Extracting Experimental Information from Large Matrixes. 1. A New Algorithm for the Application of Matrix Rank Analysis

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

Institute of Physical Chemistry, Attila József University, H-6701 Szeged, P.O. Box 105, Hungary, and Department of Chemistry, Brandeis University, P.O. Box 9110, Waltham, Massachusetts 02254-9110

Received: January 8, 1997; In Final Form: July 16, 1997[⊗]

For many, especially complex, systems, modern spectroscopic measurements can be generated as large experimental data sets in matrix form. We report a new algorithm for the application of matrix rank analysis to extract significant experimental information from these large matrixes. The algorithm may be used to detect and remove erroneous rows and/or columns from the matrixes and to monitor the most significant experimental information along the rows and/or columns of the data sets. A new method for determining the number of absorbing species and a new concept for the treatment of experimental errors are presented. The algorithm is illustrated on real experimental examples.

Introduction

Matrix rank analysis (MRA) of spectroscopic data is a widely used method to determine the number of independent absorbing species (NIAS) either in chemically reacting or in equilibrium systems.[1−6] Its importance is increasing, because of the widespread use of solid-state photodetectors in modern data acquisition systems. The large matrix of data produced by such detection systems can be a disadvantage, however, compromising MRA and causing the method to yield ambiguous results. In this paper we examine MRA and propose a new method for its reliable and unambiguous implementation on large matrixes.

MRA can be applied to any experimental data set, provided that the Beer−Lambert law (or any similar linear relation) is valid:

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

where the A_{ij}'s are the elements of the absorption matrix (**A**), absorbances normalized for unit length; n is the number of absorbing species; p is the number of samples; and q is the number of wavelengths. The symbol c_{ik} stands for the concentration of the kth absorbing species, which has a molar absorption coefficient of ϵ_{kj} at the jth wavelength. The meaning of "large matrix" to characterize the system is that $p \gg n$ and/or $q \gg n$.

Wallace[1] and Ainsworth[2] pointed out that the rank of **A** gives the number of absorbing species. They also examined how the rank changes in closed systems due to stoichiometric constraints. Since then, three different algorithms have been developed for the determination of NIAS.

(1) The algorithm developed by Wallace and Katz[3] and by Katakis[4] is based on Gauss−Jordan elimination with full pivoting.[7] The result of the calculation is a vector **P**, the ith element of which is the largest value—in the sense of absolute values—of the residual of **A** after the $(i - 1)$th elimination step. The number of nonzero elements of the vector calculated this way gives the NIAS.

(2) The method developed by Hugus and El-Awady[5] is based on the eigenvalues of $\mathbf{AA} = \mathbf{A^T} \times \mathbf{A}$ (if $p \geq q$) or $\mathbf{AA} = \mathbf{A} \times \mathbf{A}^T$ (if $p \leq q$).[8] The determination of NIAS is therefore the same problem as solving the $\mathbf{AAx} = \lambda\mathbf{x}$ equation for all λ's and finding the nonzero eigenvalues.

(3) The third method is essentially a graphic, linearized representation of the first one, developed by Coleman et al.[6] This nomographical technique is not as accurate as digital computation, and the method is not further analyzed by us.

These three procedures are mathematically equivalent. Because of unavoidable experimental errors, however, the rank calculated is always larger than n. When MRA is applied, the real problem is to decide which elements of **P** or which eigenvalues of **AA** are sufficiently small to discard them as data due to experimental errors. There are several statistical procedures developed to solve this problem:

One possibility is a calculation of standard errors for either elements of **P** or eigenvalues of **AA**. Both procedures require an initial estimation of the standard error of the measured data. The reproducibility of a measurement—which is 0.002−0.003 absorbance unit (AU) with a modern diode-array spectrometer—helps us to estimate it. The distribution of these errors is generally assumed to be Gaussian.

Wallace and Katz[3] calculated the propagation of errors[9] in parallel with the process of elimination, which handles random errors adequately. Katakis[4] took into account computational as well as experimental errors. Hugus and El-Awady[5] introduced a relation between the standard errors of eigenvalues and the original experimental errors. In each procedure, the appropriate NIAS can be estimated by comparing the values to their standard errors.

Hugus and El-Awady[5] used the χ^2 test.[7] They also used the differences between the calculated and measured absorbances (residuals). They counted as significant those values that were larger than 3 times the estimated error.

In connection with factor analysis, Malinowski and Howery[10] summarized statistical criteria found in the literature. Since the goal of factor analysis is very close to that of MRA, these criteria can also be applied for MRA.

Despite these efforts, applying statistical criteria is still the most uncertain part of MRA. The conclusion from any error treatment is highly dependent on the accuracy of the initial error estimation.[11] Different statistical criteria may lead to different

† Attila Jozsef University.
‡ Brandeis University.
[⊗] Abstract published in *Advance ACS Abstracts,* September 15, 1997.

