(14) Berthod, A. *Analusis* **1990**, *18*, 352-357.
(15) Tsai, R. S.; El Tayar, N.; Testa, B.; Ito, Y. *J. Chromatogr.* **1991**, *538*, 119-123.
(16) El Tayar, N.; Martson, A.; Bechalany, A.; Hostettmann, K.; Testa, B. *J. Chromatogr.* **1989**, *469*, 91-99.
(17) Guiochon, G. In *HPLC: Advances and Perspectives*; Horwath, Cs., Ed.; Academic Press; New York, 1980; Vol. 2.
(18) Dorsey, J. G. Chromatographic Quantitative Structure-Retention Relationships. Presented at the Analytical Chemistry Symposium, 202nd National Meeting of the American Chemical Society, New York, Aug 25-30, 1991.
(19) Melander, W. R.; Horwath, Cs. *Chromatographia* **1982**, *15*, 86-90.
(20) Krafft, M. P.; Jeannaux, F.; Le Blanc, M.; Riess, J. G.; Berthod, A. *Anal. Chem.* **1988**, *60*, 1969-1972.
(21) Ito, Y.; Oka, H.; Slemp, J. L. *J. Chromatog.* **1989**, *475*, 219-227.
(22) Mandava, N. B.; Ito, Y.; Ruth, J. M. *J. Liq. Chromatogr.* **1985**, *8*, 2221-2238.
(23) Berthod, A.; Armstrong, D. W. *J. Liq. Chromatogr.* **1988**, *11*, 1187-1204.

RECEIVED for review May 13, 1991. Accepted August 8, 1991. This work was supported by the Centre National de la Recherche Scientifique, UA 435, J. M. Mermet. Additional funding was provided by TOTAL Raffinage Distribution, Grant No. 900918E29, J. Goupy, Levallois-Perret, which is gratefully acknowledged.

Real-Time Principal Component Analysis Using Parallel Kalman Filter Networks for Peak Purity Analysis

Stephen J. Vanslyke and Peter D. Wentzell*

Trace Analysis Research Centre, Department of Chemistry, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J3

A new approach for performing principal component analysis (PCA) during data acquisition is described. The method is based on a network of multilinear models which are fit to data with the discrete Kalman filter. Application to absorbance matrices such as those obtained in chromatography with multiwavelength detection is considered. Multivariate data projected into two- and three-dimensional subspaces are fit with linear and planar models, respectively. Model deviations, detected using principles of adaptive Kalman filtering, are used to elucidate the rank of the data set. Principal component eigenvectors are then constructed from the individual models. Results of this initial work using simulated and experimental data demonstrate that extraction of the first two principal components is readily accomplished and eigenvectors obtained are in good agreement with traditional batch PCA results. Extension to more principal components should be possible although it will increase the number and complexity of models. Advantages of the new algorithm include its recursive implementation, parallel structure, and ability to indicate model errors as a function of time. The procedure should prove particularly useful for self-modeling curve resolution applications in chromatography.

The increasing availability of analytical instruments capable of producing large amounts of multidimensional data has made the use of multivariate data analysis methods more commonplace. In the case of second-order bilinear instruments such as GC/MS or HPLC coupled to photodiode array UV-visible detection, one data analysis tool that is frequently used is principal component analysis (PCA), a type of factor analysis (*1-4*). This method is at the core of most algorithms designed for multicomponent curve resolution in chromatography. The problem of mathematically resolving overlapped chromatographic profiles is an old and difficult one in chemical analysis. It consists of essentially three steps: detection of peak overlap, identification of individual analytes, and quantitation of components. A variety of methods have been proposed for the first of these problems, the simplest involving the monitoring of response ratios at two detector settings (e.g. absorbances at two wavelengths) (*5*). This approach, while straightforward, suffers from a number of disadvantages, including a susceptibility to a sloping background and the requirement of a preexisting knowledge of wavelengths to be used. Its biggest drawback, however, is that it does not identify the number and nature of coeluting analytes or provide quantitative results.

In recent years, a number of algorithms have been described for curve resolution in HPLC with UV-visible diode array detection (*3, 4, 6-13*). Most of these are based on the original method of Lawton and Sylvestre (*6*), which uses PCA coupled with nonnegativity constraints for the spectral and concentration domains. This approach requires no assumptions to be made regarding the number of components, or the shapes of component spectra or elution profiles. Additional information is often used to provide more exact solutions than can be provided by the algorithm alone, however. This self-modeling curve resolution approach has been very successful for two-component elution profiles, and commercial software is available. Numerous extensions to three or more components have been suggested (*9, 10, 13*), but effective use generally requires the availability of additional information.

One of the drawbacks of chromatographic curve resolution based on PCA is that calculations are generally performed after all of the data have been acquired. Chromatographic regions of interest must first be selected manually or automatically and subjected to PCA to determine the number of components. Thorough analysis requires interrogation of all chromatographic peaks. A simple peak purity test can be performed to screen particular areas of interest, but this suffers from the problems previously noted. A useful alternative would be the ability to carry out PCA recursively, i.e. while the data are being acquired. This would allow the determination of the number of coeluting components in real time and also act as a preprocessing step for self-modeling curve resolution. The development of a real-time PCA algorithm was the objective of this work.

The possibility of conducting recursive principal components analysis is made difficult by the fact that the usual PCA procedure is already iterative. To be capable of real-time implementation, a recursive procedure needs to maintain a static cycle time for each new data point obtained. One so-

0003-2700/91/0363-2512$02.50/0

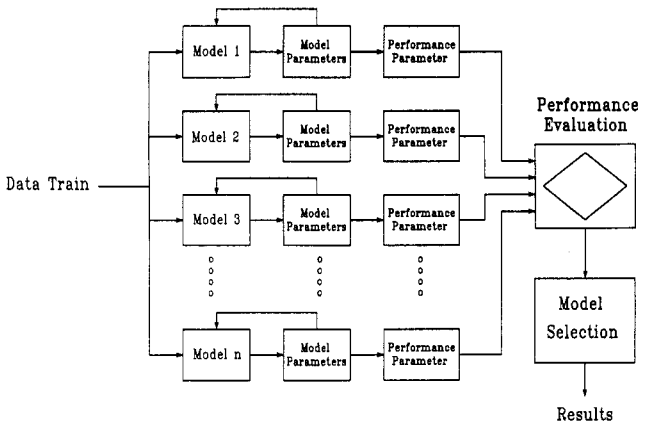

Figure 1. Strategy for implementation of parallel Kalman filter networks.

lution to the problem is to use the principles of adaptive Kalman filtering. The adaptive Kalman filter can be used as a recursive linear least-squares estimation procedure that has some built-in features to compensate for model errors. The strength of the adaptive Kalman filter is that it provides diagnostic information on model validity. If a parallel network of filters is employed, each with a different model, this information can be used to select the best model to fit experimental observations. This is the basis of the work described here, in which initial results for a recursive PCA method are presented. The use of parallel and block sequential Kalman filters for processing speech signals has been described in the engineering literature (*14*), but this is the first application of this type we have encountered. Currently, the recursive PCA procedure is limited to two-component models, but extension to higher dimensionality should be possible. Computer-simulated chromatographic profiles and experimental data from coeluting dyes are used to demonstrate the capabilities of the algorithm.

