Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Analytica Chimica Acta 489 (2003) 137–151

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

Determination of the number of light-absorbing species in the protonation equilibra of selected drugs

Milan Meloun[a,*], Tomáš Syrový[a], Aleš Vrána[b]

[a] *Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice, Namesti Cs. Legii 565, 532 10 Pardubice, Czech Republic*

[b] *IVAX Pharmaceuticals, s.r.o. 747 70 Opava, Czech Republic*

Received 6 May 2003; received in revised form 2 June 2003; accepted 11 June 2003

Abstract

The determination of the number of components in a mixture is an important tool for qualitative and quantitative analysis in spectroscopy. The accuracy of nine selected indices for an estimation of the number of components that contribute to a set of spectra was critically tested on experimental data sets of protonation equilibria of four drugs using the INDICES algorithm in S-Plus. Methods are classified into two categories: *precise methods* based on a knowledge of the instrumental error of the sabsorbance data, $s_{\mathrm{inst}}(A)$, and *approximate methods* requiring no such knowledge. Indices of precise methods predict the correct number of components, even the presence of a minor one, when the quality of data is high and instrumental error is known. Improved identification of the number of species uses the second or third derivative function for some indices, namely when the number of species in the mixture is higher than four and when, due to large variations in the indicator values even at logarithmic scale, the indicator curve does not reach an obvious point where the slope changes. The number of variously protonated components and their dissociation constants for four drugs—mycophenolate, ambroxol, silybin and silydianin—at 25 °C were determined using SQUAD(84) regression and INDICES principal component analysis of the pH-spectrophotometric data. A proposed strategy of efficient experimentation in protonation constants determination, followed by a computational strategy, is presented with the goodness-of-fit tests for various regression diagnostics enabling the reliability of parameter estimates to be accessed.

Keywords: Principal component analysis; Decomposition of absorbance matrix; Factor analysis; Rank of matrix; Number of species; Instrumental error of spectrophotometer; Determining the number of components

1. Introduction

Determining the number of compounds in mixtures is an important tool for qualitative and quantitative analysis. In the study of protonation equilibria, for instance, a reliable determination of the number of components involved will help to obtain a reasonable interpretation of variously protonated oligomers, monomers, etc. Approaches to determining the number of components that contribute to a given set of spectra are based on two different methods: pure principal component analysis PCA, and PCA combined with cross-validation [1–3]. Using PCA, a set of correlated variables are transformed into a set of uncorrelated variables, principal components, of which

[*] Corresponding author. Tel.: +420-466037026; fax: +420-466037068.
E-mail addresses: milan.meloun@upce.cz (M. Meloun), tomas.syrovy@upce.cz (T. Syrový), ales_vrana@ivax-cr.com (A. Vrána).

0003-2670/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0003-2670(03)00761-X

the first few components explain most of the variation within the data. PCA will extract some noise sources, i.e. experimental and/or random error, which will usually be represented by the principal components with the smallest size or variance. When no noise in spectra exists, the number of eigenvalues of the covariance matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ which are larger than zero is equivalent to the number of components r, assuming that the spectra of the components in the mixture are linearly independent.

As all real data contain experimental noise, however, the number of eigenvalues different from zero is usually larger than the number of components p. In order to estimate the number of components in a mixture from the eigenvalues calculated by decomposition of the covariance matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, various indices methods have been designed. Their advantages and limitations have been discussed elsewhere [1–14].

All real data sets contain experimental and/or random error and it is the level of this error which can mask the identification of the true dimensionality of a data set. Malinovski [1,3] 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 ($(p+1)$th, $(p+2)$th, $\ldots, m$th) and which can therefore be extracted from the data by retaining only the first p dimensions. Imbedded error IE is the error which mixes into the factor scheme and is contained within the first p dimensions: this error can never be completely removed from a data set but may be scaled to a minimum [3]; thus, even a data set reproduced from the true number of PCs (p) contains some error. The level of imbedded error within a data set will therefore affect the reproduction of the data space.

In this paper, a critical comparison of nine selected indices methods applied to the protonation equilibria of four various drugs will be provided and the most reliable indices will be recommended.

2. Theoretical

2.1. Notation

The following notation will be used throughout the paper: in its generalized form, $\boldsymbol{A} = \boldsymbol{\varepsilon}\boldsymbol{C}$ represents the $n \times m$ absorbance data matrix containing the n recorded spectra as rows, $\boldsymbol{\varepsilon}$ is the $m \times p$ matrix of molar absorptivities, and $\boldsymbol{C}$ is the $p \times n$ concentration matrix. Here, m denotes the number of wavelengths for which each spectrum was recorded this being equal to the number of columns in the $\boldsymbol{A}$ matrix, n is the number of solutions for which spectra have been recorded, this being equal to the number of rows in the $\boldsymbol{A}$ matrix, and p is the number of components that absorb in the chosen spectral range. The rank of the matrix $\boldsymbol{A}$ is obtained from the equation $\mathrm{rank}(\boldsymbol{A}) = \min[\mathrm{rank}(\varepsilon), \mathrm{rank}(\boldsymbol{C})] \leq \min(m, p, n)$. Since the rank of $\boldsymbol{A}$ is equal to the rank of or $\boldsymbol{\varepsilon}$ or $\boldsymbol{C}$, whichever is the smaller, and since $\mathrm{rank}(\boldsymbol{\varepsilon}) \leq p$ and $\mathrm{rank}(\boldsymbol{C}) \leq p$, then provided m and n are equal to or greater than p, it will only be necessary to determine the rank of $\boldsymbol{A}$ to find the minimum number of absorbing species [11]. The rank of this absorbance matrix is the order of the largest non-zero determinant that can be obtained from its elements. Since the determinant of $\boldsymbol{A}$ is zero if its rows and columns are linearly dependent, the $\mathrm{rank}(\boldsymbol{A})$ is equal to the number of linearly independent columns of $\boldsymbol{A}$. That is to say, the rank of or $\boldsymbol{\varepsilon}$ or $\boldsymbol{C}$ will be less than p only if (1) the concentration of one or more species is zero in all experiments, (2) the concentrations of all species are zero in more than p experiments, and (3) the concentrations of one or more species can be expressed by a linear combination of the other species in all experiments. The first two conditions are trivial, while the third can be affected by choosing different concentration levels in some experiments. The law of mass action states that this is possible even if only one λ differs from zero, $c_{ij} = \lambda_j c_j;\ j = 1, \ldots, n$ [11]. Throughout this paper the level of noise in data will be considered. The concept of the instrumental error of absorbance measured for spectrophotometer $s_{\mathrm{inst}}(A)$ is used with the signal-to-error ratio SER being defined as the ratio of the maximum signal to this instrumental error $s_{\mathrm{inst}}(A)$.