S1089-5639(97)00136-9 CCC: $14.00

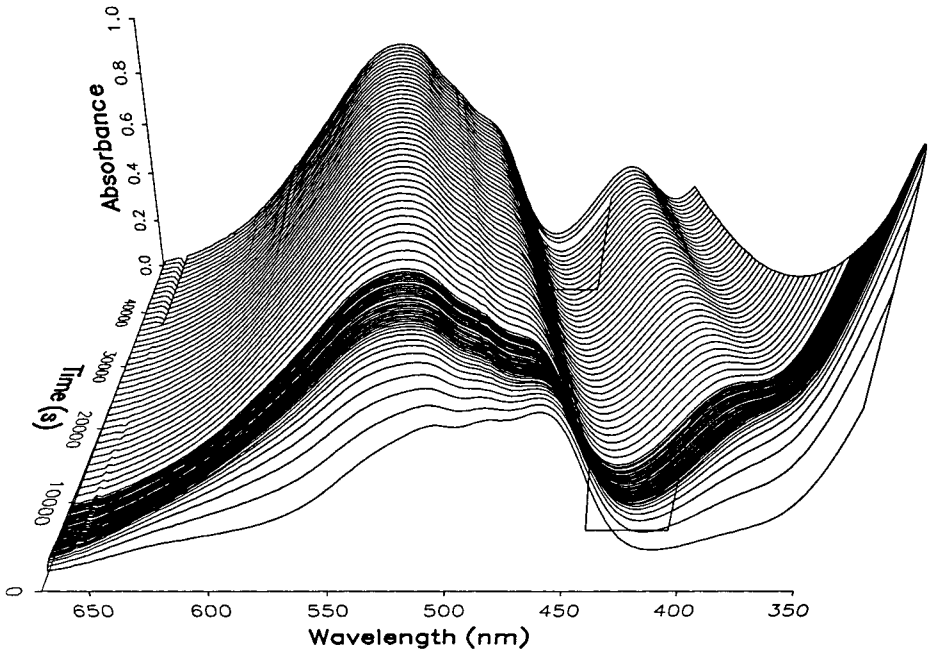

Figure 1. "Absorbance surface" of example 1 in the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction. The pH was constant at 7.5. The initial concentrations of $Co^{II}-EDTA^{2-}$ and H_2O_2 are 0.04 and 0.0012 M, respectively.

conclusions, even if the initial error estimation is the same,[10] therefore, NIAS may remain questionable.

Specific Problems of Measuring and Evaluating Large Data Sets. The advantage of using modern data acquisition systems is obvious: the collected data matrix includes much more chemical information than the data collected individually. There are, however, specific new problems that must be carefully considered during the process of evaluation.

The data collected may contain instrumental errors without any error message. The accuracy of the primary data may also change considerably in different wavelength ranges, sometimes at specific wavelength(s). The data measured at the wavelength of changing filter or light source are especially sensitive to instrumental errors. Therefore the primary experimental data must be filtered before further evaluation. In small data sets, filtering can be easily done by "eyes" and "hands". In the case of large matrixes it is much more difficult; a few erroneous data may easily lie hidden among many others.

The number of data entries used for further evaluation is usually less than the number of actually measured data values, because of computer memory limits, long computation times, overestimated range of measurements, and other factors. Reduction of the primary data matrix must be carried out without loss of any important information.

In using statistical procedures, it is assumed that the primary data are independent. This assumption is evidently valid if all the data are measured in independently prepared samples. The absorbances measured in the same sample at adjacent wavelengths, however, carry almost the same information, since the information gradually changes in small steps along the whole spectrum measured. The same is true for kinetics measurements along the time scale. In this sense the data are not independent. The interdependence may cause important information—involved in a specific range of the data—to disappear in the sink of errors when the rules for the propagation of errors are applied for large numbers of data entries.

The aim of our work is to develop a new algorithm for the application of MRA in order to solve the above problems. The new algorithm helps to check and select the elements of the data matrix and to determine the number of independent absorbing species. In addition, the new algorithm localizes that range of data which carries significant information in a large data matrix without any a priori assumption of a chemical model. The essential features of the algorithm are explained and illustrated in the following analysis of four real experimental data sets.

Experimental Section

The data sets of example 1 were collected from the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction by a Hewlett-Packard spectrophotometer, Model 8452A, in a quartz cuvette. The data set of example 2 was recorded from the $S_2O_3^{2-}-{}^{\bullet}ClO_2$ reaction by a Hi-Tech SF-61 stopped-flow instrument with built-in cuvette. The temperature and the ionic strength were constant in each solution.

The chemical information gained from these measurements will be published separately.[12,13] Therefore, the detailed experimental conditions are not given. Only two experimental facts are important for our present purposes:

(1) The photometric reproducibility—measured in separate samples—is 0.002 AU for example 1 and 0.008 AU for example 2. We used these values as the estimation of the standard deviation of the primary experimental data.

(2) It has been demonstrated experimentally[12] that *at least* four absorbing species exist in the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction.

The following sets of kinetics data are employed to illustrate use of the new algorithm we have developed.

Example 1. The first data set is a series of primary measured spectra without any selection or filtering. They were recorded between 340 and 666 nm at every second nanometer. For this example 39 spectra are between 180 and 7020 s ($\Delta t = 180$ s), and 41 additional spectra are in the range 7320−48 120 s ($\Delta t = 1020$ s). Figure 1 shows the "absorbance surface". The data matrix consists of 80 rows and 164 columns. The second data set is almost identical with the first one, but the rows and columns including erroneous data have been removed from the originally measured matrix (see below). The data matrix has 78 rows and 163 columns.