THEORY

Kalman Filter Algorithm. Details of the Kalman filter have appeared elsewhere (*15–18*) and only a brief description will be given here. Most applications of the Kalman filter to chemical measurements have implemented the algorithm as a means for performing recursive ordinary least squares. This generally simplifies mathematical and computational aspects of the algorithm. The Kalman filter requires the definition of a linear model describing how the observations (measurements) change with one or more independent variables. The adjustable parameters of this model are known as state variables, or collectively as the state vector, and estimates of these are adjusted after each measurement, ultimately converging on the final estimate which can be represented as a point in state space. State variables may be dynamic or static. Most chemical applications have employed static models, which simplifies the algorithm somewhat since the identity matrix can be used to describe how the state vector propagates between measurements. The Kalman filter algorithm defines how estimates of the state vector and its associated covariance matrix evolve with measurements and includes compensation for four factors: (1) the current parameter estimates, (2) the difference between observed and predicted measurements, (3) the estimated error in the parameters (covariance matrix), and (4) the estimated error in the measurement. Certain restrictions regarding the modeled system apply (e.g. model linearity, model validity, Gaussian white noise), but under these conditions the Kalman filter will yield estimates which are optimal in the least-squares sense.

The use of parallel Kalman filter networks has been previously described for kinetic methods of analysis (*19*). The

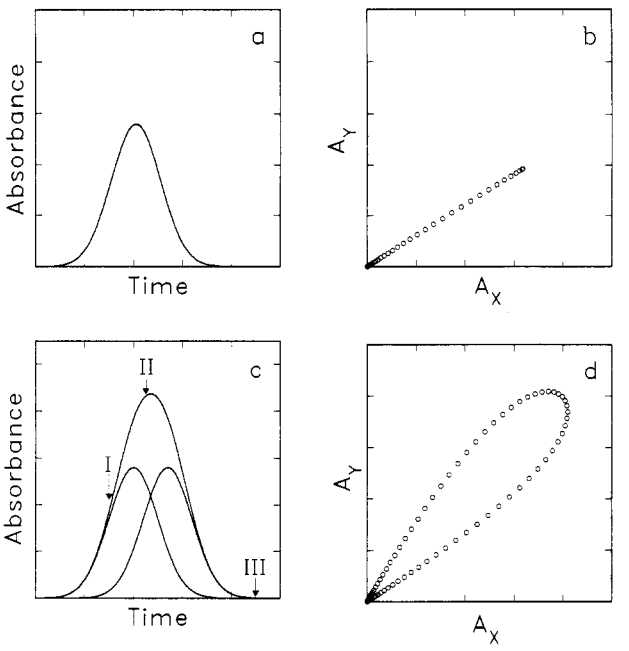

Figure 2. Chromatographic elution profiles and their projections into A^2-space: (a) and (b) one-component profile and its projection; (c) and (d) two-component profile and its projection.

general scheme is illustrated in Figure 1. The incoming data sequence is applied to the inputs of a number of Kalman filters, each with a different model. These models may be used to handle data from nonlinear systems by introducing small variations in nonlinear parameters between adjacent models (quasi-continuous case), or they may represent distinct alternatives (discrete case). The recursive PCA application described here employs the latter form. In either case, the application of each filter provides new estimates of the state parameters for the corresponding model. These parameters are used to evaluate the performance of each model in terms of its consistency with actual observations. A useful measure of the model performance is the innovations sequence which has been employed for adaptive Kalman filter algorithms (*20*, *21*). The innovation is defined simply as the difference between the actual and predicted measurement (for multiple measurements in a single cycle, the innovation is a vector). The innovation differs from the residual normally used in modeling problems in that it is calculated after each measurement on the basis of current model parameters, whereas residuals are calculated in a batch procedure. The innovations sequence is particularly useful where model deviations are localized, since it indicates regions of model validity.

The application of the Kalman filter to chromatographic peak resolution has been previously described by Brown and co-workers (*22*, *23*) and by Hayashi et al. (*24*), but these methods require a knowledge of individual component spectra or elution profiles and differ from the present work which seeks to function in the absence of this information. The principles of the recursive PCA approach are best illustrated with an example. Figure 2 shows synthetic chromatograms obtained for the elution of one- and two-component mixtures. For simplicity, Gaussian, noise-free peaks have been assumed, but this is not a requirement of the algorithm. It is also assumed that the two components are not completely overlapped and have sufficiently different spectral profiles. These are requirements of most curve resolution methods. Shown adjacent to the two representative chromatograms in Figure 2 are plots of absorbance measurements (A_x and A_y) at two wavelengths for each sampled point. The wavelengths selected are the absorbance maxima of the two hypothetical components. Plots such as these will be referred to as A^2 plots (also

known as biplots) and illustrate the principles of both the absorbance ratio and PCA methods for peak purity assessment. It is clear that the one-component mixture gives a constant A_y/A_x, while the two-component case does not. Methods based on PCA extend this principle further by recognizing that in an n-dimensional absorbance space (A^n), the one-component case will always fit a linear model within experimental error. Likewise, a two-component chromatogram will fit a planar model in A^n space. Therefore, the minimum number of components in an overlapped region can be determined from the intrinsic dimensionality (i.e. rank) of multivariate absorbance data. Obviously, the PCA method is more general than simple approaches like absorbance ratio.

