On-line multivariate statistical monitoring of batch processes using Gaussian mixture model

Tao Chen*,[a], Jie Zhang[b]

[a] *School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459*
[b] *School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.*

Abstract

This paper considers multivariate statistical monitoring of batch manufacturing processes. It is known that conventional monitoring approaches, e.g. principal component analysis (PCA), are not applicable when the normal operating conditions of the process cannot be sufficiently represented by a multivariate Gaussian distribution. To address this issue, Gaussian mixture model (GMM) has been proposed to estimate the probability density function (*pdf*) of the process nominal data, with improved monitoring results having been reported for continuous processes. This paper extends the application of GMM to on-line monitoring of batch processes. Furthermore, a method of contribution analysis is presented to identify the variables that are responsible for the onset of process fault. The proposed method is demonstrated through its application to a batch semiconductor etch process.

Key words: Batch processes, fault detection and diagnosis, mixture model, principal component analysis, probability density estimation, multivariate statistical process monitoring

1. Introduction

Batch processing is of great importance in a variety of industrial sectors for the production of low-volume, high-value added products, including pharmaceuticals, polymers, beverage and fine chemicals. With increasing commercial competition, it is crucial to ensure consistent and high product quality, as well as the safety of the processes. These requirements have helped promote the technique of multivariate statistical process monitoring (MSPM) (Kano and Nakagawa, 2008; Martin et al., 1999; Qin, 2003). The basis of MSPM is a set of historical data that have been collected when the process is running under normal operating conditions (NOC). These data are then used to establish the confidence bounds for monitoring statistics, e.g. Hotelling's T^2 and squared prediction error (SPE), to detect the onset of process deviations. The primary objective of process monitoring is to identify abnormal behavior as early as possible, in addition to keeping an acceptably low false alarm rate.

Due to the multi-way property of batch process data, special tools are required for the modelling and monitoring purposes, including multi-way principal component analysis (MPCA) (Nomikos and MacGregor, 1995b), multi-way partial least squares (MPLS) (Nomikos and MacGregor, 1995a), hierarchical PCA (Rannar et al., 1998) and their dynamic and non-linear variants (Chen and Liu, 2002; Lee et al., 2004). The methods for on-line monitoring of batch process can be classified into two categories. The first does not require measurements of the entire batch duration to be available, for example hierarchical and two-dimensional dynamic PCA (Rannar et al., 1998; Lu et al., 2005). In the other category, the entire batch data is required for the calculation of the monitoring statistics, whilst the data from a new batch is available only up to the current time. Therefore, the future data must be predicted in some way (Nomikos and

*Corresponding author. Tel.: +65 6513 8267; Fax: +65 6794 7553.
Email addresses: chentao@ntu.edu.sg (Tao Chen), jie.zhang@ncl.ac.uk (Jie Zhang)

MacGregor, 1995b,a). In this paper the latter of the two approaches is considered, and the details will be discussed subsequently in Section 2.

However, the afore reviewed conventional process monitoring methods are based on a restrictive assumption that the NOC can be represented by a multivariate Gaussian distribution. Specifically, the confidence bounds for T^2 and SPE are calculated by assuming that the PCA/PLS scores and prediction errors are Gaussian distributed. This assumption may be invalid when the process data are collected from a complex manufacturing process (Thissen et al., 2005), or when non-linear projection techniques are used to model the nominal historical data (Wilson et al., 1999). To address this issue, several semi-parametric and non-parametric statistical methods have been applied, including kernel density estimation (Martin and Morris, 1996), wavelet-based density estimation (Safavi et al., 1997), and Gaussian mixture model (GMM) (Chen et al., 2006; Choi et al., 2004; Thissen et al., 2005; Yu and Qin, 2008). Due to its solid theoretical foundation and good practical performance, GMM has been widely applied to the monitoring of continuous processes, as well as batch-wise monitoring of batch processes.

The major contribution of this paper is to extend the application of GMM to on-line monitoring of batch processes. As the first step, MPCA is applied to the nominal batch data to extract the low-dimensional representation of the process. The challenge with on-line monitoring is that the scores and SPE must be predicted based on available process measurements up to the current time step. Clearly the predicted scores and SPE are not identical to the values that are calculated from the entire batch duration, and thus the predictions may not conform to the nominal distribution even if the process is running normally. This paper follows the approach in (Nomikos and MacGregor, 1995b) to pass the nominal batches through the monitoring procedure and collect the predicted scores and SPE at each time step. Then GMM is employed to estimate the joint *pdf* of these predicted scores and SPE from MPCA at each time step. Furthermore, a contribution analysis method is proposed to investigate the influence of individual measured variable to the detected fault. The contribution analysis can facilitate the diagnosis of the source of the process fault.

The rest of this paper is organized as follows. Section 2 gives a brief overview of the PCA and GMM tools for process monitoring, followed by the discussion of the on-line monitoring strategy in Section 3. Section 4 demonstrates the application of the on-line monitoring techniques to a batch semiconductor manufacturing process. Finally Section 5 concludes this paper.

2. PCA and GMM for process monitoring

This section presents a brief overview of the PCA and GMM techniques. A number of issues related to the application to process monitoring are discussed, including model selection and the construction of confidence bound.

2.1. PCA

Principal component analysis (PCA) (Jolliffe, 2002) is a general multivariate statistical projection technique for dimension reduction, where the original data is linearly projected onto low-dimensional space such that the variance is maximized. Formally the D-dimensional data $\mathbf{x}$ is represented by a linear combination of the Q-dimensional scores $\mathbf{t}$ plus a noise vector $\mathbf{e}$: $\mathbf{x} = \mathbf{W}\mathbf{t} + \mathbf{e}$, where $\mathbf{W}$ are the eigenvectors of the sample covariance matrix having the Q largest eigenvalues ($Q \leq D$). Consequently, normal process behavior can be characterized by the first Q principal components, which capture the main source of data variability. The proper number of principal components can be selected using a number of criteria, including variance ratio, cross-validation and the "broken-stick" rule (Jolliffe, 2002). The widely used cross-validation method is adopted in this paper.

In statistical process monitoring, the next step is to define the monitoring statistics and the corresponding confidence bounds. Traditionally two metrics are used: $T^2 = \mathbf{t}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{t}$ and SPE as $r = \mathbf{e}^{\mathrm{T}}\mathbf{e}$, where $\mathbf{\Lambda}$ is a diagonal matrix comprising the Q largest eigenvalues.

As discussed previously, the first issue with T^2 and SPE is that the corresponding confidence bounds are calculated based on restrictive Gaussian distribution. Secondly, two separate metrics are required for process monitoring. Practically, the process is identified as deviating from normal operation if either T^2 or SPE moves outside the confidence bounds. This solution could potentially increase the false alarm level. (A

detailed discussion on this issue is given in Appendix A.) The technique of GMM is suitable for addressing the two issues simultaneously. In previous work (Chen et al., 2006) we have demonstrated that a unified monitoring statistic can be obtained by estimating the joint *pdf* of the PCA scores and log-SPE using GMM, i.e. the *pdf* of a $(Q+1)$-dimensional vector $\mathbf{z} = (\mathbf{t}^{\mathrm{T}}, \log r)^{\mathrm{T}}$. The logarithm operator is used to transform the non-negative SPE onto the whole real axis on which the GMM is defined.

