ONLINE PERFORMANCE MONITORING AND QUALITY PREDICTION FOR BATCH PROCESSES

Cenk Ündey and Ali Cinar *

Illinois Institute of Technology
Department of Chemical and Environmental Engineering
10W 33rd Street, Chicago, IL, 60616 USA

Abstract: Two different quality prediction techniques are incorporated with online MSPM through PLS modeling in this study. The first technique is based on unfolding a batch data array by preserving variable direction. An MPLS model between this matrix and vector of elapsed local batch times is developed to reflect the batch progress. More data partitions become available as the batch progresses and these partitions are rearranged into a matrix to develop local MPLS models predicting quality online. The second technique uses hierarchical PLS modeling in an adaptive manner resulting in a model that can be used to predict end-of-batch quality online. Neither technique requires estimation of future portions of variable trajectories and both are suitable for online multivariate statistical process monitoring and fault diagnosis. Case studies from a simulated fed-batch penicillin fermentation illustrate the implementation of the methodology.

Keywords: Online process monitoring, quality prediction, batch processes

1. INTRODUCTION

Online process performance monitoring and product quality prediction in real-time are important in batch and fed-batch process operation. Many high-value specialty chemicals are manufactured using batch processes. Early detection of excursions from normal operation that may 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 actions before the effects of disturbances ruin the batch.

Although online measurements of quality variables are not usually available, multivariate observations such as temperature, agitator power input and flow rates are recorded frequently. The measured variables are autocorrelated in time and highly correlated with each other. 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 a batch run. Multivariate statistical projection methods such as principal component analysis (PCA) and partial least squares (PLS) have been suggested in conjunction with multivariate statistical process monitoring (MSPM) (Kourti and MacGregor, 1995; Wise and Gallagher, 1996). These techniques have become an effective alternative to conventional univariate statistical process control (SPC) and statistical quality control (SQC).

Batch process measurements made on J variables at K time intervals for I batches that resulted with acceptable product quality form a three-way array $\underline{\mathbf{X}}$ of size $(I \times J \times K)$. Product quality measured at the end of batch with M variables form a matrix $\mathbf{Y}$ of size $(I \times M)$. PCA and PLS techniques have been extended to multiway PCA

* Corresponding author (e-mail : *cinar@iit.edu).*

(MPCA) and multiway PLS (MPLS) (Wold *et al.*, 1987) to account for this three-way data array decomposition of batch processes. Applications in batch/semibatch polymerization (Nomikos and MacGregor, 1995*a*) have been reported. MSPM framework including multivariate charts for both end-of-batch and online monitoring have been proposed (Nomikos and MacGregor, 1995*b*).

Online monitoring becomes challenging in batch processes because future portions of process variable trajectories are not available during the progress of the batch. Different approaches have been used to fill unmeasured parts of these trajectories based on some assumptions including the use of missing value prediction capabilities of PCA and PLS. Each assumption introduces some level of arbitrariness and should be chosen appropriately depending on the process and disturbance type. These approaches have been incorporated into MPCA and MPLS (Nomikos and MacGregor, 1995*a*).

Techniques that do not require future value estimation have also been suggested. One alternative is the adaptive hierarchical PCA (AHPCA) that develops recursive local PCA models relating previous observations in an exponentially weighted moving average manner (Rannar *et al.*, 1998). Another technique uses dynamic PCA and PLS for online batch monitoring without future value estimation (Chen and Liu, 2002). Developing local empirical models by partitioning the total run time of a batch with respect to some scheduling points has also been suggested (Louwerse and Smilde, 2000).

A different online MSPM framework can be established by unfolding the three-way data array by preserving variable direction (Henrion, 1994; Wold *et al.*, 1998; Guay, M., 2000). 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. A methodology have been proposed by developing an MPLS model between process variable matrix that is unfolded in the variable direction and local time stamp to use in the alignment of trajectories (Wold *et al.*, 1998).

Enhancements to this new MSPM framework with online quality prediction is reported in this study. Simulated data from fed-batch penicillin fermentation is used to illustrate the methodology. Remedies are proposed to trajectory alignment and modeling problems caused by the discontinuity problem due to batch/fed-batch switching. Quality prediction is also incorporated into this new MSPM framework based on data partitioning with respect to batch progress. Furthermore, an adaptive hierarchical PLS modeling framework is also developed as a benchmark. Prediction results are compared via simulated fed-batch penicillin fermentation case studies.

