Probability Density Estimation via Infinite Gaussian Mixture Model: Application to Statistical Process Monitoring

Tao Chen, Julian Morris and Elaine Martin

Centre for Process Analytics and Control Technology,
School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, U.K.

Email: e.b.martin@ncl.ac.uk; Tel. +44 191 222 6231; Fax +44 191 222 5748

Abstract

The primary goal of multivariate statistical process performance monitoring is to identify deviations from normal operation within a manufacturing process. The basis of the monitoring schemes is historical data that has been collected when the process is running under normal operating conditions. This data is then used to establish confidence bounds to detect the onset of process deviations. In contrast to the traditional approaches that are based on the Gaussian assumption, this paper proposes the application of the infinite Gaussian mixture model (GMM) for the calculation of the confidence bounds thereby relaxing the previous restrictive assumption. The infinite GMM is a special case of Dirichlet process mixtures, and is introduced as the limit of the finite GMM, that is when the number of mixtures tends to infinity. Based on the estimation of the probability density function, via the infinite GMM, the confidence bounds are calculated using the bootstrap algorithm. The proposed methodology is demonstrated through its application to a simulated continuous chemical process, and a batch semiconductor manufacturing process.

Key words: Dirichlet process mixtures, Infinite Gaussian mixture model, Markov chain Monte Carlo, Probability density estimation, Multivariate statistical process monitoring.

1 Introduction

The on-line monitoring of the performance of a manufacturing process is essential for ensuring process safety and the delivery of high quality, consistent product. With the rapid development of automatic data collection systems, the effective and efficient utilisation of the large amount of data to characterise the process has become of increasing importance to a wide range of manufacturing industries. In recent years, multivariate statistical projection approaches, such as principal component analysis (PCA) and partial least squares (PLS), have been adopted to extract relevant process information, and to attain an enhanced understanding of process behaviour (Martin et al., 1999; Qin, 2003).

The advantage of PCA and PLS is that as a consequence of the high correlation present between a number of the process measurements, the dimensionality of the original problem can be reduced whilst retaining the information inherent within the data. Statistical process monitoring representations built on the lower order latent variables are observed to be more reliable and robust (Martin et al.,1999; Qin 2003). In addition, by removing those latent variables which only explain a small percentage of the process variability, PCA and PLS can effectively remove the process measurement noise. In practice PLS is a more appropriate tool for describing the process outputs whilst PCA is applicable where it is the process itself, and its behaviour, that is of interest. Additionally, in practice, response variables are usually measured off-line and thus a time delay is incurred before they become available for inclusion in the monitoring scheme, and thus, unless inferential measurement is implemented, PLS is not an appropriate tool for process performance monitoring. For the two case studies described in this paper, the response variables are not available and hence, the focus of this paper is PCA.

Consider the case where *N* data points, $\{\mathbf{z}_n, n = 1, \cdots, N\}$, are collected when the process is running under normal operating conditions (NOC) and let $\mathbf{z}_n$ be a vector consisting of *D*-dimensional process variables. Typically the data will be pre-processed to zero mean and unit standard deviation on each dimension. The first step in PCA is to compute the sample covariance matrix, **S,** of order $D \times D$. The eigenvectors $\mathbf{u}_i$ and eigenvalues g_i of **S** are then calculated (i = 1, …, D). By retaining those eigenvectors corresponding to the q largest eigenvalues, the q-dimensional PCA score vectors, $\mathbf{t}_n$, are calculated through the linear projection of the original data onto the space spanned by the q eigenvectors: $\mathbf{t}_n = \mathbf{U}_q^{\mathrm{T}} \mathbf{z}_n$, where $\mathbf{U}_q = (\mathbf{u}_1, \ldots, \mathbf{u}_q)$. Therefore the original data can be represented as a linear combination of the scores plus a residual vector, $\mathbf{e}_n$:

$$\mathbf{z}_n = \mathbf{U}_q \, \mathbf{t}_n + \mathbf{e}_n \tag{1}$$

Consequently normal process behaviour can be characterised by the first q principal components, which capture the main sources of data variability.

In process performance monitoring, the next step is to define the confidence bounds, i.e. the thresholds that determine the normal operating region for the PCA representation determined from the process data. Traditionally two metrics are calculated to monitor the behaviour of a process, Hotelling's T^2 and the squared prediction error (SPE). Hotelling's T^2 is defined as the sum of the normalized squared scores: $T_n^2 = \mathbf{t}_n^{\mathrm{T}} \mathbf{\Lambda}^{-1} \mathbf{t}_n$, where $\mathbf{\Lambda}$ is a diagonal matrix comprising the q largest eigenvalues. The SPE is given by $r_n = \mathbf{e}_n^T \mathbf{e}_n$. Assuming the PCA scores follow a Gaussian distribution, the confidence bounds can be established for Hotelling's T^2 (Hotelling, 1947), and the SPE (Jackson and Mudholkar 1979).