In this paper the methodology in (Chen et al., 2006) is followed to establish the confidence bounds for process monitoring based on PCA and GMM techniques. GMM is described in detail in the next subsection.

2.2. GMM

As a general tool for *pdf* estimation, GMM has been used in a wide variety of problems in applied statistics and pattern recognition. A GMM is a weighted sum of M component densities, each being a multivariate Gaussian with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$:

$$p(\mathbf{z}|\boldsymbol{\theta}) = \sum_{i=1}^{M} \alpha_i G(\mathbf{z};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where the weights satisfy the constraint: $\sum_{i=1}^{M} \alpha_i = 1$. A GMM is parameterized by the mean vectors, covariance matrices and mixture weights: $\boldsymbol{\theta} = \{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i; i = 1, \ldots, M\}$.

Given a set of training data $\{\mathbf{z}_n, n = 1, \ldots, N\}$, the parameters can be estimated by maximizing the likelihood function: $L(\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\boldsymbol{\theta})$. In the context of process monitoring, $\mathbf{z}_n$ is the $(Q+1)$-dimensional vector of PCA scores and log-SPE: $\mathbf{z}_n = (\mathbf{t}_n^{\mathrm{T}}, \log r_n)^{\mathrm{T}}$. The maximization can be implemented iteratively using the expectation-maximization (EM) algorithm (Dempster et al., 1977). On each EM iteration, the following updating formulas are used to guarantee a monotonic increase in the likelihood value:

$$\alpha_i = \frac{1}{N} \sum_{n=1}^{N} p(i|\mathbf{z}_n, \boldsymbol{\theta}) \tag{2}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{n=1}^{N} p(i|\mathbf{z}_n, \boldsymbol{\theta}) \mathbf{z}_n}{\sum_{n=1}^{N} p(i|\mathbf{z}_n, \boldsymbol{\theta})} \tag{3}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{n=1}^{N} p(i|\mathbf{z}_n, \boldsymbol{\theta})(\mathbf{z}_n - \boldsymbol{\mu}_i)(\mathbf{z}_n - \boldsymbol{\mu}_i)^{\mathrm{T}}}{\sum_{n=1}^{N} p(i|\mathbf{z}_n, \boldsymbol{\theta})} \tag{4}$$

where

$$p(i|\mathbf{z}_n, \boldsymbol{\theta}) = \frac{\alpha_i G(\mathbf{z}_n; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} \alpha_k G(\mathbf{z}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{5}$$

The EM algorithm can be initialized through the K-means clustering algorithm (Choi et al., 2004; Thissen et al., 2005).

The number of mixture components, M, must be selected prior to the training of a GMM. This model selection problem can be addressed using a number of methods, including cross-validation and Bayesian information criterion (BIC) (Schwarz, 1978). BIC is widely applied in model selection problems for its effectiveness and low computational cost. According to BIC the model is selected such that $L - (H/2)\log N$ is the largest, where L is the log-likelihood of the data and H is the total number of parameters within the model. The motivation of BIC is that a good model should be able to sufficiently explain the data (the log-likelihood) with low model complexity (the number of parameters). In this study, BIC is adopted for the selection of number of mixtures.

One of the advantages of the GMM for process monitoring is that it provides the likelihood value as the single statistic for the construction of confidence bounds, as opposed to the confidence bounds for two statistics (i.e. the T^2 and SPE) in conventional process monitoring techniques. In practice a single monitoring statistic simplifies the plant operators' decision effort (Chen et al., 2006), and it may be more sensitive to some subtle process faults (Chen et al., 2004).

On the basis of the *pdf* $p(\mathbf{z}|\boldsymbol{\theta})$ for the normal operating data, the $100\beta\%$ confidence bound is defined as a likelihood threshold h that satisfies the following integral (Chen et al., 2006):

$$\int_{\mathbf{z}:p(\mathbf{z}|\boldsymbol{\theta})>h} p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z} = \beta \tag{6}$$

To determine the confidence bound, we can calculate the likelihood of all the nominal data, and then find h that is less than the likelihood of $100\beta\%$ (e.g. 99%) of the nominal data (Thissen et al., 2005). This approach is applicable to most continuous processes where the number of nominal data points can be up to several thousand; however it may be unreliable when the nominal data is very limited as in batch process monitoring. The estimation of the confidence bound based on limited batches would be very sensitive to the data, and thus a small perturbation in the data would result in very different estimation of the h.

To address this issue, we resort to numerical Monte Carlo (MC) simulation to approximate the integral in Eq. (6) (Chen et al., 2006). Specifically we generate N_s random samples, $\{\mathbf{z}^j, j = 1, \ldots, N_s\}$, from $p(\mathbf{z}|\boldsymbol{\theta})$. These samples serve as the "pseudo data" (since the real data is not sufficient) to represent the normal process behavior. Thus the MC samples, in conjunction with nominal process data, are used to calculate the confidence bound h. It has been shown that the MC method asymptotically converges to the true confidence bound of a *pdf* when the number of random samples goes to infinity (Berger, 1985). As such, MC simulation is a valid approach provided that $p(\mathbf{z}|\boldsymbol{\theta})$ is an accurate model for the process data; otherwise a better modelling method should be adopted. Therefore, a new batch $\mathbf{z}$ is considered to be faulty if $p(\mathbf{z}|\boldsymbol{\theta}) < h$ (or equivalently $-p(\mathbf{z}|\boldsymbol{\theta}) > -h$). The algorithm for the generation of random samples from a GMM can be found in, e.g. (Bishop, 2006, Chapter 9.2). The number of MC samples required (N_s) to approximate the confidence bounds is dependent on the dimension of $\mathbf{z}$, and it can be determined heuristically.

3. Monitoring of batch processes

To analyze the three-way batch data ($N \times J \times K$) (N, J and K denote the number of batches, process variables at each time instance, and time steps, respectively), multi-way analysis methods have been proposed to unfold the data array into a two-way matrix on which conventional PCA is then performed (Nomikos and MacGregor, 1995b). This study unfolds the data array into a large matrix ($N \times JK$) such that each batch is treated as a "data point". This two-way matrix is then pre-processed to zero mean and unit standard deviation on each column, prior to the application of PCA to extract the scores $\mathbf{t}_n$ and SPE r_n, $n = 1, \ldots, N$. Then a Gaussian mixture model is developed for the joint vector $\mathbf{z}_n = (\mathbf{t}_n^{\mathrm{T}}, \log r_n)^{\mathrm{T}}$, followed by the calculation of confidence bound using Monte Carlo simulation.

3.1. On-line monitoring

In the on-line monitoring stage, it is necessary to project the new batch onto the PCA space to obtain the scores and SPE, and then to calculate the likelihood value under the GMM to identify possible process anomaly. The issue is that, at time step t, the batch measurements are only available up to the current time. It is possible to develop multiple PCA and GMM models at each time step; however this strategy requires excessive computation and computer memory. A more reasonable and widely accepted method is to predict the scores and SPE using the available measurements.