2. BATCH DATA ALIGNMENT

It is a common situation in batch processes that the total duration of the batch runs and/or the duration of individual phases within a batch run are not the same due to seasonal changes in environmental variables, variations in quality and impurity concentrations of raw materials used in the recipes, etc. 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 dynamics may occur at different times, resulting in unsynchronized batch profiles.

Different techniques have been suggested to overcome unequal and unaligned batch data problem. *Dynamic time warping* (DTW) is one of these techniques which has its origins in speech recognition. It locally translates, compresses, and expands the patterns so that similar features are aligned.Recent applications for data alignment of a batch polymerization (Kassidas *et al.*, 1998) and batch fermentations are reported (Gollmer and Posten, 1996; Undey *et al.*, 2002). Recently, the *curve registration* (CR) technique has been suggested to align batch trajectories with respect to process landmarks (Undey *et al.*, 2002). It is a twofold process of identifying landmarks within a trajectory followed by warping the test trajectories to the reference trajectory containing landmark locations. The *indicator variable* (IV) technique provides a simpler alternative. An IV is selected so that process variables are sampled with respect to this variable instead of time (Kourti *et al.*, 1996; Neogi and Schlags, 1998; Undey *et al.*, 2002). This variable should be chosen such that it shows the maturity of the evolving batch, is smooth, monotonically increasing or decreasing, and should span the operation range for all of the variables. Each new observation is taken relative to the progress of this variable. The data alignment technique used in this study is a variant of IV technique and it is described in Section 3.

3. EMPIRICAL MODELING FOR ONLINE MSPM AND QUALITY PREDICTION

Two different quality prediction techniques are incorporated with online MSPM through PLS modeling in this study. Neither requires estimation of the future portions of the trajectories. Both techniques also offer efficient fault detection and diagnostic capabilities.

Online Quality Prediction via Local MPLS Modeling with Partial Data. An MPLS model developed between the process measurements matrix $\mathbf{X}$ of size $(IK \times J)$ from reference good batches 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 (Fig. 1a). This predicted maturity index can be used to align and equalize reference batches. Variable trajectories can be re-digitized based on for example, percent increments on this variable by using linear interpolation. It is advantageous to use this type of MPLS modeling for online monitoring because it is not necessary to estimate future portions of variable trajectories. However, this technique lacks online prediction of end-of-batch quality in real-time. A two-step integrated modeling approach is proposed to account for online quality prediction. After reference batch data are aligned using the IV technique, batch progress is determined according to percent increments on local batch time (or another IV) so that batches in the reference set are partitioned based on these increments that are chosen arbitrarily such as 10%, 20% of $\mathbf{z}_{\text{pred}}$ (Fig. 1a). Each partition of $\mathbf{X}$ $(IK \times J)$ is rearranged and inserted into matrix $\mathbf{X}$ $(I \times KJ)$ as shown in Fig. 1b. Whenever a partition is rearranged, i.e. some percent of the batch is completed, another MPLS model is developed between this partial data and the final product quality matrix $\mathbf{Y}$. This gives an opportunity to predict end-of-batch quality on percent progress points reflected by partitions. The number of quality predictions will be equal to the number of partitions in this case. To differentiate the two MPLS techniques depending on different types of unfolding, the one that preserves variable direction is called MPLSV (Fig. 1a) and the conventional technique that preserves batch direction is called MPLSB (Fig. 1b).

MPLSV decomposes $\mathbf{X}$ and $\mathbf{z}$ into a combination of scores matrix $\mathbf{T}$ $(IK \times R)$, loadings matrix

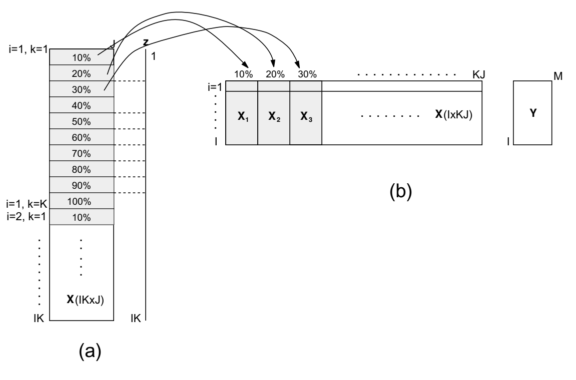

Fig. 1. (a) Partitioning of process measurements space and (b) restructuring for online quality prediction framework.

