ELSEVIER

Chemometrics and Intelligent Laboratory Systems 46 (1999) 161–180

Chemometrics and intelligent laboratory systems

Translating third-order data analysis methods to chemical batch processes

Kenneth S. Dahl [a,*], Michael J. Piovoso [a], Karlene A. Kosanovich [b,1]

[a] *Central Research and Development, DuPont, Wilmington, DE 19880, USA*
[b] *Department of Chemical Engineering, University of South Carolina, Columbia, SC 29208, USA*

Received 20 May 1998; accepted 7 August 1998

Abstract

Measurements collected from batch processes naturally produce a third-order or three-dimensional data form. The same structure also results when multiple samples are measured using hyphenated analysis techniques such as liquid chromatography with diode array detection. Analysis of third-order data by principal components analysis (PCA) is achieved by a nonunique rearrangement that produces a two-dimensional array. This preferentially models only one of the three orders present. In contrast, methods such as parallel factor analysis (PARAFAC) apply a particular decomposition that accounts for all three orders explicitly. The results from either approach should be related if data are to be interpreted reliably for applications to batch processes such as on-line monitoring and control. This work compares these two approaches from an applied point of view. To accomplish this objective, exemplary methods are selected from each type of analysis, parallel factor analysis (PARAFAC) and multiway principal components analysis (MPCA). These are employed to analyze data obtained during the manufacture of a condensation polymer in an industrial batch reactor.

Keywords: Multivariate analysis; Principal components analysis; Parallel factor analysis; Batch processes; Two-factor degeneracy

Contents

1. Introduction 162

2. Methods 163
   2.1. MPCA 163
   2.2. PARAFAC 165

3. Process description 166

* Corresponding author. Fax: +1-302-695-9811
[1] Fax: +1-803-777-8265.

0169-7439/99/$ - see front matter
PII: S0169-7439(98)00183-X

4. Results and discussion 168
4.1. Batch order 171
4.2. Variable and time orders 174

5. Summary and recommendations 177

Acknowledgements 180

References 180

1. Introduction

Extracting greater value from the extensive investment in process control technology is one of the challenges facing the manufacturing organizations of many modernized industries. Hardware and software used to gather, process, and store signals have not only become affordable, they have also increased in power and capacity (cf. the cost and sophistication of a desktop PC). With the size of the collected database growing nonlinearly, efficient methods must be found to analyze data and information. In addition, these methods and tools should yield results in real time if effective use is to made of them in a manufacturing environment.

Batch and semi-batch processes are used widely in the chemical, pharmaceutical and agricultural industries. Their flexibility to produce high-value products during short manufacturing campaigns accounts for their extensive use. The manufacture of a typical batch involves charging ingredients to the vessel, processing them under controlled conditions, and discharging the completed product. Consistency of the product is the implicit goal in repetitive operation, so minimizing batch-to-batch variability is imperative. At present, this is controlled through a plethora of simple methods, such as basic servocontrol of intrinsic measurements, to the more advanced, programmable logic controllers which enable precise sequencing of the operational steps.

Even with a high degree of automation, control of a batch process is quite challenging. This is attributed to the combination of their finite duration, nonlinear behavior, their natural nonstationarity, and multiobjective criteria. To confound the problem, insufficient in-line instrumentation is the norm rather than the exception, thus crucial information about the progress of the batch is not known in a timely manner. Feedback control of the product composition in the same sense as continuous processes is not possible (dynamic compensation), thus some form of statistical quality control is practised (steady-state compensation) which may or may not reduce batch-to-batch variability consistently and reliably.

Methods and tools for extracting information from data using multivariate statistical analysis (e.g., chemometrics) have grown rapidly in number and capability over the past decade. For batch processes, these tools include multiway principal components analysis or MPCA [1,2] and multiway partial least squares or MPLS [3]. As an example, Kosanovich et al. [4] used MPCA to analyze batch process data variability that resulted from heat transfer limitations. Compared to MPCA and MPLS, other bilinear and trilinear methods that are used routinely to convert data from hyphenated analytical instruments to analyte concentrations (e.g., for liquid chromatography with UV–visible diode array detection, or LC-DAD) can produce seemingly similar analyses [5].

A comparison of two methods for analyzing third-order data, MPCA and PARAFAC, is presented in this work. Our goal is to determine whether these methods can be used reliably in a manufacturing environment. We have chosen an application approach for this comparison. The specific application considers a homologous set of temperatures obtained during manufacture of a condensation polymer in a batch reactor. Thus, these data are a close analogue to the form of spectral data to which decomposition methods like PARAFAC have been applied. This is not intended to be an exhaustive comparison (e.g., direct comparison of computational speeds for the two

methods is not addressed explicitly), so the conclusions reached here should not be viewed as definitive. The paper is organized as follows. First, the assumptions upon which the methods were developed are presented; next the problem to which the methods are applied is introduced in sufficient detail to demonstrate that there is no bias intended to cause the analysis from one method to be favored over the other. Following this, results of applying the methods are discussed and conclusions are drawn where possible. Lastly, recommendations are offered concerning the use of the resulting models in the vital activities of detection, monitoring, and control of real-time operations.

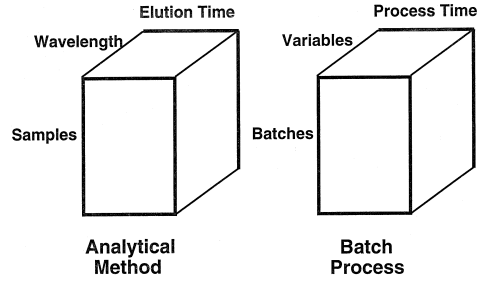

Fig. 1. Examples of third-order data arrays from a second-order analytical method (e.g., LC-DAD) and a batch process. Each sample measured by the analytical method yields a two-dimensional data matrix of absorbance readings modulated by elution time. Each batch produced yields a data matrix of process sensors' readings at discrete intervals over the course of the batch. The third order is obtained by considering data from multiple samples or multiple batches during the modeling process.

2. Methods

When two analytical techniques are combined or linked to measure a sample they create a second-order instrument [6,7]. An advantage to this approach is improved detection and quantification of desired analytes in the presence of components not accounted for in the instrument's calibration model. The measurements obtained from a single sample form a two-dimensional matrix, with each method contributing to the final signal. Using coupled LC-DAD as an example, the chromatography column changes the quantity of each analyte seen by the diode array detector over time, yielding new measurements at each channel (wavelength) at each sample time. Let the columns of the two-dimensional data array represent elution times and the rows successive channels (wavelengths) in the instrument's data matrix. Thus, each column in the matrix is the spectrum of the eluted material at a specific time while each row is the elution time profile at a specific wavelength. A similar representation can be made for process data (sensor measurements) from a single batch. Each column in this matrix contains the measured values from all sensors at a specific sampling time and each row holds the temporal information from each sensor over the time required to produce the batch.

When different samples are measured on the LC-DAD instrument, or measurements from new batches are sampled, the pooled information forms a three-dimensional array (see Fig. 1). Methods such as MPCA and PARAFAC differ in their basic form of analysis and development of the model for these 3-D arrays.

Although different terminology has been used to refer to the axes of the 3-D array, the term 'orders' will be used in this work. Additional notations used are as follows. A 3-D array composed of 2-D matrices is denoted in bold, underlined upper case letters; 2-D matrices are denoted by bold upper case letters; vectors are bold lower case letters and by convention are column vectors; row vectors are the transpose ($'$) of column vectors; and scalars are italicized letters, with i, j, and k reserved as running indices along individual orders.

2.1. MPCA

Multiway PCA (and multiway PLS) rearranges the 3-D array into a 2-D matrix that can be modeled using the conventional principal components analysis method (or partial least squares method). Two of the three orders are not modeled independently as a result, so these methods have been termed 'weakly multiway' [8]. The array is rearranged by taking 2-D slices along one of the orders (Fig. 2). Order 1 is referenced by index i ($i = 1, 2, \ldots, I$), order 2 by index j ($j = 1, 2, \ldots, J$), and order 3 by index k ($k = 1, 2, \ldots, K$). Fig. 2 shows a rearrangement in which slices are taken along order 3 and arranged to have order 1 modeled independently of the other two. This is not the only way to order 1 independently:

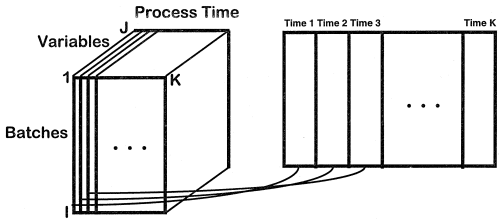

Fig. 2. Rearranging the third-order data array into a two-dimensional form for analysis by MPCA. Readings from all columns along the variable order at each interval along the process time order are considered as a blocks. These two-dimensional blocks are laid out, one after another, to form an extended two-dimensional matrix. With I batches, J variables and K times in the array, the matrix has dimensions of $(I \times JK)$.

slices can be taken along order 2 also. Overall, there are six possible rearrangements of the 3-D array into 2-D forms. To analyze batch-to-batch variability, order 2 represents process sensor values and order 3 their temporal profiles, leaving order 1 allocated to individual batches [1,2]. Thus, the 3-D array is converted to a 2-D matrix that is $I \times JK$ in size, as illustrated in Fig. 2.

For principal components analysis (PCA) to be an effective data analysis tool, two crucial assumptions about the character of the data have to be satisfied. The first is that if the data contain high frequency components then some prefiltering is necessary so that low frequency information is retained. The second is that since MPCA and its variants are linear methods, data with a significant amount of nonlinear information must be pretreated to eliminate or reduce this effect. This is accomplished by a centering operation about the mean of each of the JK columns in the 2-D matrix.

