RELIABILITY OF CHEMOMETRICAL DATA ANALYSIS IN SPECTROSCOPY

M. Meloun

Department of Analytical Chemistry, University of Chemical Technology,
532 10 Pardubice, Czech Republic,

Key words: Calibration precision, Detection limit, Determination limit, Critical level, Principal component analysis; PCA; Decomposition of absorbance matrix; Spectra analysis; Rank of absorbance matrix; Error of spectrophotometer; Number of components; Chemometrics;

The total uncertainty of some analytical quantity (the concentration, the content, etc.) is a result of a law of propagation of all kinds of errors or uncertainties concerning various experimental and instrumental operations. Calibration consists from two steps, a construction of calibration model and an inversion of calibration model when from measured signal y^* (e. g. absorbance) the unknown concentration x^* including its confidence interval is estimated. The number of chemical components in a mixture represents the first step for further qualitative and quantitative analysis in all forms of spectral data treatment. Reliability of various methods for estimation of the number of components that contribute to spectra was critically tested.

1. Reliability of Calibration Model Building

Constructing calibration model $g(x)$ a relation between the measured quantity y called a signal (potential, electric current etc.) and the quantity x called a response of system or a property (content, composition, concentration, temperature etc.) being more difficult monitored is created. Calibration types can be classified in many different ways: *Univariate-multivariate calibration* or *Linear-Nonlinear calibration* or *Selection- weighting (full spectrum) calibration* or *Direct-Indirect calibration* or *Forward (inverse)-Reverse (classical) calibration.* In order to understand the fundamental calibration problems that have to be solved, the user may benefit from studying various other methods as well.

In calibration experiment for n samples with known (or precisely adjusted) responses x_i the corresponding signal values y_i are measured. The additive model of measurements is supposed $y_i = g(x_i, \beta) + \epsilon_i$ is used. Here β is a set of adjustable parameters and ϵ_i are measurements errors which are normally distributed with constant variance σ^2. In calibration from the measured signal y^* (e. g. absorbance) the unknown concentration x^* including its confidence interval is estimated,

$$y_i = \beta_2 + \beta_1 x + \varepsilon_i, \quad i = 1, ..., n \qquad \text{and}$$

$$y_j^* = \beta_2 + \beta_1 \varkappa^* + \varepsilon_j^*, \quad j = 1, ..., M$$

In the first phase the suitable model $g(x)$ is selected and parameters β are estimated from data $\{x_i, y_i\}$ by regression analysis. For linear models this task leads to solution of linear equation system. For nonlinear regression models the minimization algorithms must be used. In the second phase a calibration model $g(x, \beta^*)$ is used that for measured value of signal y^* the mean value of response x^* and the corresponding confidence interval is estimated. Generally, the response x^* is the formal solution of equation $x^* = g^{-1}(y^*)$. On the base of the Taylor series expansion the approximate formula for variance $D(x^*)$ may be found in the form

$$D(\hat{x}^*) \approx \frac{\sigma^2}{b_1^2}\left[\frac{1}{M} + \frac{1}{n} + \frac{(y^* - \bar{y})^2}{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}\right].$$

(a) **Linear Calibration Models:** in case of linear models both steps use the calibration straight line. Besides the direct estimate $\hat{x}^*$

$$\hat{x}^* = \bar{x} + \frac{y^* - \bar{y}}{b_1}$$

where y* is the measured signal (or the average y^* for $M > 1$ repeated measurements, respectively) and b_1 is the estimate of the slope. This estimate is generally biased and a correction is made by Naszodi modified estimate,

$$\hat{x}_B^* = \bar{x} + \frac{(y^* - \bar{y})\, b_1}{b_1^2 + \dfrac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Kruchkoff proposed the inversion estimate

$$x_I^* = \bar{x} + (y^* - \bar{y})\, \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

and Schwartz proposed the nonlinear estimate

$$\hat{x}_N^* = \frac{\sum_{i=1}^{n} x_i \exp\left[\dfrac{-(y^* - b_2 - b_1 x_i)^2}{2\,\hat{\sigma}^2}\right]}{\sum_{i=1}^{n} \exp\left[\dfrac{-(y^* - b_2 - b_1 x_1)^2}{2\,\hat{\sigma}^2}\right]}$$

