Analysis of Correlated Spectral Data

Ieda Scarminio† and Mikael Kubista*,‡

Departamento de Quimica, Universidade Estadual de Londrina, 86 051 Londrina PR, Brazil, and Department of Physical Chemistry, Chalmers University of Technology, S-412 96 Gothenburg, Sweden

We describe a program for analyzing correlated spectral data by Procrustes rotation, which eliminates the need for reference samples (Kubista, M. *Chem. Intel. Lab. Syst.* 1990 *7*, 273). The experimental spectra are the only "input" required by the DATa ANalysis (DATAN) program, which calculates the number of components, their spectral profiles, their concentrations, and the ratio of their responses to two spectroscopic measurements. The DATAN program first calculates common score and loading vectors to the two data sets by the NIPALS algorithm, and the number of spectroscopically independent components is determined by a χ^2 test. The score matrices for the two measurements are then related through Procrustes rotation, which gives the spectral profiles of the components, their concentrations in the samples, and the ratios between their responses to the two measurements. We test extensively the stability of the algorithm used by the DATAN program and we discuss its limitations.

INTRODUCTION

The problem of identifying the components in a sample is one of the oldest problems in chemistry, and because of its importance, it has attracted the attention of scientists for decades. When the components cannot be separated from each other, for example when in a chemical equilibrium that would be perturbed by the separation procedure or when the components are chemically too similar to become separated, the sample has to be analyzed as a whole and the components must be identified from the responses to measurements of the entire sample. Here various spectroscopic techniques are important, because they produce a spectral response from which the components may be identified through their characteristic profiles.

A major difficulty in spectroscopic studies occurs when the component spectra overlap, and no calibration data are available. The case with two components was first discussed by Lawton and Sylvestre,[1] who provided a way to limit the number of solutions to those where the calculated concentrations and spectral profiles contain only nonnegative elements. Their approach was later extended to more components,[2-13] though these treatments usually require rather advanced programming. These methods, however, are not always applicable. Some spectra, such as dichroism and difference spectra, may contain negative elements, and also noise may provide a serious problem. Noise can always be negative, and if significant, it may be difficult to remove sufficiently without distorting the physical information in the experimental data. Imposing a nonnegative criterion in the analysis may then give incorrect results. Finally, we reemphasize, that even when these methods are applicable, they do not provide a unique solution but merely limit the number of solutions to those with nonnegative elements.

Recently, we described how this classical problem of characterizing the components in unknown samples without a priori information can be solved by recording a second spectrum of each sample, that is suitably correlated to the first spectrum.[14] The required correlation is that the contributions from the components to the two spectra have the same spectral intensity distributions, but different magnitudes, and the ratio between the magnitudes is a characteristic feature of each component. The spectral information is then sufficient to determine the number of independent components in the samples, their spectral profiles, their concentrations, and the ratios between their responses to the two measurements.

In this paper we describe the DATa ANalysis (DATAN) program we have developed to calculate the number of components, their spectral profiles, their concentrations, and the ratios between their responses to the two measurements using only the experimental data as input. We perform extensive tests of the stability of the analysis, and we discuss how the experimental design should be optimized for different situations. We also investigate how linear dependence in the experimental data affects the analysis and how such an influence can be realized from the calculated results.

THEORY

When the spectral responses of the pure components are known, their concentrations in a sample are easily quantified by deconvolution of the recorded spectrum into the individual spectral profiles.

$$\mathbf{a}(\lambda) = \sum_{i=1}^{r} c_i \mathbf{v}_i(\lambda) \qquad (1)$$

where $\mathbf{a}(\lambda)$ is the spectrum of the sample recorded as a function

* To whom correspondence should be addressed.
† Universidade Estadual de Londrina.
‡ Chalmers University of Technology.

(1) Lawton, W.; Sylvestre, E. *Technometics* 1971, *13*, 617-633.
(2) Ohta, I. *Anal. Chem.* 1973, *45*, 553-557.
(3) Ho, C.-N.; Christian, G. D.; Davidson, E. R. *Anal. Chem.* 1978, *50*, 1108-1113.
(4) Ho, C.-N.; Christian, G. D.; Davidson, E. R. *Anal. Chem.* 1980, *52*, 1071-1079.
(5) Ho, C.-N.; Christian, G. D.; Davidson, E. R. *Anal. Chem.* 1981, *53*, 92-98.
(6) Meister, A. *Anal. Chim. Acta* 1984, *171*, 149-161.
(7) Sasaku, K.; Kawata, S.; Minami, S. *Appl. Opt.* 1984, *23*, 1955-1959.
(8) Borgen, O.; Kowalski, B. *Anal. Chim. Acta* 1985, *174*, 1-26.
(9) Delaney, N.; Mauro, D. *Anal. Chim. Acta* 1985, *172*, 192-205.
(10) Kawata, S.; Komeda, H.; Sasaki, K.; Minami, S. *Appl. Spectrosc.* 1985, *39*, 610-614.
(11) Vandeginste, B.; Essers, R.; Bosman, T.; Reijnen, J.; Kateman, G. *Anal. Chem.* 1985, *57*, 971-985.
(12) Borgen, O.; Davidsen, N.; Mingyang, Z.; Oyen, O. *Mikrochim. Acta* 1986, *2*, 63-73.
(13) Burdick, D.; Tu, X. *J. Chemomet.* 1989, *3*, 431-441.
(14) Kubista, M. *Chem. Intel. Lab. Syst.* 1990, *7*, 273-279.

0003-2700/93/0365-0409$04.00/0

of wavelength, digitized into m data points ($\lambda = l, m$), and c_i and $\mathbf{v}_i(\lambda)$ are the concentrations and spectral profiles of the r ($i = l, r$) components. If the $\mathbf{v}_i(\lambda)$'s are known, the c_i's can be determined by standard least squares methods. However, if the $\mathbf{v}_i(\lambda)$'s are unknown and no calibration spectra are available, there is no way to determine the c_i's.

A somewhat better situation is when one has a number of samples that all contain the same r components, but at different concentrations.

$$\mathbf{a}_j(\lambda) = \sum_{i=1}^{r} c_{ij}\mathbf{v}_i(\lambda) \quad \text{or} \quad \mathbf{A} = \mathbf{CV}' \tag{2}$$

$a_j(\lambda)$ is the spectrum of the jth sample ($j = l, n$), and c_{ij} is the concentration of component i in sample j. In matrix notation $\mathbf{A}$ is an $n \times m$ matrix, of which the columns correspond to the different measurement variables and the rows correspond to different samples. $\mathbf{C}$ is an $n \times r$ matrix containing the component concentrations as columns, and $\mathbf{V}'$ is an $r \times m$ matrix containing the component spectra as rows. The information in $\mathbf{A}$ is, however, not sufficient to obtain a unique solution for $\mathbf{C}$ and $\mathbf{V}'$, not even when r is known.

