Components of Program for Analysis of Spectra and Their Testing

Milan Javurek [1], ***Ivan Taufer*** [2]

The spectral analysis of aqueous solutions of multi-component mixtures is used for identification and distinguishing of individual components in the mixture and subsequent determination of protonation constants and absorptivities of differently protonated particles in the solution in steady state (Meloun and Havel 1985), (Leggett 1985). Apart from that also determined are the distribution diagrams, i.e. concentration proportions of the individual components at different pH values. The spectra are measured with various concentrations of the basic components (one or several polyvalent weak acids or bases) and various pH values within the chosen range of wavelengths. The obtained absorbance response area has to be analyzed by non-linear regression using specialized algorithms. These algorithms have to meet certain requirements concerning the possibility of calculations and the level of outputs. A typical example is the SQUAD(84) program, which was gradually modified and extended, see, e.g., (Meloun et al. 1986), (Meloun et al. 2012).

Keywords: *Spectral analysis, spectrophotometry, nonlinear regression.*

Introduction

The protonation constant of reaction of a weak acid or base, $L^{z-} + H^+ \leftrightarrow HL^z$, is defined according to the Guldberg–Waage law by Eq. (1):

$$K_H = \frac{[HL^z]}{[L^{z-1}][H^+]} \tag{1}$$

where the square brackets express the equilibrium concentrations (exactly, there should be the activity concentrations there, but within the concentration range used in spectrophotometry the activity coefficients can be considered equal to 1).

Anion L can form a number of differently protonated species: HL, H_2L, H_3L, etc, hence it can generally be described by the formula H_rL. Then the number of variously protonated forms represents the number of species in the solution, r_i, whose protonation constants are defined by Eq. (2):

$$\beta_{qr} = \frac{[H_qL_r]}{[L^q][H]^r} = \frac{c}{l^q h^r} \tag{2}$$

where the so-called free concentrations are $l = [L]$, $h = [H]$ and $c = [L_qH_r]$.

Each of these species is defined by its own spectrum in the UV/VIS region, so for the solution i and the wavelength j according to the Lambert–Beer law the measured absorbance is done by Eq. (3):

$$A_{ij} = \sum_{n=1}^{n_c} \varepsilon_{j,n} c_n = \sum_{n=1}^{n_c} (\varepsilon_{r,j} \beta_r l h^r)_n \tag{3}$$

where $\varepsilon_{r,j}$ are the molar absorption coefficients of the species H_rL which are characteristic for the wavelength j and spectrophotometric path equal to one,

n_c is the number of species in the solution.

Thus the values $A_i j$ form the absorbance matrix A of the dimension n_s vs. n_w (i.e. the number of solutions with different pH values vs. the number of wavelengths). The aim of analysis of the spectral matrix is to determine the chemical model of the solution, i.e. to determine stoichiometric coefficients, protonation constants, molar absorption coefficients, and free concentrations of all species. The analysis of multi-component spectra is carried out in the

[1] *doc. Ing. Milan Javurek, CSc.*, University of Pardubice, Faculty of Electrical Engineering and Informatics, Milan.Javurek@upce.cz

[2] *Prof. Ing. Ivan Taufer, DrSc.*, University of Pardubice, Faculty of Electrical Engineering and Informatics, Ivan.Taufer@upce.cz

following way: for guessed values of protonation constants and molar absorption coefficients, the resulting absorbance A_{calc} is calculated according to Eq. (3); and with the use of the least squares method by Eq. (4):

$$U = \sum_{i=1}^{n_s} \sum_{j=1}^{n_w} (A_{exp,i,j} - A_{calc,i,j})^2 = \sum_{i=1}^{n_s} \sum_{j=1}^{n_w} (A_{exp,i,j} - \sum_{k=1}^{n_c} \varepsilon_{j,k} c_k)^2 \Rightarrow min. \tag{4}$$

is calculated the goodness of fit, i.e. the agreement between the calculated absorbances and the experimental matrix. Then the method of non-linear regression transforms the fitted parameters so that the best goodness of fit is obtained. At the same time, the concentrations of individual species are determined from the mass balance calculated from the guessed protonation constants and the known overall concentrations of the components in the solution.

The stoichiometric coefficients, i.e. the composition of individual species in solution, could also be a part of the optimized parameters, but their interdependence in the model is smaller as compared with that of the other fitted parameters; therefore, they are taken as constants for the given calculation, and more than one calculation is carried out with different stoichiometries. Finally, the most suitable model is selected on the basis of the quality of fit.