2.2. Absorbance matrix decomposition

Principal component analysis PCA performs the decomposition of an absorbance matrix into a product of two matrices $\boldsymbol{T}$ and $\boldsymbol{P}^{\mathrm{T}}$ and the residual matrix or the matrix of undescribed variability $\boldsymbol{E}$ according to $\boldsymbol{A} = TP^{\mathrm{T}} + \boldsymbol{E}$. The $n \times o$ score matrix $\boldsymbol{T}$, also called the matrix of latent variables, contains o column vectors or main components. The $m \times o$ loading matrix

$\boldsymbol{P}$ contains o column vectors which represent a measure of the contribution of a particular latent variable. The index o is the least of n and m which in spectroscopy is usually n but generally also when $\boldsymbol{E}$ is zero. The second moment of an absorbance matrix is defined $\boldsymbol{Z} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}/(n-1)$ where $\boldsymbol{A}$ is usually the absorbance matrix. The matrix $\boldsymbol{Z}$ is often called the variance-covariance matrix and contains information about the scatter of points in multi-dimensional space. In fact, it describes the elliptical covariance structure of the data. The latent root and vector decomposition is defined by two equations:

$$|\boldsymbol{Z} - g_a\boldsymbol{I}| = \boldsymbol{0} \tag{1a}$$

and

$$Zp_a = g_a\boldsymbol{p}_a \tag{1b}$$

Sometimes data can be scaled so that each variable is standardized to equal variance down the columns, in which case the matrix $\boldsymbol{Z}$ becomes the correlation matrix, the matrix $\boldsymbol{I}$ is the unit matrix, and $\boldsymbol{0}$ is a matrix of zeroes. Eq. (1b) is a constrained maximization in which g is called the Lagrange multiplier; the g_a are the p latent roots and are obtained as the roots of the polynomial equation of order m defined by the determinant. The corresponding latent vectors $\boldsymbol{p}_a$ of dimension n have two constraints: they have unit length and they are mutually orthogonal.

2.3. Exact size of the true component space

The various indicator function PC(k) techniques developed to deduce the exact size of the true component space can be classified into two general categories: (a) precise methods based upon a knowledge of the experimental error of the absorbance data, $s_{\mathrm{inst}}(A)$, and (b) approximate methods requiring no knowledge of the experimental error [8]. In general, most precise and approximate methods are based on *the first criterion* concerning the procedure on finding the point where the slope of the indicator function $\mathrm{PC}(k) = f(k)$ changes. Elbergali et al. [7] proposed a modification of index methods using derivatives to improve identification of the number of components. The *derivative criteria* SD(k) are based on the point where the slope changes and reaches a maximum. The SD(k) is defined as $\mathrm{SD}(k) = \log[\mathrm{PC}(k+1)] - 2 \times \log[\mathrm{PC}(k)] + \log[\mathrm{PC}(k-1)]$ and $p-k$ should be at the first maximum of the SD(k) function. The *third derivative* TD(k) value crosses zero and reaches a negative minimum which can be used as a criterion. The TD(k) is defined as $\mathrm{TD}(k) = \log[\mathrm{PC}(k+2)] - 3 \times \log[\mathrm{PC}(k+1)] + 3 \times \log[\mathrm{PC}(k)] - \log[\mathrm{PC}(k-1)]$ and p should be equal to k value where TD(k) has its first minimum. The change in slope can also be found by calculating the *derivatives ratio* ROD(k) by $\mathrm{ROD}(k) = \{\mathrm{PC}(k-1) - \mathrm{PC}(k)\}/\{\mathrm{PC}(k) - \mathrm{PC}(k+1)\}$. Ideally ROD($k$) should have a maximum at the point where $k = p$.

2.3.1. Precise indices

Besides the first criterion applied, indicator function PC(k) methods are also based on a comparison of an actual index PC(k) of method used with the experimental error of the instrument used, $s_{\mathrm{inst}}(A)$. These are described elsewhere [15]:

1. *Kankare's residual standard deviation*, $s_k(A)$: The $s_k(A)$ values for different numbers of components k are plotted against an index k, $s_k(A) = f(k)$, and the number of significant components is an integer $p = k$ for which $s_k(A)$ is close to the instrumental error of absorbance $s_{\mathrm{inst}}(A)$ [11,15].
2. *Residual standard deviation*, RSD(k), is used analogously as in previous method $s_k(A)$.
3. *Average error criterion*, AE(k), is used analogously as in the preceding method $s_k(A)$.
4. *Bartlett χ^2 criterion*, $\chi^2(k)$ is used when the true number of significant components corresponds to the first k value for which $\chi^2(k)$ is less than critical $\chi^2(k)_{\mathrm{expected}} = (n-k)(m-k)$.

2.3.2. Approximate methods

A more difficult problem is to deduce the number of components without relying on an estimation of the instrumental error of absorbance, $s_{\mathrm{inst}}(A)$; then the first criterion only remains. Most of the techniques presented are empirical functions [15]. Eigenvalues g_k are conventionally used as a measure of the size of a principal component [14]. The first p eigenvalues, called a set of primary eigenvalues, contain a contribution from the real components and should be considerably larger than those containing only noise. The second set, called the secondary eigenvalues contains $(o-p)$ eigenvalues and these are referred to as non-significant eigenvalues.

1. *Exner function* $\psi(k)$: The Exner $\psi(k)$ function may be used for the identification of the true dimensionality of a data. Exner proposed that $\psi = 0.3$ 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. However, the first criterion is often preferred as the more reliable.
2. *Scree test*, *RPV*(k): The scree test for the identification of 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. 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. According to the first criterion, the point where the curve begins to level off, or where a discontinuity appears, is taken to be the dimensionality of the data space [1,16].
3. *Imbedded error function*, *IE*(k): The imbedded error function IE(k) is an empirical function [1] developed to identify those k latent variables which contain 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. The behavior of the IE(k) function, as long as k varies from 1 to o, 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, the IE(k) should increase.
4. *Factor indicator function*, *IND*(k): The factor indicator function IND(k) is an empirical function which appears more sensitive than the IE(k) function to identify the true dimensionality of an absorbance data matrix [1]. This function, like the IE(k) function, reaches a minimum when the correct number of latent variables or k PC dimensions is employed in the data reproduction. However, it has been seen that the minimum is more pronounced and/or can often occur even in situations where the IE(k) function exhibits no minimum.
5. *Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA*, *RESO*(k): The recommended procedure for determining the number of components in mixtures using RESO(k) [18] contains principal components analysis for the measured spectra set using the SVD algorithm to find the eigenvalues g_i^0 which correspond to ordinary PCA. Details may be found in original paper describing RESO [6]. The testing criterion calculates the 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. It estimates the number of components by examining the $\log(\mathrm{RESO}_i^a)$ versus component number plots. RESO then locates the number of $\log(\mathrm{RESO}_i^a)$ which are very close to each other and do not change substantially with the variation of k in comparison with the remaining $\log(\mathrm{RESO}_i^a)$. This is the number of components existing in the mixture examined.

2.4. Determination of protonation/dissociation constants

For dissociation reactions realized at constant ionic strength the so-called “mixed dissociation constants” are defined as $K_{a,j} = [\mathrm{H}_{j-1}\mathrm{L}]a_{\mathrm{H}^+}/[\mathrm{H}_j\mathrm{L}]$. These constants are found in experiments where pH values are measured with glass and reference electrodes, standardized with the practical $\mathrm{pH}(s) = pa_{\mathrm{H}^+}$ activity scale recommended internationally. If the protonation equilibria between anion, L (the charges are omitted for the sake of simplicity) of a drug and the proton, H, are considered to form a set of variously protonated species L, LH, LH_2, LH_3, etc. (which have a general formula $\mathrm{L}_q\mathrm{H}_r$ in a particular chemical model and are represented by p the number of species, $(q, r)_i$, $i = 1, \ldots, p$ where index i labels their particular stoichiometry), then the overall protonation (stability) constant of the protonated species, β_{qr}, may be expressed as

$$\beta_{qr} = \frac{[\mathrm{L}_q\mathrm{H}_r]}{[\mathrm{L}]^q[\mathrm{H}]^r} = \frac{c}{l^q h^r} \qquad (2)$$

where the free concentration $[\mathrm{L}] = l$, $[\mathrm{H}] = h$ and $[\mathrm{L}_q\mathrm{H}_r] = c$. For the ith solution measured at the jth wavelength, the absorbance, $A_{i,j}$, is defined as

$$A_{i,j} = \sum_{n=1}^{p} \varepsilon_{j,n} c_n = \sum_{n=1}^{p} (\varepsilon_{qr,j} \beta_{qr} l^q h^r)_n \qquad (3)$$

where $\varepsilon_{qr,j}$ is the molar absorptivity of the L_qH_r species with the stoichiometric coefficients q, r measured at the jth wavelength. The absorbance $A_{i,j}$ is the element of the absorbance matrix $\boldsymbol{A}$ of size $(n \times m)$ measured for n solutions with known total concentrations of two basic components, c_L and c_H, at m wavelengths. The multi-component spectra analyzing program SQUAD(84) [17–19] can adjust β_{qr} and $\boldsymbol{\varepsilon}_{qr}$ for absorption spectra by minimizing the residual-square sum function, U,

