations of three components {Cr–Co–Cu} according to Beer law at $m = 27$ wavelengths was examined (Table 4).

3.3.3. *Protonation equilibria of a mixture of three sulphonephtaleins*

A mixture of 2.45×10^{-4} M Bromocresol Green, 3.29×10^{-4} M Phenol Red and 1.48×10^{-4} M Thymol Blue in 0.01 M HCl was titrated with 1 M KOH using a microburette at 298 K and ionic strength 0.001 (KCl) to adjust pH value in range 2–11 with the use of OP271 pH meter (Radelkisz, Budapest) with resolution of 0.001 pH unit. The spectra were recorded on GBC UV–VIS 916 spectrophotometer (GBC Scientific Equipment Pty Ltd., Dandenong, Australia) with $s_{\text{inst}}(A) = 0.7$ mAU in a 0.2 cm cuvette. The cell for measuring pH contained a G202B glass electrode and saturated calomel electrode (both from Radiometer, Copenhagen) and was standardized against buffers from Radiometer: pH = 4.005, 6.865 and 9.180. The Bromocresol Green, Phenol Red and Thymol Blue and other chemicals used (Lachema, Brno) were of analytical grade.

4. Results and discussion

The number of significant components r can be estimated from indices by comparing them with the experimental error, using the noise level $s_{\text{inst}}(A)$ as a threshold. This is the common criterion to determine r for precise methods. However, there are experimental situations when information about noise $s_{\text{inst}}(A)$ is not available and such comparison cannot be made and then the approximate methods are used. The point where $k = r$ was estimated from the indices of the number of principal components a. Fig. 1 shows the indices as functions of the number of principal components k for one of the simulated data sets with $r = 3$, $n = 82$, $m = 41$ and SER = 1570. Due to the large variation of some index values, a logarithmic scale is often employed. The $s(g)$ and IE are functions of the kth PC, and should change substantially when $k + 1 = r$, while the other indices reflect the cumulated effect of the first k PCs and should change when $k = r$. In these plots for simulated data $r = 3$ and a change in slope can be seen around $k = r$ for $s_k(A)$, RSD, RMS, AE, χ^2, ψ, $\log g$, IND, F-test, RESO and at $k = r + 1$ for g and IE. On base of extensive simulations a comparison of 13 indices method was made among (i) six precise indices methods — $s_k(A)$, RSD, RMS, AE, χ^2 and $s(g)$, and (ii) seven approximate indices methods — ψ, RPV, g, IE, IND, F-test, and RESO. The effect of five factors, i.e. a concentration of the minor component, homoscedastic noise, collinearity in spectra, heteroscedastic noise and sample size was investigated and the detection of the limit of the true value of number of components of each method was estimated. The indices were used to analyze real experimental data and the derivation modifications SD and ROD were also used.

4.1. Homoscedastic noise, heteroscedastic noise and concentration of the minor component

All three factors may be examined commonly using an effective resolution criterion, the SNR or the SER. Both criteria cover all three factors and therefore can be used as the common resolution factor. For simulated data sets with $r = 3$, $n = 82$, $m = 41$ there were adjusted various SER values of homoscedastic noise (Table 1). It is obvious that when SER is equal or higher than a detection limit, every index method fails. Table 1 demonstrates an estimate of detection limit for individual methods in case of homoscedastic noise: SER = 24 for χ^2, SER = 16 for IE, SER = 12 for F-test, SER = 10 for $s_k(A)$, RSD, RMS, AE, $s(g)$ and ψ, SER = 8 for g_a, SER = 6 for RESO, SER = 4 for IND. It means that for determination of the minor component two methods, RESO and IND work best and appear to be the most reliable. It is worth mentioning that most of the methods do not behave in the same way if the SER criterion decreases nearly to their detection limit. Indices $s(g)$, χ^2 and F-test are definite, they are fully based on statistic criterion and it is not complicated to predict r. The situation is simple in case of precise methods. Here we take the k for which the value of criterion is closest to the value of experimental error, $s_{\text{inst}}(A)$. However, we lose a help function of the curve shape being used here as an efficient criterion. Thus the indices ψ, and RPV which are based on detecting a break-point on the curve, are not very reliable in prediction of r in case of decreasing a SER. The g_a and RESO are reliable enough.