Components of Program

Checking of Data

The input of data is relatively complicated: it is necessary to formulate the suggested chemical model inclusive of the guess of the overall protonation constants. The checking concerns the formal, logical as well as the physical correctness of the model inclusive of the experimental values of spectrum. The stoichiometry of species and their protonation constants are guessed either on the basis of earlier experience or are sought after in literature.

Inputs and Outputs of Program

With regard to the amount of input data, the program only works in batch regime. The extent and level of inputs is controlled by one of the input parameters. The analysis of residua and the distribution diagrams of all species in the individual solutions are printed besides the input data, the course and results of refinement of parameters.

Determination of Number of Components

An important tool in the finding of chemical model is the determination of number of components from the experimental absorbance matrix. A number of mathematical procedures have been published; for a survey and their comparison, see (Meloun et al. 2000). All these procedures have a common feature: application of factor analysis to the absorbance matrix. Here, with the use of Cattel's scree plot of eigenvalues of matrix calculated by various ways is guessed the number of components. A classical procedure was formulated by Kankare (Kankare 1970): it starts from the second moment M of the absorbance matrix A by Eq. (5):

$$M = \frac{1}{n_s} A^T A \tag{5}$$

where n_s is the the rank of absorbance matrix, i.e. number of solutions.

The eigenvalues r_a of matrix M are used for determination of residual standard deviation of absorbance $s_k(A)$ by Eq. (6):

$$s_k(A) = \sqrt{\frac{tr(M) - \sum_{a=1}^{k} r_a}{n_w - k}} \tag{6}$$

where $tr(M)$ is the trace of matrix M and k is the number of latent variables, which is calculated for the resulting error of absorbance $s_k(A)$ and represents the number of light-absorbing components.

Since data are always loaded with instrumental error, the value $s_k(A)$ for the component k is compared with the instrumental error $s_{inst}(A)$, which is known for the given measurement. In the graph, the standard deviation of absorbance $s_k(A)$ is plotted against the number of components, the solution being the number k, where the curve exhibits a sharp turn – see Fig. (1).

Calculation of Free Concentrations

For the calculation of error square sum function according to Eq. (4), it is necessary to know the concentrations of individual species in solution. For the guessed protonation constant and with known overall concentrations of individual components, the roots of non-linear Eq. (2) are sought after by the Newton–Raphson method.

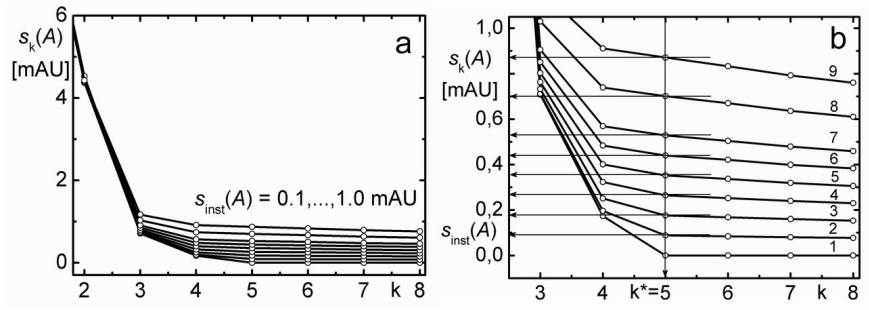

Fig. 1. (a) The Cattel's scree plot of the residual standard deviation of absorbance $s_k(A)$ depending on the number of the light-absorbing species for nine various levels of instrumental noise (b) The detail view on the Cattel's scree plot enabling an evaluation from simulated spectra of the actual instrumental standard deviation $s_{inst}(A)$ for five components $k = 5$

Optimization of Parameters

Two different methods are used for refinement of parameters assessment. The first of them, MR (Multiple Regression) uses the Gauss–Newton derivation method. The applied derivation method is fast and sufficiently precise; however, with incorrect input of initial guess of parameters it can lead to divergence. The second method used, NNLS (Non Negative Least Squares), uses penalization functions to correct the values of parameters with regard to their physical meaning. However, this method does not provide much too good results; it can only be considered as auxiliary in looking after an unknown model. The values of the first derivatives with respect to parameters (Jacobi's matrix) are calculated numerically according to the symmetrical Lagrange formula, the chosen step being in the magnitude of 0.5 % of each parameter. With regard to the fact that we have the whole matrix of data, the derivatives are added along the rows (i.e. over all the wavelengths). Besides the fitting of parameters, also calculated are the molar absorption coefficients, which characterize the species color for each wavelength. If we know some of them (e.g., from spectra of the pure components), it is suitable to input them: the calculation becomes easier and its quality improves.