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m}(A_{\mathrm{exp},i,j} - A_{\mathrm{calc},i,j})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(A_{\mathrm{exp},i,j} - \sum_{k=1}^{p}\varepsilon_{j,k}c_k\right)^2 = \text{minimum} \quad (4)$$

where $A_{i,j}$ represents the element of the experimental absorbance response-surface of size $n \times m$ and the independent variables c_k are the total concentrations of the basic components c_L and c_H adjusted in n solutions. Unknown parameters may be divided into two equal groups: (1) a hypothetical chemical model which is supplied by the user and should contain (a) an estimate of the number of light-absorbing species in solution, p, and (b) a list of variously protonated species of stoichiometry indices $(q, r)_i$, $i = 1, \ldots, p$; (2) the best estimates of the protonation constants, $\beta_{qr,i}$, $i = 1, \ldots, p$, which are adjusted by the SQUAD(84) regression algorithm. At the same time, a matrix of molar absorptivities $(\varepsilon_{qr,j}, j = 1, \ldots, m)_k$, $k = 1, \ldots, p$, as non-negative reals is estimated, based on the current values of protonation constants. For a set of current values of $\beta_{qr,i}$, the free concentrations of ligand l (as h is known from pH measurement), for each solution is calculated, followed by the concentrations of all the species in equilibrium mixture $[L_qH_r]_j$, $j = 1, \ldots, p$, forming for n solutions the matrix $\boldsymbol{C}$. Various hypotheses of chemical models with refined parameters have been tested and the statistical characteristics describing the test-of-fit of regression spectra through experimental points have been calculated. The calculated standard deviation of absorbance $s(A)$ and the Hamilton *R-factor* are used as the most important criteria for a fitness test. If, after termination of the minimization process, the condition $s(A) \approx s_{\mathrm{inst}}(A)$ is met and the *R-factor* is less than 1%, the hypothesis of the chemical model is taken as the most probable one and is accepted.

3. Experimental

3.1. Chemicals

Mycophenolate, ambroxol, silybin and silydianin were generously donated by IVAX Pharmaceuticals, Czech Republic. A silymarin extract of pharmacopoeial quality (DAB IX) was prepared from *Silybum marianum*, var. Silyb (L.) Gaertn (Asteraceae). Individual components were isolated and purified by ethylacetate extraction, crystallization and chromatography. The final purities achieved by IVAX Pharmaceuticals, Czech Republic, were: *sodium mycophenolate* was prepared by neutralization reaction with sodium methanolate. *Ambroxol hydrochloride*, was purchased from Boehringer Ingelheim, Germany, with a purity of 99.9% (titration). *Silybin*: IVAX Pharmaceuticals, company standard AB023, Batch No. 190194, 97.5% (HPLC). *Silydianin* IVAX Pharmaceuticals, company standard RD, Batch No. 090680, 99.9% (HPLC). *Perchloric acid*, 1 M, was prepared by dilution of concentrated $HClO_4$ (p.a., Lachema Brno) with redistilled water and standardization against HgO and NaI with a reproducibility better than 0.2% according to the equation $HgO+4NaI+H_2O \rightleftarrows 2NaOH+Na_2[HgI_4]$ and $NaOH + HClO_4 \rightleftarrows NaClO_4 + H_2O$. *Sodium hydroxide*, 1 M, was prepared from an exact weight of pellets (p.a., Aldrich Chemical Company) with a carbon-dioxide free redistilled water. The solution was stored for several days in a polyethylene bottle. This solution was standardized against a solution of potassium hydrogen-phthalate using the Gran method with a reproducibility of 0.1%. *Mercury oxide*, *sodium iodide*, and *sodium perchlorate* (p.a., Lachema Brno) were not further purified. *Twice-redistilled water* was used in the preparation of solutions.

3.2. Apparatus and pH-spectrophotometric titration procedure

The apparatus used and the pH-spectrophotometric titration procedure has been described previously [22].

3.3. Procedure for protonation constants estimation

The experimental and computation a scheme for the determination of the protonation constants of the multicomponent system is taken from Meloun et al, cf. [20] and five steps are described in [22]:

(1) *Instrumental error of absorbance measurements*, $s_{\text{inst}}(A)$: The INDICES algorithm should be used with solutions of potassium dichromate to evaluate $s_{\text{inst}}(A)$, cf. [15]. The scree plot of $s_k(A) = f(k)$ consists of two straight lines intersecting at $\{s_k^*(A); k^*\}$ where k^* is the matrix rank for the system. Since $k^* = 1$ for $K_2Cr_2O_7$, the value of $s_k(A)$ for $k^* = 1$ is a good estimate of the instrumental error of the spectrophotometer used, $s_{\text{inst}}(A) = s_1^*(A)$, reaching a value of $s_1(A) = 0.25$ mAU, RSD $= 0.20$ mAU and AE $= 0.18$ mAU for the Cintra 40 (GBC, Australia) spectrophotometer.

(2) *Number of light-absorbing species*: When no outliers (grossly erroneous points) are present in the spectra examined, $s_k^*(A) \leq s_{\text{inst}}(A)$ is valid. The INDICES [15] determine the number of dominant species present in the equilibrium mixture. All spectra evaluation and data simulation were performed in the S-Plus programming environment and the INDICES algorithm is available on internet, http://meloun.upce.cz/indices. Most indices methods are functions of the number of PC(k) into which the spectral data are usually plotted against k, and when the PC(k) reaches the value of the instrumental error of spectrophotometer used, $s_{\text{inst}}(A)$, the corresponding k represents the number of significant components in a mixture, $p = k$. In general, most of the methods are based on finding the point where the slope of the indicator function $\text{PC}(k) = f(k)$ changes (*the first criterion*). The dependence $f(k)$ decreases steeply with an increasing number of PCs as long as the PCs are significant. When k is exhausted the indices fall off, some of indices even displaying a minimum. At this point $p = k$ for all indices except g for which $p = k + 1$ 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 p.

(3) *Choice of experimental and computational strategy*: In a titration, the total concentration of one of the components changes incrementaly over a relatively wide range, but the total concentrations of the other components change only by dilution, or not at all if they are present at the same concentration in the titrant and titrand. The protonation equilibria of drugs are studied in the visible region, 190–760 nm. The wavelength range selected is such that every species makes a significant contribution to the absorbance. Little information is obtained in regions of great spectral overlap or where the molar absorptivities of two or more species are linearly interdependent, as the change of absorbance following changes in c_{L} and c_{H} becomes rather small. If only a small number of wavelengths is used, then maxima or shoulders should be chosen, because small errors in setting the wavelength are then less important. It is best to use wavelengths at which the molar absorptivities of the species differ greatly, or a large number of wavelengths spaced at equal intervals.

