Improved Process Understanding Using Multiway Principal Component Analysis

Karlene A. Kosanovich*

Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208

Kenneth S. Dahl and Michael J. Piovoso

Central Science & Engineering, DuPont Company, Wilmington, Delaware 19880

Producing a uniform polymer by batch processing is important for the following reasons: to improve the downstream processing performance, to enable material produced at one site to be used by another, and to remain competitive. Eliminating the sources of batch-to-batch variability and tightening the control of key variables are but two ways to accomplish these objectives. In this work, it is shown that multiway principal component analysis (MPCA) can be used to identify major sources of variability in the processing steps. The results show that the major source of batch-to-batch variability is due to reactor temperature variations arising from disturbances in the heating system and other heat-transfer limitations. Correlations between the variations in the processing steps and the final product properties are found, and recommendations to reduce the sources of variations are discussed.

1. Introduction

Multivariate statistical analysis methods can assist in the identification of process correlations, thereby supporting or improving existing process knowledge. Previous researchers have used the multivariate methods of principal component analysis (PCA), partial least squares (PLS), and canonical correlation analysis (CCA) successfully in several ways (MacGregor et al., 1994, Wold, 1978, Piovoso et al., 1992a, Kosanovich and Piovoso, 1991). They include, but are not limited to, data analysis, model development, prediction, and control variable selection. Most of the examples are taken from continuous processes. In the application investigated here, a variant of the PCA technique, multiway PCA (MPCA), is used to analyze data taken from an industrial batch process.

Batch and semibatch processes play an important role in the chemical industry, mainly because of their flexibility to produce low-volume, high-value products. Examples include reactors, crystallization, distillation, injection-molding processes, and the manufacture of polymers. Batch processing typically involves charging the vessel, processing under controlled conditions, and finally discharging the product. Successful operation means tracking a prescribed recipe and the process variables' trajectories with a high degree of reproducibility from batch to batch. Temperature and pressure profiles are implemented with servocontrollers, and precise sequencing operations are carried out by tools such as programmable logic controllers.

The main characteristics of batch processes, flexibility, finite duration, and nonlinear behavior, also make process control difficult. Control problems are complicated further by a lack of sufficient on-line instrumentation. While feedback control in the continuous process sense may not be possible, statistical quality control in some form is an option that is often used. Currently, CUSUM techniques are used to adjust the rate and duration of heating or cooling and the duration of the specific stages of the batch. CUSUM, or cumulative sum, calculates the sum of the errors over a time window and statistically charts the result. Statistical techniques such as Shewhart charts may be applied to characterize information about a single important variable, and still others such as MPCA may be used to analyze the multivariate variations.

The primary goal of this study is to demonstrate how MPCA can be used to improve process understanding. This is achieved by analyzing data taken from an industrial batch polymer reactor. It is shown that MPCA can be used effectively to identify the major sources of variability in these data and that these variations are related to product quality properties. To wit, recommendations are made on how to reduce or eliminate the variability.

Reducing the variability in the process will permit the production of a uniform, high-quality product. The economic stake for achieving this is potentially large with further ramifications in plant operations such as (a) lower energy costs, (b) lower raw material costs, (c) reduced time to transition between different products, (d) reduced off-line product testing, and (e) reduced downtime, to name a few. These opportunities exist in the batch process studied here, and since more than one manufacturing location produces either the same or analogous products, any modifications that reduce or eliminate the variability at one location will be applicable to others.

The paper is organized as follows. First, a process description is provided. Second, the MPCA method is outlined, starting with a review of conventional PCA. Third, the results obtained, from applying MPCA to data taken from two reactors producing the same polymer recipe, are presented and discussed. Finally, suggestions are made for process improvements.

2. Process Description

The charge to the reactor is an aqueous solution that is first boiled in an evaporator until the water content is reduced to approximately 20% by weight. The evaporator's contents are then discharged into a reactor

* Author to whom correspondence should be addressed. Phone: (803) 777-0143. Fax: (803) 777-8265. E-mail: kosanoka@sun.che.sc.edu.

0888-5885/96/2635-0138$12.00/0

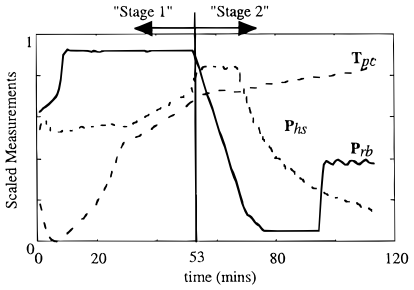

Figure 1. Example reactor profiles. P_{hs} is the heat source supply pressure, T_{pc} is the polymer center temperature, and P_{rb} is the reactor body pressure.

Table 1. Reactor Process Measurements

variable	description
T_{pc}	polymer center temperature
T_{va}	vapor temperature
T_{ps}	reactor center temperature
P_{rb}	reactor body pressure
V_{rv}	reactor vent valve position
T_{hs}	heat source supply temperature
T_{hj}	heat source jacket vent temperature
T_{hc}	heat source coil vent temperature
P_{hs}	heat source supply pressure
V_{hs}	heat source supply valve position
$P_{hs,sp}$	heat source pressure control setpoint

in which 10−20 lbs. of polymer residue may be present from the processing of the previous batch.

This batch reactor is operated according to a combination of prespecified reactor and heat source pressure profiles and timed stages. Example profiles are shown in Figure 1 (see Table 1 for nomenclature). The time to complete a batch is approximately 120 min. Key process checkpoints (e.g., attaining a specific temperature within a given time) determine when one processing stage ends and the next one begins.

In the first step of the recipe, heat is applied to the reactor to further concentrate the reactants and to supply the activation energy to start the polymerization reactions. At the outset, the reactor temperature and the pressure rise rapidly (see Figure 1). Sensor measurements indicate the existence of a temperature gradient having as much as 40 °C difference between the material at the top and at the bottom of the reactor. Shortly after the pressure reaches its setpoint, the entire mixture is boiling and the temperature gradient disappears. The solution is postulated to be well-mixed at this time. Measurements such as the cumulative amount of water removed are also used as an indication of the extent of polymerization.

In the second step, the reactor pressure is reduced (ramped down) to 0 psig to flash off any remaining water after a desired temperature is reached. Simultaneous ramping of the heat source to a new setpoint is also carried out. The duration spent at this second setpoint is monitored by a CUSUM loop so that the batch reaches a desired final reactor temperature within the prescribed batch time. In the third step, the heat source is removed and the material is allowed to continue reacting until the final desired temperature is reached. The last step involves the removal of the finished polymer as evidenced by the rise in the reactor pressure.

