Multiway Gaussian Mixture Model Based Multiphase Batch Process Monitoring

Jie Yu*,† and S. Joe Qin‡

Department of Chemical Engineering The University of Texas at Austin, Austin, Texas 78712, and The Mork Family Department of Chemical Engineering and Materials Science, Ming Hsieh Department of Electrical Engineering, Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California 90089

A novel batch process monitoring approach is proposed in this article to handle batch processes with multiple operation phases. The basic idea is to combine the Gaussian mixture model (GMM) with hybrid unfolding of a multiway data matrix to partition all the sampling points into different clusters. Then, two sequential cluster alignments are used to adjust clusters so that each of them only contains consecutive sampling instants, and all the training batches at the same sampling time belong to the same cluster. The identified multiple clusters correspond to different operation phases in the batch process. Further, a localized probability index is defined to examine each sampling point of a monitored batch relative to its corresponding operation phase. Subsequently, the occurrence and duration of process faults can be detected in this way. The proposed batch monitoring approach is applied to a simulated penicillin fermentation process and compared with the conventional multiway principal component analysis (MPCA). The comparison of monitoring results demonstrates that the multiphase based approach is superior to the global MPCA method in detecting different types of faults in batch processes with a much higher detection rate and fault sensitivity.

1. Introduction

Batch or semibatch processes have been widely used in chemical, biological, food, pharmaceutical, and semiconductor industries to produce high value and low volume products.[1−4] Large scale commercial production in batch processes requires high product quality, a safe operation environment, and stable production yield. Therefore, it is necessary to monitor and ensure the multiple process and quality variables within acceptable operation regions.[5−7] During batch operation, even small disturbances on some critical variables may degrade the final product quality and yield due to the fault propagation and accumulation. The early detection of abnormal process behavior allows operators to take immediate and corrective actions, which can prevent deteriorated product quality or serious accidents from occurring and reduce the number of rejected batches.[8,9] Compared to continuous processes, batch processes suffer from a lack of fundamental knowledge, and thus, detailed process models are often unavailable, which makes batch process monitoring quite challenging.[10,11] Furthermore, the underlying dynamic behavior, finite duration, strong nonlinearity, batch-to-batch variation, and multiplicity of operation phases all pose difficulties in batch process monitoring.[11−14]

Multivariate statistical process monitoring (MSPM) techniques such as multiway principal component analysis (MPCA) and multiway partial least-squares (MPLS) have been applied to batch process monitoring.[6,15−17] This type of method relies on the assumption that the entire process data come from a single operation phase, and the batchwise unfolded data follow a multivariate Gaussian distribution approximately. However, the multiplicity of the operating phases is an inherent nature of many batch processes, and multiple phases may exhibit significantly different variable correlation structures.[13] In addition, the operation conditions of batch processes such as feedstock, production rate, temperature, and pressure are subject to change that leads to different dynamic behaviors.[18] Such characteristics of batch processes make the global MPCA and MPLS methods ill-suited and result in high rates of false detection. Some research effort has been attempted to overcome the limitations of conventional MSPM techniques. For instance, Yoo et al. employed multiway independent component analysis (MICA) to monitor a sequencing batch reactor.[14] MICA utilizes higher-order statistics and is capable of extracting non-Gaussian components with statistical independency. It searches for the directions with maximal negentropy which is simply the entropy difference between the actual signal and a purely Gaussian signal. Although the negentropy index serves as a quantitative measure of signal non-Gaussianity,[19] it is not exactly equivalent to multimodality of batch process data with multiple operation phases and therefore cannot isolate different phases. Moreover, it is not a trivial task to determine the number of independent components to be extracted for the ICA model.[20] Lu et al. proposed a stage based sub-PCA method to deal with the multiplicity of operation stages for batch processes.[13] The *k*-means clustering technique is adopted to partition the PCA loading matrices into *C* number of clusters. Zhao et al. combined this clustering based phase division algorithm with a correlation measure index to further identify the critical phase for product quality prediction in batch processes.[21] For both methods, however, the model accuracy depends on the appropriate threshold value of minimal distance between two clusters, which determines the optimal number of clusters. A large threshold results in few clusters, some of which may include misclassified data from different operation phases. On the contrary, a small threshold leads to too many clusters and degrades the monitoring efficiency. Yoo et al. attempted to build multiple probabilistic models for batch bioprocess monitoring with shifting operation conditions.[18] The Gaussian mixture model (GMM) is adopted to discriminate different operation modes and then multiple local PCA models are established along with a new discriminant measure for online monitoring. This approach is designed to tackle batch processes with multiple operation conditions. In

* To whom correspondence should be addressed. Current address: Shell Global Solutions (US) Inc., Houston, TX 77082. E-mail: j.yu@shell.com. Tel.: +1-281-544-7629. Fax: +1-281-544-7246.
† The University of Texas at Austin.
‡ University of Southern California.

10.1021/ie900479g CCC: $40.75
Published on Web 08/03/2009

the data preprocessing step, the regular MPCA is conducted in order to reduce the high dimensionality of a timewise unfolded data matrix. Then, GMM is established with the projected data in the low-dimensional principal component subspace (PCS). However, the variance index of MPCA does not necessarily reflect the inherent multimodality of data in GMM and thus may neglect certain directions with large separability of different Gaussian components. Furthermore, the local models with an isotropic variance are estimated by the expectation and maximization (EM) algorithm that requires an exhaustive searching strategy to find the optimal number of local clusters.

In this study, a new multiway Gaussian mixture model is developed to monitor batch processes with multiple operation phases. First, the multiway batch process data are batchwise unfolded and normalized for each variable at every sampling instant across different batches. Then, the mean centered data matrix is variablewise rearranged to a two-dimensional matrix with the rows representing various batches at every sampling instant and the columns corresponding to different process variables. Further, a Gaussian mixture model is estimated from the unfolded data to isolate multiple operation phases and characterize different covariance structures. Finally, a local probability index with respect to different Gaussian regions is utilized to determine whether a new batch at various phases is under normal operation or not.

The remainder of the article is organized as follows. Section 2 introduces the procedure of building a Gaussian mixture model with the Figueiredo−Jain (FJ) algorithm for clustering the data in different Gaussian regions. Then, GMM is integrated with a hybrid unfolding strategy and extended to deal with multiway batch process data in section 3. The phase based probability index is further adopted for fault detection of batch processes in section 4. In section 5, the utility of the proposed batch monitoring approach is demonstrated with a simulated fed-batch penicillin fermentation example. Finally, section 6 concludes this paper.