However the assumption that the scores are Gaussian distributed when calculating the confidence bounds may be an invalid assumption particularly when the data is collected

from a complex manufacturing process. For example, when non-linear projection techniques are used to characterise the process (Shao et al., 1999; Wilson et al., 1999), the resulting distribution of the latent variables will typically not be multivariate Gaussian. To address this issue, a number of techniques have been proposed to estimate the probability distribution function (pdf) of the PCA scores directly, for example kernel density estimation (KDE) (Martin and Morris, 1996), where it was clearly shown that the PCA scores did not follow a Gaussian distribution. Martin and Morris (1996) focussed on bivariate monitoring plots since KDE is more challenging to implement in higher dimensional space due to the so-called *curse of dimensionality* phenomenon. That is, with increasing dimensionality, the data points become more sparsely distributed in the data space. A number of semi-parametric models have been proposed to alleviate this problem, for example, wavelet based density estimation (Safavi et al., 1997), and the Gaussian mixture model (GMM) (Choi et al., 2004; Thissen et al., 2005).

A second issue is that two separate metrics, Hotelling's T^2 and SPE, are required to monitor the performance of a process. In practice a heuristic method is adopted, whereby the operation of the process is observed to have changed from normal process operation if either Hotelling's T^2 or the SPE metric moves outside the confidence bounds. A number of techniques have been proposed to combine the two metrics into a unified statistic, for example through a weighted sum of the two metrics (Al-Alawi et al., 2005), and through kernel density estimation (Chen et al., 2004). More importantly, in practice a single monitoring metric will reduce the work load of plant operators as they will only be exposed to one monitoring chart. This is crucial for the wider acceptance of statistical process monitoring techniques in industry.

This paper proposes the application of the infinite Gaussian mixture model (IGMM) for the estimation of the probability density function for Hotelling's T^2 and the SPE. By increasing the number of mixtures in the GMM to infinity, the IGMM removes the obstacle of selecting the number of mixtures, which is a real issue with respect to the applicability of the methodology. In addition, rather than summarising the PCA representation through two metrics, Hotelling's T^2 and the SPE, the IGMM is capable of estimating the joint probability distribution of the PCA scores and the log-SPE[1] (i.e. the pdf of a q+1 dimensional vector $(\mathbf{t}_n, \log(r_n))^{\mathrm{T}}$. By adopting this approach, a unified likelihood based statistic can be constructed for process performance monitoring. Finally after the probability density function has been estimated, the confidence bounds for the unified likelihood based statistic are identified using the bootstrap. The proposed approach is demonstrated through its application for the monitoring of a simulated continuous chemical process, and a batch semiconductor manufacturing process.

2 Infinite Gaussian mixture model

This section introduces the infinite Gaussian mixture model which is subsequently used as a tool to estimate the joint pdf of the PCA scores and the log-SPE, that have been calculated from normal process operating data. The infinite GMM belongs to the family

[1] IGMM is not suitable to estimate the pdf of non-negative random variables directly, such as the SPE. Hence the logarithm operator is used to transform the SPE onto the whole real axis.

of Dirichlet process mixtures (Blackwell and MacQueen, 1973; Ferguson, 1973), and can be derived in a number of different ways. A comprehensive discussion of alternative perspectives on the Dirichlet process mixtures can be found in Teh et al. (2004). Within this paper, the concept is introduced through the finite Gaussian mixture model, whose mixing weight is given by a Dirichlet process prior. The infinite Gaussian mixture model is then derived by demonstrating that it is basically the situation where the number of mixtures tends to infinity. The inference of the infinite GMM parameters is implemented using Markov chain Monte Carlo (MCMC) methods.

2.1 Finite Gaussian mixture model

The probability distribution function of the data, *x*, can be modelled by a finite mixture of Gaussian distributions with *k* components:

$$p(x \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) = \sum_{j=1}^{k} \pi_j G(\mu_j, \tau_j^{-1}) \tag{2}$$

where $\boldsymbol{\mu} = \{\mu_1, \cdots, \mu_k\}$, $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_k\}$ and $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ are the means, precisions (inverse variances) and mixing weights (which must be positive and sum to unity), respectively. For notational simplicity, the data are assumed to be scalar. The extension to the multivariate case is presented in Appendix B.

Given a set of training data with *N* observations, $\mathbf{x} = \{x_1, \cdots, x_N\}$, the classical approach to estimating the Gaussian mixture model parameters, ($\boldsymbol{\mu}$, $\boldsymbol{\tau}$,$\boldsymbol{\pi}$), is to maximize the likelihood using the expectation-maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm guarantees convergence to a local maximum, with the quality of the maximum being heavily dependant on the random initialization of the algorithm. Alternatively, a Bayesian approach can be used to combine the prior distribution for the parameters and the likelihood, resulting in a joint posterior distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi} \mid \mathbf{x}) \propto p(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})\, p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) \tag{3}$$

However the joint posterior takes a highly complicated form. Thus it is generally not feasible to perform any analytical inference based on the above posterior distribution. MCMC approaches have typically been used to calculate the joint posterior and of the approaches proposed in the literature, Gibbs sampling is suitable for mixture models. To generate samples from the posterior distributions, Gibbs sampling updates each parameter (or a group of parameters) in turn from its conditional posterior distribution.