Each reactor is equipped with sensors that measure the relevant temperature, pressure, and the heat source variable values. These sensors are interfaced to a distributed control system that monitors and controls the processing steps. The locations of the sensors used in this study are shown in Figure 2.

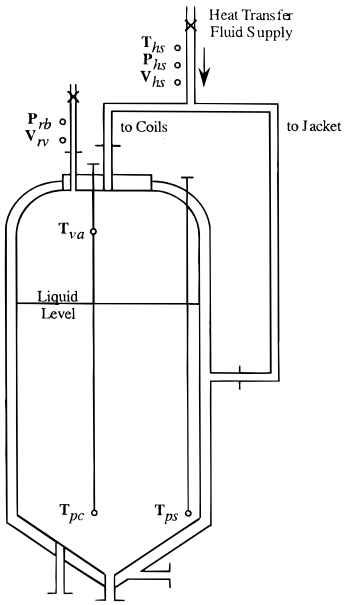

Figure 2. Schematic diagram of batch reactor showing process measurement locations. Table 1 contains a description of the process variables.

2.1. Processing Steps. Some comments on the general operations are necessary to further clarify the purpose of this study.

1. One evaporator may be used to prepare the charge to several reactors. If a reactor is unavailable (e.g., a longer processing time than usual), the contents of the evaporator are kept above its freezing point by a blanket of steam. A longer holdup time translates to a composition disturbance (more water) to the reactor which may increase both the overall batch time and energy costs.

2. The evaporator temperature is not under closed-loop control. Small deviations (1−2 °C) translate to a composition disturbance to the reactor.

3. The critical desired temperatures are monitored by CUSUM loops. Adjustment of the heat source, to compensate for failure to reach a desired temperature within a prescribed time, may occur *during* or at the *start* of a batch. It will be shown that this practice contributes to batch-to-batch process variability.

4. The product quality properties (e.g., molecular weight) are determined by off-line laboratory analysis for a selected number of batches. The analyses are usually reported 8 h or more after a batch is manufactured. This delay increases the difficulty in identifying and correcting the causes for deviations of subsequent batches.

5. The ultimate use of the product from each batch is determined by process performance and by product properties. Polymer from a batch that is not within the prespecified limits requires special processing.

3. Multiway Principal Component Analysis

The primary purpose of this work is not to demonstrate how multiway principal component analysis (MPCA) might be used for process monitoring but rather how it can be used to improve process understanding. With this focus, only a brief overview of the technique and its underlying statistical basis is presented. For a detailed discussion of MPCA including the selection of the number of principal components, development of control charts, and computation of control limits for statistical process control, the interested reader is

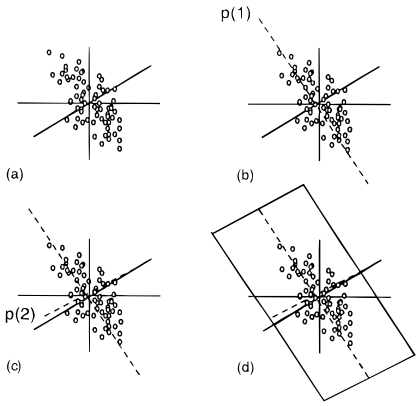

Figure 3. Geometric explanation of PCA: (a) scatter plot of the data; (b) dotted line labeled p(1) indicates the first eigenvector; (c) dotted line labeled p(2) indicates a second eigenvector that is orthogonal to the first; (d) the plane defined by p(1) and p(2) where the data are projected.

referred to the work by Nomikos (1995) and Nomikos and MacGregor (1994a, 1995).

Multiway principal component analysis (MPCA) is an extension of principal component analysis (PCA) for three-dimensional data. Relative to continuous processes, batch processes have an added dimension of the batch number in addition to the measured variables and sample times. To understand MPCA, it is necessary to review conventional PCA first.

3.1. Principal Component Analysis. PCA decomposes a single, dependent set of data into a transformed space defined by the eigenvectors of the covariance of the data. Steady-state conditions and linearity of the data are assumed to apply when PCA is used to analyze data from continuous processes. If the data are correlated, their information content can be captured by a smaller set of variables. For example, a typical process may be instrumented to collect and store hundreds of process measurements. Physicochemical relationships tell us that there are not hundreds of independent events occurring; therefore, the data are correlated. Any technique that can capture the important events based on the variability in the data will provide both a reduction in the data size and a summary of the information contained in the original data set (Kasper and Ray, 1992). PCA is one such method. It generates a set of pseudomeasurements (*scores*) that are linearly independent, each of which captures the maximum amount of variability in the data in *descending order*. Hence, the reduced data set requires fewer numbers to represent the same information found in the original data set (Wold, 1978).

As an example, consider data consisting of observations taken from three sensors. A plot of these data is shown in Figure 3a. Notice that the data are not randomly scattered in the variable space; rather they lie primarily along the dotted line drawn through the data (Figure 3b). This line, p(1), is defined by the first *eigenvector* or *loading* and represents that linear combination of the data that captures the direction of maximum variability. The projections of the data onto p(1), as defined by their distances along it, constitute the *scores*. If this approximation is not accurate enough as determined by large residuals, a second eigenvector, p(2), can be found as a function of the residual data (Figure 3c). The residuals are obtained by subtracting the previous projections from the original data. The two eigenvectors define a plane in the original variable space (Figure 3d). This process can be repeated systematically until the size of the eigenvalue associated with each new eigenvector is of such a small magnitude that it represents noise associated with the observations. In the limit where the number of significant eigenvectors equals the number of variables, there is no reduction in the dimension of the variable space, but, more importantly, it would suggest that the variables are linearly independent. This is the special case of singular value decomposition.

3.2. Background on MPCA. Analogous to continuous processes, batch data contain the time history of each measured variable. Unlike continuous processes, however, batch data must reflect the batch number from which the time histories are taken. Multiway PCA is PCA extended to deal with this extra dimension.

Batch processes are nonstationary and of finite duration. The generic processing steps are charging the batch, processing the contents according to prescribed profiles, and discharging the finished material. The degree to which the product from each batch meets quality specifications is determined, after the batch is completed, by laboratory analysis. Feedback control of the quality variables, in the same sense as that of a continuous process, is not possible; however, feedback control of other indirect measures of quality is sometimes possible (Garcia, 1984; Peterson et al., 1992). While time histories of the process variables are recorded at high frequencies, the final quality measurements are sampled at low frequencies. Statistical process control (SPC) charts based on univariate analysis usually are used to *control* product quality. This must not be confused with the servo-regulatory control layer that exists to keep the process variables tracking their prescribed profiles.

