ADCHEM 2006

International Symposium on Advanced Control of Chemical Processes
Gramado, Brazil – April 2-5, 2006

IFAC

AN ADJOINED MULTI-DPCA APPROACH FOR ONLINE MONITORING OF FED-BATCH PROCESSES

Ng Yew Seng and Rajagopalan Srinivasan*

Laboratory for Intelligent Applications in Chemical Engineering,
Department of Chemical and Biomolecular Engineering,
National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
**Email: chergs@nus.edu.sg; Tel: 65-68748041; Fax: 67791936*

Abstract: Batch processes are common in the manufacturing of high value-added products. Monitoring with the highly popular principal components analysis (PCA) approaches do not function adequately in the face of the sequential nature of batch processes, as the basic assumptions that its monitoring statistics (SPE and Hotelling's T^2) are developed upon – stationary, normal distribution of source data - are violated. Consequently, these monitoring techniques become prone to Type-I (false positives) and Type-II Errors (false negatives). In this article, an extension of PCA, called adjoined dynamic principal component analysis (ADPCA), is proposed for online monitoring of batch processes by using multiple dynamic-PCA (DPCA) models. The ADPCA models are developed by first clustering process data using fuzzy c-means algorithm and developing a DPCA model for each cluster. Each cluster is selected so that it satisfies the PCA's assumption. The problem of switching between the models which normally confounds multiple model-based approaches is overcome by allowing adjoining models to overlap and thus enabling smooth switching from one model to another during the course of batch operations. As shown in this paper, the proposed methodology reduces both Type-I and Type-II errors compared to single block methods.

Keywords: batch mode, batch control, fuzzy models, fault detection, non-stationary systems, non-linear systems.

1. INTRODUCTION

Batch/fed-batch operations are commonly used to manufacture high value-added products in the chemical, pharmaceutical, biological, and semi-conductor industries. The standard operating procedures (SOP) for fed-batch operations often occur in sequences specified in a recipe. Unlike continuous processes, batch processes have various characteristics that complicate process monitoring; these include finite time duration, nonlinearities, non-stationary (non-steady-state) behavior and run-to-run variations. Regulatory and supervisory control for batch operations is an open research area as most high-level automation applications are effective only during steady-states. Owing to the lack of effective automation and the high cognitive workload of plant operators, the occurrence of human errors during these operations is very likely (Ng and Srinivasan, 2004). A general feature of batch/fed-batch processes is that small changes in the operating conditions during critical periods may degrade the quality of the final product; this is especially obvious in biological processes. Due to the numerous complexities in these mode of operations, effective techniques for online monitoring is essential since timely corrective action may prevent fault propagation and allow a batch to be saved.

In this paper, a new monitoring technique, called adjoined dynamic principal component analysis (ADPCA), is proposed for online monitoring of fed-batch operations. ADPCA uses overlapping PCA models for monitoring batch trajectories and is inherently capable of modeling non-stationary processes more accurately. It is capable of overcoming both Type-I (false positives) and Type-II errors (false negatives) suffered by conventional single-block PCA techniques during batch operations. The organization of this article is as follows: Section 2 reviews PCA and its variants, and their shortcomings during online monitoring of batch operations. The proposed ADPCA methodology for online batch process monitoring is described in Section 3 while Section 4 presents the applications of the proposed method to a fed-batch penicillin cultivation process. Furthermore, a comparison of the proposed methodology with multiway-PCA and dynamic-PCA is also presented.

2. PCA-BASED PROCESS MONITORING : METHODS AND SHORTCOMINGS

Principal component analysis is a popular statistical technique for process monitoring (Kourti, 2002). Mathematically, PCA relies upon eigenvector decomposition of the covariance or correlation matrix to capture the major tendencies of process variables. Let $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3, ..., \boldsymbol{x}^m\}^T$ with m rows and n columns. PCA linearly decomposes the data matrix $\boldsymbol{X}$ as the sum of scores, $\boldsymbol{t}$, loadings, $\boldsymbol{p}$, and a residual matrix $\boldsymbol{e}$ in the following way (Wise & Gallagher, 1996):

$$\boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{t}_2\boldsymbol{p}_2^T + ... + \boldsymbol{t}_i\boldsymbol{p}_i^T + ... + \boldsymbol{t}_k\boldsymbol{p}_k^T + \boldsymbol{e}\,. \quad (1)$$

Here, k is the number of principal components that a user wants to retain. The scores vector, $\boldsymbol{t}$ contains information on how the samples relate to each other, while the loadings vector, $\boldsymbol{p}$ contains information regarding the correlation among the variables.