Matrix $\mathbf{A}$ can be factorized into a product of two matrices:

$$\mathbf{A} = \mathbf{TP}' \tag{3}$$

which have the same dimensions as $\mathbf{C}$ and $\mathbf{V}'$, respectively. The factorization is, however, not unique, and the matrices $\mathbf{P}$ and $\mathbf{T}'$ must not be identical to the matrices $\mathbf{C}$ and $\mathbf{V}'$. This is analogous to the fact that a scalar, a, cannot be factorized into two unique scalars, c and v, without placing additional constrains on the nature of c and v.

Such a constraint can be a second spectrum recorded on each sample, that is correlated to the first, such that the spectral responses of the components have the same profiles, but different magnitudes. The equations describing the two experiments are

$$\mathbf{A} = \mathbf{CV}' \tag{4}$$

$$\mathbf{B} = \mathbf{CDV}' \tag{5}$$

where $\mathbf{A}$ is a matrix containing the recorded spectra of the first kind, $\mathbf{a}_j(\lambda)$, as rows, $\mathbf{B}$ is a matrix containing the recorded spectra of the second kind, $\mathbf{b}_j(\lambda)$, as rows, $\mathbf{C}$ and $\mathbf{V}$ are as defined above, and $\mathbf{D}$ is an $r \times r$ diagonal matrix with the elements $d_{ii} = d_i$. These elements, as we shall see, must all be different.

The number of components and the matrices $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{V}'$ are calculated from the data matrices $\mathbf{A}$ and $\mathbf{B}$ by the DATAN program in two steps: NIPALS and ROTATION.

NIPALS. The NIPALS part calculates common loading ($\mathbf{P}$) and score matrices ($\mathbf{T}_a$ and $\mathbf{T}_b$) for the input data matrices $\mathbf{A}$ and $\mathbf{B}$ by sequentially calculating the most significant pairs of loading and score vectors. The matrices $\mathbf{A}$ and $\mathbf{B}$ are laminated to form a $2n \times m$ ($\mathbf{A}/\mathbf{B}$) matrix. The NIPALS algorithm[15,16] is as follows:

Step 1: Choose the column in the matrix ($\mathbf{A}/\mathbf{B}$) with the largest variance as a starting value for $\mathbf{t}$ (the catenated vector $\mathbf{t}_a|\mathbf{t}_b$).

Step 2: Calculate the corresponding loading vector as

$$\mathbf{p}' = \frac{\mathbf{t}'\left(\frac{\mathbf{A}}{\mathbf{B}}\right)}{\mathbf{t}'\mathbf{t}}$$

Step 3: Normalize $\mathbf{p}$ to unit length by multiplying with c:

$$c = \frac{1}{\sqrt{\mathbf{p}'\mathbf{p}}}$$

Step 4: Calculate a new value for the score vector as

$$\mathbf{t} = \frac{\left(\frac{\mathbf{A}}{\mathbf{B}}\right)\mathbf{p}}{\mathbf{p}'\mathbf{p}}$$

Step 5: Check for convergence. If convergence has been achieved go on with step 6; otherwise repeat from step 2.

Step 6: Form the residual matrix

$$\mathbf{E} = \left(\frac{\mathbf{A}}{\mathbf{B}}\right) - \mathbf{TP}'$$

Use $\mathbf{E}$ as a new ($\mathbf{A}/\mathbf{B}$) matrix and calculate the next pair of score and loading vectors by repeating the procedure. Continue until the residual matrix $\mathbf{E}$ contains only noise.

χ^2 Test. The number of independent spectral components in the samples is determined by a χ^2 test. The algorithm for the χ^2 test is as follows:[17]

Step 1: Input the average noise σ_a and σ_b in the experimental data. If not known, set σ_a to 1 and σ_b to the estimated noise level in data set B relative to data set A. The χ^2 test will still predict the correct number of independent components, though the reduced χ^2 value will be arbitrary.

Step 2: Set l to 1.

Step 3: Calculate χ^2 for the matrices $\mathbf{A}$ and $\mathbf{B}$ using the l most significant score vectors in $\mathbf{T}$ and the l most significant loading vectors in $\mathbf{P}$.

$$\chi_a^2 = \sigma_a^{-2}\sum_{j}^{m}\sum_{i}^{n}\left(\mathbf{A}_{ij} - \sum_{k}^{l}(\mathbf{T}_a)_{ik}(\mathbf{P}')_{kj}\right)^2$$

$$\chi_b^2 = \sigma_b^{-2}\sum_{j}^{m}\sum_{i}^{n}\left(\mathbf{A}_{ij} - \sum_{k}^{l}(\mathbf{T}_b)_{ik}(\mathbf{P}')_{kj}\right)^2$$

Step 4: Calculate the number of degrees of freedom ν:

$$\nu = 2\,nm - (ml + 2nl + l)$$

Step 5: Calculate the reduced χ^2:

$$\chi_\nu^2 = (\chi_a^2 + \chi_b^2)/\nu$$

Step 6: Increase l by 1 and calculate a new reduced χ^2. The number of independent spectral components, r, will be the value of l for which the reduced χ^2 is minimum. If σ_a and σ_b were known, the reduced χ^2 at minimum should be around 1.

ROTATION. The ROTATION program uses the r main loading and score vectors generated by the NIPALS program to calculate the spectra of the components, $\mathbf{v}_i(\lambda)$, their concentrations, c_{ij}, and the ratios between their responses to the two measurements, d_i. The detailed algorithm has been described elsewhere,[14] and the approach is only briefly summarized here.

The Procrustes rotation, $\mathbf{Q}$, of $\mathbf{T}_a$ relative to $\mathbf{T}_b$ is calculated[18-20]

$$\mathbf{Q} = (\mathbf{T}_a'\mathbf{T}_a)^{-1}\mathbf{T}_a'\mathbf{T}_b \tag{6}$$

(15) Fisher, R.; MacKenzie, W. *J. Agric. Sci.* **1923**, *13*, 311–320.
(16) Wold, H. In *Research papers in Statistics*; Daved, Ed.; Wiley: New York, 1966; pp 411–444.
(17) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill Book Co.: New York, 1969.
(18) Eckart, C.; Young, G. *Psychometrics* **1936**, *1*, 211–218.
(19) Schönemann, P.; Carroll, R. *Psychometrics* **1970**, *35*, 245–255.
(20) Gower, J. *Psychometrics* **1975**, *40*, 33–51.