Batch processes may exhibit batch-to-batch variability for many reasons. A few are composition disturbances, deviations from specified profiles, equipment defects, and heat-transfer limitations. Quite often, these fluctuations do not alter the product sufficiently to cause a quality problem. However, large variations can lead to the production of many off-aim batches if the problem goes undetected. A monitoring model, developed using MPCA, can be used to detect and to correct problems early in the batch cycle (MacGregor and Nomikos, 1992; Nomikos and MacGregor, 1994b).

Monitoring ideas for batch processes are plentiful. Marsh and Tucker (1991) used the notion that the trajectoriers ought to follow a certain dynamic pattern and applied a simple SPC technique for a single-variable monitoring scheme. Konstantinov and Yoshida (1992) applied temporal shapes of time profiles, while Holloway and Krogh (1990) used trajectory encoding to apply qualitative reasoning to monitor the dynamics of the batch.

Multivariate techniques can be used to identify process variability and to develop monitoring models and multivariate SPC charts for on-line process monitoring and control. Kresta et al. (1990) and MacGregor et al. (1994) use PCA and projections to latent structures (PLS) to analyze a large number of highly correlated variables that defines a continuous chemical process. By comparing new observations with a model that describes *normal* variability, simple control charts can be generated to detect data inconsistencies and processing problems.

Piovoso et al. (1992a) developed an on-line monitoring model and a control strategy based on a PLS/PCA model and implemented it on an industrial, continuous chemical process. They demonstrated the effectiveness of the on-line monitoring model to detect process upsets. In a related work, Piovoso et al. (1992b) demonstrate and

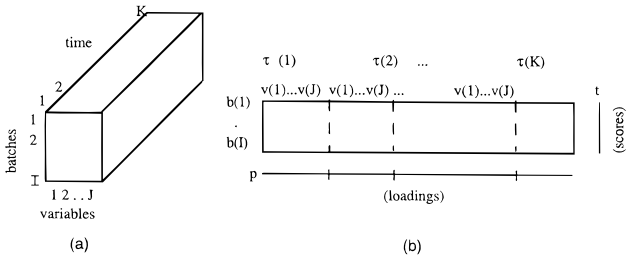

Figure 4. (a) Representation of batch data indicating the three-dimensional structure; (b) a particular unfolding where the rows are the batches and the columns are the variables, v_j, sampled at each time, τ_k.

discuss the importance of prefiltering the data (outlier and data validation) prior to model development. MacGregor and Nomikos (1992) and Nomikos and MacGregor (1994a,b, 1995) have proposed using MPCA to monitor a commercial batch process. Kosanovich et al. (1994) presented a preliminary report on the use of MPCA for improving process understanding of an industrial batch reactor.

3.3. MPCA Method. Batch data can be characterized by three parameters: batch number, process variable, and sample time. Figure 4a illustrates a typical three-dimensional data array, **X**, where the axes $i = 1, 2, ..., I$, $j = 1, 2, ..., J$, and $k = 1, 2, ..., K$ correspond to the batch number, variables, and time, respectively. MPCA is implemented by performing a PCA analysis on an unfolded form of the three-dimensional data array. There are three ways of unfolding the array to form a two-dimensional matrix; transposing each matrix when unfolding the array produces three different two-dimensional matrices, yielding a total of six possibilities. Analysis of PCA on each of the six matrices will explain a different type of variability. The unfolding used in this work is obtained by taking vertical slices along the time axis and laying the slices side-by-side to produce a two-dimensional matrix **X** of size ($I \times JK$), as shown in Figure 4b. Thus, the first J columns correspond to *all* the variables over *all* the batches sampled at time $k = 1$; the next J columns represent the same set of variables at the next sample time, and so on. For more details regarding multiway methods and their unfolding for analysis, the reader is referred to Smilde (1992) and Henrion (1994).

Analogous to PCA, MPCA performs a decomposition on the data by finding the directions of maximum variability (loadings) over all time and variables. The scores are found by projecting all the data onto the loadings. The scores, **T**, based on the entire history for all the batches are orthogonal, and the loading matrix, **P** ($=\mathbf{P}_K$), is orthonormal. Since **P** has information that is a function of time, the scores can be computed prior to obtaining the entire batch history. These projections would be the scores at the present time, k. The intermediate scores, $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_k$, and loadings, $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_k$, are, in general, *not* orthogonal. The intermediate scores provide information as to the degree to which the variability is being explained by the model. The corresponding loading matrix contains the correlations among the variables up to that time, k. The number of loading matrices and score vectors needed is generally small when there is a high degree of correlation among the data.

The data are normalized using the mean and standard deviation of each variable at each time in the batch cycle over all batches. Subtracting the average batch trajectory generally eliminates the major nonlinear and nonstationary behavior of the process (MacGregor et al., 1994). Hence, a linear, static technique such as MPCA can be used to analyze perturbations about each mean trajectory.

3.4. Statistical Measures of MPCA Analysis. There are several ways of interpreting the MPCA results. The ones used here are a combination of statistical measures and graphical analysis (Martens and Naes, 1989; Nomikos and MacGregor, 1994a). They are (a) Q-statistic, a measure of the model mismatch relative to new observations; (b) D-statistic, a measure of the fit of new observations to the model space; (c) variance plots, a measure of the batch profiles' variabilities; and (d) score plots, qualitative representations of the batch-to-batch performance, relative to the calibration model in the model space defined by the MPCA analysis. Taken together, they provide an assessment of the statistical significance of the MPCA analysis.

The Q-statistic is the sum of squares of the errors between the data and the estimates. The latter are calculated from a fixed number of principal components. The error for the ith batch is defined as:

$$Q_i = \sum_{j}^{J}\sum_{k}^{K}(x_{i,j,k} - \hat{x}_{i,j,k})^2 \qquad (1)$$

where $\hat{x}$ is the estimate based on the MPCA model. Q-statistics can also be computed for each variable (the sum in eq 1 is taken over batches and time) and at each point in time (the sum in eq 1 is taken over variables and batches). In the first case, the sum yields the amount of residual variability associated with a given variable and, in the second, the amount of residual variation at each point in time.

The D-statistic, or Hotelling statistic (Jackson, 1992), measures the degree to which data fit the calibration model. It is defined as:

$$D = \mathbf{t}'_r\mathbf{S}^{-1}\mathbf{t}_r I(I - R)/(R(I^2 - 1)) \qquad (2)$$

where **S** is the estimated covariance matrix of the scores, I represents the total number of batches, R is the total number of principal components, and $\mathbf{t}_r$ is a vector of R scores. If the calibration model data represent process operation at one setpoint and the process has shifted to a new setpoint, then the D-statistic most likely will show that data collected at this new operating condition cannot be classified with the calibration data.

