Multivariate Statistical Monitoring of Multiphase Batch Processes With Between-Phase Transitions and Uneven Operation Durations

Yuan Yao,[1]* Weiwei Dong,[2] Luping Zhao[3] and Furong Gao[2,3]

1. *Department of Chemical Engineering, National Tsing Hua University, Hsinchu, Taiwan*
2. *Center for Polymer Processing and Systems, The Hong Kong University of Science and Technology, Nansha, Guangzhou, China*
3. *Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

In order to achieve satisfactory monitoring, multivariate statistical process models should well reflect process nature. In manufacturing systems, many batch processes are inherently multiphase. Usually, different phases have different characteristics, while gradual transitions are often observed between phases. Another important feature of batch processes is the unevenness of operation durations. Especially, in multiphase batch processes, the situation becomes more complicated. In this study, a batch process modelling and monitoring strategy is proposed based on Gaussian mixture model (GMM), which can automatically extract phase and transition information for uneven-duration batch processes. The application results verify the effectiveness of the proposed method.

Keywords: multivariate statistical process control, fault detection, Gaussian mixture model, multiphase, uneven operation duration

INTRODUCTION

Due to process complexity and high dimensionality, it is difficult to build batch process models based on first principles or process knowledge within limited product-to-market time. Therefore, multivariate statistical process control (MSPC) methods, such as principal component analysis (PCA); (Jackson, 1991; Jolliffe, 2002), which require only process historical data, have attracted great attention in the field of batch process online monitoring. Conventional MSPC methods, for example, multiway PCA (MPCA); (Nomikos and MacGregor, 1994, 1995), take the entire batch data as a single object in modelling, while many industrial batch processes are multiphase in nature (Undey and Cinar, 2002; Yao and Gao, 2009b). Such batch processes are usually carried out in a sequence of steps, which lead to changes in process characteristics along operation time within each batch run. In order to reflect the unique characteristics in each phase, multiphase modelling and monitoring strategies should be applied.

A typical method for multiphase batch process monitoring is the phase-based sub-PCA developed by Lu et al. (2004a), which conducts phase division based on the changes in the variable correlation information along the operation time. In sub-PCA, each phase is separately modelled for the purpose of better monitoring. Later, this method was extended to handle gradual transitions between neighbouring phases (Zhao et al., 2007; Yao and Gao, 2009a). Fuzzy membership functions are utilised to model the transition behaviours. However, user-specified threshold parameters are required in both the algorithm of sub-PCA and its extension. As pointed out by Yu and Qin (2009), the lack of

*Author to whom correspondence may be addressed.
E-mail address: yyao@mx.nthu.edu.tw
Can. J. Chem. Eng. 90:1383–1392, 2012

DOI 10.1002/cjce.21617
Published online 19 December 2011 in Wiley Online Library (wileyonlinelibrary.com).

an objective rule in threshold setting may degrade the monitoring efficiency. Besides, multiphase PCA (MPPCA); (Camacho and Picó, 2006) provides an alternative solution for automatic phase identification and modelling. However, the transition issues are not considered.

Uneven-duration is another important feature of many batch processes. Most MSPC methods for batch process monitoring are based on an assumption that all batch runs have the same durations. However, in real systems, the batch lengths are usually not fixed because of disturbances and changes in operating conditions. In such situations, different types of trajectory synchronisation methods are often applied (Nomikos and MacGregor, 1995; Kassidas et al., 1998). But these methods may distort the correlation information and affect the efficiency of online monitoring. Meanwhile, in multiphase processes, not only the whole batch durations but also the phase durations can be different from run to run. The trajectory synchronisation methods do not consider the multiphase features. To avoid the drawbacks of trajectory synchronisation, sub-PCA method was revised to solve the uneven-duration problem (Lu et al., 2004b; Zhao et al., 2011). However, gradual transitions between phases are not identified in those research studies.

Recently, Gaussian mixture model (GMM) has been adopted in the research area of process monitoring, owing to its solid theoretical foundation and good practical performance (Choi et al., 2004; Thissen et al., 2005; Chen and Zhang, 2009). In order to deal with multiphase batch process data, Yu and Qin (2009) proposed a multiway GMM method. In their method, the batch process data are first unfolded and normalised in a batch-wise way. After that, the normalised data are re-arranged into the variable-wise format. GMM is then performed to identify the operation phases. In each phase, a local probability index is calculated for the purpose of fault detection. No subjective parameter setting is needed in the entire procedure. However, Yu and Qin's method can handle neither transitions nor uneven batch durations, thereby limiting its applications.

In this study, a new multivariate statistical process analysis, modelling and monitoring strategy is developed for multiphase batch processes. Using a two-step data normalisation method, GMM is estimated for phase division and transition identification; then, local monitoring models are built for different phases and transition periods. Comparing to the existing methods, the proposed method provides better phase division results, and is able to address the issues of transitions and uneven batch durations simultaneously.

The rest of the study is organised as following. In Problem Statement Section, the uneven-duration problem is stated for multiphase batch processes. Subsequently, the proposed strategy is presented in detail, including the fundamentals of GMM, data normalisation, phase division and transition identification, local model building and monitoring procedure. In Application Results Section, the effectiveness of the strategy is verified using the data from a benchmark penicillin fermentation process and an injection moulding process. Finally, the study is summarised with the conclusions drawn in Conclusions Section.

PROBLEM STATEMENT

In industrial processing, batch processes are not strictly repetitive. Even under the same operation conditions, disturbances may lead to uneven operation durations from batch to batch. The uneven-duration problem is of two aspects. First, the lengths of different batch runs are often unequal. Consequently, the measurement data cannot be arranged into a regular matrix, making most conventional multivariate statistical modelling methods invalid. Second, a process may reach different maturities at the same sampling interval in different rums. Especially, in multiphase batch processes, not only the total batch durations but also the lengths of each phase may vary from batch to batch.

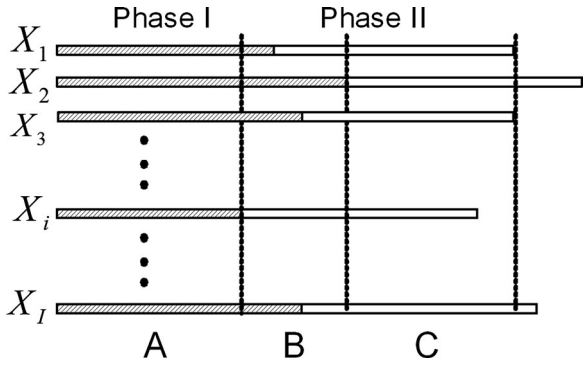

Figure 1. Illustration of an uneven-duration batch process.

Figure 1 illustrates the uneven-duration problem with a two-phase batch process. In time span A, all batches operate in the first phase. In time span B, the situation is more complicated, since some batches are still in phase I, while other batches have switched to phase II. In time span C, all batches run in phase II, and irregular tails can be observed since the uneven lengths of different cycles. It is also important to notice that, although the first batch and the third batch have the same total cycle duration, they have different lengths of both phases I and II. Such an example implies that uneven-duration problem may still exist, even if the lengths of the total batch durations seem to be equal.

It is easy to understand that the situations become more complicated when gradual transitions exist between phases.

GMM-BASED MULTIPHASE MONITORING STRATEGY

As have discussed above, batch processes in real industrial are often very complicated. The existence of multiple phases, transitions and uneven durations makes it a difficult task to model and monitor batch processes reasonably. To address these problems simultaneously, a GMM-based monitoring strategy is proposed in this section.