Difficulties of calibration task depend on calibration model used. For highly nonlinear model functions which cannot be sufficiently approximated by Taylor expansion is the variance $D(\hat{x}^*)$ biased. The generally non-symmetric distribution of quantity $\hat{x}^*$ brings difficulty. The only case of calibration line and small residual variance enables to consider a distribution of $\hat{x}^*$ as a normal one. The corresponding 95%-th confidence interval has the lower L_L and upper L_U limits

$$L_{L,U} = \hat{x}^* \mp t_{1-\alpha/2,\, n-2}\, \frac{\hat{\sigma}}{|b_1|} \sqrt{\frac{1}{n} + \frac{(y^* - \bar{y})^2}{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

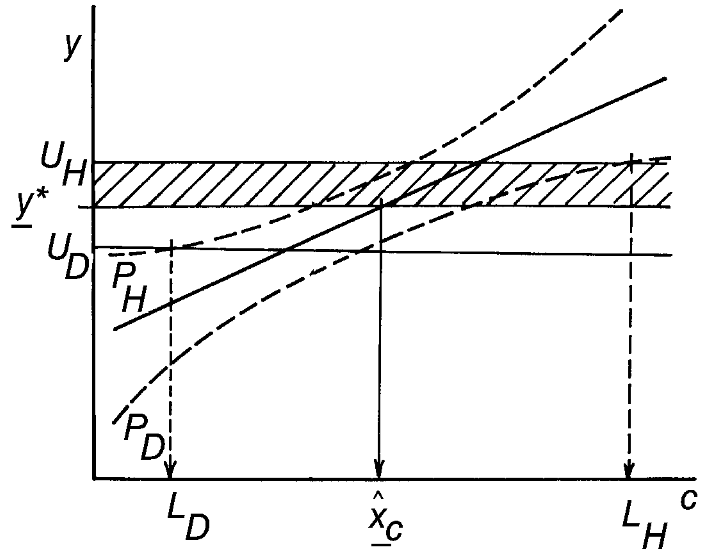

Fig. 1 Determination of the confidence interval of parameter *x* for a calibration straight line. The confidence interval of the signal is indicated by the hatched area.

(b) **Nonlinear Calibration Models**: models as polynomials which are highly nonlinear in variable x and therefore can be suitable for construction of nonlinear calibration models. Polynomials as linear regression models are attractive mainly from point of view of an application of calibration models for response computation and statistical analysis. Except to experimental data (x_i, y_i) $i = 1, ..., n$, the another set of knots are determined t_j, $j = 1, ..., k$. Knots form the boundaries of intervals in which individual piecewise function are defined. In each interval I_j bounded by knots t_{j-1}, t_j is calibration function expressed by the model $g_j(x)$.

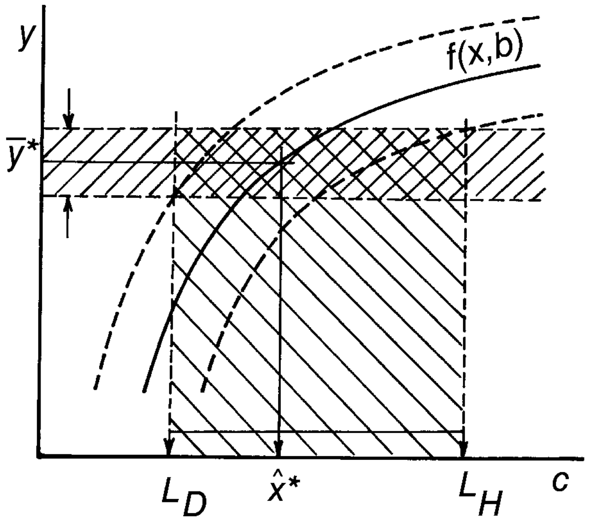

Fig. 2 Procedure for determination of concentration x^* for the mean value of signal $\bar{y}^*$.
L_L and L_U are the lower and upper limits of the confidence interval of concentration