More specifically, let $\bar{\mathbf{x}}_{1:t}$ be the vector of a new batch with available measurements from time step 1 to t. Note $\bar{\mathbf{x}}_{1:t}$ is a vector of order Jt. According to Nomikos and MacGregor (1995b), the least square prediction of the scores is:

$$\bar{\mathbf{t}}_{1:t} = \left(\mathbf{W}_{1:t}^{\mathrm{T}}\mathbf{W}_{1:t}\right)^{-1}\mathbf{W}_{1:t}^{\mathrm{T}}\bar{\mathbf{x}}_{1:t} \tag{7}$$

where $\mathbf{W}_{1:t}$ is the sub-matrix of $\mathbf{W}$ having the rows corresponding to time step 1 to t. In Eq. (7) the matrix to be inverted is well conditioned due to the orthogonality of the loading $\mathbf{W}$. Since the future measurements are not available, the prediction error can only be calculated up to time step t:

$$\bar{\mathbf{e}}_{1:t} = \bar{\mathbf{x}}_{1:t} - \mathbf{W}_{1:t}\bar{\mathbf{t}}_{1:t} \tag{8}$$

The SPE is then obtained as $\bar{\mathbf{e}}_{1:t}^{\mathrm{T}}\bar{\mathbf{e}}_{1:t}$. It was suggested to use the "instantaneous" SPE associated with the latest on-line measurements for process monitoring (Nomikos and MacGregor, 1995b), i.e. $\bar{\mathbf{e}}_t^{\mathrm{T}}\bar{\mathbf{e}}_t$, which is expected to increase the sensitivity of fault detection method. However, the instantaneous SPE leads to an excessive number of false alarms in the case study of this paper (see details in Section 4). The SPE calculated from Eq. (8), which in a sense is a smoothed version of the instantaneous SPE, may be a more appropriate monitoring metric. We will discuss this issue through the application study in Section 4. In practice, the choice between the instantaneous and smoothed SPE should be decided on a case-by-case basis.

Clearly, the predicted scores and SPE from Eqs. (7)(8), based on current available measurements, are not identical to the values that are calculated should the entire batch be available. As a result, the predicted scores and SPE may not conform to the *pdf* developed based on the entire duration of nominal batches, even if the process being monitored is running normally. This is a serious issue particularly in the initial stage of a batch processing, when only a small number of measurements are available to calculate the scores and SPE. We follow the standard approach in on-line batch process monitoring (Nomikos and MacGregor, 1995b) to pass each of the nominal batches through the monitoring procedure to collect the predicted scores and SPE at each time step from Eqs. (7)(8), and then apply GMM to estimate the joint *pdf* of these predicted scores and log-SPE at each time step, and to establish the confidence bounds as presented in Section 2. Essentially we propose to replace the confidence bounds for T^2 and SPE in (Nomikos and MacGregor, 1995b), where the process data is assumed to be Gaussian distributed, with more powerful Gaussian mixture model. For on-line monitoring of a new batch, the scores and SPE are calculated from Eqs. (7)(8), and the likelihood value is calculated under the GMM for the current time step. If this likelihood value is lower than the confidence bound, the process under monitoring is considered to be in a faulty condition.

3.2. Contribution analysis

Once an onset of fault is detected, the next step is to identify the source of the process fault. The technique of contribution analysis has become a indispensable step of process monitoring to provide valuable information for fault diagnosis (Miller et al., 1998). Contribution analysis aims to identify the variables that contribute the most to the violation of the confidence bound. In principle contribution analysis may not explicitly reveal the root-cause of the onset of faults, but it is undoubtedly helpful in pinpointing the inconsistent variables that may undergo further diagnosis procedures.

Within the context of PCA model, the conventional approach to achieving this goal is to decompose the T^2 and SPE into the sum of D contributing terms for the D process variables, and then the magnitude of these terms indicates the relative responsibility of the variables (Miller et al., 1998). However within the framework of GMM, it is not clear how the monitoring statistic, i.e. the likelihood value, can be decomposed with respect to the process variables.

This paper proposes a missing variable based contribution analysis approach that was previously utilized for fault identification in conjunction with PCA (Dunia et al., 1996; Yue and Qin, 2001). The original idea is that each process variable is treated as if it were missing and is reconstructed. The reconstruction is carried out for all variables, and the variables corresponding to the largest reconstruction errors are considered to contribute the most to the occurrence of the detected fault.

In the context of GMM based process monitoring, we do not seek to reconstruct the "missing" variable. Rather we remove each variable measured at the current time step t from $\bar{\mathbf{x}}_{1:t}$ in turn, denoted by $\bar{\mathbf{x}}_{1:t,-j}$ where $-j$ means the j-th variable is removed, and then re-calculate the scores and SPE similar to Eqs. (7)(8):

$$\bar{\mathbf{t}}_{1:t,-j} = \left(\mathbf{W}_{1:t,-j}^{\mathrm{T}}\mathbf{W}_{1:t,-j}\right)^{-1}\mathbf{W}_{1:t,-j}^{\mathrm{T}}\bar{\mathbf{x}}_{1:t,-j} \tag{9}$$

$$\mathrm{SPE} = \bar{\mathbf{e}}_{1:t,-j}^{\mathrm{T}}\bar{\mathbf{e}}_{1:t,-j} = \left(\bar{\mathbf{x}}_{1:t,-j} - \mathbf{W}_{1:t,-j}\bar{\mathbf{t}}_{1:t,-j}\right)^{\mathrm{T}}\left(\bar{\mathbf{x}}_{1:t,-j} - \mathbf{W}_{1:t,-j}\bar{\mathbf{t}}_{1:t,-j}\right) \tag{10}$$

where $\mathbf{W}_{1:t,-j}$ is obtained from $\mathbf{W}_{1:t}$ by eliminating the row corresponding to the removed variable. Finally we re-calculate the likelihood value of the scores and log-SPE under the GMM model at current time

step with the j-th variable being missing. Note that the GMM is developed based on prediction errors of all variables available at the current time. Therefore we add the average SPE from the nominal batches corresponding to the removed variable to Eq. (10). This is equivalent to replacing the SPE related to the removed variable with a nominal value.

Intuitively if one variable is responsible for the detected fault, then by eliminating it the re-calculated scores and SPE should be closer to the region of normal conditions, and thus the re-calculated likelihood should be significantly increased. This procedure is repeated for all the J process variables measured at the current time. We further denote the original likelihood value L and the re-calculated likelihood L_{-j}. Therefore the contribution of individual variables can be quantified by using $L_{-j} - L$. Furthermore, if L_{-j} is greater than the confidence bound h (or equivalently $L_{-j} - L > h - L$), the contribution of the j-th variable can be regarded as substantial, since its elimination would bring the process back to normal operating regions.

In principle, this missing variable based contribution analysis can be extended to remove a group of variables to investigate their collective contribution. However, the number of combinations of variables to be analyzed increases exponentially with the number of process variables (e.g. 10 variables have 2^{10} combinations), and thus an exhaustive search is normally infeasible. Nevertheless, if prior information is available to associate a specific fault mode with a set of process variables, then the contribution of these variables can be analyzed. If the contribution from these variables is significant, then the associated fault mode can be identified (Dunia et al., 1996; Yue and Qin, 2001).

