Mikrochim. Acta 112, 155–169 (1993)

Mikrochimica
Acta

Computer Estimation of Dissociation Constants. Part VI.* Diagnostics in Regression Analysis of Absorbance-pH Curve

Milan Meloun[1],** and Jiří Militký[2]

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

[2] Department of Textile Materials, Technical University, 461 17 Liberec, Czech Republic

Abstract. Nonlinear regression program DCMINOPT is introduced for numerical analysis of a set of {A, pH} data expressing a dependence of absorbance of a mixture of variously protonated light-absorbing species L, LH, ..., LH_R on pH. Efficiency of the program has been examined on simulated A-pH data corrupted with artificial (generated) errors namely for a case of closely overlapping protonation equilibria. An accuracy and precision of parameters estimates have been examined and compared with those determined by another three standard algorithms DCFIT, DCMINUIT and PSEQUAD. Goodness-of-fit test brings various regression diagnostics, 3D-plots and statistical measures enabling to test and prove a reliability of a regression process and accuracy and precision of parameter estimates.

Key words: consecutive protonation of LH_r, closely overlapping equilibria, nonlinear regression of A-pH curve, dissociation constants pK_a, molar absorption coefficients ε_{LH_r}, accuracy and precision of pK_a and ε_{LH_r}, regression diagnostics, goodness-of-fit test, reliability of pK_a estimation.

The analysis of an absorbance-pH curve for a protolytic acid to determine dissociation constants and molar absorptivities is not an easy task when overlapping protonation equilibria are present. The programs SPOPT and DCMINUIT [1] have been tested and compared with DCLET [2] and LETAGROP SPEFO [3] for analysis of overlapping equilibria of a triprotic acid 2-, 3- and 4-CAPAZOXS [4]. Two approaches of mathematical model formulation and several optimization algorithms were tested on absorbance-pH curve analysis of 3-CAPAZOXS and general rules for investigation were recommended [5].

Structural classification of regression programs in solution equilibria study was introduced in the ABLET system [6–9], adapted to A-pH curve analysis to deter-

* Previous Part V.: Mikrochim. Acta **1992**, *109*, 221

** To whom correspondence should be addressed mine the protonation and regression spectra analysis [10]. The content of several blocks may change and the resulting program structure was described previously [1, 15].

This paper examines the efficiency of the new program DCMINOPT and discusses the reliability of determination of two consecutive dissociation constant and corresponding molar absorptivities ε_{LH}, ε_{LH_2}, ε_{LH_3} when concerning two closely overlapping protonation equilibria of 4-CAPAZOXS at low concentration of dye in solution in which monomers prevail. The examination of parameters conditioning and an accuracy of ill-conditioned parameters using 3D-graphs of the (C-U) hyperparaboloid response-surface is introduced, some diagnostic tools as the last U contours and the correlation coefficients of parameters are estimated. Regression diagnostics and the regression process of new program DCMINOPT are compared with programs DCFIT, DCMINUIT and PSEQUAD.

Theoretical

a. Modus Operandi

The structural classification of regression program enables easy formation of the program for an analysis of A-pH curve. Besides PSEQUAD [11], the DHFIT [12] is rewritten to resulting DCFIT, and the DHMINUIT [12] to DCMINUIT and then an efficiency compared with the new program DCMINOPT. All these programs contain the following common blocks structure:

(1) Input: This block reads data $\{pH_{read}, A_{exp}\}$ and makes some correction of measured values pH_{read} for a deviation of glass electrode cell from the Nernstian slope S, for any difference in temperature from 298.16 K, and for the liquid-junction potential E_j

$$pH = ((pH_{read} - pH(st))\ 59.16\ T/(S\ 298.16)) + E_j/S + pH(st), \qquad (1)$$

where pH(st) is pa_{H^+} for the standard buffer solution used. In regression analysis, the regression model $y = f(x; \beta)$ contains the independent variable pH $(=x)$, the dependent variable A $(=y)$ and the unknown parameters $\beta_1, \ldots, \beta_m$ which are represented by dissociation constants $pK_{a,i}$ and molar absorption coefficients ε_L, ε_{LH_i}, $i = 1, \ldots, R$.

(2) Residual sum of squares $U(\beta)$: This block formulates the residual- sum of squares $U(\beta)$ which is minimized in programs DCMINOPT, DCFIT, DCMINUIT and PSEQUAD. The A-pH curve for a mononuclear acid is written with the assumption that base L is protonated to form variously protonated ions LH_1, LH_2, $LH_3, \ldots,$ $LH_r, \ldots, LH_R$, etc. of the mononuclear acid LH_R (the charges are omitted for sake of simplicity). The model $A = f(pH; pK_{a,i}, \varepsilon_L, \varepsilon_{LH_i}, i = 1, \ldots, R)$ is represented by an equation for the absorbance-pH curve at a given wavelength λ written as

$$A = d \cdot L \, \frac{\varepsilon_L + \sum\limits_{r-1}^{R} \varepsilon_{LH_r} \cdot 10^{(r \cdot \log a_H + \log \beta_{1r})}}{1 + \sum\limits_{r=1}^{R} 10^{(r \cdot \log a_H + \log \beta_{1r})}}, \qquad (2)$$

where d is the cuvette path-length, L is the total analytical concentration of LH_R, $\beta_{1r} = [LH_r]/([L][H]^r)$ and when the conventional activity pH scale is used and the mixed stepwise dissociation constant $K_{a,i} = a_H[LH_{i-1}]/[LH_i]$, it will be

$$r \cdot \log a_H + \log \beta_{1r} = \sum_{i=1}^{r} pK_{a,i} - r \cdot pH. \tag{3}$$

The program PSEQUAD [11] also enables a determination of complex-forming equilibria.

The residual sum of squares $U(\beta)$, is then formulated by

$$U(b) = \sum_{i=1}^{n} w_i[A_{exp,i} - f(pH; pK_{a,i}, \varepsilon_L, \varepsilon_{LH_r}, i = 1, \ldots, R)]^2$$

$$= \sum_{i=1}^{n} w_i(A_{exp,i} - A_{calc,i})^2 = \text{minimum}, \tag{4}$$

where $A_{exp,i}$ is the measured absorbance at a given wavelength, $A_{calc,i}$ is calculated according to Eq. (2) and w_i is the statistical weight usually taken unity. The equation $U(\beta)$ (4) contains dependent variable A, independent variable pH ($= -\log a_H$) and parameters estimated $pK_{a,i}$, ε_L, ε_{LH_i}, $i = 1, \ldots, R$.