Statistical limits on the Q-statistic and D-statistic are computed based on assuming the data are normally distributed in the multivariate sense (Jackson, 1980, 1992). When these assumptions are valid, the diagnostic limits are useful to establish when a statistically significant shift has occurred. Charts based on these statistics and used in this manner are analogous to conventional SPC charts.

The assumption of normality is rarely satisfied in practice. Non-normal data tend to inflate the variance, which, in turn, tends to reduce the D-statistic in eq 2. Typically, this increases the probability of failing to detect an outlier. In this work, since process monitoring and model development are not the objectives, this assumption of normality is not a major concern.

The explained variance (the total variance minus the residual variance) is calculated by comparing the true process data with the estimates computed from the calibration model. This measure can be computed as a function of the batch number, time, or variable number. A large explained variance indicates that the variability

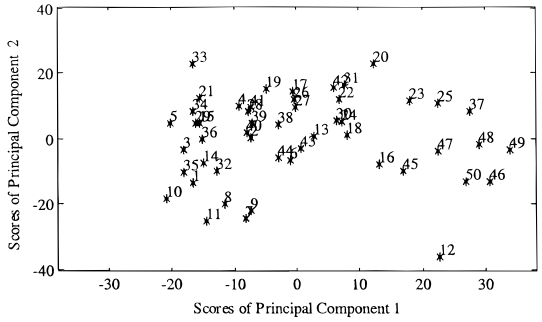

Figure 5. A two-PC score plot of reactor A's data. The numbers indicate the batch number.

in the data is captured by the model and that correlations exist among the variables. The explained variance as a function of time can be very useful to differentiate among phenomena that occur in different stages of the process operations.

The score plots provide a summary of process performance from one batch to the next. All batches exhibiting similar time histories will have scores which cluster in the same region of the principal component space. Thus, from a visual point of view, batch-to-batch process variability is readily identified.

4. Data Analyses

Process data for the same polymer recipe are analyzed for 50 nonconsecutive, sequential batches from a given reactor, reactor A, and 31 consecutive batches from another reactor, reactor B. The data are sampled at 1 min intervals during production of each batch. Normalization of the data, as discussed in section 3.3., is done prior to analysis. The final product quality property data are obtained from laboratory measurements of molecular weight and end groups, with one reading for each batch.

Figure 5 is a score plot for a two principal component (PC) calibration model developed using data taken from reactor A. The numbers in the figure indicate the batch number. Since our objective is process understanding and not monitoring, only the first two PCs are selected for analysis. Beyond two, the ability to explain the results of MPCA in terms of the physical process is extremely difficult if not impossible. Note that for monitoring purposes, two PCs are probably insufficient to capture all the process variability adequately. Also, from a monitoring perspective, some of the batches (e.g., batch 12) may be classified as outliers. However, all of the batches used for calibration have a classification of first pass quality product. Thus, from a process perspective, there is no justification for removing any data. Observe that, among the considerable scatter, there is the sequence of indexed batches, 45−50, and several others on the far right. A closer inspection of the operating conditions for these batches indicates that they were processed at a different heat-transfer rate than the others.

Figure 6 shows the variance explained by the two-PC model, as a percentage, for each of the three indices: batch number, time, and process variable. The lower set of bars in Figure 6a,c are the explained variances for the first PC, while the upper set of bars are the additional contribution from the second PC. The solid line in Figure 6b is the explained variance over time for the first PC, and the thicker line is the sum of PC 1 and 2. This format of bars and lines will be used throughout. The unexplained variance is not only noise.

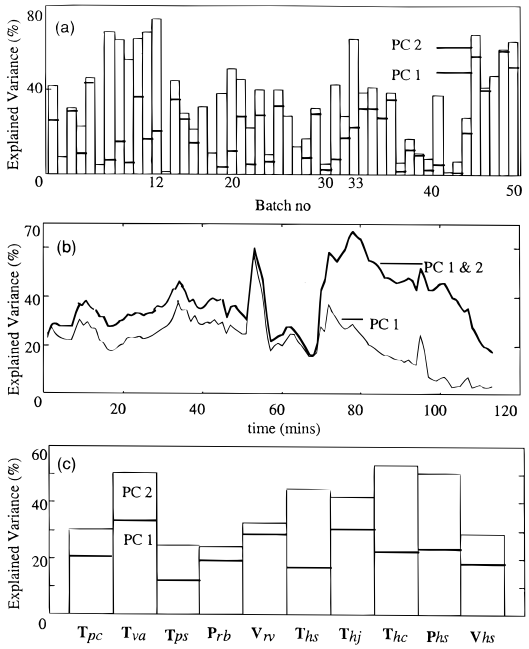

Figure 6. Explained variance (%) (a) by batches, (b) over time, and (c) over all variables for reactor A's data.

More than two principal components are needed to capture all the significant variations, but only two are selected for process understanding purposes. Figure 6a indicates, for example, that batch numbers 13 and 30 have a small explained variance, while batch numbers 12 and 33 have a greater amount of their variance captured by the calibration model after two PCs. One cannot conclude from this plot alone, however, that batches 13 and 30 are poorly modeled by the calibration model. Batches 13 and 30 may, in fact, have just small random variations about the average batch trajectory. If so, the scores for these will be nearly zero; hence, the explained variations about the mean will also be negligible. Thus, we may conclude that batches 13 and 30 are modeled adequately by the average batch trajectory.

In Figure 6b, the magnitude of the explained variance accounted for by PC 2 has noticeably increased after minute 70. This second PC also exhibits large explained variances in the heat source variables, T_{hs} and T_{hc} (Figure 6c). From process knowledge, it is known that the removal of water is the primary event in the first part of the batch recipe, while polymerization dominates in the later parts. This knowledge and these statistical results led to the conclusion that better insights into the variations may be obtained by separating the data into two time histories. The first covers time $k = 1-53$ min, while the second covers time $k = 54-113$ min to better match the physical phenomena.

