Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis

Cenk Ündey,† Sinem Ertunç, and Ali Çınar*

Department of Chemical and Environmental Engineering, Illinois Institute of Technology, Chicago, Illinois 60616

An integrated online multivariate statistical process monitoring (MSPM), quality prediction, and fault diagnosis framework is developed for batch processes. Batch data from I batches, with J process variables measured at K time points generate a three-way array of size $I \times K \times J$. Unfolding this three-way array into a two-way matrix of size $IK \times J$ by preserving the variable direction is advantageous for developing online MSPM methods because it does not require estimation of future portions of new batches. Two different multiway partial least squares (MPLS) models are developed. The first model (MPLSV) is developed between the data matrix ($IK \times J$) and the local batch time (or an indicator variable) for online MSPM. The second model (MPLSB) is developed between the rearranged data matrix in the batch direction ($I \times KJ$) and the final quality matrix for online prediction of end-of-batch quality. The problem of discontinuity in process variable measurements due to operation switching (or moving to a different phase) that causes problems in alignment and modeling is addressed. Control limits on variable contribution plots are used to improve fault diagnosis capabilities of the MSPM framework. Case studies from a simulated fed-batch penicillin fermentation illustrate the implementation of the methodology.

1. Introduction

Online process performance monitoring and product quality prediction in real time are important in batch and fed-batch process operations. Many high-value specialty chemicals in pharmaceutical, biotechnology, polymer, and semiconductor manufacturing industries are manufactured using batch processes. Early detection of excursions from normal operation (NO) that might lead to deteriorated product, diagnosis of the source cause(s) of the deviation, and prediction of product quality in real time ensure safe and profitable operation and provide the opportunity to take corrective action(s) before the effects of disturbance(s) ruin the batch.

Process variable measurements such as reactant feed rates, aeration rate, and temperatures are frequently recorded in a typical batch process run, resulting in a data set containing time-varying trajectories. These trajectories contain valuable information for monitoring the performance of the process and can also be related to product quality measurements that usually become available at the end of the batch. Multivariate statistical projection methods such as principal component analysis (PCA) and partial least squares (PLS) have been used to develop multivariate statistical process monitoring (MSPM) techniques.[1,2] These techniques have become an effective alternative to conventional univariate statistical process control (SPC) and statistical quality control (SQC). PCA and PLS techniques have been extended to multiway PCA (MPCA) and multiway PLS (MPLS) to account for three-way data array decomposition of batch processes.[3,4] Illustrative applications in batch/semibatch polymerization have been reported.[4,5] MSPM frameworks including multivariate charts for both end-of-batch and online monitoring have been proposed.[6,7] MSPM techniques based on trilinear decompositions of three-way data array such as parallel factor analysis (PARAFAC) and Tucker models have also been suggested to monitor batch processes.[8−12]

Online monitoring of batch processes was challenging because the first generation of batch MSPM techniques required complete variable trajectories to the end of a run. Because the future portions of the process variable trajectories are not available during the progress of the batch, different assumptions are made to estimate the unmeasured parts of these trajectories including the use of missing value prediction capabilities of PCA and PLS.[4] These estimation approaches have been incorporated in MPCA for online SPM and used with MPLS and multiway covariates regression models for predicting the values of end-of-batch quality variables online.[5,12] Techniques that do not require future value estimation have also been suggested. Adaptive hierarchical PCA (AHPCA) develops recursive local PCA models relating previous observations in an exponentially weighted moving average manner.[13] Another technique uses dynamic PCA and PLS for online batch monitoring without future value estimation.[14] Orthogonal function approximation theory with PCA was also proposed for online batch MSPM.[15,16] Developing local MPCA, PARAFAC, and Tucker models by partitioning the total run time of a batch with respect to some scheduling points has been suggested as well.[17] Comparative studies on the performance of these techniques are available.[18,19]

In addition to missing values and outliers, batch process data analysis has additional challenges such as unequal lengths of historical data sets and unsynchronized batch trajectories. A variety of techniques has been suggested for addressing the problem of unequal

* To whom correspondence should be addressed. Fax: (312)-567 8874. E-mail: cinar@iit.edu.

† Current address: Amgen Inc., 40 Technology Way, West Greenwich, RI 02817.

10.1021/ie0208218 CCC: $25.00
Published on Web 07/23/2003

and unsynchronized trajectories. A relatively simple approach is to use an indicator variable and resample the process variables with respect to this variable instead of time.[20] Illustrative examples have been reported for batch/semibatch industrial polymerization[20,21] and fermentations.[22,23] Another alignment method is the dynamic time-warping technique developed in the speech recognition community.[24,25] It has been applied to synchronization of batch trajectories[22,26] and detection of process phases in bioprocesses.[27] Recently, the curve registration technique was suggested to align batch trajectories and detect process landmarks; illustrative results were given for fed-batch fermentations.[22]

Most MSPM techniques for batch processes rely on unfolding the three-way data array by preserving the batch direction. A different online MSPM framework can be established by unfolding the three-way data array by preserving the variable direction.[28−30] In this MSPM framework, it is not necessary to estimate the future portions of variable trajectories. MPCA or MPLS models can be developed and used for online monitoring. Wold et al.[29] have proposed a methodology by developing an MPLS model between the process variable matrix that is unfolded in the variable direction and the local time stamp to use in the alignment of trajectories.

Enhancements to this MSPM framework with online quality prediction and variable-contribution analysis are reported in this study. Trajectory alignment and modeling problems caused by the discontinuity problem due to batch/fed-batch switching are addressed. Mean centering of nonlinear score trajectories is performed as an additional step to construct control limits resulting in better fault detection performance. Control limits on variable contributions to various MSPM statistics (scores, T^2, SPE, and average deviations) are constructed and adapted to provide improved fault diagnosis insight. A quality prediction methodology is also incorporated based on data partitioning with respect to progress of the batch. Simulated data from fed-batch penicillin fermentation are used to illustrate the method.

The paper is organized as follows. Batch data challenges are summarized in section 2. Alignment-related issues and different techniques for trajectory alignment are discussed and compared in section 2.1, followed by the discussion of unfolding methods and their effect on online SPM in section 2.2. Steps involved in the development of online and end-of-batch SPM frameworks are presented in section 3. Section 4 contains illustrative examples for monitoring a fed-batch penicillin fermentation process.

2. Batch Data Challenges and Remedies

Analysis of batch process data offers a variety of challenges. It is common in batch process operation that the total duration of the batches and/or the duration of individual phases within a batch run are not the same. These differences may be caused by seasonal changes in environmental variables (e.g., coolant temperature variations), variations in quality and impurity concentrations of raw materials, unpredictable microbial responses to slight changes in batch bioprocesses, and arbitrary termination of batches by plant operators. The unequal batch data length causes problems for vector-matrix calculations involved in empirical modeling. In addition, critical local features in process variables in each batch corresponding to certain phases of process

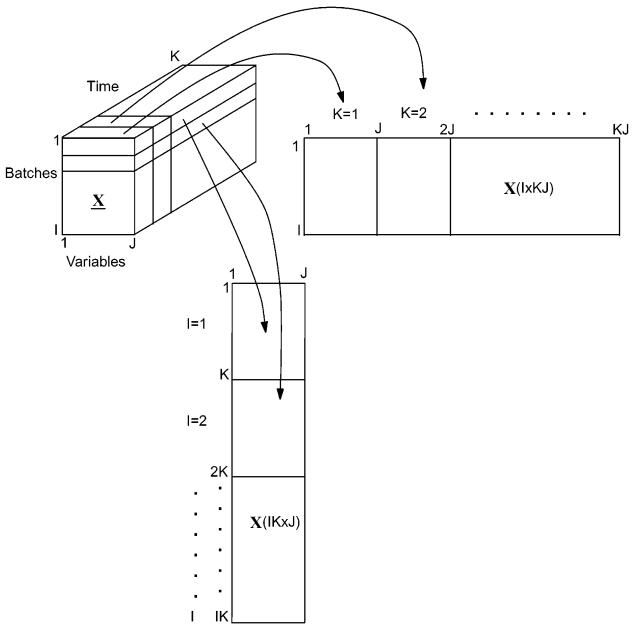

Figure 1. Three-way array formation and unfolding.

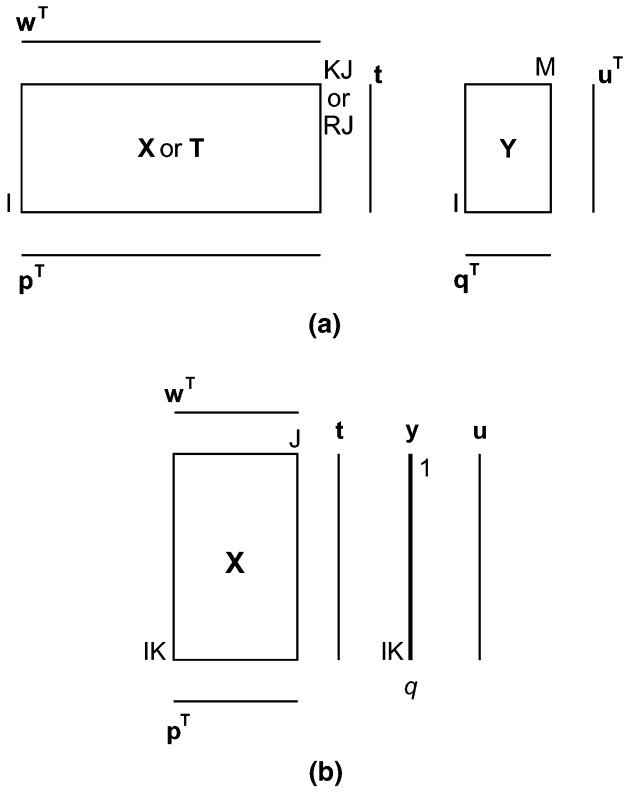

Figure 2. MPLS modeling using different unfolding approaches: (a) MPLS model blocks for predicting the product quality; (b) MPLS model blocks for predicting the progress of the batch.

dynamics may occur in different times, resulting in unsynchronized batch profiles. Various techniques to address these problems are discussed briefly in section 2.1.