(3) Minimization: The algorithm FIT [13] (in the program DCFIT), the algorithm MINUIT [14] (in the program DCMINUIT), DCMINOPT employs the algorithm MINOPT [15] and the program PSEQUAD [11] were described elsewhere.

(4) Statistical (error) analysis: This block calculates confidence intervals of parameters and correlation coefficients a description may be found in previous contribution of this series [12, 15]. PSEQUAD [11] evaluates the standard deviation of a dependent variable, $s(A) = \sqrt{U/(n - m)}$ where n is a number of points of A-pH curve and m is a number of parameters estimated; the standard deviations of parameters estimated $s(\beta_{1r})$ and $s(\varepsilon_{1r})$, and the paired r_{ij}, total ρ_{ij} and multiple R_i correlation coefficients.

(5) Goodness-of-fit test: This block contains the examination of fitness achieved by the statistical analysis of residuals. The residuals are defined as the differences

$$\hat{e}_i = A_{exp,i} - A_{calc,i}, \quad i = 1, \ldots, n, \tag{5}$$

where $A_{exp,i}$ is the i-th observation and $A_{calc,i}$ is the i-th prediction (2). As certain underlying assumptions have been outlined for the regression analysis, such as the independence of random errors ε, their constant variance (homoscedasticity), and normal (Gaussian) distribution for ε, the residuals should possess characteristics that agree with, or at least do not refute, the basic assumptions: this the residuals should be randomly distributed about the prediction A_{calc}. Systematic departures from randomness indicate that the model is not satisfactory. The goodness-of-fit test (which is also called the fitness test) analyses the residual set and examines following statistical characteristics (detailed description is in previous part [12] or ref. [16]):

(1) The *arithmetic mean of residuals* known as *the residual bias*, $E(\hat{e})$, and the robust measure of location, *the median* $\hat{e}_{0.5}$, should be equal to zero.

(2) The *mean of absolute values of residuals*, $E|\hat{e}|$, and the mean of absolute values of relative residuals $100\,E|\hat{e}_{rel}|$ in percents, with the square-root of the residual variance $s^2(\hat{e})$ known as the estimate of *the residual standard deviation*, $s(\hat{e})$, and the robust measure of scale, *the standard deviation of median* $s(\hat{e}_{0.5})$. Obviously, it is also $s(\hat{e}) \approx s_{inst}(A)$ where $s_{inst}(A)$ is instrumental error of absorbance.

(3) The *residual skewness*, $g_1(\hat{e})$, should be for normal distribution of residuals equal to zero;

(4) The *residual curtosis*, $g_2(\hat{e})$, should be for normal distribution equal to 3.

(5) The *residual variance* $s^2(\hat{e})$ is calculated from the residual sum of squares.

(6) The *determination coefficient* D^2 is computed from the relation

$$D^2 = 1 - \frac{U(b)}{\sum_{i=1}^{n} (A_{exp,i} - \overline{A_{exp,i}})^2}, \tag{6}$$

where $\overline{A_{exp}} = 1/n \sum_{i=1}^{n} A_{exp,i}$. The determination coefficient is for linear models equal to square of the multiple correlation coefficient.

(7) When determination coefficient is multiplied by 100%, we receive so called *regression rabat*, $D^2 . 100$ [%].

(8) In chemometrics the *Hamilton R-factor of relative fitness* is often used being expressed by

$$R = \sqrt{\frac{U(b)}{\sum_i A_{exp,i}^2}}. \tag{7}$$

(9) To distinguish between models the *Akaike information criterion* AIC is more suitable to apply which is defined by relation

$$AIC = -2L(b) + 2 \cdot m. \tag{8}$$

The "best" model is considered to be a model for which this criterion reaches a minimal value. Using the least-squares and models which do not belong into the same class the AIC criterion may be expressed

$$AIC = n \cdot \ln\left[\frac{U(b)}{n}\right] + 2 \cdot m. \tag{9}$$

The influential points may be easily identified on base of an one-step approximation of the Jackknife residuals $\hat{e}_{Ji}$ calculated by

$$\hat{e}_{Ji} = \frac{\hat{e}_i}{\hat{s}_{(i)}\sqrt{1 - P_{ii}}}, \tag{10}$$

where P_{ii} are elements of a projection matrix, $P = J(J^TJ)^{-1}J^T$, and $\hat{s}_{(i)}$ is residual standard deviation calculated independently on the *ith* point, cf. ref. [16].

Nonlinear measure of an influence of the i-th point on the parameter estimates is represented by the *likelihood distance*

$$LD_i = 2[\ln L(b) - \ln L(b_{(i)})]. \tag{11}$$

In case of the least-squares the likelihood distance is expressed by

$$LD_i = n \ln\left[\frac{U(b_{(i)})}{U(b)}\right]. \tag{12}$$

In both Eqs. (11) and (12) the estimates $b_{(i)}$ calculated by nonlinear regression when the i-th point was left out or the one-step approximation $b^1_{(i)}$ of the parameter estimates may be used. When $LD_i > \chi^2_{1-\alpha}(2)$ is valid the i-th point is *strongly influential.* The significance level α is usually optioned to be equal to 0.05 then $\chi^2_{0.95}(2) = 5.992$.

(6) Data simulation: This block serves for debugging a program or for an examination of reliability of parameters estimation. For optional values of parameters, the "theoretical points" along the exact curve $A = f(pH; pK_{a,i}, \varepsilon_L, \varepsilon_{LH_i}, i = 1, \ldots, R)$ are calculated. Each theoretical point is then transformed into an "experimental" one by an addition of a random error (having obviously a normal distribution) obtained with the aid of a random-number generator. All resulting "experimental points" are thus corrupted with a random error. The error set can be then tested statistically for Gaussian distribution, independence and homogeneity. Statistical measures mentioned in residual analysis, $E(\hat{\varepsilon})$, $E|\hat{\varepsilon}|$, $s(\hat{\varepsilon})$, $g_1(\hat{\varepsilon})$, $g_2(\hat{\varepsilon})$ are tested.

Corrupting the curve points with high random error may, however, decrease the accuracy and precision of the parameters estimated. When several parameters are to be refined or ill-conditioned parameters are to be adjusted, data with a low precision may result in erroneous values of the parameter estimates if a reliable minimization method is applied. In cases when a corruption is small the parameters minimizing the least-squares criterion are near the same as optioned values but for very ill-conditioned models the differences can be high.