(4) *Diagnostics indicating a correct protonation model*: When a minimization process in a regression analysis of an absorbance matrix terminates, some diagnostics are examined to determine whether the results should be accepted: the physical meaning of parametric estimates, β_{qr} and ε_{qr} should be neither too high nor too low, and ε_{qr} should not be negative. The absolute values of $s(\beta_j)$, $s(\varepsilon_j)$ give information about the last U-contour of the hyperparaboloid in vicinity of the pit, $U_{\min}$. For well-conditioned parameters, the last U-contour is a regular ellipsoid, and the standard deviations are reasonably low. High s values are found with ill-conditioned parameters and a “saucer”-shaped pit. The relation $s(\beta_j) \times F_\sigma < \beta_j$ should be met where F_σ is equal to 3. The set of standard deviations of ε_{pqr} for various wavelengths, $s(\varepsilon_{qr}) = f(\lambda)$, should have a Gaussian distribution; otherwise, erroneous estimates of ε_{qr} are obtained. The physical meaning of the species concentrations means that the calculated distribution of the free concentration of the basic components and variously protonated species of the chemical model should show molarities down to about 10^{-8} M. Since a species present at about 1% relative concentration or less in an equilibrium behaves as numerical noise in regression analysis, a distribution diagram makes

it easier to judge the contributions of individual species to the total concentration quickly. Since the molar absorptivities will generally be in the range 10^3–10^5 l mol^{-1} cm^{-1}, species present at less than ca. 0.1% relative concentration will significantly affect the absorbance only if their ε is extremely high.

The goodness-of-fit test contains the criteria for testing the correctness of a hypothetical chemical model. To identify the “best” or true chemical model when several are possible or proposed, and to establish whether or not the chemical model represents the data adequately, the residuals e should be analyzed. The goodness-of-fit achieved is easily seen by examination of the differences between the experimental and calculated values of absorbance, $e_i = A_{\mathrm{exp},i,j} - A_{\mathrm{calc},i,j}$. Examination of the spectra and the graph of the predicted absorbance response-surface through all the experimental points should reveal whether the results calculated are consistent and whether any gross experimental errors have been made in the measurement of the spectra. One of the most important statistics calculated is the standard deviation of the absorbance, $s(A)$, calculated from the set of refined parameters at the termination of minimization process. It is usually compared with the standard deviation of absorbance calculated by the INDICES program [15], $s_k(A)$, and if $s(A) \leq s_k(A)$, or $s(A) \leq s_{\mathrm{inst}}(A)$ (the instrumental error of the spectrophotometer used), the fit is considered to be statistically acceptable. Although this statistical analysis of residuals (cf. in [21], p. 62) gives the most rigorous test of the degree-of-fit, realistic empirical limits must be used. For example, when $s(A) \leq 0.002$, the goodness-of-fit is still taken as acceptable, whereas $s(A) > 0.010$ indicates that a good fit has not been obtained. Alternatively, some statistical measures of residuals e can be calculated: the mean residual $|\bar{e}|$ and the residual standard deviation $s(e) = s(A)$ should be close to the absorbance standard deviation known as the instrumental error of spectrophotometer used $s_{\mathrm{inst}}(A)$; a Hamilton *R-factor* of relative fit, expressed as a percentage, $(R \times 100\%)$, of <0.5% is taken as an excellent fit, but that one of >2% as a poor one. The *R-factor* may be used as a rigorous test of the null hypothesis H_0 (giving R_0) against the alternative H_1 (giving R_1). H_1 could be rejected at the significance level if $R_1/R_0 > R_{(k,n\text{-}k,\alpha)}$, where n is the number of experimental points, k is the number of unknown parameters, and $(n - k)$ is the number of degrees of freedom. The value of $R_{(k,n-k,\alpha)}$ can be found in statistical tables.

3.4. Software used

All spectra evaluation and data simulation were performed in the S-Plus programming environment and the INDICES algorithm is available on internet, http://meloun.upce.cz/indices [15]. Computation relating to the determination of dissociation constants was performed by regression analysis of UV-Vis spectra using the SQUAD(84) program [19].

4. Results and discussion

The UV-Vis spectra of sets of protonation equilibria of four drugs serve as an excellent example of the practical use of the proposed methodology of principal components analysis. A strategy for efficient experimentation in protonation constants determination followed by spectral data treatment is presented with the protonation equilibria of four drugs: sodium mycophenolate with two light-absorbing species L and HL in mixture, ambroxol with three, silybin with five and silydianin with six. While the simple protonation equilibria for mycophenolate are quite trivial L and HL, and were used for demonstration of methodology, for ambroxol, one dimer L_2H was indicated and in case of silybin and silydianin there are five and six variously protonated species in equilibrium mixture. pH-spectrophotometric titration enables absorbance matrix data (Figs. 1–4) to be obtained for analysis by non-linear regression. The reliability of unknown parameter estimates pK and ε may be evaluated on the basis of the goodness-of-fit test of residuals (Table 1). The SQUAD(84) program [19] analysis process starts with data smoothing followed by a factor analysis on the Kankare method using the INDICES procedure [15]. The position of a break-point on the $s_k(A) = f(k)$ curve in the scree plot is calculated and gives p with the corresponding coordinate $s_p(A)$ which also represents the instrumental error $s_{\mathrm{inst}}(A)$ of the spectrophotometer used. Due to the large variations in the indicator values, these are plotted on a logarithmic scale (Figs. 1–4). Protonation constants and molar