Data pretreatment is performed prior to analysis when batch process data include noisy and collinear data or have outliers and missing values. The collinearity problem is overcome by using subspace empirical modeling techniques such as PCA and PLS. Prediction capabilities of these techniques can also be used to

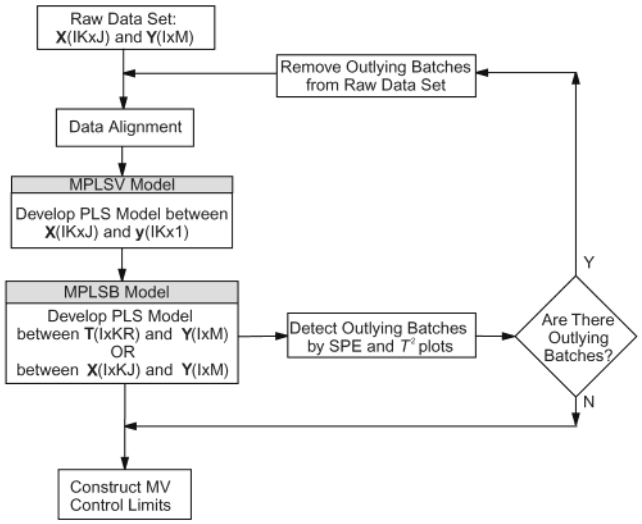

Figure 3. Overview of online SPM methodology.

detect and remove outlying data values or estimate the values of missing data.

2.1. Alignment of Variable Trajectories. Different techniques have been suggested to overcome unequal and unaligned batch process data problems. *Dynamic time warping* (DTW) locally translates, compresses, and expands the patterns so that similar features are aligned.[24,25] Recent applications for data alignment of batch polymerization[26] and batch fermentation are reported.[22,27] *Curve registration* (CR) technique[31] has been suggested to align batch trajectories with respect to process landmarks.[22,32] It is a twofold process of identifying landmarks within a trajectory, followed by warping of the test trajectories to the reference trajectory containing landmark locations.

The *indicator variable* (IV) technique provides a simpler alternative. In this technique, a variable is selected to indicate the progress of the batch instead of time. This variable should show the maturity of the evolving batch, should be smooth, monotonically increasing or decreasing, and should span the operation range for all variables. New observations are taken relative to the progress of this variable. The data alignment technique used in this study is a variant of the IV technique. DTW or CR techniques can also be used without loss of generality. The MPLS model developed between process measurements matrix $\mathbf{X}$ and local batch time vector $\mathbf{z}$ generates a predicted local time vector that can be considered as a maturity index, which has contributions from a wide range of process variable trajectories. This variable can be used to align each batch in the reference set by interpolation, resulting in aligned variable profiles with an equal number of measurements. The implementation of this approach becomes challenging when there are discontinuities in the process operation, as is the case of fed-batch fermentations. Solutions to this problem are discussed in section 3.

2.2. Unfolding Batch Data Array and Its Effect on Online SPM and Quality Prediction. Process measurements made on J variables at K time intervals for I batches are arranged into a three-way array $\underline{\mathbf{X}}$ of size $I \times K \times J$ (Figure 1) after equalization and alignment of variable trajectories. Product quality measurements at the end of the batch on M variables form a matrix $\mathbf{Y}$ of size $I \times M$. The three-way array $\underline{\mathbf{X}}$ can be unfolded into a two-dimensional matrix of the form $\mathbf{X}$ in six different ways. Only two of them, $I \times KJ$ and $IK \times J$ presented in Figure 1, are useful for MSPM. To differentiate the two MPLS techniques with different types of unfolding, the conventional MPLS technique that preserves the batch direction ($I \times KJ$ unfolding) is called MPLSB (Figure 2a) and the one that preserves the variable direction ($IK \times J$ unfolding) is called MPLSV (Figure 2b). MPLSV can be combined with MPLSB to predict the final product quality.[29] This has been implemented to predict the end-of-batch quality during the progress of the batch in section 3.3.

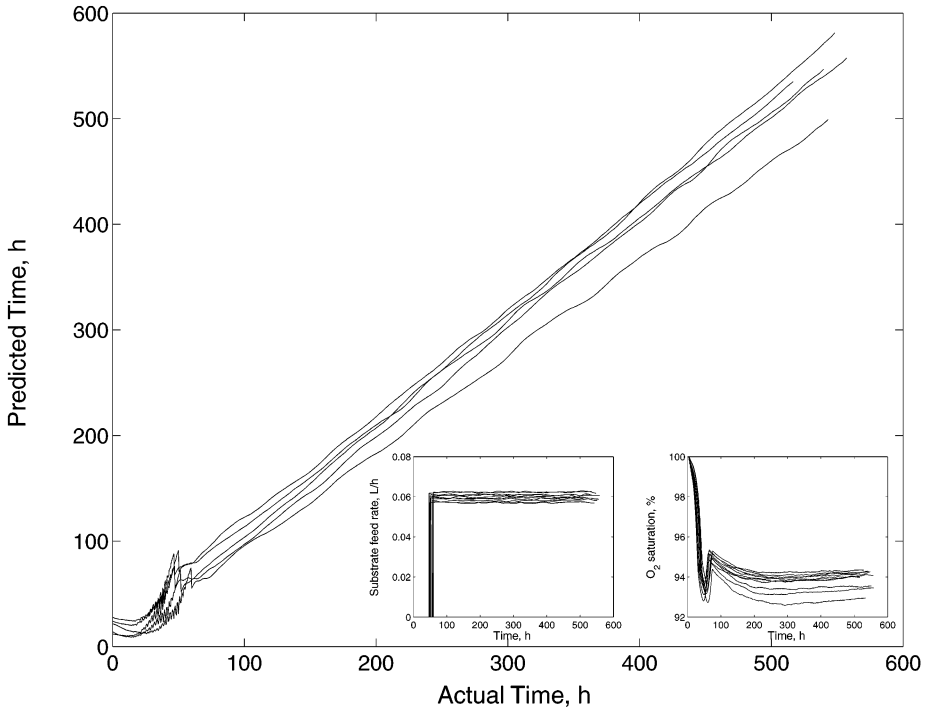

Figure 4. Predicted local batch time for the entire process duration. The peak corresponds to switching from batch to fed-batch operation.

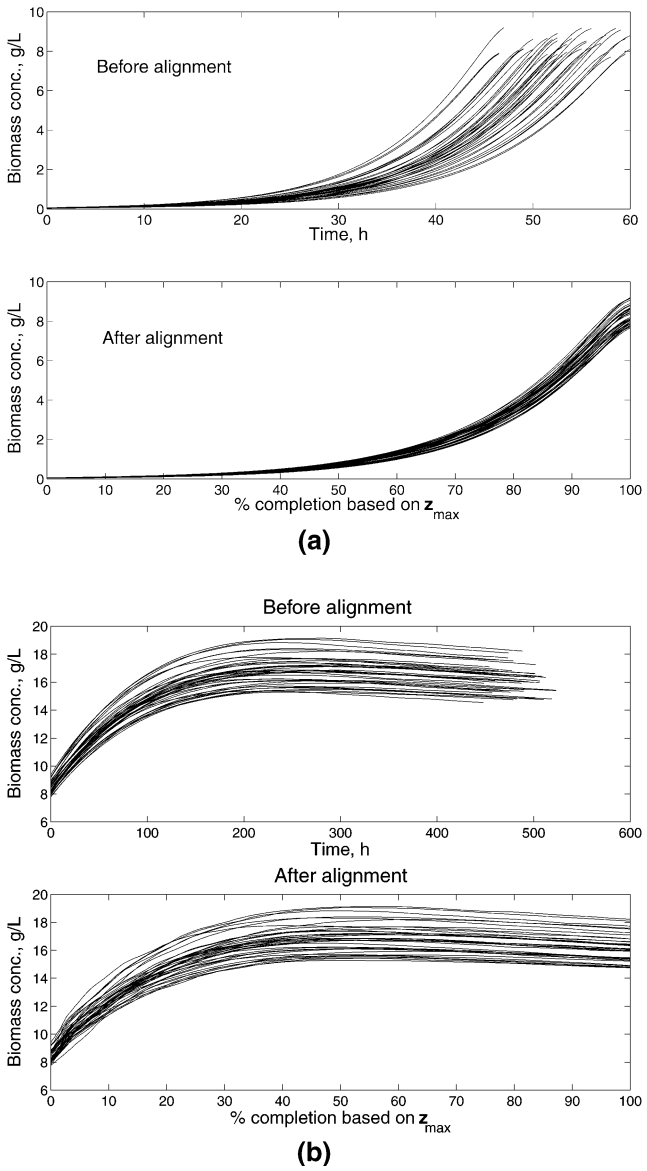

Figure 5. Results of the alignment procedure for biomass concentration profiles: (a) phase 1; (b) phase 2. The 0 on the time axis corresponds to the end of phase 1 time.

The MPLSB algorithm uses the unfolded matrix $\mathbf{X}$ ($I \times KJ$) and the quality measurements matrix $\mathbf{Y}$ ($I \times M$).[3] $\mathbf{X}$ is mean-centered to remove the nonlinear dynamic behavior of variable trajectories and is usually scaled to unit variance. An MPLS model with R latent variables, reflecting NO, is developed between the properly-scaled unfolded $\mathbf{X}$ matrix of process measurements collected from "good" batches that produced acceptable product quality and the $\mathbf{Y}$ matrix of final quality measurements of these "good" batches. This is achieved by decomposing $\mathbf{X}$ and $\mathbf{Y}$ into a combination of scores $\mathbf{T}$ ($I \times R$), loadings $\mathbf{P}$ ($KJ \times R$) and $\mathbf{Q}$ ($M \times R$), weights $\mathbf{W}$ ($KJ \times R$), and residual matrices $\mathbf{E}$ ($I \times KJ$) and $\mathbf{F}$ ($I \times M$) (Figure 2a) such that

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \quad \mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{1}$$

The data matrix $\mathbf{X}_{\mathrm{new}}$ ($K \times J$) of a new batch run that is monitored is unfolded and scaled to a $\mathbf{x}_{\mathrm{new}}$ vector of size $1 \times KJ$ to predict the scores vector $\hat{\mathbf{t}}$ ($1 \times R$) and values of the quality variables $\hat{\mathbf{y}}$ ($1 \times M$)

$$\hat{\mathbf{t}} = \mathbf{x}_{\mathrm{new}}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}, \quad \hat{\mathbf{y}} = \hat{\mathbf{t}}\mathbf{Q}^{\mathrm{T}} \tag{2}$$

$$\mathbf{e} = \mathbf{x}_{\mathrm{new}} - \hat{\mathbf{t}}\mathbf{P}^{\mathrm{T}}, \quad \mathbf{f} = \mathbf{y}_{\mathrm{new}} - \hat{\mathbf{y}} \tag{3}$$

to compare with those of reference batches.