(7) Free concentration: This block concerns PSEQUAD only. The calculation of unknown free concentrations [L], [LH], ..., $[LH_r]$ is made using a standard Newton-Raphson procedure with Choleski's algorithm to solve linear equations. The free concentrations are calculated on a logarithmic scale so no negative concentrations may occur in the course of iterations.

(8) Additional: This block contains the visualization tools of ill-conditioning: the response-surface of the $U(\beta)$ hyperparaboloid being the 3D-graph of selected parameters in the neighborhood of the "pit", U_{min}, gives a visual representation of the influence of each parameter on $U(\beta)$. For two parameters optioned in the input, the paraboloid response-surface $(C-U(\beta))$ in 3D graph is plotted by DIGIGRAPH equipment [17] where C is a numerical constant. A regular paraboloid shape proves that both parameters are well-conditioned in a model and may lead to accurate and precise estimates whereas a "saucer" shape indicates ill-conditioned parameters which lead to rather uncertain estimates.

Residual sum of squares contours may also be plotted in the space of any two variables at a time by DCMINUIT. This gives a detailed description of the shape of the U function but only when the number of variables is very few, otherwise a calculation fails. The program DCMINUIT traces contours of constant value of U as a function of the two variable when all others being fixed at their value at that time.

b. Regression Procedure

Regression analysis and an examination of adequacy of the nonlinear model proposed with data is performed using following criteria [16]:

(1) The quality of parameter estimates: The quality of found parameter estimates is considered according to their confidence intervals or according to their variances $D(b_j)$. Often in solution equilibria the empirical rule is used: the parameter is considered to be significantly differing from zero when its estimate is greater than its 3 standard deviations, $3\sqrt{D(b_j)} < |b_j|$. High values of parameters variance is often caused by termination of minimization process before reaching a minimum.

(2) The quality of achieved curve fitting: The adequacy of a proposed model with experimental data is examined by the goodness-of-fit test based on the statistical analysis of classical residuals. Following statistical characteristics for a set of classical residuals are calculated: from the residual sum of squares $U(b)_{\min}$ reached at a minimum the estimate of residual variance $s^2(\hat{e})$ and estimates of the determination coefficient D^2, the regression rabat 100 D^2 in [%], the arithmetic mean of residuals $E(\hat{e})$, the robust median $\hat{e}_{0.5}$, the mean of absolute values of residuals $E|\hat{e}|$, the mean of absolute values of relative residuals in percents $100\ E|\hat{e}_{rel}|$, the residual standard deviation $s(\hat{e})$, the robust standard deviation of median $s(\hat{e}_{0.5})$, the residual skewness $g_1(\hat{e})$, the residual curtosis $g_2(\hat{e})$, Hamilton R-factor of relative fitness in percents and Akaike Information Criterion AIC are calculated.

(3) The quality of experimental data: For examination of a quality of data the identification of influential points by regression diagnostics is used. The most suitable diagnostics are the likelihood distances LD and Jackknife residuals $\hat{e}_J$.

Software

DCMINOPT having been applied from CHEMSTAT package [18] (Trilobyte, Pardubice) on IBM PC AT while other computations (DCFIT, DCMINUIT, PSEQUAD) were performed on the EC1033 computer at the Computing Centre of the University of Chemical Technology, Pardubice, Czech Republic.

Results and Discussion

As an analysis of the absorbance-pH curve namely concerning close overlapping protonation of a ligand related from a protolytic acid LH_R is not straightforward procedure resulting always at the true values of dissociation constants and molar absorptivities, some useful diagnostic tools of regression process were proposed. For demonstration of efficiency of this process, an example, simulated data of the A-pH curve of 4-CAPAZOXS were analyzed by DCMINOPT and results compared with those determined by three another regression programs, DCFIT, DCMINUIT and PSEQUAD.

Pre-selected ("true") values of seven parameters, $\beta_1, \ldots, \beta_7$, were chosen to be close to parameters for sulphoazoxine 4-CAPAZOXS: $pK_{a1} = 2.8$ $(=\beta_1)$, $pK_{a2} = 3.0$ $(=\beta_2)$, $pK_{a3} = 7.5$ $(=\beta_3)$, $\varepsilon_L = 12000$ $(=\beta_4)$, $\varepsilon_{LH} = 9800$ $(=\beta_5)$, $\varepsilon_{LH_2} = 9000$ $(=\beta_6)$, $\varepsilon_{LH_3} = 6000$ $(=\beta_7)$. The instrumental error of absorbance expressing a noise of spectrophotometer, $s_{inst}(A)$, was chosen 0.003. For set of 35 values pH, absorbance values were calculated precisely, then corrupted with random errors. A set of random errors should ideally exhibit a normal distribution with the mean $E(\hat{e})$ equal zero, the mean error $E|\hat{e}|$ equal to 0.003 as well as the error standard deviation $s(\hat{\varepsilon})$ 0.003, the skewness $g_1(\hat{\varepsilon})$ should be 0 and the curtosis $g_2(\hat{\varepsilon})$ 3. However, due to small sample size and properties of pseudo-random variable generation procedure the real errors are obviously not exactly normal and therefore a minimum of the least-squares is not reached at optioned parameters values.

In regression analysis of a A-pH curve, the reliability of regression process and estimates found can be classified according to a precision of parameters estimated and also on the base of a goodness-of-fit achieved. To test when the regression algorithm has found the best estimates of parameters, the residuals should be randomly distributed about the predicted regression curve as the systematic departures from randomness indicate that the parametric estimates are not satisfactory. To analyze residuals, their statistics are compared with the statistics of imposed random errors; it is checked whether both distributions are Gaussian in nature and/or sign. Even the degree-of-fit achieved by all regression methods is good enough and the minimization process was assumed to have terminated successfully there are some differences in estimates pK_{a1}, pK_{a2} and ε_{LH_2} from the true values.

The purpose of this paper is to demonstrate the procedure of investigation of a reliability of parameter estimation and how much minimization methods affects the precision and accuracy of the parameter estimates when other things being equal. The *systematic deviation* and the *relative systematic deviation of the parameter estimates* from its pre-selected value β_i called also the bias and the relative bias of parameter, $e(b_i) = \beta_i - b_i$ and $e_{rel}(b_i) = 100\ e(b_i)/b_i$ [in per cents], are used to classify an accuracy (or a *bias*) of the parameter estimates caused by inaccuracy of data. Parameters precision is considered from the standard deviation of estimates.