The resulting array, $\underline{\underline{X}}$, is decomposed as sums of the products of scores and loadings for each component in the model ($\boldsymbol{u}_m$ and $\mathbf{V}_m$, respectively) and of the array of residual errors, $\underline{\underline{E}}$, not accounted for by the M components in the model:

$$\underline{\underline{X}} = \boldsymbol{u}_1 \otimes \mathbf{V}_1 + \boldsymbol{u}_2 \otimes \mathbf{V}_2 + \ldots + \boldsymbol{u}_M \otimes \mathbf{V}_M + \underline{\underline{E}}, \quad (1)$$

where $\otimes$ denotes the Kronecker product of $\boldsymbol{u}_m$ with $\mathbf{V}_m$. Fig. 3, top, shows the terms of this equation in graphical form (Smilde [11]). The interdependence of orders 2 and 3 is represented by the loadings' matrices, $\mathbf{V}_m$, while the independence of order 1 is shown as the scores' vectors, $\boldsymbol{u}_m$.

It is important to recognize that the $J \times K$ elements in each loadings *matrix* $\mathbf{V}_m$ are obtained via the PCA computation as a loadings *vector* of dimension $JK \times 1$. For an MPCA model that includes two or more components, these loadings vectors are orthogonal to each other. When a loadings vector is rearranged into a matrix of dimension $J \times K$, each row represents one sensor over the time span of a batch and each column covers all sensors at a single time during manufacture of the batch. These columns and rows can be compared among themselves for each component and between components. This permits both qualitative and quantitative assessments of how the variations among the different sensors' data contribute as a group to the individual directions of variability [4]. Note, however, that any single column or row from one components' loadings matrix $\mathbf{V}_m$ is *not* orthogonal to its corresponding column or row in another components' loadings matrix. That is to say, loadings extracted either by variable order or by time order are not necessarily orthogonal, but the scores vectors corresponding to the batch order are. The reason for pointing out this distinction will become

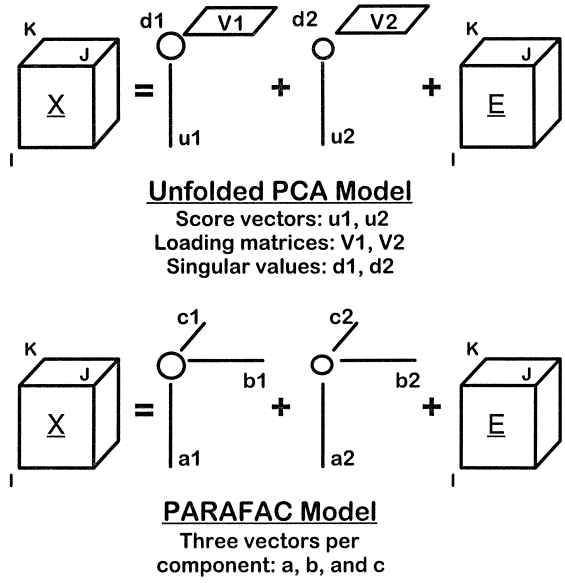

Fig. 3. Graphical representations of the MPCA (unfolded PCA) and PARAFAC models (after Smilde [11]). Readings in the X array have already been centered in this example. The size of the circles located at the junction of the vector and matrix for each dyad (top) and the three vectors that comprise each triad (bottom) represent the relative magnitudes of the successive factors.

clear when results from the PARAFAC models are presented.

2.2. PARAFAC

An extensive body of literature describes the development and evolution of three-way factor analysis and three-way generalizations of PCA and regression. Papers in the book edited by Law et al. [9] discuss these topics from the perspective of the psychometrics school. Geladi [10], Smilde [11], and Henrion [12] provide useful reviews of the differences between the MPCA procedure and that of PARAFAC. Bro [13] has provided a tutorial on the latter method. Leurgans and Ross [14] applied the PARAFAC model in a spectroscopic context where an underlying chemical and physical basis could be assigned to the form of the obtained model (i.e., their model was 'hard'). The essential features of the PARAFAC model as discussed here reflect the work of Geladi [10].

The 2-D matrix that consists of sampled measurements per batch is an order 2 tensor. This tensor can be decomposed into the outer products of two vectors, or dyads:

$$\mathbf{X} = \boldsymbol{t}_1 \boldsymbol{p}_1' + \boldsymbol{t}_2 \boldsymbol{p}_2' + \ldots + \boldsymbol{t}_R \boldsymbol{p}_R' + \mathbf{E} = \mathbf{TP}' + \mathbf{E}. \quad (2)$$

The outer product of vectors $\boldsymbol{t}_r$ and $\boldsymbol{p}_r'$, a row vector, is the dyad, $\boldsymbol{t}_r \boldsymbol{p}_r'$. The minimum number of dyads that can represent an order 2 tensor is its tensor rank or simply, rank, which is the minimum of the independent rows or columns [10]. Here, R is the rank of $\mathbf{X}$ when $\mathbf{E}$, the matrix of residuals, is not the zero matrix.

A 3-D data array, $\underline{\mathrm{X}}$, is an order 3 tensor. By analogy, an order 3 tensor can be defined in terms of a sum of triads, or Kronecker products of three vectors:

$$\underline{\mathrm{X}} = \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{c}_1 + \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{c}_2 + \ldots + \boldsymbol{a}_R \otimes \boldsymbol{b}_R \otimes \boldsymbol{c}_R + \underline{\mathrm{E}}. \quad (3)$$

$\underline{\mathrm{X}}$ is $I \times J \times K$, as is $\underline{\mathrm{E}}$, which is the array of residuals not accounted for by the R triads of the *trilinear* model. The tensor rank of $\underline{\mathrm{X}}$ is the minimum number of triads, R, necessary to describe it. Fig. 3, bottom, shows the terms of Eq. (3) in graphic form [11]. The independence of all three orders is shown by the vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$, and $\boldsymbol{c}_r$.

The results obtained by PARAFAC's direct decomposition of $\underline{\mathrm{X}}$ possess the intrinsic axis property [15]: the obtained axes are *uniquely* oriented relative to the configuration of the original data. For any PCA solution $\mathbf{X} = \mathbf{TP}'$, there exists a large class of nonsingular matrices $\mathbf{A}_n$ such that $\mathbf{X} = \mathbf{TA}_n\mathbf{A}_n'\mathbf{P}'$ (here $\mathbf{X}$ is the unfolded, 2-D form of $\underline{\mathrm{X}}$). Each member of $\mathbf{A}_n$ represents a linear transformation of the original solution, $\mathbf{TP}'$. Each of these transformed solutions has the same fitted values and residuals. A PARAFAC model of the same data yields a unique orientation of the factor axes, thus eliminating the need for additional factor rotation processes [16]. In fact, rotating a PARAFAC solution leads to loss of fit [13]. A data set that can be modeled adequately with PARAFAC can also be modeled with two-way PCA. The PARAFAC model requires fewer parameters: $(I + J + K)$ for each factor as compared with $(I + J \times K)$ for each component in two-way PCA [13].

Direct decomposition to represent a 3-D data array by PARAFAC may encounter two-factor degeneracies [17]. Large positive or negative correlations, with magnitudes of 0.9 or greater, are sometimes seen between some columns of the PARAFAC loading matrices. Characteristics for two-factor degeneracies mentioned by Kruskal et al. [17] include:

- high correlations between the two factors in all three orders,
- signs for the three correlations that are either all negative or have two positive values and one negative value, and
- magnitudes of the two factors that are almost equal and are very large.

Mitchell and Burdick [18,19] examined the connection between the presence of two-factor degeneracies and slowed progress toward a converged solution by the PARAFAC method. When a two-factor degeneracy is encountered, they advocate using an eigenanalysis-based procedure to choose different starting estimates for the PARAFAC computation and adjusting the convergence criterion to see whether a nondegenerate solution can be found. Even with these practices, 7 out of 64 arrays of fluorescence excitation and emission profiles that they studied did not reach a nondegenerate PARAFAC solution [19]. While not discussing two-factor degeneracy per se,

Harshman and Lundy [16] cite the practice of running a PARAFAC computation several times using different random starting values to start the algorithm. This practice can be useful for establishing whether a particular initial choice falls into a local minimum. It appears that no preanalysis can say conclusively whether there will be a two-factor degeneracy or not. Sensitivity to initial conditions may imply that the data to be modeled may not be trilinear or that the problem is ill-posed.

As is the case for MPCA, effective analysis of data by PARAFAC requires that the data be as linear and as static (low frequency) as possible. Additional centering along individual orders (additive adjustment) or scaling to unit variance (multiplicative adjustment) should be done on a case-by-case basis [13,15]. Scaling, in particular, will affect the form and information content of the results for multivariate models [20–22]. Preprocessing of the batch data by mean centering is all that is necessary for this comparative study. Finally, a point regarding nomenclature: in PARAFAC, preference is not given to any one order in the development of the model. Thus, rather than single out any order by referring to its weights as scores, the weights for all orders are referred to as loadings [16].

Having presented the bases for MPCA and PARAFAC, it is important to note that both methods are used in this study to obtain parsimonious descriptions of the batch reactor's temperature data. In contrast to the work of Leurgans and Ross [14], the PARAFAC models obtained here are 'soft': no chemical or physical model is assumed prior to the use of the method. The purpose behind both methods' use is to be able to identify underlying phenomena responsible for features in the data and to see how these phenomena appeared in their models' results. Based on the PARAFAC solutions not only being rotationally invariant but also explicitly modeling each of the three orders, this approach could be expected to provide more unambiguous information than MPCA for the data studied here.

3. Process description

The autoclave in this chemical process converts the aqueous effluent from an upstream evaporator into a polymer product. At least half of the cycle time is required to vaporize the incoming water from the charge, so the reactor is designed to maximize heat transfer from the heating fluid through both an external jacket and internal coils. Because mixing occurs naturally once boiling takes place throughout the autoclave, no mechanical agitation is used.