A quality of approximation is here dependent on a number and location of individual knots t_j, a form of the function $g_j(x)$ and on the class C^m from which the calibration model $g(x)$ comes. Denote that function $g(x)$ of the class C^m is continuous up to first m derivatives. Special type of piecewise polynomial functions are splines. Spline $S_{m+1}(x)$ are function of class C^m which are defined as a local polynomials of maximal degree $(m + 1)$. For calibration purposes the quadratic splines $S_1(x)$ which are continuous and smooth (continuous in first derivative) are suitable. Splines $S_1(x)$ can be simply defined as a truncated polynomials

$$S_1(x) = ß_1 + ß_2 x + ß_3 x^2 + \Sigma\, ß_{j+3}(x - t_j)_+^2$$

here $(x)_+^2 = x^2$ for $x > 0$ and $(x)_+^2 = 0$ for $x \le 0$. For known t_j the $S_1(x)$ is the linear regression model. Flexibility of regression splines may be achieved by selection of knots t_i. In program

(c) **The precision of calibration**: the limiting values of the concentration for which the measurement signal is still significantly different from the noise are defined by the three levels of signal:

1. The *critical level* y_c represents the upper limit of the 100(1 - α)% confidence interval of the predicted signal from the calibration model for the concentration equal to zero, i. e. the *blank measurement*. The signals above y_c are significantly different from the noise,

$$y_c = \bar{y} - b_1 \bar{x} + t_{1-\alpha,\, n-2}\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

The concentration x_c corresponding to this critical level y_c is determined from the calibration model from

$$x_c = \frac{y_c - \bar{y}}{b_1} + \bar{x}$$

2. The *detection limit* y_D corresponds to the concentration for which the lower 100(1 - α)% confidence interval of signal prediction from the calibration model is equal to y_c. For the linear calibration model we have

$$y_D = y_c + \hat{\sigma}\, t_{1-\alpha,\, n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_D - \bar{x}^2)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

The detection limit gives the lowest true signal level which still permits detection. The quantity x_D gives the minimum concentration which can be distinguished from zero with probability (1 - α).

3. The *determination limit* is the smallest signal level for which the relative standard deviation of prediction from the calibration model is sufficiently small and equal to the number C, where C = 0.1, usually. If the predicted value at point x_s is given $y(x_s) = \bar{y} + b_1(x_s - \bar{x})$ and the condition of determination y_s is then equal to

$$\frac{\sqrt{D(y(x_s))}}{\hat{y}(x_s)} = C$$

Substitution and rearrangement leads to the expression

$$y_s = \frac{\hat{\sigma}}{C} \sqrt{1 + \frac{1}{n} + \frac{(x_s - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and, in practice, in the chemical laboratory, an approximation is used, as follows

$$y_s \approx \frac{\hat{\sigma}}{C}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

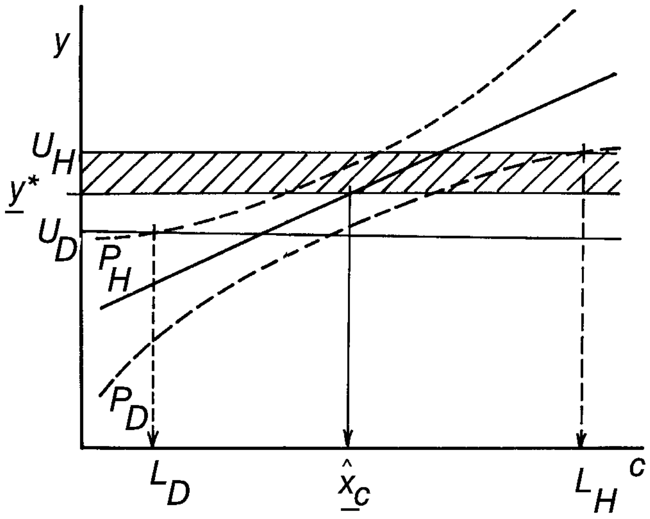

Fig. 3 Definition of the critical level y_c, the detection limit $y_D$and their corresponding concentrations x_c and x_D.