2.1 Fault detection with PCA approaches

Fault detection using PCA or its variants is usually performed through monitoring of the squared prediction error (SPE) and/or Hotelling's T^2 statistic. The SPE measures the variation of a sample $\boldsymbol{x}_i$ from the PCA model, i.e. lack-of-fit:

$$SPE_i = \boldsymbol{e}_i \boldsymbol{e}_i^T = \boldsymbol{x}_i (\boldsymbol{I} - \boldsymbol{p}_k \boldsymbol{p}_k^T) \boldsymbol{x}_i^T . \qquad (2)$$

The process is considered normal if $SPE_i < Q_{1-\alpha}$, where $Q_{1-\alpha}$ denotes the upper control limit for confidence level at $1-\alpha$ percentile (Jackson and Mudholkar, 1979). The T^2 statistic measures the variation of the sample within the PCA model:

$$T_i^2 = \boldsymbol{t}_i \boldsymbol{\lambda}^{-1} \boldsymbol{t}_i^T = \boldsymbol{x}_i \boldsymbol{p} \boldsymbol{\lambda}^{-1} \boldsymbol{p}^T \boldsymbol{x}_i^T \qquad (3)$$

where λ^{-1} is the diagonal matrix containing the inverse of the eigenvalues associated with k eigenvectors retained in the PCA model (Wise *et. al.,* 1990). An upper control limit $T^2{}_{k,m,\alpha}$ similar to $Q_{1-\alpha}$ can also be derived for the T^2 statistic (Jackson, 1991) by using a F-distribution of the training data. The SPE and T^2 monitoring statistics are complementary in nature, since the SPE measures the lack-of-fit while the T^2 statistic measures the variation of a sample within the model.

2.2 Type-I and Type-II Errors

Most of the statistical process control (SPC) literature has focused heavily on methods for handling data generated from normal distributions. The PCA-based techniques also assume that the training data follows a normal distribution. This assumption is not valid as for batch operations when analyzed *online*. There are two implications when the normal distribution assumption is used while monitoring batch processes *online*. First, the monitoring limits constructed using SPE and T^2 are prone to Type-II errors (false negatives), as the limits for monitoring statistics cover an unknown (possibly abnormal) operating region when a single model is used for online monitoring. Second, and perhaps the more common scenario, is the occurrence of Type-I errors (false positives). Even for a *perfectly normal* (data that are exactly normal distributed) process, the occurrence of Type-I errors are generally close to α (~50 false positives for 95% confidence limits with every 1000 samples analyzed) (Nomikos and MacGregor, 1994; Martin and Morris, 1996). Failure to account for the data distribution of the training data further increases the rate of false positives. The total rate of Type-I errors for a *non-normal* process, is thus equal to the sum of errors induced by the selected α, and the errors induced by the data distribution modeling process. As a result, the reliability of the supervision system is greatly reduced.

2.3 Multiway and Dynamic-PCA

PCA has been widely used for monitoring continuous operations (Kourti, 2002). However, there exist some limitations of the PCA approaches when they are used for monitoring batch processes (Nomikos and MacGregor, 1995). In practice, batch data are usually stored in a three dimensional data matrix. An extension of PCA called Multiway-PCA (Nomikos and MacGregor, 1995) was proposed for batch data analysis. MPCA organizes the batch data into time-ordered blocks by unfolding the three dimensional array into a large two dimensional matrix before they are decomposed into their corresponding principal components. In general, there exist three different ways (batch-wise unfolding; time-wise unfolding, and variable-wise unfolding) that a 3-Dimensional array $\boldsymbol{X}$ can be unfolded (Lee *et. al.,* 2004). The MPCA technique referred in this work is based on the time-wise unfolding. Such unfolding allows abnormal samples to be identified from a given batch trajectory.

PCA generates a linear static model of the data matrix $\boldsymbol{X}$. When the data contains dynamic information, as in the case with fed-batch operations, applying PCA/MPCA on the data does not capture the actual correlations between the variables, but only a linear static approximation. For most of the processes encountered, a dynamic-PCA is more appropriate (Ku *et. al.*,1995). For non-stationary systems, the current values of process variables will depend on the past values due to time-lag behavior of the chemical processes. $\boldsymbol{X(t)}$ can thus be augmented with previous observations, where

$$\boldsymbol{X}_D(t) = [\boldsymbol{X(t)} \;\; \boldsymbol{X}(t-1) \;\; \ldots \;\; \boldsymbol{X}(t-l)], \qquad (4)$$

with l being the number of previous observations that are correlated to the current sample. The extracted dynamic model is implicitly multivariate autoregressive (AR) (Ljung and Glad, 1994) if process inputs are included (Ku et. al., 1995).