4. Case study

The manufacture of semiconductors is introduced as an example of the on-line monitoring of batch processes. This study focuses specifically on an Al-stack etch process performed on the commercially available Lam 9600 plasma etch tool (Wise et al., 1999). Data from 12 process sensors, listed in Table 1, were collected during the wafer processing stage which runs for 80 s. A sampling interval of 1 s was used in the analysis. Thus for each batch, the data is of the order (12×80). A series of three experiments, resulting in three distinct data groups, were performed where faults were intentionally introduced by changing specific manipulated variables (TCP power, RF power, pressure, plasma flow rate and Helium chunk pressure). There are 107 normal operating batches and 20 faulty batches. Twenty batches were randomly selected from the normal batches to investigate the effect of false alarms. The remaining 87 nominal batches were used to build the MPCA and GMM models.

(Table 1 about here)

4.1. Off-line analysis

According to MPCA, the three-way nominal data array ($N \times J \times K = 87 \times 12 \times 80$) is unfolded into a large two-way matrix of the order (87×960), which is then mean-centered and scaled to unit standard deviation on each column. Then PCA is applied to the pre-processed data, where four principal components are retained according to leave-one-out cross-validation. Considering there are 960 columns in the unfolded matrix, it is not surprising to find that four principal components explain only 50.48% of the total variance (similar results can be found in the literature, e.g. (Nomikos and MacGregor, 1995b)).

Figure 1 gives the scatter plot of the PCA scores corresponding to the first two principal components. It is clear that the nominal data exhibits the characteristic of multiple groups, and it cannot be adequately approximated by a single multivariate Gaussian distribution. As a result, the 99% confidence bound does not capture the region of NOC accurately. (Note that the 99% bound in Figure 1 is obtained using two principal components, whilst subsequent monitoring results are based on using four components.) Clearly, more complex models are required to represent the nominal behavior of the process.

(Figure 1 about here)

To develop a GMM for the PCA scores and log-SPE, the number of mixtures was determined to be three according to BIC. Once the GMM is developed, the 95% and 99% confidence bounds is calculated using Monte Carlo simulation presented in Section 2.2, where the number of random samples is heuristically determined to be 10,000. Despite the large sample size, the CPU time for the Monte Carlo simulation was only 0.03 s (Matlab implementation under Windows XP with Pentium 2.8 GHz CPU). In the literature the 95% is treated as "warning bound" and 99% "action bound". Throughout this section the process is classified as faulty if the 99% confidence bound is violated.

Table 2 summarizes the off-line batch-wise monitoring results for both conventional PCA and the GMM approach. Both methods incur only one false alarm in this example. The large number of missing errors from T^2 appears to be the result of over-estimation of the confidence bound. The SPE statistic is more sensitive to the fault and it attains six missing errors. By combining T^2 and SPE in the way that the process is identified as faulty if either metric is exceeded, the number of missing errors is still six. Table 2 clearly indicates that GMM outperforms the conventional PCA in terms of smaller number of missing errors through the direct estimation of the joint *pdf* of the PCA scores and log-SPE.

(Table 2 about here)

4.2. On-line monitoring

The on-line monitoring results are given in Table 3. A normal testing batch is considered to be a false alarm if it is identified as faulty within the batch duration. A missing error means a faulty batch is not detected during the entire duration. The T^2 fails to detect half of the faulty batches because the scores do not conform to a multivariate Gaussian distribution. A comparison between Table 3 (a) and (b) suggests that the instantaneous SPE can detect more faulty batches than the smoothed SPE; however the increased sensitivity is at the cost of dramatically decreased robustness. The number of false alarms for instantaneous SPE is excessively large (15 out of total 20 batches), and thus the smoothed SPE is adopted for the rest of this paper. Table 3 indicates that the GMM approach gives better results than the conventional MPCA in terms of smaller number of false alarms and missing errors.

(Table 3 about here)

It should be noted that the number of missing errors in on-line monitoring is not the only index to evaluate the monitoring performance. Of greater practical importance is the time delay between the occurrence and the detection of the fault. Figure 2 illustrates the detection delay of the 20 faulty batches using MPCA and GMM. To facilitate the calculation of average delay for comparison, the detection delay is artificially set to the batch duration (i.e. 80 s) if a faulty batch is not detected by the monitoring system. Essentially this is to assume that the abnormal behavior will be identified in some way (e.g. the presence of off-specified product) when the batch finishes. In practice plant operators are often not able to identify the fault until much later than the end of batch duration. On average, the detection delay for GMM is 14.0 s that is significantly shorter than 23.5 s obtained by the PCA method. Since the process is operating relatively fast, the reduction of delay in around 10 s (equivalently 10 time steps) may not be sufficient for the operators to take appropriate actions in practice. Nevertheless if the proposed approach is applied to monitor a slow process, for example batch fermentation that takes several days to complete where data is sampled every half day, a shorter detection delay of 10 time steps would provide significant advantage in terms of reduced operational cost and improved process safety and product quality.

(Figure 2 about here)

Figure 3 illustrates the on-line monitoring charts of a normal batch using conventional PCA. Since the value of on-line SPE increases with time, we plot SPE divided by time for better illustration in the figure. The T^2 indicates that this batch is under normal operation; however T^2 is not a reliable index for the monitoring of this process as discussed previously. The SPE metric appears to be susceptible to process disturbance; it exceeds the 99% confidence bound from 18 s, despite the fact that the process is running normally. Figure 4 shows the GMM based monitoring chart, where the negative likelihood value is plotted.

The GMM approach correctly recognizes that this batch is within the region of NOC during the whole batch duration.

(Figures 3 and 4 about here)

Figure 5 and 6 give the on-line monitoring charts of a faulty batch (batch 5 as in Figure 2) using conventional PCA and the GMM approach, respectively. Both T^2 and SPE fail to detect this fault. In contrast, the likelihood value from the GMM is becoming outside the 99% confidence bound since time 2 s.

(Figures 5 and 6 about here)

Finally the contribution analysis is illustrated in Figure 7, where four faulty batches related to abnormal pressure change are investigated. The contribution of each variable is quantified as the difference between the re-calculated likelihood with the variable being missing, L_{-j}, and the original likelihood L. (All likelihood values are in log scale for better illustration.) A large value of $L_{-j} - L$ means that by removing the corresponding variable, the likelihood would significantly increase and the process would move closer to the normal operations. Furthermore, if L_{-j} is greater than the confidence bound h (i.e. $L_{-j} - L > h - L$), the contribution of the corresponding variable can be regarded as substantial, since its elimination would bring the process back to normal. On the contrary, $L_{-j} - L < 0$ means that the removal of the variable does not increase the likelihood, and thus this variable is unlikely to be responsible for the faulty behavior.

(Figure 7 about here)

It can be seen from Figure 7 that the induced pressure fault has resulted in largely similar contribution plots for all the four batches, where the valve position reading (variable 12, see Table 1) is identified as the most responsible in all cases. In the data collection procedure reported in (Wise et al., 1999), if a controlled variable (such as pressure) was moved off its set-point to induce the fault, its values were reset to have the same mean as its nominal value in the data file. The resulting data is more representative of a real process since the data look as if the controller had adjusted the controlled variable to its set-point. Therefore it is not surprising that the pressure measurement (variable 2) is not identified as responsible; rather the valve position that is changed from its nominal value to realize the change in chamber pressure, is contributing the most to the detected fault. Therefore caution must be taken when interpreting the results from contribution analysis. Although contribution analysis provides important information to facilitate fault diagnosis, it does not automatically indicate the root-cause of the process fault. In practice contribution analysis should be combined with other tools, such as expert system and/or pattern recognition techniques, to provide more reliable diagnosis of the process fault.