The rest of this section focuses on the definition of the priors and the derivation of the conditional posteriors for the GMM parameters. To facilitate the derivation, the latent indicator variable, $\mathbf{c} = \{c_1, \cdots, c_N\}$, is introduced to identify that a specific data point, x_n, belongs to mixture component, c_n. The approach is based primarily on the formulation proposed in Rasmussen (2000).

Conditional posterior distribution of the component means

The mean of each mixture component is given a Gaussian prior: $p(\mu_j \mid \lambda, \gamma) \sim G(\lambda, \gamma^{-1})$, where λ and γ are hyper-parameters that are common to all components. The conditional posterior distribution for μ_j is calculated by multiplying the prior, $p(\mu_j \mid \lambda, \gamma)$, by the likelihood (Eq. (2)), resulting in a Gaussian distribution:

$$p(\mu_j \mid \mathbf{c}, \mathbf{x}, \tau_j, \lambda, \gamma) \sim G\left(\frac{\bar{x}_j N_j \tau_j + \lambda\gamma}{N_j \tau_j + \gamma}, \frac{1}{N_j \tau_j + \gamma}\right) \tag{4}$$

where $\bar{x}_j$ and N_j are the mean and number of data points belonging to component j, respectively. The selection and updating of the hyper-parameters, including those defined in the subsequent sub-sections for the component precisions and the mixing weights, is discussed in Appendix A.

Conditional posterior distribution of the component precisions

Each component precision (inverse variance) is given a Gamma prior with hyper-parameters β and ω: $p(\tau_j \mid \beta, \omega) \sim Ga(\beta, \omega^{-1}) \propto \tau_j^{\beta/2-1} \exp(-\tau_j \omega\beta/2)$. Similarly the conditional posterior for s_j is attained by taking the product of the prior, $p(\tau_j \mid \beta, \omega)$, and the likelihood, resulting in a Gamma distribution:

$$p(\tau_j \mid \mathbf{c}, \mathbf{x}, \mu_j, \beta, \omega) \sim Ga\left(\beta + N_j, \left[\frac{\omega\beta + \sum_{i:c_i=j}(x_i - \mu_j)^2}{\beta + N_j}\right]^{-1}\right) \tag{5}$$

Conditional posterior distribution of the mixing weights

The inference of the mixing weights is more complex than the other two parameters. The mixing weights are given symmetric Dirichlet priors with concentration parameter α/k:

$$p(\pi_1, \cdots, \pi_k \mid \alpha) \sim \text{Dirichlet}(\alpha/k, \cdots, \alpha/k) \tag{6}$$

Utilising the definition of the indicator variable, the inference of the mixing weights can be indirectly realized through the inference of the indicators, whose joint conditional distribution is given by:

$$p(c_1, \cdots, c_N \mid \pi_1, \cdots, \pi_k) = \prod_{j=1}^{k} \pi_j^{N_j} \tag{7}$$

By integrating out the mixing weights as a result of the properties of the Dirichlet integral (Ferguson, 1973), the prior for the indicators is only dependent on α:

$$p(c_1,\cdots,c_N \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(N_j + \alpha/k)}{\Gamma(\alpha/k)} \tag{8}$$

where Γ(.) is the standard Gamma function. The conditional prior for a single indicator, given all the other indicators, is obtained as follows:

$$p(c_n = j \mid \mathbf{c}_{-n}, \alpha) = \frac{N_{-i,j} + \alpha/k}{N - 1 + \alpha} \tag{9}$$

where $\mathbf{c}_{-n} = \{c_1,\cdots,c_{n-1},c_{n+1},\cdots,c_N\}$, and $N_{-n,j}$ is the number of data points, excluding x_n, which belongs to mixture j. The likelihood of x_n belonging to component j is: $p(\mathrm{x_n} \mid c_n = j, \mu_j, \tau_j) = G(\mu_j, \tau_j^{-1}) \propto \tau_j^{1/2} \exp(-\tau_j (x_n - \mu_j)^2 / 2)$. Calculating the product of the prior and the likelihood, the conditional posterior for each c_n is given by:

$$p(c_n = j \mid \mathbf{c}_{-n}, \mu_j, \tau_j, \alpha) \propto \frac{N_{-n,j} + \alpha/k}{N - 1 + \alpha} \tau_j^{1/2} \exp(-\tau_j (x_i - \mu_j)^2 / 2) \tag{10}$$

2.2 Infinite Gaussian mixture model

The previous discussions have been restricted to a finite number of mixtures. The selection of the appropriate number of mixtures is a real issue in practical applications. The likelihood of the data will be a maximum when the number of mixtures is equal to the number of training data points, which results in “over-fitting". One solution is to utilise validation or cross-validation, which selects the number of mixtures by simultaneously maximizing the likelihood over the training and validation data sets. Other well-known model selection criteria include Akaike information criterion (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978).