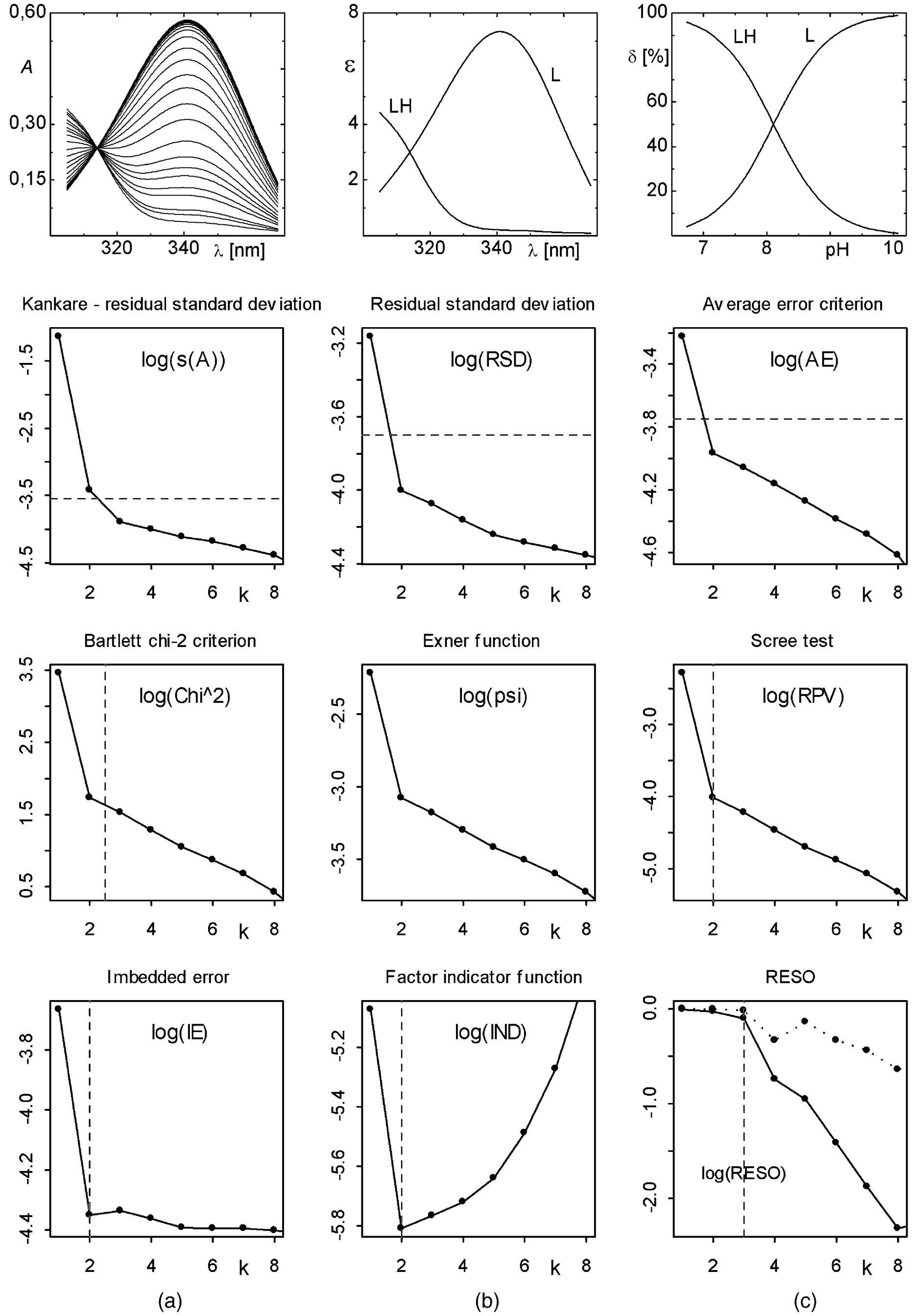

Fig. 1. Protonation of sodium mycophenolate presented on (a) pH-absorption spectra at 25 °C, (b) the spectra of molar absorptivities vs. wavelengths for all of the variously protonated species, (c) a distribution diagram of the relative concentrations of all of the variously protonated species. The logarithm dependence of 12 indices methods as a function of the number of principal components k for the pH-absorbance matrix. *Second row*: Kankare's residual standard deviation, $s_k(A)$; residual standard deviation, RSD; average error criterion, AE. *Third row*: Bartlett χ^2 criterion; Exner ψ function; Scree test RPV. *Fourth row*: imbedded error function, IE; factor indicator function, IND; RESO function.

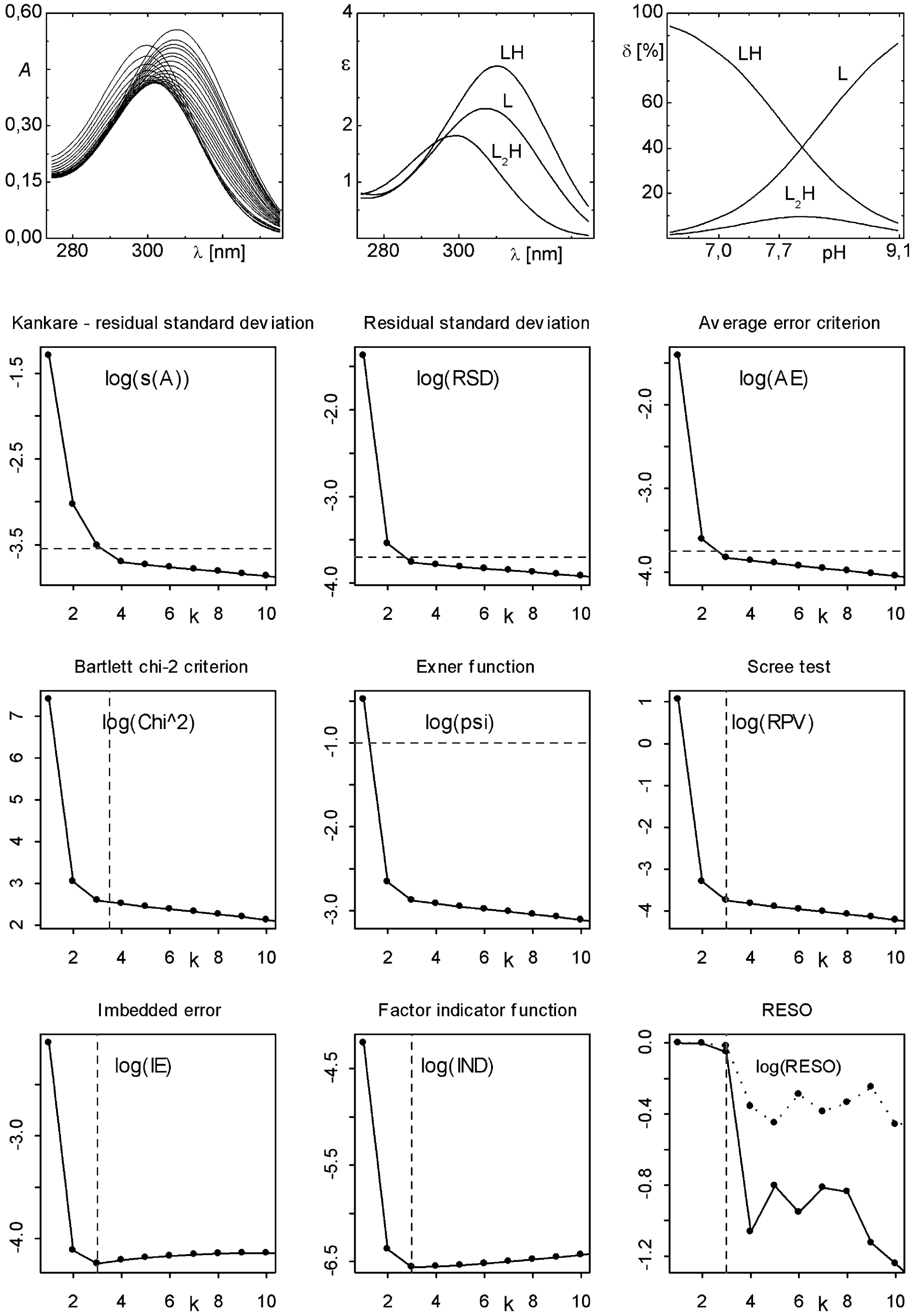

Fig. 2. Protonation of ambroxol presented according to Fig. 1.