Previous literature on online monitoring based on MPCA/DPCA for transient operations normally require the batch-length to be equal in order to manipulate the monitoring limits at each time instant to reduce the occurrence of false positives, eg: Birol *et.al.* (2002), Lee *et. al*.(2004), Nomikos and MacGregor (1995), Rännar *et. al.* (1998), *etc*. Monitoring limits which are generated based on fixed time approach might give reasonable performance if the underlying events are highly synchronized, eg: activation of fed-batch at fixed time-point, all growth of microbiology to be consistent throughout, fixed duration of batch operation. However, run-to-run variations are often significant due to environmental or human factors. Such drastic behavior cannot be adequately modeled by MPCA and DPCA techniques. To overcome these shortcomings, multiple models are needed for a more flexible monitoring of batch processes.

2.4 Need for multiple adjoined models

The use of multiple models has been a popular approach in system identification (Böling *et. al.*, 2004), advanced control (Palma and Magni, 2004), and monitoring (Bhagwat et.al., 2003). In this work, we use a *divide-and-conquer* strategy to overcome the shortcomings as outlined in the previous paragraph by using multiple PCA models for online monitoring. Though there exist several literatures (Hwang and Han,1999; Rännar *et. al.,* 1998) on performing batch monitoring with multiple models, the existing PCA approaches are limited by run-length, discontinuous in modeling, and prone to false positives. This paper extends these methods by using overlapping models in order to be *robust to run length*, give *good monitoring resolution*, and *be sensitive to new operating region* (*i.e.* detect novel fault). When multiple models are used to model transient operation, false positives are often encountered in the *border* between two or more models. In Srinivasan *et. al*., (2005), it was reported that 50% of prediction errors occurred are attributed to state change when multiple sets of neural networks are used for temporal pattern recognition. The high level of errors during model switching, similarly in multiple PCA models approaches, is due to discontinuity in modeling transient operations. Fed-batch operations usually follow a trajectory on the PC subspace. If disjoint PCA models are used, the models constructed failed to incorporate all relevant normal operating region and thus result in oversensitivity in the bordering region of each model when future samples evolve through such sensitive region. These shortcomings can be overcome when the neighboring PCA models are allowed to overlap, since overlapping PCA models allow the modeling of the smooth evolution of the process trajectory in addition to any abrupt changes that occurred. As a result, the monitoring statistics constructed from each individual model cover all relevant normal operating regions and prevent false positives from occurring during model-switching / state change.

3. ADJOINED DPCA METHOD FOR ONLINE MONITORING OF FED-BATCH OPERATIONS

In this section, we propose an adjoined multi-model DPCA-based methodology for o*nline* monitoring and supervision of fed-batch processes. We term the proposed method as adjoined-DPCA (ADPCA) as the method is developed on the basis of multiple overlapping and connected DPCA models. The proposed methodology is based on the integration of fuzzy clustering methodology with dynamic-PCA monitoring approaches. As described in the earlier section, conventional single block MPCA or DPCA monitoring approaches fail to account for the non-stationary effects in temporal signals. Fuzzy clustering of process states based on historical data can thus be used to differentiate multiple modes of operations in these temporal signals for building different DPCA models for monitoring purposes. The additional membership information obtained through fuzzy clustering provides a means to construct overlapping DPCA models. Such membership information is often not available in crisp clustering algorithm such as *k*-means. The training data from different stages/phases can be extracted and reconstructed based on a proposed *fuzzy-data reconstruction approach*. The reconstructed data groups overlap with their neighboring groups, with a DPCA model constructed for each group reconstructed. At every instant, the best-fit DPCA model is selected for monitoring purposes. The offline ADPCA model construction is described next in Section 3.1 and the steps for online monitoring are presented in Section 3.2.

3.1 ADPCA model construction

The training algorithm is based on normal operating data only, which can be obtained from the plant historian directly. Process data are often corrupted with noise. In this work, a windowed finite impulse response (FIR) filter (IEEE, 1979) is implemented to eliminate high frequency noise. Let $\boldsymbol{Y}$ be the raw data collected from a plant historian. Variable unfolding is first carried out on the 3-dimensional dataset to reduce $\boldsymbol{Y}$ to a 2-dimensional dataset for analysis. Each variable of the training data, $\boldsymbol{y}_i$, is later normalized to eliminate the varying scales of the variables:

$$\boldsymbol{x}_i = \frac{\boldsymbol{y}_i - \boldsymbol{y}^{mean}}{\sigma_{\boldsymbol{y}}}. \quad (5)$$