The corresponding concentration x_D is calculated from

$$x_D = \frac{y_D - \bar{y}}{b_1} + \bar{x}$$

The corresponding concentration x_s is given by $x_s = \frac{y_s - \bar{y}}{b_1} + \bar{x}$. Generally, it is valid that $y_c \leq y_D \leq y_s$.

(d) **The Procedure for Calibration Model Building:** the procedure consists of following steps:

(1) *Proposed model:* starting from the simplest model, models of higher power are build.

(2) *Exploratory data analysis:* using various diagnostic graphs the multicollinearity, heteroscedasticity, autocorrelation, normality of errors, influential points - outliers and extremes are investigated.

(3) *Parameter estimation:* using the least squares or the rational ranks, the best estimates of unknown regression parameters are determined. Statistical test of significance of each parameter follows. Quality of regression performed is examined by following regression characteristics: the mean error of prediction, the Akaike information criterion, the determination coefficient, the predicted determination coefficient, the standard deviation of prediction.

(4) *Regression diagnostics:* identification of influential points and examination of assumption of least squares method. Examination of the regression triplet (data, model, method).

(5) *Construction of improved model:* estimation of parameters of improved model using various modifications of the least-squares method.

(6) *Quality of calibration model:* estimation of the model parameters and the detection limit,

the determination limit and the critical level.

(7) *Determination of unknown concentration:* point and interval estimates of unknown concentration x^*.

2. Reliability of the Number of Componets in Spectra Analysis:

Procedures for determining the chemical rank of a matrix concerning a variety of empirical and statistical methods based on principal component analysis PCA have been reported. 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 the resolved spectra. This is an important aspect to consider when using these methods.

The *n* × *m absorbance data matrix* $\boldsymbol{A} = \boldsymbol{E}\,\boldsymbol{C}$ contains the *n* recorded spectra as rows being written as the product of the *m* × *r matrix of molar absorptivities* ***E*** and the *r* × *n concentration matrix* ***C***. Here *m* denotes the number of wavelengths for which each spectrum was recorded being equal to the number of columns of ***A*** matrix, *n* is the number of solutions for which spectra have been recorded being equal to the number of rows of ***A*** matrix, and *r* is the number of components that absorb in the chosen spectral range. The rank of the matrix ***A*** is obtained from the equation

$$\mathrm{rank}(\boldsymbol{A}) = \min\,[\mathrm{rank}(\boldsymbol{E}), \mathrm{rank}(\boldsymbol{C})] \le \min\,(m, r, n)$$

Since the rank of ***A*** is equal to the rank of ***E*** or ***C***, whichever is the smaller, and since $\mathrm{rank}(\boldsymbol{E}) \le r$ and $\mathrm{rank}(\boldsymbol{C}) \le r$, then provided *m* and *n* are equal to or greater than *r*, it will only be necessary to determine the rank of matrix ***A*** which is equivalent to the number of dominant components. 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, $s_{inst}(A)$ before a statistical examination. (b) *Approximate methods* requiring no knowledge of the instrumental error of the absorbance data, $s_{inst}(A)$. Many of these methods are empirical functions.

A critical comparison of various PCA methods on both simulated and experimental data is performed.

Application of precise methods: Determination of a number of light-absorbing components in mixture is based on a comparison of an actual index of method used with the experimental error of instrument used, $s_{inst}(A)$.

Residual standard deviation, $s_k(A)$*:* Kankare uses the second moment ***Z*** of an absorbance matrix ***A***. Applying eigenvalues g_j of matrix ***Z*** *the residual standard deviation of absorbance* $s_k(A)$ is estimated

$$s_k(A) = \sqrt{\frac{tr(\boldsymbol{Z}) - \sum_{j=1}^{k} g_j}{n - k}}$$

where $tr(\boldsymbol{Z})$ is a trace of the matrix ***Z*** and *r* is the estimated number of components in a mixture. The values $s_k(A)$ for different number of components *k* are plotted against an integer *k*, $s_k(A) = \mathrm{f}(k)$, and number of light-absorbing components is such integer $r = k$ for which $s_k(A)$ is close to the instrumental standard deviation of absorbance, $s_{\mathrm{inst}}(A)$.