For heteroscedastic noise (Table 2) an estimate of detection limit for individual methods are: SER = 34.5 for χ^2, SER = 17.1 for $s(g)$ and IE, SER = 13.4–13.8 for $s_k(A)$, RSD, RMS, AE, RPV, g_a and ψ, SER = 10.3 for F-test, SER = 6.7 for IND and SER = 3.4 for RESO. Once again RESO and IND work best, which demonstrates their ability in the presence of heteroscedastic noise. A possible explanation [6] might be that heteroscedastic noise does affect the eigenvalues of both SPCA and PCA but has no influence on their ratios.

4.2. Collinearity in spectra

Even severe collinearity in spectra that arranged all the indices predicted a correct number of components in mixture.

4.3. Sample size

Decreasing size from $(n \times m) = (82 \times 41)$ to (40×20) all indices found correct value of the number r. Decreasing size from (40×20) to (20×20)

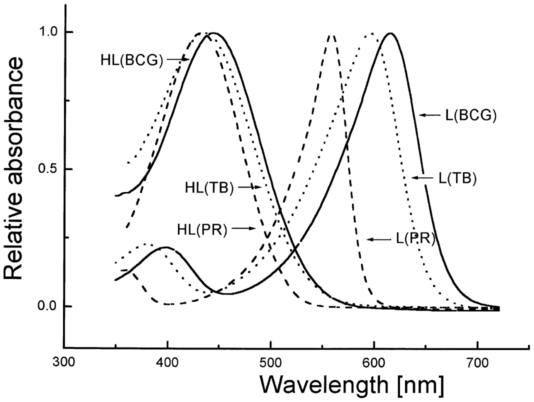

Fig. 3. Spectra of relative absorbance for six components, i.e. the protonated form HL and anion L^- of Bromocresol Green + Phenol Red + Thymol Blue in mixture with $r = 6, n = 33, m = 31$ and SER about 2780, S-Plus.

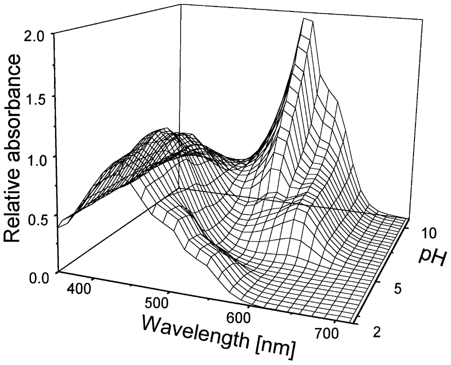

Fig. 4. 3D-relative absorption spectra for six components described as in Fig. 3.

