Multivariate Curve Resolution Applied to Spectral Data from Multiple Runs of an Industrial Process

Roma Tauler[†] **and Bruce Kowalski***

Center for Process Analytical Chemistry, University of Washington, BG-10, Seattle, Washington 98195

Sydney Fleming

duPont de Nemours & Company, Inc., Experimental Station, Building E357, P.O. Box 80357,
Wilmington, Delaware 19880-0357

A method for extracting information from spectroscopic data gathered during process monitoring is described and applied to an industrial problem. The method allows the estimation of the changes in the concentrations of the components in the process as well as their pure spectroscopic responses. Three key aspects of the new method are as follows: (1) the use of evolving factor analysis to have an initial estimation of how the concentrations of the constituents change during the process; (2) the implementation of an alternating and constrained least-squares method to optimize both the spectra and the concentrations of the components in the process; (3) the development of a new approach for the simultaneous analysis of various runs of the same process to estimate the ratio of concentrations between the common components in the different runs.

INTRODUCTION

Recent advances in process instrumentation and in data collection techniques have resulted in a rapid increase in the amount of data that can be acquired from chemical processes. Extracting the significant information from the data produced by modern instrumentation is in many circumstances a nontrivial task. The description and modeling of the evolution of a chemical process is important for both practical and economic reasons. The present work concerns the application of new tools for process monitoring and modeling. These tools can be used to extract all of the useful chemical information from process monitoring data.

The development and application of first and higher order multivariate calibration and standardization methods have allowed the solution of many problems in real-world process analytical chemistry, and much attention has been given to these methods in the field of chemometrics in recent years.[1] There are, however, important cases where these methods cannot be applied because there is no previous information available to perform a calibration of the system. Examples of this situation in process analysis are abundant and include monitoring the evolution of chemical processes where one or more parameters are changed, such as time, temperature, pH, the concentration of a reagent, or any other parameter, and there is no previous quantitative information about the evolution of the process. The multivariate data acquired with spectroscopic probes produce continuous data which can be arranged in an ordered data matrix according to the variation of the parameter changed during the process.

One way to address this problem is by curve resolution methods.[2,3] In the factor analysis framework, curve resolution decomposes a bilinear data matrix into the product of two simpler matrices which are related respectively to each one of the two orders of the original data matrix. The goal of the curve resolution methods is the determination of those decompositions which have physical and chemical meaning. In the case of process analytical chemistry, the final goal is the estimation of the matrix containing the concentration profiles of the constituents as a function of time and simultaneously the estimation of the unit responses (pure spectra) of those constituents.

To have a meaningful solution from the curve resolution decomposition, it is necessary to make some assumptions about the signals obtained such as bilinearity, nonnegativity, unimodality, and closure. Examples of such treatments with more or less success can be found in the literature.[4-7] However, in general, such treatments do not guarantee unique solutions because the rotational and intensity ambiguities inherent to curve resolution decompositions can still be present after applying the above-mentioned constraints. As it is shown in the present work, these ambiguities can be partly overcome with the use of some special techniques. One of the most interesting techniques, which has not received much attention in the field of process analysis, is evolving factor analysis (EFA).[8-12] Taking advantage of the ordered structure of the acquired data, evolving factor analysis provides valuable information concerning the windows of existence of every component in the unknown mixtures existing at any time during the process. When evolving factor analysis is applied to cases where selectivity for some component is present in any of the two orders, the determination of the concentration profile and spectroscopic response of such component can be estimated at least qualitatively without any other additional requirement. Therefore, it is especially important to detect such selectivity regions.

* Author to whom correspondence should be addressed.

[†] On leave from the Department of Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona 08028, Spain.

(1) Kowalski, Br. R.; Seasholtz, M. B. *J. Chemom.* **1991**, *5*, 129–145.

(2) Hamilton, J. C., Gemperline, P. J. *J. Chemom.* **1990**, *4*, 1–14.

(3) Windig, W., *Chemom. Intell. Lab. Syst.* **1992**, *16*, 1–16.

(4) Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617–632.

(5) Borgen, O. S.; Kowalski, B. R. *Anal. Chim. Acta* **1985**, *174*, 1–26.

(6) Spjotvoll, E.; Martens, H.; Volden, R. *Technometrics* **1982**, *3*, 173–180.

(7) Vandeginste, B. G. M.; Darks, W.; Kateman, G. *Anal. Chim. Acta* **1985**, *173*, 253–264.

(8) Gampp, H.; Maeder, M.; Meyer, Ch.; Zuberbuhler, A. D. *Talanta* **1985**, *32*, 1133–1139.

(9) Gampp, H.; Maeder, M.; Meyer, Ch.; Zuberbuhler, A. D. *Chimia* **1985**, *39*, 315–317.

(10) Gampp, H.; Maeder, M.; Meyer, Ch.; Zuberbuhler, A. D. *Talanta* **1986**, *33*, 943–951.

(11) Cartwright, H. *J. Chemom.* **1986**, *1*, 111–120.

(12) Gampp, H.; Maeder, M.; Meyer, Ch.; Zuberbuhler, A. D. *Anal. Chim. Acta* **1987**, *193*, 287–293.

0003-2700/93/0365-2040$04.00/0

In addition, if several process runs of the same process are available and as usually, at least one of the two orders is common between them (e.g., the spectral range scanned), the intensity ambiguities associated with the analysis of a single process run can be resolved. Assuming that the same component in the mixture has the same unit spectrum in the different process runs, the simultaneous analysis of different process runs gives the concentration ratios between the common components in the different runs of the process.

Finally, in the case where both orders are in common between different process runs, it is conceptually and mathematically better to take advantage of the second-order structure of the data and use multivariate higher order curve resolution methods like generalized rank annihilation method (GRAM),[13] residual bilinearization (RBL),[14] or other three-way data analysis methods.[15] However, the requirement of having the two orders coincident (synchronization) in different experiments is too strict for many practical situations in process analysis. The present work is addressed to such situations.

