Resolution of Severely Overlapped Spectra from Matrix-Formated Spectral Data Using Constrained Nonlinear Optimization

Sharon L. Neal,[1] Ernest R. Davidson,[2] and Isiah M. Warner*[,3]

Department of Chemistry, Emory University, Atlanta, Georgia 30322, and Department of Chemistry, Indiana University, Bloomington, Indiana 47405

A three-step scheme for resolving severely overlapped component spectra from bilinear matrix-formated data is reported. After the number of sample components is determined, a positive basis is first formed consisting of the most dissimilar rows and columns of the matrix. The concentration factor matrix (CFM) corresponding to this nonnegative, minimally correlated basis will be diagonal if the basis vectors happen to be feasible estimates of the component spectra. When the CFM is not diagonal, a constrained nonlinear optimization routine is used in a second step to reduce the off-diagonal elements of the CFM to zero while maintaining the nonnegativity of the estimated spectra. In many cases, the nonnegativity and feasibility constrains are not sufficient to produce a unique set of component spectra estimates. Other criteria, such as the degree of overlap of the resolved spectra, may be used as the basis of a third step to generate an arbitrary, but unique, choice among the feasible estimates of the component spectra. The performance of this scheme is evaluated by analyzing synthetic and experimental fluorescence excitation–emission matrices (EEMs) exhibiting various levels of spectra overlap and random noise. Coincident spectra can be resolved from an EEM by using this approach in the case of some EEMs of rank greater than two. Evaluations using synthetic data indicate that this scheme can be applied to EEMs that have signal-to-noise ratios above 18. Successful resolution of experimental three- and four-component mixtures is illustrated.

INTRODUCTION

The resolution of unknown component spectra from the two-dimensional spectra of mixtures is one of the most challenging mathematical problems in chemical analysis. The application of factor analysis (FA) (*1, 2*) to this problem has received quite a bit of attention in the recent literature due, no doubt, to the rather spectacular ability of these methods to extract predominant spectral features from raw data without a priori knowledge or assumptions specific to the component spectra. Having recognized that the singular vectors of the matrix, also known as principal components or abstract factors, are linear combinations of the component spectra, analysts have searched extensively for techniques to extract the component spectra from this orthonormal basis. The first FA curve resolution methods (*3, 4*) used nonnegativity and normalization constraints to resolve spectra from the orthogonal basis. Later efforts focused on finding reasonable estimates of the component spectra among the spectra forming the data matrix (*5, 6*) or on adjusting the estimated spectra iteratively (*7–9*) to meet nonnegativity or dissimilarity criteria.

The performance of these methods varies widely, but all such methods have limited utility in the case of severe spectral overlap. In this report, we describe improved resolution of overlapped spectra, by using algebraic and spectroscopic properties of the data matrix to estimate the profiles of the component spectra. We illustrate the use of this method with second-order fluorescence data, but the approach is general for any bilinear (*10*), nonnegative matrix of digitized spectral data.

THEORY

Analysis of a dilute solution of n fluorophores by videofluorometry (*11*) produces a $p \times q$ data matrix, $\mathbf{E}$. Each element, e_{ij}, represents the fluorescence intensity of the sample at wavelength λ_j, generated by excitation at wavelength λ_i. This matrix has been called the excitation–emission matrix (EEM) (*12*). The emission spectrum of a single fluorophore is independent of the excitation wavelength, so that $\mathbf{E}$ has the form $\gamma\mathbf{x}\mathbf{y}^{\mathrm{T}}$ where $\mathbf{x}$ is a normalized column vector of the digitized excitation spectrum, $\mathbf{y}$ is a normalized column vector of the digitized emission spectrum, and the constant γ is a concentration-related intensity factor. For dilute multicomponent solutions, the matrix $\mathbf{E}$ is simply the sum of the $\mathbf{E}$ matrices of the individual components. In other words, the excitation and emission spectra of the sample components are basis vectors for the column and row spaces, respectively, of the EEM. Mathematically, the multicomponent EEM is represented by the expression

$$\mathbf{E} = \mathbf{X}\boldsymbol{\Gamma}\mathbf{Y}^{\mathrm{T}} \qquad (1)$$

where the columns of the $p \times n$ matrix $\mathbf{X}$ and the $q \times n$ matrix $\mathbf{Y}$ are the normalized excitation and emission spectra, respectively. The elements of the diagonal $n \times n$ matrix $\boldsymbol{\Gamma}$ are concentration-dependent scalars. A matrix of this form, which is the sum of the outer products of the component responses, has been called a bilinear matrix (*10*).

Unfortunately, the number and form of the component spectra are not generally apparent in experimental EEMs. Therefore, the component spectra must be estimated by using linear combinations of some more easily determined basis, e.g.

$$\mathbf{X} = \mathbf{A}\boldsymbol{\Pi} \qquad (2)$$

$$\mathbf{Y} = \mathbf{B}\boldsymbol{\Theta} \qquad (3)$$

where the columns of $\mathbf{A}$ and $\mathbf{B}^{\mathrm{T}}$ are basis vectors for the column and row spaces of $\mathbf{E}$ and the columns of $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}$ are the coordinates of the component spectra with respect to those basis vectors. Without constraints, any attempt to factor $\mathbf{E}$ into the form of eq 1 usually yields an infinite number of solutions.

In this report, a scheme for choosing $\mathbf{A}$ and $\mathbf{B}$ and approximating the values of $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}$ is described. This approach

[1] Present address: Department of Chemistry, Spelman College, Atlanta, GA 30314.
[2] Indiana University.
[3] Emory University.

0003-2700/90/0362-0658$02.50/0 © 1990 American Chemical Society

imposes several fundamental properties of fluorescence spectra on the estimated spectra. The conditions under which this scheme is successful in extracting the true spectra are also discussed.