Another notable finding in this contribution analysis is that although variable 12 appears to be influential, its elimination does not bring the process back to normal operation regions in Figure 7(a)(b). The reason for this phenomenon may be that multiple variables are responsible for the detected fault and thus no single contribution term can exceed the confidence bound. As we discussed in Section 3.2, it is possible to investigate the contribution from a group of variables. However, the number of combinations of variables under analysis increases exponentially with the number of variables. This essentially becomes a combinatorial problem, and it may be addressed using genetic algorithm or other evolutionary computing methods. Currently, this topic is under investigation.

5. Conclusions

This paper extends the GMM technique for the modelling and on-line performance monitoring of batch manufacturing processes. The handling of the unobserved future batch measurements is discussed for the purpose of on-line monitoring. The GMM provides a probabilistic approach to estimating the *pdf* of the nominal process data and therefore enables more accurate calculation of the confidence bounds. Furthermore, a missing-value based contribution analysis method is proposed to facilitate the diagnosis of the detected process fault. The case study confirms that through accurate modelling of the process historical data collected from NOC, GMM is a promising approach to maintaining a low rate of both false alarms and missing errors in process performance monitoring.

A. The false alarm rate when using multiple monitoring statistics

When multiple statistics are used for process monitoring, the usual approach is to define the confidence bounds for each statistic, and the process is identified as faulty when one of the bounds is exceeded. The rationale is that multiple monitoring statistics typically characterize different aspects of the process and thus are complementary. For example, in the application of PCA model, T^2 monitors the covariance structure of process variables, whereas SPE tracks the magnitude of prediction errors.

However, interpretation of the results by using multiple statistics needs some extra care, which may affect the assumed false alarm rate. We first consider two statistics, e.g. T^2 and SPE. Suppose that 95% confidence bound is used, and thus by definition the false alarm rate is 5% for both T^2 and SPE. To facilitate the discussion, we define two random events:

- Event A: the confidence bound of T^2 is exceeded when the process is running normally;
- Event B: the confidence bound of SPE is exceeded when the process is running normally;

and the false alarm rates are $P(A)$ and $P(B)$, respectively. In the most general case these two rates may not be equal, and we assume $P(A) \geq P(B)$. Then, using two statistics simultaneously is equivalent to using a confidence level of

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A|B)P(B) \end{aligned} \tag{11}$$

Since the conditional distribution $P(A|B)$ is between 0 and 1, it can be shown that $P(A) \leq P(A \cup B) \leq P(A) + P(B)$. This result suggests that the false alarm rate will not be less than $P(A)$, which is the larger of the two individual false alarm rates. It is straightforward to generalize the reasoning to more than two statistics. Therefore, the practitioners should be warned that by using both T^2 and SPE, the actual confidence level (or false alarm rate) will change.

The actual value of $P(A \cup B)$ may be approximated in two ways. First, if we can assume that the two events are independent, then $P(A \cap B) = P(A)P(B)$ and thus Eq. (11) can be calculated analytically. The other method is to calculate T^2 and SPE of the nominal process data, and then set $P(A \cup B)$ to be the frequency of either monitoring statistic being exceeded within the nominal data set.

References

Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd Edition. Springer.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Chen, J. H., Liu, K. C., 2002. On-line batch process monitoring using dynamic PCA and dynamic PLS models. Chemical Engineering Science 57, 63–75.

Chen, Q., Kruger, U., Meronk, M., Leung, A. Y. T., 2004. Synthesis of T^2 and Q statistics for process monitoring. Control Engineering Practice 12, 745–755.

Chen, T., Morris, J., Martin, E., 2006. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. Journal of the Royal Statistical Society C (Applied Statistics) 55, 699–715.

Choi, S. W., Park, J. H., Lee, I.-B., 2004. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. Computers and Chemical Engineering 28, 1377–1387.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society B 39, 1–38.

Dunia, R., Qin, S., Edgar, T., McAvoy, T., 1996. Identification of faulty sensors using PCA. AIChE Journal 42, 2797–2812.

Jolliffe, I. T., 2002. Principal Component Analysis, 2nd Edition. Springer.

Kano, M., Nakagawa, Y., 2008. Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. Computers and Chemical Engineering 32, 12–24.

Lee, J.-M., Yoo, C., Lee, I.-B., 2004. Fault detection of batch processes using multiway kernel principal component analysis. Computers and Chemical Engineering 28, 1837–1847.

Lu, N., Yao, Y., Gao, F., 2005. Two-dimensional dynamic PCA for batch process monitoring. AIChE Journal 51, 3300–3304.

Martin, E. B., Morris, A. J., 1996. Non-parametric confidence bounds for process performance monitoring charts. Journal of Process Control 6, 349–358.

Martin, E. B., Morris, A. J., Kiparrisides, C., 1999. Manufacturing performance enhancement through multivariate statistical process control. Annual Reviews in Control 23, 35–44.

Miller, P., Swanson, R. E., Heckler, C. F., 1998. Contribution plots: a missing link in multivariate quality control. International Journal of Applied Mathematics and Computer Science 8, 775–792.

Nomikos, P., MacGregor, J. F., 1995a. Multi-way partial least squares in monitoring batch processes. Chemometrics and Intelligent Laboratory Systems 30, 97–108.

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

Qin, S. J., 2003. Statistical process monitoring: basics and beyond. Journal of Chemometrics 17, 480–502.

Rannar, H., MacGregor, J. F., Wold, S., 1998. Adaptive batch monitoring using hierarchical PCA. Chemometrics and Intelligent Laboratory Systems 41, 73–81.

Safavi, A. A., Chen, J., Romagnoli, J. A., 1997. Wavelet-based density estimation and application to process monitoring. AIChE Journal 43, 1227–1241.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Thissen, U., Swierenga, H., de Weijer, A. P., Wehrens, R., Melssen, W. J., Buydens, L. M. C., 2005. Multivariate statistical process control using mixture modelling. Journal of Chemometrics 19, 23–31.

Wilson, D. J. H., Irwin, G. W., Lightbody, G., 1999. RBF principal manifolds for process monitoring. IEEE Transactions on Neural Networks 10, 1424–1434.

Wise, B. M., Gallagher, N. B., Butler, S. W., White, D. D., Barna, G. G., 1999. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. Journal of Chemometrics 13, 379–396.

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

Yue, H., Qin, S., 2001. Reconstruction based fault identification using a combined index. Industrial and Engineering Chemistry Research 40, 4403–4414.

Table 1: Variables used for the monitoring of the semiconductor process.

1	Endpoint A detector	7	RF impedance
2	Chamber pressure	8	TCP tuner
3	RF tuner	9	TCP phase error
4	RF load	10	TCP reflected power
5	RF Phase error	11	TCP Load
6	RF power	12	Vat valve

Table 2: Off-line monitoring results.