Two notable drawbacks of usual methods based on PCA are the need for batch processing of the data and difficulty of accurately determining the dimensionality of the data. The former problem has already been mentioned. The latter results from the difficulty of distinguishing residual eigenvectors arising from experimental error from those that represent true components. This problem is futher complicated by experimental realities such as a sloping background that may appear as additional chromatographic components. Such features, while of interest in quantitation, can be misleading in the detection of peak overlap. Part of the difficulty in the determination of the true number of chromatographic components is that rank is normally assessed on the basis of one or more scalar quantities (*1*, *7*, *8*) that ignore information available in the temporal structure of the data. Clearly, deviations from an n-dimensional model due to the presence of additional components or a sloping background should exhibit characteristic patterns if the evolution of the model is examined. This behavior can be detected if PCA is performed recursively.

The principle of operation of recursive PCA is that an n-dimensional data set projected onto an $n + 1$ dimensional space (or higher) should always give a fit to a multilinear function which is satisfactory within experimental error. Thus, the one-component data in Figure 2 will give a satisfactory fit to a straight line at any two wavelengths, but the fit for the two-component data should be unsatisfactory for at least certain pairs of wavelengths. Both data sets should give a good fit to a planar model in any A^3 space, but a three-component data set should not. This strategy can be extended to higher dimensions, although visualization becomes more difficult. On this basis, the intrinsic dimensionality of the data set can be deduced by selecting a number of wavelengths and fitting data in lower subspaces comprised of various wavelength combinations. A one-component model should exhibit comparable residuals for both linear and planar models, but a two-component data set should exhibit excessive residuals for the linear model at certain wavelength combinations. Furthermore, the resulting models will serve as a means to construct good approximations to the principal component vectors. For example, the linear models should be projections of the first eigenvector, and the planar models should all contain the planes defined by the first two eigenvectors. The correspondence between the true eigenvectors and the reconstructed vectors may not be exact, since the multilinear least-squares models are generally constructed assuming no errors in the independent variable (*25*), but the correspondence should be close under the right conditions.

The use of multiple models of lower dimensionality offers no particular advantages over traditional PCA except when used in conjunction with the Kalman filter. The Kalman filter can be used to generate fits to linear and planar models recursively. This increases the potential for real-time implementation of the algorithm and observation of model evolution. Before the algorithm is initiated, n wavelengths at

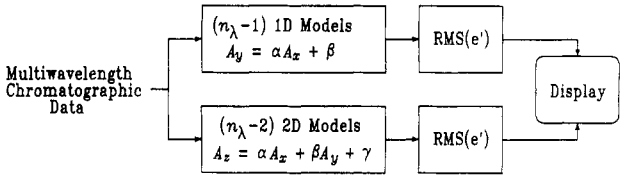

Figure 3. Parallel Kalman filter network for recursive PCA application.

conveniently spaced intervals are selected. For the one-component model, the absorbance at one wavelength (designated A_x) is selected as the independent variable and a series of $n - 1$ models of the form

$$\hat{A}_{ij} = \alpha_i A_{xj} + \beta_i \tag{1}$$

are used for the Kalman filter. In this equation, $\hat{A}_{ij}$ represents the predicted absorbance at wavelength i for measurement j, A_{xj} is the measured absorbance at the wavelength chosen as for the independent variable, and α_i and β_i are parameters associated with the one-dimensional model. The parameter estimates evolve as each measurement is processed by the Kalman filter. The model may be limited to the trivial case of single parameter (α_i) if a zero intercept is assumed, but this will not always be the case. These models are used as $n - 1$ elements of the parallel filter network. Likewise, $n - 2$ two-dimensional (planar) models of the form

$$\hat{A}_{ij} = \alpha_i A_{xj} + \beta_i A_{yj} + \gamma_i \tag{2}$$

are also used. In this case, absorbances at two wavelengths (arbitrarily designated as A_x and A_y) are needed as independent variables and three parameters are estimated. Models of higher dimensionality are also possible, but were not employed in this initial work. The $2n - 3$ models described are sufficient to indicate one, two, or more than two coeluting components.

The strategy for implementation of the models in the parallel Kalman filter network is shown in Figure 3. As each set of absorbance values for a single chromatographic point is received, a set of innovations corresponding to the $2n - 3$ models is calculated using predicted measurements according to

$$e_{ij} = \hat{A}_{ij} - A_{ij} \tag{3}$$

where e_{ij} is the innovation for measurement j at wavelength i, and $\hat{A}_{ij}$ and A_{ij} are the predicted and measured absorbances, respectively. In each cycle, there are $n - 1$ innovations calculated for the one-dimensional models and $n - 2$ for the two-dimensional models. The magnitude of the innovations should be close to the measurement noise level if the correct model is used, but significant deviations should be observed otherwise. Thus, the absolute value of the innovations sequence can be used to indicate when the dimensionality of the data does not match the dimensionality of the model. Not all of the innovations will provide this information, however, since the wavelengths used in a particular model may not be appropriate for observing model variations (e.g. if there is no absorbance for the dependent variable). One way around this problem is to examine the maximum absolute innovation for each set of models, but this will be sensitive to outliers in the data. An alternative approach is to calculate the root mean square (rms) value of innovations for each set of models:

$$\mathrm{rms}(e_k) = \sqrt{\frac{\sum e_i^2}{m_k}} \tag{4}$$

where the summation is over the number of models of dimension k, m_k ($m_k = n - k$). The rms values can be plotted in real time along with the chromatogram, and the sequence should remain fairly flat as long as the dimensionality of the

data is a subset of the model space.

A number of practical problems need to be addressed in implementing the above strategy. The first is the selection of wavelengths which will act as independent variables in the models. The selection is not entirely arbitrary since the independent variable will exhibit a certain amount of uncertainty along its axis. In the case of one-dimensional model, this means that if the wavelength selected for the x variable shows no absorbance for any component, the least-squares fit will result in a nearly vertical line ($\alpha \rightarrow \infty$) for those models whose dependent variable is non-zero. Although this may result in a valid least-squares fit, the innovations (measured in the y direction) will be excessively large, leading an erroneous assessment of model validity through eq 4. One way to minimize this problem is to ensure that the wavelengths chosen as the independent variables in an n-dimensional model exhibit significant absorbance for some portion of the data. In practice, the wavelengths exhibiting the highest absorbance when the peak is first detected are used. This does not eliminate the problem entirely since the magnitude of the innovations will still increase with the slope of the line. In the two-dimensional case, the problem is compounded by the likelihood of high correlation between the independent variables. A more robust solution uses innovations measured orthogonally from the model rather than vertically. For the one-dimensional model, it can be shown that the orthogonal innovation is given by

$$e_{ij}' = \frac{-\alpha_i A_{xj} + A_{ij} - \beta_i}{\sqrt{\alpha_i^2 + 1}} \tag{5}$$

where the prime denotes an orthogonal innovation. Likewise for a two-dimensional model

$$e_{ij}' = \frac{-\alpha_i A_{xj} - \beta_i A_{yj} + A_{ij} - \gamma_i}{\sqrt{\alpha_i^2 + \beta_i^2 + 1}} \tag{6}$$

Extension to higher dimensions is straightforward. These modified innovations were used in place of the usual values for model evaluation (eq 4) since they should more accurately reflect the true model errors. The modified values were not used in the Kalman filter algorithm, however, since they underestimate the true innovations and lead to errors in the estimation of model parameters.

