Extracting Experimental Information from Large Matrices. 2. Model-Free Resolution of Absorbance Matrices: M³

Gábor Peintler,*,† István Nagypál,† Irving R. Epstein,*,‡ and Kenneth Kustin‡

Institute of Physical Chemistry, University of Szeged, H-6701 Szeged, P.O. Box 105, Hungary, and Department of Chemistry, Brandeis University, MS 015, Waltham, Massachusetts 02454-9110

Received: November 5, 2001

We present a new method for the decomposition of an experimental absorbance matrix into concentration and molar absorption coefficient matrices. The decomposition in general is not unambiguous; therefore, the method may yield only ranges for these matrices. The experimental matrix is not changed, so deviations from the original data can be monitored element by element. Consequently, all chemical constraints (such as stoichiometry) can be taken into account. The method, which does not require an explicit chemical model, is used to analyze the reaction between a Co(II)-EDTA complex and H_2O_2 .

I. Introduction

Advanced spectroscopic instruments routinely produce large experimental data sets in matrix form. A new algorithm for handling these matrices and for determining the number of independent absorbing species (NIAS) was reported in part 1.¹ Once NIAS is known, the next step is usually the decomposition of the experimental matrix into concentration and molar absorbance matrices, based on the Beer–Lambert law (or its analogues):

$$\mathbf{A}(p \times q) = \mathbf{C}(p \times n) \times \mathbf{E}(n \times q) \tag{1}$$

where A is the experimental matrix, C is the concentration matrix, and E is the matrix of molar absorption coefficients. The dimensions of these matrices are given in parentheses, where n, p, and q are the numbers of absorbing species, samples, and wavelengths, respectively.

Methods used to perform the decomposition fall into either of two classes, which differ in principle. Model-based approaches posit a mathematical model that describes the relation among the concentrations or between the concentrations and their time derivatives. These methods most frequently use a least squares technique^{2–4} to calculate the parameters of the model, e.g., formation constants in equilibrium studies or rate constants in kinetics. Finding an appropriate model in complex kinetic or equilibrium systems requires intuition, chemical instinct, experience, and sometimes a bit of luck. In these calculations, instead of the $p \times n$ individual concentration data, one calculates only a few chemical parameters that describe the relations among them.

The second category of methods, to which the approach developed here belongs, is model-free. If a model-free solution for the concentrations is known, then further evaluation, i.e., setting up an appropriate chemical model, will be much easier than using any model-based method alone. Model-free decomposition is typically based on factor analysis (FA) and its offshoots,⁵ mainly principal component analysis (PCA).

In part 1, we detailed obstacles to the evaluation of large experimental matrices in connection with matrix rank analysis (MRA). These problems, particularly interdependence of the primary data, must be considered when FA is used as well. Additional issues should be taken into account when applying such techniques. The most important of these problems is that the solution of eq 1 is rarely, if ever, unique. One trivial example is that if a C-E matrix pair is a solution for eq 1 then the $\alpha \times$ $C-E/\alpha$ matrix pair also satisfies eq 1 where α is any positive constant. This difficulty can be remedied by introducing additional, often chemically evident, constraints (e.g., the sum of the concentrations of absorbing species is constant). However, the application of stoichiometric constraints is hardly ever sufficient to yield a unique solution, which can be calculated only if the measured matrix contains independent experimental information for every species. A unique solution cannot be extracted if the elements of the matrix A (i.e., the measured absorbances) are determined by more than one absorbing species.

Almost all variants of FA are based on an eigenvalue calculation for the matrix $\mathbf{A}\mathbf{A} = \mathbf{A}^T \times \mathbf{A}$. We have shown in part 1 that this approach may be misleading, because the statistical criteria of the eigenvalue calculation are not valid for large matrices originating from spectroscopic data acquisition systems. The nonnegativity of the elements of the absorbance matrix is an important constraint in most experimental methods (e.g., UV—vis spectrophotometry). A negative element may only be accepted if its absolute value is smaller than the experimental error.