The feed to the reactor is an aqueous salt mixture whose water content is approximately 20% by weight. The temperature of the feed is considerably lower than that of the reactor, because the latter is hotter from having just completed the previous batch. This can be seen as the difference between temperatures, T_{pc}, at minutes 1 and 7 in the example operating profile shown in Fig. 4. The reactor operates in the first stage of its complete cycle as a closed pressure vessel until the pressure vent is opened to release steam, starting its second stage (i.e., see P_{rb} at 10 min, approximately, in Fig. 4). Steam is the primary product of the boiling phase and a by-product of the condensation polymerization.

One batch takes approximately 120 min to complete the cycle from introduction of the salt mixture through discharge of the polymer product. The recipe specifies reactor and heat source pressure trajectories through five stages. The pressure profiles during the first two stages are chosen to minimize loss of a volatile reactant. The pressure reduction during the

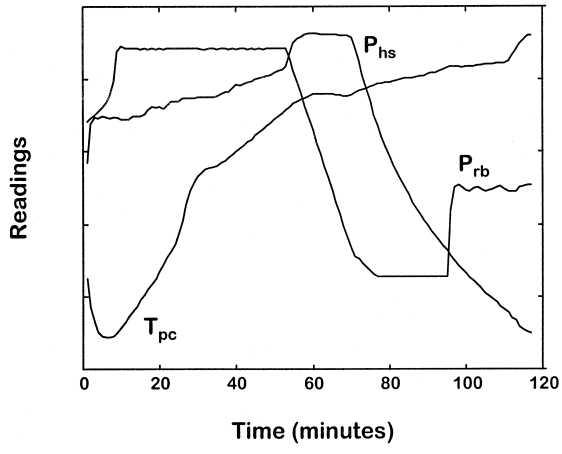

Fig. 4. Representative process trajectories from a typical batch. P_{hs}: heat source supply pressure, T_{pc}: autoclave center temperature, and P_{rb}: autoclave body pressure. Transition points between stages are evident in the trajectory of autoclave body pressure at 10, 55, 75 and 95 min, approximately.

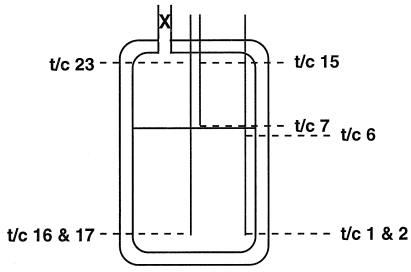

Fig. 5. Schematic diagram of the commercial autoclave showing locations of thermowells from which temperatures were obtained in this study. Sensors 1 through 6 were distributed within the thermowell near the vessel's inner wall and reported liquid phase temperatures only. Sensors 7 through 15 were located in a thermowell that spanned only the vapor phase near the center axis. Sensors 16 through 23 were arranged in another thermowell near the center axis and sampled both the vapor and liquid phases.

third stage is gradual to allow bubbles to disengage smoothly from the high viscosity melt (see P_{rb} between 55 and 75 min, approximately, in Fig. 4). Close temperature control of the reacting mass is necessary to assure a high degree of polymerization. This also avoids thermally induced side reactions which degrade the polymer.

Key process checkpoints (e.g., attaining a specific pressure or temperature within a given time) determine when one processing stage ends and the next one begins. In the first two stages of the batch cycle (i.e., between 1 and 55 min, approximately, in Fig. 4), heat is supplied initially to concentrate the reactants by removing as much of the water as possible and to provide the activation energy to start the polymerization reactions. Heating is stopped once polymer-forming reactions are self-sustaining after minute 75, approximately, in Fig. 4. For the fourth and fifth stages (i.e., after minute 75 in Fig. 4), the system operates adiabatically. Heating medium vapors in the external jacket and internal coils continue to facilitate heat transfer, either to the reactor or from it. This nonregulated action adds an important source of variability.

The polymer is discharged in the fifth stage by pressurizing the reactor (refer to P_{rb} starting at minute 95, approximately, in Fig. 4). Full discharge typically is complete after 15 min. Polymerization is quenched by cooling the extruded polymer as it is cut into small pellets.

To minimize data preprocessing, a homologous set of measurements comprised of only temperatures measured inside the autoclave are used in this analysis. Had pressure readings available been included, for example, it would have been necessary to scale the data to unit variance after mean centering due to the different measurement scales between the temperatures and pressures. Temperatures are measured in three separate thermowells inside the autoclave, shown schematically in Fig. 5. Twenty-three temperature measurements of the liquid and vapor phases constitute the data set. From operating experience, the locations of the sensors reported on phenomena that occur in distinct zones along the autoclave's vertical axis. Assignments of sensors to the four zones: lower liquid, upper liquid, lower vapor, and upper vapor, are listed in Table 1. Readings from six locations (numbers 1, 16, 17, 18, 21, and 22) are averages over the 1-min sampling interval between recorded readings.

Table 1
The twenty-three temperature sensors studied mapped to the four vertical zones in the autoclave

Sensor zone	Thermowell		
	Side	Center #1	Center #2
Upper vapor		23	15
			14
		22∗	13
			12
			11
			10
			9
			8
Lower vapor		21∗	7
Upper liquid	6		
	5	20	
		19	
	4	18∗	
Lower liquid	3		
	1∗ and 2	16∗ and 17∗	

Columns on the right represent the three thermowells in which the sensors were located. Sensors located at nearly the same distance along the vertical axis are shown in the same row. Sensors numbered 16 and 17 were two different devices located at the bottom of one center thermowell. Readings for locations 1 and 2 were obtained from the same sensor, but recorded using two different sampling methods. Readings from locations marked with an asterisk were averaged over the 1-min interval between stored values. All other readings were the instantaneous values present at each minute-to-minute interval.

All other readings are the instantaneous values obtained at each sampling interval.

4. Results and discussion

Temperature data are obtained from 39 batches of the same polymer recipe. Readings are taken at 1-min intervals for each batch and span 109 min. Fig. 6 shows the mean temperatures for all twenty-three sensors over all the batches. Three distinct features are present. The first is that the temperature trajectories are nonlinear (or would require high-order polynomials to approximate them). Next is the divergence among the temperatures in the four zones during pressurization of the autoclave in stage 1 and the early part of stage 2 (i.e., through minute 30). The final feature is that temperatures in the two liquid zones converge near the 55 min mark. They remain this way until stage 5, when the temperatures in the upper liquid increase as sensors become uncovered when the molten polymer is expelled from the reactor.

A box and whisker plot of the residuals after centering the temperature readings to remove nonlinearity for all 23 sensors is shown in Fig. 7. It is important to recognize that this operation removes information regarding temperature differences from location to location that are apparent in Fig. 6. In their place, variations around each sensor's mean trajectory become the analyzed quantities. Readings for sensors in the vapor zones span larger ranges than those in the liquid zones. Process knowledge suggests that greater turbulence in the vapor phase is the likely cause for this difference.

The middle column in Table 2 shows the ranges over which transitions from one stage to the next occur in the 39 batches. All batches are aligned with the start of pressurization at time = 0. The widths of these ranges increase from 1 min for the stage 1 to 2 transition to 7 min for the transition from stage 4 to 5. In

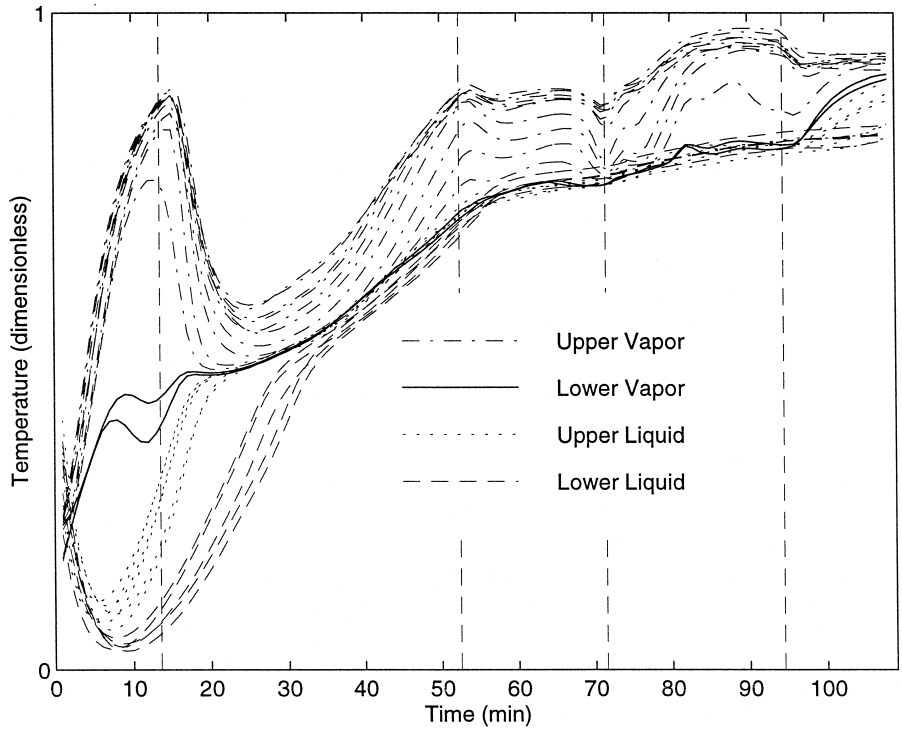

Fig. 6. Mean temperatures in the commercial autoclave from 23 temperature sensors located in four different zones. All batches were aligned to have the start of vessel pressurization as the reference point (time = 0) and then normalized by time stages. The different vertical zones in which the sensors are located are marked using different line types. The vertical lines mark the transition times from one stage to the next.