Another practical aspect of the implementation of the recursive PCA algorithm concerns its activation. In practice, the Kalman filter network is not activated until a peak is detected, although in principle the baseline region could be used if appropriate independent variables could be selected in advance. Once activated, the rms innovations for each model (eq 4) may exhibit high values for the first two or three points as the model converges on reasonable parameter estimates. Display of these points could be suppressed, but this was not done for the results presented here and is generally not necessary.

Relationship to Principal Component Analysis. In view of the importance of the connection drawn between PCA and the Kalman filter algorithm developed here, a more detailed discussion of this relationship is warranted. There are obvious computational differences between the traditional batch processing method for performing PCA and the multilinear approach presented here. This means that the resultant vectors are not necessarily identical, but the differences should be small enough to be inconsequential. It is known that the Kalman filter will provide model equations that are virtually identical to the traditional least-squares method as long as the diagonal elements of the covariance matrix are initially set to large values relative to estimated measurement error (*26*, *27*). In this work, a ratio of $>10^{10}$ was normally used for the diagonal elements in order to achieve the least-squares solution (off-diagonal elements were initially set to zero).

The relationships between the PCA eigenvectors and the Kalman filter results are as follows. For the set of one-dimensional models given by eq 1, the vector resulting from the combination of all models into A^n space corresponds to the first eigenvector obtained by traditional PCA methods if the absorbance data were mean-centered at each wavelength. As this eliminates residual eigenvectors resulting from an offset at particular wavelengths, it is preferred for peak purity analysis. If mean-centering of the absorbance vectors is not carried out prior to batch PCA, the first eigenvector will correspond to the vector generated by the models in eq 1 of the β_i's are forced to zero. In either case, the first eigenvector from the Kalman filter ($\mathbf{E}_1$) will be given by

$$\mathbf{E}_1 = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{n-1} \quad 1] \Big/ \sqrt{1 + \sum^{n-1} \alpha_i^2} \tag{7}$$

The first $n-1$ components of the vector are the projections of the $n-1$ dependent variables of the models, while the last represents the wavelength selected as the independent variable. The denominator simply serves to normalize the vector. To find the second eigenvector, both the one- and two-component models are required. This is because the two-component model only defines the plane containing the first two eigenvectors and not the vectors themselves. Generally, if it is known that two-components are present, a knowledge of the plane of the first two eigenvectors is sufficient to perform self-modeling curve resolution. Nevertheless one method of obtaining the actual eigenvectors is presented here. As in the case of the one-component models, omission of the offset parameter (in this case γ_i) will lead to the PCA result for data which are not mean-centered for absorbance. The plane containing the first two eigenvectors will also contain the vectors $\mathbf{V}_1$ and $\mathbf{V}_2$. The $(n-1)$th component of the vectors

$$\mathbf{V}_1 = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{n-2} \quad 1 \quad 0]$$

$$\mathbf{V}_2 = [\beta_1 \quad \beta_2 \quad \dots \quad \beta_{n-2} \quad 0 \quad 1]$$

corresponds to the first independent variable and nth component to the second. These vectors correspond to the intersection of the planar model with the $(n-1)$ dimensional subspaces along the axes of A_x and A_y. After normalization of $\mathbf{V}_1$ to $\mathbf{N}_1$, a vector $\mathbf{N}_2$ which is orthonormal to $\mathbf{N}_1$ is determined by projection of $\mathbf{V}_2$ onto $\mathbf{N}_1$, subtraction, and normalization. The vectors $\mathbf{N}_1$ and $\mathbf{N}_2$ are just one set of orthonormal vectors which can be used to define the plane. Ideally, the first eigenvector obtained from the one-dimensional models ($\mathbf{E}_1$) will lie on this plane, but in practice there may be a slight elevation due to minor computational differences for the two models. To ensure the integrity of the two-component plane, the first eigenvector is recalculated as its projection onto the plane defined by the orthonormal vectors ($\mathbf{E}_1'$). In all cases that we have checked, the difference

$$\mathbf{N}_1 = \mathbf{V}_1/|\mathbf{V}_1| \tag{8}$$

$$\mathbf{P} = \mathbf{V}_2 - (\mathbf{V}_1 \cdot \mathbf{V}_2)\mathbf{V}_1 \tag{9}$$

$$\mathbf{N}_2 = \mathbf{P}/|\mathbf{P}| \tag{10}$$

$$\mathbf{E}_1' = (\mathbf{E}_1 \cdot \mathbf{N}_1)\mathbf{N}_1 + (\mathbf{E}_1 \cdot \mathbf{N}_2)\mathbf{N}_2 \tag{11}$$

between $\mathbf{E}_1$ and $\mathbf{E}_1'$ has been insignificant, but calculation of the projection is more robust in cases where a two-dimensional model is employed. Calculation of the second eigenvector, $\mathbf{E}_2$, in the two-dimensional subspace is now trivial. Extension

$$\mathbf{E}_2 = -(\mathbf{E}_1 \cdot \mathbf{N}_2)\mathbf{N}_1 + (\mathbf{E}_1 \cdot \mathbf{N}_1)\mathbf{N}_2 \tag{12}$$

of these principles to systems of higher dimensionality should

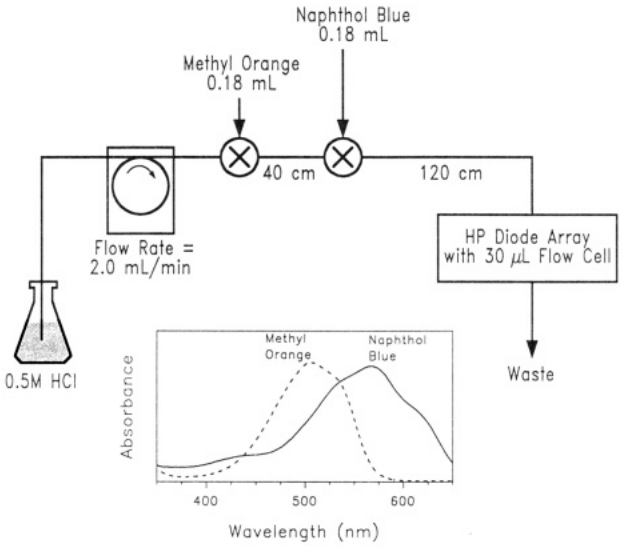

Figure 4. Merging zones continuous-flow apparatus for studies of dye coelution with spectra of dyes inset.

be straightforward, but will not be considered here.