2. Bayesian Inference Based Gaussian Mixture Model

As a kind of statistical inference based clustering method, the Gaussian mixture model has been successfully applied to continuous process monitoring, where it provides superior capability of classifying the process data from different operation conditions.[22,23] The underlying assumption of GMM lies in that the data from multiple operation modes follows different Gaussian distributions with distinct means and covariances. For an arbitrary data sample $x \in R^m$, it may come from N possible Gaussian distributions with the corresponding probabilities. Thus, the overall probability distribution of x can be expressed as a mixture of the N Gaussian components as given below:

$$p(x|\theta) = \sum_{i=1}^{N} p(x|\theta_i)P_i \tag{1}$$

where P_i represents the prior probability of the ith Gaussian component, $p(x|\theta_i)$ denotes the conditional probability density function of x following the ith Gaussian distribution, and $p(x|\theta)$ is the probability density function of x under the N-component Gaussian mixture model. For the ith Gaussian component, the density function parameter set θ_i consists of two elements, the mean μ_i and the covariance matrix Σ_i. Its corresponding probability density function $p(x|\theta_i)$ is given by

$$p(x|\theta_i) = \frac{1}{(2\pi)^{m/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x - \mu_i)\right] \tag{2}$$

and the prior probability P_i satisfies

$$\sum_{i=1}^{N} P_i = 1 \quad (0 \le P_i \le 1) \tag{3}$$

On the basis of the Bayesian inference principle, the posterior probability of x belonging to the ith Gaussian component can be computed by

$$P(\theta_i|x) = \frac{p(x|\theta_i)P_i}{\sum_{c=1}^{N} p(x|\theta_c)P_c} \tag{4}$$

Let $\{x_1, x_2, ..., x_n\}$ be a series of training samples. The distribution parameters $\theta_i = \{\mu_i, \Sigma_i\}$ and the prior probability P_i for all the Gaussian components can be estimated by the FJ algorithm,[24] where the modified expectation and maximization procedure is iterated as follows:

- E-step

$$P^{(s)}(\theta_i|x_j) = \frac{P_i^{(s)} p(x_j|\mu_i^{(s)}, \Sigma_i^{(s)})}{\sum_{k=1}^{K} P_k^{(s)} p(x_j|\mu_k^{(s)}, \Sigma_k^{(s)})} \tag{5}$$

where $P^{(s)}(\theta_i|x_j)$ denotes the posterior probability of the jth training sample within the ith Gaussian component at the sth iteration.

- M-step

$$\mu_i^{(s+1)} = \frac{\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)x_j}{\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)} \tag{6}$$

$$\Sigma_i^{(s+1)} = \frac{\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)(x_j - \mu_i^{(s+1)})(x_j - \mu_i^{(s+1)})^{\mathrm{T}}}{\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)} \tag{7}$$

$$P_i^{(s+1)} = \frac{\max\left\{0, (\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)) - \frac{V}{2}\right\}}{\sum_{i=1}^{K} \max\left\{0, (\sum_{j=1}^{n} P^{(s)}(\theta_i|x_j)) - \frac{V}{2}\right\}} \tag{8}$$

where $\mu_i^{(s+1)}$, $\Sigma_i^{(s+1)}$, and $P_i^{(s+1)}$ are the mean, covariance, and prior probability of the ith Gaussian component at the $(s + 1)$th iteration, respectively. For each Gaussian component, the mean vector $\mu_i^{(s+1)}$ and the symmetric covariance matrix $\Sigma_i^{(s+1)}$ have m and $1/2(m^2 + m)$ scalar parameters, respectively. Therefore, $V = 1/2m^2 + 3/2m$ denotes the total number of free parameters specifying each Gaussian component. The FJ algorithm can optimize the number of clusters by annihilating the Gaussian components with insignificant prior probabilities.[24] Therefore, no a priori knowledge is required for the user to specify the number of Gaussian components

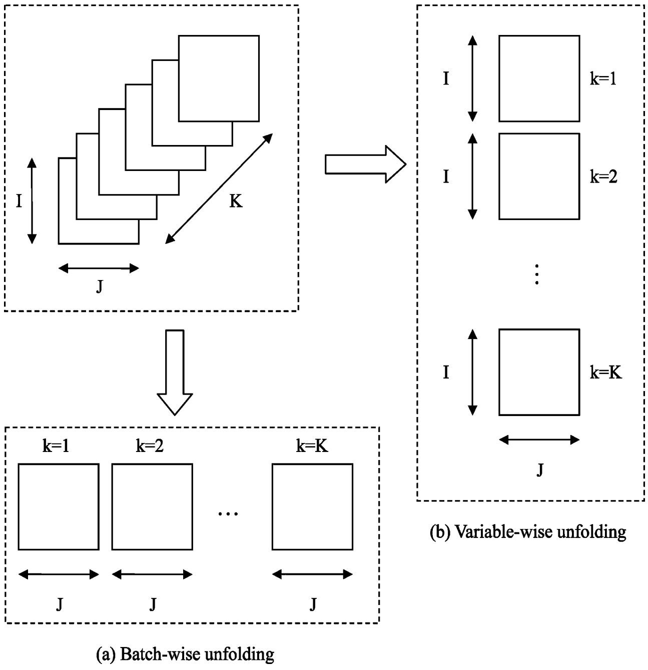

Figure 1. (a) Batchwise and (b) variablewise unfolding of three-way data matrix $X(I \times J \times K)$ into two-dimensional matrices.

or clusters. Such an advantage significantly facilitates the usage of the algorithm in practice.

3. Multiway GMM Based Segmentation of Batch Process Data

For batch processes with multiple operation phases, the process data usually follow non-Gaussian distributions and thus the traditional multivariate statistical approaches like MPCA will become inappropriate. Nevertheless, it can be assumed that the unfolded process data from each individual operation phase still follow a multivariate Gaussian distribution. Therefore, the Gaussian mixture model is applicable for clustering the batch process data from different operation phases, similar to multimode continuous process monitoring.[22,23] Unlike continuous processes, however, the data collected from batch processes are in the form of a three-way matrix that requires a preprocessing step of data unfolding prior to further analysis. Consider a three-way data matrix $X(I \times J \times K)$ from batch process, where I denotes the number of batches, J represents the number of process variables and K corresponds to the number of sampling points. Typically there are two types of data unfolding methods, i.e., batchwise unfolding and variablewise unfolding. As shown in Figure 1, the batchwise unfolded matrix $\bar{X}(I \times JK)$ has an increased number of columns JK so that the limited I batches are scattered in a high-dimensional space, which makes the estimation of Gaussian mixture model infeasible. On the other hand, the variablewise unfolded matrix $\hat{X}(IK \times J)$ neglects the dynamic trajectory along the sampling time and the inherent process nonlinearity remains in the data set.[11,16]

In this study, a hybrid unfolding strategy is employed by integrating batchwise and variablewise unfolding procedures to preprocess data matrix. First, the three-way data matrix $X(I \times J \times K)$ is transformed into a two-way matrix $\bar{X}(I \times JK)$ through batchwise unfolding. The unfolded matrix $\bar{X}$ contains sampling points from I batches within JK-dimensional space. Then, each column vector $\bar{x}_t(1 \le t \le JK)$ of matrix $\bar{X}$ is mean centered as follows

$$\hat{\bar{x}}_t = \bar{x}_t - \mu_{\bar{x}_t} \qquad (9)$$

where $\mu_{\bar{x}_t}$ is the mean value of the tth column vector $\bar{x}_t$ and $\hat{\bar{X}}$ denotes the mean centered matrix of $\bar{X}$. Assume that the block matrix format of $\hat{\bar{X}}$ is given by

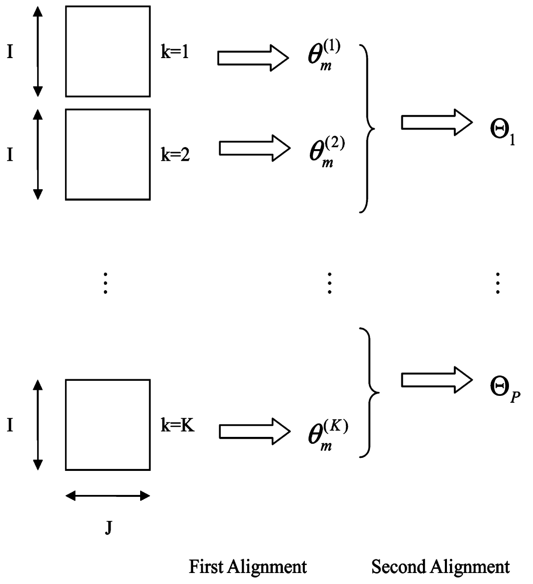

Figure 2. Schematic diagram of sequential cluster alignments across different batches and sampling instants.