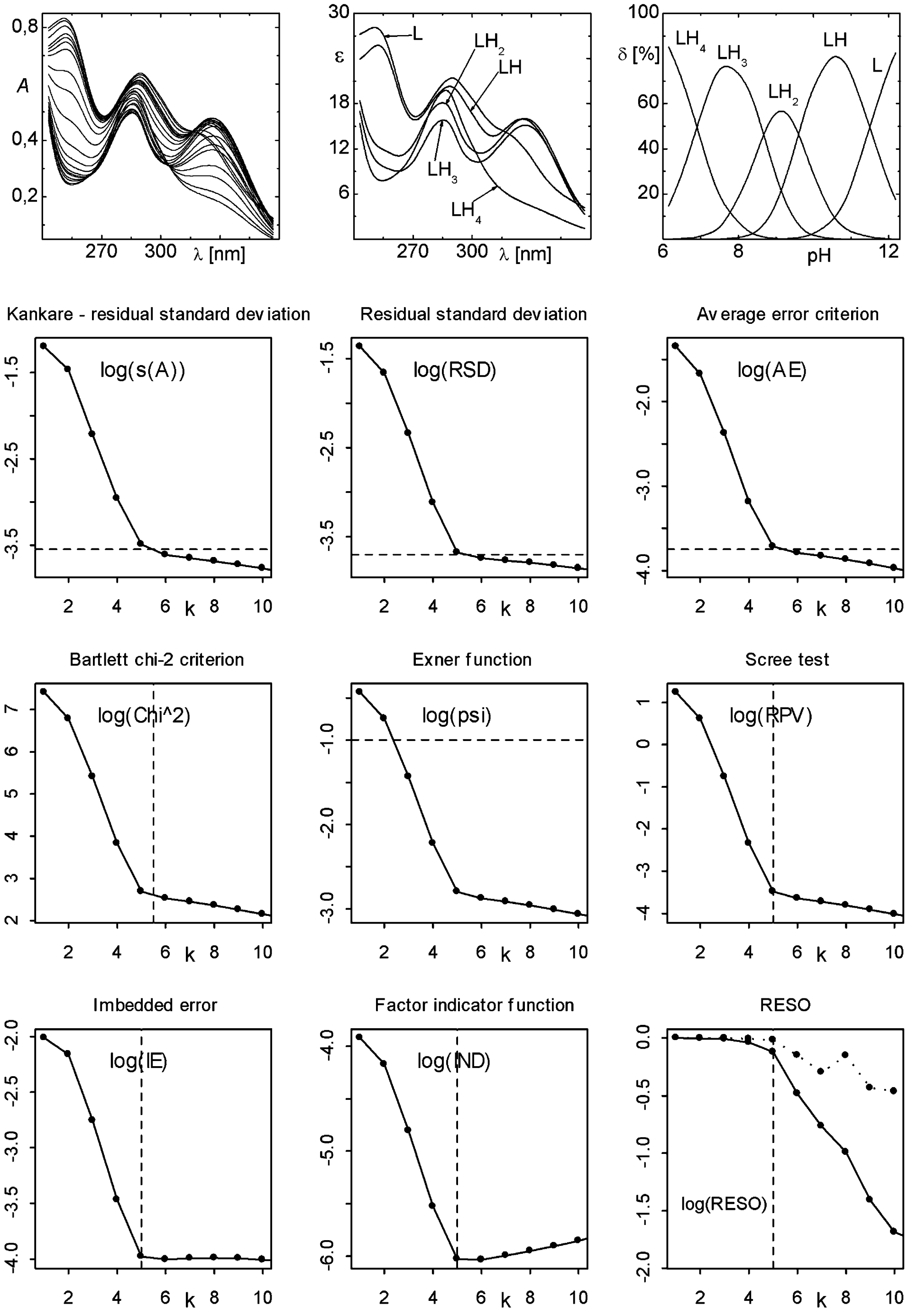

Fig. 3. Protonation of silybinin presented according to Fig. 1.

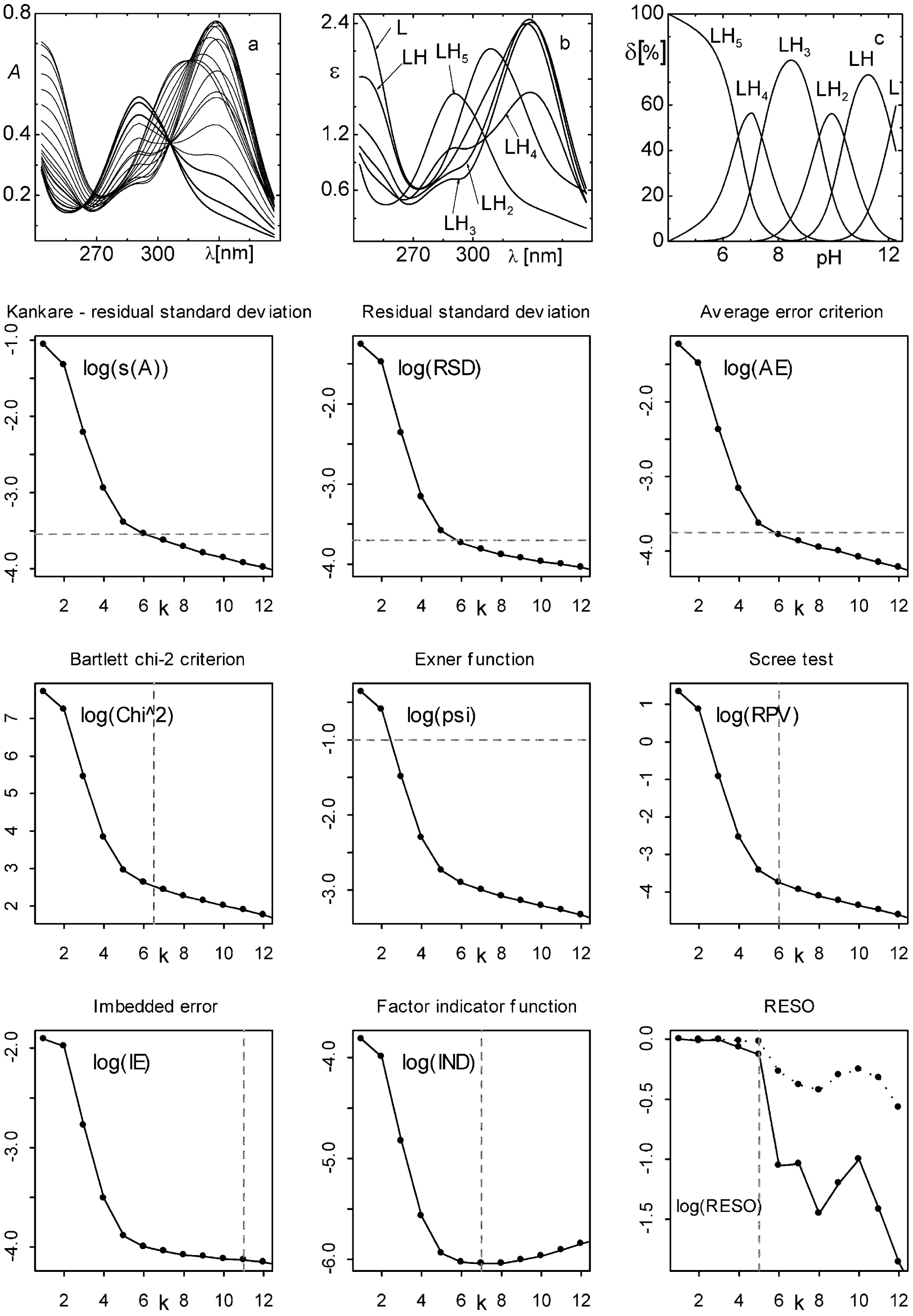

Fig. 4. Protonation of silydianin presented according to Fig. 1.

Table 1
Examination of the estimated number of light-absorbing species in a mixture using various PCA methods of the INDICES program and a search for a chemical equilibrium model of mycophenolate, ambroxol, silybin and silydianin using regression analysis of pH-spectrophotometric data with SQUAD(84)

Precise methods				Approximate methods				
$s_k(A)$	RSD	AE	χ^2	ψ	RPV	IE	IND	RESO
(1) Sodium mycophenolate[a]								
2	2	2	2	1	2	2	2	3
$s_k(A)$ (mAU) $= 0.14$		$s(A)$ (mAU) $= 1.41$		$\lvert\bar{e}\rvert$ (mAU) $= 1.03$		R-faktor (%) $= 0.45$		
(2) Ambroxol[b]								
3	3	3	3	2	3	3	3	3
$s_k(A)$ (mAU) $= 0.25$		$s(A)$ (mAU) $= 1.21$		$\lvert\bar{e}\rvert$ (mAU) $= 0.86$		R-faktor (%) $= 0.33$		
(3) Silybin[c]								
5	5	5	5	3	5	5	5	5
$s_k(A)$ (mAU) $= 0.3$		$s(A)$ (mAU) $= 1.01$		$\lvert\bar{e}\rvert$ (mAU) $= 0.67$		R-faktor (%) $= 0.20$		
Second derivative used								
5	5	5	5	5	5	5	5	–
(4) Silydianin[d]								
6	6	6	6	3	6	8	6	5
$s_k(A)$ (mAU) $= 0.23$		$s(A)$ (mAU) $= 1.14$		$\lvert\bar{e}\rvert$ (mAU) $= 0.73$		R-faktor (%) $= 0.23$		

The standard deviations of the parameter estimates in the last valid digits in brackets. The parameter reliability is proven with goodness-of-fit statistics such as the residual standard deviation $s_k(A)$ (mAU), the standard deviation of absorbance after termination of the regression process, $s(A)$ (mAU), the standard deviation of residuals (mAU) and the Hamilton R-factor (%). Indices algorithms used: $s_k(A)$ Kankare's residual standard deviation; RSD, residual standard deviation; AE, average error criterion; χ^2, Bartlett χ^2 criterion; ψ, Exner function; RPV, Scree test; IE, imbedded error function; IND, factor indicator function; RESO, the RESO procedure.