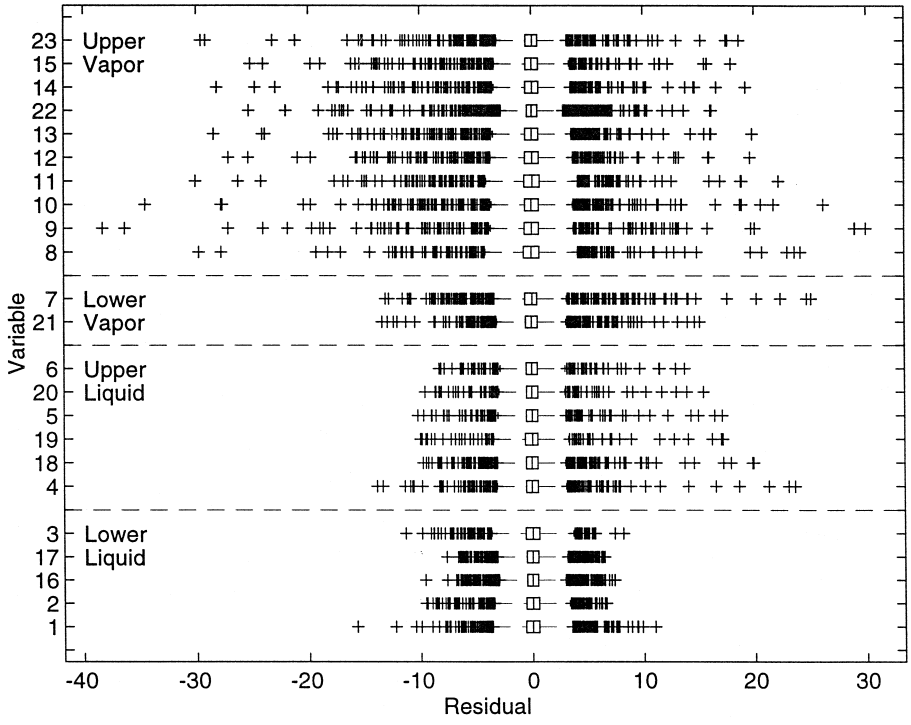

Fig. 7. Box and whisker plot showing the distribution of temperature residuals after centering, grouped by the four process zones. Each box has lines along the horizontal axis at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box by 2.5 times the range between the upper and lower quartile to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers and are marked with a + symbol. The dashed horizontal lines delineate bounds between the zones. The numbers for sensors in each zone, ordered from bottom to top, are listed along the vertical axis.

routine operations, the duration of each stage varies from batch-to-batch for several reasons. Prominent ones include delays in loading or unloading an autoclave and changes in stage times called for by the product quality control scheme. This shifting in the time order for process data is similar conceptually to retention time variations (shifts) in chromatographic methods, a phenomenon that has posed significant problems when constructing calibration models for hyphenated analytical methods such as LC-DAD [23]. (Shifts along the variable order for the process data will not occur unless a sensor's location in the thermowell changes, an extremely rare event.)

One way to correct for time shifting is to normalize the time intervals for each process stage. The duration of each stage is known for each batch, so a representative standard time can be selected for each stage. On a batch-by-batch basis, data are first interpolated and then resampled to populate the standard interval for each stage [5]. Times for transitions between respective stages after this normalization of the process data are shown in Table 2. Fig. 8 is an example that shows how normalizing the time intervals impacts the results from models. The first component loadings from MPCA models of the nine sensors in the upper vapor zone are shown for centered data as obtained (top) and after normalization (bottom). Less

Table 2
Starting and ending times, in minutes, at which transitions occurred between process stages in the data before and after normalization by time stages

Transition between stages	Readings (min)	
	'As received'	Normalized by stages
1 ⇒ 2	13.5–14.5	13.5
2 ⇒ 3	50.5–55.5	52.5
3 ⇒ 4	69.5–75.5	71.5
4 ⇒ 5	92.5–99.5	94.5

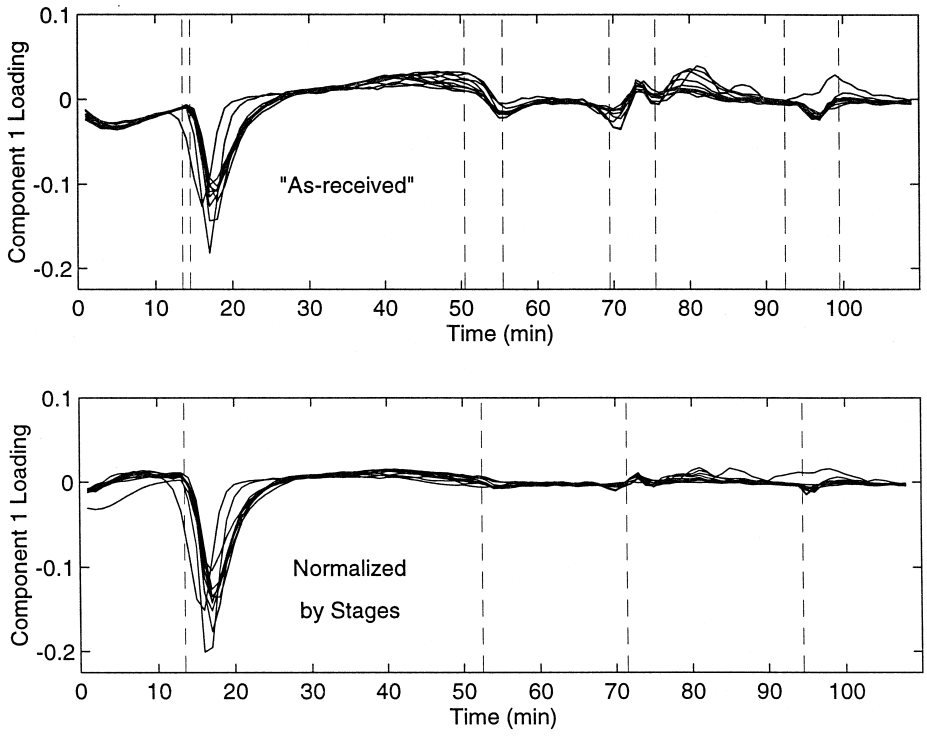

Fig. 8. Loadings for a one component MPCA model of the nine sensors in the upper vapor zone using data as received (top). The paired vertical lines mark the starting and ending times for the four transitions between process stages. Compared to loadings for a MPCA model with one component using data normalized by time stages (bottom), additional features beyond the evident peak between 15 and 20 min were more prominent at later times when the data were not normalized.

prominent features evident in the model of the data that has not been normalized are reduced significantly by normalizing to standard time intervals. These secondary features occur near the start and end points of the ranges for stage-to-stage transitions. The chief effect of the normalization focuses attention on the large loadings shortly after the start of stage 2.

The normalized and mean centered process data are modeled using routines written in MATLAB® (The MathWorks, Natick, MA, USA). The PCA routine from the PLS Toolbox® (Eigenvector Technologies, Manson, WA, USA) is used to synthesize the MPCA models. The relevant MATLAB® routines obtained from the shareware page [2] of the Royal Veterinary and Agricultural University's Department of Dairy and Food Science, Food Technology website are used to construct the PARAFAC models. Cross validation to establish the maximum number of components to include in the MPCA model is performed using The Unscrambler® software (CAMO ASA, Oslo, Norway). Eight components are selected based on the goodness-of-fit statistics from cross-validation (leaving one batch out each time) in the PCA model of the 2-D rearrangement of data used in the MPCA method. Since the objective is to compare the information content of the models developed using PARAFAC and MPCA, particularly along the time and variable orders, models containing up to eight (8) components (MPCA) or factors (PARAFAC) are synthesized. Procedurally, note that random values are used to seed the loadings for the PARAFAC algorithm, the practice recommended by Harshman and Lundy [16].

Percent variability explained by each principal component from the PCA model of the two-dimensional form of $\underline{\underline{X}}$ and by the cumulative number of factors used in corresponding PARAFAC models are

[2] http://newton.foodsci.kvl.dk/.

Table 3
Percent variance captured for the two-dimensional form of the time normalized and mean centered process data by components for the PCA model and by unconstrained PARAFAC models having the same number of factors

Component or factor number	Percent variance captured	
	MPCA	PARAFAC
1	45.6	33.4
2	67.6	53.3
3	75.7	(61.2)
4	81.4	(66.5)
5	85.5	(71.9)
6	88.3	(75.6)
7	90.2	(79.3)
8	91.4	(82.1)

Percent captured variance values in parentheses () in the PARAFAC models' column indicate that a two-factor degeneracy was present in the model.

listed in Table 3. Differences in the percent variability explained between the MPCA and PARAFAC methods reflect the objectives for the two methods. MPCA seeks to maximize the amount of variance explained with each new component used. Note that all PARAFAC models using three or more factors encountered two-factor degeneracies. Constraining the PARAFAC solution has been used elsewhere to assist interpretability or to improve its stability. A common method used is to require that loadings for one of the three orders be orthogonal across all factors in the model [13]. PARAFAC models constrained to be orthogonal along each one of the three orders—the batch, variable, and time—were examined in this study. However, results from these will not be presented since no objective basis could be established for selecting one order to be orthogonal over any other order.