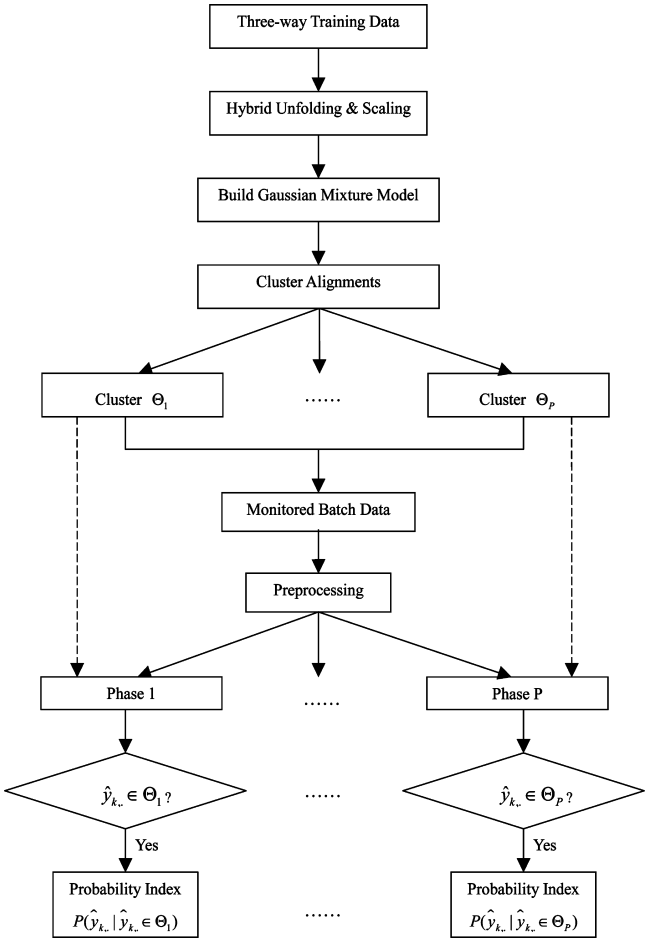
Three-way Training Data
Hybrid Unfolding & Scaling
Build Gaussian Mixture Model
Cluster Alignments
Cluster Θ_1
......
Cluster Θ_P
Monitored Batch Data
Preprocessing
Phase 1
......
Phase P
$\hat{y}_{k..} \in \Theta_1$?
......
$\hat{y}_{k..} \in \Theta_P$?
Yes
Yes
Probability Index
$P(\hat{y}_{k..} | \hat{y}_{k..} \in \Theta_1)$
......
Probability Index
$P(\hat{y}_{k..} | \hat{y}_{k..} \in \Theta_P)$

Figure 3. Flow diagram of the proposed batch process monitoring approach.

$$\widehat{\overline{\overline{X}}} = [\hat{X}_1, \hat{X}_2, ..., \hat{X}_K] \tag{10}$$

where $\hat{X}_k$ ($1 \leq k \leq K$) represents the measurement matrix of the J variables across all the I batches at the kth sampling instance. The mean centered $\widehat{\overline{\overline{X}}}$ is further rearranged to $\underline{\hat{X}}$ ($IK \times J$) through variablewise unfolding as follows

$$\underline{\hat{X}} = [\hat{X}_1{}^{\mathrm{T}}, \hat{X}_2{}^{\mathrm{T}}, ..., \hat{X}_K{}^{\mathrm{T}}]^{\mathrm{T}} \tag{11}$$

where all the J column vectors of $\underline{\hat{X}}$ also have zero mean.

With the unfolded training data matrix $\underline{\hat{X}}$, the FJ algorithm can be used to estimate the Gaussian mixture model, which consists of a number of Gaussian components corresponding to different operation phases. Let $\hat{X}_{i,k}$ be the mean centered measurement vector of the ith batch across all the J variables at the kth sampling instant and $\{\{\mu_1, \Sigma_1, P_1\}, \cdots, \{\mu_N, \Sigma_N, P_N\}\}$ be the estimated parameter values of the N Gaussian components. The posterior probability of the ith batch at the kth sampling instant belonging to the cth Gaussian component can be computed by plugging $\hat{X}_{i,k}$ into eq 4 as follows

$$P(\theta_c|\hat{x}_{i,k}) = \frac{p(\hat{x}_{i,k}|\theta_c)P_i}{\sum_{l=1}^{N} p(\hat{x}_{i,k}|\theta_l)P_l} \tag{12}$$

The decision rule for the ith batch can be expressed as

$$\hat{x}_{i,k} \in \theta_s|\{P(\theta_s|\hat{x}_{i,k}) = \max_{1\leq c\leq N}\{P(\theta_c|\hat{x}_{i,k})\}\} \tag{13}$$

which indicates that the ith batch at the kth sampling instant belongs to the Gaussian component with the maximal $P(\theta_c\hat{X}_{i,k})$ value.

It is noted that each Gaussian component corresponding to an individual operation phase should contain a series of consecutive sampling points and different batches at the same sampling instant come from the same Gaussian component. Due to estimation error, however, the GMM clustering may result in various Gaussian components even at the same sampling instant. Therefore, a first cluster alignment step is needed to assign all the I batches at each sampling instant to the most probable Gaussian component, which is defined as

$$\theta_m^{(k)} = \underset{\theta\in[\theta_1,\theta_2,...,\theta_N]}{\operatorname{argmax}} \{\omega(\hat{x}_{i,k})|\hat{x}_{i,k} \in \theta, 1 \leq i \leq I\} \tag{14}$$

where $\{\omega(\hat{x}_{i,k})|\hat{x}_{i,k} \in \theta, 1 \leq i \leq I\}$ denotes the number of batches at the kth sampling instant that belong to the Gaussian component θ. Hence, the kth sampling instant is inferred to be within the specific operation phase corresponding to the Gaussian component $\theta_m^{(k)}$.