[a] $n = 20$, $m = 10$, SER $= 4382$, p$K_{\mathrm{a1}} = 8.23 \pm 0.00$, $I = 0.008$ (KCl), 25 °C.

[b] $n = 28$, $m = 40$, SER $= 2308$, p$K_{\mathrm{a1}} = 7.97 \pm 0.01$, $\log \beta_{21} = 11.34 \pm 0.02$, $I = 0.006$ (KCl), 25 °C.

[c] $n = 20$, $m = 40$, SER $= 2784$, p$K_{\mathrm{a1}} = 6.90 \pm 0.02$, p$K_{\mathrm{a2}} = 8.67 \pm 0.02$, p$K_{\mathrm{a3}} = 9.61 \pm 0.01$, p$K_{\mathrm{a4}} = 11.50 \pm 0.01$, $I = 0.032$ (KCl), 25 °C.

[d] $n = 20$, $m = 40$, SER $= 3377$, p$K_{\mathrm{a1}} = 6.54 \pm 0.07$, p$K_{\mathrm{a2}} = 7.40 \pm 0.07$, p$K_{\mathrm{a3}} = 9.47 \pm 0.06$, p$K_{\mathrm{a4}} = 10.41 \pm 0.025$, p$K_{\mathrm{a5}} = 12.09 \pm 0.01$, $I = 0.017$ (KCl), 25 °C.

absorptivities of the drug for m wavelengths constitute $m \times p$ unknown parameters which are refined by the multiple regression (MR) algorithm in the first run of the SQUAD(84) program. In the second run, the non-negative least squares (NNLS) algorithm makes the final refinement of all of the previously found parameter estimates with all of the molar absorptivities kept non-negative. The reliability of the parameter estimates may be tested with the use of SQUAD(84) diagnostics [19]: the first diagnostic indicates whether all parametric estimates β_{qr} and ε_{qr} have physical meaning and reach realistic values. As the standard deviations $s(\log \beta_{qr})$ of parameters $\log \beta_{qr}$ and $s(\varepsilon_{qr})$ of parameters ε_{qr} are significantly smaller than their corresponding parameter estimates (Table 1), all of the variously protonated species are statistically significant. Figs. 1–4 show estimated molar absorptivities of all the variously protonated species of drugs in dependence on wavelength. Some spectra quite overlap and such cases may cause resolution difficulties given a non-linear regression approach. The second diagnostic tests whether all of the calculated free concentrations of variously protonated species on the distribution diagram have physical meaning, which proved to be the case. The diagram shows that overlapping protonation equilibria exist in the cases of ambroxol, silybin and silydianin (Figs. 2–4). The goodness-of-fit (Table 1) proves that the $s_k(A)$ value is mostly equal to 0.25 mAU and is quite close to the standard deviation of absorbance when the minimization process terminates, $s(A)$. The statistical measures of all residuals e prove that the minimum of the eliptic hyperparaboloid

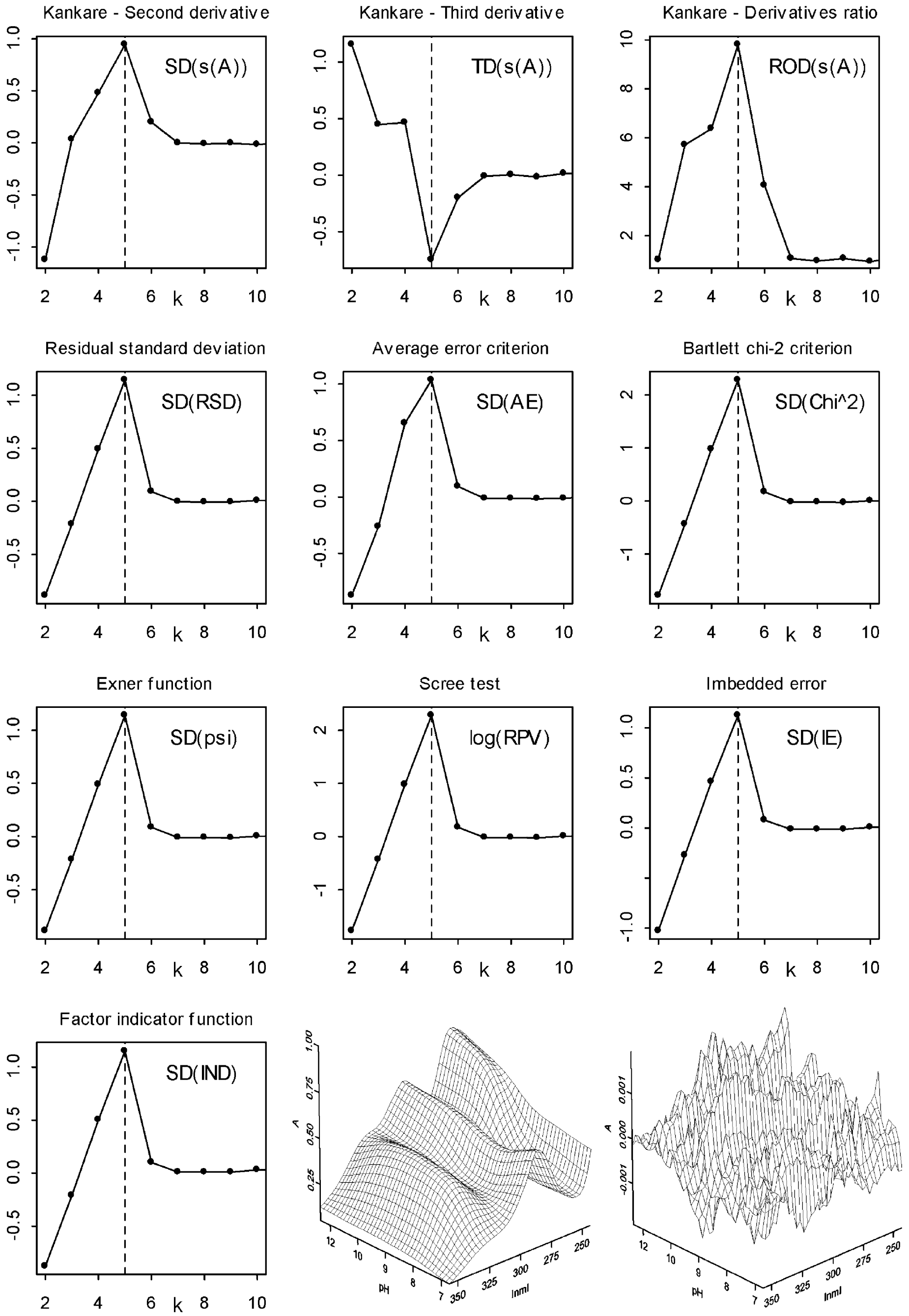

Fig. 5. The derivatives detection criteria of indices functions applied to the absorbance data set of the protonation equilibria of silybin from Fig. 3. *First row*: Kankare residual standard deviation, $s_k(k)$; the second derivative, SD($s_k(A)$) (left); the third derivative, TD($s(A_k)$) (middle); the derivatives ratio, ROD($s_k(A)$) (right). *Second row*: the second derivatives of various indices functions SD(RSD), SD(AE), SD(χ^2). *Third row*: SD(ψ), log(RPV), SD(IE). *Fourth row*: SD(IND); 3D-relative absorption spectra for variously protonated species of silybin in dependence on pH from Fig. 3; 3D-residuals map after regression fitting of absorption spectra of silybin in dependence on pH from Fig. 3.