For pre-selected values of parameters, $pK_1, \ldots, \varepsilon_{LH_3}$, the corresponding sum of squares reaches the value $U(\mathbf{b}_0) = 2.870 \cdot 10^{-4}$. The program DCMINOPT terminates at a minimum $U(\mathbf{b}_0) = 2.512 \cdot 10^{-4}$ with the point estimates which do not quite agree with pre-selected values β (Table 1a). Standard deviation of each parameter $s(b_j)$ except $s(\varepsilon_{LH_2})$ reaches small value. Bias of each parameter $e(b_j)$ are not too high. For all seven parameters the interval estimate $b_j \pm \Delta_j$ contains a pre-selected value of β_j. Statistical test says that all parameters except ε_{LH_2} are significantly different from zero.

A graphical representation of elliptic hyperparaboloid being simplified for two chosen parameters i.e. for two parametric coordinates, $m = 2$, in $(m + 1)$-dimensional space may be applied. In Fig. 1a a well-developed maximum $(1 - U(\beta))$ shows that both parameters ε_L and ε_{LH} are well-conditioned in model while the shape of a hyperparaboloid for the ill-conditioned parameters is cylindrical or flat-bottomed saucer. The cylindrical shape in Fig. 1b indicates that the parameter pK_{a1} is strongly ill-conditioned as the dependence of $(1 - U(\beta))$ on pK_{a2} is weak and nearly constant. When a dependence of $(1 - U(\beta))$ on parameters, pK_{a1} and ε_{LH_2} in Fig. 1c, is weak and an obvious maximum does not exist we say that both parameters are ill-conditioned in model. The shape of such hyperparaboloid cannot be improved and the pit also cannot be reached by any minimization method. A search for true estimates of the parameters then cannot give a certain answer, and no method is able safely to find a pit in U. Careful choice of a minimization algorithm and also of a minimization strategy is necessary because some algorithms easily fail or diverge. The hyperparaboloid response surface shows that three parameters, pK_{a1},

Table 1. Regression analysis of simulated 35 points of A-pH curve for 4-CAPAZOXS calculated for pre-selected parameters $pK_{a1} = 2.8$, $pK_{a2} = 3.0$, $pK_{a3} = 7.5$, $\varepsilon_L = 12000$, $\varepsilon_{LH} = 9800$, $\varepsilon_{LH_2} = 9000$, $\varepsilon_{LH_3} = 6000$ and corrupted with random errors generated for $s_{inst}(A) = 0.003$. Conditions: $L = 3.65 \cdot 10^{-5}$, $d = 1.000$ cm, $S = 59.16$ mV/pH, 298.16 K, pH(st) = 7.010.

(a) Point and interval estimates of parameters with their statistical characteristics calculated by DCMINOPT. Accuracy is expressed by the bias of each parameter $e(b_i)$

			Half-length of confidence interval		
Parameter β_j	Point estimate b_j	Standard deviation $s(b_j)$	Δ_j	$\Delta_{R,j}$	Bias of parameter $e(b_j)$
pK_{a3}	7.4678	0.0543	±0.1346	±0.2209	0.0322
pK_{a2}	2.8375	0.2565	±0.9784	±1.0425	0.1625
pK_{a1}	2.8380	0.0878	±0.2338	±0.3569	−0.0380
ε_{LH_3}	6013.5	105.10	±301.44	±427.10	−13.5
ε_{LH_2}	8742.4	114.30	±4645.1	±4645.2	257.6
ε_{LH}	9791.9	37.56	±119.28	±152.64	8.1
ε_L	12009.0	72.13	±245.65	±293.11	−9.0

(b) Matrix of paired correlation coefficient of parameters, r_{ij}, calculated by DCMINOPT

	pK_{a3}	pK_{a2}	pK_{a1}	ε_{LH_3}	ε_{LH_2}	ε_{LH}	ε_L
pK_{a3}	1.000	−0.180	−0.324	0.113	−0.293	0.575	0.799
pK_{a2}		1.000	0.220	−0.823	0.938	−0.315	−0.088
pK_{a1}			1.000	0.137	0.524	−0.565	−0.158
ε_{LH_3}				1.000	−0.703	0.197	0.055
ε_{LH_2}					1.000	−0.512	−0.143
ε_{LH}						1.000	0.282
ε_L							1.000

(c) Analysis of random errors with classical residuals and identification of influential points by DCMINOPT

i	Independ. variable pH	Depend. variable A_{exp}	Random error ε	Classical residual $\hat{e}$	Jackknife residual $\hat{e}_J$	Likelihood distance LD
1	1.650	0.2232	−0.0034	−2.8787E−03	−1.4158E+00	2.8354E−02
2	1.790	0.2294	0.0000	7.7375E−04	8.5502E−01	4.0363E−03
3	1.930	0.2336	0.0004	1.4463E−03	1.1072E+00	7.0172E−03
4	2.070	0.2410	0.0027	3.9933E−03	1.7521E+00	5.1190E−02
5	2.210	0.2419	−0.0031	−1.6943E−03	−1.0525E+00	1.1430E−02
6	2.350	0.2519	−0.0018	−4.2087E−04	−4.6181E−01	3.0701E−03
7	2.490	0.2614	−0.0030	−2.0252E−03	−1.1623E+00	1.6926E−02
8	2.630	0.2741	−0.0029	−2.6317E−03	−1.3369E+00	2.2450E−02
9	2.770	0.2941	0.0034	2.6616E−03	1.4440E+00	1.9853E−02
10	2.910	0.3083	0.0036	2.1141E−03	1.2976E+00	9.4042E−03