After the first cluster alignment across different batches at the same sampling time, there may still exist some outlier time instants that break a single Gaussian component into a number of disjoint segments. Thus, the second cluster alignment should be carried out at the sampling time level to make each Gaussian component consist of consecutive sampling instants only. The procedure is to force each outlier time instant into the adjoining bulk cluster that contains a series of successive sampling points. If an outlier instant $\hat{x}_k$ is between two adjoining clusters $\theta_m^{(k-1)}$ and $\theta_m^{(k+1)}$, compute the average posterior probability of the kth sampling instant across all the I batches within those two adjoining Gaussian components as

$$P(\theta_m^{(k-1)}|\hat{x}_k) = \frac{1}{I}\sum_{i=1}^{I} P(\theta_m^{(k-1)}|\hat{x}_{i,k}) = \frac{1}{I}\sum_{i=1}^{I} \frac{p(\hat{x}_{i,k}|\theta_m^{(k-1)})P_i}{\sum_{l=1}^{N} p(\hat{x}_{i,k}|\theta_l)P_l} \tag{15}$$

and

$$P(\theta_m^{(k+1)}|\hat{x}_k) = \frac{1}{I}\sum_{i=1}^{I} P(\theta_m^{(k+1)}|\hat{x}_{i,k}) = \frac{1}{I}\sum_{i=1}^{I} \frac{p(\hat{x}_{i,k}|\theta_m^{(k+1)})P_i}{\sum_{l=1}^{N} p(\hat{x}_{i,k}|\theta_l)P_l} \tag{16}$$

The outlier time instant is then classified into the cluster with the larger average posterior probability. After the second cluster alignment, all the K sampling instants are attributed to P distinct clusters Θ_1, Θ_2, ..., Θ_P, and each cluster is composed of successive time instants only. Here, the P clusters correspond to different operation phases of the batch process. It is noted that the serial numbers of break points between two adjoining clusters can be used to partition any new batch into the P operation phases. For each cluster Θ_c ($1 \leq c \leq P$), its mean μ_{Θ_c} and covariance Σ_{Θ_c} are estimated from the variablewise unfolded matrix with a subset of data at the corresponding sampling instants.

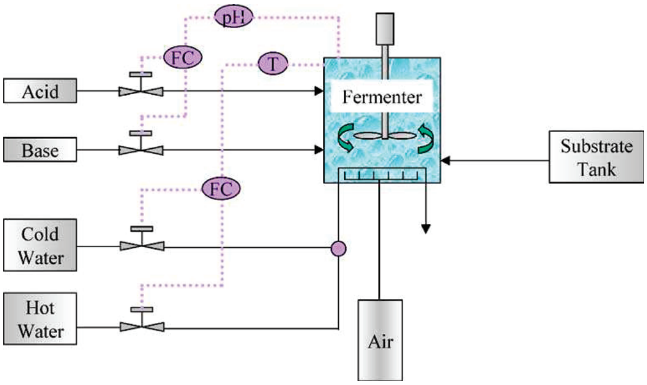

Figure 4. Flow sheet of the simulated penicillin fermentation process.[30]

Table 1. Monitored Variables in the Simulated Fed-Batch Penicillin Fermentation Process

variable no.	variable definition
1	dissolved oxygen concentration
2	culture volume
3	carbon dioxide concentration
4	pH
5	fermenter temperature
6	generated heat
7	aeration rate
8	agitator power
9	substrate feed flow rate
10	substrate feed temperature

Table 2. Simulated Faults in the Fed-Batch Penicillin Fermentation Process

fault no.	fault description	occurrence time (h)
1	10% step increase in substrate feed rate	90
2	10% step decrease in agitation power	20
3	drifting error in substrate feed rate with a slope of 0.001	60

4. Phase Based Probability Index for Batch Process Monitoring

After the various operation phases are identified by the multiway Gaussian mixture model using the training data set,

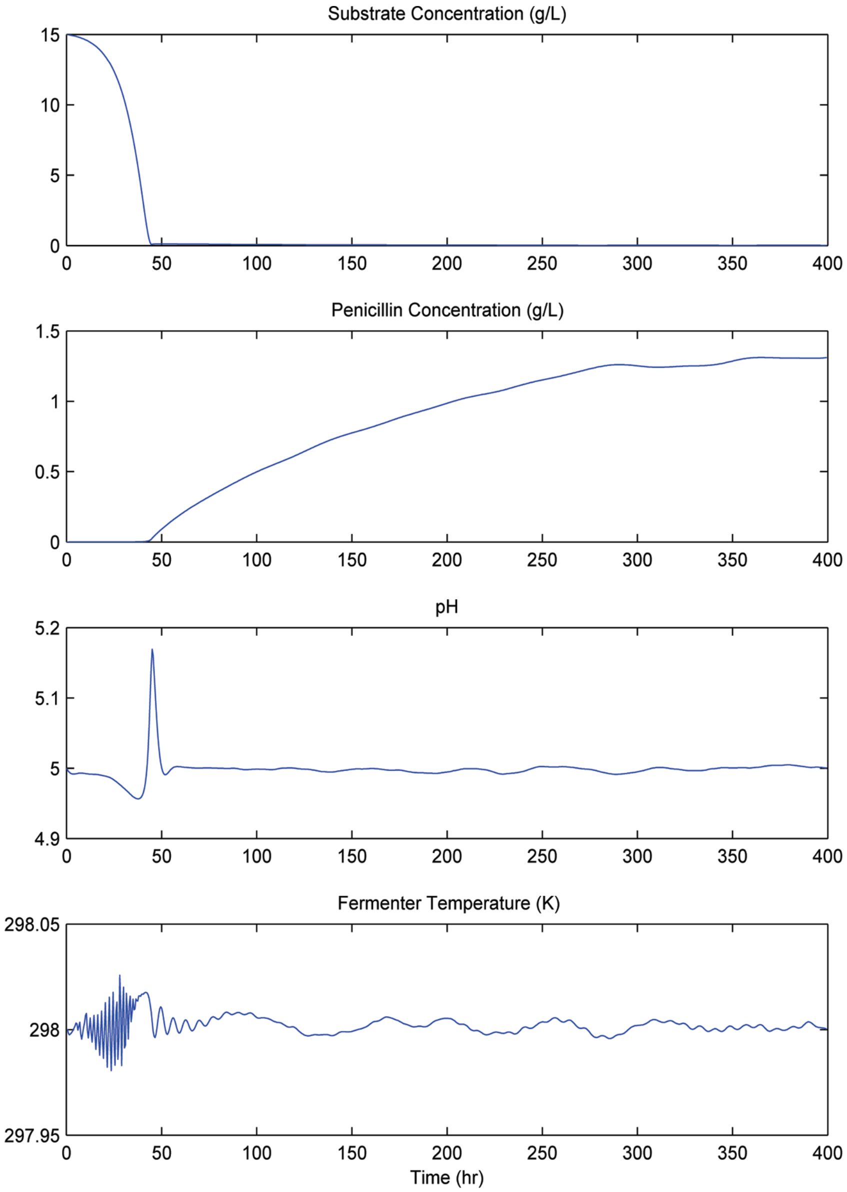

Figure 5. Trend plots of key process variables in the simulated penicillin fermentation process.

the task is to further examine the new batches and determine if they are from normal or faulty operation. As the entire batch operation consists of various phases that inherit different variable correlation structures, it becomes inappropriate to use a uniform statistical index across multiple phases. However, the unfolded process data within each single operation phase still follow a multivariate Gaussian distribution approximately. Therefore, a new test batch should first be divided into multiple phases identified by the Gaussian mixture model. Then, a local index for each operation phase can be defined to detect the faulty sampling points within the same phase.