Fundamentals of GMM

GMM is a general tool for the estimation of probability density function (*pdf*), which has been widely utilised in the area of data modelling and clustering (McLachlan and Peel, 2000; Fraley and Raftery, 2002). The basic idea of GMM is that an arbitrary probability density can be approximated with a mixture of several component Gaussian density functions. The *pdf* estimated using GMM is formulated as:

$$p(\mathbf{x}^{\mathrm{T}}|\theta) = \sum_{c=1}^{C} \alpha_c p(\mathbf{x}^{\mathrm{T}}|c) \tag{1}$$

$$p(\mathbf{x}^{\mathrm{T}}|c) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}\Sigma_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{(2\pi)^{\frac{m}{2}}|\Sigma_c|^{\frac{1}{2}}} \tag{2}$$

where $\mathbf{x}^{\mathrm{T}}$ is a sample vector, m denotes the dimension of the data, $\theta = \{\alpha_c, \boldsymbol{\mu}_c^{\mathrm{T}}, \Sigma_c;\ c = 1, 2, \cdots, C\}$ is a parameter vector, C is the total

number of the mixture components, $\boldsymbol{\mu}_c^{\mathrm{T}}$ represents the mean vector of the cth Gaussian density function, Σ_c is the corresponding covariance matrix, $p(\mathbf{x}^{\mathrm{T}}|c)$ is the cth component density of $\mathbf{x}^{\mathrm{T}}$, α_c is the prior probability of the sample belonging to the cth mixture component, satisfying $\sum_{c=1}^{C} \alpha_c = 1$ and $0 \le \alpha_c \le 1$.

The model parameters θ can be estimated using the expectation-maximisation (EM) algorithm (Dempster et al., 1977) through iteratively maximising the likelihood function. Given a set of training data, the likelihood function to be maximised is:

$$L(\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n^{\mathrm{T}}|\theta) \tag{3}$$

where N is the total number of the samples in the training set. In each iteration run, the posterior probability $p(c|\mathbf{x}_n^{\mathrm{T}})$ is calculated for each data point, indicating the probability that an object obeys a particular mixture component.

$$p(c|\mathbf{x}_n^{\mathrm{T}}) = \frac{\alpha_c p(\mathbf{x}_n^{\mathrm{T}}|c)}{\sum_{v=1}^{C} \alpha_v p(\mathbf{x}_n^{\mathrm{T}}|v)} \tag{4}$$

The above equation is known as the expectation step (E-step), followed by the maximisation step (M-step) as below:

$$\alpha_c = \frac{1}{N} \sum_{n=1}^{N} p(c|\mathbf{x}_n^{\mathrm{T}}) \tag{5}$$

$$\boldsymbol{\mu}_c^{\mathrm{T}} = \frac{\sum_{n=1}^{N} p(c|\mathbf{x}_n^{\mathrm{T}}) \mathbf{x}_n^{\mathrm{T}}}{\sum_{n=1}^{N} p(c|\mathbf{x}_n^{\mathrm{T}})} \tag{6}$$

$$\boldsymbol{\Sigma}_c = \frac{\sum_{n=1}^{N} p(c|\mathbf{x}_n^{\mathrm{T}}) (\mathbf{x}_n^{\mathrm{T}} - \boldsymbol{\mu}_c^{\mathrm{T}}) (\mathbf{x}_n^{\mathrm{T}} - \boldsymbol{\mu}_c^{\mathrm{T}})^{\mathrm{T}}}{\sum_{n=1}^{N} p(c|\mathbf{x}_n^{\mathrm{T}})} \tag{7}$$

The E- and M-steps are repeated iteratively, until the calculations converge to a stable solution which maximises the likelihood described in Equation (3).

A limitation of the basic EM algorithm is that it cannot automatically adjust the number of Gaussian mixture component during the parameter estimation. To solve this problem, an unsupervised learning algorithm can be adopted. For more detailed information about the unsupervised learning algorithm (please refer to Figueiredo and Jain, 2002).

Data Normalisation

Before multivariate process modelling, data normalisation, eliminating the effects of variable units and measuring ranges, is an important step for data pretreatment. The batch-wise normalisation (Nomikos and MacGregor, 1994) is most commonly used in the pretreatment of batch process data, which removes the average trajectories from each batch and scales the variables at each sampling interval to unit variance by dividing them with their standard deviations. By doing so, the batch-to-batch variation is emphasised. Such kind of normalisation is very much suitable for process monitoring. However, the conventional batch-wise normalisation was designed for batch processes with even operation durations. As illustrated in Problem Statement Section, in multiphase batch processes, the lengths of each phase may vary from batch to batch, even if the total batch durations are same. Under that situation, measurements on the same sampling interval in different batches may belong to different phases, making the conventional batch-wise normalisation meaningless. Therefore, reasonable phase division is necessary before performing batch-wise normalisation.

An alternative choice is the variable-wise normalisation (Wold et al., 1998), which subtracts the trajectory average of each variable over the entire batch durations in all batches, and divides the value of each variable by the overall standard deviation. The variable-wise normalisation is a ''double-edged sword.'' On one hand, the monitoring efficiency based on this kind of pretreatment is limited, since it highlights the variation along the time direction instead of the batch direction. On the other hand, the uneven-duration problem is spontaneously solved with such type of normalisation. Moreover, as indicated by Lu et al. (2003): for a multiphase batch process, different process variables dominate different phases; for phase division, it is desirable to scale process variables within a batch run to retain the inherent weights in different phases.

Based on the above discussions, a two-step normalisation is conducted in this study. First, variable-wise normalisation is used for the data pretreatment for phase division and transition identification. The normalised data form a two-way matrix $\hat{\mathbf{X}}(N \times J)$, where N is the total number of samples from all the batches in the referenced data set, J is the number of variables. Each row of this matrix is a row vector $\hat{\mathbf{x}}_n^{\mathrm{T}}$ containing a variable-wise normalised sample, where $n = 1, 2, \ldots, N$. The second normalisation is performed after phase division and transition identification, for the task of online process monitoring. In the following, ''a period'' is defined as ''either a phase or a transition period between two neighbouring phases.'' The raw measurement data in each period are re-normalised separately in the batch-wise way. The average trajectories and standard deviations in the irregular tails of uneven-duration phases or transition periods can be calculated based on the available data, using the method introduced by Lu et al. (2004b). Then, the normalised data in each period are reorganised into a two-way matrix $\tilde{\mathbf{X}}^d(N^d \times J)$, where d is the index of period and N^d is the total number of samples belonging to the dth period. Each row of $\tilde{\mathbf{X}}^d$ is a re-normalised sample $\tilde{\mathbf{x}}_n^{\mathrm{T}}$, where $n = 1, 2, \ldots, N^d$.

Phase Division and Transition Identification

In multiphase batch processes, process characteristics may change with phases, which can be indicated by the changes in the patterns of data distribution. In another word, the data belonging to the same phase often can be approximated by the same multivariate Gaussian distribution, while the data from different phases may distribute differently. During the transitions from phase to phase, process characteristics are usually similar to the previous phase at the beginning of transition periods; and then, transit to the next phase gradually. Hence, the probability densities of the data distributions during gradual transitions can be regarded as the mixtures of the Gaussian density functions in the two neighbouring phases. These findings inspire the basic idea in this research.