TABLE 1: Initial Concentrations for the Third Data Set in the $Co^{II}-EDTA^{2-}-H_2O_2$ Reaction

no. of series	$[Na_2Co^{II}-EDTA]_0$ (M)	$[H_2O_2]_0$ (M)	pH
1.7	0.04	0.0012	7.5, 7.1, 6.9, 6.7, 6.5, 6.2, 5.8
8.12	0.04, 0.035, 0.03, 0.025, 0.02	0.0008	7.5
13, 14	0.04	0.0016, 0.001	7.5

The third matrix contains filtered and selected data of 14 different runs together. The initial concentrations of the reagents are given in Table 1. The reaction was followed at 340, 350, 360, 520, 550, and 590 nm. Each run consists of 120 time points with six absorbances. The data matrix thus has 1680 rows and 6 columns. The selected points were chosen proportionally along the arc of the absorbance vs time curves. The originally measured data matrixes included 600−800 points in a run.

Example 2. The elements of this data matrix were measured in the $S_2O_3^{2-}-{}^{\bullet}ClO_2$ reaction at five different wavelengths. The reaction could only be followed at one wavelength at a time by the stopped-flow instrument used, so the different curves were recorded in physically different solutions. The absorbance values were measured 512 times between 0 and 0.4977 s (Δt = 0.000 974 s). The experimental data before 0.01 s were removed from the matrix, to avoid the influence of the dead and mixing times. These effects could only be detected during the first 4 ms of the measurement.

Calculations

All calculations were performed on IBM PC (DX-486) compatible computers. The precision of real numbers was always 19 or 20 significant figures. The programs were developed in Borland Pascal. Standard versions of the mathematical algorithms[7] were used wherever possible. Each program was tested through several artificial problems and real chemical examples from the literature.[3,5,10]

Essentially, two procedures are used for MRA in our algorithm.

The first is described by Wallace and Katz,[3] based on Gauss−Jordan elimination. The first elimination step consists of interchanging rows and columns and eliminating the nondiagonal elements of the first row and column. This transformation can be described by the operation[3,7]

$$A_{ij}' = A_{ij} - \frac{A_{ic}A_{rj}}{A_{rc}} \qquad (2)$$

performed on all elements of the **A** matrix, except A_{rc}, where A_{rc} is the element whose absolute value is the largest in **A**. Removing the *r*th row and *c*th column of **A** after applying eq 2, the rank of the remainder **A′** matrix is less by one than that of the original **A**. Now, operation 2 (eq 2) can be performed again on **A′**, and this procedure can be continued until all elements of the remainder matrix have vanished. The program collects the diagonal element of each elimination step (A_{rc}'s) into **P**. The program also retains the positions (*r* and *c* values) of these elements in the original matrix. These elements include the most important experimental information, so their significance is paramount in any further procedure. The program also calculates the standard deviations of the elements of **P** according to Wallace and Katz.[3] In the following, we refer to this procedure as the WK method.

The other method is the eigenvalue calculation detailed by Hugus and El-Awady.[5] It was performed with the help of the QL algorithm.[7] Besides the eigenvalues and their standard deviations, the eigenvectors and the χ^2 values were also calculated. The program we used also counts the number of calculated absorbances within the ranges $[A_{ij}-\sigma_{ij},A_{ij}+\sigma_{ij}]$, $[A_{ij}-2\sigma_{ij},A_{ij}+2\sigma_{ij}]$, and $[A_{ij}-3\sigma_{ij},A_{ij}+3\sigma_{ij}]$. We refer to this procedure as the HA method.

New Algorithm for Applications of MRA

Filtering Off Erroneous Rows/Columns. Filtering is based on the following feature of large matrixes: deleting one column or row cannot cause significant change in the elements of the **P** vector and in their standard deviations. If, however, a given row and/or column contains erroneous data, its omission causes significant change. Therefore—after the use of MRA on the primary data—we check if the omission of the rows and columns *belonging to the calculated elements* has any significant effect on the result. If no significant effect is found, then the primary matrix does not contain unrealistic row(s) and/or column(s). Otherwise the procedure should be continued by deleting the erroneous row(s)/column(s) until further omission does not have a noticeable effect on the results.

Filtering is illustrated through the first set of example 1. Figure 1 shows the spectra recorded in their original form. For the sake of illustration we present such a view of the surface that three defects can be noticed by eye. The absorbance drops to zero in two spectra between 396 and 434 nm. Malfunction of the data acquisition system may easily cause such effects. There is a "bump" on the surface at 654 nm. The probable source of this artifact is that the output of the deuterium lamp has a sharp peak around this wavelength.

The logic and results of the filtering outlined above can be followed in Table 2. The first run of MRA by the WK method shows the **P** vector of the original matrix. By comparing P_i values with their standard deviations, seven nonzero elements, P_1-P_7, can be identified. Their positions in the original matrix are given in the third row of run 1 in the table.