Let $Y = [\, y_{k,j}]$ $(1 \leq k \leq K,\ 1 \leq j \leq J)$ be the data matrix of an arbitrary test batch that has K sampling instants and J variables. As a preprocessing step, the test batch matrix needs to be unfolded into a row vector $[\{y_{1,1}, ..., y_{1,J}\}, \{y_{2,1}, ..., y_{2,J}\}, ...\{y_{K,1}, ..., y_{K,J}\}]$ and each element is mean shifted by $\mu_{\bar{x}_t}$ as in eq 9

$$\hat{y}_{k,j} = y_{k,j} - \mu_{\bar{x}_t} \tag{17}$$

with $t = (k - 1) \times J + j$. Then, the above normalized row vector is refolded into the matrix format as $\hat{y} = [\hat{y}_{k,j}]$. The K row vectors $\hat{y}_1, \hat{y}_2, ..., \hat{y}_K$ are further attributed to the P phases according to the serial numbers of their sampling instants and the clustering results of training data set. For each row vector $\hat{y}_k$, assume that it belongs to the cth phase Θ_c and the following regularized Mahalanobis distance[25] is adopted to determine if the kth sampling instant of a test batch is under normal operation

$$D_r((\hat{y}_k, \Theta_c)|\hat{y}_k \in \Theta_c) = (\hat{y}_k - \mu_{\Theta_c})^{\mathrm{T}}(\Sigma_{\Theta_c} + \epsilon I)^{-1}(\hat{y}_k - \mu_{\Theta_c}) \tag{18}$$

where ϵ is a small positive number to remove the ill-condition of covariance Σ_{Θ_c}.[26] As the distance metric D_r follows an approximate χ^2 distribution[27,28]

$$D_r \sim g\chi_h^{\ 2} \tag{19}$$

with

$$g = \frac{\mathrm{tr}(\Sigma_{\Theta_c}(\Sigma_{\Theta_c} + \epsilon I)^{-1})^2}{\mathrm{tr}(\Sigma_{\Theta_c}(\Sigma_{\Theta_c} + \epsilon I)^{-1})} \tag{20}$$

and

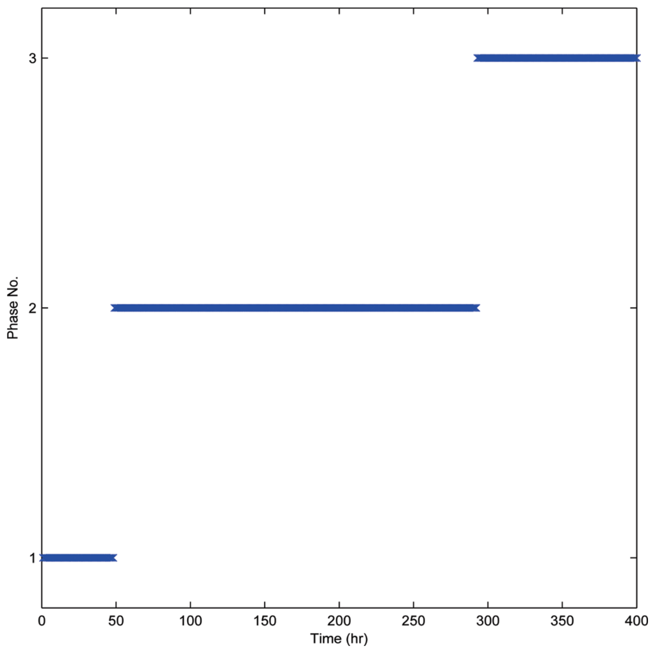

Figure 6. Operation phase segmentation of the simulated penicillin fermentation process.

$$h = \frac{[\mathrm{tr}(\Sigma_k(\Sigma_k + \epsilon I)^{-1})]^2}{\mathrm{tr}(\Sigma_k(\Sigma_k + \epsilon I)^{-1})^2} \quad (21)$$

a similar probability index as ref 23 can be defined below for fault detection

$$P(\hat{y}_k|\hat{y}_k \in \Theta_c) = \Pr\{D_r((y, \Theta_c)|y \in \Theta_c) \leq D_r((\hat{y}_k, \Theta_c)|\hat{y}_k \in \Theta_c)\} \quad (22)$$

The batch operation is considered abnormal at the *k*th sampling instant if

$$P(\hat{y}_k|\hat{y}_k \in \Theta_c) > 1 - \alpha \quad (23)$$

with 1 − α representing the confidence level.

The phase based probability index values are calculated for all the sampling instants, and thus it can be determined if the test batch operation is abnormal at any of the sampling instants. The proposed batch process monitoring procedure is illustrated in Figure 3. In the proposed approach, the initial model estimation based on iterative training may be more computationally intensive than the probability index calculations on different sampling points. However, both steps have manageable computation load for practical applications. It should also be noted that the batch lengths in real applications are probably not equal. In order to handle batches of unequal lengths, the