$\mathbf{P}$ $(J \times R)$ and vector $\mathbf{q}$ $(R \times 1)$ and weight matrix $\mathbf{W}$ $(J \times R)$ with R latent variables (model dimensions)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \qquad \mathbf{z} = \mathbf{T}\mathbf{q} + \mathbf{f} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{f}$ are the residuals matrix and vector, respectively. During the progress of a new batch, a vector $\mathbf{x}_{\text{new}}$ of size $(1 \times J)$ becomes available at each sampling time k. After applying the same scaling to new observations vector as that of reference set, scores can be predicted for time instant k by using the MPLSV model parameters in Eq. 1

$$\hat{\mathbf{t}}_{\text{k}} = \mathbf{x}_{\text{new}}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}, \tag{2}$$

$$\mathbf{e}_{\text{k}} = \mathbf{x}_{\text{new}} - \hat{\mathbf{t}}_{\text{k}}\mathbf{P}^T, \; z_{\text{pred,k}} = \hat{\mathbf{t}}_{\text{k}}\mathbf{q}. \tag{3}$$

$$\text{SPE}_k = \sum_{j=1}^{J} e_{jk}^2 \quad \sim g\chi_h^2 \tag{4}$$

z_{pred} in Eq. 3 can be calculated for each batch in the reference set and control limits (plus or minus three standard deviation of $\mathbf{z}_{\text{pred}}$ for each time interval k) can be constructed to monitor batch maturity. SPE values that are calculated for each time interval k over J variables using the residuals vector in Eq. 4 are well approximated by the chi-squared (χ^2) distribution where g is a constant and h is the effective degrees of freedom. The upper and lower control limits for new independent $\mathbf{t}$-scores under the assumption of normality are defined as

$$\pm t_{n-1,\alpha/2} s_{\text{ref}} (1 + 1/n)^{1/2}, \tag{5}$$

where $t_{n-1,\alpha/2}$ is the critical value of the t-student test with $n-1$ degrees of freedom at significance level $\alpha/2$, n and s_{ref} are the number of observations and the estimated standard deviation, respectively, of the $\mathbf{t}$-score sample at a given time interval k.

T^2 and the corresponding statistical limits are also calculated by using the mean-centered score matrix. T^2 detects small shifts and deviations from normal operation defined by the model. T^2 values for each sampling time k follow an $F_{R,I-R}$ distribution

$$T_k^2 = (\hat{\mathbf{t}}_k - \bar{\mathbf{t}}_k)^T \mathbf{S}_k^{-1} (\hat{\mathbf{t}}_k - \bar{\mathbf{t}}_k) \frac{I(I-R)}{R(I^2-1)} \tag{6}$$

where $\hat{\mathbf{t}}_k$ 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.

Online Quality Prediction with Adaptive Hierarchical PLS Modeling. HPLS is an extension of consensus PCA and is a refinement of

Table 1. Input (1-4), process (5-14) and product variables (y_{1-5}) of simulated fed-batch penicillin fermentation

Variable No.	Definition
1	Aeration rate
2	Agitator power input
3	Substrate feed rate
4	Substrate feed temperature
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
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

the PLS method with multiple **X** and one or more **Y** blocks (Wold *et al.*, 1996). In hierarchical PLS modeling, one PLS model dimension from each **X** block is arranged into a matrix called super score matrix that is then used to develop a super model to predict **Y** block(s). We have modified HPLS modeling for developing an online predictive framework in this study. In this framework, a batch data array of reference batches is divided into K time slices of two-dimensional **X** blocks of size $(I \times J)$ and local PLS models are developed recursively between the super scores of these blocks at each time interval and the product quality matrix. A weighting factor d_k balances the contributions of the new information ($\mathbf{r}_k$) and current history ($\mathbf{r}_{k-1}$) for each model dimension, playing a similar role as exponential weighting factor in an EWMA model). The adaptive HPLS model is used in both online process performance monitoring and fault detection/diagnosis, and in end-of-batch quality prediction during the progress of the batch.