Data for the present work are spectra obtained from successive runs of an industrial chemical process. The spectra are evenly spaced in time, but may represent different elapsed times from beginning to end from run to run. The spectral intensities are stored in a matrix **D** with one spectrum per row. The dimensions of **D** are the number of spectra by the number of spectral channels (e.g., wavelength). The goals of the present work are (a) to determine the number of unique chemical components included during the full process; (b) to estimate the pure (unit) spectra of these components; and (c) to estimate how the concentration profiles of these components change within a run of the process, and how they differ between different runs of the same process.

METHODS

To achieve the goals mentioned above, different multivariate data analysis techniques have been implemented and assembled in a single method. The different parts of this method are as follows:

Principal Component Analysis and Curve Resolution. Principal component analysis[16] gives a decomposition of a data matrix:

$$\mathbf{D} = \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{D}} + \mathbf{E} \qquad (1)$$

where **U** and **V** are respectively the score and loading matrices obtained for the selected number of principal components, **E** is the residual error or noise matrix not explained by them, and $\hat{\mathbf{D}}$ is the reproduced data matrix. The decomposition obtained by principal components gives orthogonal **U** and $\mathbf{V}^{\mathrm{T}}$ matrices.

The difference, **E**, between the original data matrix **D** and the reproduced matrix $\hat{\mathbf{D}}$ is calculated to know the level of residual variance not explained by the number of deduced components. For a complete bilinear data matrix, **E** should be at the level of noise or experimental error for the correct number of components.

Assuming that the spectroscopic data matrices obtained in the process analysis experiments are bilinear:

$$\mathbf{D} = \mathbf{C}\mathbf{A} \qquad (2)$$

$$d_{\mathrm{ij}} = \sum c_{\mathrm{ik}} a_{\mathrm{kj}}$$

where **D** is the data matrix which contains the spectra acquired in an ordered way along the process, for instance along the time axis. **C** and **A** are respectively the matrices of the concentration profiles in the process and the unit pure chemical component spectra for the set of components spectroscopically active in the range studied. The dimensions of these matrices are **C** (NS × NC) and **A** (NC × NW), where NS is the number of spectra acquired, NW is the number of spectroscopic channels, and NC is the number of chemical components in the mixtures.

The goal of curve resolution methods is as follows: given **D**, obtain the real physically meaningful **C** and **A**. Obviously such a task cannot be achieved directly from the principal component analysis decomposition if no additional information is provided since the equation

$$\hat{\mathbf{D}} = \mathbf{U}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{T}\mathbf{T}^{-1}\mathbf{V}^{\mathrm{T}} = \mathbf{C}\mathbf{A} \qquad (3)$$

has an infinite number of solutions for any arbitrary transformation matrix **T**. There is a rotational and an intensity ambiguity to solve for **C** and **A** if no more information is provided to constrain the number of possible solutions. This is the task, in general, of the curve resolution methods and, in particular, of the proposed method.

Intensity Ambiguity. There is an intrinsic intensity ambiguity in all curve resolution solutions since for any scalar m

$$d_{\mathrm{ij}} = \sum c_{\mathrm{ik}} a_{\mathrm{kj}} + e_{\mathrm{ij}} \qquad (4)$$

$$c_{\mathrm{ik}} a_{\mathrm{kj}} = (c_{\mathrm{ik}} m)(1/m a_{\mathrm{kj}}) = c'_{\mathrm{ik}} a'_{\mathrm{kj}}$$

This means that the estimated concentrations and spectra will be scaled by some unknown factor m indigenous to each component. This is not a serious problem in qualitative analysis (spectral identification, fingerprinting), but it is a serious problem in quantitation.

Rotation Ambiguity. More important is the rotation ambiguity inherent in curve resolution solutions which always occurs when there are two or more linearly independent overlapped components. The estimated spectrum for any of these components will be an unknown linear combination of the true components

$$a_{\mathrm{k}}' = \sum t_{\mathrm{k}} a_{\mathrm{k}} \qquad (5)$$

where t_{k} are unknown rotation constants; a_{k} are the true unit pure spectra of the components; and a_{k}' are their estimated pure unit spectra.

Selectivity. Conversely, for those time windows during a process run where there is only one component, there is no rotation ambiguity. This means that the principal component analysis or any other solution gives the correct shapes for the concentration profiles and unit spectra. Only the intensity ambiguity will still be present.

To remove the rotational ambiguity, it is important to detect the regions where selectivity is present, because in these regions the unit spectra and unit concentration profiles have the correct shapes.

Determination of Number of Components in Mixture. As the interest is centered on the investigation of the changes in the concentration of the chemical components present in the system as well as in the nature of these unknown components, the first thing that must be done is to estimate how many different components are in the data set. The determination of this number is related to the determination of the level of variance which is caused by other variance sources (e.g., light scatter). Methods such as cross validation[17] or the theory of error in factor analysis[18] will not work here

[13] Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1990**, *4*, 29–45.

[14] Ohman, J.; Geladi, P.; Wold, S. *J. Chemom.* **1990**, *4*, 135–146.

[15] Smilde, A. K. *Chemom. Intell. Lab. Syst.* **1992**, *15*, 143–157.

[16] Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

[17] Wold, S. *Technometrics* **1978**, *20*, 397–405.

[18] Malinowski, E. R. *Anal. Chem.* **1977**, *49*, 612–617.

because they provide the total number of contributions to the data variance, chemical and nonchemical baseline changes, and the background absorption.