The filtered signals, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_i, \ldots \boldsymbol{x}_m\}$, is then partitioned into different clusters through fuzzy *c*-means clustering. Fuzzy *c*-means is a generalization of the *k*-means data clustering technique where each sample belongs to one or more clusters as per a membership grade. For a dataset consisting of *m* objects and a pre-specified *c* number of clusters, the fuzzy c-means algorithm computes the optimal memberships by minimizing (Hataway and Bezdeck, 1988):

$$obj = \sum_{j=1}^{c}\sum_{i=1}^{m} \boldsymbol{u}_{ij}^{v} d(\boldsymbol{x}_i, \boldsymbol{m}_j)^2 \quad (6)$$

where $\boldsymbol{m}_j$ is the centre of cluster c, $d(\boldsymbol{x}_i, \boldsymbol{m}_j)$ is the Euclidean distance between the data point and the cluster center, $\boldsymbol{u}_{ij} \geq 0$ for all *m*, and $\sum_{j=1}^{c} \boldsymbol{u}_{ij} = 1$, for each i^{th} sample. Here, *v* is called the fuzzifier and affects the final membership distribution. $v = 1$ leads to crisp clustering solution; a value of 2 is normally used. If $\boldsymbol{u}_{ic}$ is restricted to 0 or 1, the proposed algorithm reduces to the *k*-means algorithm. The fuzzy membership grade for any given sample indicates whether there is any other cluster that is comparable to the best cluster.

The samples are then stacked into different groups based on time-wise unfolding. Consider the training data $\boldsymbol{X}$ obtained together with its clustering information, $\boldsymbol{u}_{ij}$. $\boldsymbol{X}$ is now a two dimensional array obtained by stacking the training data of different runs through multi-way approaches, it is reorganized into *c* number of groups based on a *fuzzy-data reconstruction approach*. The data reconstruction process serves as a data preparation mode for constructing adjoined-DPCA models. At any instant, a sample (measurements) has a one-to-many relation with the data groups, which means a sample can be concurrently present in one or more data groups. Since

$$\boldsymbol{u}_{ij} = \frac{1}{\sum_{r=1}^{c}\left(\frac{d(\boldsymbol{x}_i, \boldsymbol{m}_j)}{d(\boldsymbol{x}_i, \boldsymbol{m}_r)}\right)^{2/v-1}}, \quad (7)$$

where $d(\boldsymbol{x}_i, \boldsymbol{m}_j)$ is the distance from point $\boldsymbol{x}_i$ to current cluster center $\boldsymbol{m}_j$, and $d(\boldsymbol{x}_i, \boldsymbol{m}_r)$ is the distance from $\boldsymbol{x}_i$ to other cluster center $\boldsymbol{m}_r$, the highest value of $\boldsymbol{u}_{ij}$ gives the cluster which is closest to $\boldsymbol{x}_i$. The best membership of $\boldsymbol{X}$ at every i^{th} instance, $\boldsymbol{x}_i$ is thus given by:

$$\boldsymbol{b}_i^{1st} = \arg\max(\boldsymbol{u}_{ij}), \ \forall j = \{1, 2, \ldots, c\}, \quad (8)$$

since $\boldsymbol{u}_{ij}$ is inversely proportional to the distance of $\boldsymbol{x}_i$ from j^{th} cluster centroid, $\boldsymbol{m}_j$, as shown in Equation 7. The subsequent layer of cluster membership function

$\boldsymbol{u}_{ij}$ is then analyzed iteratively following the general algorithm for *fuzzy-data reconstruction* to identify the existence of r^{th} best cluster:

$$\boldsymbol{b}_i^{rth} = \arg\max(\boldsymbol{u}_{ij}),\ \forall j = \{1,2,...,c\}, \qquad (9)$$

subject to the following constraints:

$$\boldsymbol{b}_i^{rth} \neq \{\boldsymbol{b}_i^{(r-1)th}, \boldsymbol{b}_i^{(r-2)th}, ..., \boldsymbol{b}_i^{1st}\}\ \& \qquad (10)$$
$$exist(\boldsymbol{b}_i^{(r-1)th}, \boldsymbol{b}_i^{(r-2)th}, ..., \boldsymbol{b}_i^{1st})\ \&$$
$$\boldsymbol{b}_i^{rth} - \boldsymbol{b}_i^{(r-1)th} < \delta .$$