Analysis of Residua

The method of non-linear regression has only limited possibilities for verification of quality of the found solution. Primary importance belongs to the physical significance of parameters (values of protonation constants): their calculated errors from non-linear regression and the calculated free concentrations of individual species. The only tool at our disposal for evaluation of quality of fit is the statistical analysis of residua; therefore, it must not be omitted in any calculation. This is performed along the rows of absorbance matrix, i.e. over all the wavelengths. The calculation concerns the central moments (arithmetic mean, standard deviation, skewness and kurtosis coefficients). Further calculated characteristics are median (which should be equal to arithmetic mean for normal distribution of residua) and Hamilton's R–factor, which expresses the goodness of fit (Meloun and Militky 2004) by Eq. (7):

$$R_F = \sqrt{\frac{U}{\sum_{i=1}^{N} y_i^2}} \tag{7}$$

where y_i are the measured values.

In the case of good fit, the R–factor should not exceed the error of measurement. Since Jacobi's matrix is known from previous calculations, it is possible to calculate besides classical residua also the standardized residua e_{Si} and JackKnife residua e_{Ji} , which indicate outliers points (Meloun and Militky 2004), defined by Eq. (8) and by Eq. (9):

$$e_{Si} = \frac{e_i}{s_{(i)}\sqrt{1-P_{ii}}} \tag{8}$$

$$e_{Ji} = e_{Si}\sqrt{\frac{n-m-1}{n-m-e_{Si}^2}} \tag{9}$$

where $s_{(i)}$ are standard deviations of individual points by Eq. (10)::

$$s_{(i)} = \sqrt{\frac{U - \frac{e_i^2}{1-P_{ii}}}{n-m-1}} \tag{10}$$

and P_{ii} are the diagonal elements of projection matrix calculated from Jacobi's matrix J by Eq. (11):

$$P = J(J^T J)^{-1} J^T \tag{11}$$

In conclusion of residua analysis, the individual characteristics for the whole absorbance matrix are summarized. The most significant is considered to be R–factor; evaluation concerns the concordance of central moments with ideal values and the magnitude of median. It is astounding that a number of renowned commercial algorithms do not contain the residua analysis at all: hence, the user has no possibility to evaluate competently the course of calculation and the results obtained (Gampp et al. 2004).

Simulation of Data

Analysis of spectra represents a relatively complicated set of procedures, which needs to be tested and verified from the standpoint of calculation quality. The most effective procedure lies in the possibility of generating synthetic data. Precise absorbance values are calculated for given values of protonation constants and molar absorption coefficients, and these absorbance values are loaded with errors having normal distribution according to the chosen error of measurement s_{inst}. The data are then processed like experimental data. The aim is to compare the calculation results with pre-chosen parameter values. Apart from testing the algorithm itself, we also can study the behavior of variously modified experimental models. The basic condition of such procedure is the real normality of the error set used for loading the generated data. The data simulation also enables generation of random errors of chosen magnitude, which simulates various instrumental errors of measurement, i.e. the precision of measuring instrument. Besides that, this provides a reliable platform for comparison of different algorithms.

Distribution Diagrams

For the evaluation of suitability of suggested chemical model of the analyzed mixture, it is important to construct the distribution diagrams, i.e. the dependence of concentrations of the species present upon changing conditions, in this case changing pH. In this case it is sufficient to have only the graphical representation of the earlier calculated free concentrations of all individual species at individual pH values over all wavelengths. The solutions are compared for a selected wavelength. The concentrations below 5 % are usually neglected: the respective species is considered to have no physico-chemical importance.

Experimental Data

For the test system we chose the trivalent equilibrium of significant cytostatic methotrexate. For more detailed characterization of the substance and experimental conditions, see http://en.wikipedia.org/wiki/Methotrexate, (Meloun et al. 2010). For input of experimental data into SQUAD(84) program, see http://meloun.upce.cz/docs/datasets/261/squadin.txt. The system is complicated by the fact that the equilibriums are close to each other, i.e. the protonation constants of individual steps are close and cannot be differentiated in the classical dependences of absorbance upon pH (see Fig. (2)). The form of absorption spectrum is presented in Fig. (3), and the absorbance response plane in Fig. (4). The spectrum was measured for 17 values of pH and 32 values of wavelength. The measurement conditions and results of evaluation of experimental data are described in detail elsewhere (Meloun et al. 2010).