Bayesian methodology addresses the over-fitting problem by assigning a prior distribution over the number of mixtures, or as in this paper, through the placement of a Dirichlet prior over the mixing weights, this is then combined with the likelihood to give the posterior distribution for inference. In the Bayesian statistics literature, the selection of the number of mixtures has been addressed through a number of different MCMC strategies, including reversible jump MCMC (Richardson and Green, 1997) and birth-death MCMC (Stephens, 2000). This paper adopts a perspective from the Dirichlet process whereby an infinite number of mixtures is utilised. When the number of mixtures tends to infinity, there must be an infinite number of mixtures with no training data associated with them. These are termed “unrepresented” mixtures. Correspondingly, “represented mixtures” are those that have training data associated with them.

Let k_{rep} denote the number of represented mixtures. For represented mixtures, the previously derived conditional posteriors of μ_j and τ_j still hold (Eq. (4) and (5)). On the other hand, in the absence of training data, the parameters in unrepresented mixtures are solely determined by their priors ($p(\mu_j \mid \lambda, \gamma)$ and $p(\tau_j \mid \beta, \omega)$). Thus the inference of the indicators, **c**, has to incorporate the effect of infinite mixtures. Therefore letting $k \to \infty$ in Eq. (9), the conditional prior of c_n will give the limits:

$$p(c_n = j \mid \mathbf{c}_{-n}, \alpha) = \begin{cases} \dfrac{N_{-n,j}}{N-1+\alpha} & j \text{ is represented} \\ \dfrac{\alpha}{N-1+\alpha} & j \text{ is unrepresented} \end{cases} \tag{11}$$

To obtain the posterior probability of the indicators, the likelihood must be calculated. The likelihood of x_n belonging to a represented component j, takes the same form as for the finite Gaussian mixture model. Since the infinite number of unrepresented mixtures are determined by the prior, the likelihood of x_n being associated with them is an integral over the prior: $\int p(x_n \mid \mu_j, \tau_j) p(\mu_j \mid \lambda, \gamma) p(\tau_j \mid \beta, \omega) d\mu_j d\tau_j$. In summary, the conditional posteriors of the indicator variables are as follows:

$$p(c_n = j \mid \mathbf{c}_{-n}, \alpha) \propto \begin{cases} \dfrac{N_{-n,j}}{N-1+\alpha} \tau_j^{1/2} \exp(-\tau_j (x_n - \mu_j)^2 / 2) & j \text{ is represented} \\ \dfrac{\alpha}{N-1+\alpha} \displaystyle\int p(x_n \mid \mu_j, \tau_j) p(\mu_j \mid \lambda, \gamma) p(\tau_j \mid \beta, \omega) d\mu_j d\tau_j & j \text{ is unrepresented} \end{cases} \tag{12}$$

The above equation states that the training data has a certain probability of being associated with (an infinite number of) unrepresented mixtures and each represented mixture. If, in one sampling iteration, some data points are associated with unrepresented mixtures, new represented mixtures will emerge. On the other hand, if all the data points pertaining to a represented mixture are associated with other mixtures, this mixture will become unrepresented. By adopting this approach, k_{rep} will be determined according to the posterior distribution of the parameters.

2.3 Monte Carlo sampling

Based on the conditional posteriors developed in the preceding sub-sections, one iteration of Gibbs sampling is executed as follows:

1. For n = 1:N
   Sample indicators c_n, are generated according to Eq. (12).
   End.
2. Update k_{rep}, the number of represented mixtures.
3. For j = 1 : k_{rep}
   Update N_j, the number of data points belonging to mixture j.

Update mixing weights: $\pi_j = N_j / (N + \alpha)$.

End.

Update the overall mixing weight of unrepresented mixtures: $\pi_{\cdot} = \alpha / (N + \alpha)$.

4. For $j = 1 : k_{\text{rep}}$

   Sample $\mu_j \sim p(\mu_j \mid \mathbf{c}, \mathbf{x}, \tau_j, \lambda, \gamma)$ (Eq. (4)).

   Sample $\tau_j \sim p(\tau_j \mid \mathbf{c}, \mathbf{x}, \mu_j, \beta, \omega)$ (Eq. (5)).

   End.

5. Update hyper-parameters (Appendix A):

   Sample $\lambda \sim p(\lambda \mid \boldsymbol{\mu}, \gamma)$ (Eq. (14)).

   Sample $\gamma \sim p(\gamma \mid \boldsymbol{\mu}, \lambda)$ (Eq. (15)).

   Sample $\omega \sim p(\omega \mid \boldsymbol{\tau}, \beta)$ (Eq. (16)).

   Sample $\beta \sim p(\lambda \mid \boldsymbol{\tau}, \omega)$ (Eq. (17)).

   Sample $\alpha \sim p(\alpha \mid k_{\text{rep}}, N)$ (Eq. (18)).