If the chemical components have a larger contribution to the data variance than the noise, background, and baseline changes, the number of chemical components can be initially estimated simply from the comparison of their respective singular values. Singular values related to the background absorption and baseline changes are obtained in the analysis of the spectral regions where the chemical components do not absorb. Singular values related to the chemical components of interest are obtained in the analysis of the spectral regions within the absorption bands. The number of components estimated in this way is tested aftewards using evolving factor analysis (see below) and during the alternating least-squares optimization (see below), looking for those solutions which best fit the data and have physical meaning, i.e., give reasonable shapes in the concentration profiles and unit spectra.

Determination of Selectivity of System and Initial Estimation of Concentration Profiles by Evolving Factor Analysis. Evolving factor analysis (EFA)[8–12] has been applied mostly to the study of spectroscopic experiments of multiequilibria systems[19–22] and to liquid chromatography with diode array detection.[23,24] The basic idea of this procedure is to provide an initial estimation of the concentration profiles, examining how the singular values associated with these components evolve and change in magnitude along the process. In this procedure, the detected presence of selectivity ranges can solve the rotational ambiguity. Other methods used for detecting the selectivity of the system are local rank analysis,[25] window factor analysis,[26] and fixed-size moving window evolving factor analysis.[27]

Constrained Optimization of Concentration Profiles and Unitary Spectra by Alternating Least Squares. From the results of evolving factor analysis, the window or range of existence of each component as well as its concentration profile can be obtained. These concentration profiles are used as initial values in a constrained alternating least-squares optimization procedure. At each iteration of the optimization, a new estimation of the matrix of spectra **A** and of the concentration profiles **C** is obtained successively using the two following equations:

$$\mathbf{A} = \mathbf{C}^{+}\hat{\mathbf{D}} \tag{6}$$

$$\mathbf{C} = \hat{\mathbf{D}}\mathbf{A}^{+} \tag{7}$$

where the matrix $\mathbf{C}^{+}$ is the pseudoinverse of the matrix **C**, and the matrix $\mathbf{A}^{+}$ is the pseudoinverse of the matrix **A**. Obviously the selection of the correct number of components in the calculation of $\hat{\mathbf{D}}$ is essential. The use of this matrix instead of the experimental data matrix, **D**, improves the stability of the calculations, since $\hat{\mathbf{D}}$ is noise-filtered.

In order to limit the number of possible solutions to these equations, the following set of constraints can be applied:

Constraints on Concentration Profiles. (*a*) *Nonnegativity.* The concentration profiles are positive.

(*b*) *Shape.* When the shape characteristics of the concentration profiles of the different components in the process are known, they can be used as a constraint. A typical example is unimodality (only one peak per profile). In the present case, the shapes of the concentration profiles were allowed to have some small local departures from the unimodality condition.

Constraints on Unit Spectra Values. (*a*) *Nonnegativity.* The spectral values are forced to be positive. This condition is not applicable to derivative spectra.

Testing the Number of Components. If the number of components is uncertain, the complete iterative alternating least-squares optimization is performed for the different most plausible number of components. The correct number is considered to be the one which gives the best fit to the experimental data matrix **D** and physically meaningful solutions: This means solutions with reasonable shapes for the unit spectra and concentration profiles.

Uniqueness of Alternating Least-Squares Solutions and Simultaneous Analysis of Several Runs of the Process. As pointed out above, the rotational ambiguity associated with the alternating least-squares solutions can be solved for those components which have selectivity associated with at least one of the orders (spectral or time) of measurement. However there is still no guarantee that for these components the intensity ambiguity is removed. However, we have learned that the intensity ambiguity in the alternating least-squares solution can be solved by the simultaneous analysis of two or more process runs at different experimental conditions (e.g., different starting concentrations of the constituents).

Suppose there are NP different process runs of the same chemical process studied at different initial conditions or starting concentrations of the constituents. For each run of the process, a bilinear data matrix $\mathbf{D}_i$ is obtained:

$$\mathbf{D}_i = \mathbf{C}_i\mathbf{A} \qquad i = 1, 2, ..., \mathrm{NP} \tag{8}$$

$\mathbf{C}_i$ is the matrix of the concentration profiles of the chemical components in that particular run of the process, and **A** is the matrix of the unit or pure spectra of these components. In general, since the different process runs are obtained at different conditions, the concentration profiles of the different components in each process run $\mathbf{C}_i$ will differ not only in intensity but also in shape. Indeed, it is important to point out here that the procedure proposed in the present work allows the concentrations profiles of the constituents to change from process run to process run not only in intensity but also in shape. However, the spectra of the common components in the different process runs are considered to be equal and described in a unique matrix (see below for the case of noncommon components). This assumption is true whenever the external conditions like temperature and solvent composition are kept constant.

Because the number of columns (wavelengths) is the same for all the $\mathbf{D}_i$ matrices analyzed simultaneously, the complete data set can be arranged in a single augmented data matrix with the columns (wavelengths) in common and with a number of rows equal to the total number of acquired spectra in all the different process runs:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \cdot\cdot \\ \cdot\cdot \\ \cdot\cdot \\ \mathbf{D}_{\mathrm{NP}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \cdot\cdot \\ \\ \cdot\cdot \\ \mathbf{C}_{\mathrm{NP}} \end{bmatrix} \mathbf{A} = \mathbf{CA} \tag{9}$$

$$\mathbf{D} = \mathbf{CA} \tag{10}$$

Similarly to what is done in the case of the analysis of the individual process runs, the augmented data matrix **D** can be

[19] Tauler, R.; Casassas, E. *J. Chemom.* **1988**, *3*, 151–161.

[20] Tauler, R.; Casassas, E. *Anal. Chim. Acta.* **1989**, *223*, 257–268.

[21] Tauler, R.; Casassas, E.; Izquierdo-Ridorsa, A. *Anal. Chim. Acta* **1991**, *248*, 447–458.

[22] Tauler, R.; Casassas, E. *Analusis* **1992**, *20*, 255–268.

[23] Maeder, M. *Anal. Chem.* **1987**, *59*, 527–30.