MPLSB poses a problem for online MSPM in real time because the new batch data matrix $\mathbf{X}_{\mathrm{new}}$ is incomplete during the progress of the batch. To perform calculations in eqs 2 and 3 at each time interval k for online monitoring and final quality prediction in real time, future values of $\mathbf{x}_{\mathrm{new}}$ should be estimated from $k + 1$ to K. Four methods have been suggested for estimating future values of the process trajectories.[5,7,40] They introduce a certain level of arbitrariness to the MSPM performance, as reported in comparative studies.[5,7,13,18,19,40]

When process measurements array $\underline{\mathbf{X}}$ is unfolded to $\mathbf{X}$ ($IK \times J$) by preserving the variable direction,[8,28,29] it can be thought of as a combination of slices of matrices of size $K \times J$ for each batch (Figure 1). The evolution of the batch can be monitored by developing an MPLS model between $\mathbf{X}$ ($IK \times J$) and a time stamp vector $\mathbf{z}$ ($IK \times 1$) (Figure 2b).[29,33] In this case, MPLS decomposes $\mathbf{X}$ and $\mathbf{z}$ into a combination of scores matrix $\mathbf{T}$ ($IK \times R$), loadings matrix $\mathbf{P}$ ($J \times R$), vector $\mathbf{q}$ ($R \times 1$), and weight matrix $\mathbf{W}$ ($J \times R$), and residuals $\mathbf{E}$ and $\mathbf{f}$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \quad \mathbf{z} = \mathbf{T}\mathbf{q} + \mathbf{f} \tag{4}$$

During the progress of a new batch, a vector $\mathbf{x}_{\mathrm{new}}$ of size $1 \times J$ becomes available at each sampling time k. After application of the same scaling used with reference sets to the new observations vector, scores and batch progress (z_{pred}) at time k can be predicted by using the MPLS model parameters

$$\hat{\mathbf{t}}_k = \mathbf{x}_{\mathrm{new},k}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}, \quad z_{\mathrm{pred},k} = \hat{\mathbf{t}}_k\mathbf{q} \tag{5}$$

Because the dimensions of $\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}$ are $J \times R$, online monitoring of the new batch can be performed without estimation of future values.

In the preprocessing step, $\mathbf{X}$ is mean-centered and usually scaled to unit variance. The effect of this preprocessing differs from that of MPLSB because the dynamic nonlinear behavior of trajectories in $\mathbf{X}$ is retained. Mean centering in MPLSV refers to subtracting grand means of variables from the trajectories in $\mathbf{X}$. Because it is aimed to model the progress of process variable trajectories in MPLSV, variable trajectories are not linearized about the mean trajectory set.

3. Online Process Performance Monitoring Framework

The online monitoring framework in this work is based on unfolding of a three-way array by preserving the variable direction (MPLSV). In addition, MPLSB is used for online/offline quality prediction.

3.1. Model Development. Model development uses a reference data set that contains good batches presenting NO. Figure 3 summarizes the online monitoring framework that is combined with time alignment, data length equalization, and detection and removal of outlying batches from the reference set. Data alignment using an IV can be performed in different ways. If there exists a process variable that other process variables can be measured against its percent completion, it can

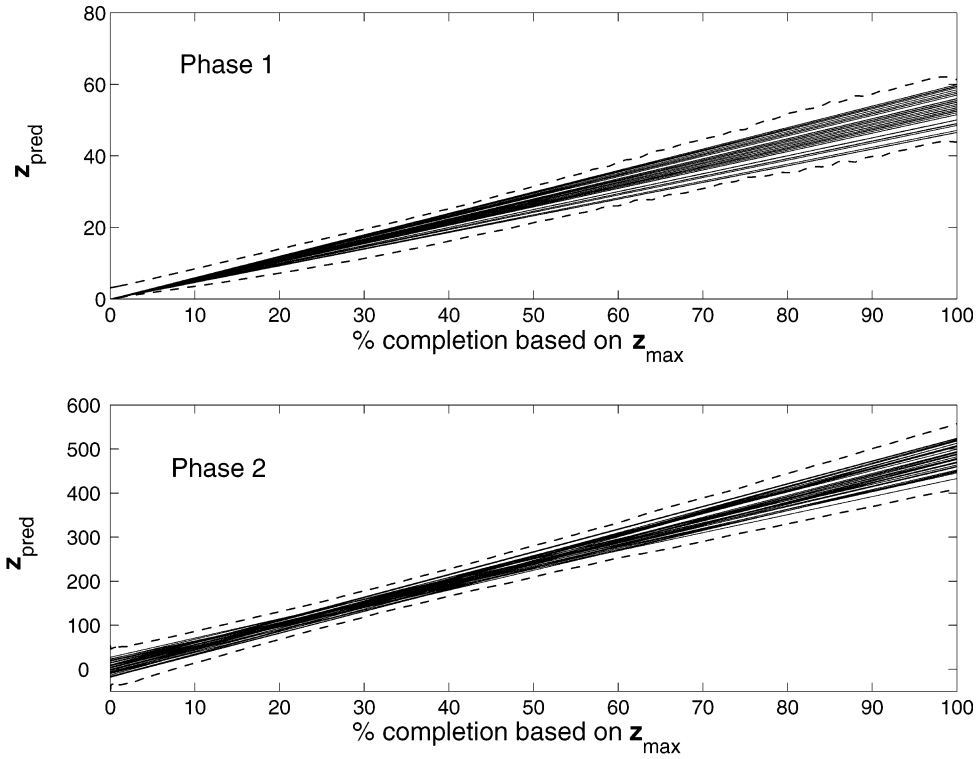

Figure 6. Predicted local batch times ($\mathbf{z}_{pred}$) in phases 1 and 2 with control limits (dashed lines).

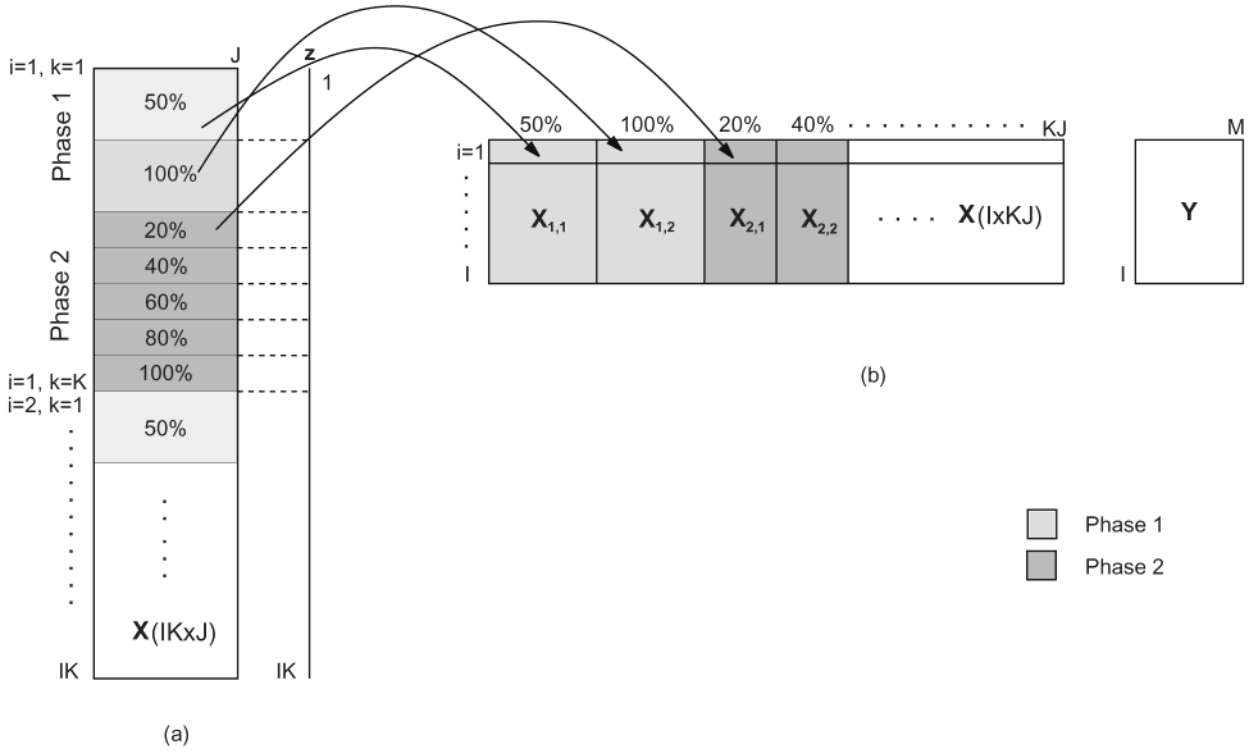

Figure 7. (a) Partitioning of process measurements space and (b) restructuring for online quality prediction framework.

be selected as the IV and trajectories of the variables in the reference set are resampled by linear interpolation techniques with respect to this IV. An alternative, especially when such an IV is not available, is to develop an MPLSV model between the process measurements matrix $\mathbf{X}$ and the local time stamps vector $\mathbf{z}$ of the individual batches in the reference set (Figure 2b; $\mathbf{y} = \mathbf{z}$). This MPLSV model provides the relationship between the local batch time ($\mathbf{z}$) and the evolution of process variable trajectories. The predictions $\mathbf{z}_{pred}$ can then be used as an IV, and process variables are resampled on percent increments of this derived variable. It is assumed that variable trajectories contain sufficient information to fairly predict batch time in MPLSV modeling. This assumption implies that variable trajectories somewhat linearly increase or decrease in each process phase. Local batch time prediction produces weak results when there are discontinuities

Table 1. Process and Product Quality Variables from Simulated Fed-Batch Penicillin Fermentation

process variable no.	definition
1	aeration rate[a]
2	agitator power input[a]
3	substrate feed rate[a]
4	substrate feed temperature[a]
5	substrate concentration
6	oxygen saturation (%)
7	biomass concentration
8	penicillin concentration
9	culture volume
10	carbon dioxide concentration
11	hydrogen ion concentration (pH)
12	temperature in the fermentor
13	generated heat
14	cooling water flow rate[a]
15	amount of substrate added (computed from variable 3)

quality variable no.	definition
y_1	final penicillin concentration
y_2	overall productivity
y_3	yield of penicillin on biomass
y_4	yield of penicillin on substrate
y_5	amount of penicillin produced