The *association threshold*, δ defines the allowable regions to overlap when computing DPCA models from the reconstructed data. Here, δ is in the range of $[1\ 0]$. A large value of δ allows fully overlapped regions while setting $\delta = 0$ prohibits the spawning of $\boldsymbol{x}_i$ totally. Each sample $\boldsymbol{x}_i$, is then duplicated and placed within their corresponding groups of *reconstructed data*, $\boldsymbol{W}_j$, where $j \in \{\boldsymbol{b}\}$ identified from the algorithm for fuzzy-data reconstruction. Upon completion of the analysis of one sample, $\boldsymbol{x}_i$, i is incremented and the subsequent sample $\boldsymbol{x}_{i+1}$ is subjected to the same analysis as described in equation 10. The end results of the above algorithm is the creation of reconstructed cluster groups $\boldsymbol{W}_j$, $\forall j = \{1,2,...,c\}$ that overlaps with their proximity clusters. Such new clusters formation is deemed important especially when generating PCA models for monitoring fed-batch processes to avoid discontinuities in modeling the normal operating region. Failure to account for such discontinuities would make the monitoring system prone to Type-I errors when new process trajectory evolve through sensitive regions.

Subsequently, a total of *c* distinct DPCA models DPM_j, $j = \{1,2,...,c\}$ are constructed for process monitoring following (Equation 1 & Equation 4). Each DPM_j corresponds to a local DPCA model generated from the reconstructed dataset $\boldsymbol{W}$. At each instant of monitoring, the DPCA model that best describe the fed-batch processes, DPM_{opt} is selected and used for process monitoring. Multivariate statistical process control charts are based on the SPE and T^2 statistics of DPM_{opt}. With the proposed approach, the effect of false positives will become less severe as the data density for each of the PCA model is locally normal distributed. Additionally, the chances of having false negatives are also greatly reduced, as monitoring limits generated from the PC models do not cover the training subspace of unknown operating region compared to single block approaches.

3.2 Online fault detection

The algorithm for online fault detection is shown in Figure 1. The main challenge in the online monitoring using the proposed approach is to identify the best DPCA models at each point in time. New process measurements, designated here as $\boldsymbol{Y}$, is first filtered and autoscaled to $\boldsymbol{X}$ (Equation 5), before they are projected to the principal component subspace as scores, **t** and loadings, **p** (Equation 1). Its distance in the principal component subspace are then evaluated against all the DPCA models, DPM_j, $\forall j = \{1,2,...,c\}$. The distance between the sample $\boldsymbol{x}_i$ and each DPCA model, DPM_j is calculated based on a *adjoined discriminant similarity factor,* $\boldsymbol{\Psi}$*'*. The original proposal of $\boldsymbol{\Psi}$ (Raich and Çinar, 1996) is solely to facilitate human operators during process monitoring by combining the two separate monitoring statistics (SPE and T^2) into one uniformed, simple index. Here, we extend the discriminant similarity factor to the selection of best DPCA model for online applications. At every instance, the distance of $\boldsymbol{x}_i$ and all constructed DPCA models, DPM_j, $\forall j = \{1,2,...,c\}$ are evaluated through:

$$\boldsymbol{\Psi'}(DPM_j, \boldsymbol{x}_i) = \sqrt{\beta({SPE_{r,i}}^2) + (1-\beta)({T_{r,i}}^2)},$$
$$\forall j = \{1,2,...,c\}. \qquad (11)$$

Here, $SPE_{r,i} = SPE_i / Q_{1-\alpha}$, ${T_{r,i}}^2 = {T_i}^2 / T^2_{k,m-k,\alpha}$, and $1-\alpha$ is the confidence level for limit evaluation. Here, β is a weighting factor between zero and one. Upon the absence of additional information, SPE and T^2 are weighted equally, where β is set to 0.5. The nearest PCA model to $\boldsymbol{x}_i$,

$$DPM_{opt} = \arg\min(\boldsymbol{\Psi'}(DPM_j, \boldsymbol{x}_i)),$$
$$\forall j = \{1,2,...,c\}, \qquad (12)$$

is then selected for monitoring the current stage of the fed-batch operations. Monitoring statistics, *i.e.*: SPE and T^2, are generated based on DPM_{opt} and compared with their upper control limit, $Q_{1-\alpha}$ and $T^2_{k,m-k,\alpha}$ for anomaly detection.