EXPERIMENTAL SECTION

Simulation Studies. Generation of absorbance vs wavelength vs time data to test the Kalman filter algorithm was carried out with a program that allowed a variety of conditions to be simulated. For simplicity, Gaussian shapes were assumed for spectral and concentration profiles. Normally distributed random values were added to simulate measurement noise. Further details of conditions used accompany results presented in the Results and Discussion.

Dye Coelution. Experimental results for coeluting components were obtained using a continuous-flow system in a merging zones configuration, as shown in Figure 4. Dye solutions employed for the results presented here were 0.312 μM methyl orange (acid orange 52, color index 13025; Fisher, Fair Lawn, NJ) and 1.17 μM naphthol blue (Meldola's blue, basic blue 6, c.i. 51175; Pfaltz & Bauer, Waterbury, CT), both prepared in 0.5 M HCl. These concentrations were found to give a noise level suitable for testing the algorithm. The samples were injected simultaneously into the stream using two six-port two-way valves (Rheodyne Model 5020, Cotati, CA) with 180 μL sample loops. The injected samples were transported to the detector through 0.8 mm i.d. PTFE tubing by an 8-roller Ismatec SA peristaltic pump (Cole-Parmer, Chicago, IL). A Hewlett-Packard Model 8452A photodiode array spectrometer (Hewlett-Packard, Palo Alto, CA) with a 30-μL flow cell (Hellma Cells, Jamaica, NY) was used to acquire spectra at 1-s intervals for about 100 s after injection.

Computational Aspects. All calculations were carried out on a 16-MHz IBM PC/AT compatible computer with a math coprocessor. Programs were written in Microsoft QuickBASIC (Microsoft Corp., Redmond, WA). Implementation of the Kalman filter employed the standard algorithm (*18*) with double precision arithmetic. Principal components analysis was carried out using subroutines written in our laboratory and based on procedures outlined by Malinowski and Howery (*1*). Three-dimensional displays of experimental data were generated with the program SURFER (Golden Software, Golden, CO).

RESULTS AND DISCUSSION

Simulation Studies. To provide an initial evaluation of the Kalman filter algorithm for recursive PCA, simulated chromatographic data were used. Because a large number of parameters will affect the performance of the algorithm (chromatographic peak shapes, spectral profiles, spectral and chromatographic resolution, number of components, component ratios, noise level, background absorbance, number of wavelengths used, etc.), only a limited subset of results is presented here to demonstrate the principles of the method. More complete studies to investigate the limitations of the algorithm are ongoing.

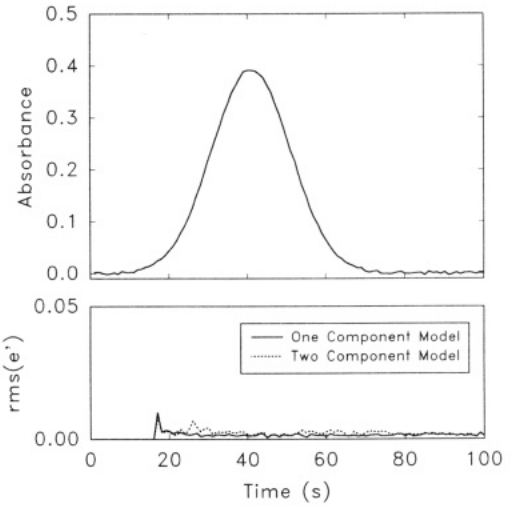

Figure 5. Results of the application of the Kalman filter PCA algorithm to a single-component elution profile (simulated). The top trace shows the chromatographic signal at the wavelength of maximum absorbance. The bottom trace shows the sequence of rms orthogonal innovations for each model type.

The simulated experimental data presented here utilized Gaussian profiles in both the chromatographic and spectral domains for simplicity. Component 1 was assigned a wavelength maximum of 400 nm and component 2 a maximum at 500 nm. The width of both spectral peaks was σ = 100 nm, and equivalent molar absorptivities were assumed. The concentration ratio and chromatographic resolution of components were varied between studies. Chromatographic peak widths of σ = 10 s were used with a simulated spectral sampling rate of 1 s. Chromatographic resolution was defined in the usual manner ($R = \Delta t_R/4\sigma$). The noise level (typically 0.5% RSD) was computed as a percentage of the maximum of the entire absorbance matrix. For all of the results presented here, 10 wavelengths at equally spaced intervals were used.

Figure 5 shows typical results obtained with the parallel Kalman filter network for the elution of a single component. The top part of the figure shows the chromatographic trace at the absorbance maximum, while the bottom portion shows the rms orthogonal innovations for the one- and two-component models. Note that both models indicate acceptable performance, verifying that there is only one component present. In contrast, Figure 6 shows results for two eluting components (3:1 ratio, 0.35 resolution, 0.5% noise). Under these conditions, the two-component model gives a fairly flat innovations trace, while the trace for the one-component model indicates significant model deviations. Furthermore, the point at which the innovations sequence begins to diverge for the one-component model reveals where the second component begins to appear. This information is not provided with batch PCA and is significant because it allows a key set of factors to be identified for target transformation (*7–9, 12*). This could expedite the generation of component elution profiles considerably.