Runs 2−8 show the results of MRA when the rows are temporarily removed one by one. Omitting the 37th or 77th spectrum significantly changes the **P** vector, while the other deletions leave the **P** vector virtually unmodified. Removing both the 37th and 77th spectra decreases the rank of the data matrix by 2 as can be seen in run 9. This procedure proves that these two spectra contain erroneous data, so they should be deleted permanently from the original matrix. The procedure was repeated with the modified matrix. Runs 10−14 show the results. Since the **P** vectors from these runs are very similar to that of run 9, no more rows need be deleted. The process was continued by deleting the appropriate columns. Runs 15−19 show the results. While removing columns 1, 91, 62, or 26 does not change the rank of **P**; erasing column 158 decreases the rank by 1.

Repeating the calculation with the new matrix—from which rows 37 and 77 as well as column 158 were permanently omitted—clearly shows that no more columns need be deleted. The **P** vectors of runs 20−23 do not differ significantly from that of run 18. The procedure may be repeated for the rows again if necessary. We did so for our example, but no further significant change in **P** was detected.

The HA method is not suitable for this procedure, because the eigenvalues cannot be assigned unequivocally to the rows and columns of the original matrix. Following the reviewer's proposal, we carried out the same calculations using the HA method. The required computing time was longer by 2 orders of magnitude, and it was impossible to find the erroneous rows/columns unambiguously. We plan to compare the methods in this respect in a separate study.

TABLE 2: Results of MRA Applied to Example 1[a]

no. of runs	deleted rows	deleted columns		P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
1			LAV	1.2742	0.7122	−0.3905	0.1911	0.0366	0.0300	-0.0206	-0.0042	0.0035
			σ	±0.0020	±0.0024	±0.0029	±0.0031	±0.0029	±0.0052	±0.0041	±0.0032	±0.0042
			r,*c*	6,1	80, 91	37, 29	1, 62	61, 158	77, 42	25, 26	68, 159	15, 8
2	6		LAV	1.2734	0.7063	−0.3913	0.1936	0.0367	0.0301	−0.0193	−0.0043	−0.0042
3	80		LAV	1.2742	0.7103	−0.3907	0.1911	0.0340	0.0298	−0.0206	−0.0041	0.0035
4	37		LAV	1.2742	0.7122	−0.3708	0.1879	0.0367	−0.0215	−0.0042	0.0035	0.0033
5	1		LAV	1.2742	0.7122	−0.3905	0.1151	0.0365	0.0301	−0.0200	−0.0042	0.0033
6	61		LAV	1.2742	0.7122	−0.3905	0.1911	0.0359	0.0301	−0.0206	−0.0043	0.0035
7	77		LAV	1.2742	0.7122	−0.3905	0.1911	0.0366	−0.0205	−0.0042	0.0035	0.0034
8	25		LAV	1.2742	0.7122	−0.3905	0.1911	0.0366	0.0300	−0.0203	−0.0042	0.0042
9	37, 77		LAV	1.2742	0.7122	0.1879	0.0367	−0.0215	−0.0042	0.0035	0.0033	−0.0012
			σ	±0.0020	±0.0024	±0.0030	±0.0029	±0.0036	±0.0032	±0.0041	±0.0048	±0.0053
			r, *c*	6, 1	80, 91	1, 62	61, 158	25, 26	68, 159	15, 8	35, 160	75, 96

[a] P_i denotes the *i*th element of **P**. LAV (largest absolute value) means this element itself, σ denotes its standard deviation, *r* and *c* are the row and column positions of P_i in the original matrix. The second column of the table shows which rows and columns were removed from the primary matrix before using the WK method. Boldface entries indicate a significant change of **P**.

Reduction of the Size of the Matrix. The aim is to reduce the size but to keep those rows and columns that carry the most important information. MRA is an excellent tool for doing this if we use the WK method.

Let us examine the 18th run of Table 2 in detail. The largest deviation from zero is P_1 if we try to describe the matrix without any absorbing species, i.e., $n = 0$ in (1). If one absorbing species is assumed, P_2 gives the largest deviation. Similarly, P_3 is the largest unexplained datum in the remainder matrix of the $\mathbf{A}'$ matrix, which describes the experiments by a linear combination of two absorbing species, etc. Generally, if P_i is not zero, then the matrix cannot be described by (i - 1) absorbing species, and its row and/or column contains the most important information on the *i*th species.

The positions of the elements of **P** are known in the original **A** matrix (third row of run 18 in Table 2). The rows and/or columns in these positions carry the most important experimental information about the species. When the size of matrix **A** has to be reduced, the rows and columns belonging to the elements of **P** must not be removed. For example, if the second data set of example 1 is too large, the 1st, 91st, 62nd, and 26th columns (the absorbance vs time curves at 340, 520, 462, and 390 nm, respectively) as well as the 6th, 80th, 1st, and 25th rows (the spectra at 1080, 48120, 180, and 4500 s, respectively) must not be deleted.

If, for a moment, we delete these rows and columns and use the same procedure, the second most important rows and columns are found, and so on. In this way, an optimal size can be obtained, containing the most important information for further evaluation.

Comparison of the WK and HA Methods for MRA. We have analyzed the examples with both methods. Table 3 summarizes the results. It can be seen from the table that different conclusions can be drawn from different data for the same reaction and from different methods.