As introduced in Fundamentals of GMM Section, GMM not only provides estimations of Gaussian densities, but also approximates non-Gaussian densities with weighted sums of two or more Gaussian density functions. Meanwhile, GMM is also able to

provide suggestions on data partitions, using the posterior probability information. Such properties make GMM a favourable method in phase division, transition identification and multiphase batch process modelling and monitoring. The detailed procedure of phase division and transition identification is like follows.

After variable-wise normalisation, a GMM called phase division model can be estimated from $\hat{\mathbf{X}}$, using the EM algorithm or the unsupervised learning algorithm as introduced in Fundamentals of GMM Section. In such a mixture model, each Gaussian component corresponds to a phase in batch operation. The membership degree of each sample belonging to the cth component is reflected by the value of the posterior probability $p(c|\hat{\mathbf{x}}_n^{\mathrm{T}})$ calculated in Equation (4). To identify the phase boundaries, a confidence bound ϕ (e.g. 95% or 99%) is specified. If $p(c|\hat{\mathbf{x}}_n^{\mathrm{T}}) \geq \phi$, then the nth sample statistically belongs to the cth phase, where $c = 1, 2, \ldots, C$. Otherwise, if there is no phase solely dominating a sample, that is, $p(c|\hat{\mathbf{x}}_n^{\mathrm{T}}) < \phi$ for all Gaussian components, the sample is in a transition period between two phases. Practically, the number of phases (Gaussian components) is often known in advance based on process knowledge. Otherwise, if such knowledge is unavailable, the unsupervised learning algorithm can be adopted to automatically determine the optimal component number.

One point to emphasise is that the determination of a GMM is difficult or even impossible when the data are high-dimensional and collinear. Under that situation, PCA technique (Jackson, 1991; Jolliffe, 2002) is suggested to be employed before performing GMM, for dimension reduction and data orthogonalisation:

$$\hat{\mathbf{X}} = \hat{\mathbf{T}}\hat{P}^{\mathrm{T}} + \hat{\mathbf{E}} \tag{8}$$

where $\hat{\mathbf{X}}$ is the variable-wise normalised data matrix, $\hat{\mathbf{T}}$ is the PCA score matrix, $\hat{\mathbf{P}}$ is the loading matrix and $\hat{\mathbf{E}}$ is the residual matrix. Consequently, the GMM is estimated using the PCA scores stored in $\hat{\mathbf{T}}$, instead of the normalised measurements $\hat{\mathbf{x}}_n^{\mathrm{T}} (n = 1, 2, \cdots, N)$.

One may have noticed that, if the phase division model is built based on the measurement data instead of PCA scores, the normalisation step is not necessary. The GMM algorithm has already contained a hidden step of normalisation by removing the means and equalising the variances, as shown in Equation (2). However, if the PCA scores are to be utilised, the variable-wise normalisation is important, which ensures the PCA model to extract correct systematic information.

Phase and Transition Modelling

After phase division and transition identification, the process data are re-normalised as described in Data Normalisation Section, in order to highlight the variations between cycles which are more significant for process monitoring. According to Undey and Cinar (2002), local models are advantageous for monitoring, when different phases exist. Therefore, each phase/transition period is modelled separately. The idea of GMM is modified for phase and transition modelling.

The modelling procedure follows three major steps: phase modelling, transition modelling and control limits derivation.

According to the phase division, the measurement data in each phase should follow similar distribution pattern. Therefore, it is not necessary to describe the phase probability density with a mixture of several Gaussian density functions. A more reasonable way is to model each phase with a single component density function corresponding to that phase. Based on such an idea, the phase model can be derived as following. With a batch-wise normalised phase data matrix $\tilde{\mathbf{X}}^d$, the corresponding model parameters, including the mean vector and the sample covariance matrix, are calculated as:

$$\tilde{\boldsymbol{\mu}}_d^{\mathrm{T}} = \frac{\sum_{n=1}^{N^d} \tilde{\mathbf{x}}_{n^d}^{\mathrm{T}}}{N^d} \tag{9}$$

$$\tilde{\Sigma}_d = \frac{\sum_{n=1}^{N^d} (\tilde{\mathbf{x}}_{n^d}^{\mathrm{T}} - \tilde{\boldsymbol{\mu}}_d^{\mathrm{T}})(\tilde{\mathbf{x}}_{n^d}^{\mathrm{T}} - \tilde{\boldsymbol{\mu}}_d^{\mathrm{T}})^{\mathrm{T}}}{N^d - 1} \tag{10}$$

The probability densities $p(\tilde{\mathbf{x}}_{n^d}^{\mathrm{T}}|d)$ of the training data are then derived using Equation (2).

Then, the transition periods are modelled. A transition period is a connection of two phases neighbouring to it. Consequently, the probability densities in a transition period should be described as the weighted sums of only two-phase Gaussian density functions, while the information of other phases is not necessary to be introduced into the transition model. Without losing generality, assume the gth period is a transition period. Therefore, the $(g-1)$ and $(g+1)$th periods are two operation phases before and after this transition period. The formula of transition probability densities $p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g)$ are:

$$p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g) = \alpha_{g-1}^g p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g-1) + \alpha_{g+1}^g p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g+1) \tag{11}$$

where n^g is the sample index in the transition period, $p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g-1)$ and $p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|g+1)$ are estimated from Equation (2), using the mean vectors $\tilde{\boldsymbol{\mu}}_{g-1}^{\mathrm{T}}$, $\tilde{\boldsymbol{\mu}}_{g+1}^{\mathrm{T}}$ and the covariance matrices $\tilde{\Sigma}_{g-1}$, $\tilde{\Sigma}_{g+1}$ derived during the phase modelling step. The prior probability parameters α_{g-1}^g and α_{g+1}^g are achieved using the iterative EM algorithm:

$$p(h|\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}) = \frac{\alpha_h^g p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|h)}{\sum_{v \in h} \alpha_v^g p(\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}|v)} \tag{12}$$

$$\alpha_h^g = \frac{1}{N^g} \sum_{n=1}^{N^g} p(h|\tilde{\mathbf{x}}_{n^g}^{\mathrm{T}}) \tag{13}$$

where $h = g-1$ or $g+1$ and N^g is the total number of samples belonging to the transition period. Different from the EM algorithm for the standard GMM estimation, it is not necessary to update the mean vectors and the covariance matrices in the iteration runs.

Similar to the situation in phase division and transition identification, the high-dimensionality and collinearity of the data may affect the GMM estimation. Therefore, again, PCA can be performed to solve the problem. In phase modelling, PCA is conducted on each phase data matrix $\tilde{\mathbf{X}}^d$. Then, the parameters of the phase models are calculated using the PCA scores instead of $\tilde{\mathbf{x}}_{n^d}^{\mathrm{T}}$. In transition modelling, PCA is applied to the augmented data matrix $\tilde{\mathbf{X}}_{\mathrm{aug}}$:

$$\tilde{\mathbf{X}}_{\mathrm{aug}} = \begin{bmatrix} \tilde{\mathbf{X}}^{g-1} \\ \tilde{\mathbf{X}}^{g} \\ \tilde{\mathbf{X}}^{g+1} \end{bmatrix} = \tilde{\mathbf{T}}\tilde{P}^{\mathrm{T}} + \tilde{\mathbf{E}} \tag{14}$$

where $\tilde{\mathbf{X}}^g$ is the batch-wise normalised data matrix of a transition period; $\tilde{\mathbf{X}}^{g-1}$ and $\tilde{\mathbf{X}}^{g+1}$ are data matrices of the phases neighbouring to the transition period; $\tilde{\mathbf{T}}$, $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{E}}$ are the PCA score matrix,

loading matrix and residual matrix, respectively. The score matrix $\tilde{\mathbf{T}}$ can be represented as:

$$\tilde{\mathbf{T}} = \begin{bmatrix} \tilde{\mathbf{T}}^{g-1} \\ \tilde{\mathbf{T}}^{g} \\ \tilde{\mathbf{T}}^{g+1} \end{bmatrix} \tag{15}$$

where $\tilde{\mathbf{T}}^{g-1}$, $\tilde{\mathbf{T}}^{g}$ and $\tilde{\mathbf{T}}^{g+1}$ correspond to $\tilde{\mathbf{X}}^{g-1}$, $\tilde{\mathbf{X}}^{g}$ and $\tilde{\mathbf{X}}^{g+1}$, respectively. Then, $\tilde{\boldsymbol{\mu}}_{g-1}^{\mathrm{T}}$, $\tilde{\Sigma}_{g-1}$ and $\tilde{\boldsymbol{\mu}}_{g+1}^{\mathrm{T}}$, $\tilde{\Sigma}_{g+1}$ are calculated as the mean vectors and the sample covariance matrices of $\tilde{\mathbf{T}}^{g-1}$ and $\tilde{\mathbf{T}}^{g+1}$; and the parameters of the transition model are derived using Equations (10)–(12), where the score vectors in $\tilde{\mathbf{T}}^{g}$ are utilised instead of the batch-wise normalised data $\tilde{\mathbf{x}}_{ng}^{\mathrm{T}}$.

After estimating the *pdf*, thresholds should be defined for each phase and each transition period, in order to perform online monitoring. If a $100\beta\%$ control limit is to be set, the threshold h should satisfy Equation (16) (Chen et al., 2006):

$$\int_{\tilde{\mathbf{x}}^{\mathrm{T}}:p(\tilde{\mathbf{x}}^{\mathrm{T}}|d)>h} p(\tilde{\mathbf{x}}^{\mathrm{T}}|d)d\tilde{\mathbf{x}}^{\mathrm{T}} = \beta \tag{16}$$

In order to achieve efficient monitoring, it is better to set different thresholds $h_{k^d}^d$ for different sampling intervals in each phase or transition period, where the symbol d in Equation (16) is the index of period and k^d is the sample index in this period. Therefore, the likelihoods of the normal operation data over all the batches at the k^dth sampling interval in the dth period are calculated. When the number of samples is large, $h_{k^d}^d$ can be directly identified, which is less than the likelihood of $100\beta\%$ of the nominal data (Thissen et al., 2005). Otherwise, when the amount of the normal operation data is limited, numerical Monte Carlo (MC) simulations can be utilised to approximate the integral in Equation (16) (Chen et al., 2006; Chen and Zhang, 2009). More specifically, utilising the algorithm introduced by Bishop (2006), random samples can be generated from a GMM based on MC simulation. These samples are called the ''pseudo data,'' which represent the normal process behaviour. Then, the pseudo data are used together with the real data calculate the threshold h in Equation (16). Since the MC method asymptotically converges to the true confidence bound of a *pdf* when the number of random samples goes to infinity (Berger, 1985), it is a reasonable choice when the real data are insufficient.

Online Monitoring

In online applications, the procedure of monitoring with the phase/transition models is described as follows:

1 First, the new measurements are variable-wise normalised. If the PCA scores have been used in the modelling step instead of the normalised data, the same kind of PCA decomposition should also be conducted on the normalised new measurements to calculate the scores.
2 In order to call a proper phase or transition model, the status of each sample should be determined. Using the phase division model, the posterior probabilities of each sample are calculated and compared to the confidence bound ϕ. Consequently, the new sample can be allocated to a certain phase or transition period.
3 The measurements are re-normalised separately in each phase or transition period in batch-wise way. Again, PCA decomposition may be performed, if the PCA scores have been utilised in monitoring model building.
4 Using a proper phase or transition model, the probability density of each sample is derived. If a probability density value is smaller than the corresponding threshold $h_{k^d}^d$, a faulty point is identified.

If one wants to do fault diagnosis after an abnormality is detected, the missing variable-based contribution analysis can be conducted. The details of such method can be found in the literature (Chen and Zhang, 2009).

APPLICATION RESULTS

Penicillin Fermentation

In this subsection, a well-known benchmark penicillin fermentation process is used to demonstrate the effectiveness of the proposed method. The penicillin fermentation is a typical multiphase batch process, which starts with a batch preculture for biomass growth. During the batch preculture, most of the initially added substrate is consumed by the microorganisms and the carbon source (glucose) is depleted. After the glucose concentration reaches a threshold value, the process switches to the fed-batch operation with continuous substrate feed. The flow diagram of the penicillin fermentation process is plotted in Figure 2.

A process simulator will be used to generate the operation data for process modelling and monitoring purposes. The simulation time for each batch is 400 h with a sampling interval of 0.5 h. Three hundred batches of normal operation data are utilised in phase division and modelling. Although the total batch durations are selected to be same, the durations of each phase change from batch to batch, as revealed later. Ten process variables are measured, as listed in Table 1. The initial conditions and parameter

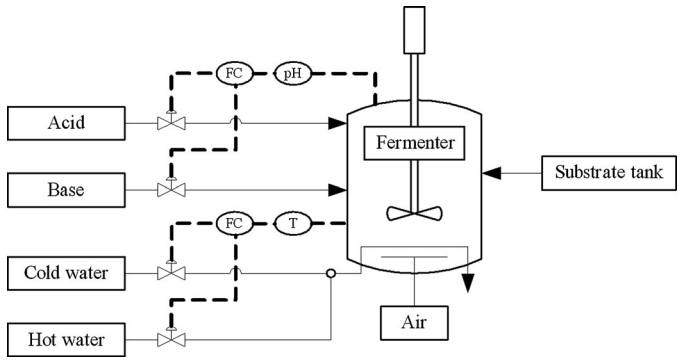

Figure 2. Flow diagram of the penicillin fermentation process.

Table 1. Monitored variables in the penicillin fermentation process

Variable no.	Variable definition
1	Dissolved oxygen concentration
2	Culture volume
3	Carbon dioxide concentration
4	pH
5	Fermentor temperature
6	Generated heat
7	Aeration rate
8	Agitator power
9	Substrate feed flow rate
10	Substrate feed temperature

Table 2. Initial conditions and controller parameters for nominal operations

Variable	Value
Initial substrate concentration (g/L)	15
Initial dissolved O_2 concentration (g/L)	1.16
Initial biomass concentration (g/L)	0.1
Initial penicillin concentration (g/L)	0
Initial fermentor volume (L)	100
Initial CO_2 concentration (mmol/L)	0.5
Initial Hydrogen ion concentration (mol/L)	10^{-5}
Initial fermentor temperature (K)	298
Initial heat generated (kcal)	0
Nominal value of aeration rate (L/h)	8.6
Nominal value of agitator power (W)	30
Nominal value of substrate feed rate (L/h)	0.042
Nominal value of feed temperature (K)	296
Temperature set point (K)	298
pH set point	5
Cooling controller gain	70
Cooling integral time (h)	0.5
Cooling derivative time (h)	1.6
Heating controller gain	5
Heating integral time (h)	0.8
Heating derivative time (h)	0.05
Acid controller gain	0.0001
Acid integral time (h)	8.4
Acid derivative time (h)	0.125
Base controller gain	0.0008
Base integral time (h)	4.2
Base derivative time (h)	0.2625

settings for normal operations are listed in Table 2. More details about the simulator can be found by Birol et al. (2002).

Before process monitoring, phase division and transition identification are performed firstly. For comparison, the results achieved by the proposed method and the multiway GMM method are both plotted in Figure 3. In the plots, the dash, dash/dot, dash/dot/dot curves are the curves of posterior probabilities corresponding to three different Gaussian components, respectively, while the solid lines in Figure 3a and b are the 99% confidence bound for phase division using the proposed method. The detailed descriptions of Figure 3 are provided as following.

Based on the proposed method, Figure 3a and b show the posterior probabilities of the samples in two arbitrary training batches. A confidence bound of 99% is defined for phase division and transition identification. Obviously, each batch is divided into four periods. From the process knowledge, it is easy to know that period I is a phase corresponding to the batch preculture step, while the fed-batch operation is further divided into two phases (II and IV) connected by a gradual transition period (III). Period II is dominated by the production of penicillin, while the penicillin production is saturated in period IV. It is also noticed that there is no transition between the first two phases, which is because the switch from the batch preculture to the fed-batch production happens in sudden instead of gradually. Therefore, the results coincide with the process nature very well.

Comparing the locations of the phase division points in Figure 3a and b, variations in operation progress between batches can be observed in the results of phase division and transition identification, which reveals that the process is of uneven durations, although the entire batch lengths are equal. Specifically, the lengths of the second phase vary from to 95.5 to 111 h, while the transitions usually take 11.5–24.5 h. Such extracted information could be potentially very useful for process or product quality improvements. The issues of process improvement and quality control utilising uneven-duration information will be a focus in future research.

The posterior probabilities calculated based on multiway GMM are plotted in Figure 3c and d for comparison, which provide less clear or stable results for phase division. An important reason is that the multiway GMM performs batch-wise normalisation before phase division. When the uneven-duration problem exists, such normalisation is not reasonable, which distorts the phase division results.

Two faults are generated to verify the fault detection ability of the proposed method. In the first case, a 10% step increase is exerted on the variable of substrate feed rate, which starts from the 80th hour. The fault detection results based on the proposed method are plotted in Figure 4, where the 95% control limit is utilised in online monitoring. For better illustration, all likelihood values are in log scale. The circle curve in the figure shows the negative log likelihood values, while the solid curve shows the corresponding thresholds. As the control chart indicates, the fault is efficiently detected, almost without any delay. The second fault is a slow drift in the trajectory of substrate feed rate, which starts from the 200th hour with a slope of 0.001. Considering the small slope of the drift, the monitoring results are also very good, as shown in Figure 5. Figure 6 shows the monitoring results of the second fault based on the conventional MPCA method. The variable trajectories have been synchronised before performing MPCA. As can be observed, the *SPE* control chart provides a clear indication of the fault only at the end of this batch, while the T^2 control chart cannot detect the fault at all. The comparison between Figures 5 and 6 demonstrates the effectiveness of the proposed method.

Injection Moulding

Another example is an injection moulding process (Rubin and Rubin, 1972), which is a well-established technique widely applied in the polymer processing industry. Injection moulding is also a multiphase batch process, where the main operation phases include filling, packing–holding and cooling. During the filling phase, the screw moves forward, injecting polymer melt into the mould cavity. Once the cavity is fully filled, the packing–holding phase starts. During this phase, additional material is packed into the cavity to compensate for the material shrinkage caused by the material cooling and solidification. Once the gate freezes, the process switches to the cooling phase, during which the material in the mould cavity is cooled down, until the part in the mould is sufficiently solidified and becomes rigid enough to be ejected from the mould without damage. Plastication happens in the early part of this phase. The screw rotates to shear and melt the material in the barrel, and conveys the polymer melt to the front of the screw, preparing for next cycle. Therefore, the cooling phase can be further divided into two sub-phases.

In the experiments, the feed material is high-density polyethylene (HDPE). The setpoint of the injection velocity is 35 mm/s. The three-band barrel temperatures are set to be at 210°C, 210°C and 160°C, respectively. The packing pressure is set to be 30 bar. The packing–holding time is fixed to be 5 s. The sampling interval is 50 ms. Under the above operation condition, eight process variables are measured as listed in Table 3. Fifty normal cycles are collected for phase division and process modelling.

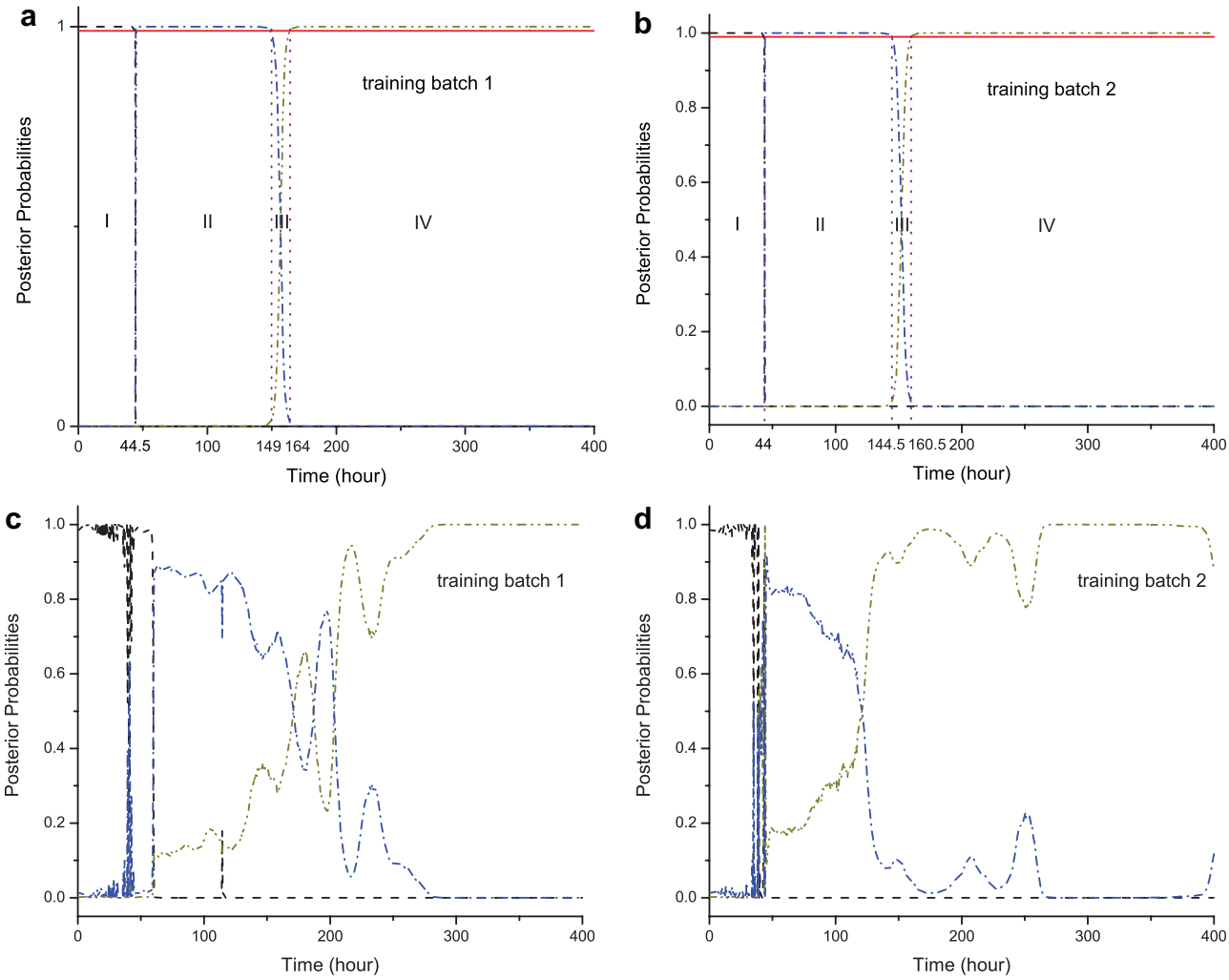

Figure 3. Phase division and transition identification results: (a) of training batch 1 based on the proposed method; (b) of training batch 2 based on the proposed method; (c) of training batch 1 based on multiway GMM; (d) of training batch 2 based on multiway GMM. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

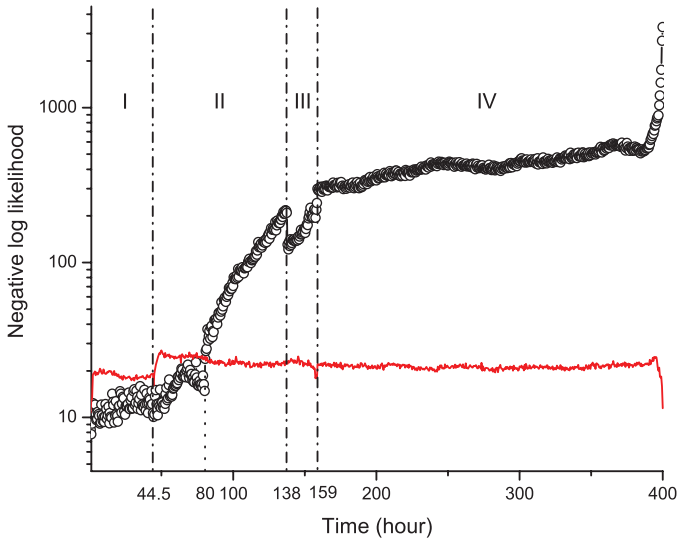

Figure 4. Online detection result of a step-change fault based on the proposed method. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

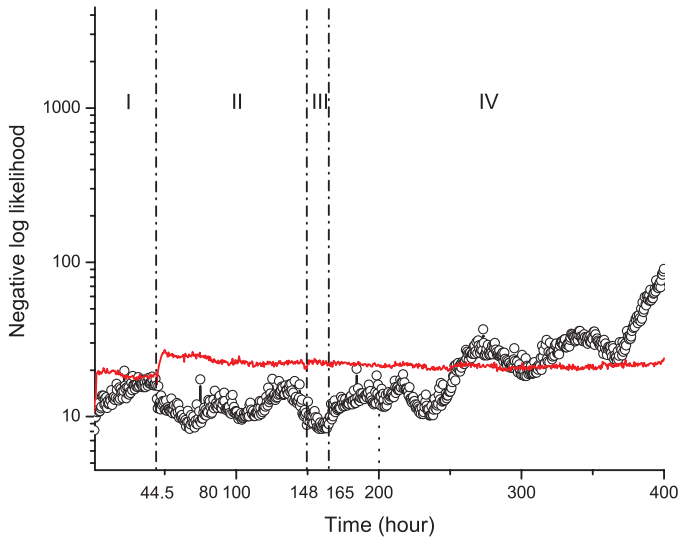

Figure 5. Online monitoring results of a process drift based on the proposed method. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

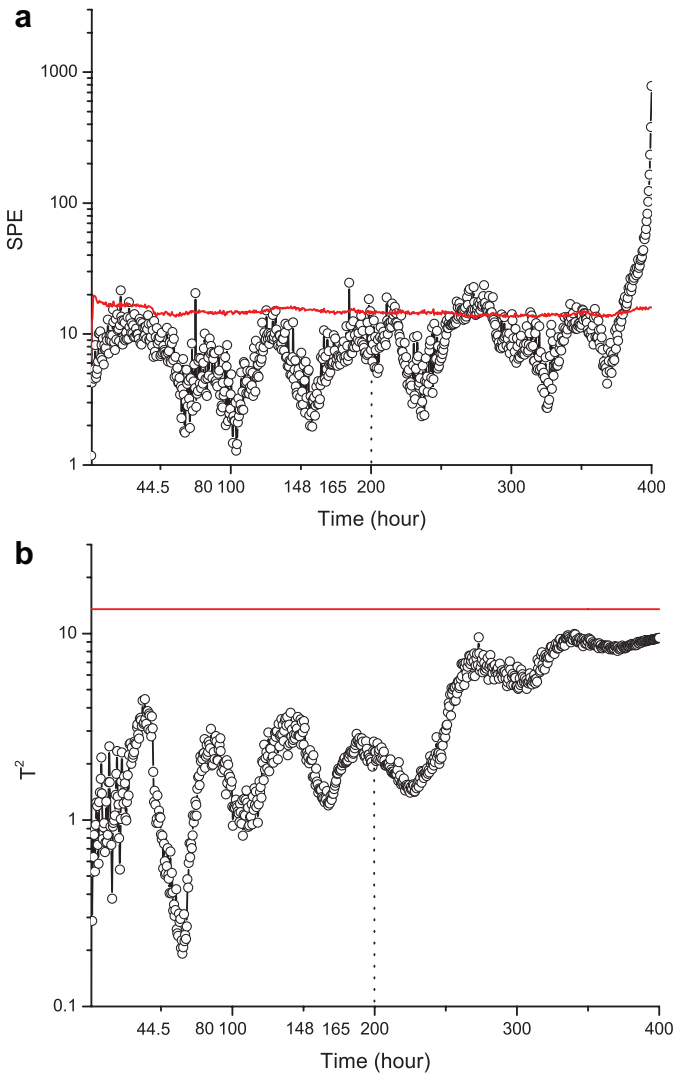

Figure 6. Online monitoring results of a process drift based on the MPCA. (a) SPE control chart; (b) T^2 control chart. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

Phase division is conducted based on the proposed method. Figure 7 summarises the phase division results of two arbitrary batches. The figure shows that the phase division model consists of five Gaussian component functions, where the components 1–4 correspond to the phases of filling, packing and holding, plastication, and cooling after plastication, respectively. The 5th component function describes the spikes at the beginning of some

Table 3. Measured variables in the injection moulding process

Variable no.	Variable definition
1	Valve opening 1
2	Valve opening 2
3	Screw stroke
4	Injection velocity
5	Injection pressure/back pressure
6	Barrel temperature (zone 1)
7	Barrel temperature (zone 2)
8	Barrel temperature (zone 3)

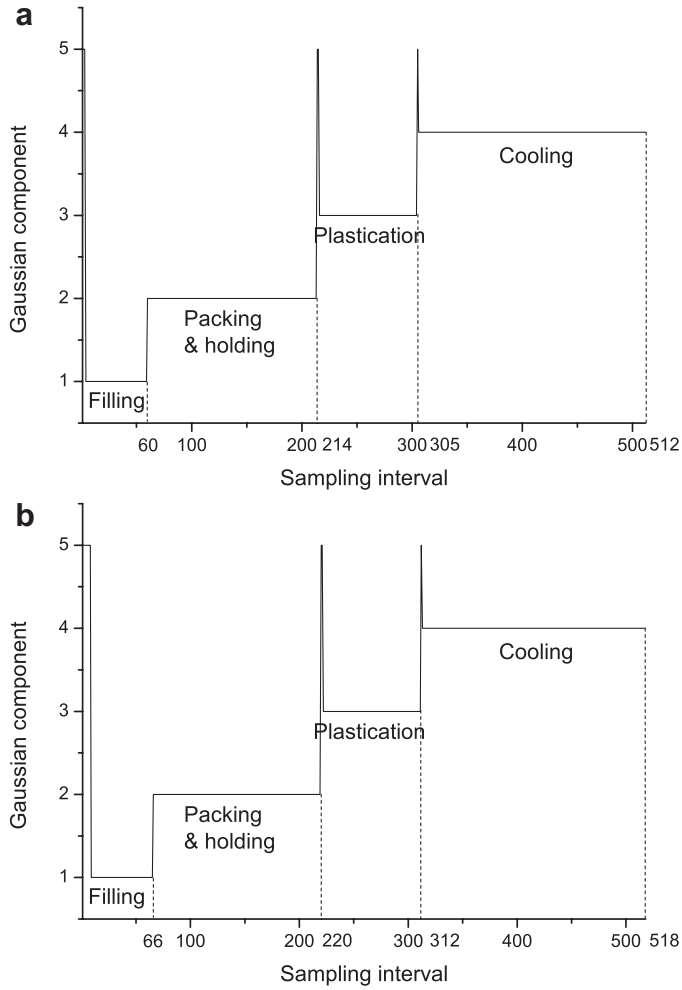

Figure 7. Phase division results for an injection moulding process. (a) Phase division results for arbitrary batch 1; (b) phase division results for arbitrary batch 2.

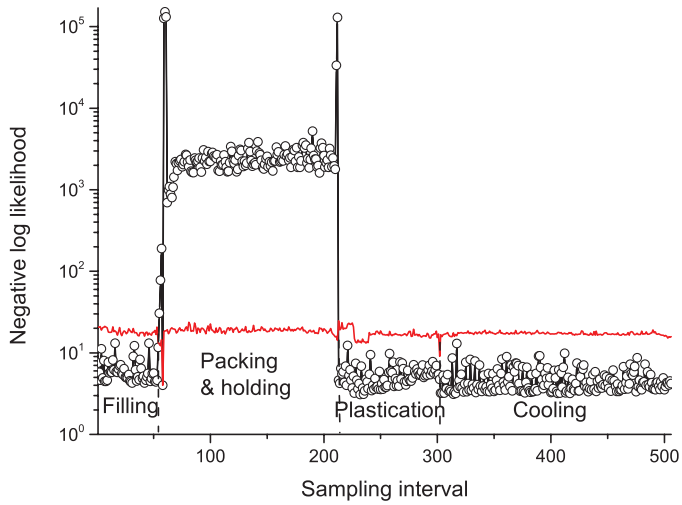

Figure 8. Online monitoring results of a false caused by lower packing pressure based on the proposed method. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

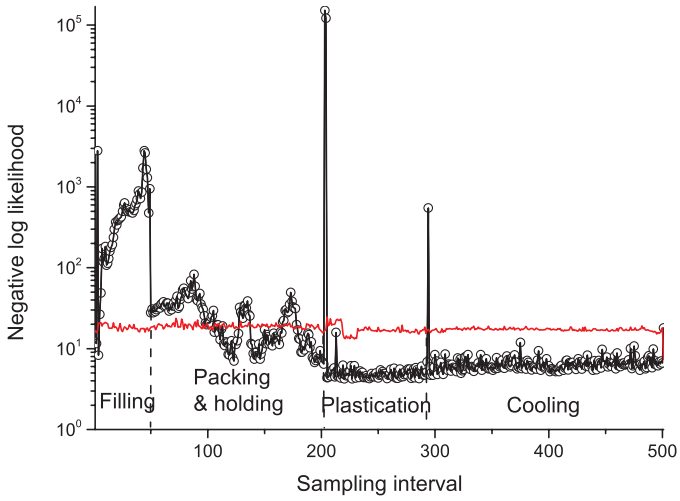

Figure 9. Online monitoring results of a false caused by higher injection velocity based on the proposed method. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

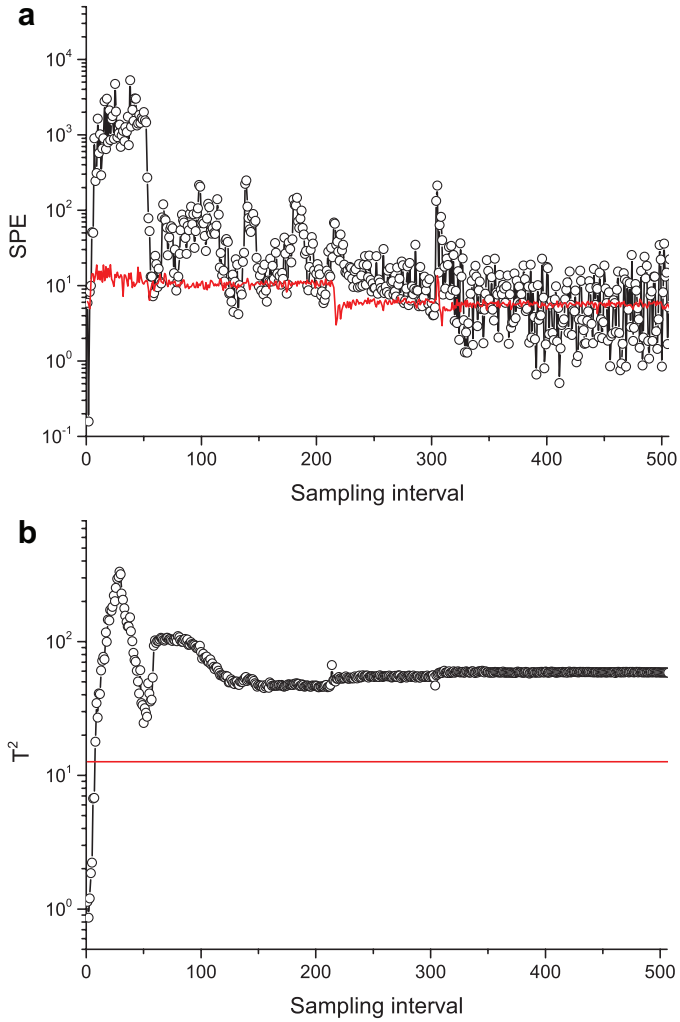

Figure 10. Online monitoring results of a false caused by higher injection velocity based on MPCA. (a) SPE control chart; (b) T^2 control chart. [Color figure can be seen in the online version of this article, available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-019X]

phases. From the figure, it can be noticed that there is no gradual transition between the neighbouring phases, which is in good agreement with the nature of the injection moulding process. The spikes are caused by the sudden transition from one phase to another. Meanwhile, uneven durations are observed in the figure, which is mainly attributed to the varying lengths of the first phase. Again, such results reveal process characteristics. In this process, the variations in injection velocity lead to uneven durations of the filling phase.