In order to illustrate how the Kalman filter network functions, the evolution of a single one-dimensional model for a two-component data set is shown in Figure 7. One A^2 data space for the data in Figure 2c is shown. Lines in Figure 7 correspond to the one-dimensional model at various points throughout the elution of the peak and are labeled to correspond to points indicated in Figure 2c. Initially, when only one component is present, the linear model fits the observed data relatively well and the innovations are small (case I). As the second component introduces curvature into the data, the least-squares fit must accommodate this nonlinearity and the model begins to track the measurements more poorly, leading

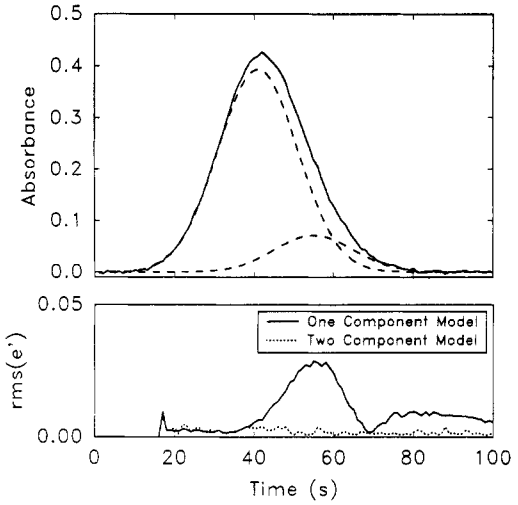

Figure 6. Results of the application of the Kalman filter algorithm to a simulated two-component elution profile.

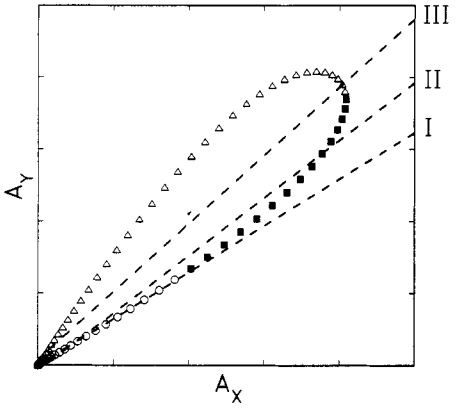

Figure 7. Evolution of the one-component model for the data in Figure 2c. The model equations (dashed lines) are shown at three stages: initial (after the open circles, I), intermediate (after the filled squares, II), and final (III). Corresponding points on the elution profile are indicated in Figure 2.

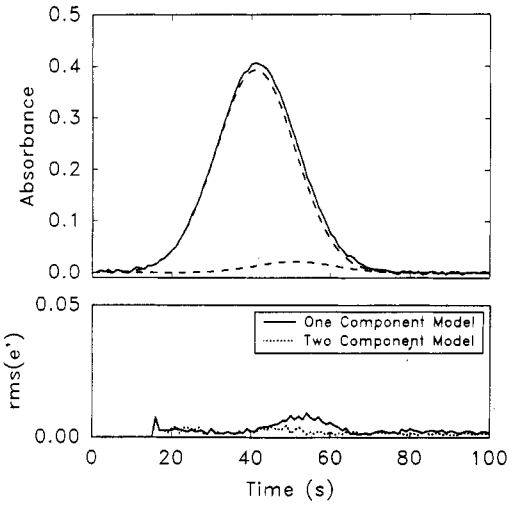

Figure 8. Results of the application of the Kalman filter algorithm to a simulated two-component elution profile near limiting conditions.

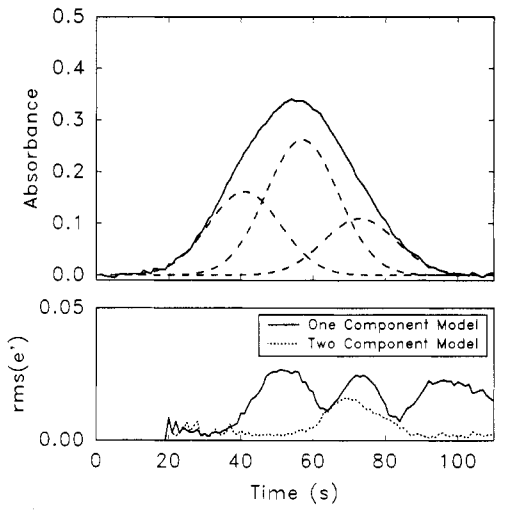

Figure 9. Results of the application of the Kalman filter algorithm to a simulated three-component elution profile.

to larger innovations (case II). Although the model is no better when the signal returns to the baseline (case III), measurements near the origin do not exhibit large deviations and so the innovations return to their original level. Note that in this sense the innovations do not behave the same as a residual sum of squares.

As an indication of the limitations of the Kalman filter PCA method, Figure 8 shows results obtained with a 10:1 component ratio and a resolution of only 0.25. The second component can still be detected in this case. Although the statistical significance of the small perturbation to the one-dimensional model might be contested, the algorithm is intended mainly to provide an indication of the possibility of a second component, so this point will not be addressed here. The ability of the algorithm to detect minor components is very dependent on the noise level, as expected. Generally, it was found that when the recursive algorithm failed to distinguish a second component, visual inspection of the data in the plane resulting from the first two eigenvectors also suggested only one component. As anticipated, performance of the method also improves with chromatographic and spectral resolution and with the number of wavelengths used. The latter effect arises from the increased likelihood of selecting wavelengths with maximum discriminating ability, and smoother traces for the rms innovations. Improvements in results achieved by increasing the number of wavelengths are quickly limited by the spectral correlation of the two components, however. The order of component elution (i.e. minor component first or second) affects the shape of the innovations trace but does not significantly diminish the ability of the algorithm to determine the dimensionality of the data set in most cases. In some extreme cases, the rms innovations of the two-component model exhibit a small disturbance when the second component is detected (i.e. the reverse of the usual case) but this is due to the fact that the planar model "floats" around its primary axis until the necessary points are obtained to more rigidly define the second eigenvector. The shape of the elution profile should also affect the performance of the algorithm, but this aspect has not been investigated in detail.

An example of a simulated three-component mixture is shown in Figure 9. In this case, the third component was assigned a wavelength maximum of 300 nm with σ = 100 nm. The concentration ratio (c_1:c_2:c_3) is 1:3:1, and components elute in order with a resolution of 0.4 between adjacent peaks. Other conditions are as previously given. Note that the trace of rms innovations indicates the successive failure of the one- and two-component models. It should be pointed out, however, that failure of the two-component model was not always observed for three-component mixtures, depending on the relationships among spectra and elution profiles. It is believed that this problem has its roots in the correlation between wavelengths selected for independent variables. Solutions include a more careful selection of wavelengths or imposition of a complete set of models with all wavelength combinations. Other options also exist but may not be necessary as the

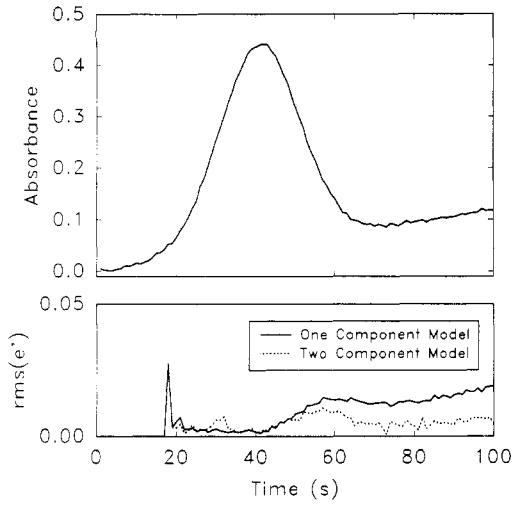

Figure 10. Effect of a sloping baseline on the Kalman filter algorithm.

pattern of the innovations for the one-dimensional model will indicate the presence of a third component in most cases.