The aim of the present work is to introduce a new algorithm to overcome the difficulties outlined above. The new method is called \mathbf{M}^3 (model-free modeling with matrices). As we shall see, the method is capable of determining NIAS and calculating a large number of possible \mathbf{C} and \mathbf{E} matrix pairs. The algorithm is illustrated on a real example, the reaction of a cobalt(II)— EDTA (EDTA = ethylenediaminetetraacetate) complex with hydrogen peroxide.

II. The M³Algorithm

The goal of M³ is to minimize the target function

$$\mathbf{Sn} (\mathbf{C}(n), \mathbf{E}(n)) = \sum_{i=1}^{p} \sum_{j=1}^{q} |A_{ij} - \sum_{k=1}^{n} c_{ik} \epsilon_{kj}|$$
 (2)

^{*} To whom correspondence should be addressed.

[†] University of Szeged.

[‡] Brandeis University.

where c_{ik} and ϵ_{kj} are the concentration and molar absorption coefficient, respectively, of the *k*th species in the *i*th sample and at the *j*th wavelength. **Sn** indicates that the assumed NIAS is n.

The proposed procedure uses the data of the original experimental matrix, avoiding any transformation, to assign deviations to individual points; it is numerically stable, even in the case of singular matrices; and it allows for the incorporation of chemical constraints and any independent information such as known molar absorbances.

To meet these requirements, the following two-level procedure is employed. Initially, the elements of the **E** matrix are estimated. The estimation requires some experience to choose an appropriate initial guess, but it is not crucial. Formally, the c_{ik} values are the only unknowns in eq 2. Because different rows of the **A** matrix depend on different concentration values and each concentration value occurs in only one row, **Sn** is the sum of the following mathematically independent expressions

$$\mathbf{Sn}(\mathbf{C}(n)) = \sum_{i=1}^{p} Sn_{i}(c_{i1}, c_{i2}, ..., c_{in}),$$

$$Sn_{i}(c_{i1}, c_{i2}, ..., c_{in}) = \sum_{i=1}^{q} |A_{ij} - \sum_{k=1}^{n} c_{ik}\epsilon_{kj}| \quad (3)$$

To minimize Sn_i, concentrations for the following linear equation system (LES) are calculated:

$$A_{ij} = \sum_{k=1}^{n} c_{ik} \epsilon_{kj}, \quad j \in \{1, ..., q\}$$
 (4)

Additional equation(s) must be added to eq 4 if there are stoichiometric constraints on the concentrations. For example, if the sum of the concentrations of the absorbing species is constant, the following equation must also be taken into account:

$$T_i = \sum_{k=1}^{n} c_{ik} \tag{5}$$

where T_i is the total concentration in the *i*th sample.

If the number of equations in eqs 4 and 5 is equal to NIAS, then the concentrations are given "exactly" and Sn_i is zero. Otherwise, the LES to be solved is overdetermined, and Sn_i is calculated after solving the LES.

Once Sn_i has been determined for all rows, Sn and the matrix C are known at a fixed E. Thus, Sn can be regarded as a function of E (Sn = Sn(E)) only, because the concentrations are now fixed. By finding the minimum of Sn(E), the elements of the E matrix can also be calculated. The minimization can be performed by any numerical method of nonlinear parameter estimation, because Sn is a continuous function of the elements of E. During parameter estimation, any element of E can be fixed or omitted. If statistical features, especially the correlation matrix, are taken into consideration, any relationships among the elements of E can also be found.

Alternatively, if **A** has more columns than rows, the concentration matrix can be fixed first, and then the elements of matrix **E** can be calculated column by column.

III. Calculations

Programs for M³ were developed in Borland Pascal for IBM PC (DX-486) compatible computers. Standard mathematical

procedures² were applied wherever possible. The calculations were performed with at least 19 significant digits.