Neither the WK nor the HA method gives unambiguous results for NIAS in every example. In general, a 95% confidence interval is the most frequently accepted criterion, but the large number of data in our examples would require the use of a 99.7% interval. However, comparing P_i or λ_i values to their standard deviations led to different NIAS with different confidence intervals. This fact alone suggests that a simple MRA is not always suitable for the determination of NIAS.

Another apparent contradiction is that different data sets for the same reaction suggest different NIAS. Namely, different data sets for the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction give slightly different NIAS, as shown in Table 3. This apparent inconsistency can easily be explained by taking into account that the third set contains 14 independent series of measurements, while the second set contains only one.

The most serious problem is that the different methods of MRA lead to different NIAS, clearly shown by the third data set in Table 3. The result of the WK method is consistent with experiment, since the presence of at least four absorbing species has been shown independently.[12] However, the HA method failed to find even the possible range of NIAS.

To check the HA method in detail, we also calculated some additional statistical criteria suggested by Hugus and El-Awady.[5] Since the qualitative consequences were the same, only the check of the widely used χ^2 test is detailed here. It uses the data matrix and the initial error estimation of its elements as input data. A guess-value for NIAS must also be given. The output is a probability between 0 and 1, which gives the likelihood that the assumed NIAS is equal to or larger than the correct value. Decreasing the initial error estimation for examples 1 and 4 by 0.000 16 and 0.0006 AU, respectively, increases the calculated NIAS by 1. It is evident that the error estimation cannot be controlled within the requirements of successful use of the χ^2 test. Therefore, this statistical criterion is practically useless in the case of large experimental data sets.

Different interpretations of the degrees of freedom were found in the literature[5,10] regarding the χ^2 test. We used the more restrictive Malinowski definition.[10] The other definition was also employed, but the qualitative result remained the same.

Assuming the same error for all experimental data may be too crude an approximation. Therefore, we also used other methods for creating estimated errors. We tried absolute and relative estimations and combinations of these. The qualitative results were the same in each case.

In conclusion, the simple MRA alone is not a foolproof method for determining the number of absorbing species from large data matrixes. The use of the HA method is especially dangerous for three reasons:

(1) The statistical procedures involved in the HA method presuppose the independence of the experimental data. In any practical work with data acquisition systems, data next to each other are highly dependent.

(2) The method requires much more calculation than the WK method. Therefore, the computational error propagation may distort the experimental data and their real errors.

(3) The HA method carries out matrix multiplication for creating a smaller matrix from the original one. It may easily happen that some of the information appears only in a small

TABLE 3: Determination of NIAS in Data Matrices Using Different Methods for MRA[a]

data set		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	method	NIAS
second one of example 1	P_i	1.2742 ± 0.0020	0.7122 ± 0.0024	0.1879 ± 0.0030	0.0216 ± 0.0035	0.0035 ± 0.0040	-0.0036 ± 0.0034	WK	4
	λ_i	3554 ± 0.01941	43.91 ± 0.01948	1.568 ± 0.01660	0.05738 ± 0.01913	0.003012 ± 0.01897	0.0003282 ± 0.01964	HA	3, 4(?)
third one of example 1	P_i	1.2890 ± 0.0020	0.8732 ± 0.0023	-0.1223 ± 0.0032	0.0197 ± 0.0039	-0.0101 ± 0.0037	0.0021 ± 0.0039	WK	4, 5(?)
	λ_i	2170 ± 0.07447	1095 ± 0.06909	2.267 ± 0.07346	0.02850 ± 0.4913	0.004827 ± 0.05567	0.003458 ± 0.04960	HA	3
example 2	P_i	0.6302 ± 0.0080	-0.0248 ± 0.0107	-0.0128 ± 0.0155	0.0042 ± 0.0116	0.0029 ± 0.0121		WK	1, 2(?)
	λ_i	38.14 ± 0.0426	0.02425 ± 0.0431	0.005948 ± 0.0483	0.004563 ± 0.03028	0.0003006 ± 0.01568		HA	1

[a] λ_i denotes the ith largest eigenvalues of AA. If NIAS is uncertain, a question mark indicates the less probable value. The other abbreviations are defined in the text.

part of the primary matrix. The multiplication hides the information content of that part.

Calculation and Interpretation of Residual Absorbance Curves. We have seen that the ith element of vector **P** has the largest absolute value of the remainder **A** after the $(i - 1)$th elimination step and that a serial number of a row and column can be assigned to it. This fact opens a new approach for the application of MRA.

Let us denote the assumed NIAS by m and calculate P_{m+1}. The serial number of this element gives the row (or column) that contains the most important information on the possible existence of the $(m + 1)$th species. If this row (or column) is omitted and MRA is carried out again, the second most important row (or column) is found, and so on. This gradual omission, one by one, can be continued until only m rows (or columns) remain. Then the calculated P_{m+1} elements are plotted as a function of their serial numbers. We may also plot the P_{m+1} values as a function of an independent variable assigned to the serial numbers.