[a] Input variables.

or there exist instances in which variables have simultaneous piecewise linear dynamics during the evolution of the batch. As illustrated in Figure 4 with fed-batch penicillin fermentation data, the predicted time shows nonincreasing or decreasing behavior in the region around the discontinuity (Figure 4; batch/fed-batch switching at about 55 h, indicated by peaks), which makes it inappropriate for data alignment. Similar results were also reported for industrial data.[18]

This problem can be solved by partitioning the entire process into major operational phases. Two different data alignment methods are used. When batches in the reference data set are of unequal length and there is no appropriate IV that spans the whole batch run, an MPLSV model is developed between **X** and the local batch time stamps **z** *for each process phase*. Process variable trajectories are then resampled with respect to the percent completion of $\mathbf{z}_{\text{pred}}$. A vector $\mathbf{z}_{\max}$ containing predicted termination times of reference batches is used to calculate the percent completion on $\mathbf{z}_{\text{pred}}$. An alternative is to select appropriate IVs in each phase of operation. The discontinuity occurs in the transition from batch to fed-batch operation in penicillin fermentation (Figure 4). Consequently, there are two operational phases and two IVs are used. In this case, process termination is determined according to a maturity indicator such as a preset percent conversion level or a certain amount of a component is fed. The use of two different IVs for fed-batch penicillin fermentation is discussed in section 4.2. Figure 5 shows aligned biomass concentration profiles of the reference batches in each phase of the batch run. As a result of alignment, the number of measurements in each batch on each variable is equalized and temporal variation of process events is minimized so that similar events can be compared. $\mathbf{z}_{\text{pred}}$ of a new batch can be used as a maturity indicator. If its value is smaller than the observed value, the process is progressing more slowly than the reference batches. Limits are used to detect an unusual deviation from the expected time course of the batch. $\mathbf{z}_{\text{pred}}$ profiles calculated from the MPLSV model in each phase of the reference batches are plotted against actual **z** values in Figure 6 along with their control limits.

Once the reference data set of good batches is aligned to give an equal number of measurements in each batch and synchronized variable profiles, an MPLSV model is developed between the aligned process variables set and the percent completion of the batch run, **z**. Model parameters from this step are used to construct MSPM charts as outlined in section 3.2.

3.2. MSPM Charts. It is advantageous to use the MPLSV model for online monitoring of batch evolution because no future value estimation is required. However, this technique may produce weak results for small disturbances when the goal is detecting deviations from the mean trajectories. Nonlinear estimated score trajectories are obtained as a result of MPLSV modeling because nonlinear variable trajectories are used in the **X** matrix. Predicted score vectors of reference batches are gathered to form the reference scores matrix $\hat{\mathbf{T}}_{R,k}$. Each of the I reference batches is passed through the online MPLSV calculations (eq 5) as if they were new batches. Their scores ($\hat{\mathbf{t}}_{r,k}$, $r = 1, R$) forms the $\hat{\mathbf{T}}_{R,k}$ matrix at each sampling time k resulting in I observations on R scores

$$\hat{\mathbf{T}}_{R,k} = [\hat{\mathbf{t}}_{I1,k}, \hat{\mathbf{t}}_{I2,k}, ..., \hat{\mathbf{t}}_{Ir,k}, ..., \hat{\mathbf{t}}_{IR,k}] \tag{6}$$

The nonlinear dynamic behavior is removed from $\hat{\mathbf{T}}_{R,k}$ by subtracting mean score trajectories obtained from the MPLSV model and constructing multivariate statistical control limits by using this mean-centered score matrix. Average score traces $\bar{\mathbf{t}}_k$ and the covariance matrix of mean-centered reference scores $\mathbf{S}_k$ at time k are calculated from the $\hat{\mathbf{T}}_{R,k}$ matrix of size $I \times R$

$$\mathbf{S}_k = \frac{\hat{\mathbf{t}}_k{}^{\mathrm{T}}\hat{\mathbf{t}}_k}{I-1} \tag{7}$$

The *score plots* of latent variables are used to detect any departure from the in-control region defined by the confidence limits calculated from the reference set. Control limits for scores can be computed using Student's t distribution statistics after checking that the scores have almost normal distribution. The control limits for new independent **t** scores under the assumption of normality are defined as[34]

$$\pm t_{n-1,\alpha/2} s_{\text{ref}} (1 + 1/n)^{1/2} \tag{8}$$

where $t_{n-1,\alpha/2}$ is the critical value of the Student's t test with $n - 1$ degrees of freedom at significance level $\alpha/2$ and n and s_{ref} are the number of observations and the estimated standard deviation, respectively, of the **t** score at time k.

If the distribution of scores is significantly different than normal, $\bar{\mathbf{t}} \pm 3\boldsymbol{\sigma}$ should be used where $\bar{\mathbf{t}}$ are average estimated scores and $\boldsymbol{\sigma}$ their standard deviations[29]

$$\bar{\mathbf{t}}_k = \frac{1}{I}\sum_{i=1}^{I} \hat{\mathbf{T}}_{iR,k}, \quad \sigma_k = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I} (\hat{\mathbf{T}}_{iR,k} - \bar{\mathbf{t}}_k)^2} \tag{9}$$

The T^2 *chart* detects small shifts and deviations from NO defined by the model. T^2 and its statistical limits

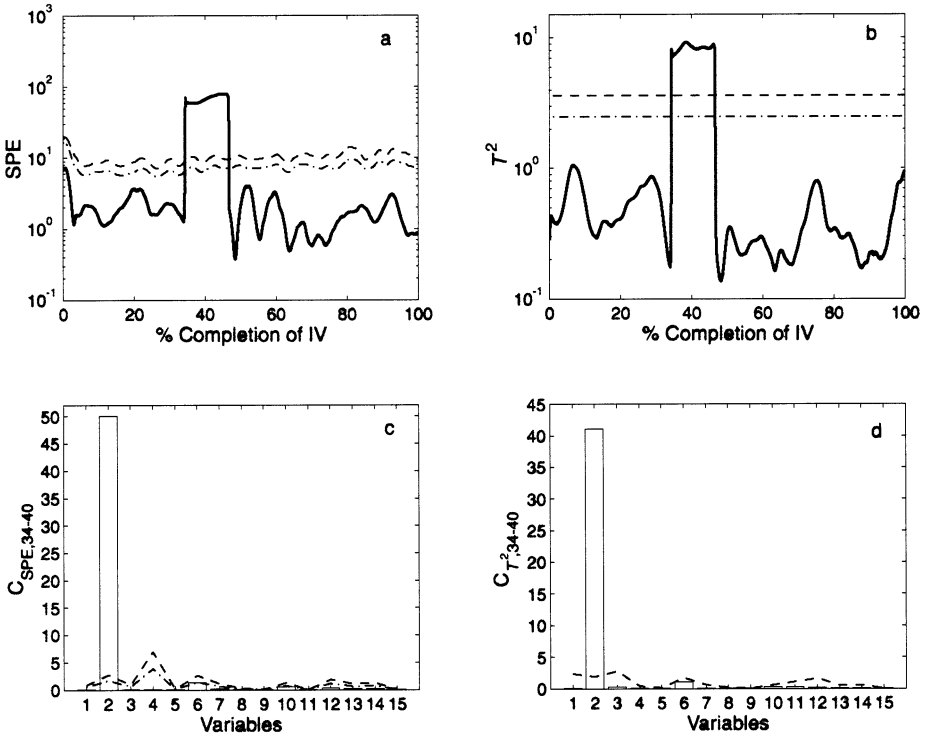

Figure 8. Control charts for SPE and T^2 for the entire process duration and contributions of variables to SPE and T^2 for a selected interval after an out-of-control signal is detected, for case 1 in phase 2 with 95% and 99% control limits (dash-dotted and dashed lines).

are also calculated by using the mean-centered score matrix. T^2 values at each time k follow an F distribution[6]

$$T_k^2 = (\hat{\mathbf{t}}_{\text{new},k} - \bar{\mathbf{t}}_k)^{\mathrm{T}}\mathbf{S}_k^{-1}(\hat{\mathbf{t}}_{\text{new},k} - \bar{\mathbf{t}}_k)\frac{I(I-R)}{R(I^2-1)} \sim F(R,I-R) \quad (10)$$

where $\hat{\mathbf{t}}_{\text{new},k}$ is the predicted score vector of the new batch calculated using eq 2, I the number of batches in the reference set, and R the number of latent variables retained in the model.

The *squared prediction error* (SPE) chart shows large variations and deviations from NO that are not defined by the model. SPE values that are calculated at each time k using eq 3 are well approximated by

$$\text{SPE}_k = \sum_{j=1}^{J} \mathrm{e}_{jk}^2 \sim g\chi_h^2 \quad (11)$$

where g is a constant and h is the effective degrees of freedom of the χ^2 distribution.[35]

Contribution plots are used for fault diagnosis. T^2 and SPE charts and score plots produce an out-of-control signal when a fault occurs, but they do not provide any information about the fault and its cause. Contribution plots for T^2, SPE, and scores show which variable(s) are responsible for inflating T^2, SPE, or scores to indicate deviation from NO. Contributions to SPE can be calculated by

$$C_{\text{SPE},ijk} = \mathrm{e}_{ijk}^2 \quad (12)$$

where $C_{\text{SPE},ijk}$ is the contribution of batch i to the SPE value for process variable j at time k.[36,37]

Variable contributions to T^2 and mean-centered scores are calculated using a modification of the formulation in Nomikos[38] for MPLS

$$C_{T_k^2,j} = \sum_{r=1}^{R} \mathbf{S}_{rr}^{-1} t_{\text{new},r_k} \mathbf{x}_{\text{new},k} \mathbf{W}_{r,j}^{*} \quad (13)$$

where matrix $\mathbf{W}^*$ of size $R \times J$ is defined as $\mathbf{W}^* = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}$.