4. ILLUSTRATIVE CASE STUDIES

A data set of 40 batches of penicillin fermentation containing 14 process variables and 5 end-of-batch quality variables was simulated under normal operating conditions (NOC) with small perturbations in input and parameter spaces to mimic real fermentation processes. Variable definitions and numbers are shown in Table 1. Each batch has a different completion time resulting in unequal number of measurements on each variable and there also exist temporal variations in the occurrence of local features. Variable trajectories are equalized and aligned using the IV technique. Because penicillin fermentation has two operational phases (batch and fed-batch, respectively), there is no IV that spans the entire batch operation. MPLS modeling for entire batch evolution generated poor results on local batch time prediction due to this operational discontinuity. We overcome these problems by dividing process data into two parts phase-wise. Batch alignment is performed based on the predicted local batch time stamps in each phase. We have arbitrarily chosen each 50% and 20% batch completion points calculated by MPLS model in phase 1 and phase 2, respectively, to rearrange into new matrix partitions that are used to develop predictive models with quality matrix **Y**. This gives an opportunity to predict end-of-batch quality on percent progress points reflected by these partitions. The number of quality predictions will be equal to the number of partitions that is 7 in this case.

The same aligned reference batch data set is used in testing both online quality prediction schemes. The adaptive HPLS model is developed with an arbitrary choice of weight coefficient $d = 0.38$. Two cases are considered for illustration. A normal batch is investigated first. As expected, the SPE plot produced no out-of-control signal and final product quality on all five variables (shown as a solid star) is successfully predicted by both models (Figs. 2 and 3). The prediction capability of the partial MPLS model 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 -0.067% h^{-1} is introduced into substrate feed rate at the beginning of the fed-batch phase until the end of operation, SPE plot signaled the out-of-control status right after the third quality prediction point at 179 h for MPLS model. Because models are not valid beyond out-of-control points no further confidence limits are plotted on y_{1-5} in Fig. 4. Although model predictions might not be accurate beyond out-of-control points, MPLS-based framework generated fairly close predictions of the inferior quality. In another fault scenario, a small downward drift of magnitude -0.01% h^{-1} into substrate feed rate is introduced and it is then corrected. Monitoring and prediction results are presented in Fig. 5. Fault is detected successfully in this case. Since it is corrected its effect on product quality is minimal hence giving close estimates for end-of-batch quality values.

T^2 and mean-centered latent variable (LV) plots are also used to investigate the disturbance (Fig. 6). Because the fault is introduced in fed-batch phase, charts are plotted only for this phase against the IV used in the alignment. Detection of out-of-control times are depicted with a vertical dotted line. Results of fault detection and diagnosis performances are summarized in Table 2. LVs have signaled out-of-control situation in different times. Although SPE has detected out-of-control

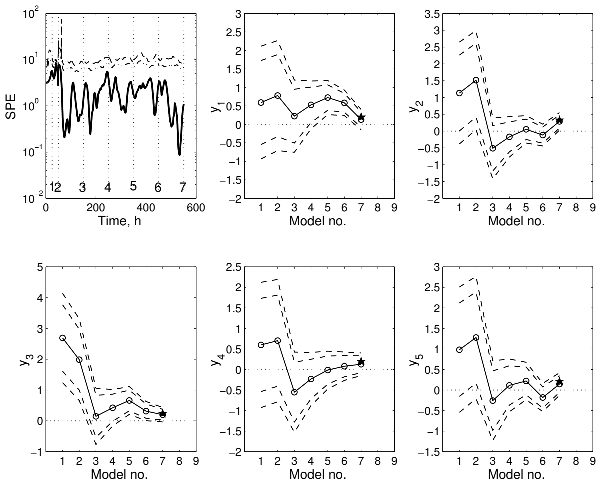

Fig. 2. SPM and quality estimation of an NOC batch with partial MPLS modeling.

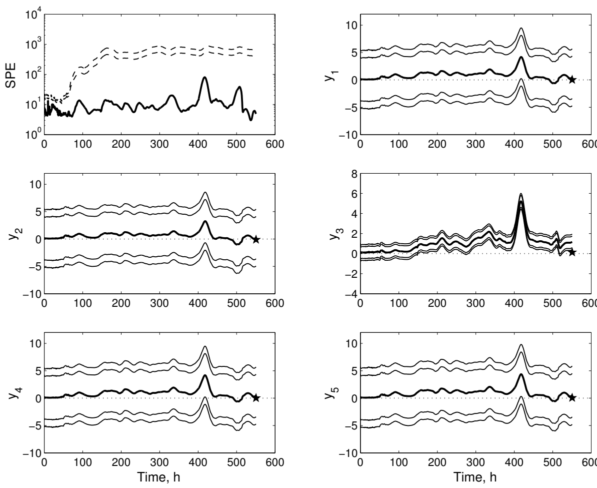

Fig. 3. SPM and quality estimation of an NOC batch with adaptive HPLS modeling.