A. Preparation of the Experimental Matrix. Before applying M³, one must estimate a range of plausible values for NIAS, adjust the experimental data to unit cell length if necessary, and remove any erroneous rows or columns. The size of the matrix should be kept as small as possible. The optimal number of wavelengths is one or two greater than the largest possible NIAS. In this way, the LESs to be solved are overdetermined, and the number of parameters to be fitted is not unnecessarily high. Those wavelengths of the original matrix evaluation that carry the important experimental information should be kept for further evaluation. Preparation of the matrix for M³ may easily be carried out by using the MRA technique¹.

B. Solving LESs: A combination of two numerical methods is used to calculate the concentrations at the initially guessed or iteratively refined E. First, the singular value decomposition (SVD) algorithm² is applied. It is fast, can handle singular matrices, and yields residuals that may be either positive or negative. The price paid for speed is that the procedure is not stable when **E** is far from the final solution. The instability may be manifested in negative concentrations and/or in failure to find a solution. Should a negative value occur, we may either (a) reject it and replace the singular value method by the slower linear optimization, which guarantees a positive concentration or (b) accept a negative value for c_{ik} if the calculated negative concentrations are small and $|c_{ik}\epsilon_{kj}|$ is negligible compared to the experimental errors at all wavelengths j. The final calculated concentrations should be same within experimental error for either choice.

If the above condition for negative concentrations is not fulfilled or if no solution is found, then instead of the LES defined in eq 4 the following system of inequalities is employed:

$$A_{ij} \ge \sum_{k=1}^{n} c_{ik} \epsilon_{kj}, \quad j \in \{1, ..., q\}$$

$$c_{ik} \ge 0.0, \quad k \in \{1, ..., n\}$$
(6)

The inequalities (6) may be supplemented by chemical constraints such as eq 5. The problem is then to find the maximum of the target function

$$Zn_{i}(c_{i1}, c_{i2}, ..., c_{in}) = \sum_{j=1}^{q} \sum_{k=1}^{n} c_{ik} \epsilon_{kj} = \sum_{k=1}^{n} (c_{ik} \sum_{j=1}^{q} \epsilon_{kj}),$$

$$i \in \{1, ..., p\} \quad (7)$$

subject to the conditions in eq 6.

The task defined by eqs 6 and 7 is a standard linear optimization problem. Our programs use the simplex algorithm² for calculating the concentrations. This algorithm is slow (the SVD algorithm runs about five times faster), and all of the residuals are positive, but the procedure is very stable and is independent of the elements of **E**. It therefore always finds an appropriate solution.

The "mixing" of two different target functions for solving the same problem raises the question of whether the results are the same, or at least compatible with further steps. From the experimental side, it is evident that an increase in experimental accuracy decreases the difference between the results. Computationally, we apply the following procedure. When E is still a crude approximation, only the simplex algorithm is used. Near the end of the fitting process, when SVD is also stable, we switch to this algorithm. The final results are calculated by SVD.

TABLE 1: Determination of the Stoichiometric Constraint in the Co^{II}-EDTA²⁻-H₂O₂ Reaction by Using Different Applications of MRA^a

used series	i	1	2	3	4	5	6
1-14	$abs(P_i)$	1.2890	0.8732	0.1223	0.0197	0.0101	0.0021
1-14, one row subtracted	mean of $1680 \text{ abs}(P_i)$	0.9157	0.5776	0.0868	0.0205	0.0107	0.0025
1-8, 13, 14	$abs(P_i)$	1.2890	0.8732	0.1223	0.0197	0.0101	0.0019
1-8, 13, 14, one row subtracted	mean of $1200 \text{ abs}(P_i)$	0.9394	0.4969	0.0197	0.0149	0.0035	0.0023

^a P_i denotes the *i*th element of the **P** vector which includes the largest residuals of the successive elimination steps (see part 1 for the detailed definition).