This method is illustrated in example 2, where it is assumed that NIAS = 1. The calculated elements of **P** are in the first row of Table 4. The time value and the position of P_2 are also indicated in the table. The row belonging to P_2 is removed from the original matrix, and MRA is carried out again with the remainder. Now, the new P_2 contains the most information about the second species, and its row in the original matrix is also known.

This process was continued until two rows remained in the continuously reduced data matrix. The last two steps are also indicated in Table 4. The result is called the residual absorbance curve (RAC), indicating that it represents that part of the whole time scale which cannot be described by assuming one absorbing species.

Figure 2 contains the residual absorbance curve of example 2. The points clearly show a systematic deviation. This deviation proves the existence of the second absorbing species. Furthermore, one may definitely conclude that this species is an intermediate. The course of this RAC also helps to estimate the rate of the chemical processes in which this intermediate is produced and transformed into a nonabsorbing species.

It should be mentioned that the formation and decomposition of the absorbing intermediate cannot be observed visually from the original absorbance vs time points.

Further residual absorbance curves can also be created by increasing the assumed NIAS. We also carried out the necessary calculations when NIAS was 2. The points of the residual curve in this case were randomly distributed around zero.

These observations show that two absorbing species (one reagent and one intermediate) are necessary and sufficient for describing the primary data matrix within experimental error.

The procedure outlined can also be carried out in a modified way: the columns are removed instead of the rows. In this case, the unexplained residual absorbance is a function of the wavelength. We call this function the wavelength-dependent residual curve (WRC), while the previous function is called the time-dependent residual curve (TRC).

We also calculated WRCs and TRCs in the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction. It was emphasized above that different methods lead to different values of NIAS in this chemical system. The possible values were 3, 4, and 5 if a simple MRA was used (see Table 3).

At first, four WRCs were produced from the second data set of example 1. The assumed values for NIAS were 3, 4, 5, and 6. Figure 3 shows the results. When the assumed value of NIAS is 3, the calculated WRC clearly shows that three

TABLE 4: Calculation of the Residual Absorbance Curve by Assuming NIAS = 1 in Example 2

removed rows	P_1	P_1	position of P_1 (r, c)	time at P_2 (s)	P_3	P_4	P_5
	0.6302	−0.0248	76, 3	0.08379	−0.0128	0.0042	0.0029
76	0.6302	−0.0229	80, 3	0.08769	−0.0114	0.0044	−0.0029
76, 80	0.6302	−0.0228	89, 3	0.09645	−0.0109	0.0044	−0.0035
76, 80, 89	0.6302	−0.0225	108, 3	0.11496	−0.0114	0.0049	−0.0031
all but 1, 2, 278, 450	0.6302	0.0018	450, 3	0.44806	−0.0012	−0.0099	0.0003
all but 1, 2, 278	0.6302	−0.0016	278, 4	0.28053	−0.0010	−0.0099	0.0003

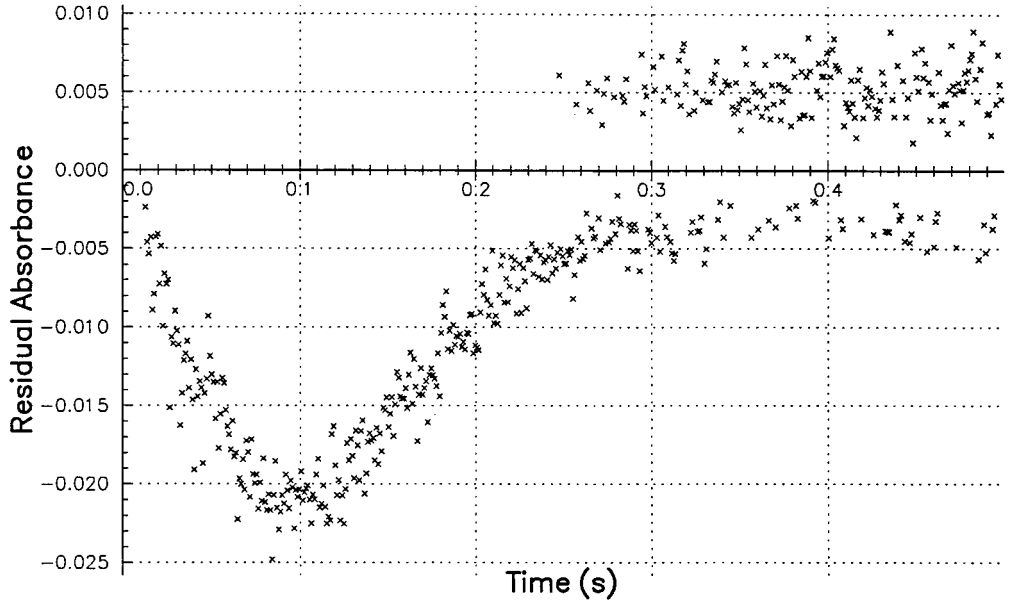

Figure 2. Time-dependent residual absorbance curve in the $S_2O_3^{2-}-{}^{\bullet}ClO_2$ reaction. One absorbing species was assumed for the data matrix of example 2.