Choosing a Nonnegative Basis. The orthonormal basis vectors of the column and row spaces of the EEM provided by the singular value decomposition (SVD) (*13*) form an accessible basis that is frequently used in spectra estimation (*3–9*). In addition to the relative ease of computing this basis (*14*), the structure of the decomposition provides a mechanism for estimating the number of fluorescent components in the sample. In the orthonormal basis, the ideal EEM is given by the expression

$$\mathbf{E} = \mathbf{R}\boldsymbol{\Psi}\mathbf{C}^{\mathrm{T}} \tag{4}$$

where the n columns of **R** and **C** are the excitation and emission (column and row) singular vectors, respectively. The diagonal elements of $\boldsymbol{\Psi}$ are the singular values which reflect the fraction of the matrix variance associated with the corresponding pair of column and row basis vectors. Noise increases the rank of experimental EEMs to the smaller dimension of **E** so that m ($m = \min(p,q)$) rather than n basis vectors are required to describe the matrix. Determination of the number of sample components is trivial in samples that are sufficiently concentrated, because there are singular values associated with the spectra signal which are significantly larger than those that merely reflect noise in the data. Alternative indicators, such as the autocorrelation coefficients of the singular vectors (*15*) or cross-validation (*16*) can be used for more dilute samples. Once the number of sample components is determined, the n significant column and row singular vectors may be used to construct a least-squares approximation to the experimental EEM. These n primary singular vectors also provide a reduced basis in which the matrices **X** and **Y** can be expanded. Therefore, many matrix calculations can be performed more efficiently in the reduced basis, yet yield identical results to calculations utilizing the entire least-squares matrix.

However, the singular vectors are not usually good estimates of the component spectra. They are merely mathematical devices that generally lack significant properties of true fluorescence spectra, most notably nonnegativity. On the other hand, the n most dissimilar rows and columns of **E** will also form bases for the row and column spaces of **E**, with the additional advantage of being positive. Unfortunately, choosing these n vectors from the set of all p rows and q columns is a complicated combinatorial problem which requires examination of all n-element combinations of the p rows and q columns.

In this paper, a sequential (noncombinatorial) procedure for selecting n independent columns of **E** is presented. The pair of columns with the smallest scalar product are selected as the first two members of the basis. The columns $\{\mathbf{a}_1,\mathbf{a}_2\}$, which meet this criterion are found by comparing the off-diagonal elements of the column correlation matrix, **Z**, which are the scalar products of the normalized columns of **E**, given by $\bar{\mathbf{E}} = \mathbf{EN}$, where **N** is a diagonal matrix of the inverse square roots of the norms of the columns of **E**. The column correlation matrix is efficiently computed from the singular vectors by using the expression

$$\mathbf{Z} = \bar{\mathbf{E}}^{\mathrm{T}}\bar{\mathbf{E}} = \bar{\mathbf{Q}}^{\mathrm{T}}\bar{\mathbf{Q}} \tag{5}$$

where $\bar{\mathbf{Q}}$ is $\boldsymbol{\Psi}\mathbf{C}^{\mathrm{T}}\mathbf{N}$.

Once the first two columns have been chosen, an iterative procedure is used to chose the remaining columns of **A**. The overlap matrix $\mathbf{S}_t$ (where t is the number of columns chosen for **A**) is the product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ obtained by using the existing columns of **A**. The matrix $\mathbf{T}_i$ is the overlap matrix which would result if the ith column of **E** were chosen as the $t + 1$ column of **A**. The determinant of $\mathbf{T}_i$ is

$$|\mathbf{T}_i| = |\mathbf{S}_t \|\mathbf{e}_{\mathrm{res}}\|^2 \tag{6}$$

where

$$\|\mathbf{e}_{\mathrm{res}}\|^2 = \|\mathbf{e}_i - \mathbf{S}_t^{-1}\mathbf{e}_i\|^2 \tag{7}$$

and $\mathbf{e}_{\mathrm{res}}$ is the residual of expanding the ith column of **E** in terms of the existing columns of **A**. The column of **E** which produces the largest $|\mathbf{T}_i|$ had the largest residual when expanded in terms of the existing columns of **A**. Therefore, adding this column to the basis will account for the largest portion of the variance not represented in the set $\{\mathbf{a}_k\text{: } k = 1...t\}$. This column is added to the columns of **A** and the procedure is repeated until n columns of **E** have been chosen as basis vectors for the excitation spectra. This procedure, which we are calling best measured basis (BMB), can be applied similarly to the columns of $\mathbf{E}^{\mathrm{T}}$ to determine the nonnegative basis vectors for the row space of **E**.

Imposing Bilinearity. In the columns of **A** and **B**, there are now basis vectors which span the spaces of **X** and **Y** and are nonnegative, as are fluorescence spectra. These properties make these vectors particularly well-suited for use as basis vectors in estimating the component spectra by eq 2 and 3. In fact, they may occasionally themselves, be the component spectra. However, these basis vectors are generated independently, i.e. without regard to the bilinear structure of **E**. Matrices that span the column and row spaces of **E**, **V**, and $\mathbf{W}^{\mathrm{T}}$, for example, are consistent with bilinear structure, if $\mathbf{E} = \mathbf{W}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ where $\boldsymbol{\Sigma}$, called the concentration factor matrix (CFM) in this report, is a positive diagonal matrix. Pairs of matrices that meet this condition are feasible estimates component spectra and will be referred to as compatible spectra.

There is no reason to expect that the columns of **A** and **B** are compatible with each other in this sense, though they span their respective spaces and are positive. Large off-diagonal elements of the CFM associated with the nonnegative basis indicate that the basis vectors are incompatible. The CFM is given by the expression

$$\Delta = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{E}\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1} \tag{8}$$

In some cases, the CFM generated by using eq 8 is not diagonal, even though the columns of **A** and **B** are feasible estimates of the component spectra. In this case, the matrix Δ can be made diagonal by permutation of the columns of **A** and **B**. The columns of **A** and **B** are always permuted to place the largest elements of Δ onto the diagonal at this point in the resolution scheme. Even when **A** and **B** are not compatible, this procedure orders the basis vectors so that they are more amenable to the rest of the scheme.