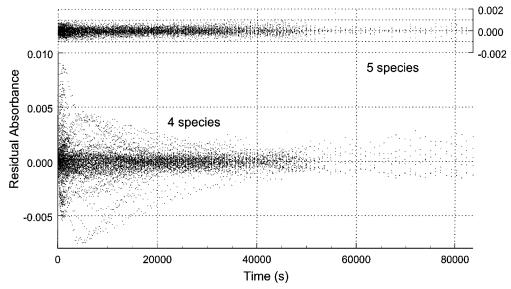


Figure 1. Differences between measured and calculated absorbances in the Co^{II}-EDTA²⁻-H₂O₂ reaction. The lower curves were calculated with four absorbing species, and the upper curves were calculated with five. The constraint given in eq 8 was applied.

C. Parameter Estimation. Once the concentrations are known with the initial or a refined E, the minimum of Sn(E) is sought by the downhill simplex method.^{2,6} This method is slow but requires the weakest set of assumptions about the surface to be searched. It is probably the most reliable method for seeking minima. When the minimum is reached, the final iteration and calculation of statistical parameters are carried out by the Gauss-Newton-Marquardt method^{2,7} with a numerical Jacobian matrix.

IV. An Example

We illustrate the application of M³ with a real example, the reaction of a cobalt(II)-EDTA complex with hydrogen peroxide, which is treated in detail in part 1. One absorbing species in the system is Co^{II}-EDTA²⁻. Its molar absorption coefficients have been determined independently⁸ as 0.6403, 0.3910, 0.3349, 9.579, 4.705, and 2.184 M^{-1} cm⁻¹ at 340, 350, 360, 520, 550, and 590 nm, respectively. These data were used as fixed parameters in E.

A. MRA as a Preparatory Method for M³. In part 1, we found that MRA alone cannot distinguish between NIAS = 4and NIAS = 5. The residual absorbance curves (obtained from a series of MRA), however, suggested that NIAS is 5. The use of M^3 , therefore, was restricted to NIAS = 4 and to NIAS = 5. The example matrix includes six columns, so the LESs to be solved are always overdetermined.

The relevant stoichiometric constraint in the present case is that a weighted sum of the concentrations of the absorbing species is equal to the initial Co(II) concentration.9 More generally, MRA can be used to find the stoichiometric con-

Wallace¹⁰ and Ainsworth¹¹ elaborated a method to detect the presence of stoichiometric constraints. If one row is subtracted from all of the other rows of the absorbance matrix, then a stoichiometric constraint is manifested as a decrease of rank. If no decrease of rank is found, then there is no stoichiometric constraint among the absorbing species in the system.

We carried out this analysis on our matrix. Table 1 summarizes the results. The first row of the table shows the absolute values of the largest residuals when (i-1) absorbing species are assumed. The second row contains the results of MRA after the subtraction. All of the 1680 rows were subtracted from the others, one by one, and the mean of the calculated 1680 largest absolute residuals is indicated in the table. These means really characterize their ranges, e.g., when $abs(P_5)$ and $abs(P_6)$ were calculated for the second row, the ranges were 0.008-0.0207 and 0.0016-0.0052 a.u. (absorbance units), respectively. It is clear from these nonoverlapping ranges that the subtraction did not decrease the rank of the experimental matrix, no matter which row was subtracted.

This calculation was repeated using experimental data measured in samples of 0.04 M initial concentration of Na₂-Co^{II}-EDTA (runs 1-8, 13 and 14). Comparison of the third and fourth rows of Table 1 reveals that the rank of the matrix decreased by one. The chemical meaning of the decrease of the rank at constant initial concentration is that all Co(II)containing species present in the system absorb light in the wavelength range selected, i.e.

$$[Na_2Co^{II}-EDTA]_{0i} = \sum_{k=1}^{NIAS} c_{ik}$$
 (8)

We note that $P_5 = 0.0035$ a.u. is slightly higher than the usual experimental error (0.002 a.u.). Experimental uncertainty in [Na₂Co^{II}-EDTA]₀ can easily cause this increase. Equation 8 must be incorporated into the LESs when M³ is used.