As a final example, Figure 10 demonstrates the effect of a sloping background on the recursive PCA algorithm. Conditions used were the same as for Figure 5 (one component) but a gradient ramped up to 30% of the maximum absorbance was added at all wavelengths. A gradient of this type should show up as an additional component with batch PCA procedures with no indication of its source. With the Kalman filter method, deviations of the one-component model are also seen, but the divergence of the innovations trace is more gradual, indicating that the second component is due to changes in the background rather than coelution. Such information about the temporal structure of model variations is a particular advantage of the Kalman filter approach.

The computational performance of the parallel Kalman filter network is currently limited by its implementation in serial fashion but is still quite acceptable. Cycle times of about 0.1 s are not difficult to achieve with one- or two-component models at 10 wavelengths. This is in a range suitable for most chromatographic applications. The efficiency of the serial implementation will diminish as the models of higher dimensionality are added and the number of wavelengths is increased. The highly parallel nature of the algorithm can exploit trends in computing toward vector processing, however, and this should dramatically reduce computation time.

Experimental Results. Since simulated experimental data often imposes deterministic and stochastic characteristics which are not observed in practice (e.g. Gaussian profiles, uncorrelated noise), the recursive PCA algorithm was also applied to experimental data from the coelution of organic dyes. One of the data sets used in this study, obtained with the apparatus in Figure 4, is shown in Figure 11. Dye concentrations were reduced to a level which gave a relatively noisy signal (approximately 3% baseline noise relative to the absorbance maximum). It also appears from the figure that the noise exhibits some correlation, possibly due to pump pulsations. The ratio of peak heights (methyl orange to naphthol blue) was 2:1 and the resolution (determined by individual injection) was about 0.4. Ten wavelengths at equally spaced intervals were used. Results of the application of the Kalman filter are shown in Figure 12. The presence of two components in the elution profile is clearly indicated by the rms innovations sequence even though the noise level is quite high. The rms innovations are likely higher in this case than for simulated results due to the high correlation along the wavelength axis.

Comparison with PCA Results. A comparison between the eigenvectors computed by the usual batch PCA procedure

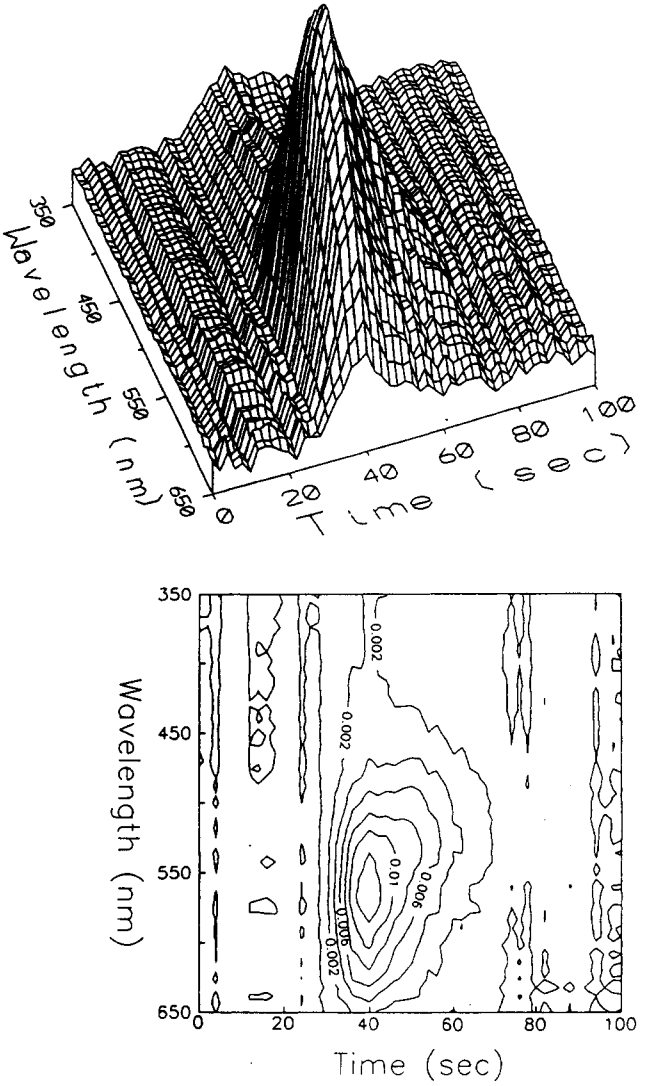

Figure 11. Absorbance matrix obtained from the coelution of organic dyes.

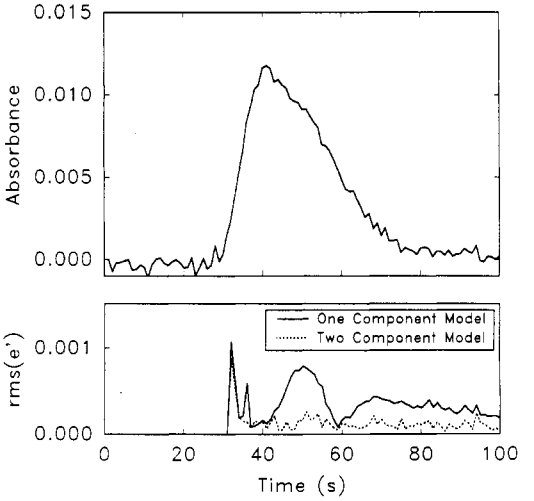

Figure 12. Results of the application of the recursive PCA algorithm to the data in Figure 11.