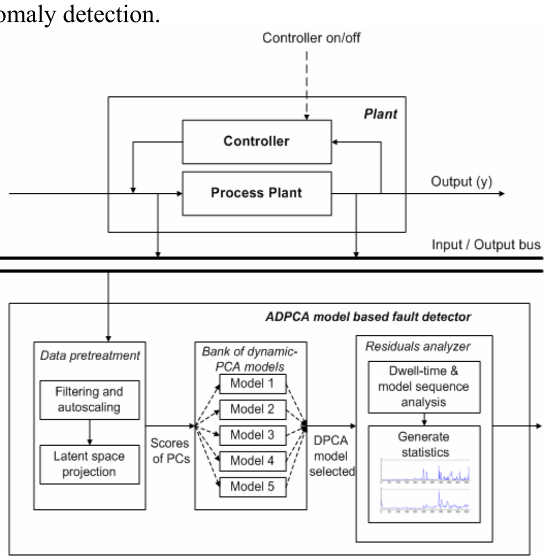

Figure 1: Online architecture for adjoined-DPCA fault detection

4. MONITORING OF FED-BATCH PENICILLIN CULTIVATION PROCESS

In this section, the proposed adjoined-DPCA (ADPCA) method is tested on a fed-batch penicillin cultivation process. In this work, we use the penicillin fed-batch simulator PenSim v2.0 (Birol *et. al*., 2002a) that is based on the mathematical model of Bajpai and Reuss (1980). The simulator captures the dynamics in sixteen process variables, namely: flow rates of input streams, temperature, pH, heat generated, aeration rate of fermenter, etc. The final quality and quantity of the final product is very much affected by pH and temperature. They are therefore controlled at specified setpoint. In our simulation, the pH was controlled at 5.0 and the temperature was maintained at 25°C to promote cell growth. In all runs, an initial batch culture is followed by a fed-batch operation based on the depletion of the carbon source (glucose). The process switches to the fed-batch mode when the level of glucose concentration reaches 0.3g/l. Detailed description of the fed-batch simulator including the state equations and simulation parameters is given by Birol *et. al.* (2002). In this work, a total of 50 normal batches were simulated to create the reference datasets, we have used an integration step size of 0.02h and a sampling interval of 0.5h.

The variables selected for the monitoring of penicillin cultivation process are similar to the study of (Lee *et al*., 2004). An association threshold, δ of 0.1 was chosen for constructing the adjoined-DPCA models. 10 adjoined-DPCA models were built based on the methods proposed in Section 3.1 by setting the time-lag parameters, l=0. By ignoring the time-lag behavior of the ADPCA, the method is actually reduced to adjoined-PCA approaches. The proposed technique is not prone to false positives as each of the PCA model constructed can model the different phases of the fed-batch process more accurately. Occasionally, more than one adjoined PCA models were used to monitor a single phase of the fermentation. The oscillatory nature of some process variables due to process controllers (pH and Temperature) caused some models to be used quite frequently throughout the cultivation period, eg: model 1 and model 5. On the other hand, the use of multiway-PCA and dynamic-PCA techniques, which are based on a single model, is prone to false positives at t~36.5. The results observed are consistent with the analysis of Lee *et. al.* (2004). Seven process disturbances have been tested as summarized in Table 1.

Table 1: Summary of fault scenarios considered

Case	Fault type	Occurrence time (h)
1.	pH controller failure	0.5
2.	Temperature controller failure	0.5
3.	-15% step in aeration rate	60
4.	-15% step in agitation power	30
5.	Ramp increase in aeration rate	70
6.	Ramp increase in agitation power	40
7.	Ramp increase in substrate feed rate	30

4.1 Monitoring of high agitation power

PEN06 corresponds to a ramp increase in agitation input power, P_w with a slope of +0.05. The increment in P_w results in positive deviation from nominal mass transfer rate and causes an over supply of oxygen to the biomass. Multiway-PCA technique detects the anomaly at t=252.0h (Figure 2) when the SPE exceeds the 99% confidence limit. Dynamic-PCA technique gives slightly better monitoring results by being able to detect the fault at t=241.0h when unusual variation are observed through T^2 statistic. The best result is observed from the proposed ADPCA technique (Figure 3) which detects the fault at t=233.0h, which is 19 hours (38 samples) earlier than multiway-PCA and 8 hours (16 samples) earlier than dynamic-PCA technique. The method is also not prone to false positives in comparison to multiway-PCA (11 false alarms) and dynamic-PCA techniques (19 false alarms). In general, dynamic-PCA is more sensitive in detecting process drift/ramp errors in comparison to multiway-PCA, as the method uses time-lag information from previous samples. However, the improvement in sensitivity of dynamic-PCA has been at the cost of having more false alarms since any misclassified samples are included as time-lag information in the future samples. These effects fade off with time. In contrast, with the proposed ADPCA technique, the improved fault detection sensitivity is not correlated with any increase in Type-I errors