Alternative methods: *Residual standard deviation, Root mean square error, Average error criterion,* χ^2 *criterion, Standard deviation of eigenvalues.*

Application of approximate methods: If no knowledge of the experimental error associated with the data is available then one of the empirical function has to be applied to approximate the true dimensionality of the data.

Eigenvalues (EV): Eigenvalues *EV(k)* or g_k are conventionally used as a measure of the size of a principal component. Eigenvalues are calculated as the sum of squares of the score vectors

$$EV(k) = g_k = \sum_{i=k}^{n} t_{ki}^{2}, \qquad k = 1, 2, ..., r, ..., q$$

The first *r* eigenvalues being 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.

Alternative methods: *Logarithms of eigenvalues (*log g_k*), Exner function (*psi), *Scree test (RPV), Imbedded Error (IE*), *Factor Indicator (IND), F–test, Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA (RESO).*

Signal-to-noise ratio *SNR* (or *SER*) and detection limit: *SNR* are typically based on the ratio of the maximum signal to maximum noise value. The *signal-to-error ratio SER* is defined similarly but for an error the instrumental standard deviation of absorbance, $s_{inst}(A)$ is used. The *detection limit* is equivalent to the amount of "detectable impurity" or the smallest relative concentration of the minor component. 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) signal-to-noise ratio *SNR* with respect to the minor component.

Analysis of 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}.

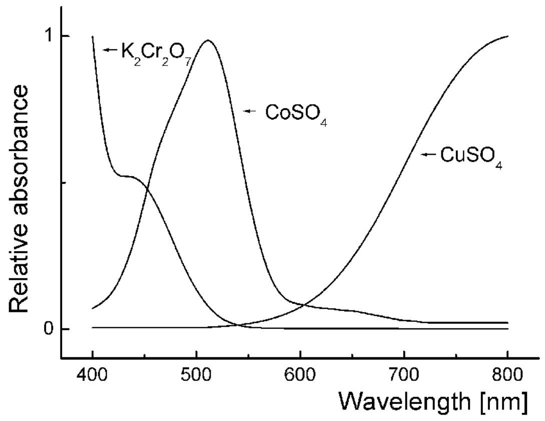

Fig. 4a Spectra of relative absorbance for three components

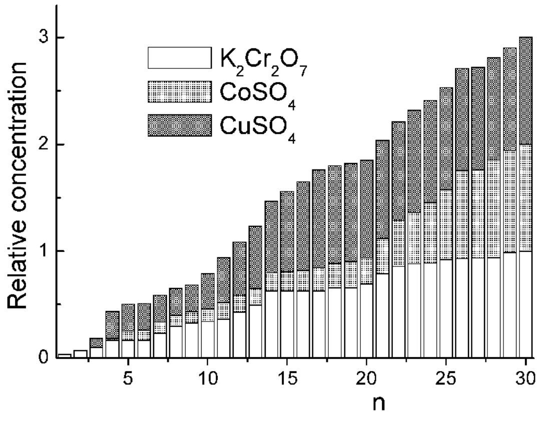

Fig. 4b Diagram of a relative concentration of three components in mixture for a simulated data set of three components