and diagonalized to give the d values:

$$\mathbf{UQU}^{-1} = \mathbf{D} \tag{7}$$

The concentrations (c_{ij}) and spectral profiles, $\mathbf{v}_i(\lambda)$, are finally calculated from

$$\mathbf{C} = \mathbf{T}_a\mathbf{U}^{-1} \tag{8}$$

$$\mathbf{V}' = \mathbf{UP}' \tag{9}$$

Meaning of the Procrustes Rotation. The effect of Procrustes rotation and the diagonalization can be understood by defining $\hat{\mathbf{C}} = \mathbf{CD}$. Equations 4 and 5 become

$$\mathbf{A} = \mathbf{CV}' \tag{10}$$

$$\mathbf{B} = \hat{\mathbf{C}}\mathbf{V}' \tag{11}$$

These have identical $\mathbf{V}'$ matrices, and the $\mathbf{C}$ matrices are related such that each column in $\mathbf{C}$ is the same as in $\hat{\mathbf{C}}$ but for a factor (the d value). From NIPALS we obtained

$$\mathbf{A} = \mathbf{T}_a\mathbf{P}' \tag{12}$$

$$\mathbf{B} = \mathbf{T}_b\mathbf{P}' \tag{13}$$

The problem is to tranform $\mathbf{T}_a$, $\mathbf{T}_b$, and $\mathbf{P}'$ to $\mathbf{C}$, $\hat{\mathbf{C}}$, and $\mathbf{V}'$. Since $\mathbf{CV}' = \mathbf{A} = \mathbf{T}_a\mathbf{P}'$ and $\hat{\mathbf{C}}\mathbf{V}' = \mathbf{B} = \mathbf{T}_b\mathbf{P}'$, the relations between $\mathbf{T}_a$ and $\mathbf{C}$, $\mathbf{T}_b$ and $\hat{\mathbf{C}}$, and $\mathbf{P}'$ and $\mathbf{V}'$ are correlated:

$$\mathbf{C} = \mathbf{T}_a\mathbf{U}^{-1} \tag{14}$$

$$\hat{\mathbf{C}} = \mathbf{T}_b\mathbf{U}^{-1} \tag{15}$$

$$\mathbf{V}' = \mathbf{UP}' \tag{16}$$

where $\mathbf{U}$ is an $r \times r$ square matrix. It can be determined from the correlation beween $\mathbf{C}$ and $\hat{\mathbf{C}}$ by calculating the matrix $\mathbf{Q}$, which upon multiplication by $\mathbf{T}_a$ becomes as much like $\mathbf{T}_b$ as possible, i.e. $\min\|\mathbf{T}_b - \mathbf{T}_a\mathbf{Q}\|$. This matrix is called "the Procrustes rotation of $\mathbf{T}_a$ relative to $\mathbf{T}_b$", after Procrustes, who in the Greek tale lodged travellers in his bed and during their sleep either cut their legs or elongated them to make them fit precisely into the bed. In analogy with Procrustes himself, the Procrustes rotation changes the elements of $\mathbf{T}_a$ to become as close as possible to the corresponding elements in $\mathbf{T}_b$. $\mathbf{Q}$ is calculated by least squares criterion by eq 6. But since $\mathbf{D}$ in $\mathbf{C} = \hat{\mathbf{C}}\mathbf{D}$ is diagonal and $\mathbf{Q}$ is not, it must be diagonalized, which is done in eq 7.

Normalization. The d values calculated by the ROTATION program have their correct values, but the concentrations and spectral profiles must be normalized. The normalization is arbitrary, and one can choose among five alternative ways:

1. Concentrations are calculated relative to those in sample 1:

$$c_{i1} = 1 \qquad i = l, r$$

2. Concentrations are calculated as fractions of the total concentrations of each component:

$$\sum_{j=1}^{n} c_{ij} = 1 \qquad i = l, r$$

3. Concentrations are calculated as fractions of the total concentration in each sample:

$$\sum_{i=1}^{r} c_{ij} = 1 \qquad j = l, n$$

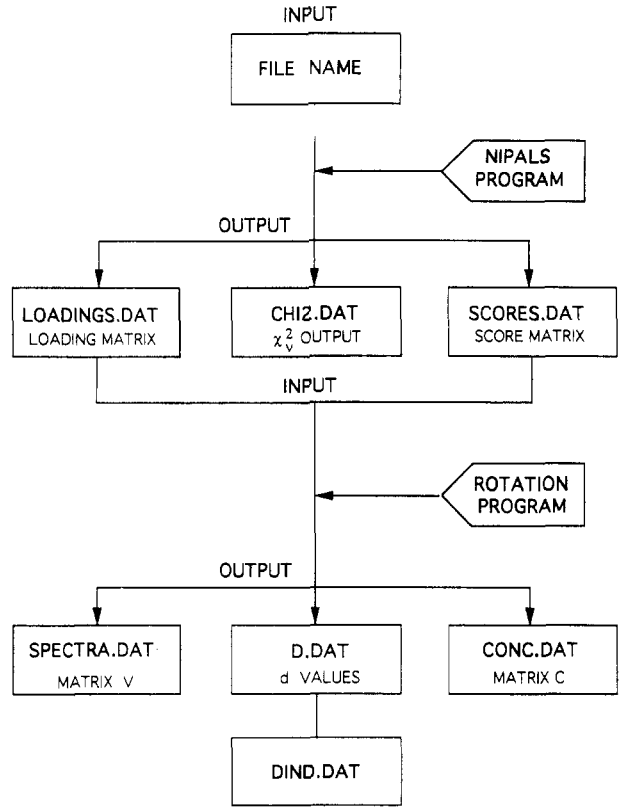

Figure 1. Flow chart of the DATAN program.

4. The areas of the spectral responses are set equal to 1:

$$\sum_{\lambda=1}^{m} \mathbf{v}_i(\lambda) = 1 \qquad i = l, r$$

5. The lengths of the spectral vectors are set equal to 1:

$$\sum_{\lambda=1}^{m} \mathbf{v}_i(\lambda)^2 = 1 \qquad i = l, r$$

The normalization does not affect the accuracy of the analysis, and the data can always be renormalized, if desired.