The conditional posteriors of μ_j and τ_j are Gaussian and Gamma distributions respectively, from which samples can be generated using standard procedures. The sampling of the indicators requires the evaluation of the integral in Eq. (12), which is only analytically feasible if the conjugate prior is used (for example, the Gaussian-Inverse-Gamma prior for the joint distribution of μ_j and τ_j^{-1}). This paper follows the approach proposed by Rasmussen (2000) whereby independent priors are assigned to μ_j and τ_j^{-1} respectively, and these are not conjugate to the likelihood. To approximate this integral using a Monte Carlo approach, Neal (1998) proposed generating samples of (μ_j, τ_j) from their prior. This strategy is adopted in this paper. Further details are given in Algorithm 8 of Neal (1998). Alternative sampling methods have also been proposed in the literature (MacEachern and Muller, 1998; Walker and Damien, 1998).

2.4 Prediction

The calculation of the predictive probability of new data will be averaged over a number of MCMC samples, which are selected from those where the algorithm tends to stabilize. Stabilization will be assessed heuristically based on the value of the log-likelihood. Additionally to eliminate the auto-correlation, one sample will be selected from each consecutive set of 10 iterations.

For a particular MCMC sample, the predictive probability is attained from two components: the represented and the unrepresented mixtures. In a similar manner to that adopted in the sampling stage, the probability from unrepresented mixtures will be approximated by a finite mixture of Gaussians, whose parameters, (μ_j, τ_j), are drawn from the prior.

3 Confidence bounds

Once the probability distribution has been derived that reflects normal process operation, confidence bounds, i.e. action and warning limits, are required to identify any departure of the process from nominal behaviour. For example, a confidence bound of 100b% ($0<b<1$) defines a region that encompasses 100b% of the nominal process data as the sample size tends to infinity. A process is classified as statistically deviating from normal behaviour when new data, superimposed on the nominal representation, lies outside the nominal region. Dependent on the confidence level, b, two types of errors are potentially present: false alarms (a normal data point is classified as faulty) and missing errors (failure to observe a non-conforming data point). A small value of b would result in an unacceptable number of false alarms, whilst a confidence level close to one would fail to identify the onset of process faults in a timely and acceptable manner. In practice in process performance monitoring, the confidence level is normally assumed to be 0.99 for the action limit and 0.95 for the warning limit.

Based on the probability distribution $p(x \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$, the 100$b$% confidence bound can be defined as a likelihood threshold, h, that satisfies the following integral:

$$\int_{x:p(x)>h} p(x \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) dx = b \tag{13}$$

Hence a new data point, x^*, is identified as non-conforming if $p(x^* \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) < h$. For the infinite Gaussian mixture model, the above integral is not analytically tractable and therefore it is not possible to obtain the threshold directly. One possible solution is to approximate this integral by generating Monte Carlo samples from the probability distribution function:

1. Generate M samples, x_i, $i = 1, \ldots, M$, from $p(x \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$.
2. Calculate the likelihood of these samples as $p(x_i \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$.
3. Sort $p(x_i \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$ in descending order.
4. The confidence bound is given by $h = p(x_{\text{lim}} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$, where lim=$Mb$.

The issue with this approach is that as the model parameters are averaged over a number of MCMC iterations, the resultant probability density is relatively smooth with a heavy tail. Therefore the confidence bound may be smaller in magnitude than required, and thus will fail to identify non-conforming process behaviour. A more robust approach is to use the bootstrap (Efron, 1981). First a large number of samples, say 1000, are drawn with replacement from nominal process data. Then these samples are used to calculate a confidence bound following the algorithm described above. The procedure is repeated a number of times (e.g. 100) and an averaged value is obtained for the confidence bound.

4 Case studies

This section applies the proposed approach to the monitoring of two manufacturing processes. The first example is that of the simulated Tennessee Eastman continuous stirred tank reactor which was presented in Downs and Vogel (1993) as a benchmark for testing new methodologies in advanced process control and process performance monitoring. The second process is a batch semiconductor etch process that comprised three modes of operation (Wise et al., 1999). This data set is publicly available from Eigenvector Research, Inc. (http://software.eigenvector.com/Data/Etch/index.html).

4.1 Tennessee Eastman continuous process

The Tennessee Eastman process comprises a set of unit operations (reactor/separator/stripper/compressor) with two simultaneous exothermic reactions and two by-product reactions. In this study, the simulation software is run with a decentralized control strategy (Ricker, 1996). The process has 12 manipulated variables and 41 measurements. However a number of the quality measurements, such as product concentration, are only available infrequently in industrial scale plant and hence were removed from the analysis. Thus the final data set that was used to build the model comprised 22 measurements, plus 12 manipulated variables. The details of these 34 variables can be found in Downs and Vogel (1993). The sampling interval was 0.02 hrs.

The process was initially run for 20 hours under normal operating conditions, giving 1000 data points. The first 500 points were selected to define the nominal operating region, and the remaining 500 data points were reserved to assess the false alarm rate. The process was then run under process conditions that simulated faulty behaviour. A total of four faults were considered. In all cases the faults were introduced by adding a disturbance to the process manipulated variables, and/or by simulating a device malfunction (Table 1). The specific details of the faults are discussed in Downs and Vogel (1993). For each fault scenario the process was run under abnormal behaviour for a further 6 hours, giving 300 faulty data points. From previous analysis that have been reported in the literature, it is acknowledged that fault “IDV(1)” results in a direct step change in two process measurements, and thus is relatively easy to detect. In contrast fault “IDV(14)” is more subtle as it disturbs the reactant temperature which was not directly measured. Finally “IDV(12+15)” is the most complicated, as it reflects the simultaneous onset of two faults, a disturbance in an unmeasured variable and a device failure and thus it is extremely challenging to detect.