status the earliest, averaged variable contributions to SPE right after a few observations from the out-of-control time diagnose only the root cause (variable 3) as it exceeds its contribution limits (dashed lines) (bottom-left portion of Fig. 7). Investigating T^2 and LV plots provide further interpretation on the variables that are affected by the deviation of the disturbed variable. Averaged variable contributions in Fig. 7 along with their control limits (dashed line) between the first out-of-control signal point following six consecutive observations indicated that different variables are diagnosed as affected by the root cause. Had we used contribution plots for SPE and T^2 alone, we would have diagnosed only the root cause (variable 3) and, 6 and 9 as affected variables. Based on our process knowledge, we infer that these variables are affected by the deviation in substrate feed rate (variable 3). Variable contributions to LVs at different out-of-control points indicates the entire spectrum of the variables that are affected by the disturbance.

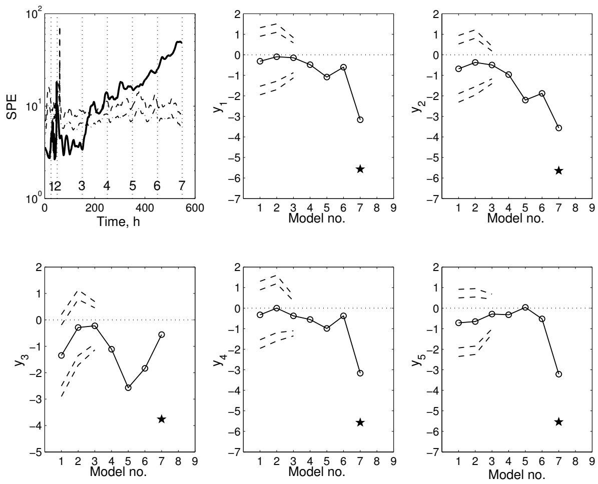

Fig. 4. SPM and quality estimation of a faulty batch with partial MPLS modeling.

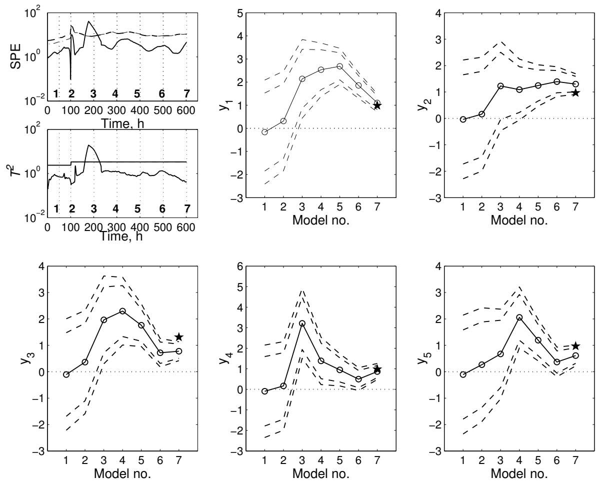

Fig. 5. SPM and quality estimation of a corrected faulty batch with MPLS modeling.

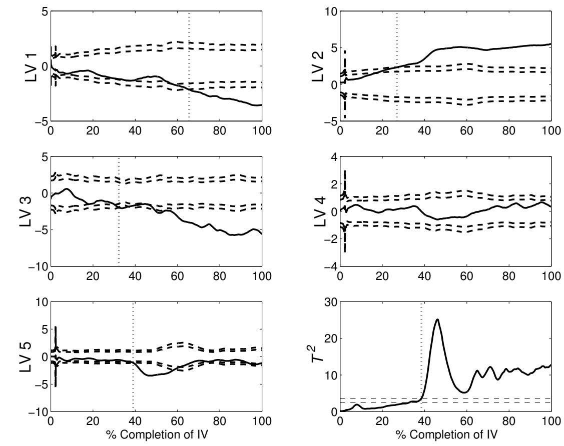

Fig. 6. Latent variable (LV) plots and T^2 chart of the faulty batch based on MPLS technique.