[24] Maeder, M.; Zilian, A. *Chemom. Intell. Lab. Syst.* **1988**, *3*, 205–213.

[25] Geladi, P.; Wold, S. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 273–281.

[26] Malinowski, E. R. *J. Chemom.* **1992**, *6*, 29–40.

[27] Keller, H. R.; Massart, D. L. *Anal. Chim. Acta* **1991**, *246*, 379–390.

decomposed using principal component analysis

$$\mathbf{D} = \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{D}} + \mathbf{E} \qquad (11)$$

where now **U** and $\mathbf{V}^{\mathrm{T}}$ are respectively the score and loading matrices of **D** for the preselected number of components, **E** is the residual error matrix containing the variance not explained by these numbers, and $\hat{\mathbf{D}}$ is the reproduced data matrix. Under the assumption of linearity, the number of correct components will give a residual error matrix close to the noise or experimental error.

As a first step in the curve resolution method, the matrix of the concentration profiles **C** is estimated from the initial estimation of the $\mathbf{C}_1$, $\mathbf{C}_2$, ..., $\mathbf{C}_{\mathrm{NP}}$ submatrices obtained by evolving factor analysis of the $\mathbf{D}_1$, $\mathbf{D}_2$, ..., $\mathbf{D}_{\mathrm{NP}}$ submatrices, as described before. An initial estimation of the augmented **C** matrix is then obtained simply by setting the estimation of the $\mathbf{C}_i$ matrices one on top of each other in the same order as they are in **D**.

As done in the single-process data analysis, at each iteration of the alternating least-squares method, a new estimation of the matrix of spectra **A** and of the concentration profiles **C** is obtained. The same procedure is used but is now applied to the augmented matrix obtained by principal component analysis for the considered number of components, $\hat{\mathbf{D}}$. In addition to the constraints present in the single-process analysis, in the multiple-process analysis there are two more constraints to apply:

(*a*) *Common Components Have Unique Spectra in All Process Runs.* When the set of unit spectra in **A** are obtained for the different process runs analyzed, the components which are in common in the different process runs are forced to have the same unit spectra. This constraint has an important effect on resolving the concentration profiles and unit spectra in terms of quantitation. The scale and intensity ambiguities can be removed in this way.

(*b*) *Zero Concentration Components.* When a component is known not to be present in a specific run, then the concentration of such component in **C** is forced to be equal to zero. The question about the presence or absence of a certain component in a run can be answered by looking at the pseudo rank of the associated data matrix and also by looking at the results of the individual analysis over that particular run to see whether there is coincidence between the shapes of the recovered spectra (fingerprint matching). Visual inspection by the analyst is required in this step.

The alternating least-squares procedure is repeated until convergence is achieved or until a predetermined number of cycles has occurred.

Determination of Global Equation To Describe the Process at Any Point in Time During Process Run. The procedure described above explains how the process evolves in the several runs analyzed. However it is also of great practical interest to have a simple way to describe the process just from its spectrum at any time and use such a method to predict the state of the process in future process runs. Assuming that the conditions of the process will be similar to the ones analyzed (which is reasonable because they are supposed to represent the different situations found in practice), it is possible to provide a set of coefficients which can be used on line to give the concentrations of the components at any time during the process. These coefficients are found from the pseudoinverse of the unit spectra obtained in the least-squares optimization. As will be shown later, better results are obtained when the experimental spectra are subtracted by the first spectrum at time zero, because then baseline differences between runs are partly removed. The equations to obtain these coefficients are derived from the bilinear model defined in eq 2.

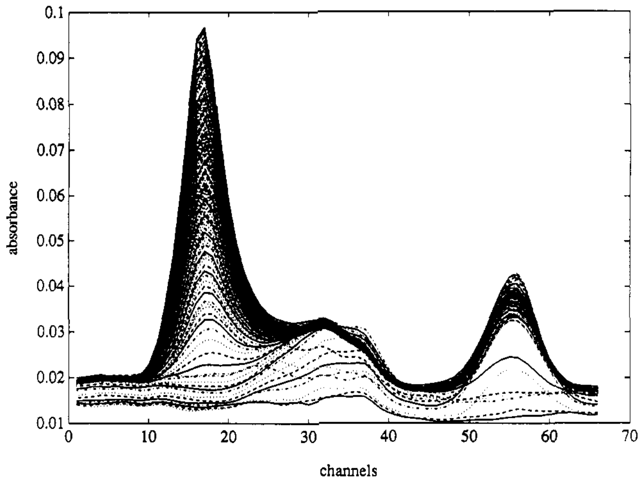

Figure 1. Example of the spectra acquired in one process run.

$$\mathbf{K} = \mathbf{A}^{+} \qquad (12)$$

$$r_{\mathrm{s}} = r_{\mathrm{t}} - r_0 \qquad (13)$$

$$c_{\mathrm{t}} = r_{\mathrm{s}}\mathbf{K} \qquad (14)$$

where **K** is the matrix of coefficients, $\mathbf{A}^{+}$ is the pseudoinverse of the unit spectra matrix, r_{t} is the spectrum at any time of the process, r_0 is the first spectrum at time 0, r_{s} is the subtracted spectrum, and c_{t} is the calculated concentration of the components at any time of the process.

EXPERIMENTAL SECTION

Eight different runs of the same industrial chemical process at different days of production were analyzed. Every run generated between 75 and 125 spectra, 795 in total, measured along a selected IR spectral range of 66 channels. An example of the data collected in one of the runs of the process is given in Figure 1. The entire IR spectrum was not used because it contained information on process components that were not of interest to this study. The spectra change with time, starting from a very weak and flat background absorption, increasing the absorption to give two main absorption bands at channel numbers of 10–20 and 50–60 and a broader absorption band around channel number 30, and finally decreasing the absorption very rapidly on all bands when the process is terminated. All runs show a similar pattern but with slight differences in the timing and in the position of the maxima (shifting), which show that there are some differences in the chemistry of the different runs of the process. The problem to solve is as follows: Given the set of spectra collected along the different runs of the process determine (a) how many chemical components are responsible for the observed spectral changes; (b) how the concentrations of these components change within every run of the process, and also how these concentrations differ from one run to another; (c) what are the spectral features of these components; and (d) what the method is that describes the process on line, at any time during a future process run.