and those determined by the parallel Kalman filter network is given in Table I. The basis of the comparison is the angle between the eigenvectors calculated through the batch PCA procedure and the Kalman filter network. Results under various simulation conditions for a ten-dimensional space (ten wavelengths) are shown. The first eigenvector used from the

Table I. Angles between Eigenvectors Produced by Traditional Principal Component Analysis and the Kalman Filter Method

no. of components	concn ratio	resoln	noise level, %	angle between eigenvectors, deg: first eigenvector	angle between eigenvectors, deg: second eigenvector
1			0.5	0.01	53.0
2	3:1	0.35	0.5	0.06	1.37
2	3:1	0.35	2.0	0.49	7.99
2	10:1	0.25	0.5	0.02	6.39
2	5:1	0.2	0.5	0.08	3.12
2	1:5	0.2	0.5	0.06	3.02

Kalman filter network was that obtained from the combination of one- and two-component models rather than from the one-component model alone, but differences were insignificant. Agreement between the two PCA methods, while not perfect, is very good in most cases. One exception is where an attempt is made to determine the second eigenvector for the one-component data set. In this case, the second eigenvector is defined purely by noise and is of no real consequence, however. Certainly, the agreement between batch and recursive procedures should be good enough to permit further calculations, such as self-modeling curve resolution, to be carried out.

The recursive PCA method does not directly provide eigenvalues or the row and column matrices associated with batch PCA, but these can be easily determined (if necessary) once the eigenvectors are assigned. Eigenvalues are not as essential for determination of rank with the recursive algorithm since this information is provided by the rms innovations sequence.

CONCLUSIONS

The initial studies presented here have demonstrated the viability of performing principal components analysis recursively through the use of a parallel Kalman filter network. Application to the problem of chromatographic peak purity analysis has shown how the rank of a data matrix can be deduced while the data are being acquired. Although more extensive studies are required to fully explore the potential and limitations of this approach, several important advantages are apparent. First, because the algorithm is recursive and parallel with a fixed cycle time, it should be significantly faster than traditional PCA methods, especially when implemented on parallel computing architectures. The speed advantage does not result from a more computationally efficient algorithm, but rather because data analysis is performed while data are being acquired. A second advantage of the recursive approach is that it provides information on the temporal evolution of models. This is particularly useful in cases such as chromatography and titrimetry where certain types of behavior can be anticipated. To obtain equivalent information by batch PCA, numerous subsets of the data would have to be processed independently. The information provided by recursive PCA should be particularly useful in resolving ternary component mixtures by evolutionary factor analysis (*13*) since it identifies regions in which certain models are valid. Furthermore, it can help diagnose model deviations arising from factors such as a sloping background. Absolute information on model deviations is readily provided by the rms innovations sequence, which should approximate measurement noise when the model is valid. Finally, the flexibility of the Kalman filter models allows for a variety of processing options to be exercised, simultaneously if desired. Inclusion of the offset term in the models, for example, will have the same effect as mean-centering the absorbance data prior to batch PCA. Unlike some approaches (*3*), however, the absorbance data are not normalized, so the measurement noise information is retained at its original magnitude (*11*).

In spite of these advantages, it is clear that the Kalman filter network will not be useful in those cases where real-time data processing is not required. It is also likely to become less useful as the number of factors to be extracted becomes large, since the number and complexity of models become more difficult to handle. Nevertheless, it may allow techniques such as self-modeling curve resolution to be more readily implemented in real time.

LITERATURE CITED

(1) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley: New York, 1980.
(2) Sharaf, M. A.; Kowalski, B. R. *Anal. Chem.* **1981**, *53*, 518–522.
(3) Osten, D. W.; Kowalski, B. R. *Anal. Chem.* **1984**, *56*, 991–995.
(4) Ramos, L. S. Ph. D. Thesis, University of Washington, Seattle, WA, 1988.
(5) Yost, R.; Stoveken, J.; MacLean, W. *J. Chromatogr.* **1977**, *134*, 73–82.
(6) Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617–633.
(7) McCue, M.; Malinowski, E. R. *Appl. Spectrosc.* **1983**, *37*, 463–469.
(8) Gemperline, P. J. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 206–212.
(9) Vandeginste, B. G. M.; Derks, W.; Kateman, G. *Anal. Chim. Acta* **1985**, *173*, 253–264.
(10) Vandeginste, B.; Essers, R.; Bosman, T.; Reijnen, J.; Kateman, G. *Anal. Chem.* **1985**, *57*, 971–985.
(11) Lindberg, W.; Ohman, J.; Wold, S. *Anal. Chem.* **1986**, *58*, 299–303.
(12) Gemperline, P. J. *Anal. Chem.* **1986**, *58*, 2656–2663.
(13) Schostack, K. J.; Malinowski, E. R. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 121–141.
(14) Azimi-Sadjadi, M. R.; Lu, T.; Nebot, E. M. *IEEE Trans. Signal Process.* **1991**, *39*, 137–147.
(15) Kalman, R. E. *J. Basic Eng.* **1960**, *181*, 35–45.
(16) Brown, S. D. *Anal. Chim. Acta* **1986**, *181*, 1–26.
(17) Rutan, S. C. *J. Chemomet.* **1987**, *1*, 7–18.
(18) Wentzell, P. D.; Wade, A. P.; Crouch, S. R. *Anal. Chem.* **1988**, *60*, 905–911.
(19) Wentzell, P. D.; Vanslyke, S. J. *Anal. Chim. Acta*, in press.
(20) Brown, S. D.; Rutan, S. C. *Anal. Chim. Acta* **1984**, *160*, 99–119.
(21) Brown, S. D.; Rutan, S. C. *J. Res. Natl. Bur. Stand. (U.S.)* **1985**, *90*, 403–407.
(22) Barker, T.; Brown, S. D. *J. Chromatogr.* **1989**, *469*, 77–90.
(23) Redmond, M.; Brown, S. D.; Wilk, H. R. *Anal. Lett.* **1989**, *22*, 963–979.
(24) Hayashi, Y.; Yoshioka, S.; Takeda, Y. *Anal. Chim. Acta* **1988**, *212*, 81–94.
(25) Williamson, J. H. *Can. J. Phys.* **1968**, *46*, 1845–1847.
(26) Sorenson, H. W. *IEEE Spectrum* **1970**, *7* (7), 63–68.
(27) Poulisse, H. N. J. *Anal. Chim. Acta* **1979**, *112*, 361–374.

RECEIVED for review May 30, 1991. Accepted August 14, 1991. We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada.