Bayesian Inference and Joint Probability Analysis for Batch Process Monitoring

Zhiqiang Ge and Zhihuan Song
Dept. of Control Science and Engineering, State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, Zhejiang, China

DOI 10.1002/aic.14119
Published online May 3, 2013 in Wiley Online Library (wileyonlinelibrary.com)

A new probabilistic monitoring method for batch processes that have multiple operating conditions is described. Particularly, for multiphase batch processes, a phase-based Bayesian inference strategy is introduced, which can efficiently combine the information of multiple operation modes together into a single model in each specific phase. Therefore, without any process knowledge, local monitoring results in different operation modes can be automatically integrated. Besides, the information of the operation mode can be obtained through joint probability analysis under the Bayesian monitoring framework. Potential extensions of the proposed method for fault diagnosis and identification are also discussed. A benchmark case study on the penicillin fermentation process is given to evaluate the feasibility and efficiency of the proposed method. It is demonstrated that the monitoring performance and the process comprehension have both been improved. © 2013 American Institute of Chemical Engineers *AIChE J*, 59: 3702–3713, 2013
Keywords: multiphase batch process, multimode, Bayesian inference, process monitoring, mode identification

Introduction

Due to the great requirement of producing low volume, high-value products, batch and semibatch processes play more and more important roles in modern industries. As a key technology for process control and management, process monitoring is an efficient tool for safety analysis and product quality improvement of batch processes. With the wide use of modern instrumental tools, a huge number of process data have been collected, depending on which data-based modeling method has become very popular in recent years. Among those developed methods, multivariate statistical process control (MSPC)-based monitoring methods such as principal component analysis, partial least squares, and independent component analysis have received much attention.[1–9]

However, typical characteristics of the batch process such as batch-to-batch variations and nonsteady-state behavior may complicate the implementation of many on-line monitoring techniques that were successfully applied in continuous processes. By extending the conventional MSPC methods to batch process, the multiway counterparts of traditional MSPC methods have been developed, such as multiway principal component analysis (MPCA), multiway partial least squares (MPLS), and multiway independent component analysis (MICA).[10–18] Particularly, for those batch processes which have multiple phases, various data-based modeling and monitoring approaches have been developed.[19–25]

To our best knowledge, most of those developed techniques have assumed that the process uncertainties are only introduced by batch-to-batch variations. In fact, due to great changes of modern market demands, different kinds of products should be produced, which means that significant changes of process operating conditions are required. So far, unfortunately, there are few methods that have been developed for monitoring batch processes with varying or multimode behaviors,[26,27] especially for those batch processes which simultaneously have multiple phases.

In multimode batch processes, batch datasets generated under different operating conditions always have different characteristics from each other. If the traditional multiway MSPC model is used for monitoring, the performance may be degraded. A straightforward solution for this problem is to build a separate monitoring model for each operation mode. When the operation mode change has been detected by the process, or known previously, the monitoring model is switched to the one corresponding to the current operating condition. However, in some cases, if we do not know the exact information of the operation mode change, there may be some risk in switching the monitoring model. Besides, the automation requirement of the process control system also demands the monitoring model to work through an unsupervised manner. Therefore, carrying out the process monitoring method automatically is important to the process control system. In this case, no model switch is need, less process knowledge is incorporated, and satisfactory monitoring performance can also be obtained.

The main aim of this article is to develop an efficient approach for monitoring multiphase batch processes under different operating conditions. It is assumed that we have no exact information of the operation mode in the process, and

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to Z. Ge at gezhiqiang@zju.edu.cn.

the datasets collected from different operation modes are mixed together in the historical database. Therefore, a data clustering method is used to divide the dataset into several subsets, which correspond to different operation modes. Without any process knowledge or expert experience, it is difficult to decide the exact operation mode for the current monitored batch. To this end, the Bayesian inference strategy is introduced, which can softly assign the current monitored batch to different operation modes. Due to the multiphase behavior of the batch process, different monitoring models are built for corresponding phases. Hence, a phase-based Bayesian combination strategy can be constructed. Another contribution of this article is the development of a new mode identification method in the batch process, which is based on joint probability analysis. Although the mode information is not required for online monitoring, it is important for product classification, process improvement, and so forth.

The rest of this article is organized as follows. First, a detailed demonstration of the proposed method is given in the next section, which includes operation mode clustering and phase division, phase-based Bayesian combination strategy, and the operation mode identification method. In section Results and Illustration, a case study on the penicillin fermentation benchmark process is illustrated to evaluate the viewpoint proposed in section Methodology. Finally, conclusions are made.

Methodology

Operation mode clustering and phase division

Before development of the monitoring model, the batch dataset should be clustered into groups, which represent different operation modes. In practice, datasets collected from different operation modes are probably mixed together in the historical database. To construct a multiphase model structure, the batch process should also be divided into several different phases, which are dominated by different physical and chemical phenomena. As a result, batch-to-batch variations are small within each operation mode and will become larger in different operation modes. On the other hand, among each batch, process characteristics are considered to be similar within the same phase and dissimilar over different phases.

To divide the batch process data into different groups, lots of unsupervised clustering methods can be used, such as K-means, Fuzzy-C means, their variants, and so forth.[28–30] For simplicity, the K-means method is used for model clustering in this article. In the past years, several phase division approaches have been developed, including the expert knowledge-based method, process analysis approaches, automatic recognition methods, optimization and pattern recognition schemes, and so forth,[25] among which the expert knowledge-based method is used here. When the batch process has been divided into different groups and phases, feature extraction and modeling procedures can then be carried out, details of which are illustrated as follows.

Two-step ICA-PCA for feature extraction and dimension reduction

Before the construction of the monitoring model, a two-step ICA-PCA (principal component analysis) feature extraction strategy is introduced to reduce the variable dimension of the process data, which can facilitate further statistical modeling and analysis. We first denote the original collected batch process data as $\mathbf{X}(I \times J \times K)$, where I is the total batch number, J is the total variable number, and K is the total data samples during each batch. Suppose a total of Q operation modes are identified, and the process has been divided into PH phases. Then, we represent the process subdatasets as $\mathbf{X}_q\left(I_q \times J \times K_{ph}\right)$, where $q=1,2,\ldots,Q$, $ph=1,2,\ldots,\mathrm{PH}$, and $\sum_{q=1}^{Q} I_q=I$, $\sum_{ph=1}^{\mathrm{PH}} K_{ph}=K$, where I_q is the batch number is each operation mode, and K_{ph} is the number of data samples in each phase. Through the variable direction, the three-way batch process datasets can be unfolded into two-dimensional datasets, denoted as $\mathbf{X}'_{q,ph}\left(K_{ph} I_q \times J\right)$, $q=1,2,\ldots,Q$, $ph=1,2,\ldots,\mathrm{PH}$. Here, we have assumed that all batches are of the same batch length and have the same number of data samples within the corresponding phases. In practice, however, different batches may have quite different batch length, and the number of data samples within each phase may also differentiate from each other. In this case, we can just rearrange the data samples of each batch in a stack manner. Suppose the length of each batch is represented as $K_i\left(i=1,2,\ldots,I_q\right)$, the modeling dataset will be $\mathbf{X}_q\left(\sum_{i=1}^{I_q} K_i \times J\right)$. Similarly, the unequal phase length problem can also be easily addressed. Therefore, without loss of generality, the numbers of data samples in each batch and within each phase are both assumed to be constant in the present article.

Then, ICA is carried out in the first step to extract the high-order information, therefore, $\mathbf{X}'_{q,ph}$ is decomposed as follows[31]

$$\begin{aligned} \mathbf{X}'_{q,ph} &= \mathbf{A}_{q,ph} \cdot \hat{\mathbf{S}}_{q,ph} + \mathbf{X}''_{q,ph} \\ \hat{\mathbf{S}}_{q,ph} &= \mathbf{W}_{q,ph} \mathbf{X}'_{q,ph} \\ \mathbf{X}''_{q,ph} &= \mathbf{X}'_{q,ph} - \mathbf{A}_{q,ph} \cdot \hat{\mathbf{S}}_{q,ph} \end{aligned} \tag{1}$$

where $q=1,2,\ldots,Q$, $ph=1,2,\ldots,\mathrm{PH}$, $\mathbf{A}_{q,ph}$, and $\mathbf{W}_{q,ph}$ is the mixing and demixing matrix, $\mathbf{X}''_{q,ph}$ is the residual matrix after the ICA step. In the second step, PCA is used to model the Gaussian information of the process data. The decomposition is carried out as

$$\mathbf{X}''_{q,ph} = \mathbf{T}_{q,ph} \mathbf{P}^T_{q,ph} + \mathbf{R}_{q,ph} \tag{2}$$

where $\mathbf{T}_{q,ph}$ and $\mathbf{P}_{q,ph}$ are score and loading matrices of the PCA decomposition, and $\mathbf{R}_{q,ph}$ is the residual matrix after the analysis of PCA. To determine the numbers of independent components and principal components in the ICA and PCA models, various methods can be applied, such as variance-based approach, non-Gaussianity testing method, among others.[32,33]

Monitoring statistic construction and confidence limit determination

After feature extraction and dimension reduction, the phase-based monitoring models can be built for each phase under different operation modes. Traditionally, I^2, T^2, and SPE monitoring statistics and their corresponding confidence limits can be developed as follows[3,31]

$$I^2 = \hat{\mathbf{s}}^T_{q,ph,i} \hat{\mathbf{s}}_{q,ph,i} \le I^2_{q,ph,\lim} \tag{3}$$

$$T^2 = \sum_{i=1}^{k_{q,cph}} \frac{t^2_{q,ph,i}}{\lambda_i} \le T^2_{q,ph,\lim} = \frac{k_{q,ph}\left(K_{ph} I_q - 1\right)}{K_{ph} I_q - k_{q,ph}} F_{k_{q,ph},\left(K_{ph} I_q - k_{q,ph}\right),\alpha} \tag{4}$$

$$\mathrm{SPE} = \mathbf{r}_{q,ph,i} \mathbf{r}^T_{q,ph,i} \le \mathrm{SPE}_{q,ph,\lim} = g_{q,ph} \chi^2_{h_{q,ph},\alpha} \tag{5}$$

where $q=1,2,\ldots,Q$, $ph=1,2,\ldots,PH$, $\hat{\mathbf{s}}_{q,ph,i}$, $\mathbf{t}_{q,ph,i}$, and $\mathbf{r}_{q,ph,i}$ are vectors of $\hat{\mathbf{S}}_{q,ph}$, $\mathbf{T}_{q,ph}$, and $\mathbf{R}_{q,ph}$, λ_i is the eigenvalue corresponding to each PC, $k_{q,cph}$ is the number of PCs, α is the selected significance level, and $g_{q,ph}=v_{q,ph}/\left(2m_{q,ph}\right)$, $h_{q,ph}=2m_{q,ph}^2/v_{q,ph}$, in which $m_{q,ph}$ and $v_{q,ph}$ are the mean and variance values of SPE within operation mode q. The confidence limit of the I^2 statistic can be determined by kernel density estimation. However, a more appropriate and efficient method to determine the confidence limit of the I^2 statistic is the support vector data description (SVDD) method, which has recently been introduced to approximate the IC distributions.[34] To construct the minimum volume of the hypersphere, SVDD solve the following optimization problem

$$\begin{aligned}&\min_{R,a,\xi} R_{q,ph}^2+C_q\sum_{i=1}^{K_{ph}I_q}\xi_{q,ph,i}\\ &\mathbf{s}.t.\left\|\Phi\left(\hat{\mathbf{s}}_{q,ph,i}\right)-\mathbf{a}_{q,ph}\right\|^2\leq R_{q,ph}^2+\xi_{q,ph,i},\ \xi_{q,ph,i}\geq 0\end{aligned}\tag{6}$$

where $\mathbf{a}_{q,ph}$ is the center of the hypersphere, C_q gives the trade-off between the volume of the hypersphere and the number of errors. $\xi_{q,ph,i}$ represents the slack variable which allows the probability that some of the training samples can be wrongly classified, $\Phi(\cdot)$ is a nonlinear transform function. Suppose $K(\cdot)$ is the kernel function which is often selected as the Gaussian kernel function, the center $\mathbf{a}_{q,ph}$ and the radius $R_{q,ph}$ can be determined by[34]

$$\begin{aligned}&\mathbf{a}_{q,ph}=\sum_{i=1}^{K_{ph}I_q}\alpha_i\Phi\left(\hat{\mathbf{s}}_{q,ph,i}\right)\\ &R_{q,ph}=\sqrt{1-2\sum_{i=1}^{K_{ph}I_q}\alpha_iK\left(\hat{\mathbf{s}}_{q,ph,0},\hat{\mathbf{s}}_{q,ph,i}\right)+\sum_{i=1}^{K_{ph}I_q}\sum_{j=1}^{K_{ph}I_q}\alpha_i\alpha_jK\left(\hat{\mathbf{s}}_{q,ph,i},\hat{\mathbf{s}}_{q,ph,j}\right)}\end{aligned}\tag{7}$$

where $\hat{\mathbf{s}}_{q,ph,0}$ is referred to a support vector. Then, the new non-Gaussian monitoring statistic can be defined as[35]

$$NG\mathbf{S}_i=d^2\left(\Phi\left(\hat{\mathbf{s}}_{q,ph,i}\right)\right)=\left\|\Phi\left(\hat{\mathbf{s}}_{q,ph,i}\right)-\mathbf{a}_{q,ph}\right\|^2\leq NG\mathbf{S}_{q,ph,\lim}=R_{q,ph}^2\tag{8}$$

Phase-based Bayesian monitoring approach

After data scaling and rearrangement of the new data sample in the current batch, represented as $\mathbf{x}'_{\text{new},k}(1\times J)$, the monitoring statistics can be calculated as follows

$$\hat{\mathbf{s}}_{q,\text{new},k}=\mathbf{W}_{q,cph}\mathbf{x}'_{\text{new},k}\tag{9}$$

$$NG\mathbf{S}_{q,\text{new},k}=d^2\left(\Phi\left(\hat{\mathbf{s}}_{q,\text{new},k}\right)\right)=\left\|\Phi\left(\hat{\mathbf{s}}_{q,\text{new},k}\right)-\mathbf{a}_{q,cph}\right\|^2\tag{10}$$

$$\mathbf{x}''_{\text{new},k}=\mathbf{x}'_{\text{new},k}-\mathbf{A}_{q,cph}\hat{\mathbf{s}}_{q,\text{new},k}\tag{11}$$

$$\mathbf{t}_{q,\text{new},k}=\mathbf{x}''^{T}_{\text{new},k}\mathbf{P}_{q,cph}\tag{12}$$

$$T_{q,\text{new},k}^2=\sum_{i=1}^{k_{q,cph}}\frac{t_{q,\text{new},k,i}^2}{\lambda_i}\tag{13}$$

$$\mathbf{r}_{\text{new},k}=\mathbf{x}''_{\text{new},k}-\mathbf{t}_{q,\text{new},k}\mathbf{P}_{q,cph}^T\tag{14}$$

$$\text{SPE}_{q,\text{new},k}=\mathbf{r}_{\text{new},k}\mathbf{r}_{\text{new},k}^T\tag{15}$$

where $q=1,2,\ldots,Q$, cph represents the current phase that the monitored data sample belongs to. Before the introduction of Bayesian inference strategy for posterior probability calculation, a transformation from the monitoring statistic to the probability value should be made, which is defined as follows

$$P_{\text{NGS}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)=\exp\left\{-\frac{\text{NGS}_{q,\text{new},k}}{\text{NGS}_{q,cph,\lim}}\right\}\tag{16}$$

$$P_{T^2}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)=\exp\left\{-\frac{T_{q,\text{new},k}^2}{T_{q,cph,\lim}^2}\right\}\tag{17}$$

$$P_{\text{SPE}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)=\exp\left\{-\frac{\text{SPE}_{q,\text{new},k}}{\text{SPE}_{q,cph,\lim}}\right\}\tag{18}$$

where $q=1,2,\ldots,Q$, based on the Bayesian inference,[36] the posterior probabilities of each operation mode corresponding to the three monitoring statistics are given as

$$\begin{aligned}P_{\text{NGS}}\left(q,cph|\mathbf{x}'_{\text{new},k}\right)&=\frac{P_{\text{NGS}}\left(q,cph,\mathbf{x}'_{\text{new},k}\right)}{P_{\text{NGS}}\left(\mathbf{x}'_{\text{new},k}\right)}\\ &=\frac{P_{\text{NGS}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)}{\sum_{q=1}^{Q}\left[P_{\text{NGS}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)\right]}\end{aligned}$$

$$\begin{aligned}P_{T^2}\left(q,cph|\mathbf{x}'_{\text{new},k}\right)&=\frac{P_{T^2}\left(q,cph,\mathbf{x}'_{\text{new},k}\right)}{P_{T^2}\left(\mathbf{x}'_{\text{new},k}\right)}\\ &=\frac{P_{T^2}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)}{\sum_{q=1}^{Q}\left[P_{T^2}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)\right]}\end{aligned}\tag{20}$$

$$\begin{aligned}P_{\text{SPE}}\left(q,cph|\mathbf{x}'_{\text{new},k}\right)&=\frac{P_{\text{SPE}}\left(q,cph,\mathbf{x}'_{\text{new},k}\right)}{P_{\text{SPE}}\left(\mathbf{x}'_{\text{new},k}\right)}\\ &=\frac{P_{\text{SPE}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)}{\sum_{q=1}^{Q}\left[P_{\text{SPE}}\left(\mathbf{x}'_{\text{new},k}|q,cph\right)P(q,cph)\right]}\end{aligned}\tag{21}$$

where $P(q,cph)$, $q=1,2,\ldots,Q$ are prior probabilities at the specific phase for the monitored data sample, which can be simply determined as

$$P(q,cph)=\frac{K_{cph}I_q}{K_{cph}I}\tag{22}$$

After the posterior probabilities of the new data sample $\mathbf{x}'_{new,k}$ have been obtained, we should decide its fault probability in each operation mode, which can be calculated as follows

$$P_{f,\text{NGS}}^{q,cph}\left(\mathbf{x}'_{\text{new},k}\right)=\Pr\left\{\text{NGS}_{q,cph}\left(\mathbf{x}_{tr,q,cph}\right)\leq\text{NGS}_{q,cph}\left(\mathbf{x}'_{\text{new},k}\right)\right\}\tag{23}$$

$$P_{f,T^2}^{q,cph}\left(\mathbf{x}'_{\text{new},k}\right)=\Pr\left\{T_{q,cph}^2\left(\mathbf{x}_{tr,q,cph}\right)\leq T_{q,cph}^2\left(\mathbf{x}'_{\text{new},k}\right)\right\}\tag{24}$$

$$P_{f,\text{SPE}}^{q,cph}\left(\mathbf{x}'_{new,k}\right)=\Pr\left\{\text{SPE}_{q,cph}\left(\mathbf{x}_{tr,q,cph}\right)\leq\text{SPE}_{q,cph}\left(\mathbf{x}'_{\text{new},k}\right)\right\}\tag{25}$$

where $q=1,2,\ldots,Q$, $\mathbf{x}_{tr,q,cph}$ is the training samples in operation mode q and phase cph. The values of these three probabilities can be simply determined by measuring the number of the training samples whose statistic values are smaller than that of the new data sample. Alternatively, they can also be determined more precisely by density estimation method with appropriate level of significance.

Then, the new phase-based Bayesian combination monitoring statistics (PBC) can be constructed based on the posterior probabilities and the fault probabilities as

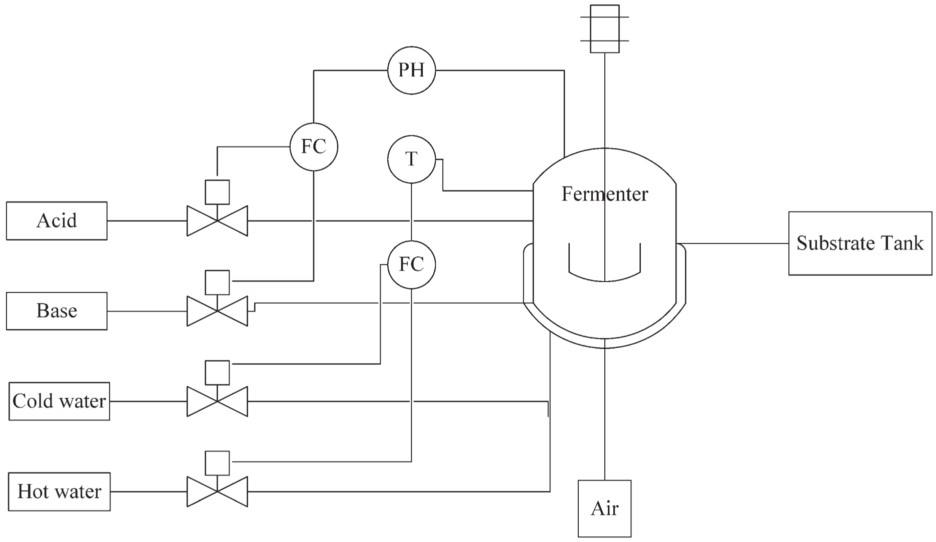

Figure 1. Penicillin fermentation process.

$$\mathrm{PBC}_{\mathrm{NGS}}\left(\mathbf{x}'_{\mathrm{new},k}\right)=\sum_{q=1}^{Q}\left[P_{\mathrm{NGS}}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)P_{f,\mathrm{NGS}}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)\right] \tag{26}$$

$$\mathrm{PBC}_{T^2}\left(\mathbf{x}'_{\mathrm{new},k}\right)=\sum_{q=1}^{Q}\left[P_{T^2}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)P_{f,T^2}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)\right] \tag{27}$$

$$\mathrm{PBC}_{\mathrm{SPE}}\left(\mathbf{x}'_{\mathrm{new},k}\right)=\sum_{q=1}^{Q}\left[P_{\mathrm{SPE}}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)P_{f,\mathrm{SPE}}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)\right] \tag{28}$$

Because the values of $P_{f,\mathrm{NGS}}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)$, $P_{f,T^2}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)$, and $P_{f,\mathrm{SPE}}^{q,cph}\left(\mathbf{x}'_{\mathrm{new},k}\right)$ are all ranged from zero to one, and the posterior probabilities $P_{\mathrm{NGS}}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)$, $P_{T^2}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)$, and $P_{\mathrm{SPE}}\left(q,cph|\mathbf{x}'_{\mathrm{new},k}\right)$ have been normalized, hence, the bounds of all phase-based Bayesian combination monitoring statistics are also ranged from zero to one. Under a prespecified significance level α, the new data sample $\mathbf{x}'_{\mathrm{new},k}$ is determined to be normal if all of the $\mathrm{PBC}_{\mathrm{NGS}}\left(\mathbf{x}'_{\mathrm{new},k}\right)$, $\mathrm{PBC}_{T^2}\left(\mathbf{x}'_{\mathrm{new},k}\right)$, and $\mathrm{PBC}_{\mathrm{SPE}}\left(\mathbf{x}'_{\mathrm{new},k}\right)$ values are not larger than $1-\alpha$. Otherwise, this data sample should be treated as an abnormality.

Mode identification and updating

To know the mode information of the monitored batch, a new mode identification scheme is developed based on the Bayesian monitoring framework. Having calculated the posterior probabilities of the monitored batch correspond to different operation modes, the mode identification scheme can be simply implemented through these posterior probabilities, thus the mode with the biggest posterior probability value should be determined as the current operation mode. However, based on the normalization of the posterior probability, any operation batch would be assigned to its most similar one of the known operation modes. A possible pitfall of this method is that, if an unknown operation mode or some fault has happened, it will also be assigned to one of the known operation modes. Hence, if we use the posterior probability for mode identification, there will be inevitable false identifications and misunderstandings of the process which may cause further risks.

To reduce the risk and also to enhance the process reliability, the joint probability analysis method is used in this work. Similar to the posterior probability, successful

Table 1. Variables Used in the Monitoring of the Penicillin Simulation Benchmark

No.	Variables
1	Aeration rate (L/h)
2	Agitator power (W)
3	Glucose feed temperature (K)
4	Dissolved oxygen concentration (% saturation)
5	Culture volume (L)
6	Carbon dioxide concentration (mmol/L)
7	pH
8	Temperature (K)
9	Cooling water flow rate (L/h)

Table 2. Fault Descriptions in the Penicillin Fermentation Process

Batch number	Fault type	Fault presentation
1	Step	The process initially runs under the first operation mode, then a step increased of aeration rate by 1% at 200 h happens.
2	Step	The process initially runs under the second operation mode, then a step increased of agitator power by 1% at 200 h happens.
3	Ramp	The process initially runs under the first operation mode, then a slow varying of aeration rate from 200 h to the end of the batch with a slope of 0.2 happens.
4	Ramp	The process initially runs under the second operation mode, then a slow varying of agitator power from 200 h to the end of the batch with a slope of 0.2 happens.
5	Step	The process initially runs under the third operation mode, then a step increased of aeration rate by 1% at 200 h happens.
6	Ramp	The process initially runs under the third operation mode, then a slow varying of agitator power from 200 h to the end of the batch with a slope of 0.2 happens.

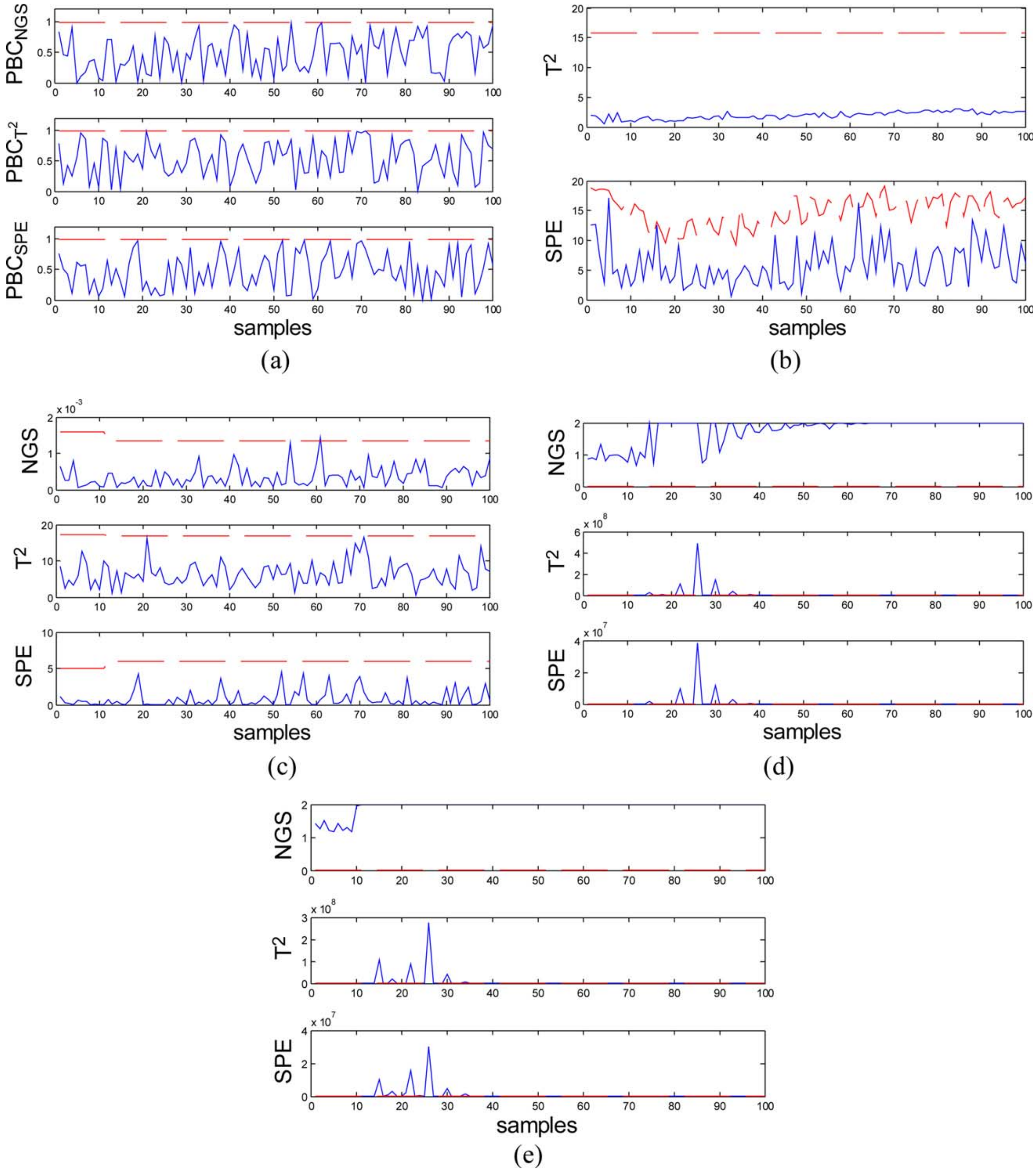

Figure 2. Monitoring results of the first normal batch, (a) PBC; (b) MPCA; (c) first single model; (d) second single model; and (e) third single model.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

identification will be obtained if the process was operated under the operation mode that was previously known. However, although the posterior probability fails to identify the correct operation mode in case of the new operation mode and fault, the joint probability can successfully identify the change of the process. That is, if a new operation mode or a fault happens, the joint probabilities of the monitored data sample with all operation modes will decrease to zero. Compared to the posterior probability, the joint probability can reflect the process change more accurately; therefore, it can bring more comprehensive understandings to the process. If a new operation mode has been detected, we can build the new local monitoring model for this new operation mode. By adding the new local model to the model pool, the process mode cases can be updated.

Depending on the calculated probabilities of the monitored data sample in each operation mode, which are given in Eqs. 16–18, the joint probabilities of this monitored data sample with each operation mode can be calculated as follows

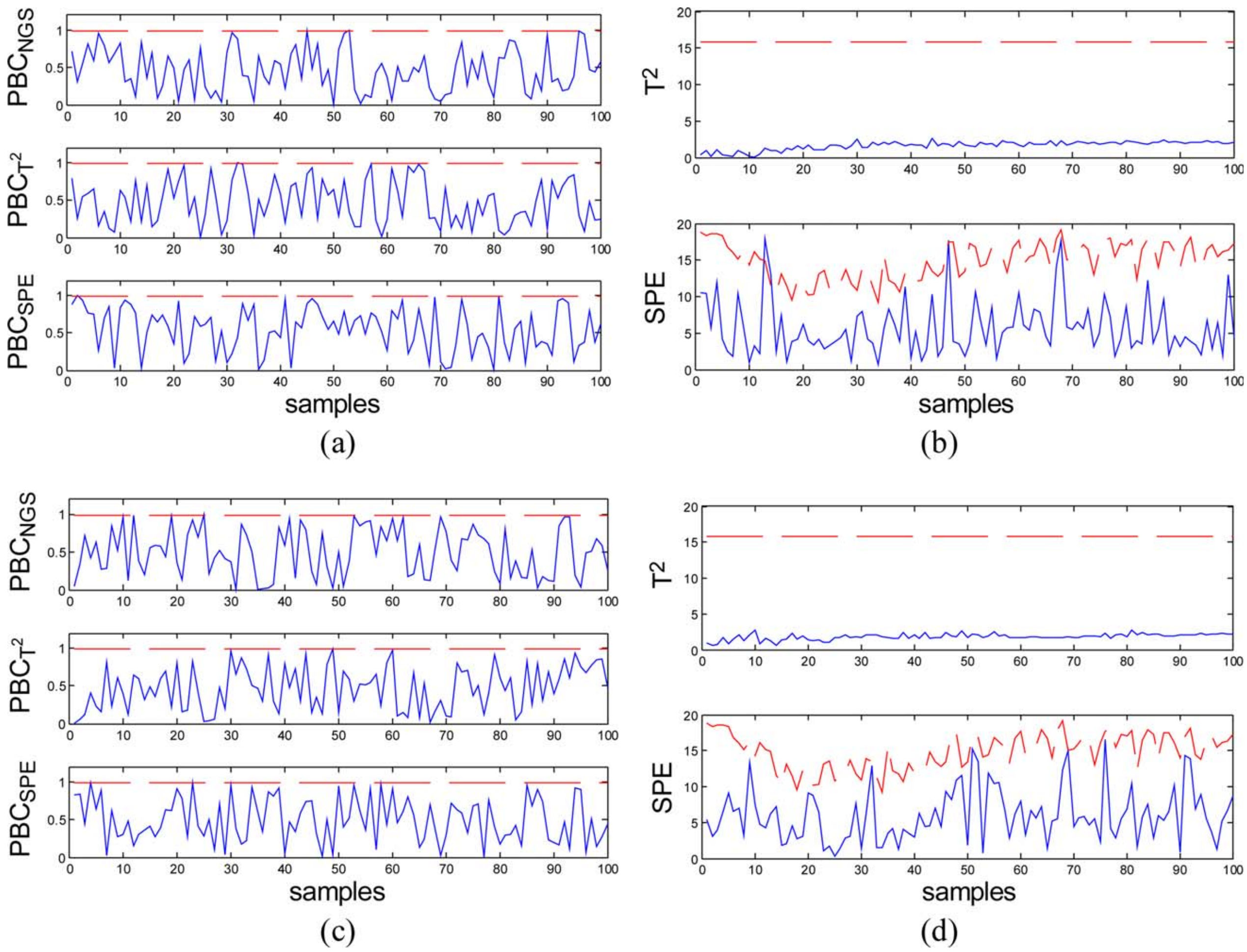

Figure 3. Monitoring results of the second and third normal batches, (a) PBC for second normal batch; (b) MPCA for second normal batch; (c) PBC for third normal batch; and (d) MPCA for third normal batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$P_{\mathrm{NGS}}\left(\mathbf{x}'_{\mathrm{new},k}, q, cph\right) = P_{\mathrm{NGS}}\left(\mathbf{x}'_{\mathrm{new},k}|q, cph\right)P(q, cph) \quad (29)$$

$$P_{T^2}\left(\mathbf{x}'_{\mathrm{new},k}, q, cph\right) = P_{T^2}\left(\mathbf{x}'_{\mathrm{new},k}|q, cph\right)P(q, cph) \quad (30)$$

$$P_{\mathrm{SPE}}\left(\mathbf{x}'_{\mathrm{new},k}, q, cph\right) = P_{\mathrm{SPE}}\left(\mathbf{x}'_{\mathrm{new},k}|q, cph\right)P(q, cph) \quad (31)$$

where *a priori* probability of each operation mode in the current phase is given in Eq. 22.

Results and Illustrations

In this section, the feasibility and efficiency of the proposed phase-based Bayesian combination monitoring method is evaluated through the well-known penicillin benchmark process, the detailed description of which was given in Birol et. al.[37]

Process description and data preparation

In typical penicillin fermentation, most of the necessary cell mass is generated during the initial preculture stage. The penicillin starts to be produced at the exponential growth phase and continues to be produced until cell growth reaches the stationary phase. Cell growth must continue at a certain minimum rate to maintain high penicillin productivity. It is for this reason that glucose is fed continuously into the system during fermentation instead of being added all at once at the beginning. In this study, batch process data was generated using a simulator (PenSim v2.0) developed by the monitoring and control group of the Illinois Institute of Technology.[37] The flow sheet of the penicillin cultivation process is illustrated in Figure 1. The system switches itself to the fed-batch phase of operation when the glucose concentration reaches a certain threshold value, which is chosen as 0.3 g L^{-1} in this study. The duration of a whole batch is selected as 400 h, and the system switches to the fed-batch phase after about 45 h. Therefore, this batch process has two phases. The sampling interval is chosen as 4 h, thus 100 data samples can be generated during each batch running. The selected monitoring variables are listed in Table 1.

To simulate the multimode behavior of this batch process, the initial culture volume is artificially set to different values. In this study, three operating conditions are generated, corresponding to 100, 105, and 110 L of the initial culture volumes, which are denoted as the first, second, and third operation mode in the following illustrations. Under each operating condition, 60 batches are generated, with 50 batches for modeling training purpose and 10 batches for testing. Therefore, a total of 180 normal batches have been obtained $\mathbf{X}(180\times9\times100)$, with 150 batches for model development $\mathbf{X}_{tr}(150\times9\times100)$ and 30 batches for testing $\mathbf{X}_{te}(30\times9\times100)$. The training batches can be grouped into three different clusters, which represent three operation modes, and the batch process is divided into two phases,

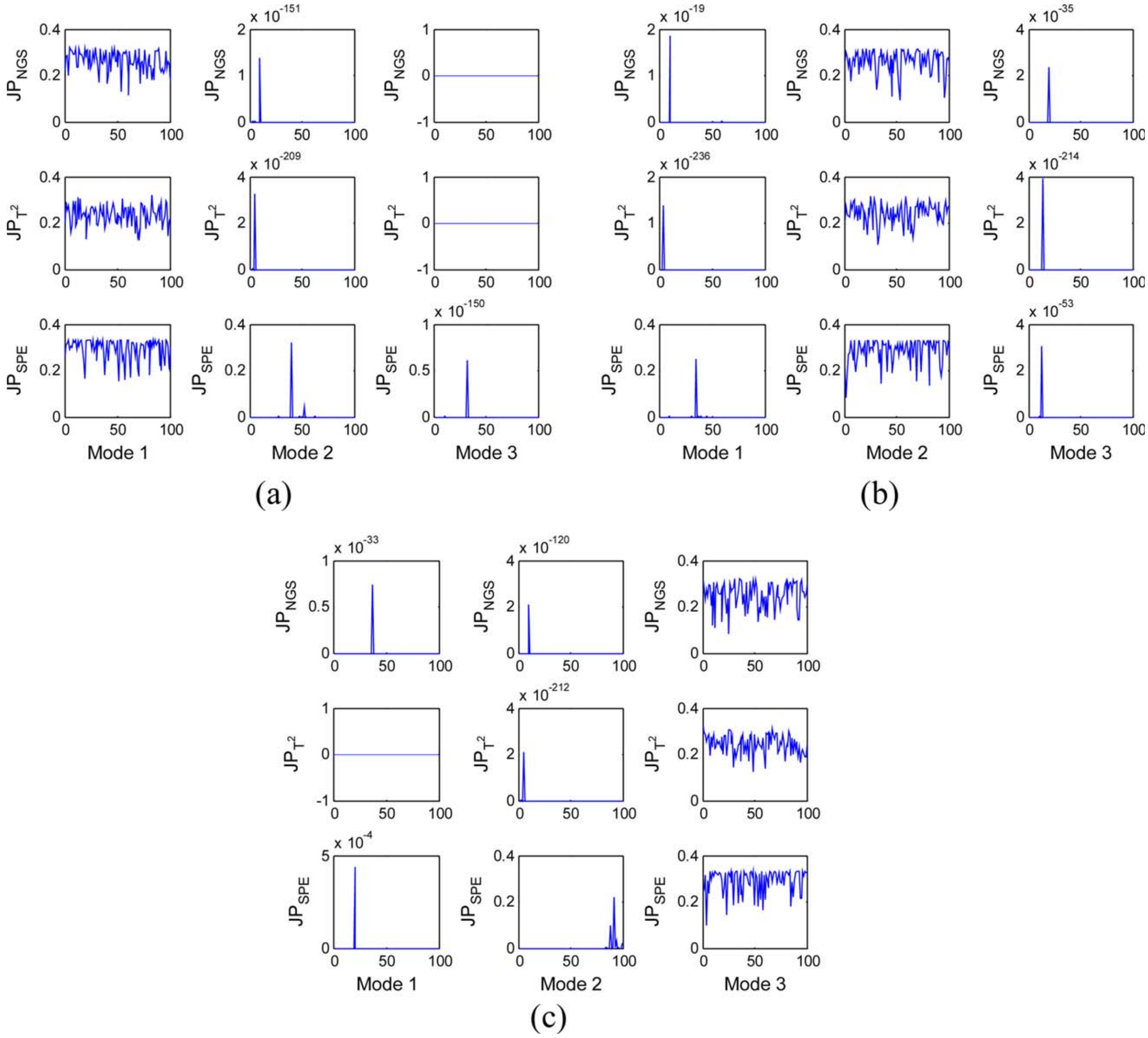

Figure 4. Joint probability results of the three normal batches in different operation modes, (a) first normal batch; (b) second normal batch; and (c) third normal batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

with 11 data samples in the first phase, and 89 in the second one. Therefore, the phase-wise datasets are generated as $\mathbf{X}_{tr,ph1}(150\times9\times11)$, $\mathbf{X}_{tr,ph2}(150\times9\times89)$, $\mathbf{X}_{te,ph1}(30\times9\times11)$, and $\mathbf{X}_{te,ph2}(30\times9\times89)$. To test the fault detection ability of the proposed method, several faults are introduced to the process, which are listed in Table 2. The faulty datasets are represented as $\mathbf{X}_{\text{fault}}(6\times9\times100)$.

Result illustrations and discussions

After the dataset has been generated, the phase-based Bayesian combination monitoring model is developed. The parameters for SVDD model development are tuned that the false classification rate is controlled as 1%, thus the 99% confidence limit can be obtained. The component number of each PCA or ICA model can be determined by the cumulative percentage variance (CPV) method when $\text{CPV} > 85\%$. To evaluate the feasibility and efficiency of the proposed method, three normal batches collected in different operation modes are tested. The monitoring results of the first normal batch collected from the first operation mode are given in Figure 2. It can be seen that both PBC and MPCA methods indicate that this batch is operated under normal process condition, because all monitoring statistics are under their corresponding control limits. However, if we build separate models of different operation modes for monitoring, only the model which corresponds to the monitored batch will give the right results, which are shown in Figures 2c–e. Therefore, when we switch the wrong model for process monitoring, false alarms could be generated. Similar testing results of the second and the third normal batches can be obtained, which are demonstrated in Figure 3. To identify the mode information of these three normal batches, joint probability analyses are carried out, the results of which are illustrated in Figure 4. From this figure, one can easily determine correct operation modes for these normal testing batches.

To evaluate the fault detection capability of the proposed method, six faults (listed in Table 2) are generated, which are introduced on different operation modes. First, a step change of the aeration rate is introduced under the first operation mode at 200 h. The monitoring results of both PBC