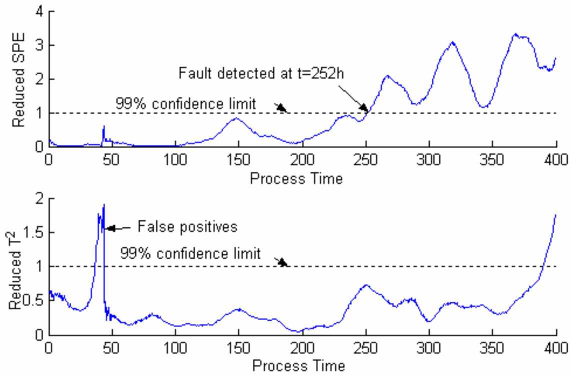

Figure 2: Monitoring results of PEN06 through MPCA

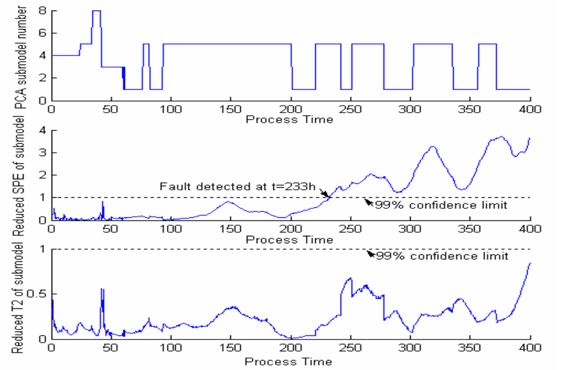

Figure 3: Monitoring results of PEN06 through ADPCA

4.2 Summary of monitoring

The summary of the monitoring results for all disturbances is presented in Table 2. All disturbances are successfully detected by all the methods utilized. Multiway-PCA and dynamic-PCA approaches are very prone to false positives. Multiway-PCA gives a total of 36 false alarms throughout the monitoring of 5600 samples (7 batches), and dynamic-PCA 75 false alarms. Dynamic-PCA gives better performance compared to multiway-PCA in terms of speed of detection as in cases PEN01 (-0.5h), PEN05(-4.0h), PEN06(-11.0h) and PEN07(-52.5h). As a whole, the proposed ADPCA technique gives the best monitoring resolution by being (1) able to detect all the disturbances in short duration (PEN01, PEN02, PEN05, PEN06), and (2) not prone to false positives. No false alarms is observed in all the fault cases studied. The ADPCA technique also offers the convenience of using lesser number of PCs to model batch operations. In this case study, 4 out of the 10 PCA models constructed used only 2 PCs to model a local phase of the fed-batch process (retaining > 95% variance of each of the local phase), with the rest ranging from 1

PC (min) to 5 PCs (max). On the other hand, the use of multiway-PCA technique requires 6 PCs to be retained while the number of PCs used is much higher in dynamic-PCA technique.

Table 2: Summary of monitoring results (Number of false alarms indicated in parenthesis)

Fault ID	Multiway-PCA	Dynamic-PCA	Adjoined-DPCA
	Time Fault Detected (hr)	Time Fault Detected (hr)	Time Fault Detected (hr)
PEN01	2.5 (0)	2.0 (0)	1.0 (0)
PEN02	4.0 (0)	4.0 (0)	2.0 (0)
PEN03	60.5 (10)	60.5 (24)	60.5 (0)
PEN04	30.5 (0)	30.5 (0)	30.5 (0)
PEN05	90.5 (10)	86.5 (24)	89.5 (0)
PEN06	252.0 (11)	241.0 (19)	233.0 (0)
PEN07	112.0 (5)	59.5 (8)	68.5 (0)

5. CONCLUSIONS AND DISCUSSIONS

In this paper, the shortcomings of MPCA and DPCA during online monitoring of batch processes are outlined and addressed. MPCA and DPCA are unable to adequately model fed-batch processes as they are based on the use of single block PCA for monitoring a transient trajectory. Such methods are usually prone to Type-I and Type-II errors as their applications violate the principle of *normal data distribution* based on which their monitoring limits (*SPE* and T^2) are developed upon. Exact estimation of the real distribution in multivariable system is difficult and multiple models could be used for modeling these transient, non-stationary and non-normal system. However, use of multiple PCA models are also prone to Type-I errors especially in the border region when model switching/ state change occurs as these models are discontinues whereas fed-batch operations are characterized by smooth evolution of process trajectories. In this article, we overcome these shortcomings through overlapping PCA models. The proposed adjoined-DPCA technique (ADPCA) uses multiple overlapping PCA models which allows the data densities in each DPCA model to be locally *normal distributed*. The overlap of neighboring DPCA models enforces continuity even when modeling batch-type operations. An optimal PCA model is selected at every instant for process monitoring during online application. Extensive testing of the proposed method shows its ability to reduce both Type-I and Type-II errors as it classifies the normal operating region more accurately in the principal component subspace. Such classification reduces the occurrence of Type-I errors as all relevant NOR is included in the model. Furthermore, the improvements in modeling also increases the methods sensitivity as unknown regions are also better excluded, subsequently reducing the chances of Type-II errors. Such performance is not achievable from the single block methods such as MPCA or DPCA where the well-known tradeoff between *selectivity and sensitivity* prevents their concurrent improvement. The automatic tracking of batch processes across phases based-on the criteria of *minimum-distance model selection* also allows the method to be applicable to operations of unequal batch length; with the most appropriate PCA models selected at each instant for process monitoring.