Score plots developed from separately calibrated PC models for each reactor are shown in Figure 7. Observe the presence of clusters in Figure 7a,c for data from time $k = 1-53$ min and the scatter in Figure 7b,d for $k = 54-113$ min. Figure 7a exhibits three clusters: in the upper right-hand corner, labeled by the symbol ⊕; in the lower center, labeled by the symbol ⊗; and a nearly linear one labeled by the symbol ○. Figure 7c has three clusters as well: batches 13, 15, and 22 (symbol ⊕); a linear cluster (symbol ⊗); and another linear cluster (symbol ○). Three batches, 49, 23, and 26, are underscored in Figure 7a. These batches, will be referred to subsequently to explain certain phenomena occurring within the batch operation (also see Figure 11). Additionally, the symbols ⊗ and ⊕ are used to define those

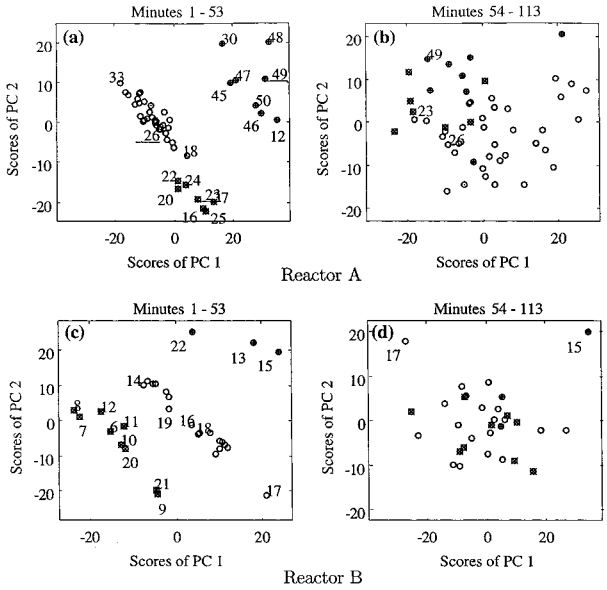

Figure 7. (a and b) A two-PC score plot for reactor A's data. (c and d) Reactor B. Symbol ⊕ is used to indicate a cluster in the upper right-hand corner, symbol ⊗, the cluster in the lower center, and symbol ○, the third and nearly linear cluster.

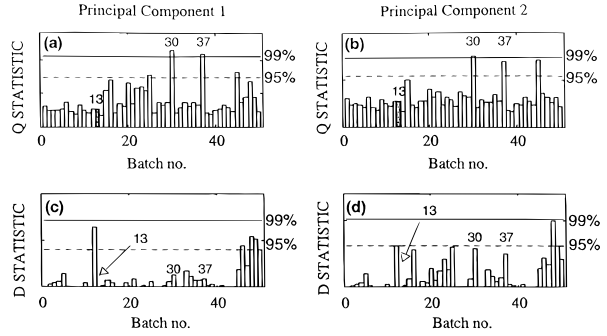

Figure 8. Q- and D-statistics by batch number for $k = 1-53$ min for reactor A's data: (a) Q-statistic for PC 1; (b) Q-statistic for PC 2; (c) D-statistic for PC 1; and (d) D-statistic for PC 2. The solid horizontal line indicates the 99% confidence limit, and the dashed horizontal line, the 95% confidence limit in all the plots. Observe that batches 30 and 37 exceed both the 99% and 95% confidence limits in both PC 1 and 2 of the Q-statistic but not the D-statistic; batch 13 is within the 95% confidence limit of both the Q- and D-statistics.

batches in the clusters that have high residuals in the process variables.

One conclusion that can be drawn from the clustering that is found during $k = 1-53$ min but not in the later processing time is that the later stage is not influenced by the same factors which led to the formation of clusters in the earlier stage, and conversely, that the processing steps in the later stage remove whatever led to the differentiation in the earlier stage.

The Q- and D-statistics for reactor A, as a function of batch number, are shown in Figure 8 for time $k = 1-53$ min. The Q-statistic for batches 30 and 37, for example, exceeds the 95% limit for *both* PCs, indicating that the calibration model does not capture well the variations in these batches. In the D-statistics plot, batch 48 exceeds the 95% limit for both PCs. The conclusion that can be reached is that the magnitudes of the variations in this batch are larger than those captured in the calibration model. The Q-statistic for 50 batches produced in reactor A as a function of time, depicted in Figure 9, shows that deviations from the model subspace occur primarily in the first 35 min.

Recall that batches 13 and 30 had small explained variances. Observe, that the Q- and D-statistics for

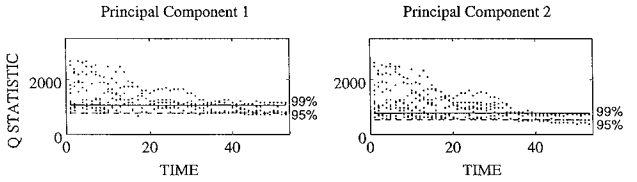

Figure 9. Q-statistic for $k = 1-53$ min for reactor A's data illustrating larger deviations in both principal components during the first 20 min of the operation.

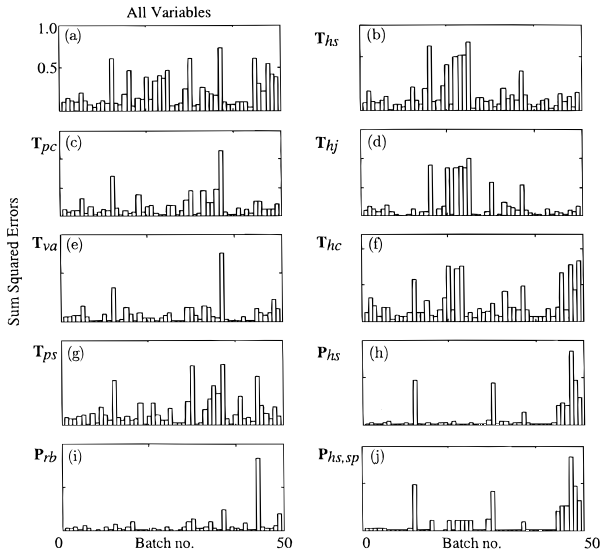

Figure 10. (a) Fraction of the total error over all variables; (b−j) sum of the squares of the errors by variable, for Reactor A's data. See Table 1 for a description of each variable and Figure 2 for their sensor location.

batch 13 indicate that it is within the 95% limit for both PCs. In contrast, the Q-statistic of batch 30 is not, while its D-statistic is below the 95% limit. The conclusion that can be drawn is that the variations in batch 13 are small random deviations about the average batch. In the case of batch 30, a small component of the data fit the calibration model, but the majority do not. These variations are either large random fluctuations or variations that are orthogonal to the model subspace.

The sum of squares of the errors (SSE) for each variable is shown in Figure 10. Observe that nearly all batches in the clusters, labeled by ⊗ and ⊕ in Figure 7a, have at least 30% of the total error (Figure 10a). Many of the same batches have significantly higher SSE values for the individual variables as well. Closer inspection of the three clusters in Figure 7a reveals that a different value of the heat source pressure setpoint, $P_{hs,sp}$ (Figure 10j), was used for each of the three clusters. Thus, heat source pressure setpoint differences are the origin for the clusters. Figure 11 confirms this finding. Figure 11a shows a representative batch from each cluster (marked by the underscore) in Figure 7a. There is a different value of $P_{hs,sp}$ for prolonged periods during the first 53 min. Parts b and c of Figure 11 show that the effects of $P_{hs,sp}$ are evident as well in the variables P_{hs} and T_{hj}.