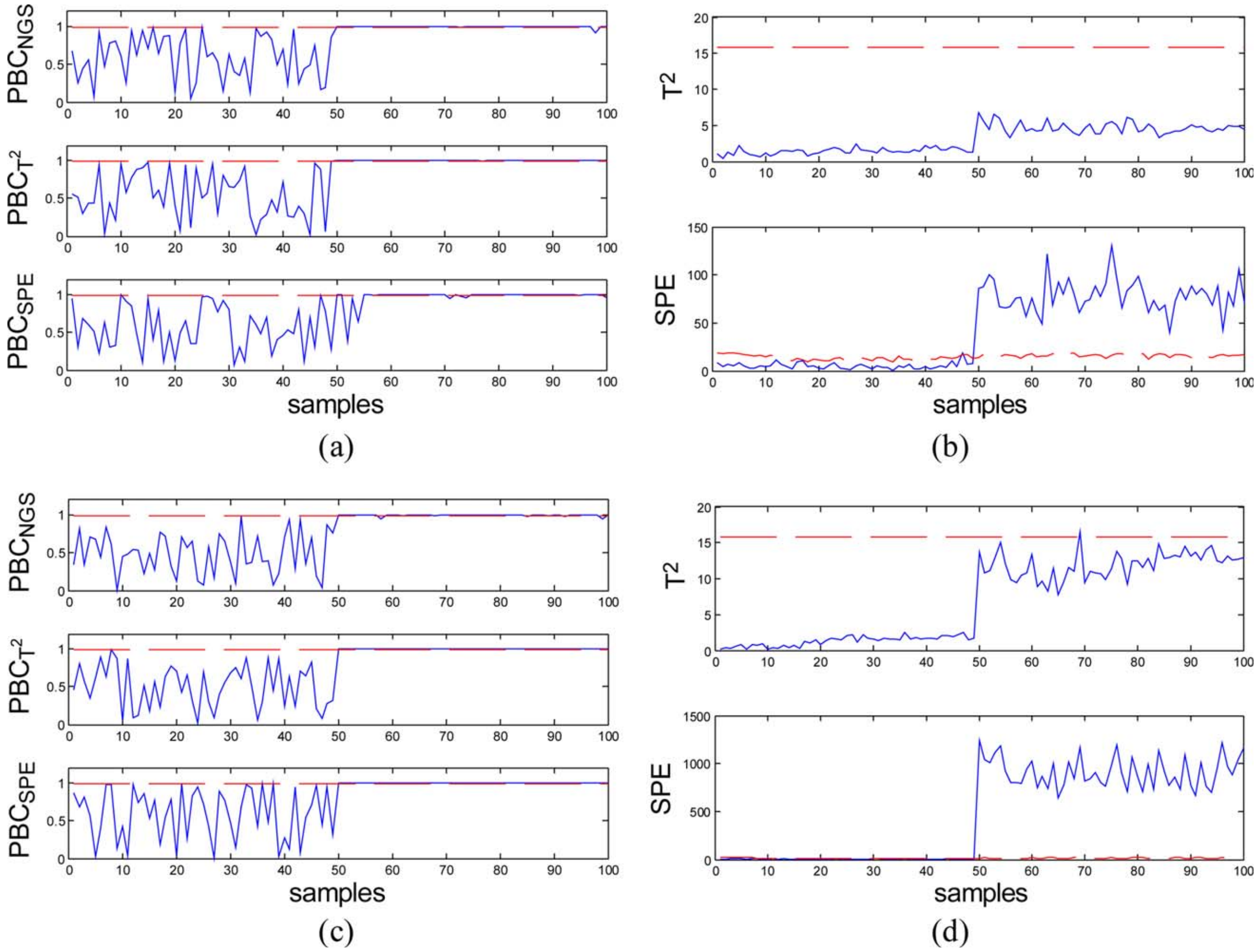

Figure 5. Monitoring results of the first two faulty batches, (a) PBC for first faulty batch; (b) MPCA for first faulty batch; (c) PBC for second faulty batch; and (d) MPCA for second faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and MPCA are given in Figure 5a, b. It is very clear that all monitoring statistics of the PBC method can successfully detect this fault. Although there is a little detection decay by the $\mathrm{PBC}_{\mathrm{SPE}}$ statistic, both of the $\mathrm{PBC}_{\mathrm{NGS}}$ and PBC_{T^2} statistics can detect the fault immediately when it happened. However, only the SPE statistic of the MPCA method can

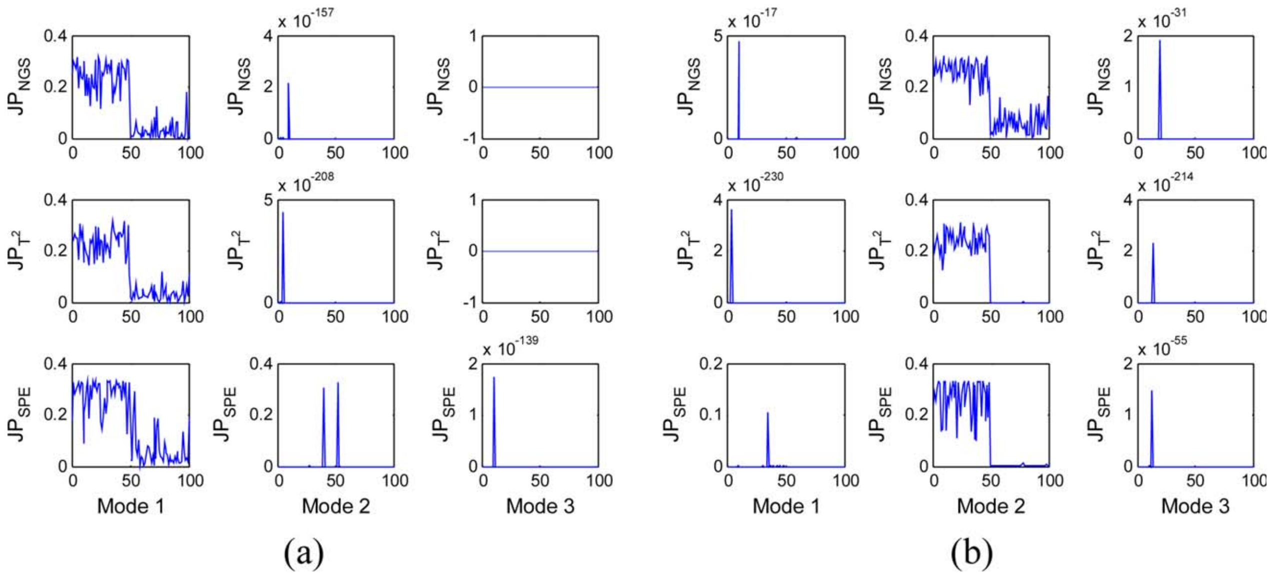

Figure 6. Joint probability analysis results, (a) first faulty batch and (b) second faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

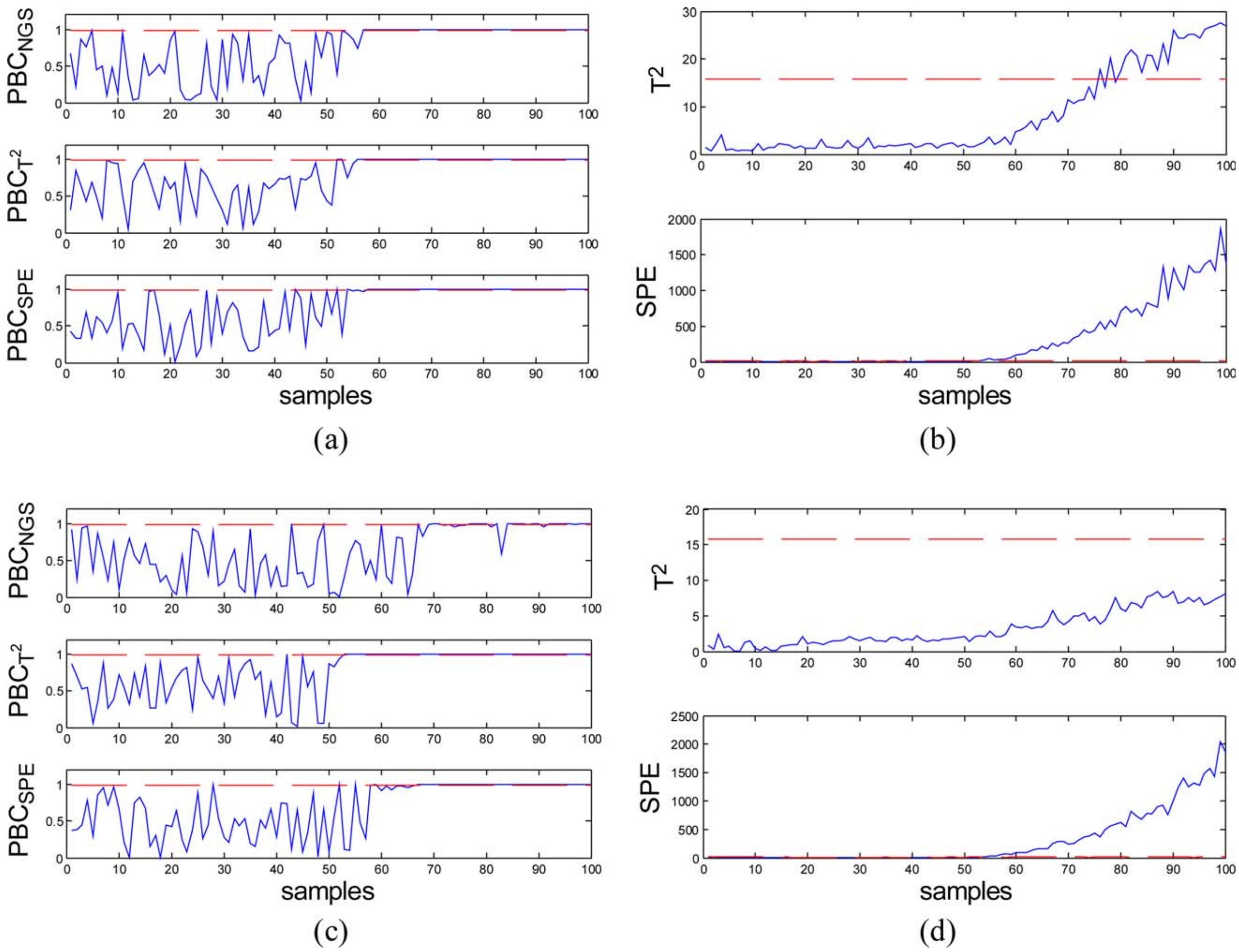

Figure 7. Monitoring results of the third and fourth faulty batches, (a) PBC for third faulty batch; (b) MPCA for third faulty batch; (c) PBC for fourth faulty batch; and (d) MPCA for fourth faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

detect this fault, which can be seen in Figure 5b. Therefore, comparing the monitoring results of these two methods, one can infer that the PBC is more reliable. By examining the joint probability analysis results in Figure 6a, it can be inferred that this batch is operated under the first operating condition until some fault has been introduced at 200 h to

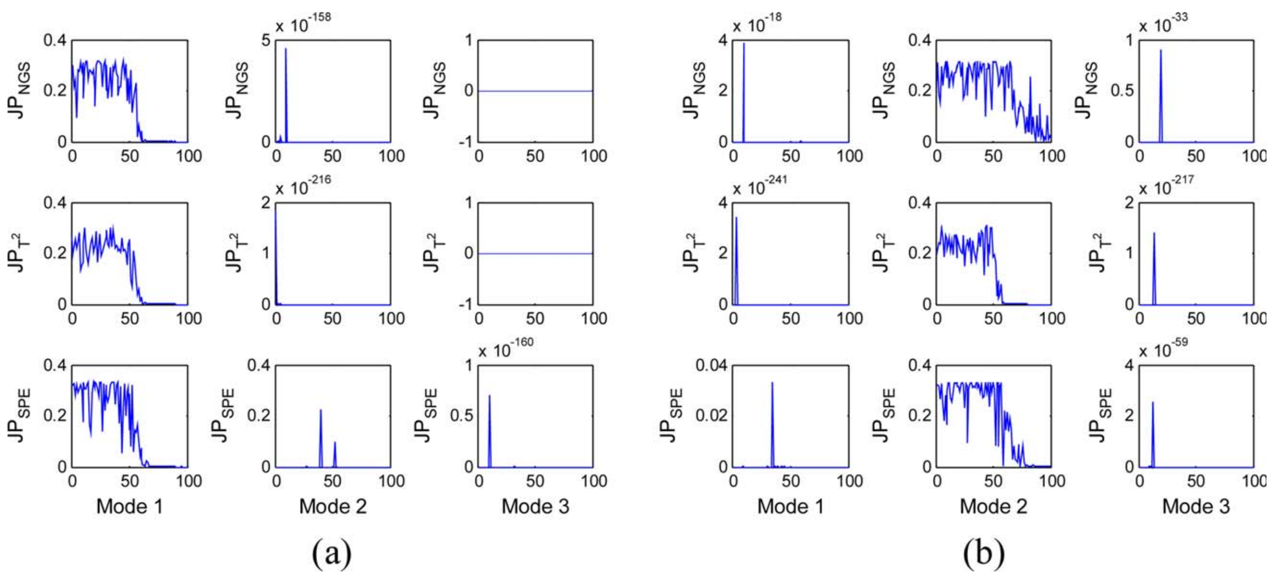

Figure 8. Joint probability analysis results, (a) third faulty batch and (b) fourth faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

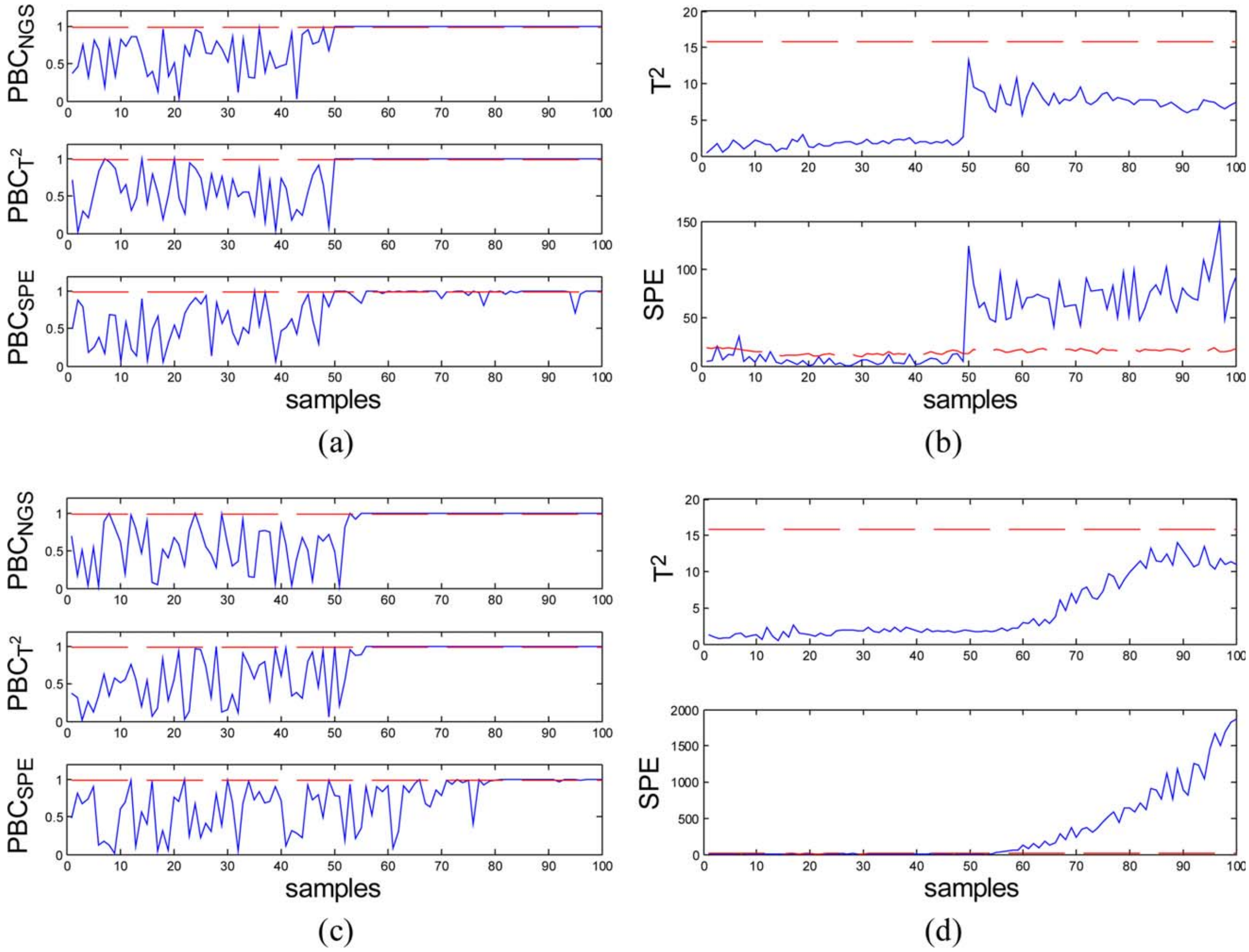

Figure 9. Monitoring results of the last two faulty batches, (a) PBC for fifth faulty batch; (b) MPCA for fifth faulty batch; (c) PBC for sixth faulty batch; and (d) MPCA for sixth faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