Recently, control limits have been suggested for contribution plots.[33,37] These limits are adapted in this work for contributions to T^2 and SPE. Control limits for variable contributions to SPE (eq 12) also follow the χ^2 distribution; therefore, they can be calculated as defined in eq 11. For contributions to T^2, limits are computed by means of a jackknife procedure in which each batch in the reference set is left out once, and variable contributions are calculated for the batch that is left out. The next step is to calculate the mean and variance of these contributions from I batches for each jth variable at a kth time period. Westerhuis et al.[37] proposed to use an upper control limit (UCL) for contributions that is calculated as the mean of the variable contributions at each time interval plus 3 times the corresponding standard deviation. Charting variable deviations from average trajectories at each time instant can also be used as a diagnostic tool.[29] The same jackknife procedure is also used to construct control limits for these deviations. They provide information on how variable trajectories are deviating about the mean trajectories, but their univariate nature hinders effective diagnosis especially in the case of drift types of disturbances.

When used as is, MPLSV modeling produces nonlinear estimated scores along with control limits described earlier. When a new batch is monitored with the model parameters of MPLSV, estimated scores of this new batch will also be nonlinear. After mean centering of

Table 2. Fault Detection Times for Case 1

type	% completed IV	time, h
SPE	34.4	200
T^2	34.4	200
linear score LV4	34.4	200
linear score LV5	34.4	200
nonlinear score LV4	34.4	200
nonlinear score LV5	34.4	200
linear score LV3	36	206

Table 3. Fault Detection Times for Case 2

type	% completed IV	time, h
T^2	48.4	269
linear score LV2	50	276
linear score LV5	51.6	283
nonlinear score LV2	51.6	283
nonlinear score LV5	52	285
linear score LV4	59.4	319
nonlinear score LV4	60.6	324
linear score LV3	70.2	368
nonlinear score LV3	84.2	433
SPE	–	–

these scores to reduce their nonlinearity, it is possible to construct tighter control limits by using eq 8. This modification allows faster fault detection, as discussed in section 4. When an out-of-control status is detected with either type of score plot, variable contributions are checked for fault diagnosis.

3.3. Online Prediction of the Product Quality. MPLSV-type models lack the capability to predict online end-of-batch quality in real time. A two-step integrated modeling approach can provide online quality prediction. The first step uses MPLSV models for data partitioning. Data alignment followed by partitioning for quality prediction can also be performed using DTW or CR techniques. Because of its simplicity, the IV technique is chosen in this study. After reference batch data are aligned, batch data are partitioned according to percent increments of batch progress (for example, based on the local batch time or another IV) so that batches in the reference set are partitioned based on arbitrarily chosen increments such as 10%, 20% of $\mathbf{z}_{\text{pred}}$ (Figure 7a). Each partition of $\mathbf{X}$ ($IK \times J$) is rearranged into a matrix $\mathbf{X}$ ($I \times KJ$) as shown in Figure 7b. This provides a transition between MPLSV and MPLSB modeling to permit the development of an MPLSB model between this partial data and the final product quality matrix $\mathbf{Y}$. This enables the prediction of end-of-batch quality on percent progress points reflected by partitions. The number of quality predictions during the progress of the batch will be equal to the number of partitions.

Confidence intervals at significance level α are also suggested for the predicted quality values,[5]

$$\hat{\mathbf{y}} \pm t_{I-R-1,\alpha/2}(\text{MSE})^{1/2}[1 + \hat{\mathbf{t}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\hat{\mathbf{t}}^{\mathrm{T}}]^{1/2} \quad (14)$$

where $t_{I-R-1,\alpha/2}$ is the critical value of the Student's variable with $I - R - 1$ degrees of freedom at significance level $\alpha/2$, $\mathbf{T}$ and MSE are the score matrix and mean-squared error from the MPLSB model, and $\hat{\mathbf{t}}$ is the estimated scores vector of the new batch. An illustration of the quality prediction method is presented in case 3 of section 4.2.

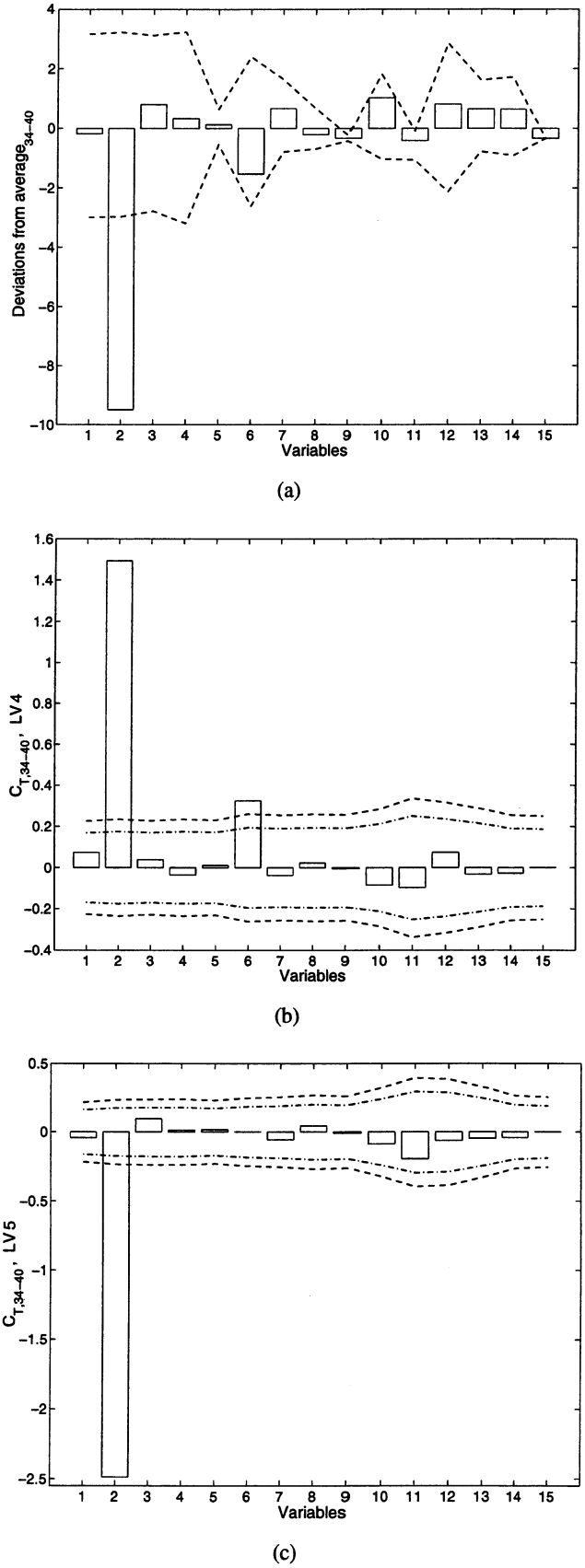

Figure 9. Variable contributions in phase 2 of case 1 to (a) deviations from the average batch behavior, (b) to linearized LV4 score, and (c) to linearized LV5 score, all calculated at the interval of 34−40% completion of IV with control limits (dash-dotted and dashed lines).

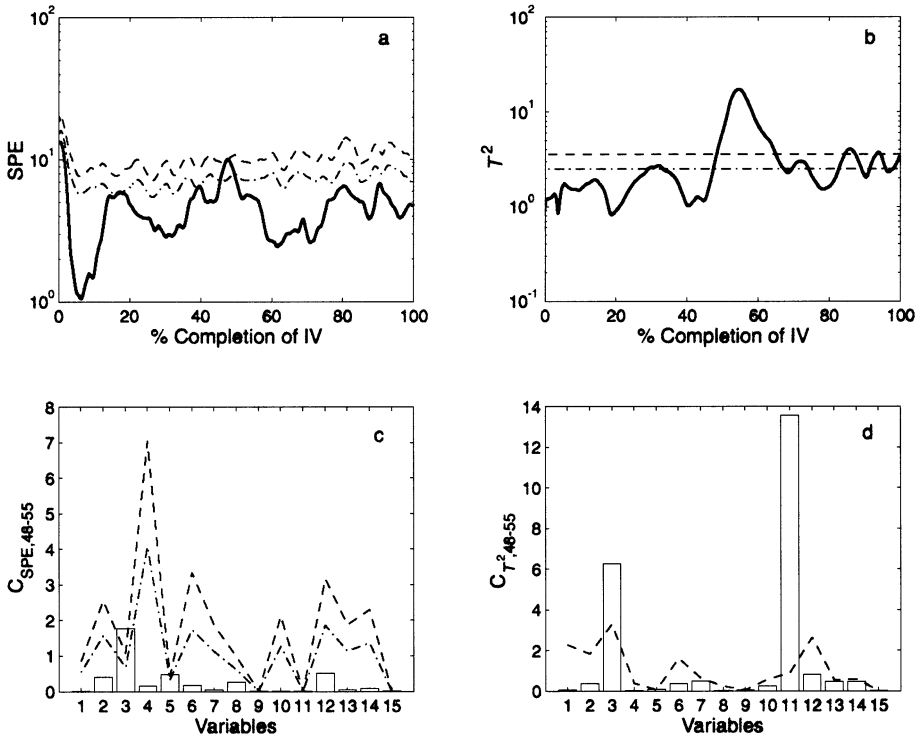

Figure 10. Control charts for SPE and T^2 for the entire process duration and contributions of variables to SPE and T^2 for a selected interval after an out-of-control signal is detected, for case 2 in phase 2 with 95% and 99% control limits (dash-dotted and dashed lines).