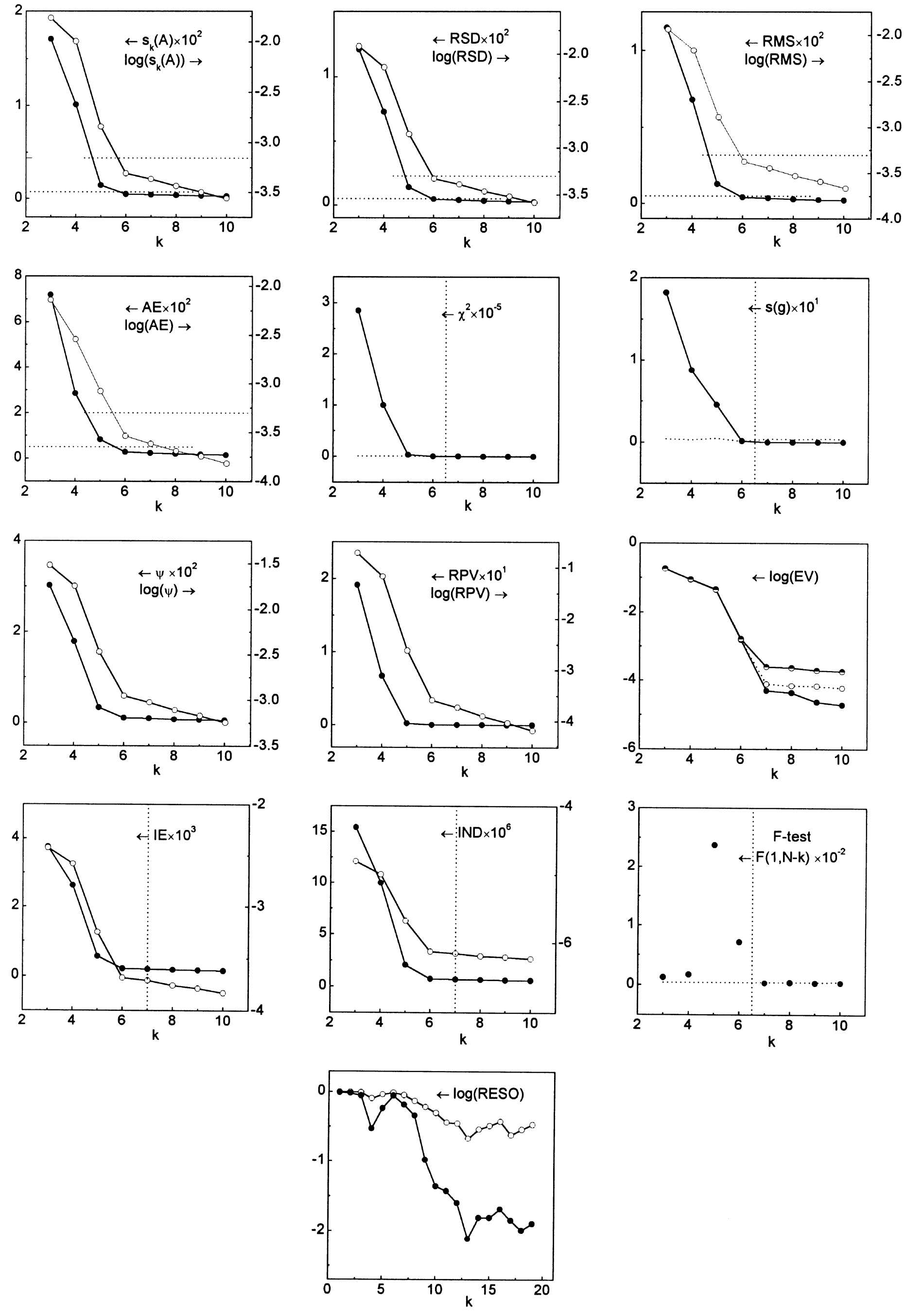

Fig. 5. The indices (full circles) and logarithm of the indices (empty circles) of 13 methods as a function of the number of principal components k for a experimental data set of 6 components described as in Fig. 3.