To show the effectiveness of online monitoring using the proposed method, two different types of faults are generated, which are caused by a lower packing pressure and a higher injection velocity, respectively. In the first faulty cycle, the set point of the packing pressure is assigned to be 25 bar. Such fault only affects the packing and holding phase. Since the packing–holding time is fixed, such fault does not affect the length of the cycle. Figure 8 shows that the monitoring is very efficient and well reflects the process feature. The second fault occurs to the injection velocity. In the filling phase, the set point of the velocity is 45 mm/s. Such fault has influences on the filling phase. Besides, because of the inertia of the screw, the initial value of the stroke in the next phase is smaller than usual, changing the process characteristics in the packing and holding phase. Because the injection velocity is increased, the duration of the filling phase becomes shorter than the phase duration in normal operations, causing the uneven-duration problem. The control chart in Figure 9 clearly detects the fault in the first two phases; and since the fault does not affect the plastication phase and the cooling phase, the monitoring results in these two phases show that the process operation is normal. For comparison, the MPCA method based on trajectory synchronisation is also utilised to monitor the same faulty cycle. The number of PCs is selected as 4, according to the cross-validation. Figure 10 shows that, although both the *SPE* and the T^2 control charts are able to detect the fault, they cannot correctly identify the abnormal phases. All the four phases are considered to be faulty by MPCA, which is not true.

CONCLUSIONS

In industrial processing, many batch processes are inherently multiphase and of uneven durations. Such nature increases the complication of the process characteristics, and makes many multivariate statistical methods invalid in batch process modelling and monitoring. In this study, a GMM-based method is developed to handle both the multiphase and uneven-duration issues simultaneously, while the transitions between phases can also be identified and modelled well. The phase division ability and the monitoring efficiency of the proposed method are verified with a benchmark penicillin fermentation process and an injection moulding process.

ACKNOWLEDGEMENTS

This work was supported in part by Grant No. 100N2072E1 of the Advanced Manufacturing and Service Management Research Center of National Tsing Hua University, Taiwan.

REFERENCES

Berger, J. O., ''Statistical Decision Theory and Bayesian Analysis,'' (2nd ed.) Springer, New York (1985).

Birol, G., C. Undey and A. Çinar, ''A Modular Simulation Package for Fed-Batch Fermentation: Penicillin Production,'' Comput. Chem. Eng. **26**, 1553–1565 (2002).
Bishop, C. M., ''Pattern Recognition and Machine Learning,'' Springer, New York (2006).
Camacho, J. and J. Picó, ''Online Monitoring of Batch Processes Using Multi-Phase Principal Component Analysis,'' J. Process Control **16**, 1021–1035 (2006).
Chen, T., J. Morris and E. Martin, ''Probability Density Estimation Via an Infinite Gaussian Mixture Model: Application to Statistical Process Monitoring,'' J. R. Stat. Soc. Ser. C Appl. Stat. **55**, 699–716 (2006).
Chen, T. and J. Zhang, ''On-Line Multivariate Statistical Monitoring of Batch Processes Using Gaussian Mixture Model,'' Comput. Chem. Eng. **34**, 500–507 (2009).
Choi, S. W., J. H. Park and I. B. Lee, ''Process Monitoring Using a Gaussian Mixture Model Via Principal Component Analysis and Discriminant Analysis,'' Comput. Chem. Eng. **28**, 1377–1387 (2004).
Dempster, A., N. Laird and D. Rubin, ''Maximum Likelihood From Incomplete Data Via the EM Algorithm,'' J. R. Stat. Soc. Ser. B Methodol. **39**, 1–38 (1977).
Figueiredo, M. and A. Jain, ''Unsupervised Learning of Finite Mixture Models,'' IEEE Trans. Pattern Anal. Mach. Intell. **24**, 381–396 (2002).
Fraley, C. and A. Raftery, ''Model-Based Clustering, Discriminant Analysis, and Density Estimation,'' J. Am. Stat. Assoc. **97**, 611–632 (2002).
Jackson, J., ''A User's Guide to Principal Components,'' Wiley, New York (1991).
Jolliffe, I., ''Principal Component Analysis,'' Springer, New York (2002).
Kassidas, A., J. MacGregor and P. Taylor, ''Synchronisation of Batch Trajectories Using Dynamic Time Warping,'' AIChE J. **44**, 864–875 (1998).
Lu, N., F. Gao and F. Wang, ''Sub-PCA Modelling and On-Line Monitoring Strategy for Batch Processes,'' AIChE J. **50**, 255–259 (2004a).
Lu, N., F. Gao, Y. Yang and F. Wang, ''PCA-Based Modelling and On-Line Monitoring Strategy for Uneven-Length Batch Processes,'' Ind. Eng. Chem. Res. **43**, 3343–3352 (2004b).
Lu, N., Y. Yang, F. Gao and F. Wang, ''Stage-based multivariate statistical analysis for injection molding,'' in ''Proc. 7th International Symposium on Advanced Control of Chemical Processes,'' Hong Kong (2003), pp. 639–644.
McLachlan, G. and D. Peel, ''Finite Mixture Models,'' Wiley, New York (2000).
Nomikos, P. and J. MacGregor, ''Monitoring Batch Processes Using Multiway Principal Component Analysis,'' AIChE J. **40**, 1361–1375 (1994).
Nomikos, P. and J. MacGregor, ''Multivariate SPC Charts for Monitoring Batch Processes,'' Technometrics **37**, 41–59 (1995).
Rubin, I. and A. Rubin, ''Injection Molding: Theory and Practice,'' Wiley, New York (1972).
Thissen, U., H. Swierenga, A. de Weijer, R. Wehrens, W. Melssen and L. Buydens, ''Multivariate Statistical Process Control Using Mixture Modelling,'' J. Chemom. **19**, 23–31 (2005).
Undey, C. and A. Cinar, ''Statistical Monitoring of Multistage, Multiphase Batch Processes,'' IEEE Control Syst. Mag. **22**, 40–52 (2002).
Wold, S., N. Kettaneh, H. Fridén and A. Holmberg, ''Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments,'' Chemom. Intell. Lab. Syst. **44**, 331–340 (1998).
Yao, Y. and F. Gao, ''Phase and Transition Based Batch Process Modelling and Online Monitoring,'' J. Process Control **19**, 816–826 (2009a).
Yao, Y. and F. Gao, ''A Survey on Multistage/Multiphase Statistical Modelling Methods for Batch Processes,'' Annu. Rev. Control **33**, 172–183 (2009b).
Yu, J. and S. Qin, ''Multiway Gaussian Mixture Model Based Multiphase Batch Process Monitoring,'' Ind. Eng. Chem. Res. **48**, 8585–8594 (2009).
Zhao, C., S. Mo, F. Gao, N. Lu and Y. Yao, ''Statistical Analysis and Online Monitoring for Handling Multiphase Batch Processes With Varying Durations,'' J. Process Control **21**, 817–829 (2011).
Zhao, C., F. Wang, N. Lu and M. Jia, ''Stage-Based Soft-Transition Multiple PCA Modelling and On-Line Monitoring Strategy for Batch Processes,'' J. Process Control **17**, 728–741 (2007).

Manuscript received August 28, 2011; revised manuscript received November 2, 2011; accepted for publication November 10, 2011.