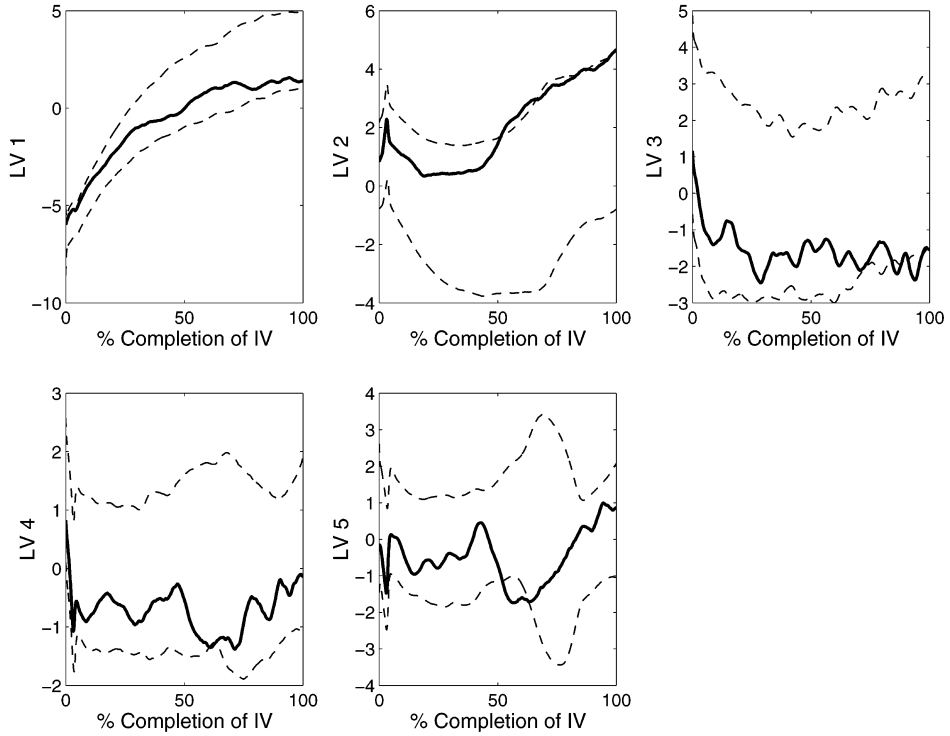

Figure 11. Nonlinear scores for case 2 in phase 2 with control limits (dashed lines).

4. Case Study: Monitoring the Fed-Batch Penicillin Fermentation Process

4.1. Fed-Batch Penicillin Fermentation Model. Fed-batch penicillin fermentation process data are generated using a detailed mathematical model and a simulator.[39] The model has five input variables (1−4 and 14), nine process variables (5−13), and five quality variables (Table 1). Feedback controllers regulate variables 11 and 12. Penicillin fermentation has four physiological phases (lag, exponential cell growth, stationary, and cell death) and two operational phases. The first two physiological phases are conducted as batch operation (first operational phase) while the last two are conducted as fed-batch operation. In the first operational phase, fermentation is carried out in a batch mode to promote biomass growth resulting in high cell densities. The second operational phase is a fed-batch operation where glucose is fed until the end of the fed-batch

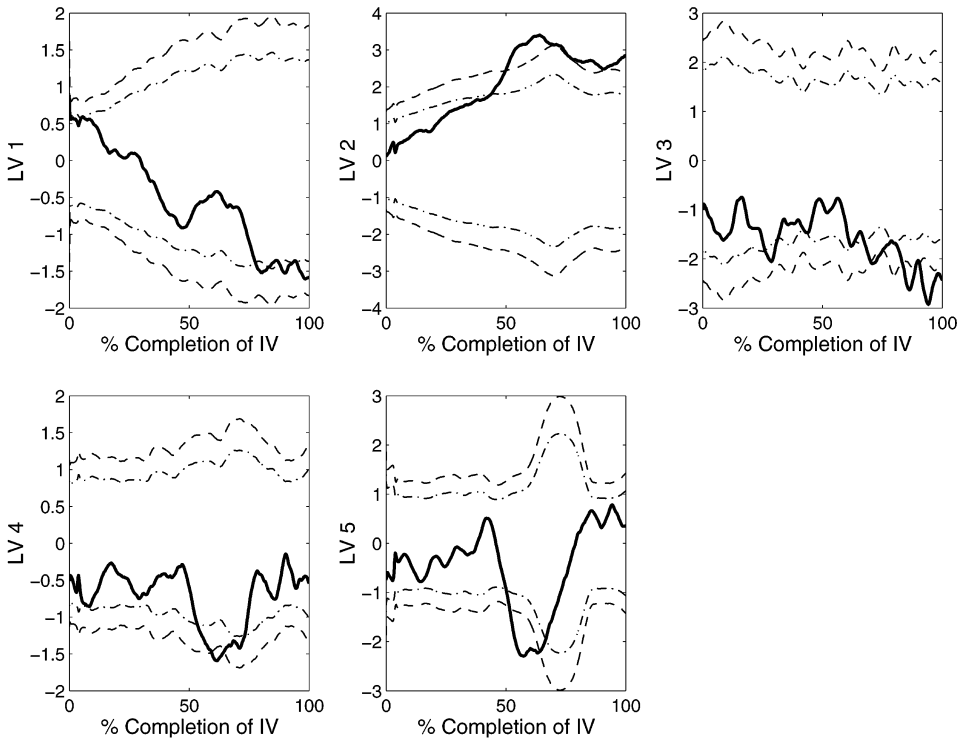

Figure 12. Linear scores for case 2 in phase 2 with 95% and 99% control limits (dash-dotted and dashed lines).

operation. In a batch fermentation process that lasts several days, some microorganisms may have different generation times. Slight changes in operating conditions during critical periods may have a significant influence on the growth and differentiation of microorganisms and may impact the final product quality and yield. To simulate the physical uncertainty present in each batch due to variable metabolic responses, very small perturbations are introduced into the parametric space and input variables used in the simulator while generating the batch data set.

4.2. Online Monitoring with IV-Based Alignment. The data set has 40 batches containing 14 variables (Table 1; variable 15 is computed from variable 3). Each batch has a different completion time resulting in an unequal number of measurements on each variable. The data set is preprocessed for partitioning according to operational phases. First, a batch/fed-batch switching point is found for each batch and data are divided into two sets as phases 1 and 2. Every step of the flowchart in Figure 3 is applied to these data sets separately. Because variable 3 (substrate feed) is zero in the batch phase, only 13 variables are left in this first set. Data alignment is performed by using the IV technique. Because an IV is not available for the entire batch run, separate IVs are selected for each phase. Variable 9 (culture volume) is a good IV candidate for phase 1. A new variable called "percent substrate fed" (variable 15) is calculated from variable 3 and used as IV for phase 2. It is assumed that the fed-batch phase is completed when 25 L of substrate is added to the fermenter. Data are resampled by linear interpolation at each 1% completion of volume decrease for phase 1 and at each 0.2% of total substrate amount added for phase 2. Data alignment results in an equal number of data points in each phase such that data lengths in each phase are $K1 = 101$ and $K2 = 501$.

Table 4. Explained Variance of MPLSB Models for Online Quality Prediction

model no.	**X** block	**Y** block	model no.	**X** block	**Y** block
1	61.57	68.85	5	63.10	97.31
2	61.27	71.27	6	63.35	98.39
3	58.85	89.21	7	63.39	98.89
4	60.62	95.07			

An MPLSV model is developed for phase 1 between autoscaled **X1** ($IK1 \times J1$) and the IV vector **z1** ($IK1 \times 1$) by using five latent variables. Cross validation is used to determine the number of latent variables. **X1** ($IK1 \times J1$) can be rearranged into matrix **X1** ($I \times K1J1$) to develop an MPLSB model to obtain an estimate of end-of-batch quality at the end of phase 1. Because all $K1$ measurements of phase 1 have been recorded by the beginning of the second phase, there would be no estimation of variable trajectories required and $I \times KJ$ unfolding can be used for modeling and quality prediction (section 3.3). Autoscaled **X1** ($I \times K1J1$) and product quality matrix **Y** ($I \times M$) are used as predictor and predicted blocks, respectively, as illustrated in case 3. Similarly, another MPLSV model is developed for phase 2 between autoscaled **X2** ($IK2 \times J2$) and IV vector **z2** ($IK2 \times 1$).

Process variables for the new batch are sampled at percent increments of volume decrease for phase 1. After the completion of phase 1, the sampling time is switched to percent completion of the amount of substrate added. New data vector $\mathbf{x}_{new}$ ($1 \times J$) is monitored by using the following algorithm at each sampling time k.

For $k = 1, ..., K$ ($K = K1$ or $K2$), perform the following:

1. Obtain new batch data: $\mathbf{x}_{new}$ ($1 \times J$).

2. Calculate new batch scores, SPE, T^2, and variable contributions to these statistics by using the information generated by the MPLSV model (eqs 5 and 10−13).