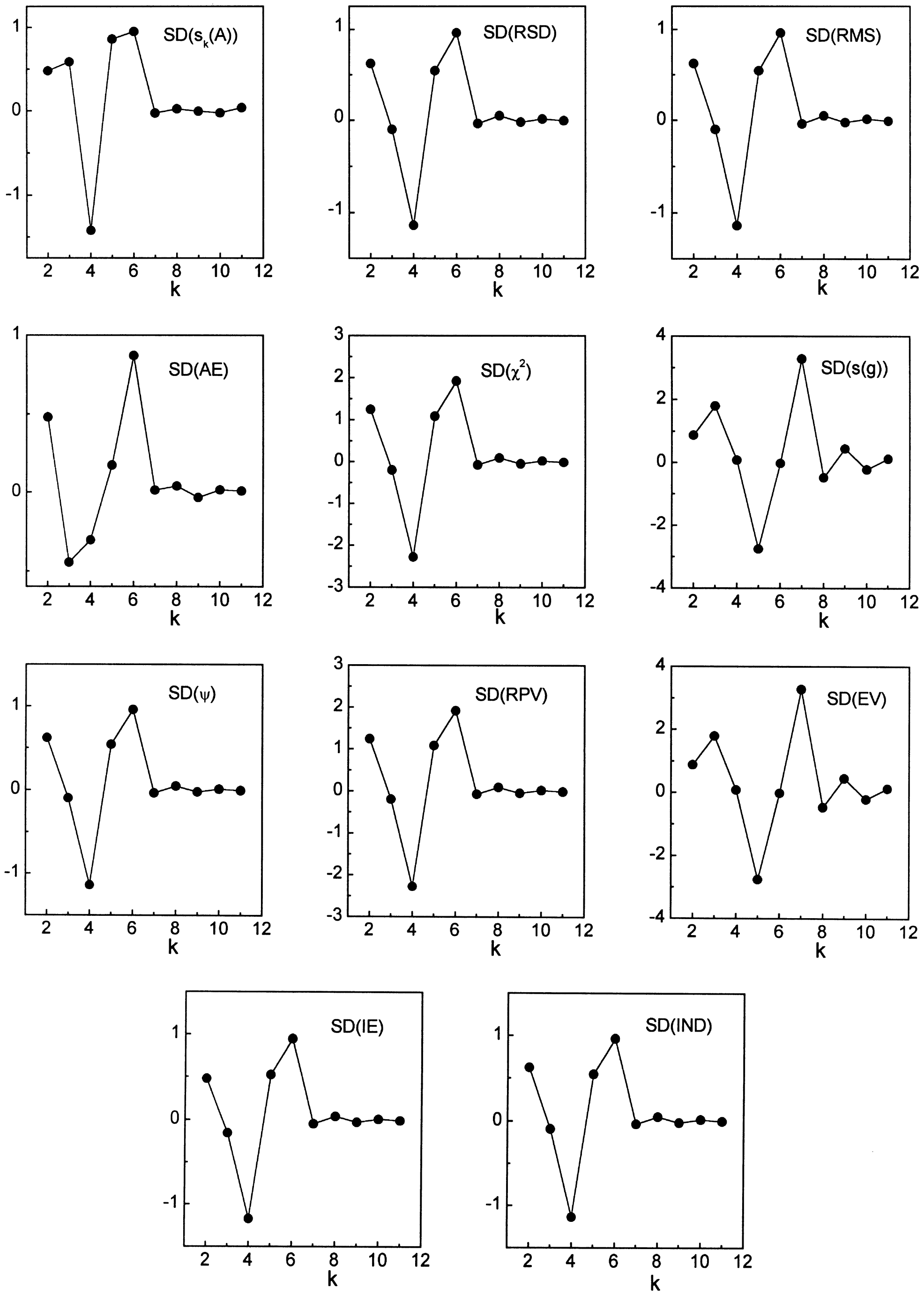

Fig. 6. The second derivative detection criterium applied on 11 indices methods described as in Fig. 3.