4.1. Batch order

Loadings along the batch order for the four factors that explain the most variance in the data of the

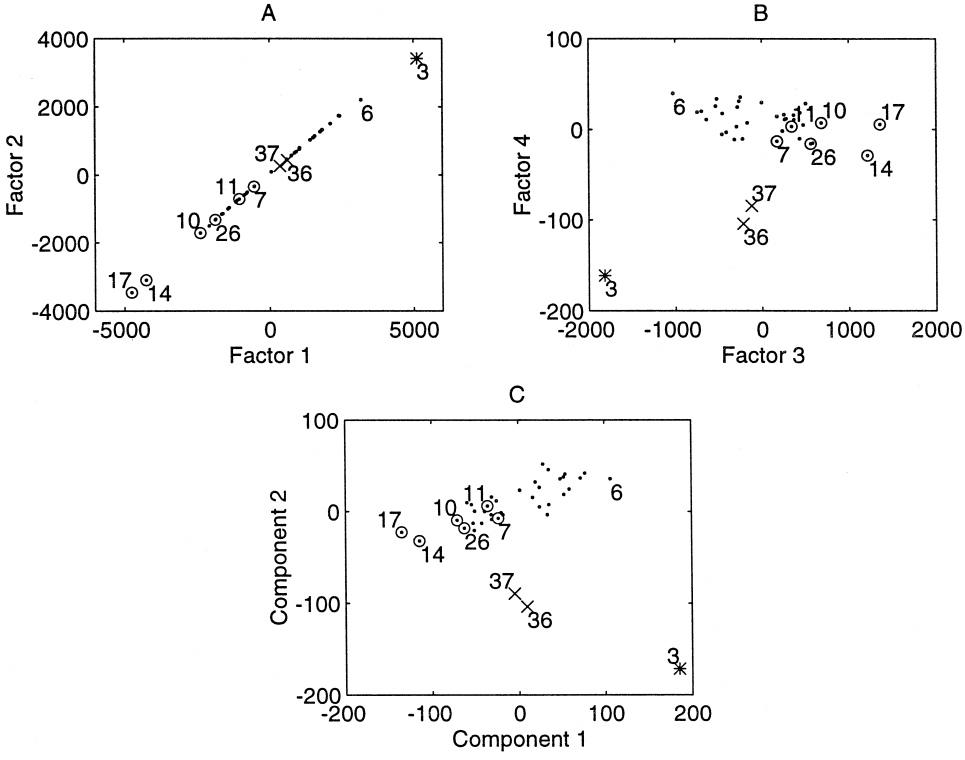

Fig. 9. Loadings along the batch order from the PARAFAC (graphs A and B) and MPCA (graph C) models. Eight factors are used in the PARAFAC model. The high positive correlation between factors 1 and 2 in graph A is due to the two-factor degeneracy in the PARAFAC model. Some batches are represented by their sequence number. The symbols used correspond to the number of idle time periods experienced by the batch prior to processing in the autoclave (see text) with ∗, ×, ⊙, and · standing for 7, 4, 1, and no periods, respectively.

eight-factor PARAFAC model and for the first two components from the MPCA model are presented in Fig. 9. A strong correlation exists between loadings from factors 1 and 2 in the PARAFAC model (Fig. 9A). This correlation is positive and, in conjunction with strong correlation between loadings for these same two factors in the other two orders, will be shown to be evidence that a two-factor degeneracy is present.

Observe that batches 3, 36 and 37 are separated from the remaining batches in Fig. 9B and C. Goodness-of-fit statistics for these batches confirmed their outlier status. Process data show that batch 3's temperature readings differ significantly from those of the other batches, especially during the first stage and first half of the second stage. Reduced temperature values at locations 7 and 21 in the lower vapor zone for this batch suggest that a larger quantity of liquid was present at its start than for other batches. Other information about these batches reveals that processing of the salt solution discharged from the upstream evaporator did not begin immediately for these three batches (and for several others). Operating practices are adjusted to compensate for these 'idle time' periods. In terms of time periods of constant length, batch 3 waited through seven (7) such periods, while batches 36 and 37 were held through four (4) periods. Batches 7, 10, 11, 14, 17 and 26 were idled for one (1) period. All other batches were processed immediately after discharge from the evaporator.

Using axes based in the space in which the data were obtained, Fig. 10 illustrates how different lengths of idle time periods are associated with batch-to-batch temperature differences during auto-

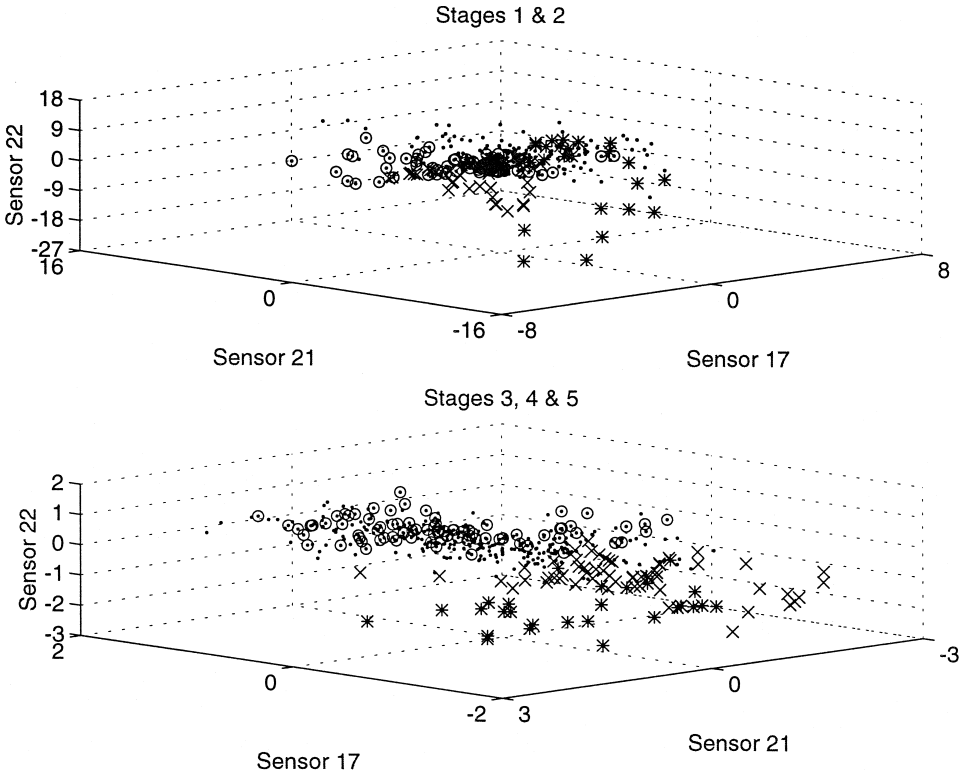

Fig. 10. Temperature residuals after centering for three out of the twenty-three measured temperatures, plotted at 2-min intervals by batch during process stages 1 and 2 (top) and stages 3 through 5 (bottom). Sensors chosen for the axes were in the lower liquid, lower vapor, and upper vapor zones (numbers 17, 21, and 22, respectively). The symbols used correspond to the number of idle time periods experienced by the batch prior to processing in the autoclave (see text) with $*$, $\times$, $\odot$, and $\cdot$ standing for 7, 4, 1, and no periods, respectively. Points plotted for idle periods 7 and 4 are from batch 3 and from batches 36 and 37, respectively. Data were plotted from only twelve out of thirty-six batches having idle periods of 1 or 0 to improve clarity.

clave operation. Observe that points from batches 3, 36 and 37 are separated from those of the other batches. The same effect of the idle time periods is seen in the relative locations of batch 3 and of batches 36 and 37 relative to other batches in the multivariate models' plots in Fig. 9B and C.

The assembled centered temperature data have a characteristic form and distribution in the space described by the original measurements' axes (see Fig. 10). Both modeling methods attempt to reconstruct this using different representations. MPCA is variance driven, that is, its aim is to account for as much variability as possible with each new component. The resulting successive components' scores and loadings are orthogonal to one another. Unconstrained PARAFAC does not have this aim, so the resulting factors' loadings used in the model usually will not be mutually orthogonal within any order. That is to say, a PARAFAC model's representational axes are oblique rather than orthogonal. Nevertheless, the optimization goals for these two methods are similar enough that the largest contributions to underlying variation in the original data appear in loadings' patterns for both multivariate models.

Loadings along the batch order for factors five through eight of the eight-factor PARAFAC model and for the third and fourth components from the MPCA model are presented in Fig. 11. The symbols × and + mark the two separate clusters evident in the plot of PARAFAC loadings for factors 7 and 8 (Fig. 11B). If points from batches that had delayed starts are ignored (i.e., the circled (○) symbols and batches 3, 36, 37), the same two clusters are seen in the MPCA model (Fig. 11C). The phenomenon responsible for this is identified in the subsequent discussion of the variable and time orders.

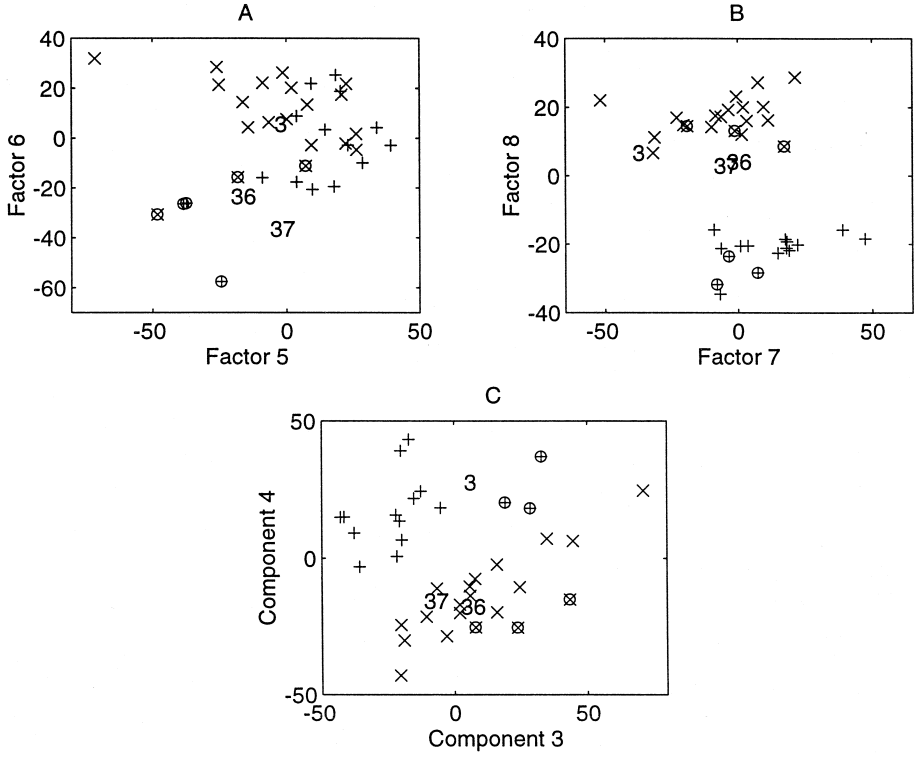

Fig. 11. Loadings along the batch order from the PARAFAC (graphs A and B) and MPCA (graph C) models. Eight factors are used in the PARAFAC model. Batch 3 and batches 36 and 37 corresponding to idle periods 7 and 4, respectively, are marked by their number. Circled points (○) mark batches having idle periods of 1. The symbols × and + mark batches belonging to the two distinct clusters seen in the PARAFAC model (graph B). Approximately the same clusters are seen in the MPCA model (graph C); this point is discussed further in the text.