This model, inclusive of the protonation constants, wavelength range, concentrations of components, and the found values of molar absorption coefficients of methotrexate was taken as a basis for verification of quality of the calculations performed by means of the SQUAD(84) program. The values of protonation constants found by analysis of experimental data (Meloun et al. 2010) are pK_{13} = 3.086; pK_{12} = 4.403; pK_{11} = 5.675; the difference between the second and the third protonation step is 1.2 of pH unit; these are near equilibriums. For testing the program, we selected a number of values of instrumental errors for generating the simulated data: s_{inst} = 1.0E-8, 0.0001, 0.0004, 0.0008, 0.001.

Tab. (1) shows that the determination of parameters of chemical model is reliable and corresponds to instrumental error of input data. This is best seen on the resulting standard deviation of absorbance $s(A)$, which never exceeds the value s_{inst}. However, a problem is encountered in the case of determination of the first dissociation constant, because the range of pH from 3.332 to 6.499 does not sufficiently cover the needed area. The value of the first constant as well as its error are markedly worsened with increasing instrumental error. Therefore, the data matrix was extended to the range of pH from 2.665 to 6.499. The results are presented in Tab. (2).

The trend in improvement of fitting of values of protonation constants is univocal. The residua characteristics for both data matrices are comparable; the only problem appears in insufficient normality of the generated errors which are

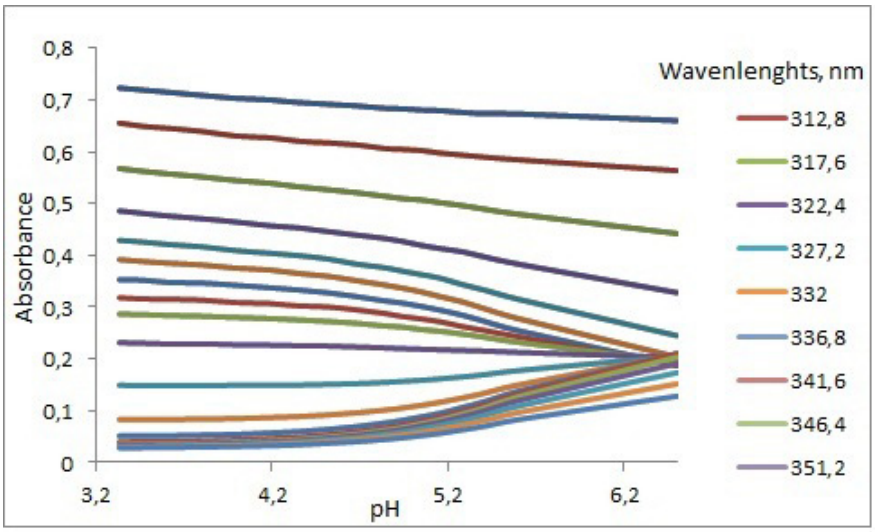

Fig. 2. Dependences A–pH of Methotrexate.

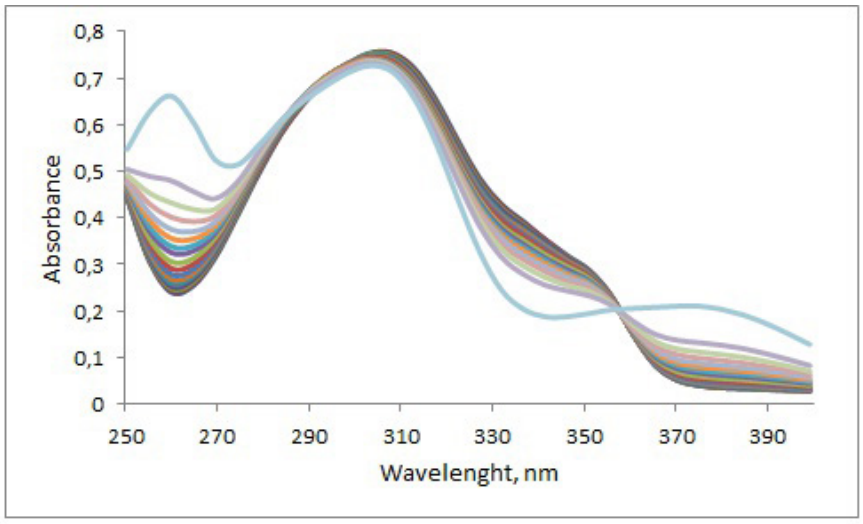

Fig. 3. Absorption spectrum of Methotrexate.