RESULTS

Ten ($n = 10$) spectra of each kind are generated using three components ($r = 3$), each represented by 100 data points ($m = 100$), and an artificial noise of 25% of the average signal intensity is added: $\mathbf{A} = \mathbf{CV} + 25\%$ noise, $\mathbf{B} = \mathbf{CDV} + 25\%$ noise. The generated spectra are very similar and overlap extensively (Figure 2). Score and loading vectors are calculated for the joint matrix (**A**/**B**) by NIPALS, and reduced χ^2 values are evaluated. The reduced χ^2 decreases steeply with increasing l, reaching a minimum value at $l = 3$, whereafter it slowly increases (Figure 3). From the minimum we can conclude that the number of independent spectral components is three ($r = 3$), and only the three most significant score and loading vectors will be used by the ROTATION program. In our simulations and in analyses of real data, we have found that the χ^2 test very accurately predicts the correct value of r. Indeed, even in simulations with considerably more noise (>200%), where calculated spectra and concentrations have lost most of their features, the χ^2 test predicts the correct number of spectral components. Although the χ^2 test can be used to automatically predict the number of independent spectral components, we have chosen to input r to the ROTATION program to allow complete analysis for any arbitrary number of components. The number of independent components predicted by the χ^2 test can often be

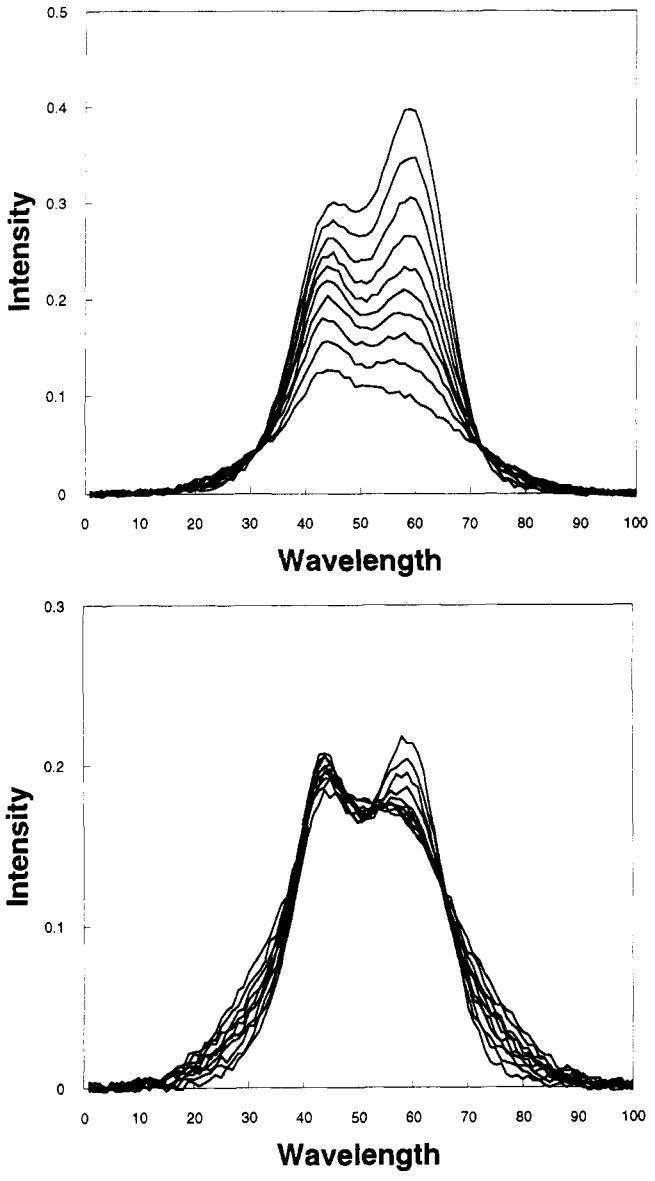

Figure 2. Generated test spectra of type A (a, top) and of type B (b, bottom). There are 10 spectra of each kind (n = 10), each being represented by 100 data points (m = 100). A random noise of 25% of the average signal intensity has been added to each spectrum.

confirmed by visual inspection of the main score and loading vectors. Only the r most significant ones should contain physical features, the remaining ones should contain only noise.[14]

The three main score and loading vectors are passed on to the ROTATION program, which calculates the Procrustes rotation and performs the diagonalization (eq 6 and 7). The diagonal elements are 1.96, 1.02, and 0.52, which should be compared to the d values used in the construction of the test data: 2, 1, and 0.5. The accuracy in the determination of the d values is thus excellent, and the d values are often useful in identifying the unknown components. The ROTATION program also calculates the spectral profiles of the components and their concentrations (eqs 8 and 9). The calculated spectra, $\mathbf{v}_i(\lambda)$, and concentrations, c_{ij}, normalized by alternative no. 1, are compared to those used in construction of the test data in Figures 4 and 5. Despite the extensive overlap between the spectral profiles of the components, the calculated profiles and concentrations are in good agreement with those used in the construction.

Effect of Noise. For reliable data analysis one must know how experimental noise affects the precision in the determination of the various parameters. We define the mean errors as

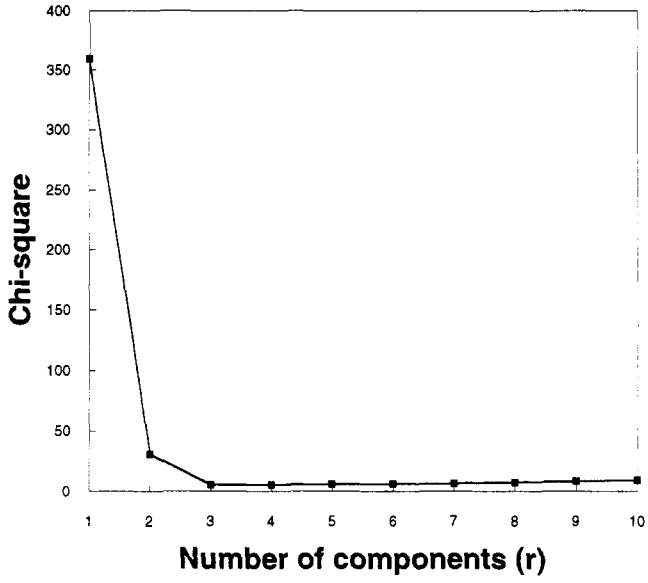

Figure 3. Reduced χ^2 as a function of the number of principal components (r) used in the analysis.