4.1. Data Analyses of Reactors A and B. Data from both reactors (each producing the same recipe) are analyzed using MPCA. Figure 12 shows a score plot for a two-PC model. Each reactor forms a distinct cluster. Possible reasons for this segregation are biases in instrument calibration and time in service for each reactor, but there may be other reasons as well. The heat source pressure setpoint-induced clusters, visible when each reactor's data are modeled separately (*cf.*

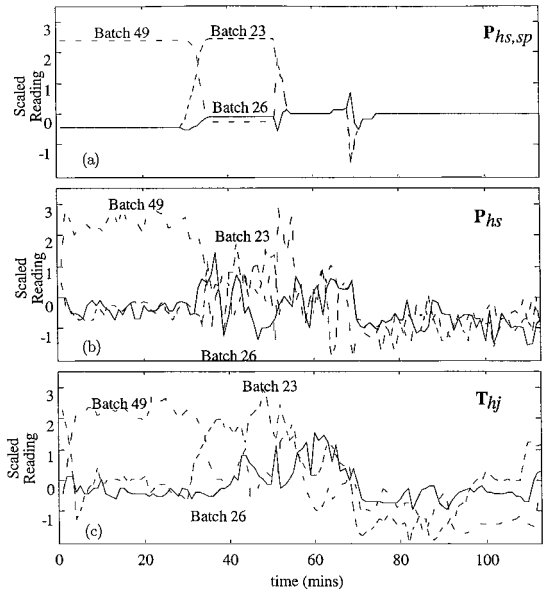

Figure 11. Temporal history of selected batches to demonstrate the differences among the clusters depicted in Figure 7a: (a) heat source pressure setpoint, $P_{hs,sp}$; (b) heat source supply pressure, P_{hs}; (c) heat source jacket vent temperature, T_{hj}, for reactor A's data.

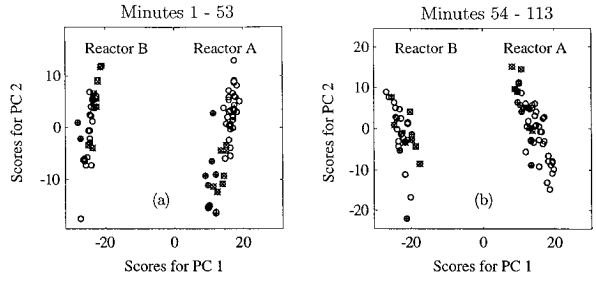

Figure 12. Two-PC score plot for reactors A and B's data partitioned to match the approximate time of the physicochemical phenomena. The separation is due primarily to biases in instrument calibration.

parts a and c of Figure 7), are not evident in Figure 12a. Kosanovich et al. (1994) report similar findings.

Based on the previous discussion, the heat source pressure setpoint variable is removed from the data set. Two other variables, valve positions V_{hs} and V_{rv}, are also excluded, as they provide no additional insights into the data analysis. Because they are manipulated variables, they did not correlate well with other variables and hence are not well explained by a two-PC MCA model. To remove individual instrument or setpoint-derived biases, the remaining data for each reactor are mean-centered by their individual variables' mean trajectories over all batches in each of the six clusters in Figures 7a,c. The results are then combined and analyzed using MPCA.

Figure 13a is a two-PC score plot from $k = 1-53$ min for both reactors. The ellipses show control limits at 95% (− −) and 99% (−) confidence levels for the calibration model, assuming an approximate normal distribution. The two separate clusters (*cf.* Figure 12) no longer appear in the score plot nor do clusters associated with different setpoints (cf. Figures 7a,c). Hence, scaling by the individual variables' mean trajectories over all batches is effective. Figure 13b indicates that the dominant variations are associated with the reactor temperature variables (magnitude of the explained variance by PC 1). The use of the value of the amount of explained variance to determine important variables is considerably easier than analyzing the loadings. This

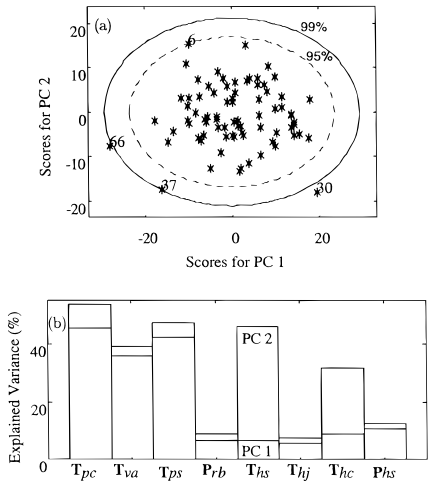

Figure 13. (a) Score plot of reactors A and B's data corrected for heat source pressure setpoint differences for $k = 1-53$ min. The solid ellipse represents the 99% confidence limit and the dotted ellipse the 95% confidence limit. (b) Explained variance (%) of each variable over all batches and time. Observe the large explained variance contributed by variable T_{hs}, the heat source supply temperature, to PC 2.

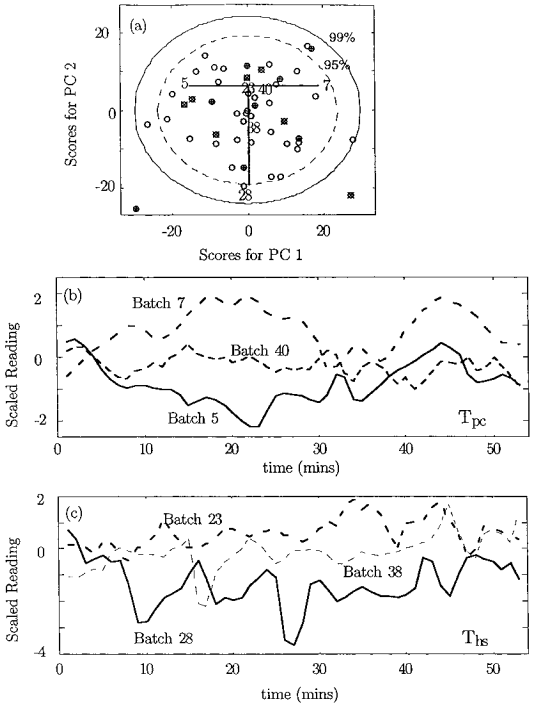

Figure 14. (a) Two-PC score plot for $k = 1-53$ min for reactor A's data corrected for heat source pressure setpoint differences. The solid ellipse represents the 99% confidence limit and the dotted ellipse the 95% confidence limit. (b) Temporal history of selected batches, delineated in plot a, of the reactor center temperature, T_{pc}. (c) Temporal history of selected batches, delineated in plot a, of the heat source supply temperature, T_{hs}. Plots b and c illustrate that the primary difference among these batches is contributed by variations in temperature.

is because, in MPCA, the loadings are functions of time; hence, the analysis is more involved. Contrast that to the evaluation of a single number, the explained variance value. A high value of this quantity, for a given variable, provides a reliable indication of its contribution to a particular score.