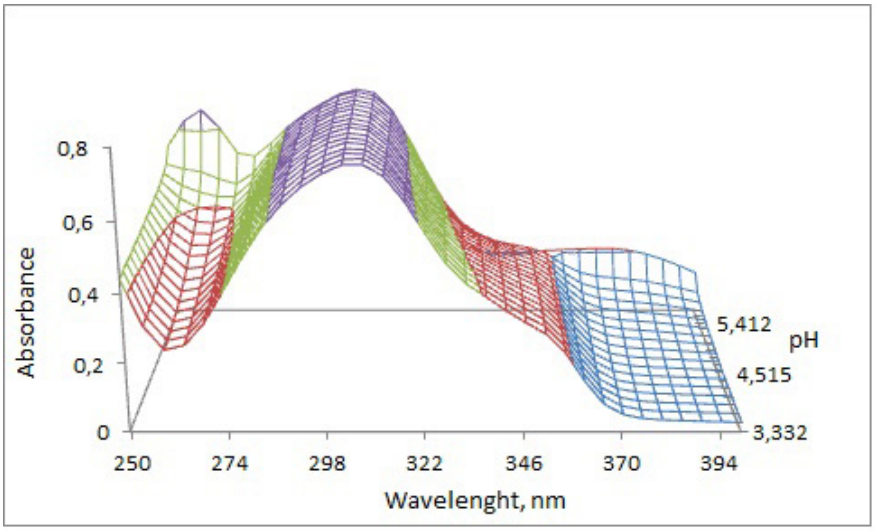

Fig. 4. Absorbance response plane of Methotrexate.

Tab. 1. Results of analysis of spectra generated for various values of instrumental error s_{inst} (17 solutions, 32 wavelengths)

s_{inst}	1.00E-08	1.00E-4	4.00E-4	8.00E-4	1.00E-3
pK_{13}	3.0841	3.1163	3.1922	3.2260	3.2395
pK_{12}	4.4039	4.4215	4.4821	4.5185	4.5367
pK_{11}	5.6750	5.6768	5.6847	5.6900	5.6931
$s(pK_{13})$	0.0041	0.0243	0.1031	0.1494	0.1754
$s(pK_{12})$	0.0015	0.0097	0.0483	0.0756	0.0919
$s(pK_{11})$	0.0001	0.0009	0.0051	0.0087	0.0110
$s(A)$	0.000150	9.63E-05	4.71E-04	7.55E-04	9.44E-04
Analysis of Residua					
Arithmetic mean	-1.4920E-15	-1.0670E-16	-1.0930E-16	-1.4730E-16	-9.6320E-17
Median	-9.2710E-10	-4.1220E-09	-2.7410E-08	-3.3990E-08	-4.2850E-08
Average residuum	3.0700E-06	6.4360E-05	3.1750E-04	5.0790E-04	6.3490E-04
Standard deviation	1.5520E-05	9.6290E-05	4.7140E-04	7.5470E-04	9.4370E-04
Skewness	-2.8760E-01	3.2640E-02	1.3750E-01	1.3690E-01	1.3610E-01
Kurtosis	1.3300E+02	2.7620E+00	2.6430E+00	2.6420E+00	2.6420E+00
Resid. sum of squar.	9.9420E-08	3.8290E-06	9.1790E-05	2.3520E-04	3.6780E-04
R-factor	3.0040E-05	1.8650E-04	9.1290E-04	1.4610E-03	1.8280E-03