(Table 1 about here.)

PCA was performed on the nominal data set (500 data points) and the dimensionality of the problem was reduced to 12 principal components, which explained 70.2% of the total variance. One thousand iterations were performed from which the infinite Gaussian mixture model parameters were sampled thereby enabling the estimation of the joint pdf of the PCA scores and the log-SPE extracted from the nominal data. Based on the log-likelihood, the algorithm tended to stabilize after the first 500 iterations. Figure 1 (a) shows the number of represented mixtures (k_{rep}) versus the number of MCMC iterations. The frequency of k_{rep}, computed from the final 500 iterations, is illustrated in Figure 1(b). Both figures show that approximately 15 to 30 mixtures were automatically inferred from the data. Of the final 500 iterations, one sample was

selected from each consecutive set of 10 iterations, resulting in a total of 50 samples being selected. The probability of the data points was then calculated based on an average over these 50 samples. The bootstrap technique described in Section 3 was then used to determine the 99% confidence bound.

(Figure 1 about here.)

The process monitoring charts for fault IDV(12+15), introduced at time point 20 hrs, are shown in Figures 2 and 3. Figure 2 illustrates the use of the traditional confidence bounds for Hotelling's T^2 and SPE. It can be seen that Hotelling's T^2 is not sensitive to this fault in the initial stage with the process being identified as normal prior to 22 hrs, that is a detection delay of 2 hrs. The SPE statistic, in Figure 2(b), is capable of identifying this fault at approximately 20.8 hrs. Figure 3 shows the case where the confidence bounds using the infinite GMM approach are considered. To ensure a fair comparison with Hotelling's T^2, Figure 3(a) was obtained by only estimating the pdf of the PCA scores when calculating the confidence bound. In this case, the process abnormality is detected at approximately 21 hrs, significantly more rapidly than when Hotelling's T^2 was applied. When the joint pdf of the PCA scores and the log-SPE was estimated using the infinite GMM (Figure 3(b)), the obtained confidence bound provides the best result, detecting the onset of the fault at around 20.4 hrs.

(Figure 2 and Figure 3 about here.)

Table 2 examines two types of potential errors: false alarm and missing error, and summarizes the results in terms of error rates (number of errors divided by number of test data points), for different fault scenarios. For the traditional confidence bound, the data is classified as faulty if it exceeds the bound either for Hotelling's T^2 or the SPE. The false alarm rate for both the traditional approach and the infinite Gaussian mixture model are close to 1%. This is consistent with the concept of the 99% confidence bound, which states that statistically 1% of normal operating data will fall outside this bound. Since fault IDV(1) results in a dramatic change in the magnitude of the process variables, it is relatively easy to identify and thus has a low missing error rate for both methods. For the other three faults, the infinite Gaussian mixture model is consistently superior to Hotelling's T^2 and the SPE, in terms of lower missing error rates.

(Table 2 about here.)

Figure 4 shows the quantile-quantile (Q-Q) plots for the PCA scores of the nominal process data versus the standard Gaussian distribution. If the PCA scores are distributed as univariate Gaussian, the Q-Q plots would be linear. This is not the case, especially for the scores corresponding to the largest two eigenvalues. It is known that if the data is not normal in a univariate sense, it will also not be normally distributed in the multivariate case. Therefore the Gaussian assumption that underpins the construction of the confidence bounds for Hotelling's T^2 and SPE is indeed problematic and needs to be addressed to ensure effective process performance monitoring.

(Figure 4 about here.)

4.2 A batch semiconductor process

The manufacture of semiconductors is introduced as an example of the monitoring of batch processes. Although there are many steps in this process, 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 3, was collected during the wafer processing stage which was of 80 second duration. A sampling interval of 1 second was used in the analysis. Thus for each batch, the data is of the order (80 x 12). 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 one batches, seven from each group, were selected from the normal batches to investigate the false alarm rate. The remaining 86 nominal batches were used to build the nominal PCA representations.

(Table 3 about here.)

To analyse three dimensional data ($N_{batch} \times N_{variable} \times N_{time}$), "multi-way" analysis methods have been proposed to "unfold" the three-dimensional array into a two-dimensional matrix and conventional PCA is then applied to the unfolded data matrix (Nomikos and MacGregor, 1994). This study unfolds the data array ($N_{batch} \times N_{variable} \times N_{time}$) into a large two-dimensional matrix ($N_{batch} \times N_{variable}\ N_{time}$) on which PCA is performed. It was observed that the initial three principal components explain 32.8%, 12.9%, and 2.7% of the total variance, respectively, which supports the selection of only 2 principal components. In a similar manner to the previous example, MCMC sampling was performed for one thousand iterations and again it tended to stabilize after 500 iterations. The probability of the data was obtained based on an average being calculated over 50 samples selected from the final 500 iterations, with one sample being selected from each consecutive set of 10 iterations. The bootstrap technique was again used to calculate the 99% confidence bound.