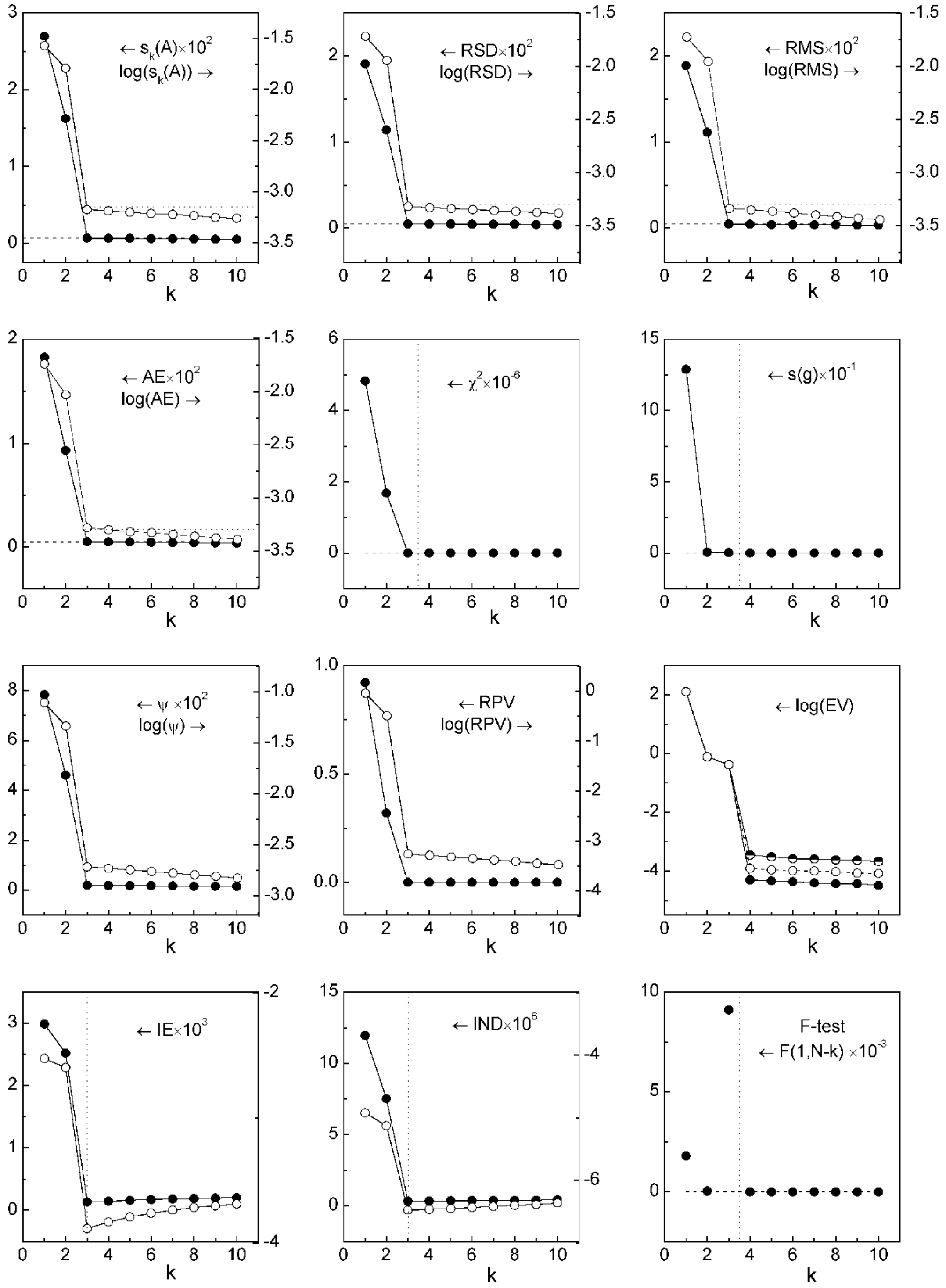

Fig. 5 The indices (full circles) and logarithm of the indices (empty circles) of 13 methods as a function of the number of principal components *k* for a simulated three-components system in mixture, potassium bichromate - cobalt(II) sulphate - copper(II) sulphate, with *r* = 3, *n* = 82, *m* = 41 and *SER* = 1570, *S-Plus*

An absorbance matrix was created by multiplying absorptivity spectra of three components (Fig. 4a) by their simulated concentration profiles (Fig. 4b) 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, $s_{inst}(A)$, to reach an optioned *SER* value. Most of simulated spectra sets for examination of five factors, (i. e. *concentration, homoscedasticity noise, collinearity in spectra, heteroscedasticity noise and sample size*), are of sample size $n = 82$ spectra and $m = 41$ wavelengths.