Table 1 (continued)

i	Independ. variable pH	Depend. variable A_{exp}	Random error $\hat{\varepsilon}$	Classical residual $\hat{e}$	Jackknife residual $\hat{e}_J$	Likelihood distance LD
11	3.050	0.3224	0.0047	2.8809E−03	1.4993E+00	1.6520E−02
12	3.190	0.3269	−0.0017	−3.5572E−03	−1.5668E+00	4.2023E−02
13	3.330	0.3363	−0.0010	−2.4381E−03	−1.2815E+00	1.8957E−02
14	3.470	0.3413	−0.0021	−3.3529E−03	−1.5188E+00	4.8987E−02
15	3.610	0.3508	0.0029	2.0726E−03	1.2949E+00	8.8326E−03
16	3.750	0.3527	0.0018	1.2157E−03	1.0370E+00	4.0654E−03
17	3.890	0.3566	0.0035	3.2580E−03	1.5963E+00	1.9786E−02
18	4.030	0.3573	0.0028	2.6998E−03	1.4652E+00	1.2899E−02
19	4.170	0.3530	−0.0025	−2.4623E−03	−1.2825E+00	1.2234E−02
20	4.700	0.3618	0.0046	4.8013E−03	1.9137E+00	8.4731E−02
21	5.230	0.3542	−0.0037	−3.5135E−03	−1.5555E+00	3.9657E−02
22	5.760	0.3598	0.0008	8.8557E−04	9.1879E−01	4.0516E−03
23	6.290	0.3595	−0.0027	−2.9304E−03	−1.4096E+00	1.9412E−02
24	6.475	0.3621	−0.0024	−2.7643E−03	−1.3652E+00	1.5015E−02
25	6.600	0.3642	−0.0035	−2.8599E−03	−1.3900E+00	1.5356E−02
26	6.845	0.3777	0.0055	4.7239E−03	1.9012E+00	6.7013E−02
27	7.030	0.3800	0.0021	9.6248E−04	9.5007E−01	4.0385E−03
28	7.215	0.3841	−0.0010	−2.3098E−03	−1.2439E+00	1.5466E−02
29	7.400	0.3976	0.0045	2.8871E−03	1.5037E+00	2.6048E−02
30	7.585	0.4018	0.0001	−1.4918E−03	−9.7771E−01	7.2642E−03
31	7.770	0.4121	0.0023	6.9906E−04	8.5737E−01	3.2193E−03
32	7.955	0.4151	−0.0019	−3.3463E−03	−1.5139E+00	2.9689E−02
33	8.140	0.4261	0.0032	1.9672E−03	1.2645E+00	1.1547E−02
34	8.325	0.4315	0.0041	3.0442E−03	1.5357E+00	3.3908E−02
35	8.510	0.4291	−0.0016	−2.4951E−03	−1.3092E+00	2.0889E−02

Goodness-of-fit test

	Errors	Residuals
Bias, $E(\hat{e})$	4.3E−4	−2.4E−6
Median, $\hat{\varepsilon}_{0.5}$ or $\hat{e}_{0.5}$	0.0003	0.0007
Standard deviation of median, $s(\hat{e}_{0.5})$	0.0050	0.0048
Mean of absolute values of ..., $E\lvert\hat{e}\rvert$	0.0026	0.0025
Mean of abs. values of relative ..., $100\lvert\hat{e}\rvert$, [%]	0.776	0.775
Variance, $s^2(\hat{\varepsilon})$ or $s^2(\hat{e}) \cdot 10^6$	8.564	8.972
Standard deviation, $s(\varepsilon)$ or $s(\hat{e})$	0.0029	0.0030
Skewness, $g_1(\hat{\varepsilon})$ or $g_1(\hat{e})$	0.140	0.171
Kurtosis, $g_2(\hat{\varepsilon})$ or $g_2(\hat{e})$	1.543	1.573
Sum of squares, $ESS \cdot 10^4$ or $RSS \cdot 10^4$	2.398	2.512
Regression rabat, $100 \cdot D^2$, [%]	*	99.805
Akaike Information Criterion, AIC	*	−392.75
Hamilton R-factor, [%]	0.81	0.71
Normality test, H_0: $\{\hat{\varepsilon}\}$ or $\{\hat{e}\}$ have normal distribution, $\chi^2_{1-\alpha}(2) = 5.992$		
χ^2_{exp}:	3.635	3.550
Independence test, H_0: $\{\hat{\varepsilon}\}$ or $\{\hat{e}\}$ are independent, $t_{1-\alpha/2}(35 + 1) = 2.028$		
t_{exp}:	0.007	0.850