5. CONCLUSION

The integration of online process monitoring techniques with quality prediction has been investigated. MPLS using a batch data matrix unfolded by preserving variable direction and adaptive hi-

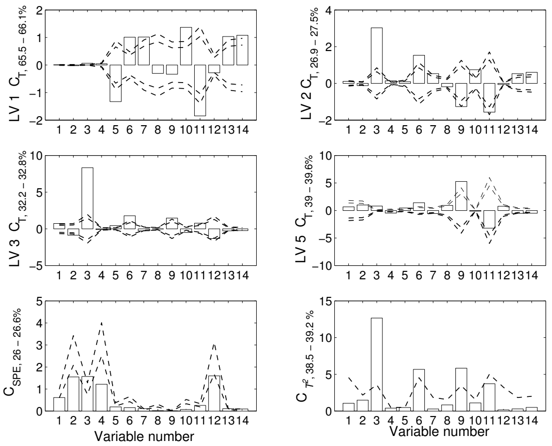

Fig. 7. Variable contributions to latent variable plots, SPE and T^2 based on MPLS technique.

Table 2. Fault detection and diagnosis results

Statistic	Detection	Diagnosis
SPE	26 %	3
LV 2	26.9 %	3,6,7,10,14
LV 3	32.2 %	3,5,6,10
T^2	38.5 %	3,6,9
LV 5	39 %	3,5,9
LV 1	65.5 %	3,5,6,7,10,11,13,14

erarchical PLS are used to develop online process monitoring, and fault detection/diagnosis tools as well as predicting end-of-batch quality. These two techniques were applied to monitoring and quality prediction in a simulated fed-batch penicillin fermentation. Both techniques were useful in providing information for an early assessment of final product quality, indicating trends that may cause inferior quality and effective detection and diagnosis of process disturbances. An advantage of adaptive HPLS is its capability to estimate final product quality at each time instant during batch progress. MPLS with partial modeling provides quality estimates on the scheduled points. To obtain estimates more frequently the number of local MPLS models can be increased. Variable contributions to latent variables are also investigated for further diagnosis of trends. These contributions provided an opportunity for thorough trend analysis.

6. REFERENCES

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

Gollmer, K. and C. Posten (1996). Supervision of bioprocesses using a dynamic time warping algorithm. *Control Eng. Practice* **4**(9), 1287–1295.

Guay, M. (2000). Personal communication.

Henrion, R. (1994). N-way principal component analysis. Theory, algorithms and applications. *Chemo. Intell. Lab. Sys.* **25**, 1–23.

Kassidas, A., J.F. MacGregor and P.A. Taylor (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE J.* **44**(4), 864–875.

Kourti, T. and J.F. MacGregor (1995). Process analysis, monitoring and diagnosis using multivariate projection methods. *Chem. Intell. Lab. Sys.* **28**, 3–21.

Kourti, T., J. Lee and J.F. MacGregor (1996). Experiences with industrial applications of projection methods for multivariate statistical process control. *Comput. Chem. Eng.* **20**(Suppl. A), 745.

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

Neogi, D. and C. Schlags (1998). Multivariate statistical analysis of an emulsion batch process. *Ind. Eng. Chem. Res.* **37**(10), 3971–3979.

Nomikos, P. and J.F. MacGregor (1995*a*). Multiway partial least squares in monitoring batch processes. *Chemo. Intell. Lab. Sys.* **30**, 97–108.

Nomikos, P. and J.F. MacGregor (1995*b*). Multivariate SPC charts for monitoring batch processes. *Technometrics* **37**, 41–59.

Rannar, S., J.F. MacGregor and S. Wold (1998). Adaptive batch monitoring using hierarchical PCA. *Chemo. Intell. Lab. Sys.* **41**, 73–81.

Undey, C., B.A. Williams and A. Cinar (2002). Monitoring of batch pharmaceutical fermentations: Data synchronization, landmark alignment, and real-time monitoring. In: *Proc. 15th IFAC World Congress on Automatic Control.* Barcelona, Spain.

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

Wold, S., N. Kettaneh, H. Friden and A. Holmberg (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemo. Intell. Lab. Sys.* **44**, 331–340.

Wold, S., N. Kettaneh-Wold and K. Tjessem (1996). Hierarchical multiblock PLS and PC models, for easier model interpretation, and as an alternative to variable selection. *J. Chemometrics* **10**, 463–482.

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