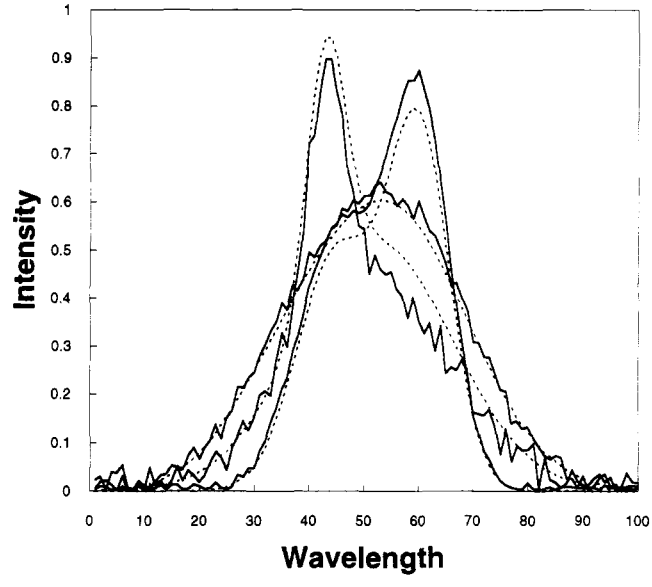

Figure 4. Calculated (—) and original (- - -) spectral profiles.

$$\text{error in V} = \frac{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{r}(v_{ij}^{\text{calc}} - v_{ij}^{\text{orig}})^2}}{m \times r}$$

$$\text{error in C} = \frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{r}(c_{ij}^{\text{calc}} - c_{ij}^{\text{orig}})^2}}{n \times r}$$

$$\text{error in D} = \frac{\sqrt{\sum_{i=1}^{r}(d_i^{\text{calc}} - d_i^{\text{orig}})^2}}{r}$$

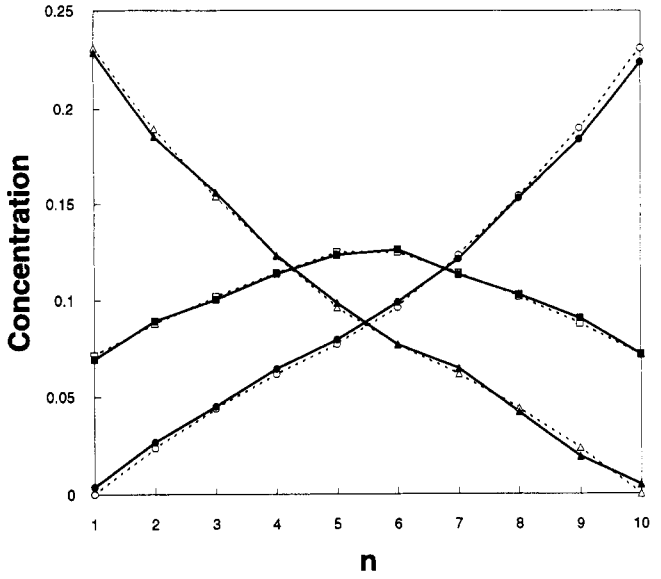

Figure 5. Calculated (filled symbols) and original (open symbols) concentration profiles.

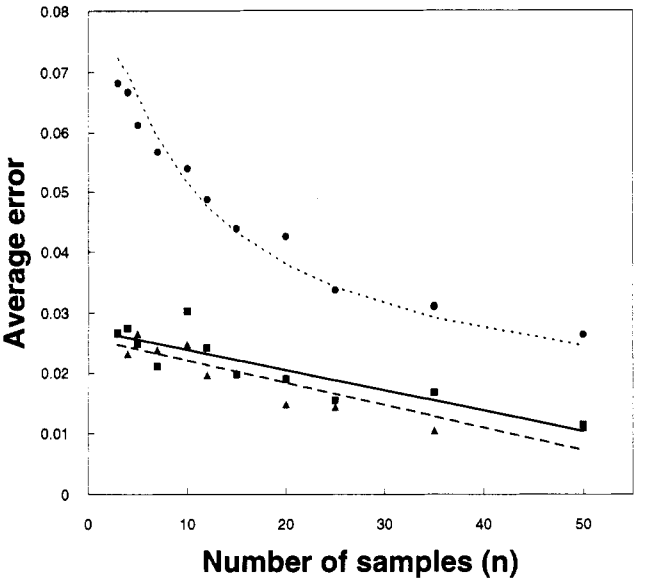

Figure 6. Average errors in calculated **C** (■), **D** (▲), and **V** (●, ×10) as a function of the number of samples used in the analysis. The solid (—) and dashed (- - -) straight lines are fitted to the errors in **C** and **D**, respectively. The dotted line (···) is drawn according to (noise) × $(2n - 3)^{0.5}/2n$, where noise = 0.25 and n = 3.

In all cases, the mean errors are calculated from at least five independent simulations. Simulations show that the mean errors in **C**, **V**, and **D** increase essentially in a linear fashion with increasing noise, as expected intuitively.

Optimizing the Experimental Design—Effect of *n* and *m*. When designing the experiment, one can sample the spectra at different resolutions (collecting different number of data points, m, per spectrum) and one can often vary the number of samples (for example, in a titration experiment one can add the titrand in arbitrary amounts, thereby controlling the number of samples, n, to be analyzed). Since increasing the number of data points per spectrum and the number of samples is time consuming, it is important to know how this affects the precision in the determined parameters. The number of samples (n) and the number of data points per sample (m) affect different dimensions of the data matrices **A** and **B** (increasing n increases the number of rows, increasing m increases the number of columns), and they have different influences on the various calculated parameters. Figure 6 shows the errors in **C**, **V**, and **D** as a function of the number of analyzed samples. The precision in the determinations of **C** and **D** improves only moderately, in an essentially linear fashion, with increasing number of samples, whereas the accuracy in the calculated spectral profiles (**V**) improves more substantially. This is a direct consequence of how n affects the dimensions of **A** and **B**, and thus of **C** and **V**. Increasing n increases the number of data points to be determined in **C**, whereas the number of data points in **V** is unchanged. The dotted line in Figure 6, fitting rather well to the errors in **V**, is drawn according to $0.25(2n - 3)^{0.5}/2n$, suggesting that the error at large n decreases as $n^{-0.5}$. This is the same improvement as when a single spectrum is recorded n times, and we conclude that instead of collecting each spectrum several times, to improve the signal to noise ratio, it is in this respect equivalent to increase the number of samples.

The situation is reciprocal when m is increased: **C** is improved more than **V**. Here, however, there is usually a threshold value below which all results are bad, since when m is too small, the component spectra are hard to separate in the analysis. Of course, the threshold value depends on the degree of spectral overlap between the components and on the noise level. With the test data above, very bad results were obtained with $m < 50$. Because of the reciprocal effect of n and m on the dimensions of **A** and **B**, there should be a threshold value also for n. However, in most real situations the overlap between the spectral profiles of the components is far greater than the overlap between their concentration profiles, the threshold value for n is usually very low, and reasonable results can be obtained when the number of samples is just a few more than the number of components. This is rarely the case for the number of data points per spectrum, which generally has to be considerably larger than the number of components.