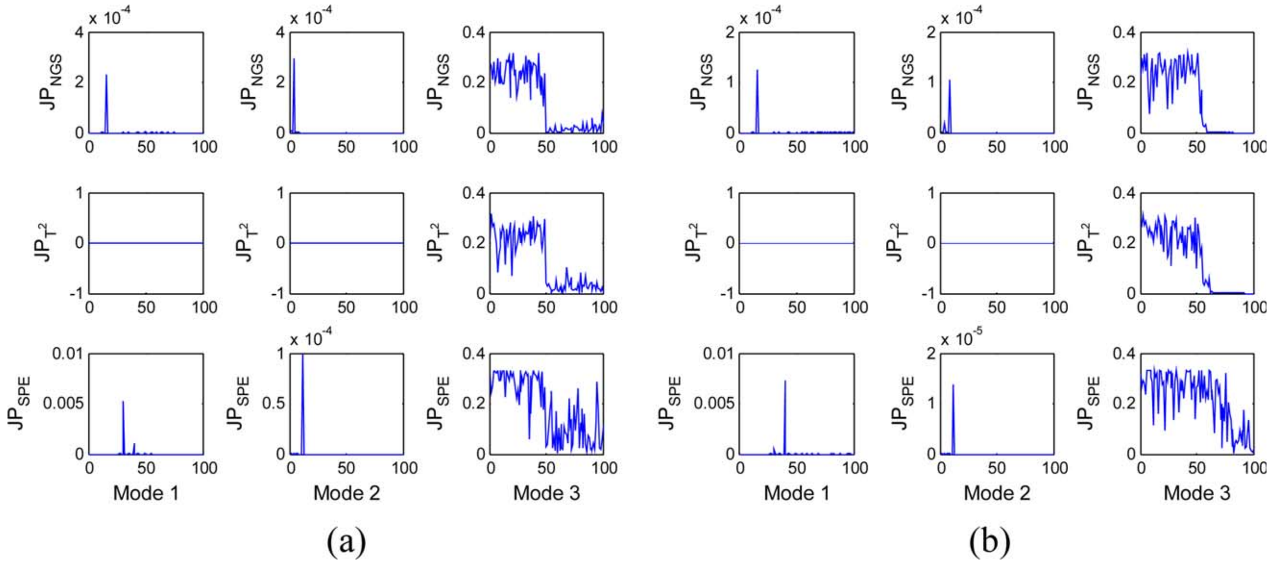

Figure 10. Joint probability results of the last two faulty batches in different operation modes, (a) fifth faulty batch; (b) sixth faulty batch.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the process. Second, a similar step change of the aeration rate is introduced under the second operation mode at 200 h. Similar results can be obtained. The superiority of the PBC method to the conventional MPCA method is illustrated in Figures 5c, d, with its corresponding mode identification result given in Figure 6b. Next, two ramp faults are introduced to the process, which correspond to Faults 3 and 4 in Table 2. Monitoring and mode identification results of these two ramp faults are demonstrated in Figures 7a–d and 8a, b, respectively. Due to the slow change characteristic of these two process faults, they cannot be detected in the first several samples by both PBC and MPCA methods. However, the detection delay of MPCA for the third fault is much larger than that of PBC. Besides, the fourth fault cannot be detected by the T^2 statistic of MPCA, whereas it can be successfully detected by all of the three statistics of PBC, which again shows that PBC is more reliable than MPCA. In addition, two more faulty batches which are initially operated under the third mode are tested, the results of which are given in Figures 9 and 10. One can find that good results have been obtained by PBC in both monitoring and mode identification aspects.

According to the modeling principal of PBC, it can be easily extended for fault diagnosis and identification. Thus, if we have obtained enough faulty batch data, a fault pool can be easily constructed for fault identification. Depending on the results of this example, it can be inferred that the proposed PBC method is more reliable for process monitoring than the MPCA method. Another important issue which should be concerned is how to differentiate the new operation mode and the new process fault. Although the joint probability can detect the operation mode that happened in the process, it cannot tell us whether this new mode is normal process change or not. Generally speaking, this issue is difficult to be addressed without appropriate process knowledge. Any endeavor which is carried out to differentiate these two kinds of process changes automatically is interesting, and therefore should be put significant efforts in the future work.

Conclusions

A phase-based Bayesian combination method has been proposed for monitoring multiphase batch processes with the multimode behavior in this article. The contributions of this study are summarized as follows. First, a Bayesian monitoring method has been proposed, which differentiates the existing works in monitoring batch processes with multimode behaviors. Second, an efficient mode identification method has been developed to locate the operation mode of the monitored batch. A successful application study of the penicillin fermentation benchmark process has been demonstrated, and several important issues have also been illustrated and discussed. In conclusion, the probabilistic implementation of the method has greatly improved the monitoring performance and comprehension of the batch process.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) (61004134), National Project 973 (2012CB720500), and the Fundamental Research Funds for the Central Universities (2013QNA5016).

Literature Cited

1. Ge ZQ, Song ZH, Gao FR. Review of recent research on data-based process monitoring. *Ind Eng Chem Res.* 2013;52:3543–3562.
2. Kano M, Tanaka S, Hasebe S, Hashimoto I, Ohno H. Monitoring independent components for fault detection. *AIChE J.* 2003;49:969–976.
3. Lee JM, Yoo CK, Lee IB. Statistical process monitoring with independent component analysis. *J Process Control* 2004;14:467–485.
4. Ge ZQ, Yang CJ, Song ZH. Improved kernel PCA-based monitoring approach for nonlinear processes. *Chem Eng Sci.* 2009;64:2245–2255.
5. Wang X, Kruger U, Irwin GW, McCullough G, McDowell N. Nonlinear PCA with the local approach for diesel engine fault detection and diagnosis. *IEEE T Contr Syst Tech.* 2007;16:122–129.
6. Ge ZQ, Song ZH. Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Contr Eng Pract.* 2008;16:1427–1437.
7. Wang J, He QP. Multivariate statistical process monitoring based on statistics pattern analysis. *Ind Eng Chem Res.* 2010;49:7858–7869.
8. Ge ZQ, Song ZH. Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE J.* 2010;56:2838–2849.
9. Yu J. Localized fisher discriminant analysis based complex chemical process monitoring. *AIChE J.* 2011;57:1817–1828.
10. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J.* 1994;44:1361–1375.
11. Nomikos P, MacGregor JF. Multiway partial least square in monitoring batch processes. *Chem Intel Lab Syst.* 1995;30:97–108.
12. Yoo CK, Lee JM, Vanrolleghem PA, Lee IB. On-line monitoring of batch processes using multiway independent component analysis. *Chem Intel Lab Syst.* 2004;71:151–163.
13. Chen JH, Chen HH. On-line batch process monitoring using MHMT-based MPCA. *Chem Eng Sci.* 2006;61:3223–3239.
14. Zhang YW, Qin SJ. Fault detection of nonlinear processes using multiway kernel independent analysis. *Ind Eng Chem Res.* 2007;46:7780–7787.
15. Choi SW, Morris AJ, Lee IB. Dynamic model-based batch process monitoring. *Chem Eng Sci.* 2008;63:622–636.
16. Ge ZQ, Song ZH. Semiconductor manufacturing process monitoring based on adaptive substatistical PCA. *IEEE Semicond Manuf.* 2010;23:99–108.
17. Wang D. Robust data-driven modeling approach for real-time final product quality prediction in batch process operation. *IEEE T Ind Inform.* 2011;7:371–377.
18. Yu JB. Fault detection using principal components-based Gaussian mixture model for semiconductor manufacturing processes. *IEEE T Semiconduct Manuf.* 2011;24:432–444.
19. Undey C, Cinar A. Statistical monitoring of multistage, multiphase batch processes. *IEEE Contr Syst Mag.* 2002;22:40–52.
20. Muthuswamy K, Srinivasan R. Phase-based supervisory control for fermentation process development. *J Process Control* 2003;13:367–382.
21. Lu NY, Gao FR, Wang FL. A sub-PCA modeling and online monitoring strategy for batch processes. *AIChE J.* 2004;50:255–259.
22. Liu J, Wong DSH. Fault detection and classification for a two-stage batch process. *J Chemom.* 2008;22:385–398.
23. Doan XT, Srinivasan R. Online monitoring of multi-phase batch processes using phase-based multivariate statistical process control. *Comput Chem Eng.* 2008;32:230–243.
24. Camacho J, Pico J, Ferrer A. multiphase analysis framework for handling batch process data. *J Chemom.* 2008;22:632–643.
25. Ge ZQ, Zhao LP, Yao Y, Song ZH, Gao FR. Utilizing transition information in online quality prediction of multiphase batch processes. *J Process Control* 2012;22:599–611.
26. Yoo CK, Villez K, Lee IB, Rosén C, Vanrolleghem PA. Multimodel statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotech Bioeng.* 2007;96:687–701.
27. Zhao CH, Wang FL, Gao FR, Zhang YW. Enhanced process comprehension and statistical analysis for slow-varying batch processes. *Ind Eng Chem Res.* 2008;47:9996–10008.
28. Sebzalli YM, Wang XZ. Knowledge discovery from process operational data using PCA and fuzzy clustering. *Eng App Artif Intel.* 2001;14:607–616.
29. Chen WC, Wang MS. A fuzzy c-means clustering-based fragile watermarking scheme for image authentication. *Exp Syst App.* 2009;36:1300–1307.
30. Bicego M, Figueiredo MAT. Soft clustering using weighted one-class support vector machines. *Pattern Recognit.* 2009;42:27–32.

31. Ge ZQ, Song ZH. Process monitoring based on independent component analysis-principal component analysis (ICA-PCA) and similarity factors. *Ind Eng Chem Res.* 2007;46:2054–2063.
32. Valle S, Li W, Qin SJ. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind Eng Chem Res.* 1999;38:4389–4410.
33. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Network* 2000;13:411–430.
34. Tax DMJ, Duin RPW. Support vector data description. *Mach Learn.* 2004;54:45–66.
35. Ge ZQ, Xie L, Song ZH. A novel statistical-based monitoring approach for complex multivariate processes. *Ind Eng Chem Res.* 2009;48:4892–4898.
36. Ge ZQ, Song ZH. Multimode process monitoring based on Bayesian method. *J Chemom.* 2009;23:636–650.
37. Birol G, Undey C, Cinar A. A modular simulation package for fed-batch fermentation: Penicillin production. *Comput Chem Eng.* 2002;26:1553–1561.

Manuscript received Aug. 7, 2012, and revision received Mar. 16, 2013.