From the score plot for a model of reactor A alone from $k = 1-53$ min (Figure 14a), batches 5, 40, and 7 have constant values in the direction of PC 2, while they cover much of the range of values associated with PC 1. The labeling is as before to indicate the clustering that was

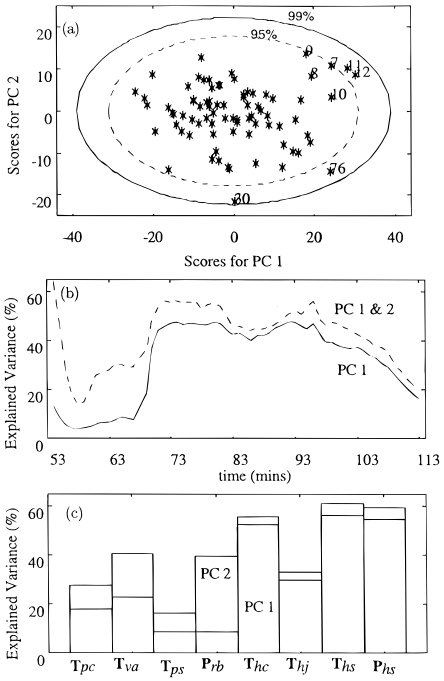

Figure 15. (a) Two-PC score plot for $k = 54-113$ min for both reactors' data corrected for heat source pressure setpoint differences. The solid ellipse represents the 99% confidence limit and the dotted ellipse the 95% confidence limit. (b) Explained variance (%) over time shows that the contribution to PC 2 occurs in the first 15 min. The solid line represents PC 1 and the dotted line the cumulative effect of PC 1 and 2. (c) Explained variance (%) over the variables (see Table 1 for a description of each variable). Observe that the large contributions to PC 1 come from the variables that describe the state of the heat source.

found in Figure 7. By comparing the reactor temperature values for these batches, it is found that a negative value of PC 1 is associated with low temperatures, and positive values correspond to high temperatures (Figure 14b). Analogous analyses of batches 23, 38, and 28 indicate that a negative value of PC 2 is associated with low heat source temperatures, and positive values are associated with higher heat source temperatures (Figure 13c).

Results of the MPCA analysis on data from $k = 54-113$ min are shown in Figure 15 for both reactors. The score plot shows a single cluster (Figure 15a), while the explained variance as a function of time shows a dramatic rise in the first principal component around the 70 min mark (Figure 15b). This increase is found to be associated with variations in the heat source variables (Figure 15c). From process knowledge it is known that, at or near $k = 70$ min, the external heat source is abruptly turned off to both the jacket and coil (cutoff point). This cutoff is a required step in the current process recipe. This affects the rate at which the residual vapors, in the jacket and coil, leave the system. The residual vapor variations are a function of the cutoff time and the state (energy) of the reactor proper contents. The majority of the explained variance by PC 2 occurs between 54 and 70 min (Figure 15c). A cumulative rise of 40% in the explained variance of the reactor pressure, P_{rb}, is the greatest increase associated with this principal component.

Since the principal components are determined in an ordered fashion, reducing the variability in the reactor temperature which is found to be the major source of variability, will reduce batch-to-batch variability in the processing steps and, possibly, final product quality properties. Currently, the reactor temperature is not controlled; a study by Kosanovich and Schnelle (1995),

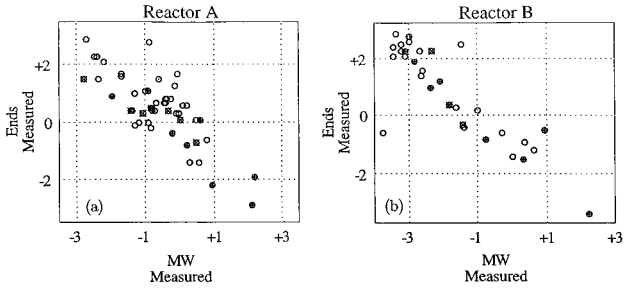

Figure 16. Laboratory results for end groups versus molecular weight. The inverse relationship between the two is apparent. Symbols ⊕, ⊗, and ○ are used to represent the clusters found in Figure 7. Some of the batches marked by ⊕ have high molecular weight and low end groups, while most of those marked by the symbol ⊗ have low molecular weight and high end groups.

investigating different control schemes to reduce temperature variations, indicates that significant reduction in batch-to-batch variability can be achieved but that no one scheme can compensate for all the known, measurable disturbances. Nevertheless, this operability change is being pursued aggressively at one manufacturing site.

4.2. Quality Data Analysis. The final polymer quality is determined by laboratory analysis of molecular weight and end groups. The results are available usually 8 h or more after the batch is completed. The size of the production volume and laboratory costs make quality measurements for every batch impossible. However, for this study, final product quality properties are available, one for each batch; thus, analyses of the correlations between the quality properties and process variables are possible. Parts a and b of Figure 16 show the relationship between the molecular weight and end groups for data taken from both reactors. The expected inverse relationship between these two polymer properties is evident; that is, a high molecular weight implies a low number of end groups and vice versa.

Two MPCA calibration models are developed for each time history, using process data from both reactors, uncorrected for setpoint changes. This was done to determine if setpoint changes correlated to quality variations. The variations in the molecular weight and end groups are correlated to the process variables by regressing them onto the scores of the MPCA model. This is effectively multiway principal component regression (MPCR). A similar analysis can be performed using multiway partial least squares (MPLS). However, since the primary objective of this work is to demonstrate that MPCA can identify variations related to product quality variations, there is no reason to compare one method to another.

Figure 17 shows the predicted versus measured values of the molecular weight and end groups. The *r*-squared number (r^2) represents the fraction of the variance explained in the measurements by the model. A substantial fraction of the variations in the quality measurements is correlated to the variations captured by the calibration model: 48−58% for molecular weight and 36−49% for end groups. These results show that product quality variations are correlated to the variations in the process measurements used in the model and also imply that product quality can be improved by reducing the variability in the process data.