The PCA scores plot of the process data is shown in Figure 5, where the contours of the 99% confidence bounds were defined using the infinite Gaussian mixture model and the standard Gaussian based approach of Hotelling's T^2. The multi-modal property in this data set invalidates the underlying Gaussian assumption with respect to the traditional confidence bounds. Therefore the global Hotelling's T^2 metric fails to identify many of the non-conforming data points. On the other hand, the infinite Gaussian mixture model approach provides a more appropriate confidence bound that identifies the non-conforming batches, and effectively recognizes the distinct clusters in the data.

(Figure 5 about here.)

It could be argued that the "local" modelling strategy of Wise et al., (1999), which calculates Hotelling's T^2 for each local group, can address the multi-modal problem. However the determination of the number of clusters is still an issue. An alternative

approach would be to utilise a finite GMM using the EM algorithm to estimate the joint pdf of the PCA scores and the log-SPE. Again it is necessary to identify the appropriate number of mixtures. For comparison, both Bayesian information criterion (BIC) and cross validation were used to determine the number of mixtures in the Gaussian mixture model. Figure 6 shows the BIC value and the log likelihood of 5-fold cross validation with different numbers of mixtures, where both criteria indicate that a Gaussian mixture model with 3 mixtures achieves the largest likelihood. Using 3 mixtures appears to be an optimal choice as there are 3 distinct groups in the data. However, Table 4 shows that a Gaussian mixture model with 3 mixtures results in 3 false alarms and 5 missing errors. In contrast, the infinite Gaussian mixture model incurs only 1 false alarm and 2 missing errors. This result implies that, even if the intuitively 'correct' number of mixtures (clusters) is determined, each local cluster may not be adequately modelled by one Gaussian distribution. This result justifies the application of the infinite Gaussian mixture model which automatically selects approximately 6 to 9 represented mixtures during the MCMC iterations, in this example.

(Table 4 about here.) (Figure 6 about here.)

5 Conclusions and discussions

This paper introduces the infinite Gaussian mixture model as a tool for calculating confidence bounds for statistical process performance monitoring. Although previous research has focused on extracting information from multivariate process data for monitoring process performance, many algorithms still rely on the Gaussian assumption to build the confidence bounds for both Hotelling's T^2 and SPE for the calculated principal components. The infinite Gaussian mixture model provides a Bayesian approach to estimating the probability density function of the nominal process data, and therefore enables the more accurate calculation of the confidence bounds.

Furthermore, the infinite Gaussian mixture model is capable of combining the principal component scores and the log-SPE into a unified likelihood based statistic to provide improved and more simplistic process monitoring results. The proposed framework was evaluated on a simulation of an industrial continuous process and a batch manufacturing process of semiconductors. Promising results were achieved. The proposed approach can be applied to other multivariate statistical projection techniques, by estimating the joint probability distribution of all possible source of information.

Acknowledgments

T. Chen would like to acknowledge the financial support of the EPSRC KNOW-HOW (GR/R19366/01) and Chemicals Behaving Badly II (GR/R43853/01), and the UK ORS Award for his PhD study.

Appendix A: Updating hyper-parameters

The selection of the hyper-parameters that determine the prior distributions of the infinite GMM parameters has an important impact on the inference of these parameters. Given hyper-priors, the hyper-parameters can also be updated. This hierarchical structure tends to be more robust than the approach whereby the hyper-parameters are simply selected. The updating of the hyper-parameters requires the derivation of their conditional posterior distributions. This aspect is presented below.

The hyper-parameters for the component means, λ and γ, are given vague Gaussian and Gamma hyper-priors[2]: $p(\lambda) \sim G(\mu_x, \sigma_x^2)$, where μ_x and σ_x^2 are the mean and variance of the training data respectively. The shape parameter of the Gamma hyper-prior is set to unity, corresponding to a vague distribution. The conditional posterior for λ and γ are obtained by calculating the product of the hyper-priors and $\prod_{j=1}^{k_{rep}} p(\mu_j \mid \lambda, r)$, and can be simplified to give:

$$p(\lambda \mid \boldsymbol{\mu}, \gamma) \sim G\left(\frac{\mu_x \sigma_x^{-2} + \gamma \sum_{j=1}^{k_{rep}} \mu_j}{\sigma_x^{-2} + k\gamma}, \frac{1}{\sigma_x^{-2} + k\gamma} \right) \tag{14}$$

$$p(\gamma \mid \boldsymbol{\mu}, \lambda) \sim Ga\left(k+1, \left[\frac{\sigma_x^2 + \sum_{j=1}^{k_{rep}} (\mu_j - \lambda)^2}{k+1} \right]^{-1} \right) \tag{15}$$