4.2. Variable and time orders

Loadings along the variable and time orders from the eight-factor PARAFAC model are plotted in Figs. 12 and 13. Loadings for the first and second factors along the variable and time orders show strong correlations with each other, just as they did along the batch order in Fig. 9A. The correlation is negative for variables and positive for time in Fig. 12A and B, respectively.

Loadings from the MPCA models, grouped by process zones, are plotted in Figs. 14–17. The multiple curves in each plot represent loadings from each sensor vs. time. Loadings which differ visibly from others in a group are labeled with the corresponding sensor's number.

Considered as a whole, loadings along the time order in Figs. 12–17 usually have their largest magnitudes in process stages 1 and 2. This is true especially for the earliest factors (Fig. 12B and D) and components (Figs. 14 and 15). With a requirement to bring the autoclave pressure up to target value as rapidly as possible at the start of a batch (see Fig. 4), it is not surprising that variability is high during this time and that the models captured this phenomenon.

The rapid increase in the temperatures located in the vapor zone during pressurization, followed by their sharp drop once the vent valve is opened and steam is allowed to leave the vessel, can be expected to vary considerably due to natural turbulence that exists in the vapor zone. The magnitudes of MPCA loadings in the upper vapor zone are large throughout stage 1 for the second component (Fig. 15), when pressurization of the autoclave occurs, and during the opening minutes of stage 2 in the first component (Fig. 14). The first four factors in the PARAFAC model show the same phenomenon: time order loadings for the same four factors in the PARAFAC

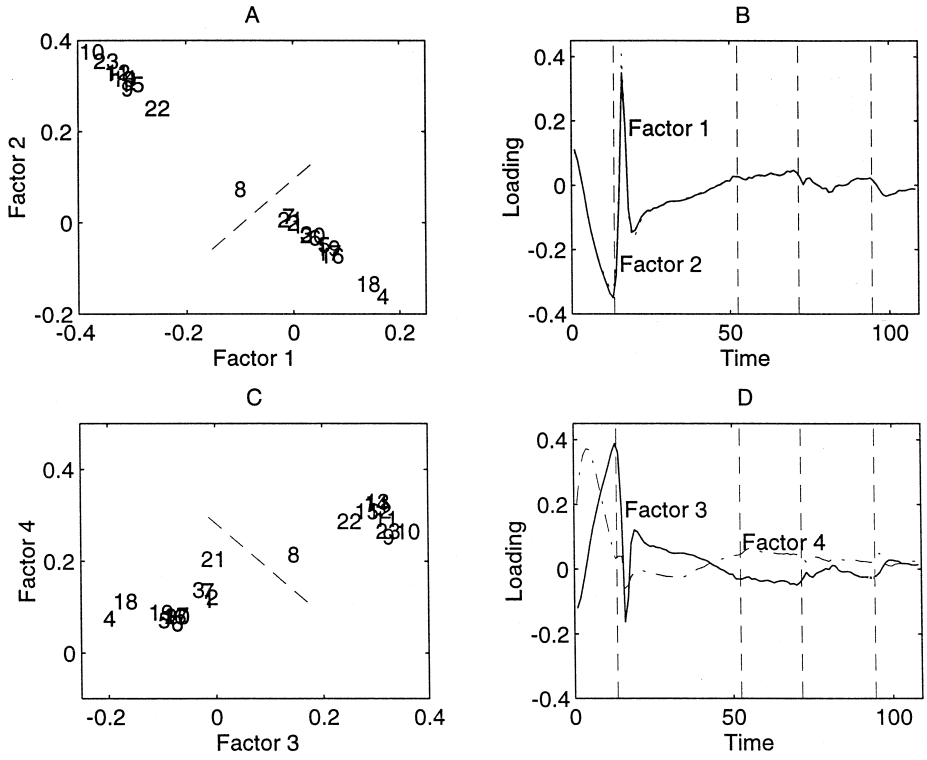

Fig. 12. Loadings for the first through fourth factors of the eight-factor PARAFAC model. Variable order loadings are in graphs A and C, where individual sensors are represented by the numbers. The dashed lines mark the boundary between sensors in the upper vapor zone and all other zones. Time order loadings are in graphs B and D, in which the vertical lines mark transition points between stages of the batch process. The high correlation between factors 1 and 2 (negative in graph A and positive in graph B) is due to the two-factor degeneracy in the PARAFAC model.

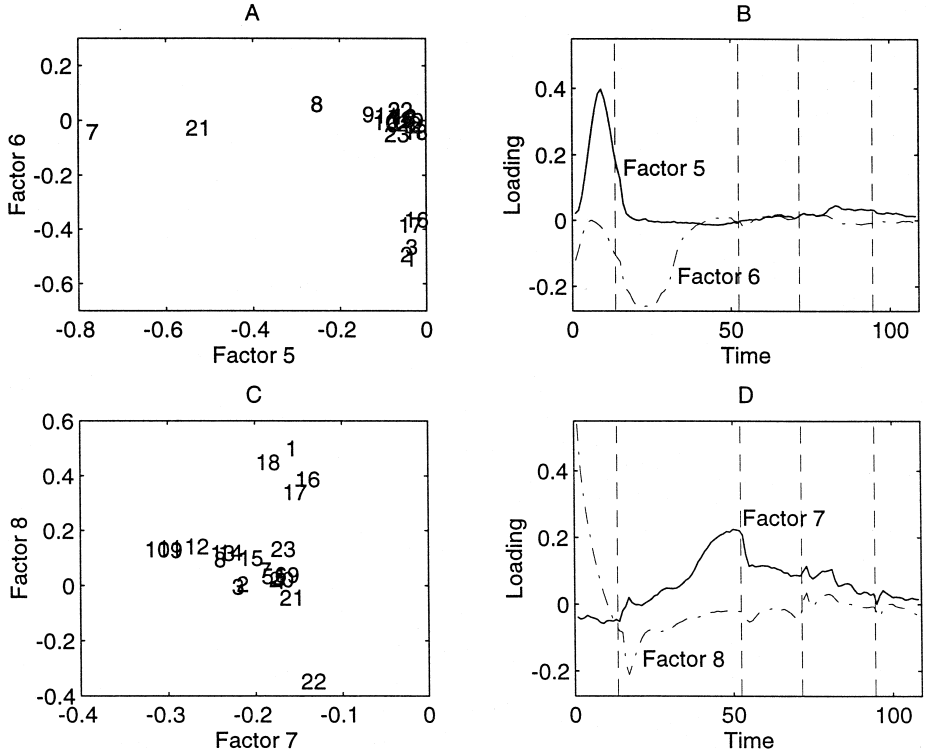

Fig. 13. Loadings for the fifth through eighth factors of the eight-factor PARAFAC model. Variable order loadings are in graphs A and C, where individual sensors are represented by the numbers. Time order loadings are in graphs B and D, in which the vertical lines mark transition points between stages of the batch process.

model have their largest magnitudes during process stage 1 and the start of stage 2 (see Fig. 12B and D). Upper vapor zone temperature sensors are separated from those in the other zones (see Fig. 12A and C). The exception is at location #8, which is the upper vapor zone sensor that is located closest to the lower vapor zone.

The two sensors located in the lower vapor zone are distinguished from others by the fifth factor in the PARAFAC model. Here, a separation occurs along the variable order in conjunction with large loadings during process stage 1 along the time order (see Fig. 13A and B). The third component of the MPCA model captures this phenomenon (Fig. 16). The probable reason for this behavior at the outset of the batch process is that the liquid level fluctuates in the autoclave as boiling begins during pressurization. By covering and uncovering the lower vapor zone's sensors, these level changes contribute to variability of signals from batch to batch.

Uneven heating in the large volume of liquid mass will contribute to greater variation during the early portion of each batch. Uniform boiling throughout the liquid is established only after approximately 35 min have elapsed. Loadings' magnitudes for the MPCA model are large in both zones of the liquid during process stages 1 and 2 (see Figs. 14 and 15). By process stage 3 and lasting through the end of the batch, MPCA loadings in the liquid zones for all four components shown are much smaller (Figs. 14–17). This reflects the effect of the regulatory control scheme to minimize temperature variations during the development of polymer chains (molecular weight). Factor 6 of the PARAFAC model is associated with the separation of lower liquid zone sensors from all others that occurs predominantly in process stage 2 (see Fig. 13A and B).

Significant nonzero loadings during stage 4 are evident in the first through fourth components of the MPCA model in the upper vapor zone (Figs. 14–17).