LITERATURE CITED

Bajpai, R., and Reuss, M., (1980). A mechanistic model for penicillin production, Journal of Chemical Technology and Biotechnology 30, 330-344.

Bhagwat, A., Srinivasan, R., Krishnaswamy, P.R., (2003). Multi-linear model-based fault detection during process transitions, Chemical Engineering Science 58, 1649-1670.

Birol, G., Ündey, C., Çinar, A., (2002). A modular simulation package for fed-batch fermentation: penicillin production, *Computers & Chemical Engineering 26*, 1553-1565.

Böling, J.M., Häggblom, K.E., Nyström, R.H., (2004). Multivariable uncertainty estimation based-on multi-model output matching, Journal of Process Control 14, 293-304.

Hathaway, R. J., and Bezdek, J.C., (1988). Recent convergence for the fuzzy c-means clustering algorithm, *Journal of Classification 5*, 237-247.

Hwang, D.H., Han, C, (1999). Real-time monitoring for a process with multiple operating modes, *Control Engineering Practice 7*, 891-902.

IEEE Digital Signal Processing Committee (eds), (1979). Programs for digital signal processing, IEEE Press, New York.

Jackson, J.E., (1991). A user's guide to principal components, Wiley Interscience, New York.

Jackson, J.E., and Mudholkar, G., (1979). Control procedures for residuals associated with principal component analysis, *Technometrics 21*, 341-349.

Kourti, T., (2002). Process analysis and abnormal situation detection: From theory to practice, IEEE Control System Magazine, October Issue 2002.

Ku, W., Storer, R.H., and Georgakis, (1995). Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems 30*, 179-196.

Lee, J.M., Yoo, C.K., Lee, I.B., (2004). Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis, *Journal of Biotechnology 110*, 119-136.

Ljung, L., and Glad, T., (1994). Modeling of dynamic systems, *Prentice Hall*, Englewood Cliffs.

Martin, E.B., and Morris, A.J., (1996). Non-parametric confidence bounds fro process performance monitoring charts, *Journal of Process Control 6(6),* 349-358.

Ng, Y.S, and Srinivasan, R., (2004). Transitions in the Process Industries: Opportunities and Prospective Solution, *IEEE International Symposium on Intelligent Control*, (Invited Paper), Taipei, Taiwan.

Nomikos, P., and MacGregor, J.F., (1994). Monitoring batch processes using multiway principal component analysis, *AIChE Journal 40*, 1361-1375.

Nomikos, P., and MacGregor, J.F., (1995). Multivariate SPC charts for monitoring batch processes, *Technometrics 37(1*), 41-59.

Palma, F.D., and Magni, L, (2004). A multi-model structure for model predictive control, Annual Reviews in Control 28, 47-52.

Raich, A.C., and Çinar, A., (1996). A statistical process monitoring and disturbance diagnosis in multivariate continuous processes, *AIChE Journal 42*, 995-1009.

Rännar S., MacGregor, J.F., Wold, S., (1998). Adaptive batch monitoring using hierarchical PCA, *Chemometrics and Intelligent Laboratory Systems 41*, 73-81.

Srinivasan, R., Wang, C., Ho, W. K., Lim, K. W., (2004). Dynamic PCA based methodology for clustering process states in agile chemical plants, *Industrial and Engineering Chemistry Research 43*, 2123 – 2139.

Wise, B.M., and Gallagher, B., (1996). The process chemometrics approach to process monitoring and fault detection, Journal of Process Control 6, No.6, 329-349.

Wise, B.M., Ricker, N.L., Veltkamp, D.J., and Kowalski, B.R., (1990). A theoretical basis for the use of principal components model for monitoring multivariate processes, *Process Control and Quality 1(1),* 41-51.