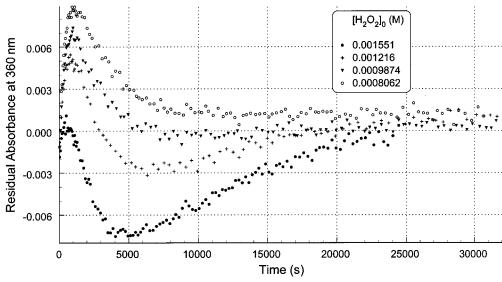


Figure 2. Residual absorbances as a function of time at different initial hydrogen peroxide concentrations in the Co^{II} -EDTA²⁻- H_2O_2 reaction. Runs 13, 1, 14, and 8 from Table 1 of part 1^1 were used.

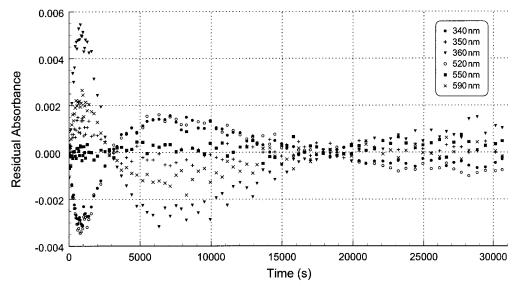


Figure 3. Residual absorbances at several wavelengths in the Co^{II}-EDTA²-H₂O₂ reaction. Run 1 from Table 1 of part 1¹ was used.

B. Calculations with M³. To determine the number of absorbing species and their possible concentration vs time curves, calculations were carried out with the linear equation system defined by eqs 4 and 8.

Four absorbing species were assumed initially. The lower part of Figure 1 shows the differences between the measured and calculated absorbances for all 10 080 experimental data. The figure shows many differences that are too large relative to the experimental photometric reproducibility.

It can be seen from Figure 1 that the large deviations are functions of time. Moreover, there are systematic relations between these deviation vs time curves and some chemical parameters (Figures 2 and 3). Figure 2 shows the residual vs time curves at 360 nm as a function of the initial hydrogen peroxide concentration. (Similar figures are obtained at other wavelengths.) Changing $[Na_2Co^{II}-EDTA]_0$ instead of $[H_2O_2]_0$ yields curves that represent similar relations.

Figure 3 illustrates another feature of the residuals. Residual vs time curves at different wavelengths but for the same samples show a systematic change in the shape of the curves.

These examples demonstrate both that four absorbing species are insufficient to describe the absorbance matrix of our example

and that the residuals can be monitored element by element in the ${\bf A}$ matrix, which provides useful information for determining the chemical model

In the calculations with NIAS = 4, we did not obtain unique C and E matrices. On the contrary, many C-E matrix pairs could be calculated from significantly different initial estimates of the E matrix. However, all of them gave the same calculated absorbance matrix. The experimental matrix carries enough information to calculate the smallest possible residuals, but it does not include enough information for a unique solution of the deconvolution. Mathematically, the Sn function defined by eq 2 has numerous minima, and the function values are the same at these minima. The products A_1 , A_2 , A_3 , and A_4 of four different C and E matrix pairs are shown in Figure 4. The elements of A_1 are taken as the independent variable, whereas the elements of the A_2 , A_3 , and A_4 matrices are the dependent variables. The plotted points have the same positions in their matrices. The figure clearly shows that all of the 30 240 plotted points lie on the y = x straight line, demonstrating that the products of different C and E matrix pairs are identical whenever the calculated absorbance matrices correspond to a minimum of Sn.

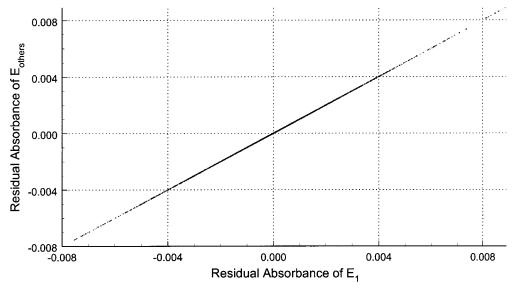


Figure 4. Calculated absorbance residuals of three different C and E matrix pairs as a function of residuals belonging to a fourth C and E matrix pair. Positions of the dependent and independent variables are the same in all matrices. NIAS = 4.