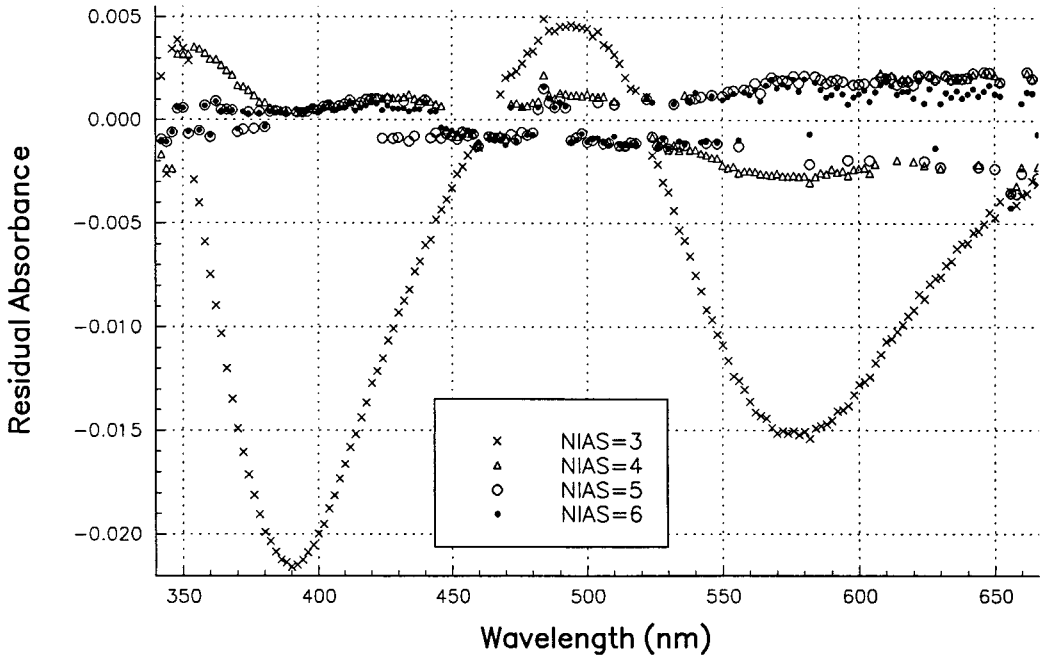

Figure 3. Four wavelength-dependent residual absorbance curves in the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction. The second data matrix of example 1 was used in the calculations.

absorbing species are not sufficient for describing the experiments. The unexplained residual values are large and the variation of the function is not random.

It is also clear that five absorbing species are sufficient to describe the data matrix within experimental error. There are no significant differences between WRCs calculated with NIAS = 5 or 6 values. Furthermore, the distribution of the points is random within the experimental deviations. What is uncertain from this figure is the existence of the fifth absorbing species. There are some portions of WRCs where a nonrandom deviation can be seen (340−390 nm and 530−610 nm). However, these differences are small.

To decide between the existence of four or five absorbing species, we calculated TRCs for the third set. This data set contains much more experimental information about the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction. The results are illustrated in Figure

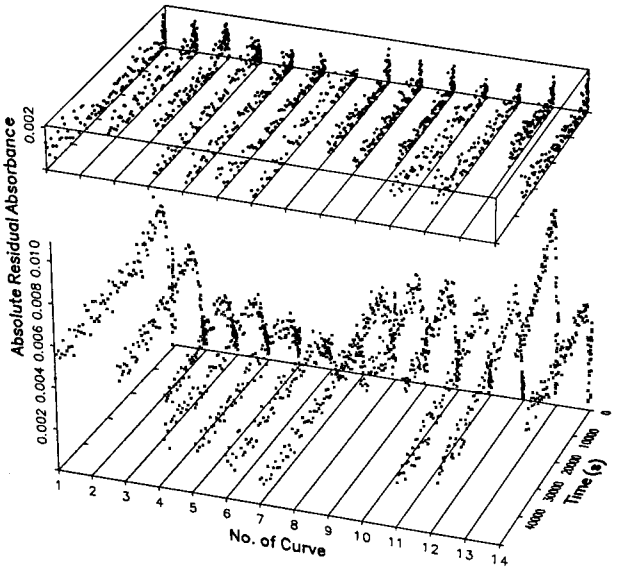

Figure 4. Time-dependent residual absorbance curves in the $Co^{II}-EDTA^{2-}-H_2O_2$ reaction. The third data matrix of example 1 was used in the calculations. The lower series of curves was created by assuming four absorbing species; five absorbing species were used in the upper series of curves. The box around the upper series delineates the largest residual.

4. The absolute RAC is drawn for the sake of clarity. The lower series of curves clearly demonstrates that four absorbing species are not sufficient for describing the data matrix. There are large residual values, and their behavior is not random. One may also find correlations between the height of the peaks and initial conditions given in Table 1.

The upper series of curves in Figure 4 was calculated by taking NIAS = 5. The values of these curves are small; the largest is 0.002 08 absorbance unit. This value is roughly equal to the estimated error. The points are randomly distributed. One might be inclined to conclude that these curves have peaks around $t = 1000$ s, but 22% of the experimental data are found in the range 0−2000 s, which is only 4% of the time range. The higher density of experimental points at the beginning of the curves causes the illusion of unexplained peaks.