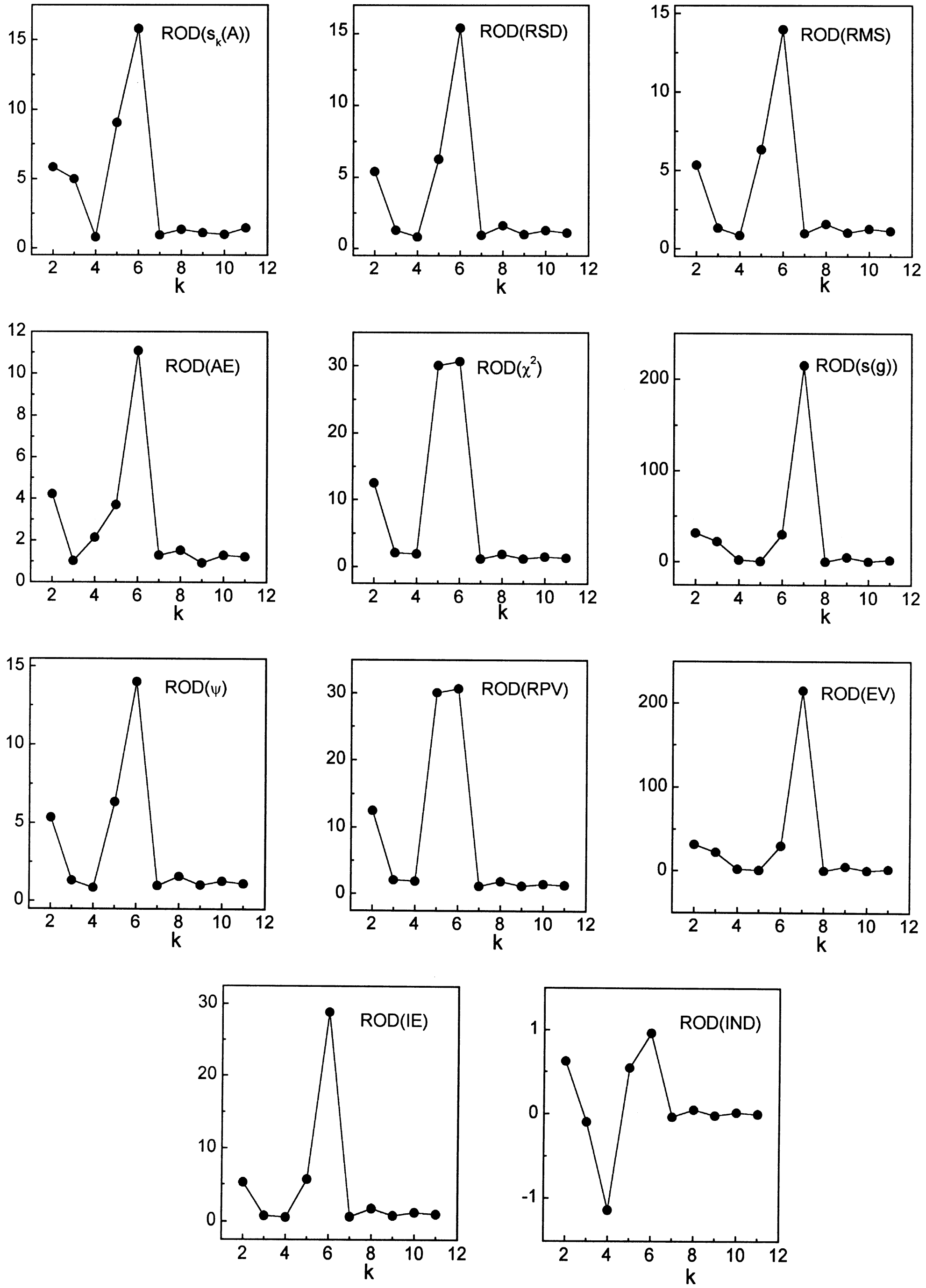

Fig. 7. The ratio of derivatives detection criterium applied on 11 indices methods described as in Fig. 3.

RPV, χ^2, F-test and $s(g)$ methods fail to find the correct value of the number r (Table 3). When $n = 12$ most approximate indices fail, the correct value of number r is estimated by precise methods only. Some methods are more sensible to the changes of size of the absorbance matrix. This is especially the case of RESO. On the other hand, precise methods are sensitive neither to the decreasing size nor to the incorrect dimension of data, i.e. highest number of wavelength in columns than spectra in rows. In our study it was found that sufficient size of absorbance matrix seems to be about 30 spectra and a little bit smaller number of wavelengths. In an agreement with Elbergali et al. [8] we can conclude that a higher number of data points collected in a given wavelength range improve ability of the indices to predict the number of light absorbing components. Recorded spectra should be digitized into the maximum number of data points, especially for data set with low SER value and many components.

4.4. Real experimental data

The Wernimont–Kankare procedure estimates the instrumental standard deviation of spectrophotometer used, $s_{\mathrm{inst}}(A) = 0.7\,\mathrm{mAU}$ in range 380–650 nm. This value can be used for a prediction of SER value for experimental data. Only two indices, IND and RESO, predict correct number of components for a concentration of minor component 0.5% (corresponding to SER $= 7.9$). Except χ^2 and $s(g)$ all other indices predict correct number for a concentration 1% (corresponding to SER $= 15.7$) and χ^2 and $s(g)$ for 1.5% (corresponding to SER $= 23.6$) (Table 4). The extensive simulations and experimental data treatment showed that most of the indices accurately predict the number of components that contribute to a set of absorption spectra. For the simulated spectra all indices performed well, all being absolutely correct for data sets with SER of at least 10. For data with higher noise the g_a, RESO and IND performed best.

To examine severe collinearity in spectra and ill-conditioned model of similar spectra for individual components, an experimental system of protonation equilibria of three sulphonephthaleins in mixture, Bromocresol Green + Phenol Red + Thymol Blue (BCG + PR + TB) was tested. Protonated forms HL of all three sulphonephtaleins are nearly of same colour and their $\lambda_{\max}$ of their absorption bands are very close. This is valid also for their anions L^- (Fig. 3). No sulphonephtalein was in minor concentration and a concentration ratio BCG:PR:TB = 1:1:1 was used. For various values of pH 2–11, the spectra of a protonation equilibria of a mixture BCG + PR + TB was monitored and digitized into absorbance matrix of $n = 33$ at $m = 31$ wavelengths with SER about 2780 (Fig. 4). Most indices on Fig. 5 indicated correct number of six component present even an ill-defined problem of very similar spectra of individual components was solved.

Elbergali et al. [8] proposed a modification of index methods using derivatives to improve identification of the number of components. The derivative criteria are based on the point where the slope changes and reaches a maximum (Fig. 5). Even the second derivative criterion SD(k) function for some indices had about the same magnitude for two successive points around $k = r$ we can say that except two indices $s(g)$ and g_a, all indices predicted the correct number of components, $r = 6$ (Fig. 6). In the case of the ratio of derivatives criterion ROD(k) the same phenomenon for only indices χ^2 and RPV is valid. For $s(g)$ and g_a the correct value r is $r = k - 1$ for SD(k) and also for ROD(k) plot (Fig. 7).