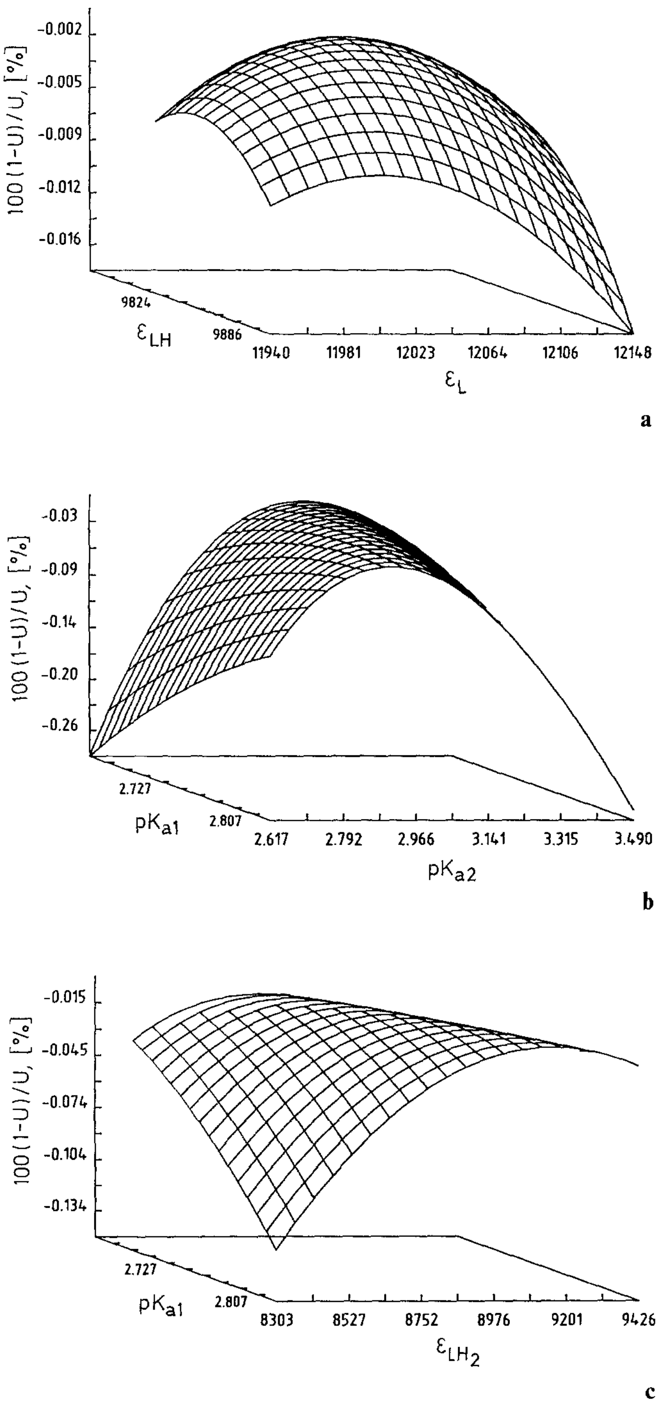

Fig. 1. The 3D graph of the $(1 - U(\beta))$ response surface for A-pH data from Table 1 indicates (**a**) that ε_L and ε_{LH} are well-conditioned in model because the surface exhibits an obvious maximum; (**b**) two ill-conditioned parameters pK_{a1} and ε_{LH_2}. For both cases, (**b**) and (**c**), there is no well-developed obvious maximum $(1 - U(\beta))$

pK_{a2} and ε_{LH_2} are ill-conditioned because the minima are broad and indefinite so that these parameters cannot be determined accurately.

The last contour (so called D-boundary in Sillen's terminology [3]) expressed as the supercurve $U = U_{min} + s^2(A)$ serves as estimation of the standard deviation in each parameter b_i. The statistic $\Delta_{R,j}$ represents the maximum difference between the value for b_i at any point on the D-boundary, and the value for b_i at the minimum. Because for the ill-conditioned parameters the response surface resembles a large

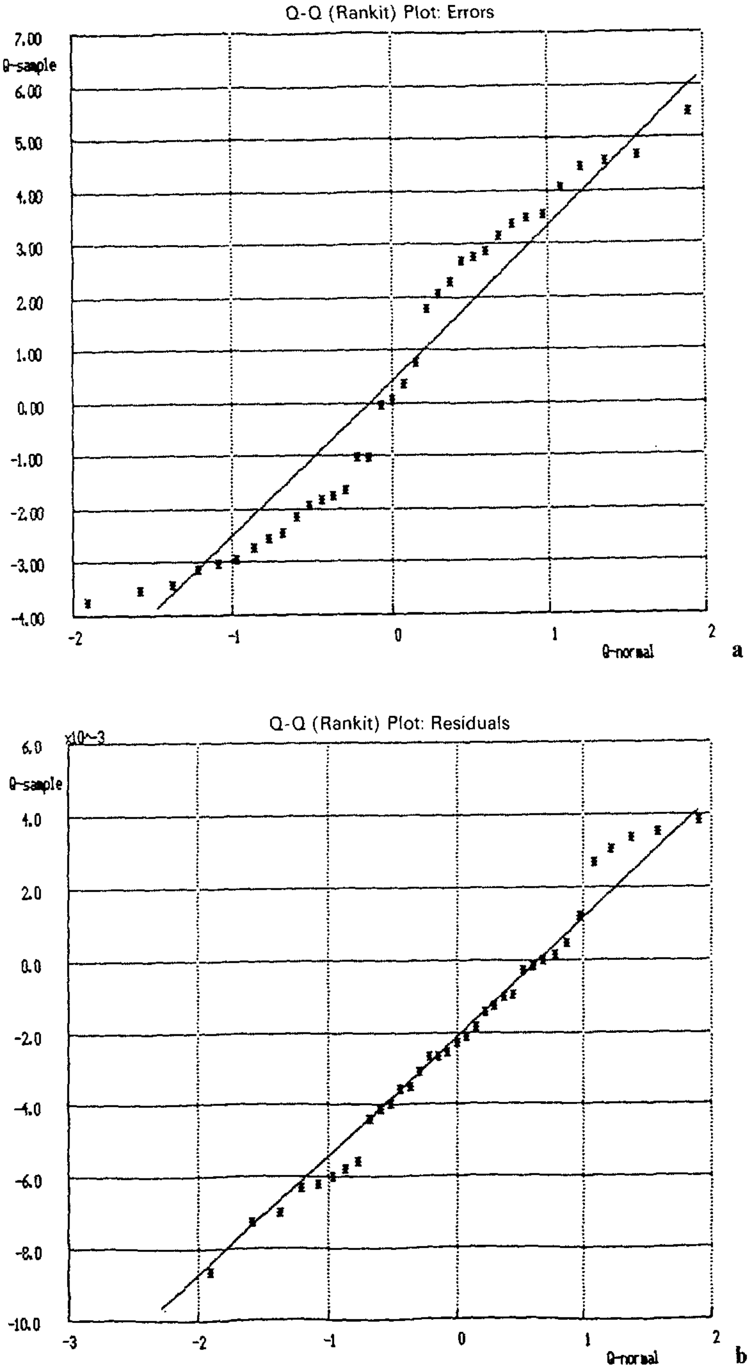

Fig. 2. Quantile-quantile (rankit) plot of the sample of (**a**) generated random errors, and (**b**) residuals proves that both samples come from the one common population

flat-bottomed saucer, the standard deviations will have significantly greater values than those for the well-conditioned parameters. It may be therefore concluded, the larger values of s(ε_{LH_2}), s(pK_{a1}) and s(pK_{a2}) express a large amount of uncertainty in a location of the pit while s(ε_L), s(ε_{LH}), s(ε_{LH_3}) and s(pK_{a3}) concern well-conditioned parameters which lead to a pronounced maximum (1 − U(β)).