Based on the residual absorbance curves, the existence of five linearly independent absorbing species is the most reasonable conclusion. The example also illustrates two important facts.

(1) The residual absorbance curves can also be calculated if more experimental runs are involved in one data matrix.

(2) Without sufficient experimental information, any application of MRA fails to give acceptable results. However, inappropriate evaluation techniques can also obscure the extractable information.

The wavelength-dependent residual absorbance curve can also be useful if the size of the original matrix must be reduced. Larger values in a WRC indicate the wavelengths where the measurements carry more information. For example, it can be seen in Figure 3 that the measured absorbances at 390, 490, and 575 nm carry the most important information about the fourth absorbing species. In addition, the values at 350 nm give us information about the fifth absorbing species.

Another technique in the literature[15−17] is somewhat similar to the calculation of residual absorbance curves. Gampp et al. call their method evolving factor analysis (EFA). Through EFA, the rows or columns are also removed systematically and the eigenvalues of **AA** are calculated. However, the order of removal is determined before any calculation and is not an inherent part of the calculation process. For example, the actual first (or last) row (or column) is always deleted before a new eigenvalue calculation. The authors had to choose this approach because the determination of the eigenvalues does not give the position of the element carrying the most important information. We have applied EFA to all of our examples. We have found that much more, and localized, information can be gained from RAC than from EFA.

Discussion

Our calculations on and analysis of real examples clearly show that the use of MRA requires special care and a special algorithm for data given by modern data acquisition systems. The advantages of the procedures presented are as follows:

Erroneous experimental data may be readily filtered out.

The size of the data matrix can be adjusted for the requirements of a further evaluation procedure in a way that preserves the most important information. The extent of the information content can be monitored along the rows and/or columns.

The number of independent absorbing species may be determined unambiguously. We have shown that the HA method for MRA may give misleading information in the case of large matrixes, because its statistical criteria are not valid for the data produced by modern data acquisition systems.

Calculation of the residual absorbance curves is an especially valuable tool to decide on the number of independent absorbing species, to detect the presence of intermediates, and even to localize their appearance in time or along the wavelength scale.

A systematic comparison of the different and mathematically equivalent methods for MRA shows that only the procedure suggested by Wallace and Katz[3] is appropriate for MRA in the case of large matrixes.

Conclusion

On the basis of the results presented, the following algorithm for the application of MRA is suggested.

Filter out the rows and columns containing instrumental errors.

Calculate the residual absorbance curves along the rows and columns assuming increasing numbers of independent absorbing species until the residuals are random and are in accord with the accuracy of the instrument.

Reduce the size of the matrix according to the requirements of the further evaluation procedure.

Acknowledgment. The authors thank Attila Horváth for providing the data matrix of example 2. This work was supported by the Hungarian Academy of Sciences (Grant No. OTKA/F-020863), by the National Science Foundation (Grant No. CHE-9615834), and by a U.S.-Hungarian cooperative grant.

Supporting Information Available: The programs used for MRA are available through anonymous FTP or by WWW at ftp://ftp.jate.u-szeged.hu/pub/chem/mra/mraXXX.exe, where XXX is the version number. The data sets of the examples can also be found at the same location in the file mrapaper.exe. Both files are in zipped and self-extracting form.

References and Notes

(1) Wallace, R. M. *J. Phys. Chem.* **1960**, *64*, 899.
(2) Ainsworth, S. *J. Phys. Chem.* **1961**, *65*, 1968.
(3) Wallace, R. M.; Katz, S. M. *J. Phys. Chem.* **1964**, *68*, 3890.
(4) Katakis, D. *Anal. Chem.* **1965**, *37*, 876.
(5) Hugus, Z. Z.; El-Awady, A. A. *J. Phys. Chem.* **1971**, *75*, 2954.
(6) Coleman, J. S.; Varga, L. P.; Mastin, S. H. *Inorg. Chem.* **1970**, *9*, 1015.

(7) Press: W. H; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in Pascal*; Cambridge University Press: New York, 1992.

(8) Margenau, H.; Murphy, G. M. *The Mathematics of Physics and Chemistry*; Ronald Press: New York, 1964.

(9) Norris, A. C. *Computational Chemistry*; Wiley & Sons Ltd.: New York, 1981.

(10) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley & Sons Ltd.: New York, 1980.

(11) Duewer, D. L.; Kowalski, B. R.; Fasching, J. L. *Anal. Chem.* **1976**, *48*, 2002.

(12) Peintler, G.; He, X.; Kustin, K.; Nagypál, I. To be published.

(13) Horváth, A.; Nagypál, I. To be published.

(14) Schöneshöfer, M. *Int. J. Radiat. Phys. Chem.* **1973**, *5*, 375.

(15) Gampp, H.; Maeder, M.; Meyer, C. J.; Zuberbühler, A. D. *Talanta* **1985**, *32*, 1133.

(16) Gampp, H.; Maeder, M.; Meyer, C. J.; Zuberbühler, A. D. *Talanta* **1986**, *33*, 943.

(17) Gampp, H.; Maeder, M.; Meyer, C. J.; Zuberbühler, A. D. *Comm. Inorg. Chem.* **1987**, *6*, 41.