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

1. Instrumental error, $s_{inst}(A)$: for determination of the instrumental error of spectrophotometer used, $s_{inst}(A)$, Wernimont-Kankare method was applied. One light-absorbing component in solution means that a rank of 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 $k = 1$: for potassium bichromate $s_1(A) = 0.0007$.

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 compo-nents {Cr-Co-Cu} according to Beer law at $m = 27$ wavelengths was examined (Table 4).

Fig. 5 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 in the index value besides normal scale in each graph the logarithmic scale is often used. The *s(g)* and *IE* are functions of the *k*th 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, *REV, log g, IND, F*-test, *RESO* and at $k = r + 1$ for *g* and *IE*. On base of extensive simulations a comparison of thirteen 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, EV, IE, IND, F-test,* and *RESO*.

(a) *Homoscedastic noise, heteroscedastic noise* and *concentration of the minor component*: all three factors may be examined commonly using an effective resolution criterion, the *signal-to-noise ratio SNR* or the *signal-to-error ratio SER*. Both criteria cover all three mentioned 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 *EV*, $SER = 6$ for *RESO*, $SER = 4$ for *IND*. It means that for determination of the minor component two methods, *RESO* nad *IND* work best and 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, $s_{inst}(A)$. 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 signal-to-error ratio

SER. The *EV* 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, EV* 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 sufficiently demonstrates their ability to resist heteroscedastic noise.

(b) *Collinearity in spectra:* even severe collinearity in spectra was arranged the all indices predicted a correct number of components in mixture.

(c) *Sample size*: decreasing size from ($n \times m$) = (82 × 41) up to (40 × 20) all indices found correct value of number *r*. Decreasing size from (40 × 20) up to (20 × 20) *RPV, CHI, F-test* and *s(g)* methods fail to find a correct value of number *r*, (Table 3). When $n = 12$ most approximate indices fail. Correct value of number *r* may be found 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 sensible 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. 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.

(d) *Real experimental data*: Wernimont-Kankare procedure estimates the instrumental standard deviation of spectrophotometer used, $s_{inst}(A) = 0.0007$ 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), *cf.* Table 4. All 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 predicted highly accurate, all being absolutely correct for data sets with *SER* of at least 10. For data with higher noise the *EV, RESO* and *IND* predicted best.

Conclusions

1. The detection limit is the reliable limiting value for which the signal is still different from noise.
2. Two indices, *RESO* and *IND*, are stable in many situations and correctly predict a minor component in a mixture event 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.
3. Most indices predict the correct number of components for data sets with the *signal-to-error ratio SER* of at least 10 but *RESO* and *IND* of at least 6.
4. Wernimont-Kankare procedure is a reliable method for determination of the instrumental standard deviation of spectrophotometer used.
5. In case of real experimental data the *RESO, IND* and methods based on knowledge of instrumental error should be preferred.

Acknowledgement

Financial support of Internal Grant Agency of the Czech Ministry of Health (Grant No 4290-3) is thankfully acknowledged.

Recommended references:

[1] E. R. Malinowski, Factor Analysis in Chemistry, 2nd edn., Wiley, New York 1991.
[2] M. Meloun, J. Havel and E. Högfeldt, Computation of Solution Equilibria, 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] Zeng-Ping Chen, Yi-Zeng Liang, Jian-Hui Jiang, Yang Li, Jin-Ye Qian and Ru-Qin Yu, J. Chemom. 13 (1999) 15.
[7] Zeng-Ping Chen, Jian-Hui Jiang, Yang Li, Hai-Ling Shen, Yi-Zeng Liag and Ru-Qin Yu, Anal. Chim. Acta, 381 (1999) 233.
[8] A. K. Elbergali, J. Nygren and 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 and R. G. Brereton, Chemom. Intell. Lab. Syst. 27 (1995) 55.
[11] Y-Z. Liang, O. Kvalheim, A. M. Rahmani and R. G. Brereton, J. Chemom. 7 (1993) 15.
[12] S. Wold, C. Albano, W. J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström, 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 and 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, Brit. 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 and T. J. Klingen, Anal. Chem. 46 (1974) 982.
[24] R. D. Catell, Multivariate Beahavioral Research 1 (1966) 245.
[25] E. R. Malinowski, J. Chemom. 1 (1987) 49.