Tab. 2. Results of analysis of spectra generated for various values of instrumental error s_{inst} (30 solutions, 32 wavelengths)

s_{inst}	1.00E-08	1.00E-4	4.00E-4	8.00E-4	1.00E-3
pK_{13}	3.0856	3.0843	3.0795	3.0751	3.0728
pK_{12}	4.4026	4.3989	4.3855	4.3754	4.3694
pK_{11}	5.6749	5.6746	5.6738	5.6730	5.6725
$s(pK_{13})$	0.0011	0.0069	0.0276	0.0547	0.0677
$s(pK_{12})$	0.0007	0.0046	0.0185	0.0366	0.0453
$s(pK_{11})$	0.0001	0.0005	0.0019	0.0037	0.0046
$s(A)$	1.47E-05	9.35E-05	3.76E-04	7.56E-04	9.45E-04
Analysis of Residua					
Arithmetic mean	-2.4710E-17	-1.0670E-16	-2.4900E-17	7.8170E-18	-8.1900E-17
Median	-5.4180E-10	-4.1220E-09	-1.7920E-08	-3.1230E-08	-3.8630E-08
Average residuum	3.6860E-06	6.4360E-05	2.7600E-04	5.5480E-04	6.9350E-04
Standard deviation	1.4650E-05	9.6290E-05	3.7610E-04	7.5570E-04	9.4470E-04
Skewness	1.2120E+00	3.2640E-02	-1.1740E-02	-5.7690E-03	-5.7490E-03
Kurtosis	2.6380E+00	2.7620E+00	2.6910E+00	2.6770E+00	2.6770E+00
Resid. sum of squar.	1.7800E-07	3.8290E-06	1.1730E-04	4.7340E-04	7.3980E-04
R-factor	3.0520E-05	1.8650E-04	7.8330E-04	1.5740E-03	1.9670E-03

used to load the calculated absorbance values. We failed to solve this problem; but in the whole context, its importance seems to be negligible. It has to be noted here that the characteristics of residua are added along the rows, i.e. for all the wavelengths. For the data loaded by virtually no error, the obtained results fully correspond with the pre-chosen values, while the quality of determination is proportionately lowered with the error-loaded data. It can be concluded that the processing of data is fully reliable; hence, the determination of parameters of chemical model is exclusively given by experimental data quality.

Conclusion

Analysis of spectra represents a very useful tool in studies of chemical equilibriums, i.e. in determination of chemical model of the given substance. Important factor is not only the composition of solution i.e. the content of

individual species depending on pH change, but also (and foremost) correct determination of protonation constants, which give us basic information about acid-base behavior of the substance. Collecting of experimental data is relatively easy nowadays; available are sophisticated and highly precise spectrophotometers that measure absorbances to six decimal places. Then the key role belongs to the evaluation proper – it is impossible to perform it without the computer and corresponding algorithm. Literature describes a number of algorithms, out of which the SQUAD(84) program used in our workplace was supplemented with a number of tools important for the user evaluating the quality of calculation. It was also compared with other newer programs (see [4]): even here it provided quite comparable results.

Acknowledgement: *The problem was dealt with in the framework of research plan MŠM 0021627505 "Control, Optimizing and Diagnostics of Complex Systems".*

References

Gampp, H., Maeder, M., Mayer, C.J., Zuberbuhler, A.D., 2004. Specfit/32. Spectrum Software Associates .

Kankare, J.J., 1970. Computation of equilibrium constants for multicomponent system from spectrophotometric data. Anal. Chem .

Leggett, D.J., 1985. Computational Methods for the Determination of Formation Constants. Plenum Press. ISBN:.

Meloun, M., Capek, M., Miksik, P., Brereton, R.G., 2000. Critical comparision of methods predicting number of components in spectrophotometric data. Analytica Chimica Acta .

Meloun, M., Ferencikova, Z., Javurek, M., 2012. Reliability of dissociation constants and resolution capability of squad(84) and specfit/32 in the regression of multiwavelength spectrophotometric ph-titration data. Spectrochimica Acta Part A .

Meloun, M., Ferencikova, Z., Vrana, A., 2010. The thermodynamic dissociation constants of methotrexate by non-linear regression and factor analysis of multiwavelength spectrophotometric ph-titration data. Cent. Eur. J. Chem .

Meloun, M., Havel, J., 1985. Computation of Solution Equilibria 1. Spectrophotometry. Folia Fac. Sci. Nat. Univ. Purkyn. Brunensis. ISBN:.

Meloun, M., Javùrek, M., Havel, J., 1986. Multiparametric curve fitting 10. a structural classification of programs for analysing multicomponent spectra and their use in equilibrium model determination. Talanta .

Meloun, M., Militky, J., 2004. Statistical Analysis of Experimental Data (in Czech). Academia. ISBN: 8020012540.