The background absorption and baseline can also change during the process and are different for different runs. Therefore, some pretreatment of the data is needed. Two methods were used.

First, to account for the differences in the initial baseline absorption among different process runs, subtraction of the first spectrum of each run from the following spectra in the same run removes these differences. This is true because in the first spectrum only the baseline or initial background absorption is present and the spectral bands of interest have not appeared. With this treatment the first spectrum in each run will be zero.

Second, to account for the changes in the baseline or background during a particular run, the first and second derivatives of the raw spectra are calculated. That pretreatment allows the minimization of the contributions which are constant along a

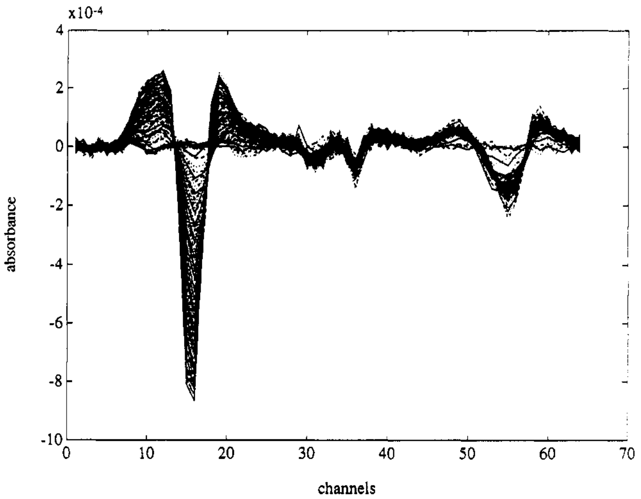

Figure 2. Second derivative spectra of one process run.

Table I. Comparison of Reduced Singular Values[a]

795 spectra, 64 channels	795 spectra, 5 channels
1.5381559×10^{-6}	1.4242273×10^{-7}
4.4253437×10^{-7}	1.1659144×10^{-7}
1.3784506×10^{-7}	1.3244814×10^{-7}
9.2429634×10^{-8}	1.6341295×10^{-7}
4.8231603×10^{-8}	2.0548819×10^{-7}

[a] Obtained in the analysis of the complete second derivative data set to the reduced singular values obtained in the analysis of the spectral regions where there is no contribution of the components of interest.[28]

particular spectrum (first derivative) as well as those contributions which produce constant slopes within every spectrum (second derivative). In Figure 2, the second derivative spectra of the data set in Figure 1 are given. Most of the baseline changes are now removed.

Both pretreatment, second derivative, and subtraction of the first spectrum were analyzed by the methods described herein and also compared with the analysis of the raw experimental spectra without any pretreatment.

RESULTS AND DISCUSSION

(1) Determination of Number of Components. The estimation of the singular values related to noise is performed using the first channels of the second derivative spectra where no band is present (see Figure 2). As mentioned before, in the second derivative spectra, the background and baseline contributions to the data variance are considerably diminished. The singular values obtained in this narrow spectral range are estimated for all the process runs together to include the variation between process runs. At the same time, the singular values of the complete data set comprising the 64 channels of measurement (process runs individually, and all together) are also calculated. For the comparison (Table I), the dimensions of the data matrix in one case and another were not the same, and therefore the reduced singular values[28] were used. When the complete set of runs and spectra are analyzed, the maximum number of different chemical components was estimated as either two or three, since the value of the third singular value is similar to the first singular value associated with the noise in the nonabsorbing parts of the spectra. When the analysis is performed over the individual process runs, it was found that the number of chemical components was always between two and three.

The number of components obtained in this way is used only to start the procedure. Three components were confirmed from evolving factor analysis and the alternating least-squares optimization procedures (see below).

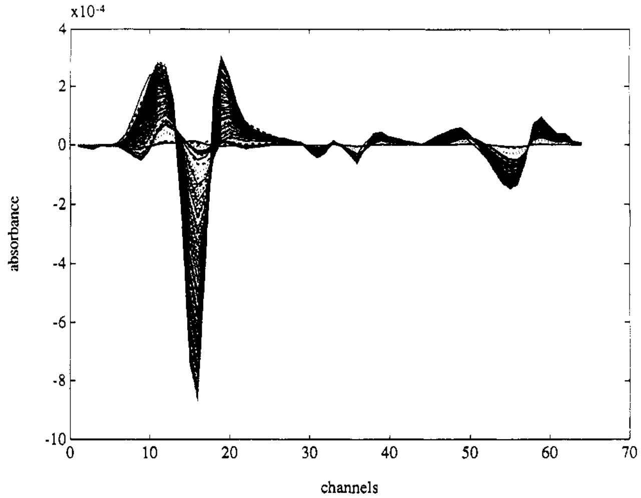

Figure 3. Reproduced second derivative spectra of the process run given in Figure 2 using three principal components.

Table II. Standard Deviation of Residuals Obtained in Data Analysis[a]

	2 components	3 components	4 components
exp – PCA[b]	2.82×10^{-3}	1.18×10^{-3}	1.16×10^{-3}
PCA – calc[c]	5.50×10^{-3}	3.30×10^{-3}	6.76×10^{-3}
exp – calc[d]	6.19×10^{-3}	3.80×10^{-3}	6.86×10^{-3}

[a] Results are given for two, three, and four components in the simultaneous analysis of the eight runs of the process. [b] Standard deviation of the residuals between the experimental data matrix and the principal component analysis reproduced data matrix. [c] Standard deviation of the residuals between the principal component analysis reproduced data matrix and the calculated data matrix using the optimized set of unit spectra and concentration profiles given in Figures 7 and 8. [d] Standard deviation of the residuals between the experimental data matrix and the calculated data matrix using the optimized set of unit spectra and concentration profiles given in Figures 7 and 8.