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 0.0003 to 0.0028 and various concentrations of third minor component; (bold digit means correct value found)

Noise level	*SNR*		*SER*		Precise methods						Approximate methods						
	All	Minor	All	Minor	$s_k(A)$	*RSD*	*RMS*	*AE*	χ^2	*s(g)*	ψ	*RPV*	*EV*	*IE*	*IND*	*F-test*	*RESO*
Concentration of minor component 0.25%																	
0.0003	932	2.3	3670	9.2	2	2	2	2	2	2	2	2	2	2	**3**	2	**3**
0.0007	485	1.2	1594	4.0	2	2	2	2	2	2	2	2	2	2	2	2	2
0.0014	226	0.6	802	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2
0.0028	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%																	
0.0003	932	4.7	3670	18.3	**3**	**3**	**3**	**3**	4	2	**3**	**3**	2	**3**	**3**	**3**	**3**
0.0007	485	2.4	1594	8.0	2	2	2	2	**3**	2	2	2	2	2	**3**	2	**3**
0.0014	226	1.1	802	4.0	2	2	2	2	2	2	2	2	2	2	2	2	2
0.0028	107	0.5	402	2.0	2	2	2	2	2	2	2	2	2	2	2	2	2
Concentration of minor component 1%																	
0.0003	932	9.3	3670	36.7	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.0007	485	4.9	1594	15.1	**3**	**3**	**3**	**3**	2	2	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.0014	226	2.3	802	8.0	2	2	2	2	2	2	2	2	**3**	2	**3**	2	**3**
0.0028	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%																	
0.0003	932	13.8	3670	55.0	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.0007	485	7.3	1594	23.9	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
0.0014	226	3.4	802	12.0	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	2	**3**	**3**	**3**
0.0028	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%																	
0.0003	932	23.3	3670	91.7	3	3	3	3	3	3	3	3	3	3	3	3	3
0.0007	485	12.1	1594	39.9	3	3	3	3	3	3	3	3	3	3	3	3	3
0.0014	226	5.7	802	20.0	3	3	3	3	2	3	3	3	3	3	3	3	3
0.0028	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

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 0.00001 to 0.0003; (b) from 0.00001 to 0.0007; (c) from 0.00001 to 0.0014; (d) from 0.00001 to 0.0028 and various concentrations of third minor component; (bold digit means correct value found)

Noise level	*SNR*		*SER*		Precise methods						Approximate methods						
	All	Minor	All	Minor	$s_k(A)$	*RSD*	*RMS*	*AE*	χ^2	*s(g)*	ψ	*RPV*	*EV*	*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	2	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

Table 3 Number of components predicted by 13 indices for simulated three-component system for various size of absorbance matrix, homoscedastic noise level 0.0007, *SER* = 15.7 and relative absorbance of all three components 1 : 1 : 0.02; (bold digit means correct value found)

Matrix size $n \times m$	Precise methods						Approximate methods						
	$s_k(A)$	*RSD*	*RMS*	*AE*	χ^2	*s(g)*	ψ	*RPV*	*EV*	*IE*	*IND*	*F-test*	*RESO*
10 × 10	2	2	2	2	**3**	2	2	2	2	7	5	2	1-**3**
12 × 10	**3**	**3**	**3**	**3**	**3**	2	2	2	**3**	7	**3**	**3**	1-**3**
20 × 10	**3**	**3**	**3**	**3**	4	2	2	2	**3**	7	**3**	2	1-**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

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 0.0007 was used while for experimental data a value $s_{inst}(A)$ = 0.0007 was found; (bold digit means correct value found)

		Precise methods						Approximate methods						
c_3	*SER*	$s_k(A)$	*RSD*	*RMS*	*AE*	χ^2	*s(g)*	ψ	*RPV*	*EV*	*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**

(-) means that a value can not be estimated