Effect of D. *Some d values are similar.* The d values are crucial to the analysis, since it is because of them being different for each component that we can solve equation system 4 and 5. When designing the experiment, one can usually affect the values of the d_i's. For example, when analyzing pairs of emission spectra, the d values are the ratios of the molar absorptivity coefficients at the excitation wavelengths, $d_i = \epsilon_i(\lambda_{ex}{}^b)/\epsilon_i(\lambda_{ex}{}^a)$, and these can be chosen arbitrarily. Figure 7 shows the errors in **C**, **D**, and **V** as a function of the difference between the two most similar d values. The errors in **C** and **V** are proportional to the inverse of the difference, whereas the error in **D** is independent of the values of its elements. Closer inspection of the calculated concentration and the spectral profiles reveals that only those of the two components having similar d values are erroneous; the calculated concentration and spectral profiles of the third component, with a d value significantly different from the others, are correct. The similar d values of the two components result in a mixing of their contributions in the analysis: $\mathbf{C}_{12}{}^{calc} = \mathbf{C}_{12}{}^{orig}\mathbf{R}^{-1}$ and $\mathbf{V}_{12}{}^{calc} = \mathbf{R}\mathbf{V}_{12}{}^{orig}$, where the subscript denotes the submatrices containing the elements from the two components with similar d values. Since the mixing of **C** and **V** is reciprocal, the correct concentration profiles can be calculated if the correct spectral profiles can be obtained (i.e., if the components can be identified). In general, the calculated concentration and spectral profiles are mixed in the analysis pairwise to a degree that is proportional to the inverse of the difference between their d values. Therefore, when designing the experiment, one should maximize the difference between the most similar d values. In special cases, when a component is more interesting than others, one can maximize the difference between its d value and the most similar d value of the other components.

One component does not contribute to one of the two sets of spectra. If one component has no contribution to the

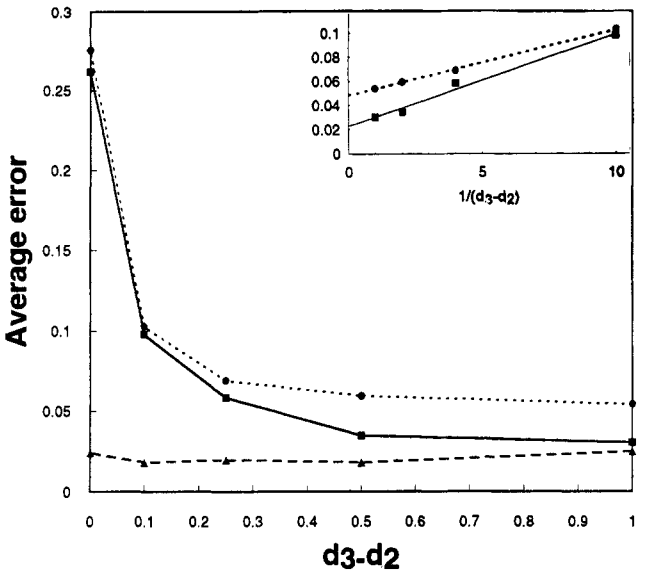

Figure 7. Average errors in calculated **C** (■), **D** (▲), and **V** (●, ×10) as a function of the difference between d_3 and d_2. $d_1 = 0.5$; $d_2 = 1$; noise level 25%. The inset shows the errors in **C** and **V** plotted versus $(d_3 - d_2)^{-1}$.

B-type spectra its d value is zero, whereas if it has no contribution to the A-type spectra its d value is infinite. The analysis works for $d = 0$, but it fails for $d = \infty$. When $d = \infty$ some of the calculated d values are complex numbers having no meaning, and all other d values are erroneous. It is important that the diagonalization routine checks the imaginary parts of the d values and gives a warning message if any is significant.

Checking the validity of the analysis. The ratios of the responses of the components to the two types of spectra, the d values, are effectively calculated by a least squares approach inherent to the Procrustes rotation (eq 6). By postmultiplying $\mathbf{T}^a$ and $\mathbf{T}^b$ by $\mathbf{U}^{-1}$, one obtains the contributions of the components to the experimental spectra (eqs 14 and 15). $(\mathbf{T}^a\mathbf{R})_{ij}/(\mathbf{T}^b\mathbf{R})_{ij}$ is the ratio between the contributions to the A- and B-type spectra of component i in sample j. For each one of the components these ratios should be the same for all samples, and they should equal the d values. If a ratio differs significantly from the corresponding d value, the calculated concentration of that component in the particular sample has probably a large error. This happens, for example, when a component is not present in a sample (zero concentration). Further, if there is a large spread among these ratios for a certain component, it probably does not fulfill the requirements for the analysis, and the results should be interpreted with care. The $(\mathbf{T}^a\mathbf{R})_{ij}/(\mathbf{T}^b\mathbf{R})_{ij}$ ratios are calculated by the DIND routine in the DATAN program.

Different Signal to Noise Ratios in the Two Types of Spectra. The result of the analysis depends on which set of spectra is treated as A and which is treated as B. When the noise levels in the two sets are different, which is the usual case owing to their different natures, this choice is important. In general, it is better to treat the set with less noise as A. For example, with 5% noise in **A** and 50% noise in **B**, the errors in calculated **C**, **D**, and **V** are 0.03, 0.03, and 0.006. If the sets are interchanged (50% noise in **A** and 5% noise in **B**), the errors are 0.48, 0.01, and 0.029. Although the interchange has no effect on the d values, the errors in the calculated spectral profiles become significantly larger, and the errors in the calculated concentration profiles become very large. Comparing with the errors when the noise levels in **A** and **B** are the same, the errors with 5% noise in **A** and 50% noise in **B** correspond to an overall noise level of the average, 27%, in **A** and **B**. On the other hand, with 50% noise in **A** and 5% noise in **B**, the error in **V** corresponds to an overall noise level of about 110% in **A** and **B**, and the error in **C** corresponds to a noise level of about 325%!