U is reached: the mean residual $|\bar{e}|$ and the residual standard deviation $s(e) = s(A)$ have sufficiently low values. The Hamilton *R-factor* of relative fitness proves an excellent achieved fitness, and therefore the parameter estimates may be considered reliable. For all four drugs studied the most efficient tools, such as the Hamilton *R*-factor, the mean residual and the standard deviation of residuals are applied: as the *R*-factor in all cases reaches a value of less then 0.5%, an excellent fitness and reliable parameter estimates are indicated. The standard deviation of absorbance $s(A)$ after termination of the minimization process is always lower than 2 mAU, the proposal of a good chemical model and reliable parameter estimates are proven.

The first problem in the evaluation of the protonation equilibria of three drugs (ambroxol, silybin and silydianin) concerns the strongly overlapping equilibria, because the difference of two consecutive dissociation constants is less than 3. Such close equilibria are always difficult to evaluate and therefore the user should carefully prove the true number of variously protonated species in the mixture and the reliability of each dissociation constant estimate. A distribution diagram of the relative concentrations of all of the variously protonated species also demonstrates the overlapping protonation equilibria for close consecutive dissociation constants.

The second problem concerns small differences of molar absorptivities in variously protonated species within a spectrum (Figs. 2–4). It may happen that non-linear regression can fail when small differences of absorbance are of the same magnitude as instrumental noise, $s_{\mathrm{inst}}(A)$.

The number of light-absorbing species p can be predicted from the indices function values by finding the point $p = k$ where the slope of indices function $\mathrm{PC}(k) = f(k)$ changes or comparing $\mathrm{PC}(k)$ values with the instrumental error $s_{\mathrm{inst}}(A)$. This is the common criterion for to determining p. For a comparison of effectivity of selected indices methods of PCA in searching the number of light-absorbing species, the different sets of spectra concerning the protonation equilibria of the four drugs were applied. Very low values of $s_{\mathrm{inst}}(A)$ have proven that quite reliable spectrophotometer and experimental technique were used. Due to the large variations in the indices values, instead of indices, their logarithms of 12 selected methods as a function of the number of principal components k for every drug analyzed were used. For precise indices methods in Fig. 1 (as the Kankare's residual standard deviation $s_k(A)$, the residual standard deviation *RSD* and the average error criterion, AE), the horizontal line denotes the value of the instrumental error, $s_{\mathrm{inst}}(A)$. The best approximation of $s_{\mathrm{inst}}(A)$ for sodium mycophenolate was found for $k = 2$, while higher values of k do not lead to any significant decrease of $s_k(A)$. For the Bartlett χ^2 criterion the horizontal line denotes a magnitude of χ^2_{krit} and a vertical line separates values of k for which H_0 was accepted. In the case of the approximate indices methods for the Exner ψ function the value $\psi \leq 0.1$ is achieved for $k = 2$ while higher values k do not bring a significant decrease, in the value ψ. For the scree test RPV the curve of dependence $\mathrm{RPV}(k) = f(k)$ begins to level off at some point of k. This k value is considered to be the dimensionality of the absorbance data space. For the imbedded error function IE there is a minimum of $k = 2$ on the curve of the function $\mathrm{IE} = f(k)$. Similarly, for the factor indicator function, a minimum of $k = 2$ on the curve of the function $\mathrm{IND} = f(k)$ is reached.

A critical comparison of all of the methods of determining the number of light-absorbing species in solution based on the first criterion was carried out, and the results are given in Table 1. Our test showed that most of the indices accurately predict the number of variously protonated species that contribute to a set of spectra. When there are more than four components derivative methods are recommended: the curve $\mathrm{PC}(k) = f(k)$ does not exhibit an obvious break-point, and the second or third derivative localize this break more reliably. Fig. 5 shows reliable determination of the number of components using derivative methods of the protonation equilibria of silybin.

5. Conclusion

Indices methods are all based on finding the point where the slope of the indices function changes. Generally, the most reliable indices methods seem to be methods based on a knowledge of the instrumental error of absorbance, $s_{\mathrm{inst}}(A)$ which are usually preferred.

Acknowledgements

The financial support of the Internal Grant Agency of the Czech Ministry of Health (Grant No NB/7391-3) and of the Ministry of Education (Grant No MSM253100002) is gratefully acknowledged.

References

[1] E.R. Malinowski, Factor Analysis in Chemistry, second ed., Wiley, New York 1991.

[2] E.R. Malinowski, Abstract factor analysis of data with multiple sources of error and a modified Faber–Kowalski *f*-test, J. Chemom. 13 (1999) 69.

[3] E.R. Malinowski, Determination of the number of factors and the experimental error in a data matrix, Anal. Chem. 49 (1977) 612.

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

[5] Z.-P. Chen, Y.-Z. Liang, J.-H. Jiang, Y. Li, J.-Y. Qian, R.-Q. Yu, Determination of the number of components in mixtures using a new approach incorporating chemical information, J. Chemom. 13 (1999) 15.

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

[7] A.K. Elbergali, J. Nygren, M. Kubista, An automated procedure to predict the number of components in spectroscopic data, Anal. Chim. Acta 379 (1999) 143.

[8] 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.

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

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

[11] J.J. Kankare, Computation of equilibrium constants for multicomponent systems from spectrophotometric data, Anal. Chem. 42 (1970) 1322.

[12] M.S. Bartlett, Brit. J. Psychiatr. Stat. Sec. 3 (1950) 77.

[13] Z.Z. Hugus Jr., A.A. El-Awady, The determination of the number of species present in a system: a new matrix rank treatment of spectrometric data, J. Phys. Chem. 75 (1971) 2954.

[14] T.M. Rossi, I.M. Warner, Rank estimation of excitation—emission matrices using frequency analysis of eigenvectors, Anal. Chem. 58 (1986) 810.

[15] M. Meloun, J. Čapek, P. Mikšík, R.G. Brereton, Critical comparison of methods predicting the number of components in spectroscopic data, Anal. Chim. Acta 423 (2000) 51–68.

[16] R.D. Catell, Multivariate Beahavioral Res. 1 (1966) 245.

[17] D.J. Leggett (Ed.), Computational Methods for the Determination of Formation Constants, Plenum Press, New York, 1985, pp. 99–157, 291–353.

[18] D.J. Leggett, W.A.E. McBryde, General Computer Program for the Computation of Stability Constants from Absorbance Data, Anal. Chem. 47 (1975) 1065.

[19] M. Meloun, M. Javůrek, J. Havel, Multiparametric curve fitting. X. A structural classification of program for analysing multicomponent spectra and their use in equilibrium-model determination, Talanta 33 (1986) 513–524.

[20] M. Meloun, J. Havel, E. Högfeldt, Computation of Solution Equilibria, Ellis Horwood, Chichester, 1988, p. 226.

[21] M. Meloun, J. Militký, M. Forina, Chemometrics for analytical chemistry, PC-Aided Regression and Related Methods, vol. 2, Ellis Horwood, Chichester, 1994; PC-Aided Statistical Data Analysis, vol. 1, Ellis Horwood, Chichester, 1992.

[22] M. Meloun, T. Syrový, A. Vrána, The thermodynamic dissociation constants of ambroxol, antazoline, naphazoline, oxymetazoline and ranitidine by the regression analysis of spectrophotometric data, Talanta, in press.