	T^2	SPE	T^2 + SPE	GMM
False alarms	0	1	1	0
Missing errors	15	6	6	3

Table 3: On-line monitoring results. (a) SPE is calculated based on process measurements at current time step (instantaneous SPE); (b) SPE is calculated based on process measurements from batch beginning to current time step (smoothed SPE).

(a)

	T^2	SPE	T^2 + SPE	GMM
False alarms	0	15	15	10
Missing errors	10	0	0	0

(b)

	T^2	SPE	T^2 + SPE	GMM
False alarms	0	4	4	1
Missing errors	10	5	4	2

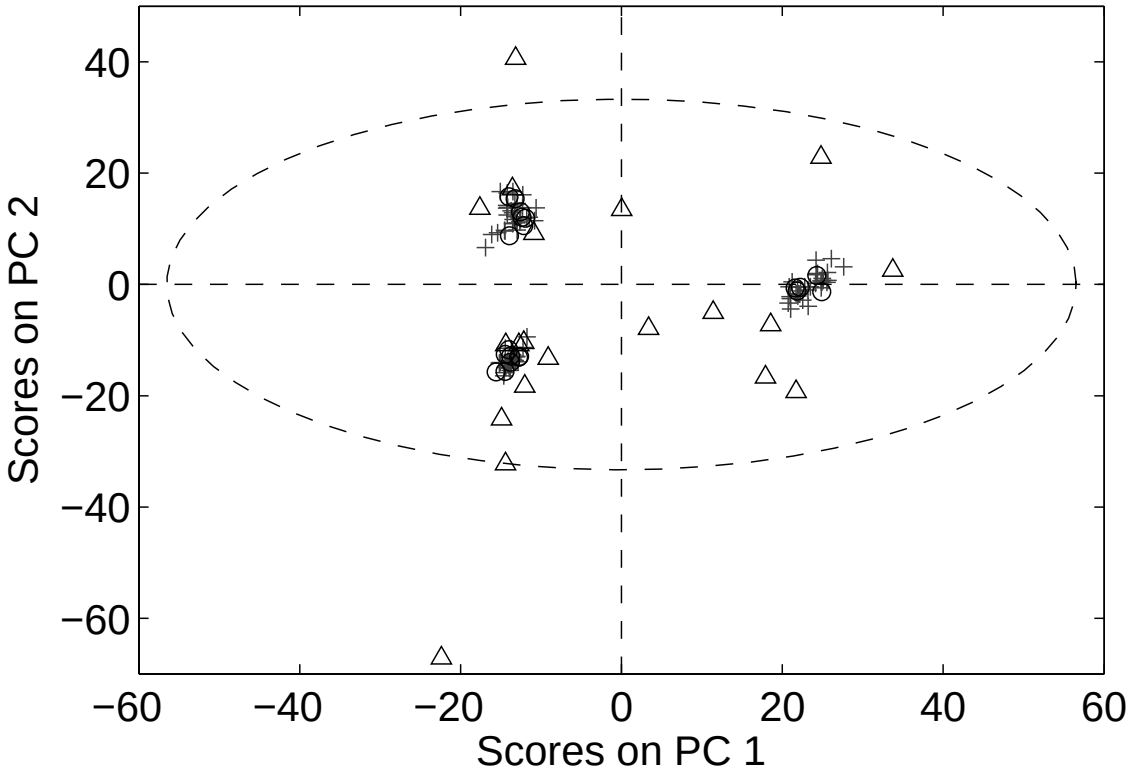

Figure 1: Bivariate scores plot for principal components 1 and 2 with 99% confidence bound (− − −−): nominal (+), normal (∘) and faulty (△).

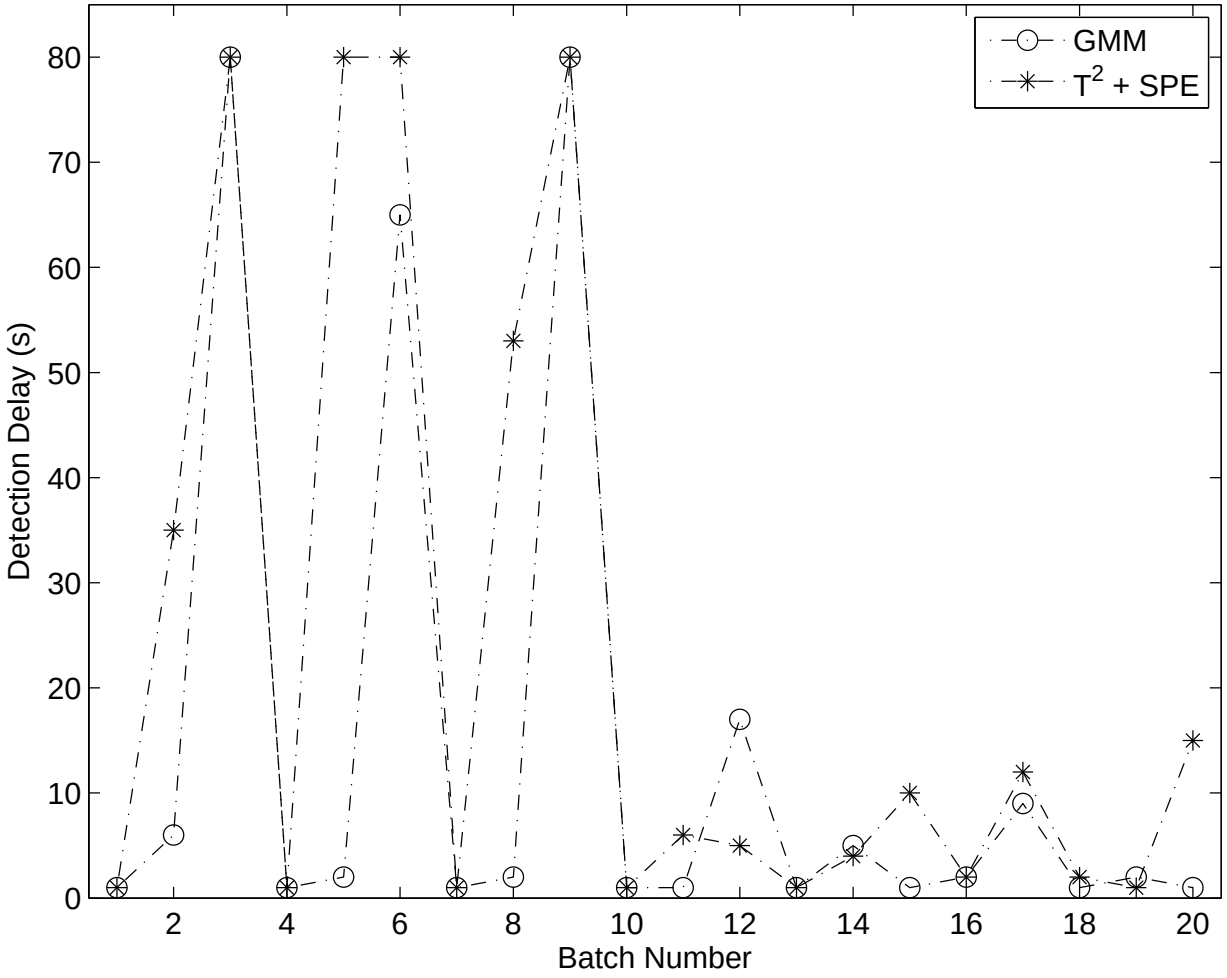

Figure 2: Delay in the detection of the faulty batches.