3. Compute z_{pred} to determine the batch progress rate (eq 5).

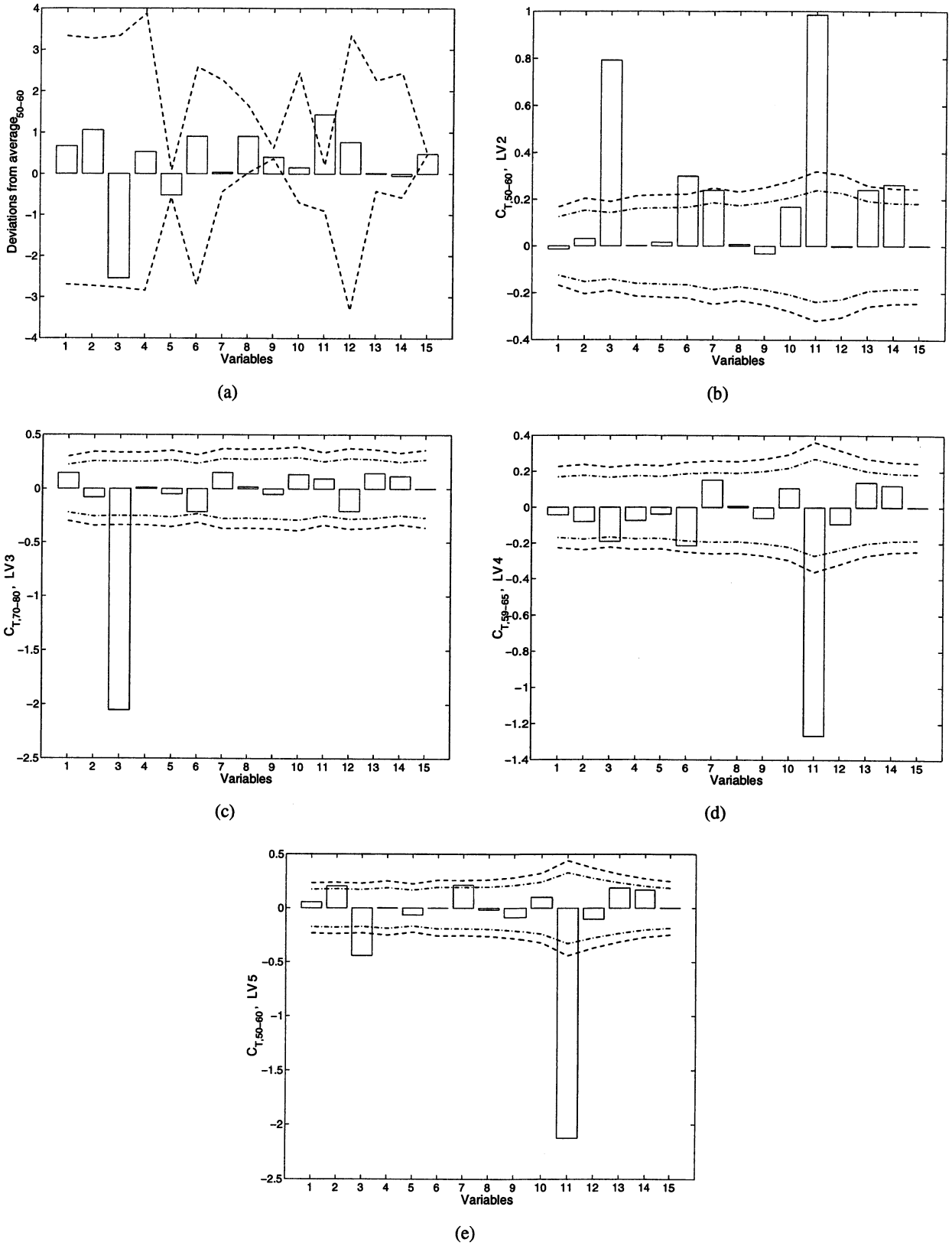

Figure 13. Variable contributions for case 2 in phase 2 to (a) deviations from the average batch behavior at the interval of 50−60% completion of IV, (b) to linearized LV2 score at the interval of 50−60% completion of IV, (c) to linearized LV3 score at the interval of 70−80% completion of IV, (d) to linearized LV4 score at the interval of 59−65% completion of IV, (e) to linearized LV5 score at the interval of 50−60% completion of IV, all with control limits (dash-dotted and dashed lines).

4. Plot data to MSPM charts and check for abnormalities.

End.

Three cases with different types of faults are presented: a step decrease in agitator power (bias change) and two drift cases of magnitudes −0.018% and −0.05%

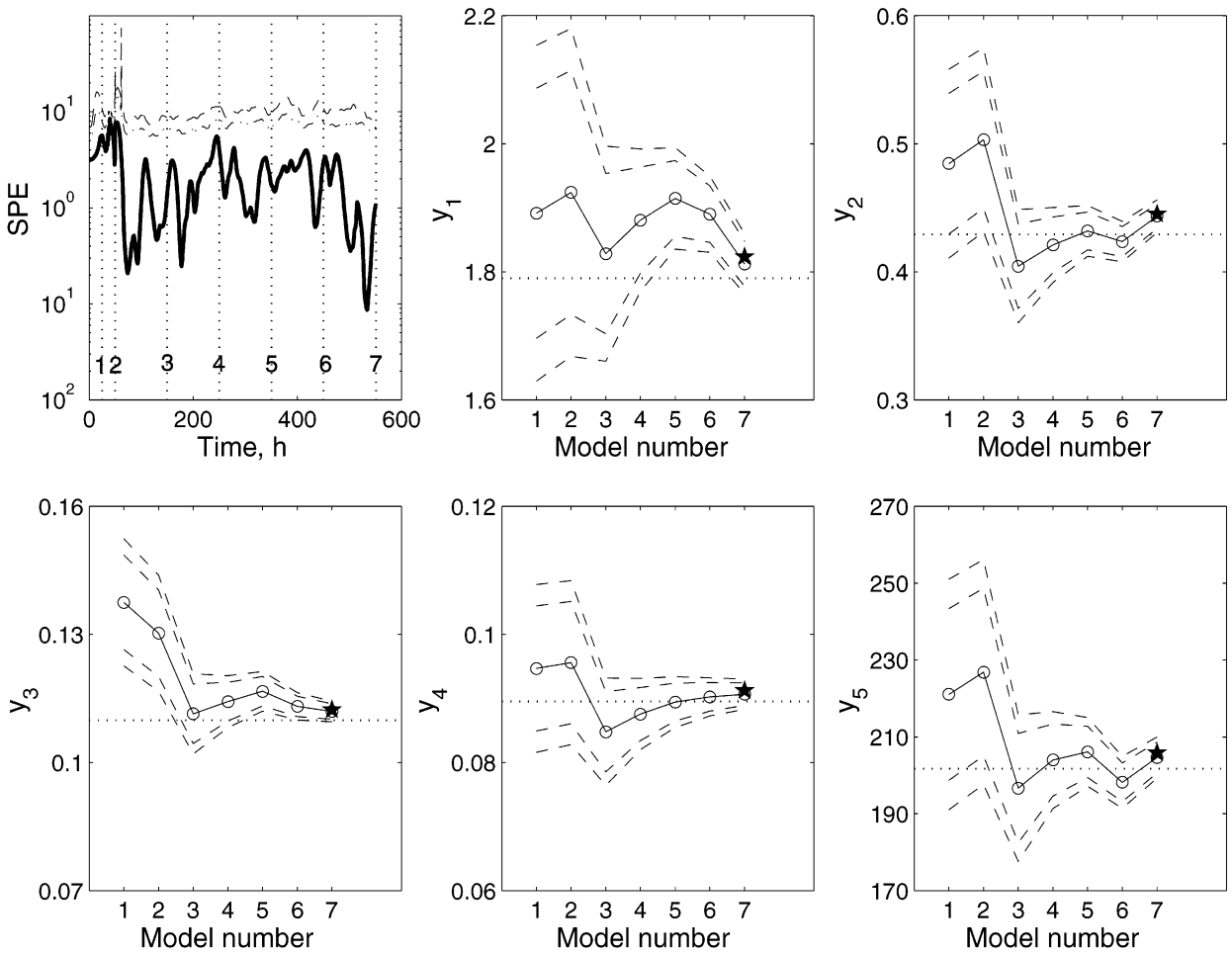

Figure 14. Online predictions for end-of-batch quality values for a normal batch.

h^{-1} in the substrate feed rate (drift), respectively. Fault detection times for different multivariable control charts are recorded and compared. The process is assumed to be out-of-control if three consecutive points are out of the 99% confidence limit.

Case 1. The first test case is a step decrease in agitator power input (variable 2) of magnitude 25% introduced starting at 200 h from the beginning of the batch until 250 h. The fault is detected in both SPE and T^2 plots (Figure 8). Latent variables 4 and 5 of nonlinear score plots and latent variables 3−5 of linear score plots (obtained by subtracting mean trajectories) detected the fault nearly at the same time (Table 2). Contributions to SPE and T^2 (Figure 8c,d), deviation from the average batch behavior (Figure 9a), and contributions to latent variables (Figure 9b,c) are inspected to identify process variables that have affected the batch. $C_{\cdot,34-40}$ denotes the contribution of each process variable to statistic "$\cdot$" (T^2, SPE, or scores) in the interval 34−40% completion of the IV. Variable 2, the agitator power input, is identified correctly to be the source of the deviation in all contribution plots. In addition, the contribution plot for latent variable 4 indicates variable 6 (dissolved O_2) as an influential variable, which is consistent with the fundamentals of the process.

Case 2. The second test case is a small drift of magnitude −0.018% h^{-1} introduced into the substrate feed rate (variable 3) from the start of fed-batch operation at 50 h until the end of the batch run. There are significant differences in fault detection times and out-of-control signal generation in this case (Table 3). T^2 detected the fault fastest (Figure 10). The second fastest detection is obtained by the linear score control chart of latent variable 2 (Figures 11 and 12). The last four latent variables give out-of-control signals for both linear and nonlinear score matrices. Although SPE is in control throughout the course of the batch, the contribution plot for SPE signals an unusual situation for variable 3 (Figure 10c). Variables 3 and 11 (pH) are found to be the most affected variables by the fault according to the T^2 contribution plot. Deviation from the average batch behavior plot is ineffective in indicating the most affected variable(s) in this case (Figure 13a).

Case 3. In this case, the online quality prediction method presented in section 3.3 is tested. Two different IVs are used to align the two distinct operational phases (batch and fed-batch). To develop the MPLSB model for the first phase, data are collected in 50% increments of phase 1, resulting in two data partitions $\mathbf{X}_{1,1}$ and $\mathbf{X}_{1,2}$ (Figure 7b). A similar approach in phase 2 for every 20% increase results in five data partitions ($\mathbf{X}_{2,n}$, $n = 1, ..., 5$). MPLSB modeling is performed between the rearranged $\mathbf{X}$ matrix, which is augmented as a new partition becomes available, and the $\mathbf{Y}$ matrix containing end-of-batch product quality measurements on five variables for each batch in the reference set listed in Table 1. Table 4 summarizes the variance of both $\mathbf{X}$ and $\mathbf{Y}$ blocks explained as new MPLSB models are developed when more data partitions are appended to $\mathbf{X}$. The variance