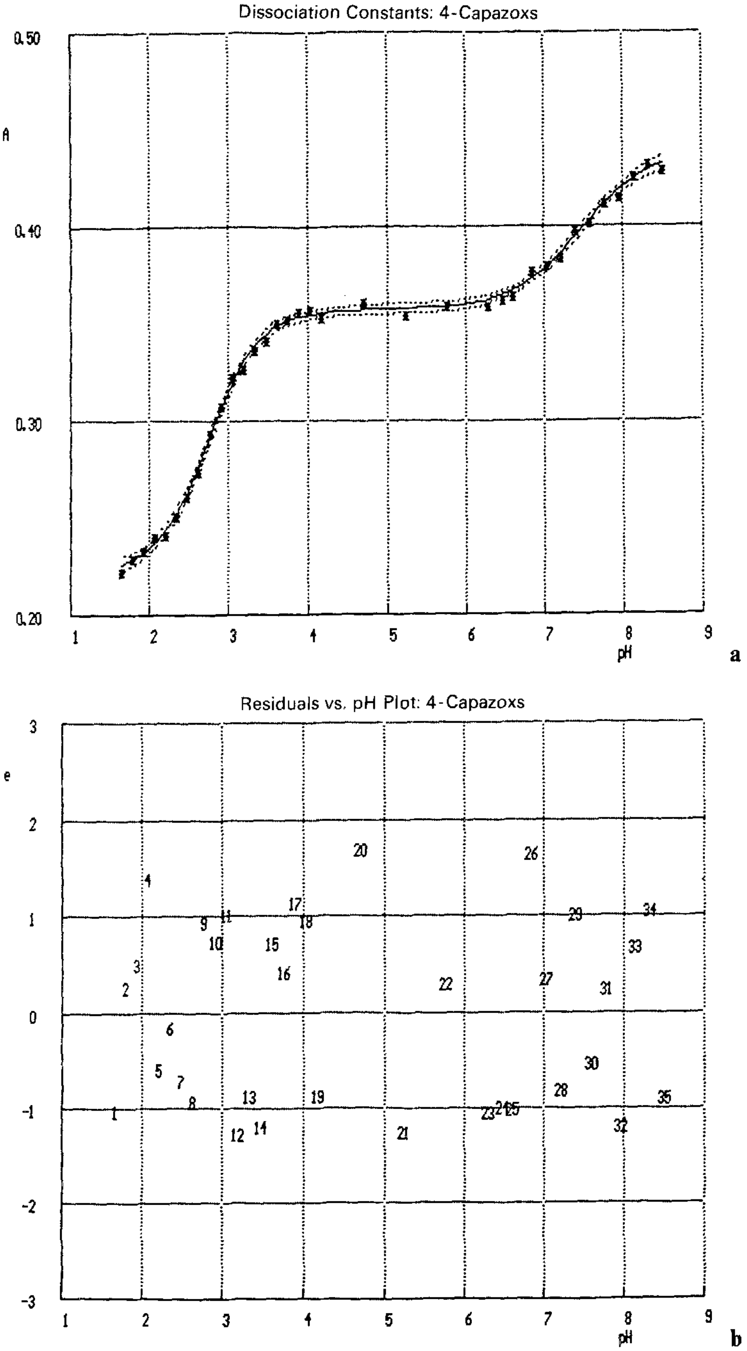

Fig. 3. a Curve-fitting for the A-pH dependence, and (b) scatter plot of residuals on the independent variable pH

The paired correlation coefficients of two parameters in Table 1b indicate quite strong correlation of following pairs: $\varepsilon_L - pK_{a3}$ being 0.799, $pK_{a2} - \varepsilon_{LH_2}$ being 0.938, $pK_{a1} - \varepsilon_{LH_2}$ being 0.524, $pK_{a2} - \varepsilon_{LH_3}$ being −0.823. A high correlation may be elucidated as a flat shape of the maximum $(1 - U(\beta))$ in Fig. 1 while a small correlation between two parameters proves their independence and correspondence to a well-developed maximum $(1 - U(\beta))$.

Goodness-of-fit test (Table 1c) analyses random errors and residuals and indicates that sufficiently close fit was achieved: the statistical measures of residuals are close to those of random errors. Moreover, the residual standard deviation $s(\hat{e}) = 0.0030$ are of same magnitude as the instrumental error $s_{inst}(A) = 0.003$ leading to $s(\hat{\varepsilon}) = 0.0029$. Certain underlying assumptions of regression analysis as an independence of random errors and residuals $(t_{exp} < t_{1-\alpha/2}(35 + 1))$, normal distribution for errors and residuals $(\chi^2_{exp} < \chi^2_{1-\alpha}(2))$, skewness $g_1(\hat{\varepsilon})$ or $g_1(\hat{e})$ should be zero and curtosis $g_2(\hat{\varepsilon})$ or $g_2(\hat{e})$ should be 3. The residuals should possess all these statistics that agree or at least do not refute characteristics of errors. Quantile-quantile (rankit) plot of random errors (Fig. 2a) and residuals (Fig. 2b) indicates some deviation from a normal distribution of both quantities. Due to small sample size the errors do not exhibit the correct straight line. The effect of "supernormality" (cf. ref. [16]) causes that residuals are more normal than errors.

Hamilton R-factor of relative fitness, regression rabat D^2 and Akaike Information Criterion AIC in Table 1c also enable to monitor the regression process. In minimum U_{min} the R-factor and AIC reach a minimal value while D^2 the maximal one. No influential points (i.e. outliers and high-leverages) were detected by $\hat{e}_J$ and LD as no jackknife residuals $\hat{e}_J$ is higher than 3 and no points fulfilled a condition that $LD_i > \chi^2_{1-\alpha}(2) = 5.992$.

Confidence interval of prediction A_{calc} (Fig. 3a) and the scatter plot of residuals in dependence on the independent variable pH (Fig. 3b) proves sufficiently close fitting calculated regression A-pH curve through experimental points.

Comparing regression of simulated data by four different programs in Table 2, two criteria were applied: (a) the relative systematic deviation of each parameter, (or the *relative bias*) $e_{rel}(b_j)$ in [%], and (b) the goodness-of-fit test.

The lowest bias from an pre-selected value of each parameter cannot be used for identification of accuracy due to non-idealities of random error corruption. This is evident from value $U(\mathbf{b}_0) = 2.87 \cdot 10^{-4}$ for pre-selected values of parameters which is greater than a minimum of sum of squares $U(\mathbf{b}_0) = 2.5121 \cdot 10^{-4}$. Programs DCFIT, DCMINUIT and PSEQUAD lead to inaccurate selection of minimum (Table 2). For all programs the same initial guess of parameters have been used.