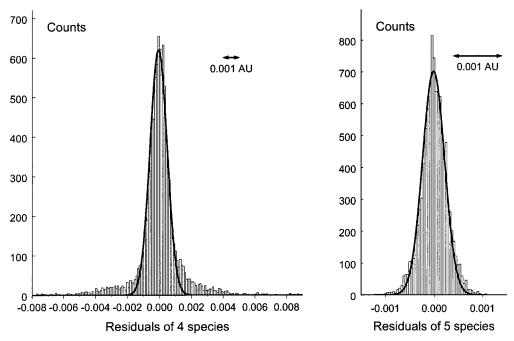


Figure 5. Histograms of the residuals for NIAS = 4 and 5. The thick line shows fitted density function for a Gaussian distribution. Relative widths of intervals, width of interval divided by the half width, are the same for both cases. Axes are normalized.

We carried out further calculations assuming five independent species. The upper part of Figure 1 shows the residuals for all individual data. No systematic deviation can be detected, and each residual is smaller than the photometric reproducibility. Therefore, five linearly independent absorbing species are necessary and sufficient to describe the primary absorbance matrix.

We also checked the distribution of the residuals for both four and five absorbing species. It is clear from Figure 5 that a normal distribution is appropriate if NIAS is 5. However, this distribution of residuals cannot be described by the theoretical Gaussian histogram if NIAS is 4. This result also supports the existence of five absorbing species.

The concentrations calculated by M3 (or any model-free method) are not unambiguous, as was pointed out explicitly by Lawton and Sylvestre¹² and Ohta. ¹³ In spite of their warnings, this crucially important limitation has often been forgotten. The shapes of the concentration vs. time curves, however, carry important information. For example, Figure 6 shows two different "optimal" time series. Both contain an important feature in common, a minimum in [CoII-EDTA²⁻]. These minima were found in every one of the 100 or so calculated C matrices, which suggests that the reaction mechanism must have step(s) in which Co^{II}-EDTA²⁻ is regenerated in the later stages of the reaction. Without M³, this feature does not emerge from the primary experimental data, and there is no indication of it in the literature. $^{12-15}$

V. Discussion and Conclusion

In this article, we have demonstrated that M³ is a powerful tool for determining the number of linearly independent absorbing species. Like MRA, described in part 1, M³ avoids sophisticated statistical considerations and is simpler and more reliable than other methods using eigenvalue calculations. The results can be interpreted easily, and the interdependence of the

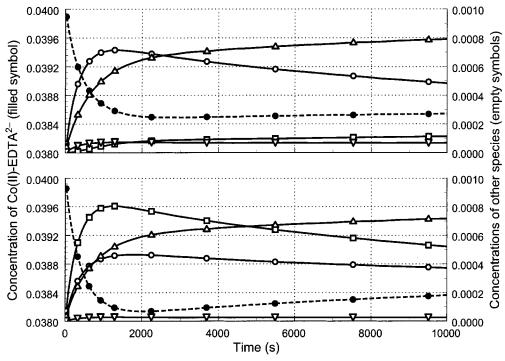


Figure 6. Two possible, and significantly different, series of calculated concentration vs time curves in the first run of the Co^{II} -EDTA²⁻- H_2O_2 reaction. Concentration of Co^{II} -EDTA²⁻ is on left axis, whereas the right axis shows other (unidentified) absorbing species.

measured data does not influence the conclusions. M³ has several advantages over MRA: it does not require transformation of the primary experimental data, it can incorporate the nonnegativity of the absorbance matrix, and it monitors the residuals element by element, whereas MRA considers the deviations only row by row and/or column by column. However, MRA requires much less computation time and computer memory than does M³, and the residuals derived from MRA can be evaluated more easily. MRA is necessary for filtering the matrix before using M³.