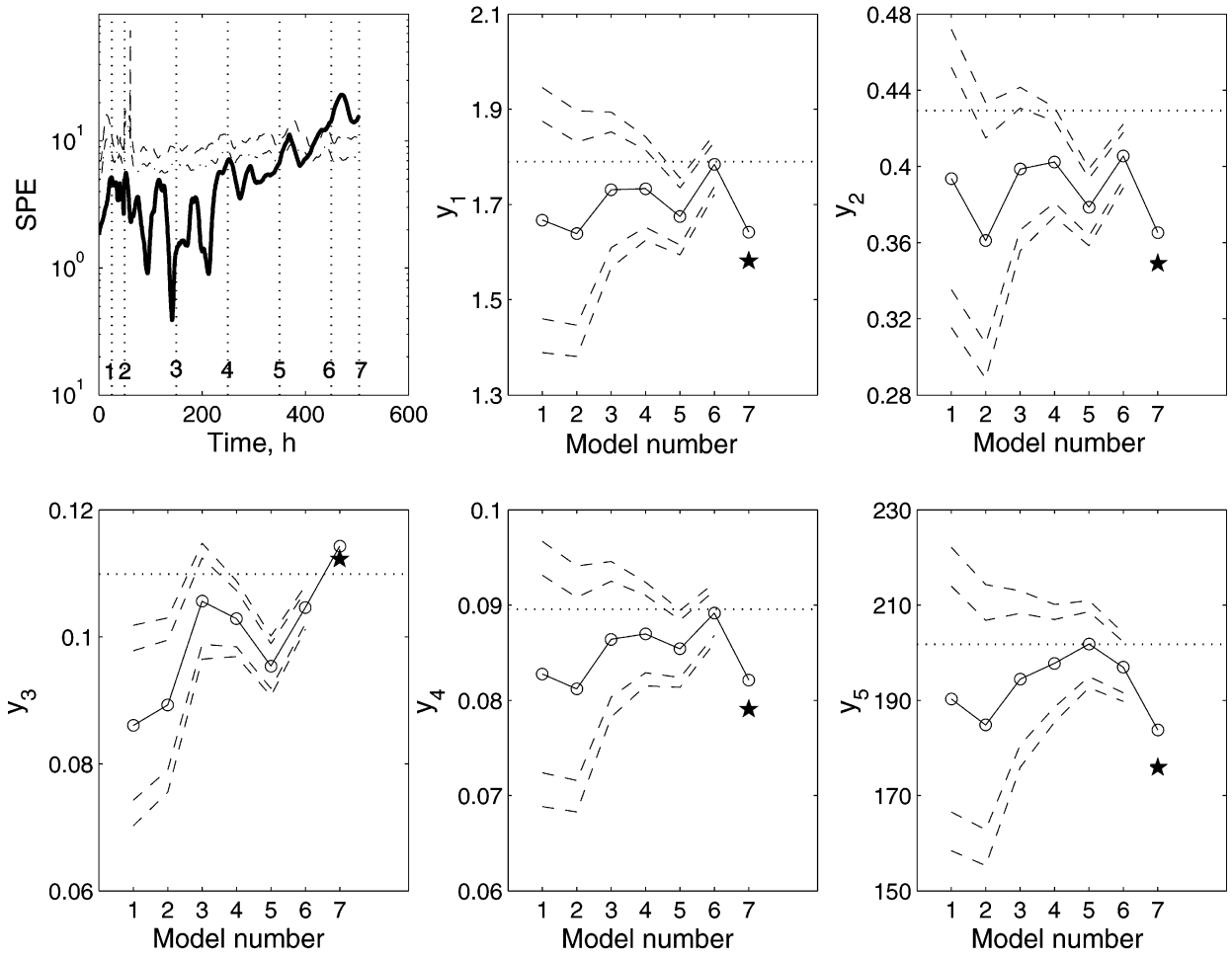

Figure 15. Online predictions for end-of-batch quality values for the faulty batch in case 3.

of **Y** explained increases as expected because more process information becomes available with each additional model.

Two cases are considered to test this integrated framework. A normal batch is investigated first. As expected, the SPE plot produced no out-of-control signal and the final product quality on all five variables (shown as a solid star) is successfully predicted (Figure 14). The prediction capability is somewhat poor in the beginning because of limited data, but it gets better as more data become available. In the second case, where a drift of magnitude $-0.05\%\ h^{-1}$ is introduced into the substrate feed rate at the beginning of the fed-batch phase until the end of operation, the SPE plot signaled out-of-control right after the sixth quality prediction point (80% completion of phase 2). Because the MPLSB model is not valid beyond this point, no further confidence limit is plotted (Figure 15). Although the predictions of the MPLSB model might not be accurate for the seventh (and final) value, the framework generated fairly close predictions of the inferior quality.

5. Conclusions

An MSPM framework is presented for online batch process performance monitoring and fault diagnosis. Unfolding the three-way data array by preserving the variable direction allowed online monitoring without requiring future value estimation. Alignment of variable trajectories and online quality prediction are integrated into MSPM. Predicting the values of the end-of-batch quality during the progress of the batch provided a useful insight to anticipate the effects of excursions from NO on the final quality. Control limits on contribution plots enhance the identification of variables that contribute to the inflation of the SPM statistics and improve the diagnosis capabilities.

Acknowledgment

Financial support provided by NSF Award No. BES-0084749 is gratefully acknowledged.

Literature Cited

(1) Kourti, T.; MacGregor, J. F. Process analysis, monitoring and diagnosis using multivariate projection methods. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 3−21.

(2) Wise, B.; Gallagher, N. The process chemometrics approach to process monitoring and fault detection. *J. Process Control* **1996**, *6*, 329−348.

(3) Wold, S.; Geladi, P.; Esbensen, K.; Ohman, J. Multi-way principal component and PLS analysis. *J. Chemom.* **1987**, *1*, 41−56.

(4) Nomikos, P.; MacGregor, J. F. Monitoring batch processes using multiway principal component analysis. *AIChE J.* **1994**, *40*, 1361−1375.

(5) Nomikos, P.; MacGregor, J. F. Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97−108.

(6) Tracy, N. D.; Young, J. C.; Mason, R. L. Multivariate control charts for individual observations. *J. Qual. Control* **1992**, *24*, 88−95.

(7) Nomikos, P.; MacGregor, J. F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41−59.

(8) Smilde, A. K. Three-way analysis. Problems and prospects. *Chemom. Intell. Lab. Syst.* **1992**, *15*, 143−157.

(9) Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149−171.

(10) Dahl, K. S.; Piovoso, M. J.; Kosanovich, K. A. Translating third-order data analysis methods to chemical batch processes. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 161−180.

(11) Wise, B. M.; Gallagher, N. B.; Butler, S. W.; White, D.; Barna, G. G. A comparison of principal components analysis, multi-way principal components analysis, tri-linear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *J. Chemom.* **1999**, *13*, 379−396.

(12) Boque, R.; Smilde, A. K. Monitoring and diagnosing batch processes with multiway regression models. *AIChE J.* **1999**, *45*, 1504−1520.

(13) Rannar, S.; MacGregor, J. F.; Wold, S. Adaptive batch monitoring using hierarchical PCA. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 73−81.

(14) Chen, J.; Liu, K. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chem. Eng. Sci.* **2002**, *57*, 63−75.

(15) Chen, J.; Liu, J. On-line piecewise monitoring for batch processes. *Proceedings of the International Symposium on Advanced Control of Chemical Processes (ADCHEM 2000)*, Elsevier Science: Pisa, Italy, 2000; pp 635−640.

(16) Zheng, L. L.; McAvoy, T. J.; Huang, Y.; Chen, G. Application of multivariate statistical analysis in batch processes. *Ind. Eng. Chem. Res.* **2001**, *40*, 1641−1649.

(17) Louwerse, D. J.; Smilde, A. K. Multivariate statistical process control of batch processes based on three-way models. *Chem. Eng. Sci.* **2000**, *55*, 1225−1235.

(18) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemom.* **1999**, *13*, 397−413.

(19) van Sprang, E. N. M.; Ramaker, H. J.; Westerhuis, J. A.; Gurden, S. P.; Smilde, A. K. Critical evaluation of approaches for on-line batch process monitoring. *Chem. Eng. Sci.* **2002**, *57*, 3979−3991.

(20) Kourti, T.; Lee, J.; MacGregor, J. F. Experiences with industrial applications of projection methods for multivariate statistical process control. *Comput. Chem. Eng.* **1996**, *20*, S745.

(21) Neogi, D.; Schlags, C. Multivariate statistical analysis of an emulsion batch process. *Ind. Eng. Chem. Res.* **1998**, *37*, 3971−3979.

(22) Ündey, C.; Williams, B. A.; Çınar, A. Monitoring of batch pharmaceutical fermentations: Data synchronization, landmark alignment, and real-time monitoring. Proceedings of the 15th IFAC World Congress on Automatic Control, Barcelona, Spain, 2002.

(23) Jorgensen, P.; Pedersen, J. G.; Jensen, E. P.; Esbensen, K. On-line batch fermentation monitoring—Prediction of biological time. *Preprints of the Scandinavian Symposium on Chemometrics*, Copenhagen, Denmark, 2001; p A29 (Proceedings published as a special issue of *Journal of Chemometrics*; 2002).

(24) Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process.* **1978**, *2*, 43−49.

(25) Rabiner, L. R.; Rosenberg, A. E.; Levinson, S. E. Consideration in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. Acoust., Speech, Signal Process.* **1978**, *6*, 575.

(26) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. Synchronization of batch trajectories using dynamic time warping. *AIChE J.* **1998**, *44*, 864−875.

(27) Gollmer, K.; Posten, C. Supervision of bioprocesses using a dynamic time warping algorithm. *Control Eng. Pract.* **1996**, *4*, 1287−1295.

(28) Henrion, R. *N*-way principal component analysis. Theory, algorithms and applications. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 1−23.

(29) Wold, S.; Kettaneh, N.; Friden, H.; Holmberg, A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331−340.

(30) Guay, M. Personal communication, 2000.

(31) Ramsay, J. O.; Silverman, B. W. *Functional Data Analysis*; Springer-Verlag: New York, 1997.

(32) Williams, B. A.; Ündey, C.; Çınar, A. Detection of process landmarks using registration for on-line monitoring. *Preprints of IFAC DYCOPS6*, Elsevier Science: Cheju Island, Korea, 2001; pp 257−263.

(33) Çınar, A.; Parulekar, S. J.; Ündey, C.; Birol, G. *Batch Fermentation: Modeling, Monitoring and Control*; Marcel Dekker: New York, 2003.

(34) Hahn, G.; Meeker, W. *Statistical Intervals. A Guide to Practitioners*; John Wiley: New York, 1991.

(35) Box, G. Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in one-way classification. *Ann. Math. Stat.* **1954**, *25*, 290−302.

(36) Miller, P.; Swanson, R. E.; Heckler, C. F. Contribution plots: The missing link in multivariate quality control. *Int. J. Appl. Math. Comput. Sci.* **1998**, *8*, 775−792.

(37) Westerhuis, J. A.; Gurden, S. P.; Smilde, A. K. Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 95−114.

(38) Nomikos, P. Detection and diagnosis of abnormal batch operations based on multiway principal components analysis. *ISA Trans.* **1996**, *35*, 259−266.

(39) Birol, G.; Ündey, C.; Çınar, A. A modular simulation package for fed-batch fermentation: Penicillin production. *Comput. Chem. Eng.* **2002**, *26*, 1553−1565.

(40) Cho, H.; Kim, K. A method for predicting future observations in the monitoring of a batch process. *J. Qual. Technol.* **2003**, *57*, 63−75.

Received for review October 21, 2002
Revised manuscript received May 4, 2003
Accepted May 16, 2003

IE0208218