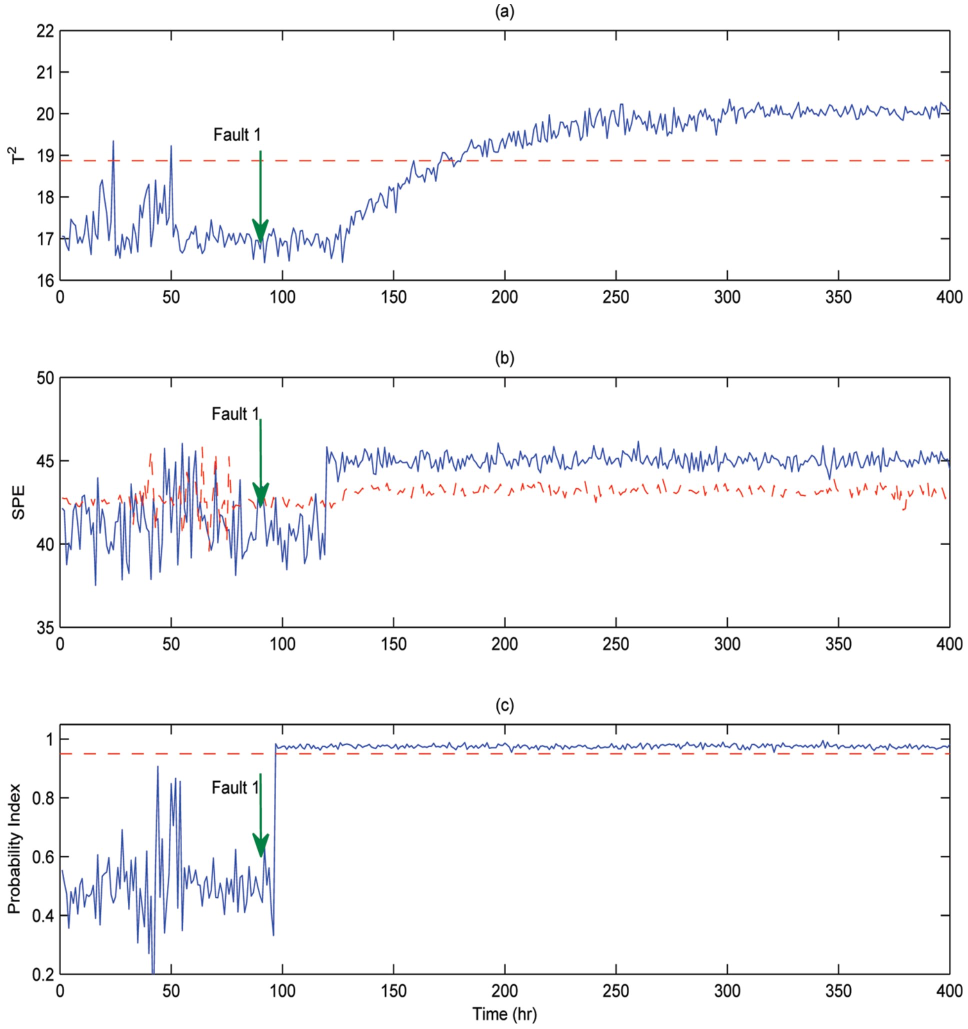

Figure 7. Comparison of batch monitoring results using MPCA based (a) T^2, (b) SPE, and (c) the proposed probability index in the first faulty case.

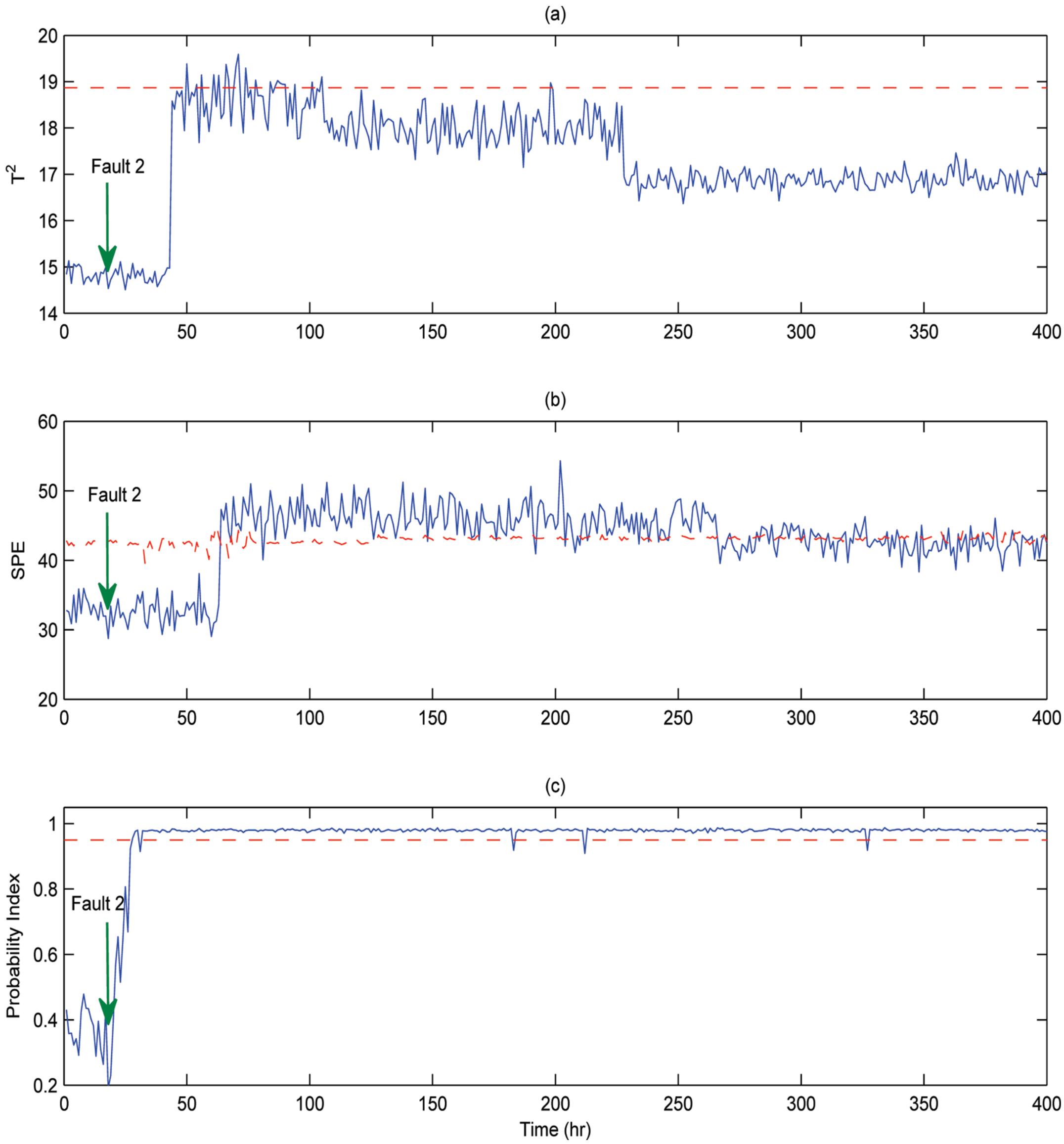

Figure 8. Comparison of batch monitoring results using MPCA based (a) T^2, (b) SPE, and (c) the proposed probability index in the second faulty case.

multivariate dynamic time warping (DTW) technique[10,29] can be used in a data preprocessing step to equalize the batch trajectory and synchronize time series.

5. Case Studies

5.1. Simulated Fed-Batch Penicillin Fermentation Process. A simulated fed-batch penicillin fermentation process is used to demonstrate the validity and effectiveness of the proposed multiphase batch process monitoring approach. The process simulator has been developed by Birol et al. for education and research purposes and is available at the website: http://www.chbe.iit.edu/∼cinar/. The flow sheet of the fermentation process is shown in Figure 4. As described in ref 30, the fermenter starts with a batch culture for biomass growth. After about 40 h, the process is switched to fed-batch operation mode when the cells enter the stationary phase with penicillin production and the substrate of glucose is being added continuously. The process simulation is based on a mechanistic model that has been extended to characterize the dynamic behavior of the batch operation. The entire duration of each batch is 400 h with a sampling interval of 0.5 h, and the ten monitored process variables are provided in Table 1. In the fermenter, two proportional, integral, and derivative (PID) loops are adopted to control the pH and temperature by adjusting acid/base and cold/hot water flow rates, respectively. On the other hand, the substrate of glucose is continuously fed into the fermenter under open-loop operation during the fed-batch mode.

To estimate the normal operation region, a total of 50 training batches are simulated with small variations in the process variables. The same initial conditions and parameter settings as the nominal operation in ref 30 are used throughout all the training batches. Then different scenarios of process faults, as listed in Table 2, are designed to examine the effectiveness of the proposed batch monitoring method. The nominal batch data of several key process variables, i.e., fermenter temperature, pH, and substrate and penicillin concentrations, are shown in Figure 5. Note that the confidence levels used in both the MPCA based method and the proposed batch monitoring approach are set to 95%.