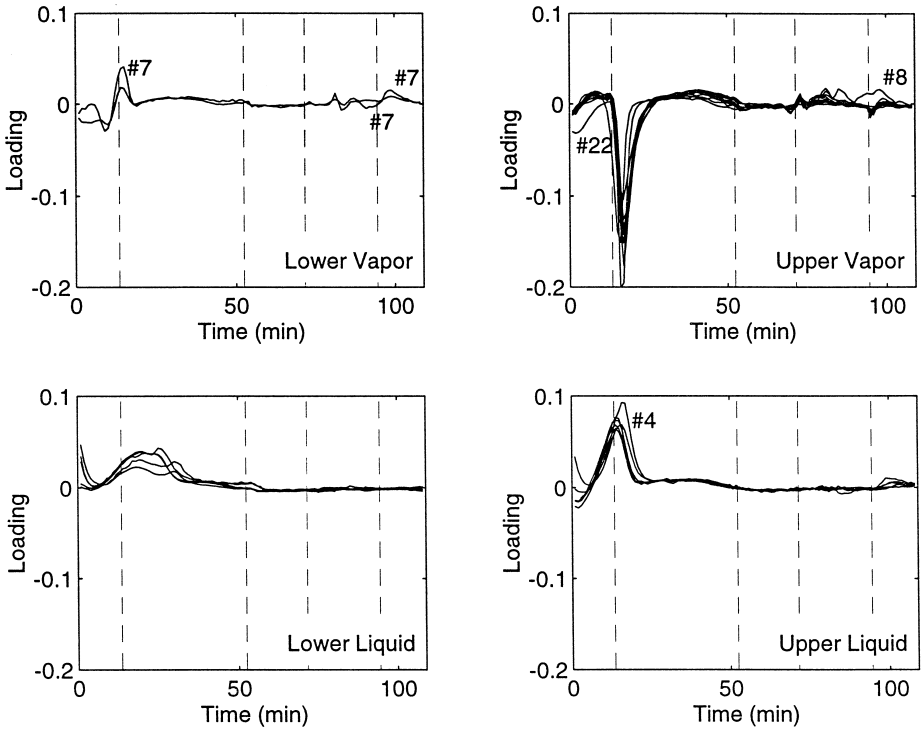

Fig. 14. Loadings for component 1 of the MPCA model, arranged by process zones. The labeled curves mark sensors whose loadings differ noticeably from others in the same zone. The vertical lines mark transition points between stages of the batch process. Loadings' magnitudes are greatest just after the stage 1 to stage 2 transition.

The liquid level apparently rose in response to the lower pressure in the autoclave during this process stage (see Fig. 6).

Interesting differences for MPCA loadings for all three orders are seen in the fourth component (Figs. 17 and 11C) and for factor 8 of the PARAFAC model (Figs. 13C, D and 11B). The six referenced loadings curves in Fig. 17 correspond to those sensors whose readings are averages obtained between intervals due to the process historian's sampling method. Five of the same six locations—1, 16, 17, 18 and 22—account for the spreading of the PARAFAC model's loadings for factor 8 in Fig. 13C. Differences in loadings between the these sensors and all others (which are sampled instantaneously) are prominent during process stages 1 and 2. Features of the respective signals were investigated to determine why this phenomenon occurred. As expected, rounding the averaged sensors' readings to the nearest whole unit did not eliminate the differences (the averaged readings had been recorded to the nearest hundredth of a degree). Auto-covariance analyses of the averaged signals against themselves and cross-covariance analyses with those of other sensors show that prior to averaging, the signal from these sensors had been filtered, thereby imposing an autocorrelation structure not seen in the instantaneous sensors.

Characteristics of two-factor degeneracies were discussed in Section 2.2. For the unconstrained PARAFAC model presented here, high correlations between loadings of factors 1 and 2 are found for all three orders (see Fig. 9A, Fig. 12A and B). The signs of the correlations are consistent with a two-factor degeneracy: two are positive and one is negative. Finally, magnitudes of the two factors' loadings in all orders are large and nearly equal, as required for such degeneracies.

In an effort to understand better the origin of the two-factor degeneracies, unconstrained PARAFAC models are developed to represent not only individ-

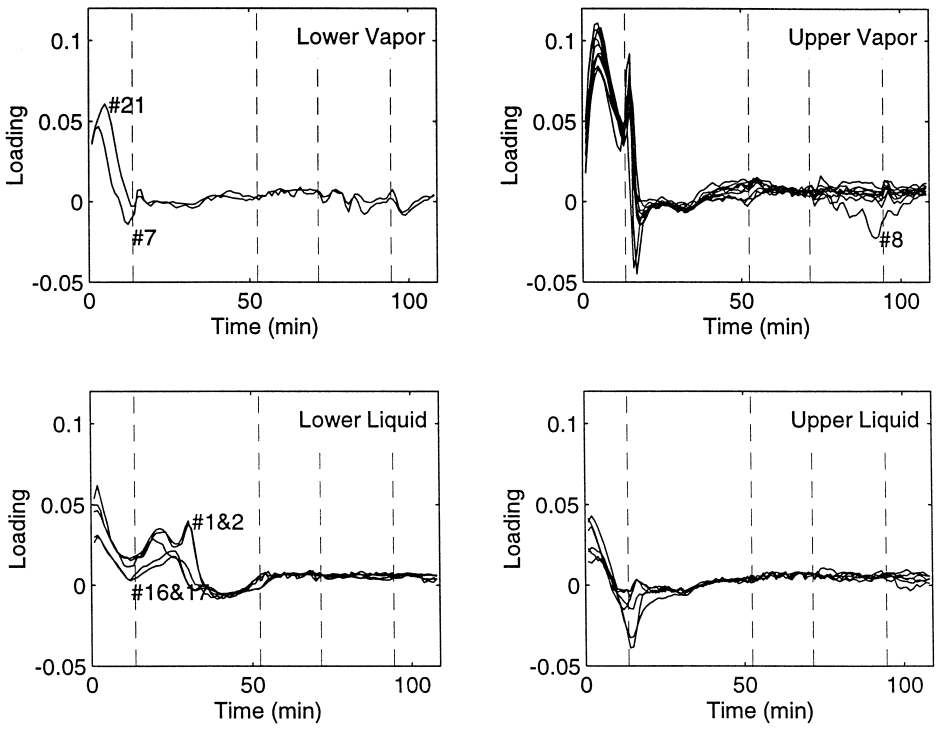

Fig. 15. Loadings for component 2 of the MPCA model, arranged by process zones. The labeled curves mark sensors whose loadings differ noticeably from others in the same zone. The vertical lines mark transition points between stages of the batch process. Loadings' magnitudes are greatest around the stage 1 to stage 2 transition for all zones.

ual stages of the process, but also adjacent stages. These models used from one to four factors. Table 4 shows the increases in explained variance as more factors are included for six different cases modeled by PARAFAC. Four of these are specific to process stages 1 through 4, considered individually. Only the four-factor stage 2 model and the three-factor stage 3 model encountered two-factor degeneracies. Of the two cases where data from adjacent stages are combined, PARAFAC models for stages 1 and 2 containing two or more factors encountered two-factor degeneracies. The presence of such degeneracies in PARAFAC models obtained by joining smaller subsets of data cannot be predicted from their presence or absence in PARAFAC models of the individual subsets.

5. Summary and recommendations

Based on the data set used, both the MPCA and PARAFAC models were able to resolve combinations of locations and times where significant variation occurred in the measured process temperatures. Both methods found more variability at the start of each batch, a finding that reflects the product's control goals. This was also consistent with process operations' knowledge, since the requirement to reach the autoclave pressure setpoint as quickly as possible almost always resulted in vigorous heating of the contents at the start of a batch.

The PARAFAC models' results along the variable order were striking: loadings were clustered distinctly for three of the four different process zones. Such classification is helpful in off-line analyses as were performed here, particularly when combined with process knowledge. While they reflected the same phenomenon, the assignment of locations to zones used for the MPCA loadings plots were first imposed from a priori process knowledge (i.e., the differences in mean trajectories evident in Fig. 6). For both models, the selection of some sensors by their sampling method pointed out the importance of

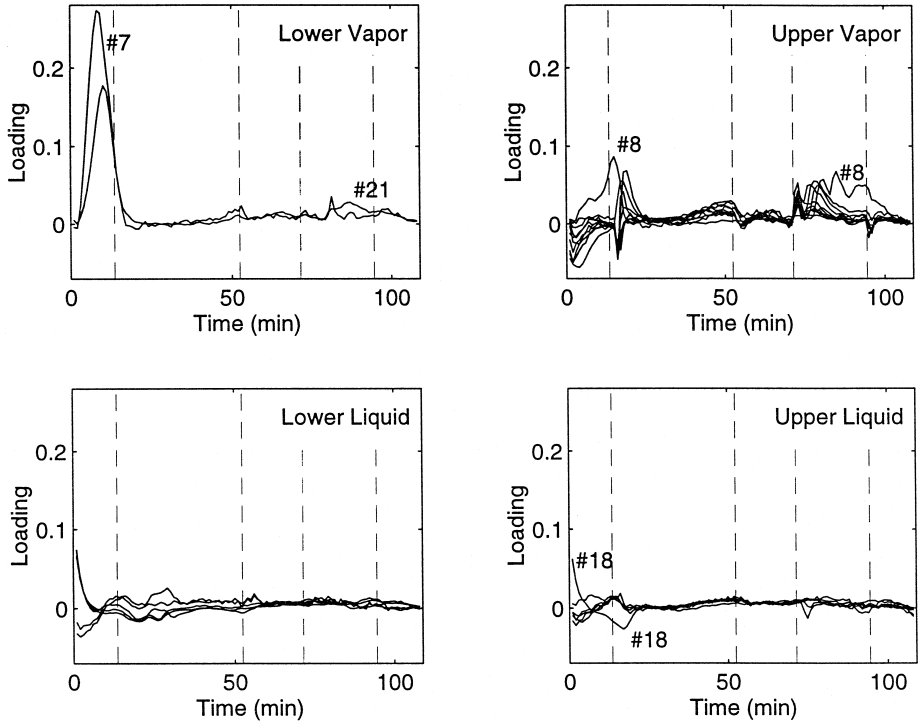

Fig. 16. Loadings for component 3 of the MPCA model, arranged by process zones. The labeled curves mark sensors whose loadings differ noticeably from others in the same zone. The vertical lines mark transition points between stages of the batch process. Loading magnitudes are greater in the vapor phase than in the liquid phase.

knowing what pretreatment occurs to process signals prior to their collection and storage. In this instance, process knowledge suggested that this phenomenon was *not* due to sensor location, so less obvious reasons for the differences needed to be found.