In the instance of incompatibility, a constrained nonlinear optimization method can be used to adjust the values of $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}$ so that the estimated spectra are compatible as well as nonnegative. One of the disadvantages of constrained optimization is the capacity of these algorithms to collapse on the steep response surface grades formed by the constraints rather than converge on local optima. This problem is exacerbated as the number of constraints and variables increases; therefore, the program NLPQL (*17*), which is an implementation of the projected Lagrangian methodology (*18, 19*), was used, rather than the more commonly used simplex algorithm (*20*). One feature of these methods is that the performance of optimization methods is sensitive to the initial estimates of the variables. This is another reason that the nonnegative basis is more suitable than the singular vectors as basis vectors for the component spectra.

The CFM formed by **A** and **B** is diagonalized subject to the constraints that the estimated spectra remain nonnegative and normalized to unit length. This optimization is stated mathematically as

$$\text{minimize} \qquad \Sigma_{k\neq 1}\hat{\gamma}^2_{\mathrm{kl}} \tag{9}$$

$$\text{where} \qquad \hat{\Gamma} = \Pi^{-1}\Delta(\Theta^{-1})^{\mathrm{T}}$$

subject to

$$\hat{\gamma}_{kk} \geq 0,$$

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\Pi} \geq 0,\ \langle\hat{\pi}_k,\hat{\pi}_k\rangle = 1$$

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\Theta} \geq 0,\ \langle\hat{\theta}_k,\hat{\theta}_k\rangle = 1$$

$$\Pi_{\mathrm{init}} = \Theta_{\mathrm{init}} = \mathbf{I}_n$$

For an input matrix **E**, which can be written in the form of eq 1, the minimum of $\Sigma_{k\neq 1}\gamma^2_{\mathrm{kl}}$ will always be zero. The resulting Π and Θ will produce compatible sets of factors on the columns of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ when used in eq 2 and 3. This algorithm, referred to hereafter as the concentration factor matrix diagonalization (CFMD), is a generalization of Lawton and Sylvestre's self-modeling curve resolution (SMCR) (*3*) to *n* components. The constraints on the norms of the transforms were included to maintain the separation of the concentration and spectra information so that unique solutions can be generated. These constraints also reduce the number of degrees of freedom in Π and Θ from $2n^2$ to $2n(n-1)$.

Several authors (*3, 4, 21, 22*) have described the geometric significance of the nonnegativity constraints on the estimated spectra for two-, three-, and four-component systems. Accurate graphical representations of the constraints require elaborate iterative methods even for three-component systems (*21*). Therefore, a general algorithm for depicting the regions of feasible estimates of the component spectra has not been described. The generalization to *n* components is simpler to describe algebraically. The nonnegativity constraints on the excitation spectra are given by the equations

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\Pi} \geq 0 \tag{10}$$

Those on the emission spectra are given by

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\Theta} \geq 0 \tag{11}$$

From $\mathbf{E} = \mathbf{X}\Gamma\mathbf{Y}^{\mathrm{T}} = \mathbf{A}\Delta\mathbf{B}^{\mathrm{T}}$

$$\hat{\Theta} = \Delta^{\mathrm{T}}(\hat{\Pi}^{-1})^{\mathrm{T}}\hat{\Gamma}^{-1} \tag{12}$$

so that there are additional constraints on the coordinates of the excitation spectra

$$\hat{\mathbf{Y}} = \mathbf{B}\Delta^{\mathrm{T}}(\hat{\Pi}^{-1})^{\mathrm{T}}\hat{\Gamma}^{-1} \geq 0 \tag{13}$$

These constraints on both $\hat{\Pi}$ and $\hat{\Pi}^{-1}$ may lead to complete specification of some elements of $\hat{\Pi}$ in favorable cases. similarly, constraints on $\hat{\Theta}$ and $\hat{\Theta}^{-1}$ can specify elements of $\hat{\Theta}$.

Choosing Solutions from the Feasible Regions. The relationship between the compatible solutions generated by CFMD and the true component spectra is a function of the overlap of the component spectra. As the degree of overlap increases, the region between the row and column nonnegativity boundaries is widened. In these cases, there are an infinite number of feasible spectral estimates and simple diagonalization of $\hat{\Gamma}$ will not generate a unique solution for eqs 2 and 3.

Table I. Objective Functions That Impose Common Features on Resolved Spectra

objective function	features imposed
$\max[\hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}}) \times \det(\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{Y}})]$	maximizes dissimilarity
$\max[\sum_k\hat{\pi}_{kk}{}^2 + \hat{\theta}_{kk}{}^2]$	maximizes similarity to **A** and **B**
$\min[\sum_k(\sum_i\hat{x}_{ik} + \sum_j\hat{y}_{jk})]$	maximiizes simplicity
$\min[\sum_k(\sum_i\hat{x}_{ik}\ \ln\ \hat{x}_{ik} + \sum_j\hat{y}_{jk}\ \ln\ \hat{\gamma}_{jk})]$	maximizes simplicity (entropy minimization)