5.2. Multiphase Batch Process Monitoring Results. For all the designed faulty cases, the proposed multiphase batch monitoring approach is applied to isolate the different operation phases and then detect the process faults. With the training data set of 50 normal operation batches, a multiway Gaussian mixture

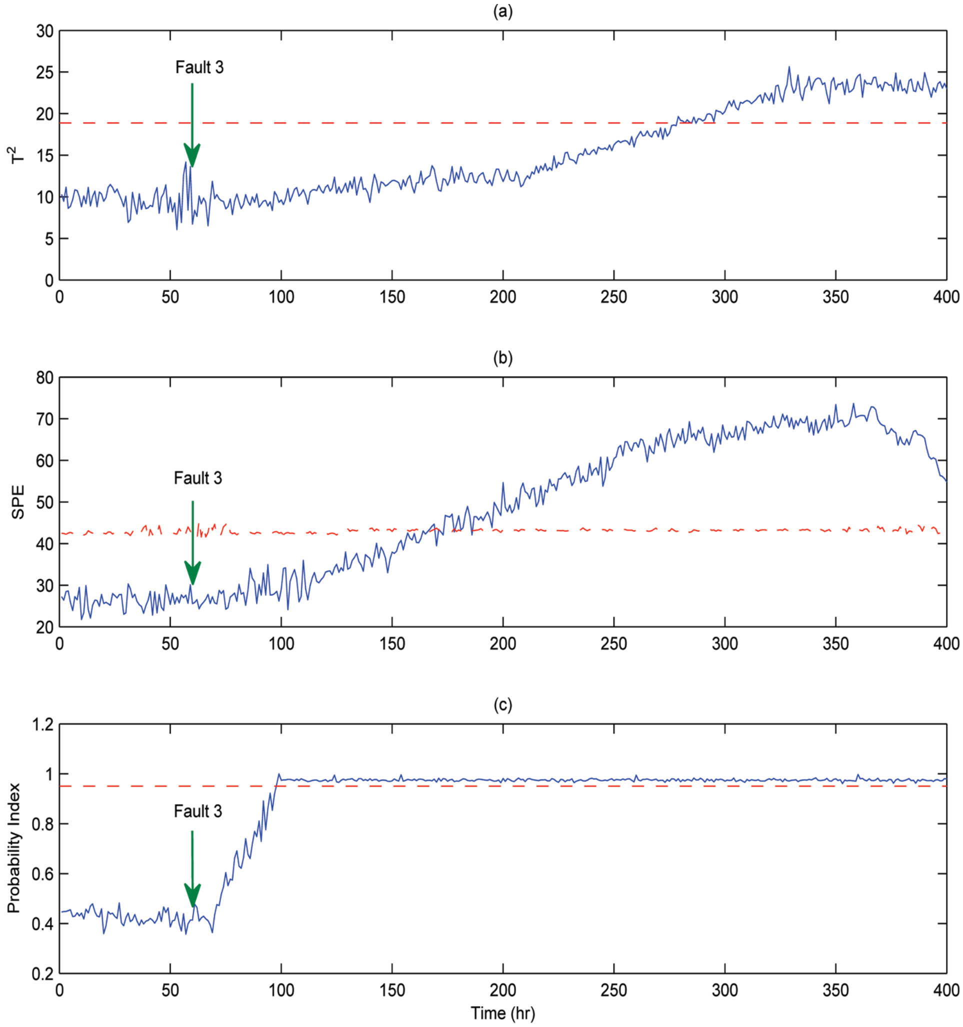

Figure 9. Comparison of batch monitoring results using MPCA based (a) T^2, (b) SPE, and (c) the proposed probability index in the third faulty case.

model is built and each batch at every sampling instant is classified into one specific Gaussian component. Further, distinct operation stages can be obtained by two sequential cluster adjustments, which are utilized to force different batches at the same instant to a single cluster and then merge the outlier clusters into the bulk ones (see Figure 2). As illustrated in Figure 6, three operation phases have been identified for this fermentation process. The first phase starts from the initial point until the 48th hour, when the cells grow fast in the fermenter. Then, it is switched to the second operation phase and ends at the 292th hour. During this stage, the operation of the bioreactor is changed from batch to fed-batch mode and the penicillin production dominates. The last phase coincides with the production saturation of penicillin when the cells are in a stationary status. The phase partition using multiway GMM matches well the operation shifts of the cell fermentation process.

The proposed monitoring approach is further applied to detect different scenarios of process faults. In the first faulty case, the step increase of substrate feed rate occurs from the 90th hour, which causes a higher production rate of penicillin. The fault detection results of the proposed multiphase batch monitoring approach are compared to those of the regular MPCA method in Figure 7, where the starting time instant of the process fault is marked. One can see that the MPCA based T^2 index is very insensitive to the step change of substrate feeding. The bias fault is not detected until about the 170th hour with significant delay. The SPE index appears to be more sensitive to this bias fault than the T^2 index with earlier detection at around the 120th hour. However, it still delays over 30 h before triggering a fault alarm. Furthermore, some normal sampling instants between the 40th and 75th hours are misclassified as faulty operations. On the other hand, the proposed multiphase batch monitoring approach leads to early detection of the bias fault starting from the 98th hour when the probability index value jumps above the confidence limit 0.95. The delay of fault detection is largely reduced to only 8 h, and the control chart in Figure 7c indicates the continuous fault alarm since the 98th hour until the end of batch simulation without any misdetection. The superior monitoring results of the proposed approach over those of the conventional MPCA method demonstrate that the multiway GMM based approach can effectively capture the dynamic changes of variable correlations. On the contrary, the MPCA method assumes that the correlation structure remains the same throughout the entire batch operation, which lessens the sensitivity to process faults.

For the second fault, a 10% step decrease is applied to agitation power from the 20th hour. The reduced agitation power then causes the decrease of dissolved oxygen, which further restricts the biomass growth as the oxygen level is one of the key factors to determine the specific growth rate. As a result, the penicillin production will be affected adversely with lower product concentration. It can be observed from Figure 8a that the MPCA based T^2 index value exceeds the 95% control limit at the 45th hour but does not give an effective fault alarm because it just moves around the threshold line until the 105th hour. Then, it falls below the control limit and cannot detect the bias fault of agitation power. The SPE index in Figure 8b, on the other hand, first alerts the fault occurrence with a 43-h delay. From the 272nd hour, the index value drops and moves around the control limit line, which indicates the high type-II error. In contrast, the proposed probability index can detect the abnormal operation much earlier at the 29th hour with only a 9-h delay, which is probably due to the fault propagation time from the agitation power to other process variables. Since then, the index value remains above the threshold line except for only three false detection points, which means that the type-II error is as low as 0.8%. Compared to the MPCA method, the multiphase based approach has largely improved monitoring accuracy and sensitivity that can be attributed to its phase dependent correlation structures.