Normalization of the process data by time stages was shown to be necessary to avoid generating erroneous, large loadings in the models. Such loadings reflected the asynchronous operations from batch to batch rather than actual variation in the process at those times, judging by their diminished presence in the models after the data were normalized (see Fig. 8). Since it appears that normalization of the data is crucial for either modeling method to be effective (i.e., sensitive to true process variability), application of such models (and their product quality property-predictive counterparts) for monitoring or control must reflect this requirement. This imposes a limitation for real-time applications of these methods since all the data for an entire stage are required a priori for the normalization process. One means of circumventing this is to develop an approximate dynamic model of the process to predict the relevant batch data in time. With these data, the normalization process is iterative. That is, the predictions from the model are updated at each sample time; the normalization proceeds with the updated information to improve the predictions, and the MPCA and PARAFAC model development is repeated with these new normalized data. As a consequence, the monitoring and control activities are always improving as the batch proceeds.

From literature accounts, the occurrence of two-factor degeneracies is a difficult problem to resolve in the PARAFAC models. Whether using a trilinear model to describe the data in this study is not appropriate or whether a two-factor degeneracy cannot be avoided for a particular data set when three or more factors are used are questions that need to be addressed. However, their effect to complicate interpretations for this exploratory data analysis was less troublesome than their typical effect to lengthen the

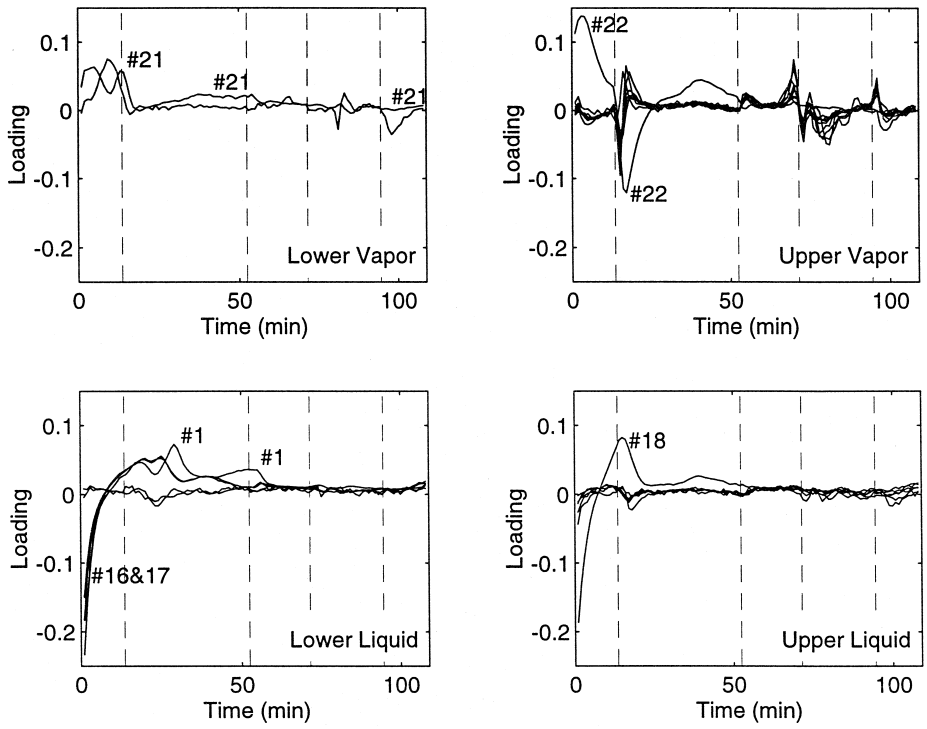

Fig. 17. Loadings for component 4 of the MPCA model, arranged by process zones. The labeled curves, whose appearances differ from others in each zone, are from sensors whose sampling method differed from that of all others (see text). The vertical lines mark transition points between stages of the batch process.

time dramatically for the method to converge to a solution. This increased time and the potential ambiguity introduced by two-factor degeneracies offset any advantage that PARAFAC attained from using fewer parameters to reconstruct the data. The suggestions of Mitchell and Burdick [19] to experiment with eigenanalysis-based procedures to obtain different starting points and to terminate the computations and start with new initial estimates when a two-factor degeneracy is encountered provide a useful guide for further work to apply PARAFAC methods on process data correctly.

The industrial community's manufacturing sector is interested in having data analysis tools be available that can condense and summarize the vast quantity of data it generates throughout its operations. To

Table 4
Percent variance captured by factors of PARAFAC models of different process stages for the time normalized and mean centered process data

Factor number	Percent variance captured					
	Stage 1	Stage 2	Stages 1 and 2	Stage 3	Stage 4	Stages 3 and 4
1	48.9	59.6	37.2	59.6	33.3	33.3
2	65.1	72.2	(58.8)	67.3	49.3	50.4
3	74.3	80.7	(65.3)	(72.2)	59.1	60.3
4	86.5	(85.0)	(72.4)	76.5	64.1	64.2
Initial sum of squared error	105,960	170,300	276,260	9146	15,112	24,258

No constraints were applied in the models. Percent captured variance values in parentheses () mark models in which there were two degenerate factors. The initial sum of squared error values are from the residuals obtained after mean centering.

be applied broadly and used to good advantage, such tools must be robust in operation and should minimize ambiguity of interpretation. A tool that is robust will produce a model that can accommodate on-line data that may include missing values, out-of-calibration sensors, and other common data-gathering problems and produce meaningful information. Or, failing in that, it will offer diagnostic information which will assist plant personnel to locate and fix conditions not previously encountered. By keeping ambiguity to a minimum, the data analysis tool will hasten the process of separating operational fact from experiential hearsay. Of the two methods compared here, MPCA is further down the path toward these goals than PARAFAC, due to it having been tested sooner in the manufacturing environment. Both warrant further study using data from actual operations so that ultimately they can be tested on-line and, it is hoped, be found useful to keep there.

Acknowledgements

Ricard Boqué kindly provided the MATLAB code for normalizing the process readings by time stages. Routines for creating PARAFAC models were developed by Rasmus Bro, who also provided helpful advice regarding their use when models contained highly correlated factors.

References

[1] P. Nomikos, J.F. MacGregor, Monitoring of batch processes using multi-way principal component analysis, AIChE Journal 40 (1994) 1361.
[2] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41.
[3] P. Nomikos, J.F. MacGregor, Multi-way partial least squares in monitoring batch processes, Chemom. Intell. Lab. Syst. 30 (1995) 97.
[4] K.A. Kosanovich, K.S. Dahl, M.J. Piovoso, Improved process understanding using multiway principal component analysis, Ind. Eng. Chem. Res. 35 (1996) 138.
[5] R. Boqué, A.K. Smilde, Monitoring and diagnosing batch processes with multiway regression methods, submitted to AIChE Journal, 1998.
[6] B.R. Kowalski, M.B. Seasholtz, Recent developments in multivariate calibration, J. Chemometrics 5 (1991) 129.
[7] K.S. Booksh, B.R. Kowalski, Theory of analytical chemistry, Anal. Chem. 66 (1994) 782A.
[8] B.M. Wise, N.B. Gallagher, Multi-way analysis in process monitoring and modeling, Chemical Process Control V, Tahoe, CA, 1996.
[9] H.G. Law, C.W. Snyder Jr., J.A. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984.
[10] P. Geladi, Analysis of multi-way (multi-mode) data, Chemom. Intell. Lab. Syst. 7 (1989) 11.
[11] A.K. Smilde, Three-way analyses. Problems and prospects, Chemom. Intell. Lab. Syst. 15 (1992) 143.
[12] R. Henrion, N-way principal component analysis. Theory, algorithms and applications, Chemom. Intell. Lab. Syst. 25 (1994) 1.
[13] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[14] S. Leurgans, R.T. Ross, Multilinear models: applications in spectroscopy, Stat. Sci. 7 (1992) 289.
[15] J.B. Kruskal, Multilinear methods, in: Law, Snyder, Hattie, McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984, p. 36.
[16] R.A. Harshman, M.E. Lundy, The PARAFAC model for three-way factor analysis and multidimensional scaling, in: Law, Snyder, Hattie, McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984, p. 122.
[17] J.B. Kruskal, R.A. Harshman, M.E. Lundy, How 3-MFA data can cause degenerate PARAFAC solutions, among other relationships, in: R. Coppi, S. Bolasco (Eds.), Multiway Data Analysis, Elsevier, Amsterdam, 1989, p. 115.
[18] B.C. Mitchell, D.S. Burdick, An empirical comparison of resolution methods for three-way arrays, Chemom. Intell. Lab. Syst. 20 (1993) 149.
[19] B.C. Mitchell, D.S. Burdick, Slowly converging PARAFAC sequences: swamps and two-factor degeneracies, J. Chemometrics 8 (1994) 155.
[20] P. Paatero, A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis, Chemom. Intell. Lab. Syst. 38 (1997) 223.
[21] R.A. Harshman, M.E. Lundy, Data preprocessing and the extended PARAFAC model in: Law, Snyder, Hattie, McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984, p. 216.
[22] S.N. Deming, J.A. Palasota, J.M. Nocerino, The geometry of multivariate object preprocessing, J. Chemometrics 7 (1993) 393.
[23] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Standardization of second-order chromatographic/spectroscopic data for optimum chemical analysis, Anal. Chem. 70 (1998) 218.