In the usual event that a range of compatible solutions exists, it is useful to have some criterion for choosing one solution from that range. The nonlinear optimization routine provides a flexible mechanism for incorporating a number of spectral features into a unique solution. Algorithms that minimize the dissimilarity (*7, 8*) or entropy (simplicity) (*9*) of the estimated spectra have been used for this purpose. By use of the constrained nonlinear optimization routine used in CFMD, these spectra features can be imposed on the resolved spectra by careful choice of the objective function used to generate the solution. Moreover, this approach will always produce a compatible, nonnegative solution. The general form of the optimization is given by the expressions

$$\text{optimize} \qquad f(\hat{\Pi},\hat{\Theta}), \tag{14}$$

subject to

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\Pi} \geq 0,$$

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\Theta} \geq 0,$$

$$\hat{\gamma}_{kk} \geq 0$$

$$\hat{\gamma}_{kl} = 0 \text{ for } k \neq 1,$$

$$\hat{\Pi}_{\mathrm{init}} = \hat{\Pi}_{\mathrm{CFMD}};\ \hat{\Theta}_{\mathrm{init}} = \hat{\Theta}_{\mathrm{CFMD}}$$

where $f(\hat{\Pi},\hat{\Theta})$ is the objective function that imposes the desired spectra features. The entropy (or simply the sum of the spectra elements) measures the degree of dispersion of information across the spectral elements. Minimizing the entropy localizes the spectral bands. The correlation between vector similarity and distance is the basis of the objective function for the most dissimilar solution. The coordinates of the component spectra form *n* simplices inside which the coordinates of the rows and columns of the EEM must fall, because the rows and columns of **E** form convex sets. The square of the volume of the simplex formed by the coordinates of the estimated spectra in the row and column spaces of *E* is given by the determinants of the overlap matrices (eq 6) of the estimates. Maximizing those volumes produces the most dissimilar set of feasible component spectra. The objective function, which is maximized as the transforms $\hat{\Pi}$ and $\hat{\Theta}$ are diagonalized, implements the principles of target factor analysis methods such as key set analysis (*6*) or the best measured basis algorithm described above, by seeking the feasible solutions most like the columns of **A** and **B**. The justificaiton for incorporating properties such as dissimilarity into the estimated spectra is primarily intuitive and heuristic, but the flexibility of the objective function approach, permits the analyst to adjust the degree of incorporation of spectral features to the information at his or her disposal. Table I lists four typical objective functions and the features they impose on the spectral estimates.

Solution Simplex Minimization. When the analyst has information concerning the data which warrants the choice of a particular unique solution, the use of specialized objective functions can produce a more useful solution than CFMD. For example, if the analyst is confident that the component spectra are not all severely overlapped, the least dissimilar solution, subject to compatibility and nonnegativity constraints, will frequently be more like the component spectra than the most dissimilar. This solution can be generated by aligning the coordinates of the spectral estimates with as many of the row and column coordinates as possible subject to the other constraints.

Several objective functions can be used to align the solution simplex with the row and column coordinates. Theoretically, the entropy function and the sum of the squares of the elements of the estimated spectra generate this solution. Minimizing the volume of the solution simplices is an alternative function which has several attraction features. First, minimizing the volume avoids the tendency of the sums, and to a lesser extent the entropy, to weight the elements of the spectra unevenly, resulting in an estimated spectra with a larger average bandwidth than the optimal solution. Moreover, the function describing the volume is easier to differentiate than the entropy function, permitting the use of analytic gradients in the optimization routine which decreases the computation time significantly. In order to utilize the resolution information in both vector spaces, the product of the volumes of the excitation and emission solution simplices is used as the objective function in this third step. In this algorithm, which is referred to as solution simplex minimization (SSM), the objective function is given by the expression: $\det(\hat{\mathbf{X}}^T\hat{\mathbf{X}}) \times \det(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}})$.

EXPERIMENTAL SECTION

Reagents. The compounds 9,10-dimethylanthracene, 2,3-benzanthracene, anthracene, perylene, and fluoranthene (Aldrich Chemical Co., Milwaukee, WI) were acquired at 99%+ purity and used without further purification. Various combinations of these compounds were dissolved in glass-distilled cyclohexane (Burdick and Jackson, Muskegon, MI).

Apparatus. A Hewlett-Packard 9845B desktop minicomputer, equipped with a 7908 hard disk (Hewlett-Packard, Palo Alto, CA) was used to control the acquisition of second-order fluorescence spectra by the videofluormeter. The details of the spectrometer specifications and operation are described elsewhere (*23*). The data were transferred by HP terminal emulation software to a MicroVAX II (Digital Equipment Corp., Maynard, MA) for analysis and storage. The resolution programs are coded in FORTRAN 77.

RESULTS AND DISCUSSION

The original description of SMCR (*3*) assumed that resolvable spectra lie on the nonnegativity boundary and have characteristic responses to the experimental parameters. Yet, it is clear from convex set theory (*24*, *25*) that in mixtures of more than two components, characteristic bands place row and column coordinates on the vertices of the simplex formed by the component spectra and they need not lie on the nonnegativity boundary. On the other hand, characteristic base lines produce coordinates that lie on the nonnegativity boundary. It follows, then, that it is not the presence of characteristic bands that ensures the unique resolution of component spectra but the existence of base-line responses to wavelengths to which the remaining components are responsive. For two-component mixtures, these two conditions are identical because a characteristic base line for one component must occur at the wavelength of a characteristic band for the other component.

A synthetic three-component EEM was generated from hypothetical excitation and emission spectra that all exhibited characteristic base lines yet lacked characteristic bands. Figure 1 depicts the coefficients for expanding each row and column of E in terms of the singular vectors. For this graph, each row and column of $\mathbf{E}$ was normalized so that the sum of their elements was unity. The coefficients of these rows and columns are similarly normalized and consequently also sum to unity. With this convention, the coefficients are coordinates of a simplex that has the pure spectra as the extreme points (vertices) and the columns (and rows) of $\mathbf{E}$ as interior points. Only the coefficients of the rows and columns with respect to the second and third singular vectors are shown in Figure 1, but the first can be determined by using the fact that the coefficients sum to unity. The asterisks are the coordinates of the rows and columns of the matrix, the circles are the coordinates of the component spectra, and the crosses are the coordinates of the rows and columns selected by the CFMD procedure as the columns of $\mathbf{A}$ and $\mathbf{B}$. Coordinates that coincide with the face of the component simplex indicate that those columns are weighted averages of just two-component emission spectra. This feature reflects the presence of a characteristic base line in the excitation spectra of the third component. The simplex formed by the most dissimilar rows and columns would have excluded several of the matrix coordinates because $\mathbf{A}$ and $\mathbf{B}$ do not form a feasible solution. Not only does the simplex of the compatible solution found by CFMD, plotted with solid lines, include all the matrix coordinates but its vertices coincide with the coordinates of the component spectra despite the absence of row and column coordinates at these points. This example confirms the superfluity of characteristic bands; they are not necessary because the positions of the cooresponding vertices are implied by coordinates on the opposite face of the component simplex.