5. Conclusion

Two indices, RESO and IND, are stable in many situations and correctly predict a minor component in a mixture even if its relative concentration is about 0.5–1% relatively to remaining components. Both can detect minor components and solve the ill-defined problem with severe collinearity in spectra. Most indices predict the correct number of components for data sets with the SER of at least 10 but RESO and IND of at least 6. For more than four components in the mixture, the modification of Elbergali et al. seem to be useful resolution tool enabling the correct prediction of the number of components in spectra for all indices except $s(g)$ and g_a. The Wernimont–Kankare procedure is a reliable method for determination of the instrumental standard deviation of spectrophotometer used. In case of real experimental data the RESO, IND and methods based on knowledge of instrumental error should be preferred.

Acknowledgements

Financial support of Grant Agency of the Czech Republic (Grant No 303/00/1559) is thankfully acknowledged. Karel Kupka is acknowledged for his help with INDICES written in S-Plus.

References

[1] E.R. Malinowski, Factor Analysis in Chemistry, 2nd Edition, Wiley, New York, 1991.

[2] M. Meloun, J. Havel, E. Högfeldt, Computation of Solution Equilibria, Ellis Horwood, Chichester, 1988.

[3] E.R. Malinowski, J. Chemom. 13 (1999) 69.

[4] E.R. Malinowski, Anal. Chem. 49 (1977) 612.

[5] J.M. Deane, H.J.H. MacFie, J. Chemom. 3 (1989) 477.

[6] Z.-P. Chen, Y.-Z. Liang, J.-H. Jiang, Y. Li, J.-Y. Qian, R.-Q. Yu, J. Chemom. 13 (1999) 15.

[7] Z.-P. Chen, J.-H. Jiang, Y. Li, H.-L. Shen, Y.-Z. Liag, R.-Q. Yu, Anal. Chim. Acta 381 (1999) 233.

[8] A.K. Elbergali, J. Nygren, M. Kubista, Anal. Chim. Acta 379 (1999) 143.

[9] J.M. Dean, Data reduction using principal components analysis, in: R.G. Brereton (Ed.), Multivariate Pattern Recognition in Chemometrics Illustrated by Case Studies, Elsevier, Amsterdam, 1992.

[10] A.K. Elbergali, R.G. Brereton, Chemom. Intell. Lab. Syst. 27 (1995) 55.

[11] Y.-Z. Liang, O. Kvalheim, A.M. Rahmani, R.G. Brereton, J. Chemom. 7 (1993) 15.

[12] S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström, in: Proceedings of the IUFOST Conference, Food Research and Data Analysis, Applied Science Publishers, London, 1983, p. 147.

[13] M.A. Saraf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley, Chichester, 1986.

[14] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester 1989.

[15] D.R. Cox, D. Oakes, Analysis of Survival Data, Chapman & Hall, London, 1984.

[16] D.L. Massart, R.G. Brereton, R.E. Dessy, P.K. Hopke, C.H. Spiegelman, W. Wegscheider (Eds.), Chemometrics Tutorials, Elsevier, Amsterdam, 1990.

[17] D.L. Massart, W. Wegscheider, B.G. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1990.

[18] J.J. Kankare, Anal. Chem. 42 (1970) 1322.

[19] M.S. Bartlett, Br. J. Psych. Stat. Sec. 3 (1950) 77.

[20] Z.Z. Hugus Jr., A.A. El-Awady, J. Phys. Chem. 75 (1971) 2954.

[21] T.M. Rossi, I.M. Warner, Anal. Chem. 54 (1986) 810.

[22] H.F. Kaiser, Educ. Psych. Meas. 20 (1966) 141.

[23] J.H. Kindsvater, P.H. Weiner, T.J. Klingen, Anal. Chem. 46 (1974) 982.

[24] R.D. Catell, Multivariate Behav. Res. 1 (1966) 245.

[25] E.R. Malinowski, J. Chemom. 1 (1987) 49.

[26] M. Meloun, P. Miksik, Karel Kupka, Critical comparison of factor analysis methods in spectra analysis determining the number of light-absorbing species, in: Proceedings of the XIIIth Seminar on Atomic Spectrochemistry, Podbánské, September 1996, pp. 307–329, ISBN 80-967325-7-9.

[27] Algorithm INDICES: http://meloun.upce.cz/indices.