In addition to determining NIAS, M³ is able to resolve the absorbance matrix into matrices of concentrations and molar absorption coefficients. In the literature, many variants of factor analysis are used for this deconvolution. M3 offers several advantages over FA-based methods. Because M3 does not transform the primary matrix, the residuals are not distorted. They can be interpreted easily, and they can be monitored during the whole calculation process element by element. Stoichiometric constraints can readily be incorporated, as can any independent chemical information by fixing or omitting elements of the E matrix. Relations among the calculated molar absorption coefficients can also be studied by any reliable method of parameter estimation. If multiple solutions exist, M3 and parameter estimation together may provide the most important features of the system. Many series of concentration vs time curves can be calculated systematically. Their shapes are helpful in developing a chemical model of the system. In contrast, the variants of FA calculate only abstract matrices having no chemical meaning. To obtain real C and E matrices, a matrix transformation must be carried out.⁵ Such a procedure requires intuition, sophisticated knowledge of matrix algebra, experience, and a great deal of calculation. This transformation can easily result in the loss of important experimental information.

The use of M³ makes it possible to extract much essential experimental information from large matrices without assuming a chemical model. It is rarely possible, however, to obtain a unique solution without a chemical model. M³ can give much information about the system examined, but unless a total

characterization of the experimental data is feasible, neither M^3 nor FA can replace completely the usual evaluation procedures based on a postulated chemical model. Using M^3 or FA is, however, important for carrying out further calculations efficiently, and M^3 extracts information from large matrices more simply and more reliably than FA.

Acknowledgment. This work was supported by the Hungarian Academy of Sciences (Grant Nos. OTKA/TB029838 and OTKA/TB030446), by the Chemistry Division of the National Science Foundation, and by a U.S.—Hungarian cooperative grant.

Supporting Information Available: Experimental data for this example. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

- (1) Part 1. Peintler, G.; Nagypál, I.; Jancsó, A.; Epstein, I. R.; Kustin, K. J. Phys. Chem. A **1997**, 101, 8013.
- (2) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in Pascal*; Cambridge University Press: New York, 1992
- (3) Leggett, D. J. Computational Methods for the Calculation of Stability Constants; Plenum: New York, 1985.
- (4) Bittrich, H. J.; Haberland, D.; Just, G. Methoden chemisch-kinetischer Berechnungen; VEB Deutscher Verlag für Grundstoffindustrie: Leipzig, 1979.
- (5) Malinowski, E. R.; Howery, D. G. Factor Analysis in Chemistry; Wiley: New York, 1980.
 - (6) Nelder, J. A.; Mead, R. Comput. J. 1965, 7, 308.
- (7) Bard, Y. Nonlinear Parameter Estimation; Academic Press: New York, 1974.
 - (8) Peintler, G.; He, X.; Kustin, K.; Nagypál, I. To be published.
- (9) The weighting factor is the number of Co atoms in a species, if polynuclear species are formed as intermediate(s) or product(s).
 - (10) Wallace, R. M. J. Phys. Chem. **1960**, 64, 899.
 - (11) Ainsworth, S. J. Phys. Chem. 1961, 65, 1968.
 - (12) Lawton, W. H.; Sylvestre, E. A. Technometrics 1971, 13, 617.
 - (13) Ohta, N. Anal. Chem. 1973, 45, 552.
 - (14) Schwarzenbach, G. Helv. Chim. Acta 1949, 32, 839.
 - (15) Shimi, I. A. W.; Higginson, W. C. E. J. Chem. Soc. 1958, 260.
 - (16) Yalman, R. G.; Warga, M. B. J. Am. Chem. Soc. 1958, 80, 1011.
- (17) Yalman, R. G. J. Phys. Chem. 1961, 65, 556.