(2) PCA Analysis. Figure 3 shows the recalculated spectra shown in Figure 2 using principal component analysis. The principal components used in reproduction are the three more significant found by the principal component analysis of the whole augmented data matrix. Comparison of Figures 2 and 3 shows that most of the dominant spectral features are described by the three principal components obtained in the analysis of the 795 spectra from the eight different runs from the process. Noise filtering is also achieved, and very little information is lost by using only three components. If only two components were used, a poor reproduction of the original data is observed. Four components yield very little improvement. The standard deviations of the residuals between the eight-run experimental matrix and the reproduced matrix considering two, three, and four principal components are given in Table II.

(3) Evolving Factor Analysis. Figure 4 is a plot of the concentration profiles from evolving factor analysis (see refs 8–12, and above) applied to one run of the process. From evolving factor analysis, three components are detected and differentiated from the other contributions. The fourth and fifth components emerge significantly from the error contributions only at the very end of the process when the reaction was terminated and are, therefore, not of interest for the present study. Similarly, the evolving factor analysis of each of the eight other process runs provides an initial estimation of the concentration profiles of the components in each process run. These concentration profiles are used as initial values

[28] Malinowski, E. R. *J. Chemom.* **1987**, *8*, 33–40.

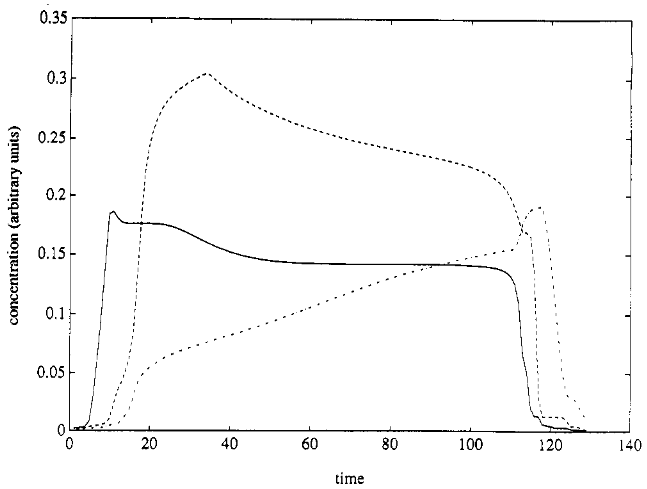

Figure 4. Initial concentration changes estimated by using evolving factor analysis of a single-process run.

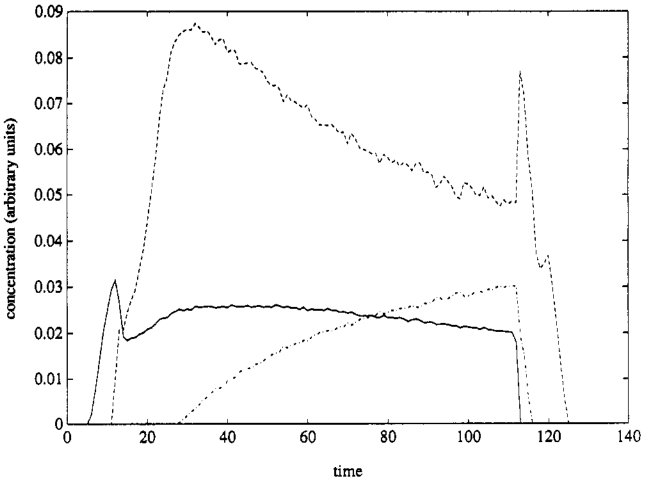

Figure 5. Concentration changes calculated by constrained alternating least squares of a single-process run.

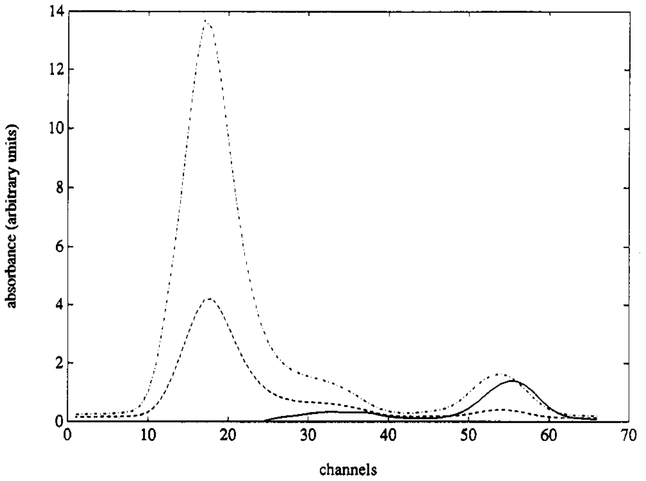

Figure 6. Unit (pure) spectra calculated by constrained alternating least squares of a single-process run.

in the alternating least-squares optimization. They can be used in the individual analysis of each set of data, or better, to build up the initial estimation of the augmented concentration matrix to be used in the simultaneous analysis of the eight runs of the process.

(4) Alternating Least-Squares Optimization Applied to Data from Single-Process Run. Figures 5 and 6 give respectively the optimized concentration profiles and unit spectra obtained in the analysis of an individual process run. These concentration profiles and unit spectra are obtained after applying the constrained optimization described before when data from a single-process run are analyzed. For this process run, some selectivity exists for both the first and third components and because of the effect of the applied constraints.