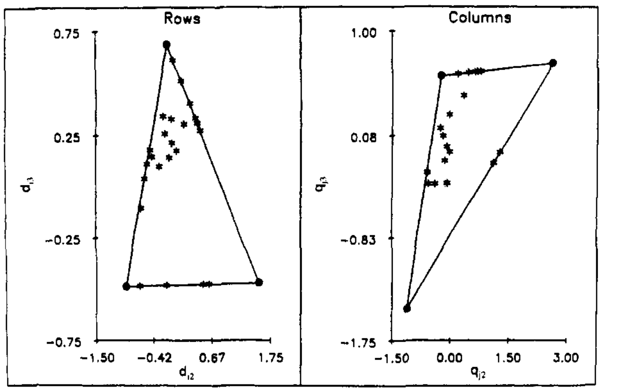

Figure 1. Coordinates of the columns and rows of a synthetic ternary EEM that has no characteristic bands. The coordinates of the columns, $\mathbf{e}_{\cdot j}$(*), are q_{kj} where $\mathbf{e}_{\cdot j} = \Sigma_k \mathbf{r}_{\cdot k} q_{kj}$. The coordinates of the rows, $\mathbf{e}_{i\cdot}$(*), are d_{ik} where $\mathbf{e}_{i\cdot} = \Sigma_k d_{ik} \mathbf{c}^T_{k\cdot}$. The CFMD solution lies at the vertices of the solid simplex (——). The coordinates of the component spectra (O) coincide with the CFMD solution.

Three criteria were used to evaluate the performance of solution simplex minimization (SSM). A series of ideal three-component EEMs that have systematic variations in the component spectra overlap were generated and analyzed. Then, the robustness of the algorithm was tested by analyzing ideal matrices that have unique, compatible factorizations to which various levels of random noise had been added. Finally, the results of these evaluations were compared to the resolution of experimental multicomponent EEMs acquired by using the videofluorometer.

The number of unique combinations of characteristic base lines in the spectra of n components is given by Pascal's triangle. In the case of three components, there are eight (1 + 3 + 3 + 1) hypothetical configurations of spectra overlap for three components. In the first all three-component spectra have characteristic base lines. There are three degenerate configurations consisting of two characteristic base lines. There are also three degenerate configurations exhibiting one characteristic base line. In the last configuration, the three-component spectra are completely overlapped.

Sixty-four synthetic EEMs were generated producing all possible combinations of overlap of excitation and emission spectra. The spectra were designed so that the number of characteristic bands equals the number of characteristic base lines. Synthetic spectra for each of the three components were simulated by using three adjacent Guassian peaks of various heights. The various overlap conditions were simulated by varying the wavelength range of the simulated spectra. The EEMs were generated from the weighted outer products of the component spectra (eq 1). The same set of weights (concentration factor) were used for all 64 EEMs: $\gamma_{11} = 5312$, $\gamma_{22} = 3253$, $\gamma_{33} = 1883$.

The effectiveness of the resolution scheme consisting of BMB, CFMD, and SSM is excellent for the synthetic EEMs. The data matrices comprised of component spectra that each exhibits a characteristic band can be uniquely resolved by using the BMB algorithm. Fifteen EEMs comprised of com-

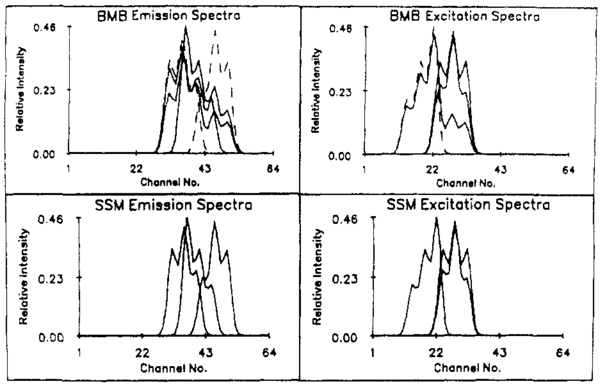

Figure 2. Spectra resolved from a synthetic ternary EEM that has *n* rather than 2*n* characteristic base lines using BMB and SSM (——). The reference spectra (– – –) are included for comparison.

Table II. Scalar Products of Emission Spectra Resolved from Third Component and Ideal Spectrum

S/N	BMB[a]	CFMD[b]
15		0.9894 ± 0.0131
10	0.9968 ± 0.0026	0.9714 ± 0.0300
9	0.9973 ± 0.0004	0.9508 ± 0.0557
8	0.9970 ± 0.0019	0.9697 ± 0.0051
7	0.9897 ± 0.0105	0.9344 ± 0.0194
6	0.9878 ± 0.0048	0.8340 ± 0.0499
5	0.9651 ± 0.0057	
4	0.6935 ± 0.1506	

[a] BMB, best measured basis. [b] CFMD, concentration factor matrix diagnalization.

Table III. Scalar Products of Emission Spectra Resolved from Third Component and Ideal Spectra

S/N	replicates	SSM[a]
50	5	0.9918 ± 0.0183
38	4	0.9963 ± 0.0023
25	4	0.9709 ± 0.0092
18	4	0.8855 ± 0.0550
13	5	0.9366 ± 0.0244

[a] SSM, solution simplex minimization.

ponent spectra which have a total of $[2n(n - 1)]/n$ characteristic base lines are uniquely resolved using CFMD, despite the absence of characteristic bands.