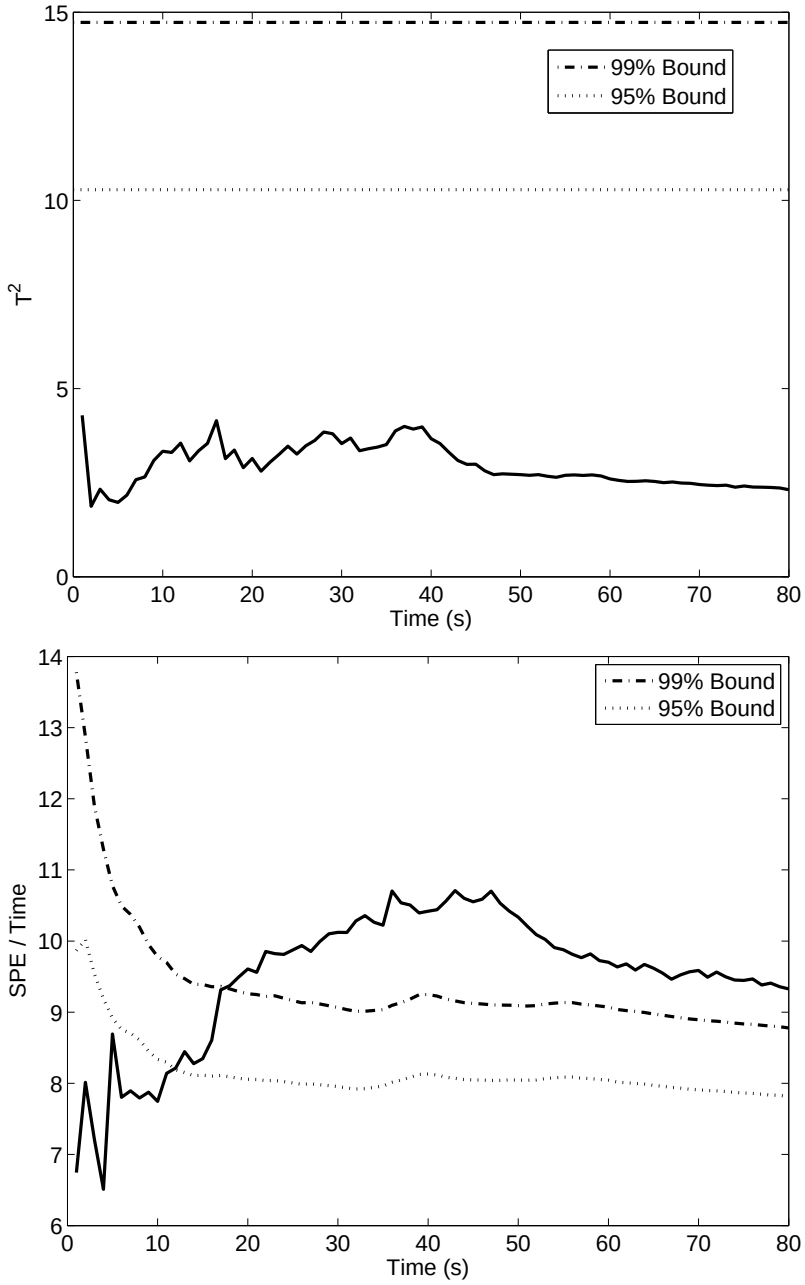

Figure 3: On-line monitoring of a normal batch using T^2 and SPE.

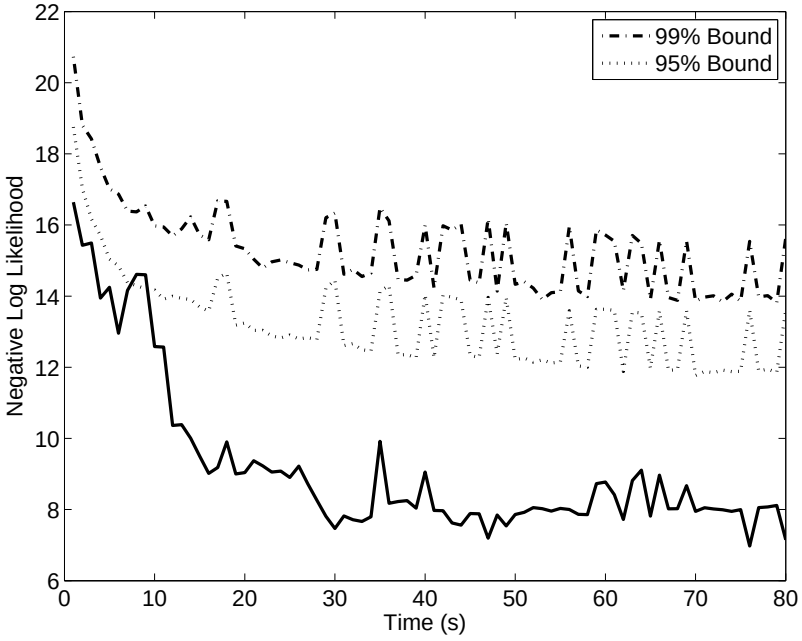

Figure 4: On-line monitoring of a normal batch using GMM.

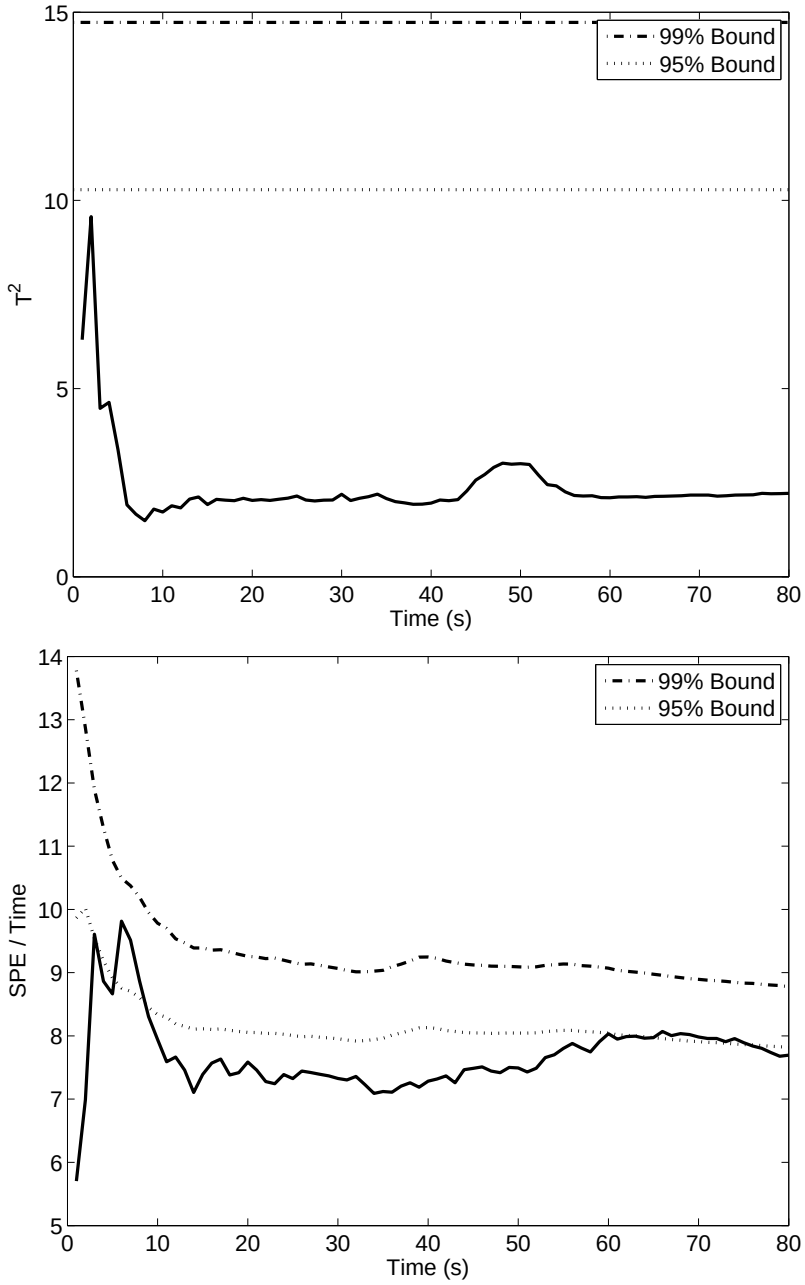

Figure 5: On-line monitoring of a faulty batch using T^2 and SPE.

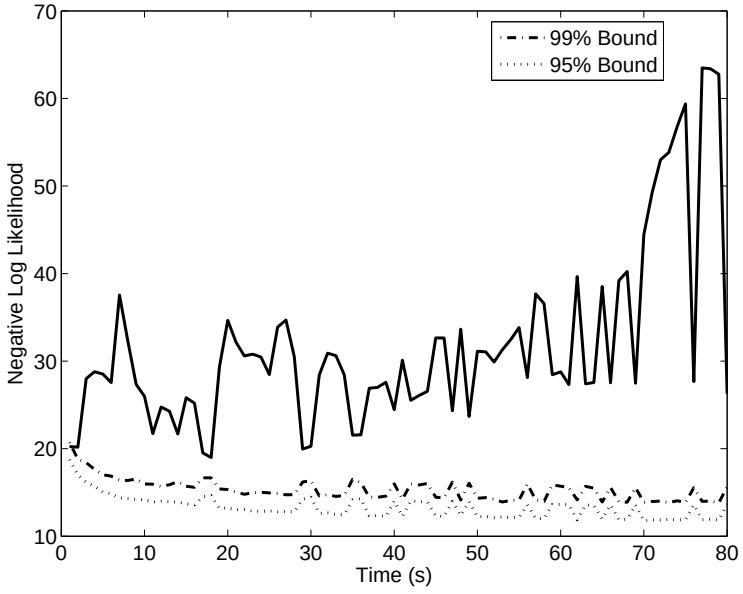

Figure 6: On-line monitoring of a faulty batch using GMM.

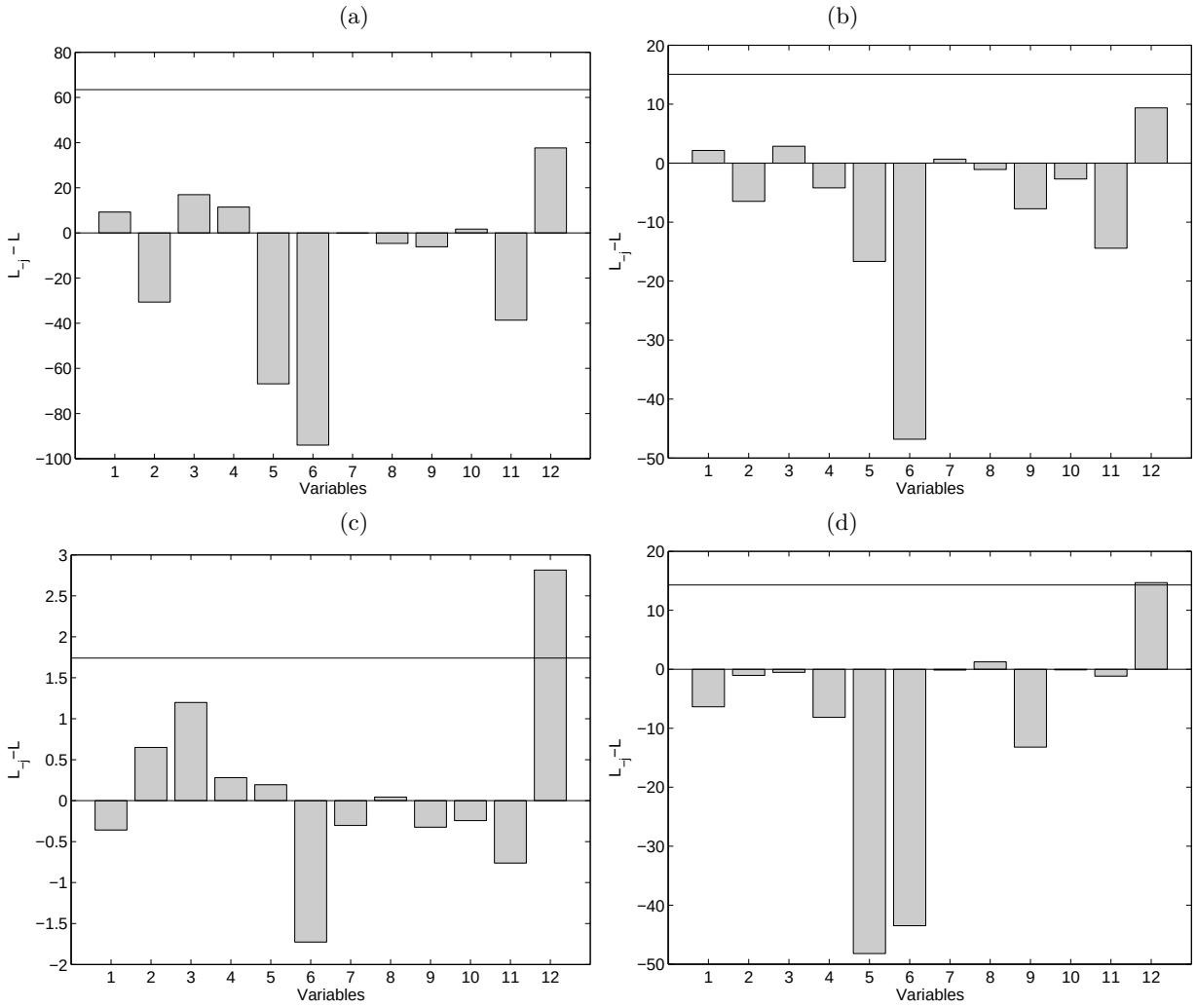

Figure 7: Contribution analysis with 99% confidence bound in terms of $h - L$ (solid line). The faults were induced by moving the chamber pressure away from its set-point: (a)-(c): the pressure is increased; (d) the pressure is decreased.