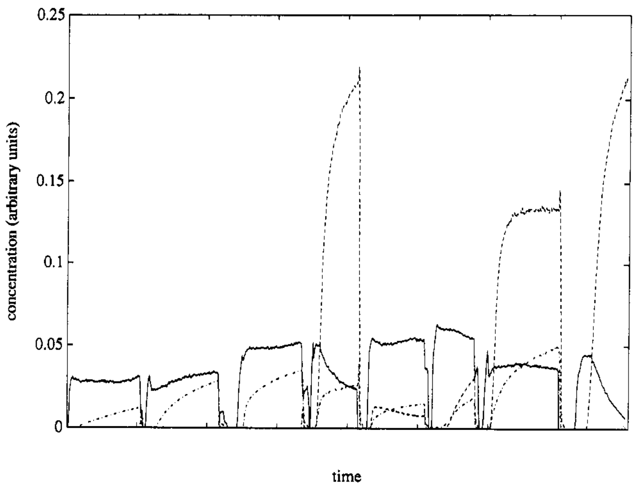

Figure 7. Concentration changes calculated by constrained alternating least squares of eight process runs. *X*-axis is a time axis and gives an indication of the spectra number within the total number of spectra analyzed (795 in total).

(5) Alternating Least-Squares Optimization Applied To Simultaneous Analysis of Data Acquired from Eight Runs of the Process. The alternating and constrained least-squares optimization method was applied to the augmented data matrix containing the eight runs of the process arranged in the three following forms: (a) without any pretreatment, (b) with the first spectrum of each run subtracted from all spectra, and (c) to the second derivative augmented data matrix. The concentration matrix used initially in the optimization was the augmented concentration matrix containing the concentrations obtained in the individual evolving factor analysis of each run. Of the three arrangements of the data matrix, the one which gave the best results is the second case where the spectra of each run are corrected by subtracting the first spectrum of the same run. The reason for the superior results in this case is that the subtraction of the first spectrum of each run removes the arbitrary offset between data from different process runs. When the optimization is performed using the second derivative spectra, the nonnegativity constraint is lost and cannot be applied over the unit spectra. While the results are still in agreement with those obtained with the subtracted data matrix, the shapes of the recovered concentration profiles and unit spectra are less reliable. Of the three cases, the poorest results were obtained when no pretreatment was performed. The reason for this degradation of the resolution is because of the effects of the baseline (background absorption) differences between process runs.

In order to summarize the large amount of calculations performed during the present work, only the results obtained from the simultaneous analysis of the eight runs of the process will be given. In the alternating least-squares optimization of thè complete data set, the number of three components is again reconfirmed. If another number of components is used, not only is it the fit worse, (see Table II) but also the shapes of the recovered unit spectra and concentration profiles do not make chemical sense.

Figure 7 shows the concentration profiles obtained in the analysis of the eight runs of the process after applying constrained alternating least-squares optimization. Component one (solid line) rises very fast at the beginning of the process, decreases a bit afterward, and then remains stable until the process is terminated. Component two (dashed line)

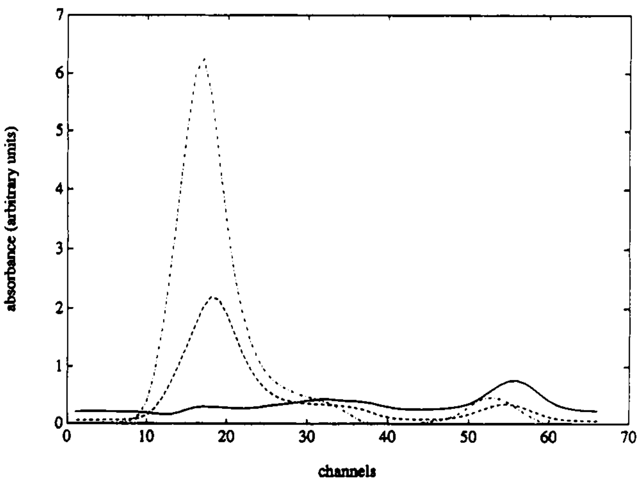

Figure 8. Unit (pure) spectra calculated by constrained alternating least squares of eight process runs.

does not appear at appreciable concentrations in all the runs. It is nearly nonexistent in process runs 1, 2, and 3, but it is the dominant contribution in process runs 4, 7, and 8. Component three is always present except in process run 8, showing always the same pattern of growth. The reason why this component does not appear in process run 8 is because the data analyzed pertain only to the first part of the process before it became present in appreciable concentrations.

The two constraints applied to the shape of the concentration profiles, nonnegativity and unimodality, have an important role during the optimization. It was more difficult to apply the unimodality constraint because of small random oscillations and changes in the concentration of the components during the process. This means that although the global shape must be unimodal and smooth, locally some small oscillations of the unimodality condition had to be allowed (see the concentration profiles of Figures 5 and 7).

Figure 8 shows the unit spectra of the three common components deduced from the analysis of the eight runs of the process. The shapes obtained for these three unit spectra explain very well the changes observed in the shapes of the raw process experimental spectra of every run. For instance, the first component only has the band around channels 50–60; this is in agreement with the first experimental spectra of every run which only show that band. Conversely, the second and third components have two bands approximately in the same locations but shifted between them. The strong absorption around channels 10–20 is more important for the third component than for the second, but the later becomes the dominant contribution in some experiments. This is in agreement with what is observed from the detailed comparison of the spectra of the different runs. Moreover, the recovered unit spectra show different baseline as a consequence of the observed baseline changes in the experimental spectra.

When the analysis is performed using the second derivative spectra, the results are very similar to the ones given in Figures 7 and 8, but with a poorer description of the concentration profiles in some parts of the process, especially for component 2 in the first three runs. The reason for this degradation of the resolution is because the nonnegativity constraint cannot be applied over the second derivative unit spectra.