The CFMD algorithm does not produce unique factorizations for the remaining 48 overlap combinations. In these cases, specialized objective functions can be used to take advantage of information imbedded in the data matrix. Figure 2 illustrates the successful resolution of all six-component spectra from one of the 64 typical EEMs that had just three characteristic base lines using SSM to produce the unique solution. The BMB solution, which is the initial estimate to the component spectra used by the optimization routine, and the component spectra are depicted in Figure 2 for reference. The existence of two characteristic base lines in the emission spectra and one in the excitation spectra provided sufficient information to compensate for the very severe overlap of the remaining excitation spectra. The BMB–CFMD–SSM sequence was able to successfully resolve, i.e. factor the EEM into the component spectra from which it was generated, 24 of the remaining 48 EEMs.

To compare the effect of noise on the performance of the optimization methods, various levels of random noise were added to three ideal matrices, which were successfully resolved by BMB alone, BMB with CFMD, and the BMB–CFMD–SSM sequence, respectively. The random noise was calculated by using the expression

$$n_{ij} = \mathrm{RND}(e_{ij}^{1/2} + b) \qquad (15)$$

where the random variable RND was equally distributed on the interval $[-1,1]$, e_{ij} is the corresponding element of the ideal EEM, and b approximates the mean of the background. The first term of this expression approximates photon statistical noise and the second the dark current. The mean background value was subtracted from the degraded matrix before analysis. Each analysis was performed in triplicate unless otherwise indicated.

Table II lists the maximum signal-to-noise ratios (S/N) for the third component of the synthetic matrices used to evaluate the resolution algorithm and the average of the scalar products of the emission spectrum of the third component resolved from each noisy matrix by BMB and BMB–CFMD with the ideal spectrum. The third spectrum was used because it has the smallest S/N. An identifiable resolved spectrum is indicated by a scalar product equal to 0.95. This threshold was determined heuristically. The data in Table II indicate that BMB will extract recognizable spectra from matrices that it could resolve in the absence of noise as long as the S/N is above 5, and the BMB–CFMD is similarly useful for data that has S/N above 8. Table III lists the average scalar products of ideal spectra with the emission spectra of the third component resolved by SSM from EEMs generated by adding noise to the ideal data used to produce Figure 2. The data indicate that the effects of noise on the SSM are more severe than on the CFMD. Not only is the S/N threshold higher for this algorithm, but the variability in the results is much higher at all noise levels. In fact, the optimization routine frequently converged on nonoptimal solutions. Eight replicate analyses were required to produce the data used to calculate the averages in Table III. The cause of this phenomenon has not been determined but may stem from limitations in the routines used to determine the best search vector. Another source of difficulty could be discontinuities that are introduced to the constraints with the addition of high-frequency noise.

To evaluate the validity of our models, the EEMs of several mixtures of polynuclear aromatic hydrocarbons (PNAs) were acquired by using the videofluorometer and analyzed by using both optimization-based algorithms. The average S/N of the experimental matrices was estimated by using the expression

$$S/N = \frac{\sum_{k=1}^{n}\psi_{kk}}{\sum_{i=n+1}^{m}\psi_{ii}} \qquad (16)$$

where $m = \min(q,p)$.

The isometric projection of a mixture of anthracene, 9,10-dimethylanthracene, perylene, and 2,3-benzanthracene is shown in Figure 3. The resolution of the component spectra from this matrix using CFMD is illustrated by the vectors drawn with solid lines. The standard spectra of the components are included in Figure 3 for comparison and plotted with dashed lines. The estimated S/N for this matrix was 35 and the spectra of all four components are successfully resolved, illustrating that the diagonalization procedure is applicable to experimental data. Figure 4 illustrates the resolution of the component spectra from a mixture of 9,10-dimethylanthracene, perylene, fluoranthene, and 2,3-benzanthracene. The S/N of this matrix was approximately 25. In this case, the spectra of all four components are also successfully resolved, but a two-dimensional Savitsky–Golay filter was applied to the least-squares estimate of the matrix before the application of the resolution algorithm. The constraints also were relaxed by permitting small negative values (1%) in the resolved spectra.

CONCLUSIONS

Several extensions or generalizations of Lawton and Sylvestre's self modeling curve resolution have been previously