In the third case, a drift error with the slope of 0.001 is added to the substrate feed rate starting from the 60th hour. As the glucose is increasingly fed into the fermenter during the fed-batch operation mode, the biomass and penicillin concentrations grow faster than their normal trajectory. Subsequently, the increased heat and carbon dioxide generation is accompanied with higher cooling water flow and more oxygen consumption. The monitoring results based on different control charts are shown in Figure 9. The T^2 index responds to the drifting error of glucose very slowly with a long delay of over 200 h before it initiates the out-of-control signal. Though the SPE based control chart reports this fault significantly faster with earlier drifting and larger slope, the index value does not reach the control limit line until the 177th hour that reflects a 117-h detection delay. Both T^2 and SPE based control charts are sluggish in tracking the drifting type of fault. The proposed multiphase approach, however, leads to a much more efficient control chart in Figure 9c, where the probability index quickly responds to the introduced fault by drifting from the 69th hour and then rising across the control limit line at the 99th hour. The early response and large drifting slope of the probability index shorten the detection delay and mitigate the type-II error substantially. Since the first indication of fault occurrence, the rest of the sampling points are also determined as abnormal operation correctly because their probability index values remain above the threshold without any misdetection. Similar to the previous faulty cases, the proposed approach is also more sensitive and reliable than the MPCA method in detecting the drift type of fault.

6. Conclusions

In this article, a multiway Gaussian mixture model based approach has been developed to monitor a batch process with multiple operation phases. The changes of variable correlation structures across different phases make the global MPCA method unable to characterize the multiple operation regions of the batch process. In the proposed approach, however, the Gaussian mixture model is integrated with hybrid unfolding and performed on the three-way data matrix to classify different samples into several Gaussian components. Then, two sequential cluster alignments are conducted to merge the consecutive sampling points into each single cluster, which corresponds to an operation phase of the batch process. A multiphase based local probability index is further adopted to determine if every sampling point of a monitored batch is normal or faulty.

The application example of a simulated fed-batch Penicillin fermentation process shows that the presented approach outperforms the conventional MPCA method in detecting various types of process faults with much lower detection errors and shorter delays. Furthermore, the following two advantages of the proposed approach are worthwhile to be highlighted: (i) It is a completely data-driven technique and does not require a priori knowledge of the batch process in order to partition different operation phases. This feature alleviates the algorithm requirements and makes the approach appealing in practice. (ii). The single probability index based on multiple phases instead of two complementary indices (T^2 and SPE) serves as a uniform indicator of process faults. In this research, the Gaussian mixture model based monitoring strategy is successfully extended from continuous to batch processes. Future work can be focused on fault diagnosis of multiphase batch processes in order to identify the underlying root causes.

Acknowledgment

The authors appreciate the valuable comments and suggestions of the anonymous reviewers.

Literature Cited

(1) Wise, B. M.; Gallagher, N. B.; Butler, S. W.; White, D. D.; Barna, G. G. A Comparison of Principal Component Analysis, Multiway Principal Component Analysis, Trilinear Decomposition and Parallel Factor Analysis for Fault Detection in a Semiconductor Etch Process. *J. Chemom.* **1999**, *13*, 379–396.

(2) Yue, H.; Qin, S.; Markle, R.; Nauert, C.; Gatto, M. Fault detection of plasma etchers using optical emission spectra. *IEEE Trans. Semicond. Manuf.* **2000**, *13*, 374–385.

(3) Gunther, J. C.; Conner, J. S.; Seborg, D. E. Fault detection and diagnosis in an industrial fed-batch cell culture process. *Biotechnol. Prog.* **2007**, *23*, 851–857.

(4) Doan, X.; Srinivasan, R. Online monitoring of multi-phase batch processes using phase-based multivariate statistical process control. *Comput. Chem. Eng.* **2008**, *32*, 230–243.

(5) Nomikos, P.; MacGregor, J. F. Monitoring of batch processes using multi-way principal component analysis. *AIChE J.* **1994**, *40*, 1361–1375.

(6) Kourti, T.; Nomikos, P.; MacGregor, J. F. Analysis, monitoring, and fault diagnosis of batch processes using multi-block and multi-way PLS. *J. Process Control* **1995**, *5*, 277–284.

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

(8) Lennox, B.; Hiden, H. G.; Montague, G. A.; Kornfeld, G.; Goulding, P. R. Process monitoring of an industrial fed-batch fermentation. *Biotechnol. Bioeng.* **2001**, *74*, 125–135.

(9) Yoo, C. K.; Lee, J. M.; Vanrolleghem, P. A.; Lee, I. B. On-line monitoring of batch processes using multiway independent component analysis. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 151–163.

(10) Çinar, A., Parulekar, S., Ündey, C., Birol, G. *Batch Fermentation: Modeling, Monitoring, Control*; CRC Press: New York, 2003.

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

(12) Flores-Cerrillo, J.; MacGregor, J. F. Multivariate monitoring of batch processes using batch-to-batch information. *AIChE J.* **2004**, *50*, 1219–1228.

(13) Lu, N.; Gao, F.; Wang, F. Sub-PCA Modeling and On-line Monitoring Strategy for Batch Processes. *AIChE J.* **2004**, *50*, 255–259.

(14) Yoo, C. K.; Lee, D. S.; Vanrolleghem, P. A. Application of multiway ICA for on-line process monitoring of a sequencing batch reactor. *Water Res.* **2004**, *38*, 1715–1732.

(15) Ündey, C.; Çinar, A. Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Syst. Mag.* **2002**, *10*, 40–52.

(16) Kourti, T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemom.* **2003**, *17*, 93–109.

(17) Ündey, C.; Ertunç, S.; Çinar, A. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. *Ind. Eng. Chem. Res.* **2003**, *42*, 4645−4658.

(18) Yoo, C. K.; Villez, K.; Lee, I. B.; Rosén, C.; Vanrolleghem, P. A. Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnol. Bioeng.* **2007**, *96*, 687–701.

(19) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley & Sons, Inc.: New York, 2001.

(20) Lee, J. M.; Qin, S. J.; Lee, I. B. Fault detection and diagnosis based on modified independent component analysis. *AIChE J.* **2006**, *52*, 3501–3514.

(21) Zhao, C.; Wang, F.; Mao, Z.; Lu, N.; Jia, M. Improved Knowledge Extraction and Phase-Based Quality Prediction for Batch Processes. *Ind. Eng. Chem. Res.* **2008**, *47*, 825–834.

(22) Choi, S. W.; Park, J. H.; Lee, I. B. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Comput. Chem. Eng.* **2004**, *28*, 1377–1387.

(23) Yu, J.; Qin, S. J. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J.* **2008**, *54*, 1811–1829.

(24) Figueiredo, M. A. F.; Jain, A. K. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Machine Intell.* **2002**, *24*, 381–396.

(25) De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.

(26) Mao, J.; Jain, A. K. A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Trans. Neural Networks* **1996**, *7*, 16–29.

(27) Box, G. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-way Classification. *Ann. Math. Stat.* **1954**, *25*, 290–302.

(28) Qin, S. J. Statistical process monitoring: basics and beyond. *J. Chemom.* **2003**, *17*, 480–502.

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

(30) Birol, G.; Ündey, C.; Çinar, A. A modular simulation package for fed-batch fermentation: penicillin production. *Comput. Chem. Eng.* **2002**, *26*, 1553–1565.

Received for review March 24, 2009
Revised manuscript received June 10, 2009
Accepted July 21, 2009

IE900479G