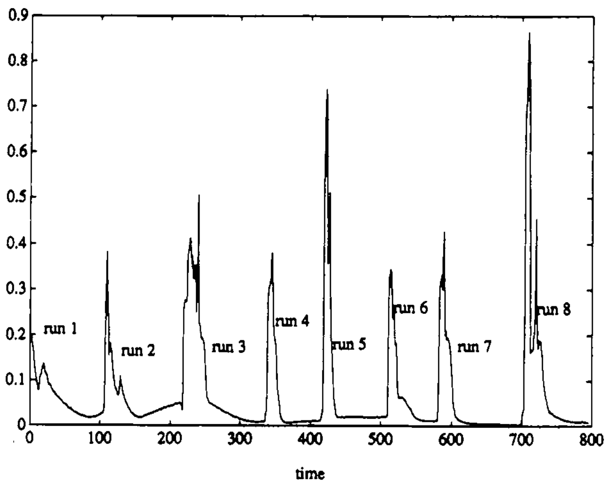

Figure 9. Relative standard deviations of the time residuals between the experimental data of the eight process runs and the recalculated values using the concentration changes and unit spectra given in Figures 7 and 8. *X*-axis is a time axis, and it gives an indication of the spectra number within the total number of spectra analyzed (795 in total).

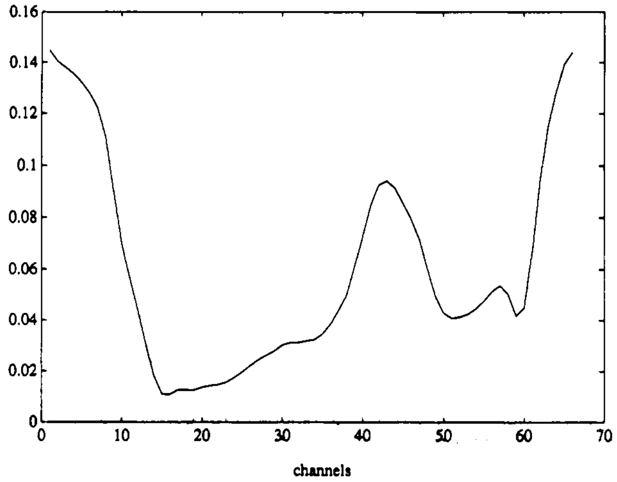

Figure 10. Relative standard deviations are given of the spectra channel residuals between the experimental data of the eight process runs and the recalculated values using the concentration changes and unit spectra given in Figures 7 and 8.

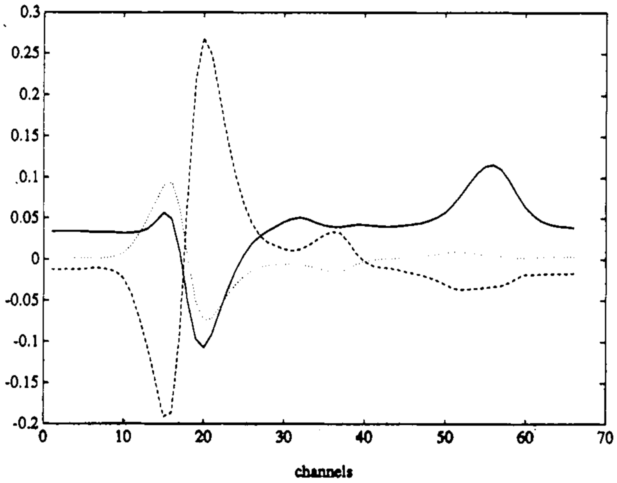

Figure 11. Prediction coefficients for the estimation of the changes in concentration of the three components during the process.

In Figures 9 and 10, the relative standard deviation of the residuals along both axes of measurement, spectra number and spectral channel, are given. These residuals correspond to the difference between the experimental data matrix and the recalculated data matrix using the set of concentration profiles and unit spectra derived from the alternating least-squares optimization (Figures 7 and 8). In order to get the relative values, the standard deviations of the residuals were divided by the maximum absorbance for every particular spectrum (Figure 9) or for every spectral channel (Figure 10), respectively. From Figure 9, it is seen that the fitting error is larger at the two extremes of every process, which is in agreement with the fact that at that point the conditions of the process are less under control (the process is terminated or initiated with a low absorbance value) and are not of interest for the present work. The spectral channels with larger relative standard deviations are the ones where there is almost no absorption of interest (Figure 10). This means that the experimental spectra are described appropriately by the three unit spectra obtained in the constrained least-squares optimization.

(6) Determination of Concentrations of Components at Any Stage of the Process from Its Time Spectrum. In Figure 11, the three sets of coefficients calculated from the unit spectra are plotted. The three have a spectral shape, each one with independent features. They do not describe the noise but the relevant changes in the spectra. When they are applied, the concentrations at any time of the process can be determined. In particular, if they are applied to the augmented data matrix with the eight process runs, the concentration profiles of Figure 7 are obtained. The concentrations predicted from these coefficients when applied to new spectra not used in the building of the model have to be compared in relation to those given in Figure 7 and not in absolute terms. In order to have absolute values, external calibration information must be provided.

CONCLUSIONS

The proposed method allows the description of the chemical changes produced during process runs. The method is suited to those cases where no previous information about the system is available and only a model-free approach can be used. The method takes advantage of the selectivity present in the system, as well as the information gained when several runs of the process under different conditions are analyzed simultaneously.

From the results obtained from the analysis of a selected group of process runs, it is possible to predict the course of future similar process runs in real time, thereby allowing optimal control.

At the present time, the method used requires decisions regarding noise levels, appropriate spectra shapes, and use of various constraints made by the analyst. Future work will be aimed at providing a more automatic and robust tool to the process analytical chemist.

ACKNOWLEDGMENT

R.T. has received financial support from the DGICYT (Spain) for research at the University of Washington.

RECEIVED for review November 30, 1992. Accepted April 13, 1993.