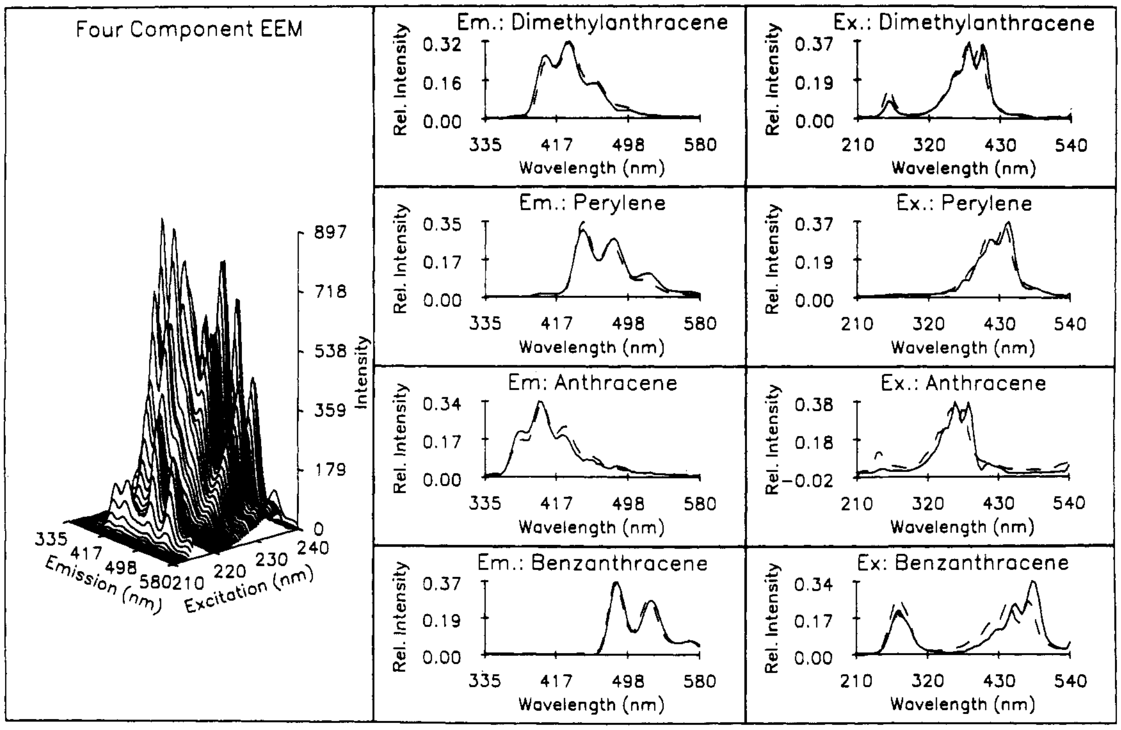

Figure 3. EEM and spectra resolved by CFMD (—) from a mixture of 7.64 × 10^{-6} M 2,3-benzanthracene, 5.56M × 10^{-7} M perylene, 2.09 × 10^{-6} M 9,10-dimethylanthracene, and 6.47 × 10^{-6} M anthracene. Reference spectra are indicated by dashed lines (- -).

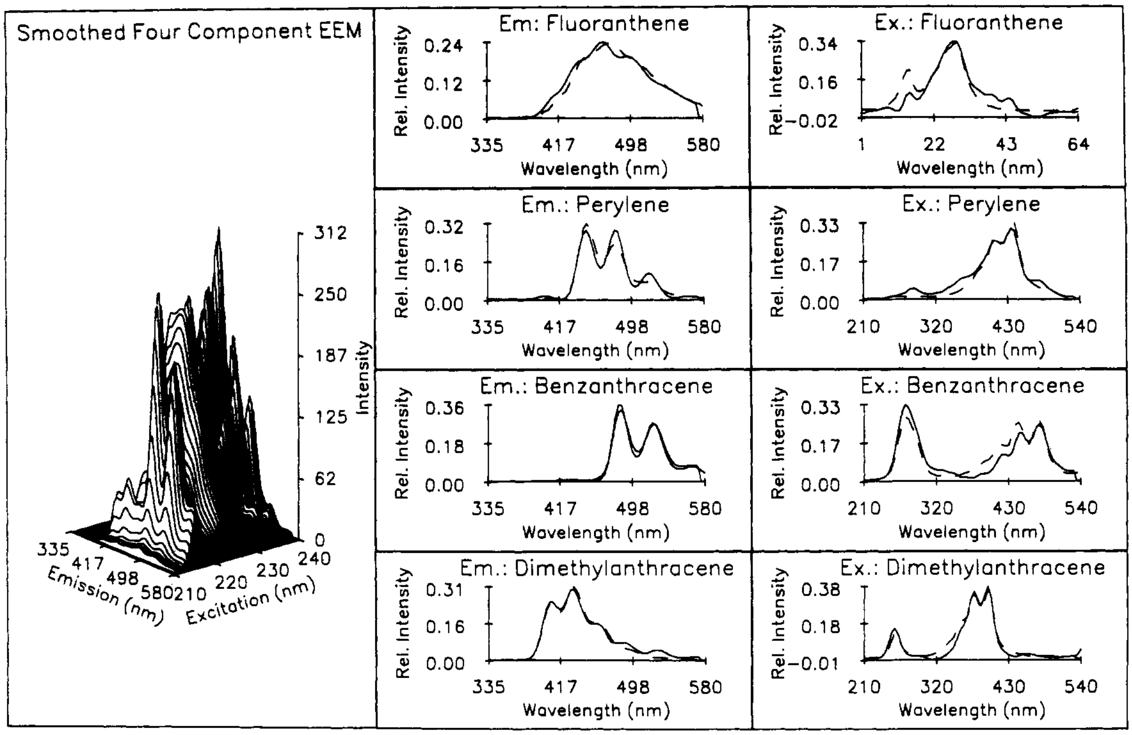

Figure 4. EEM and spectra resolved by CFMD (—) from a mixture of 2.30 × 10^{-6} M 2,3-benzanthracene, 1.67 × 10^{-8} M perylene, 6.29 × 10^{-7} M 9,10-dimethylanthracene, and 2.74 × 10^{-6} M fluoranthene. Reference spectra are indicated by dashed lines (- -).

described (*4, 9, 21, 22, 26*). The algorithm presented here differs in several respects. First, the CFMD uses the complementarity of the component spectra, as well as their nonnegativity and convexity, to estimate the component spectra. Most importantly, the flexibility of the nonlinear optimization approach permits the analyst to evaluate a wide variety of solutions as well as choose which features are incorporated into the solution. The scheme described in this report also allows the analyst to insert related data analysis tasks into the curve resolution scheme easily. For example, optimizing the volume function without the compatibility constraints is theoretically a nongraphical alternative to evaluating the range of feasible solutions. The spectra that lie on the inner boundaries, i.e. those generated by the restrictions on the sign of the coefficients of the rows and columns expanded in terms of the component spectra, can be found by minimizing the volume, while those that lie on the nonnegativity boundaries can be found by maximizing the volume. Similarly, curve resolution can be coupled to calibration of known components by inserting the coordinates of standards into the CFMD and SSM routines (*27*).

This resolution scheme also differed from earlier algorithms in the use of the SVD. The SVD was used in this scheme as an aid to rank estimation and calculation. All of the steps, can be performed, albeit more slowly, on **E** directly. The basis vectors chosen by the BMB method are used to estimate the spectra, so that matrices which are very large or intractable to the SVD may be analyzed by using this approach.