5. Summary

This study demonstrates that MPCA can be used to analyze historical data from a commercial batch reactor,

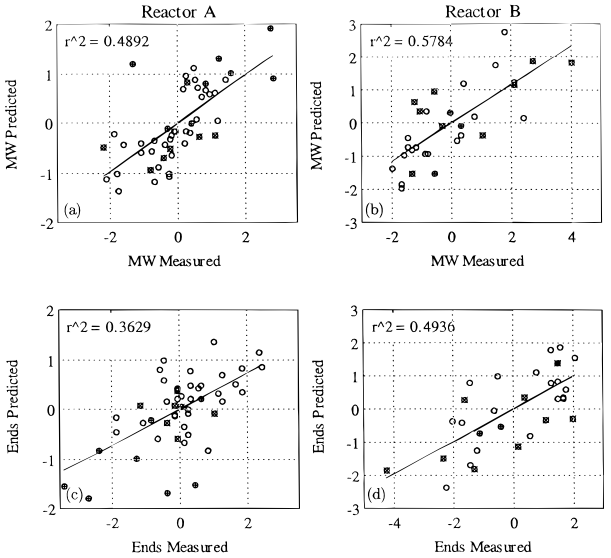

Figure 17. Predicted versus laboratory results for end groups and molecular weight as determined by MPCR. The r^2, fraction of the variations explained by the model, is approximately 50%.

leading to improved process understanding. Additionally, the analysis shows a correlation between the variations (~50%) in the process variables and those of the quality measurements. By using specific process knowledge with the data analysis results, a more useful interpretation of the analyses can be obtained.

Major sources of process variability are identified as (1) differences in instrument calibration and/or reactor age, (2) betweem reactors, varying heat rate (variable pressure setpoint) during and from batch to batch, and (3) the uncontrolled residual vapor decay rate of the heat source after 70 min of processing (cutoff point). The major change, that of direct control of the reactor temperature to a prescribed trajectory, is recommended; plans to achieve this are underway at at least one manufacturing site.

Acknowledgment

The authors acknowledge the assistance of Paul Nomikos and John MacGregor of McMaster University, Hamilton, Ontario, Canada; Mark Sibley and Keith Marchildon, of DuPont Canada, and Mahmud Rahman and J. C. Chang, of the DuPont Co., U.S.A.

Literature Cited

Garcia, C. E. Quadratic/Dynamic Matrix Control of Nonlinear Processes. An Application to a Batch Reaction Process. AIChE November Meeting, San Francisco, CA, 1984.

Henrion, R. *N*-way Principal Component Analysis: Theory, Algorithms and Applications. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 1−23.

Holloway, L. E.; Krogh, B. H. On-Line Trajectory Encoding for Discrete-Observation Process Monitoring. *Proceedings of the On-Line Fault Detection and Supervision in the Chemical Processing Industry*, IFAC Symposium, Newark, DE, 1990.

Jackson, J. E. Principal Components and Factor Analysis: Part I—Principal Analysis. *J. Qual. Technol.* **1980**, *12*, 201−213.

Jackson, J. E. *A User's Guide to Principal Components*; John Wiley & Sons: New York, 1992.

Kasper, M. H.; Ray, W. H. Chemometric Methods for Process Monitoring and High-Performance Controller Design. *AIChE J.* **1992**, *38*, 1593−1608.

Konstantinov, K. B.; Yoshida, T. Real-Time Qualitative Analysis of the Temporal Shapes of (Bio)process Variables. *AIChE J.* **1992**, *38*, 1703−1715.

Kosanovich, K. A.; Piovoso, M. J. Process Data Analysis using Multivariate Statistical Methods. *Proc. Am. Control Conf.* **1991**, *1*, 721−723.

Kosanovich, K. A.; Schnelle, P. D., Jr. Improved Regulation of an Industrial Batch Reactor. Presented at the Session on Novel Applications in Process Control, AIChE Spring Meeting, Houston, TX, 1995.

Kosanovich, K. A.; Piovoso, M. J.; Dahl, K. S.; MacGregor, J. F.; Nomikos, P. Multi-Way PCA Applied to an Industrial Batch Process. *Proc. Am. Control Conf.* **1994**, *2*, 1294−1298.

Kresta, J. V.; MacGregor, J. F.; Marlin, T. E. Multivariate Statistical Monitoring of Process Operating Performance. *Can. J. Chem. Eng.* **1990**, *69*, 35−47.

MacGregor, J. F.; Nomikos, P. Monitoring Batch Processes. *NATO Advanced Study Institute for Batch Processing System Engineering*; Springer-Verlag: Heidelberg, 1992.

MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multi-Block PLS Methods. *AIChE J.* **1994**, *40*, 826−838.

Marsh, C. E.; Tucker, T. W. Application of SPC Techniques to Batch Units. *ISA Trans.* **1991**, *30*, 39−47.

Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1989.

Nomikos, P. Statistical Process Control of Batch Processes. Ph.D. Dissertation, McMaster University, Hamilton, Ontario, Canada, 1995.

Nomikos, P.; MacGregor, J. F. Monitoring of Batch Processes Using Multi-way Principal Component Analysis. *AIChE J.* **1994a**, *40*, 1361−1375.

Nomikos, P.; MacGregor, J. F. Multi-way Partial Least Squares in Monitoring Batch Processes. First International Chemometrics InterNet Conference, Sept 26−Nov 18, 1994b.

Nomikos, P.; MacGregor, J. F. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* **1995**, *37*, 41−59.

Peterson, T.; Hernandez, E.; Arkun, Y.; Schork, F. J. A Nonlinear DMC Algorithm and its Application to a SemiBatch Polymerization Reactor. *Chem. Eng. Sci.* **1992**, *47*, 737−753.

Piovoso, M. J.; Kosanovich, K. A.; Yuk, J. P. Process Data Chemometrics. *IEEE Trans. Instrum. Meas.* **1992a**, *41*, 262−268.

Piovoso, M. J.; Kosanovich, K. A.; Pearson, R. K. Monitoring Process Performance in Real-Time. *Proc. Am. Control Conf.* **1992b**, *3*, 2359−2364.

Smilde, A. K. Three-way Analyses, Problems and Prospects. *Chemom. Intell. Lab. Syst.* **1992**, *15*, 143−157.

Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397−405.

Received for review April 24, 1995
Revised manuscript received September 18, 1995
Accepted October 3, 1995[⊗]

IE9502594

[⊗] Abstract published in *Advance ACS Abstracts,* December 1, 1995.