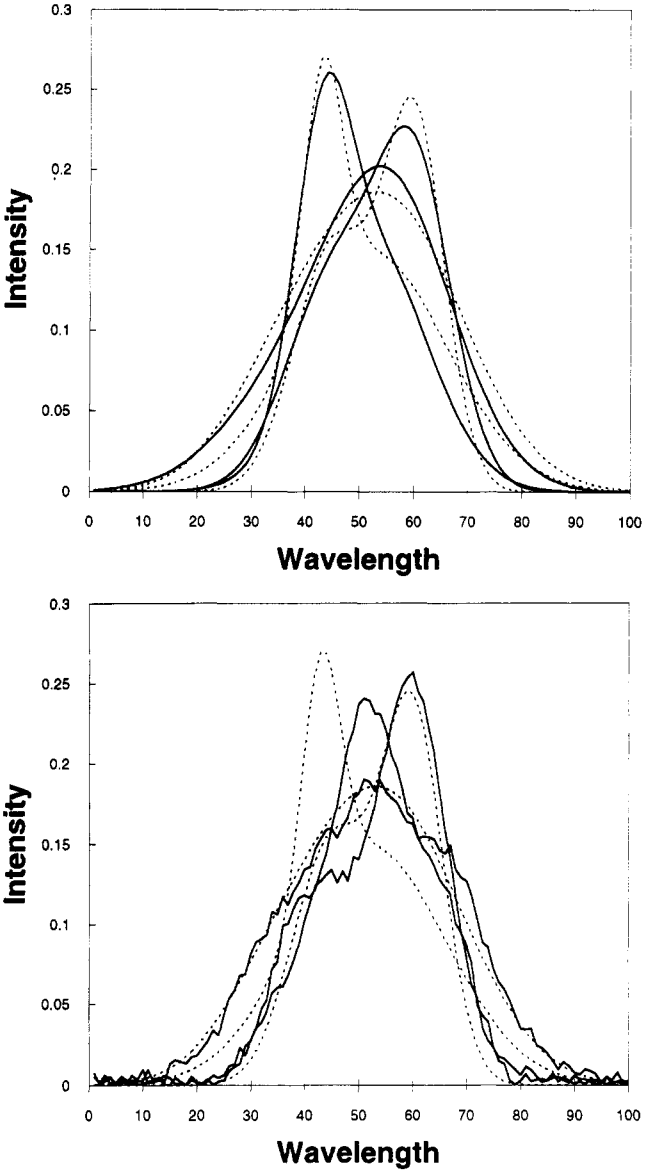

Figure 8. (a, Top) two sets of spectral profiles used to generate **A** and **B**, to which the components have somewhat different contributions. (b, Bottom) Calculated spectral profiles (—) compared those used to generate **A** (- - -).

Variations in Spectral Profiles. *The spectral profiles of the components are different in the two kinds of spectra.* Although many kinds of spectra give identical profiles for each component,[14] some spectra have different sensitivities to the underlying contributions to the spectral intensities and the profiles in set A and set B may be somewhat different. For example, all vibronic effects are positive in absorption spectra but some may give negative contributions to circular dichroism spectra. As a consequence, the spectral profiles of the components in the two sets of spectra will be slightly different. Unfortunately, the analysis is very sensitive to such effects. Figure 8a shows two slightly different sets of component spectral profiles used to generate new data matrices **A** and **B**. The calculated profiles are significantly different, and they are not averages of the pairs used in the construction (Figure 8b). Although their shapes bear some similarities with the original profiles, their magnitudes are clearly different and they are considerably shifted. The calculated concentration profiles and d values were even worse

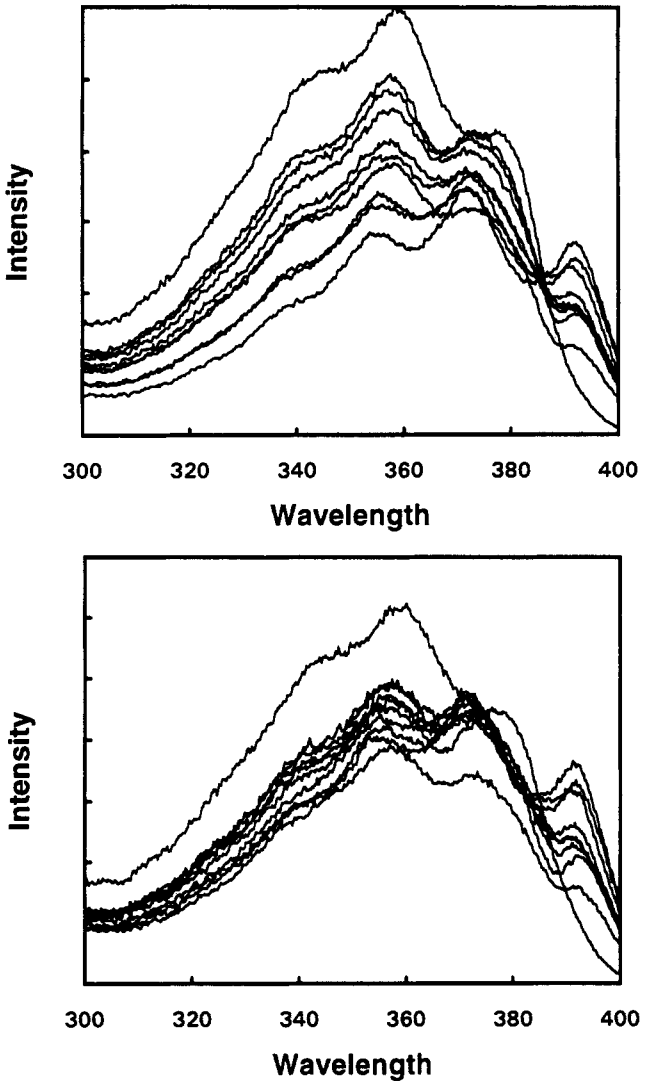

Figure 9. Fluorescence excitation spectra of mixtures containing 1,4-bis[(5-phenyloxazol-2-yl)]benzene (POPOP), dimethyl POPOP (DMPOPOP), and diphenylanthracene (DPA) in cyclohexane. Spectra were recorded with 410 (a, top) and 430 nm (b, bottom) emission. Excitation and emission spectral bandwidths were 1 nm, scanning rate 50 nm/min, time constant 1 s. Spectra are quantum corrected and corrected for the inner-filter effect. The data were kindly provided by Drs. Svante Eriksson and Bo Albinsson.