In the course of the development of this algorithm, it was determined that it is the existence of a characteristic base line rather than a characteristic response which governs spectral resolution. This observation is the crux of the solution simplex methodology. These evaluations also show that the optimization-based methods are more sensitive to noise than the target factor analysis method (BMB) presented here. Evaluations using synthetic data indicate that S/N thresholds of at least 8 are required for the resolution of recognizable spectra from nonideal EEMs using CFMD and those of at least 18 are needed for SSM. Additionally, the solution simplex minimization is more prone to nonconvergence than the CFMD.

Evaluations with experimental EEMs acquired by the videofluorometer confirm the conclusions drawn from the synthetic data for the diagonalization of the concentration factor matrix. The evaluations of synthetic data indicate that resolution of sets of spectra that do not have a $2n$ characteristic base line may be possible but will require higher S/N.

Finally, it has been stated that this scheme is directly applicable to the results of bilinear multiparametric methods. It should also be noted that because concentration is nonnegative and convex, this method can be applied to matrices formed from collections of sample responses to one parameter.

ACKNOWLEDGMENT

We thank Dr. Henry Wolkowisc for providing us with a copy of NLPQL and the code for the nonlinear least-squares subproblem.

LITERATURE CITED

(1) Harman, H. H. *Modern Factor Analysis*; University of Chicago Press: Chicago and London, 1976.
(2) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; John Wiley: New York, 1980.
(3) Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617.
(4) Ohta, N. *Anal. Chem.* **1973**, *45*, 533.
(5) Knorr, F. J.; Futrell, J. H. *Anal. Chem.* **1979**, *51*, 1236.
(6) Malinowski, E. R. *Anal. Chim. Acta* **1982**, *134*, 129.
(7) Gemperline, P. J. *Anal. Chem.* **1986**, *58*, 2656.
(8) Vandeginste, B. G. M.; Derks, W.; Kateman, G. *Anal. Chim. Acta* **1985**, *173*, 253.
(9) Sasaki, K.; Kawata, S.; Minami, S. *Appl. Opt.* **1984**, *23*, 1955.
(10) Wilson, B. E.; Lindberg, W.; Kowalski, B. R. *J. Am. Chem. Soc.* **1989**, *111*, 3797-3804.
(11) Johnson, D. W.; Callis, J. B.; Christian, G. D. *Anal. Chem.* **1977**, *49*, 747A.
(12) Weber, G. **1961**, *Nature 190*, 27.
(13) Searle, s. R. *Matrix Algebra Useful for Statistics*; John Wiley: New York, 1983.
(14) Golub, G. H.; VanLoan, C. F. *Matrix Computations*; Johns Hopkins University Press: Baltimore, MD, 1983.
(15) Shrager, R. I.; Hendley R. W. *Anal. Chem.* **1982**, *54*, 1147.
(16) Wold, S. *Technometrics* **1978**, *20*, 397.
(17) Schittowski, K. *Design, Implementation and Test of a Nonlinear Programming Algorithm*; Springer-Verlag: New York, in press.
(18) Scales, L. E. *An Introduction to Nonlinear Optimization*; Springer-Verlag: New York, 1985.
(19) Gill, P. E.; Murray, W.; Wright, M. H. *Practical Optimization*; Academic Press: London and New York, 1981.
(20) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308.
(21) Borgen, O. S.; Kowalski, B. R. *Anal. Chim. Acta* **1985**, *174*, 1.
(22) Borgen, O. S.; Davidsen, N.; Mingyang, Z.; Oyen, O. *Mikrochim. Acta* **1986**, *2*, 63.
(23) Warner, I. M.; Fogarty, M. P.; Shelly, D. C. *Anal. Chim. Acta* **1979**, *109*, 361.
(24) Lay, S. R. *Covex Sets and Their Applications*; John Wiley: New York, 1982.
(25) Bronsted *An Introduction to Convex Polytopes*; Springer-Verlag: New York, 1983.
(26) Meister, A. *Anal. Chim. Acta* **1984**, *161*, 149.
(27) Nelson, G.; Neal, S. L.; Warner, I. M. *Spectroscopy* **1988**, *3*, 24.

RECEIVED for review July 14, 1989. Revised manuscript received December 19, 1989. Accepted December 29, 1989. This work was supported in part by grants from the National Science Foundation (CHE-8609372) and the Office of Naval Research. I. M. Warner acknowledges support from an NSF Presidential Young Investigator Award (CHE-8351675). S. L. Neal acknowledges the support of a graduate fellowship from the National Organization of Black Chemists and Chemical Engineers sponsored by the Procter and Gamble Company.

Locally Weighted Regression and Scatter Correction for Near-Infrared Reflectance Data

Tormod Naes* and Tomas Isaksson

MATFORSK, Oslovegen 1, 1430 Ås, Norway

Bruce Kowalski

Laboratory of Chemometrics, University of Washington, Seattle, Washington 98195

This paper investigates the effect of multiplicative scatter correction (MSC) and nonlinear regression based on the first two and three principal components for near-infrared reflectance (NIR) spectroscopy data. The focus will be on linearity/nonlinearity as well as treatment of outliers. The main contribution of the paper is the presentation of and testing of a calibration method based on classification and local least-squares regression. The theory is illustrated by three examples from NIR analysis. The local linear calibration outperformed traditional methods in two of the examples.

INTRODUCTION

Most calibration techniques used in near-infrared reflectance (NIR) spectroscopy are linear in the sense that they produce models that are linear functions of the spectral reflectances or absorbances (*1, 2*), i.e. log (1/reflectance). Such methods, for instance principal component regression (PCR) (*3*), partial least-squares (PLS) regression (*4*), Fourier regression (*5*), and different versions of stepwise multiple linear regression (SMLR), have proved useful in practice.

In many studies, other functions or corrections of the spectral reflectances combined with linear calibrations are tested (*5–10*). The idea behind these approaches is to improve linearity, to optimize predictions, and to eliminate or reduce

* Author to whom correspondence should be sent.

0003-2700/90/0362-0664$02.50/0 © 1990 American Chemical Society