Conclusion

In case of closely overlapping protonation equilibria, an estimation of near consecutive dissociation constants is not straightforward and easy. Regression diagnostics enable to examine reliability of refined parameters even for cases of near dissociation constants which are always ill-conditioned in model. A bias of parameters estimates from pre-selected values may be considered from a deviation of each estimate from

Table 2. Regression analysis of simulated A-pH curve from Table 1 using various regression algorithms and examination of reliability of estimated ill-conditioned parameters. Standard deviations of parameters estimates are in parentheses being expressed in last valid digits. Accuracy is expressed by the bias of each parameter from its given value, e(b$_i$), in percents

(a) Parameters estimates refined by various regression algorithms:

Parameters are:	Kept constant	Refined	Refined	Refined	Refined
Algorithm used:	MINOPT	MINOPT	FIT	MINUIT	PSEQUAD
Found $U_{min} \cdot 10^4$	2.870	2.512	10.582	10.480	10.453
Parameters given Relative bias, [%]	Parameters estimates				
pK_{a3} (=7.500)	7.500 (60)	7.468 (54)	7.465 (110)	7.469 (110)	7.442 (109)
e_{rel} (pK_{a3})	0	−0.43	−0.47	−0.41	−0.77
pK_{a2} (=3.000)	3.000 (335)	2.837 (256)	2.920 (550)	2.766 (485)	3.079 (578)
e_{rel} (pK_{a2})	0	−5.50	−2.67	−7.80	2.63
pK_{a1} (=2.800)	2.800 (87)	2.838 (88)	2.849 (153)	2.816 (208)	2.862 (129)
e_{rel} (pK_{a1})	0	5.43	5.03	0.53	2.07
ε_L (=12000)	12000 (80)	12009 (72)	12011 (146)	12002 (145)	11990 (142)
e_{rel} (ε_L)	0	−1.28	0.09	0.02	−0.08
ε_{LH} (=9800)	9800 (41)	9792 (38)	9786 (78)	9799 (74)	9769 (81)
e_{rel} (ε_{LH})	0	−1.43	−0.14	−0.01	−0.11
ε_{LH_2} (=9000)	9000 (1106)	8742 (1143)	9101 (1905)	8374 (2649)	9570 (1244)
e_{rel} (ε_{LH_2})	0	−4.29	1.12	6.96	6.33
ε_{LH_3} (=6000)	6000 (107)	6013 (105)	5990 (200)	6031 (228)	5959 (180)
e_{rel} (ε_{LH_3})	0	−1.12	−0.17	0.52	−0.68

(b) Goodness-of-fit test for various regression algorithms:

Parameters are: Algorithm used:		Kept constant MINOPT	Refined MINOPT	Refined FIT	Refined MINUT	Refined PSEQUAD
	Random errors	Residuals				
$\hat{e}_{0.5}$	0.0003	0.0000	0.0007	−0.0043	−0.0043	−0.0046
$s(\hat{e}_{0.5})$	0.0050	0.0048	0.0048	0.0046	0.0046	0.0051
$E(\hat{e})$	4.3E−4	3.6E−4	−2.4E−6	−4.7E−3	−4.7E−3	−4.7E−3
$E\|\hat{e}\|$,	0.0026	0.0026	0.0025	0.0048	0.0048	0.0047
$100\ E\|\hat{e}\|$, [%]	0.775	0.775	0.742	1.390	1.392	1.373
$s^2(\hat{e}) \cdot 10^6$	8.564	10.251	8.972	37.792	37.430	37.331
$s(\hat{e})$	0.0029	0.0032	0.0030	0.0061	0.0061	0.0061
$g_1(\hat{e})$	0.140	0.164	0.171	0.114	0.082	0.182
$g_2(\hat{e})$	1.543	1.553	1.573	1.735	1.736	1.758
$RSS \cdot 10^4$	2.398	2.870	2.512	10.582	10.480	10.453
$100 \cdot D^2$, [%]	*	99.777	99.805	99.178	99.186	99.188
Akaike AIC	*	−395.90	−392.75	−350.23	−350.57	−350.66
Normality test, H_0: $\{\hat{\varepsilon}\}$ or $\{\hat{e}\}$ have normal distribution, $\chi^2_{1-\alpha}(2) = 5.992$						
χ^2_{exp}:	3.635	3.630	3.550	2.728	2.683	2.762
Independence test, H_0: $\{\hat{\varepsilon}\}$ or $\{\hat{e}\}$ are independent, $t_{1-\alpha/2}(35 + 1) = 2.028$						
t_{exp}:	0.007	0.032	0.850	0.115	0.078	0.060

its pre-selected value while the precision from its standard deviation. A reliability of regression process being examined by the goodness-of-fit test seems to be best when DCMINOPT is applied.

References

[1] M. Meloun, M. Javůrek, *Talanta* **1985**, *32*, 973.

[2] M. Meloun, J. Čermák, *Talanta* **1979**, *26*, 569.

[3] L. G. Sillén, B. Warnqvist, *Ark. Kemi* **1969**, *31*, 377.

[4] M. Meloun, J. Chýlková, *Collect. Czech. Chem. Commun.* **1979**, *44*, 2815.

[5] M. Meloun, J. Chýlková, M. Bartoš, *Analyst* **1986**, *111*, 1189.

[6] M. Meloun, J. Čermák, *Talanta* **1984**, *31*, 947.

[7] M. Meloun, J. Havel, *Computation of Solution Equilibria, Part 1, Spectrophotometry*, Folia UJEP, 1985.

[8] J. Havel, M. Meloun, in: *Computational Methods for the Determination of Formation Constants* (D. J. Leggett, ed.), Plenum, New York, 1985, p. 221.

[9] M. Meloun, J. Havel, E. Hoegfeldt, *Computation of Solution Equilibria, A Guide to Methods in Potentiometry, Extraction and Spectrophotometry*, Ellis Horwood, Chichester, 1988, p. 81.

[10] M. Meloun, M. Javůrek, J. Havel, *Talanta* **1986**, *33*, 513.

[11] L. Zékány, I. Nagypal, in: *Computational Methods for the Determination of Formation Constants* (D. J. Leggett, ed.), Plenum, New York, 1985, p. 291.

[12] Previous part of this series: M. Meloun, M. Javůrek, J. Militký, *Mikrochim. Acta* **1992**, *109*, 221.

[13] J. Lang, R. Muller, *Comp. Phys. Commun.* **1971**, *2*, 79.

[14] F. James, and M. Ross, *Comp. Phys. Commun.* **1976**, *10*, 343.

[15] J. Militký, M. Meloun, *Talanta* **1993**, *40*, 269 and 279.

[16] M. Meloun, J. Militký, *Chemometrics for Analytical Chemistry, Part 2, Interactive Model Building and Testing*, Ellis Horwood, Chichester, 1993.

[17] M. Javůrek, *PhD Thesis*, University of Chemical Technology, Pardubice, 1988.

[18] J. Militký, J. Čáp, *Proceedings Conf. CEF'87*, Taormina, Sicilia, May 1987.

Received October 16, 1991.