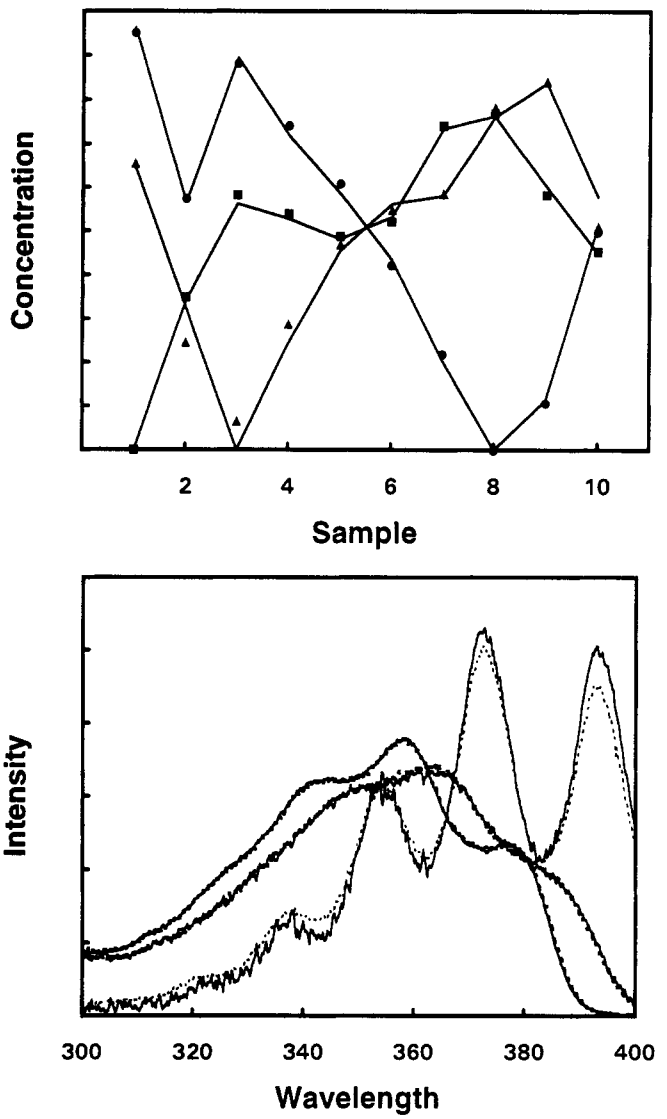

Figure 10. (a, Top) calculated concentrations of POPOP (●), DMPOPOP (▲), and DPA (■), compared with their correct concentrations (straight lines), and (b, bottom) their calculated spectral profiles (solid lines) compared with excitation spectra recorded on the free dyes (dashed lines).

(data not shown). Interchanging the data matrices **A** and **B** has no effect on the result. This kind of problem can usually be realized from the χ^2 test: the reduced χ^2 value decreases steeply until $l = 3$ but does not attain a minimum value, and it is considerably larger than 1. It continues to decrease, reporting a larger number of components.

Experimental drift results in small wavelength shifts. A similar situation arises if there is an experimental drift resulting in small wavelength shifts in the spectra. Also here the analysis predicts a too large number of components, and the result is erroneous. This problem can usually be avoided by decreasing the resolution of the measurements making the drift negligible. Because of the very high resolving power of the analysis, it is rarely necessary to push the resolution of the experiments to the limit. In contrast, a drift in intensity causes very little problem. The analysis provides the correct spectral profiles of the components, though there will be a small effect on the calculated **C** and d values.

Similar Concentration Profiles. For the analysis to work properly, the concentration profiles of the components must all be different. If two components have the same concentration profiles, that is when the ratios between their concentrations in all samples are the same, the results will be erroneous. This may occur, for example, when a species is invoked in a partition equilibrium.[21] We simulated the case by calculating a new concentration profile of component 2, making it progressively more similar to the profile of component 1:

$$c_{2i}^{\text{new}} = c_{1i} + x(c_{2i}^{\text{old}} - c_{1i})$$

When the degree of mixing is increased (decreasing x), the result becomes progressively worse. For $x = 0$ the concentration profiles of components 1 and 2 are identical, and some of the results were completely wrong. The calculated spectral profiles of the two components having the same concentration profiles were mixed, and their calculated d values were wrong. Their concentration profiles were calculated correctly, but they were scaled erroneously:

$$c_{1i}^{\text{calc}} = \text{const} \times c_{1i}^{\text{orig}}$$

$$c_{2i}^{\text{calc}} = \text{const}' \times c_{2i}^{\text{orig}}$$

Still, the important conclusion, that the ratios between their

(21) Chiesa, M.; Domini, I.; Samori, B.; Eriksson, S.; Kubista, M.; Nordén, N. *Gazz. Chim. Ital.* **1990**, *120*, 667–670.

concentrations in all samples are the same, can be made. The spectral profile and the d value of the third component, having a unique concentration profile, are calculated correctly, but its calculated concentration profile is completely wrong.

Example with Experimental Data. The Procrustes approach was used to analyze samples containing mixtures of 1,4-bis[(5-phenyloxyazol-2-yl)]benzene (POPOP), dimethyl POPOP (DMPOPOP), and diphenylanthracene (DPA) in cyclohexane. Fluorescence excitation spectra were recorded on the samples using 410 (Figure 9a) and 430 nm (Figure 9b) emissions. The dyes obey the Kasha–Vavilov rule,[22] having identical spectral profiles in the two measurements. The magnitudes of their responses are proportional to their concentrations, fluorescence quantum yields, and the fraction of the total emission observed at the emission wavelength of the experiments. The d values, being the ratios of their responses to the two measurements, are the ratios between the fractions of their total fluorescence at the two emission wavelengths: $d_i = I(\lambda_{em,2})_i / I(\lambda_{em,1})_i$. The spectra were digitized into 501 data points each and analyzed by the DATAN program. The χ^2 test correctly predicted three independent spectral components, and three loading and score vectors were used in the ROTATION. The calculated d values for the three dyes were 0.77, 0.43, and 0.62, and the calculated concentrations and spectral profiles are shown in Figure 10.

DISCUSSION

We have described the DATa ANalysis (DATAN) program to analyze correlated spectroscopic data. The program, which is based on the Procrustes rotation method,[14] calculates the number of independent spectral components (r), their spectral responses ($\mathbf{V}'$) and concentrations ($\mathbf{C}$), and the ratios between their responses to two spectroscopic measurements ($\mathbf{D}$), using only the experimental spectra as input. From extensive tests of the program, we conclude that the results are highly accurate and reliable when the spectral profiles, the concentration profiles, and the d values of all the components are different. The results can be summarized in a few points:

- The number of independent spectral components is accurately predicted by the χ^2 test.
- The accuracy in the calculated parameters can be improved by either increasing the number of samples (n) or by increasing the number of data points per spectrum (m). Increasing n improves mainly the accuracy in $\mathbf{V}'$; increasing m improves mainly the accuracy in $\mathbf{C}$.
- The d values are crucial to the analysis, and the experiment should be designed to make them as different from each other as possible.
- The result depends on which data set is treated as A and which is treated as B. In general, the set with less noise should be treated as A.
- A d value may be zero, but not infinite.
- If some calculated d values have significant imaginary parts, the analysis has gone wrong and the result is unreliable.
- The analysis is very sensitive to changes in spectral profiles, and the experiments should be designed to minimize spectral shifts.

The DATAN program is available from the authors (M.K.).

ACKNOWLEDGMENT

We thank Dr. Björn Sjögren at the department of scientific computing at Uppsala University, Sweden, for valuable discussions of how to perform the error analysis, and Drs. Svante Eriksson and Bo Albinsson for valuable discussions of experimental design and for providing us with Figures 9 and 10. This project is supported by Stiftelsen Wilhelm och Martina Lundgrens Vetenskapsfond and the Swedish Natural Research Council.

RECEIVED for review July 28, 1992. Accepted November 5, 1992.

(22) Turro, N. J. *Modern Molecular Photochemistry*; The Benjamin Cummings Publishing Co.: Menlo Park, CA, 1978.