The hyper-parameters for component precisions, β and ω, are given Gamma hyper-priors: $p(\beta^{-1}) \sim Ga(1,1)$, $p(\omega) \sim Ga(1, \sigma_x^2)$. Similarly, the conditional posterior for β and ω are obtained by multiplying the hyper-priors with $\prod_{j=1}^{k_{rep}} p(\tau_j \mid \omega, \beta^{-1})$, and can be simplified giving:

$$p(\omega \mid \boldsymbol{\tau}, \beta) \sim Ga\left(k\beta + 1, \left[\frac{\sigma_x^{-2} + \beta \sum_{j=1}^{k_{rep}} \tau_j}{k\beta + 1} \right]^{-1} \right) \tag{16}$$

$$p(\beta \mid \boldsymbol{\tau}, \omega) \propto \Gamma\left(\frac{\beta}{2}\right)^{-k_{rep}} \exp\left(\frac{-1}{2\beta}\right)\left(\frac{\beta}{2}\right)^{\frac{k_{rep}\beta - 3}{2}} \prod_{j=1}^{k_{rep}} \left[(\tau_j \omega)^{\beta/2} \exp\left(-\frac{\beta \tau_j \omega}{2}\right) \right] \tag{17}$$

$p(\beta \mid \boldsymbol{\tau}, \omega)$ is not in the form of a simple probability distribution function but as it is log-concave, the samples can be generated using adaptive rejection sampling (Gilks and Wild, 1992).

[2] In a strict Bayesian hierarchical analysis, the priors should not depend on the training data. The current specification of priors is essentially a empirical Bayesian hierarchical approach. Other reasonable priors will result in similar results.

Finally the concentration parameter for Dirichlet distribution, α, is given an inverse Gamma prior, $p(\alpha^{-1}) \sim Ga(1,1)$. The posterior of α given the number of represented mixtures, k_{rep}, and the number of data points, N, is:

$$p(\alpha \mid k_{\text{rep}}, N) \propto \frac{\alpha^{k_{\text{rep}}-3/2} \exp(-1/(2\alpha))\Gamma(\alpha)}{\Gamma(N+\alpha)} \tag{18}$$

$p(\alpha|k_{rep},N)$ is log-concave, and can be sampled using the adaptive rejection sampling method as above.

Appendix B Multivariate generalization

The extension to multivariate observations is straightforward. The component means and precisions become vectors and matrices respectively, and their prior and posterior distributions become multivariate Gaussian and Wishart respectively. Similar modifications apply to the hyper-parameters and their priors.

Alternatively diagonal covariance matrices for the Gaussian mixtures can be selected. This strategy ignores the correlation between the variables, but this limitation can be largely overcome by using more mixtures than required if the full covariance matrices had been utilised. The use of diagonal covariance matrices considerably simplifies the inference of the mixture models, and reduces the number of parameters. For D-dimensional data, a full covariance matrix introduces $D(D+1)/2$ free parameters, whereas a diagonal matrix only requires D parameters. Since selecting the appropriate number of mixtures is not an issue in infinite Gaussian mixtures, diagonal covariance matrices were utilised in this paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 1973.

Al-Alawi, A., Morris, A. J., and Martin, E. B. (2005). Enhanced fault detection using canonical variate analysis. *7th World Congress of Chemical Engineering*, Glasgow, Scotland, July 2005.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *Annals of Statistics*, **1**, 353–355.

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

Choi, S. W., Park, J. H., and 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., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, **39**, 1–38.

Downs, J. J. and Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers and Chemical Engineering*, **17**, 245–255.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, **68**, 589-599.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, **41**, 337–348.

Hotelling, H. (1947). Multivariate quality control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis (Eds.), *Techniques of Statistical Analysis*. New York: McGraw-Hill.

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

MacEachern, S. N. and Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

Martin, E. B. and 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., and Kiparrisides, C. (1999) Manufacturing performance enhancement through multivariate statistical process control, *Annual Reviews in Control*, **23**, 35-44.

Neal, R. M. (1998). Markov chain sampling methods for Dirichlet process mixture models. Technical Report No. 9815, Department of Statistics, University of Toronto, Canada.

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

Qin, S. J. (2003) Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, **17**, 480-502.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In S. A. Solla, T. K. Leen, and K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems 12*. MIT Press.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, **39**, 731-792.

Ricker, N. L. (1996). Decentralized control of the Tennessee Eastman challenge process. *Journal of Process Control*, **6**, 205–221.

Safavi, A. A., Chen, J, and 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.

Shao, R., Jia, F., Martin, E. B., and Morris, A. J. (1999). Wavelets and non-linear principal components analysis for process monitoring. *Control Engineering Practice*, **7**, 865–879.

Stephens, M (2000) Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *Annals of Statistics*, **28**, 40-74.

Teh, Y.W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004) Hierarchical Dirichlet processes. *Technical Report 653*, UC Berkeley Statistics, 2004.

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

Walker, S. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes. In D. Dey, P. Muller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, 243–254. New York: Springer.

Wilson, D. J. H., Irwin, G. W., and 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., and 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.

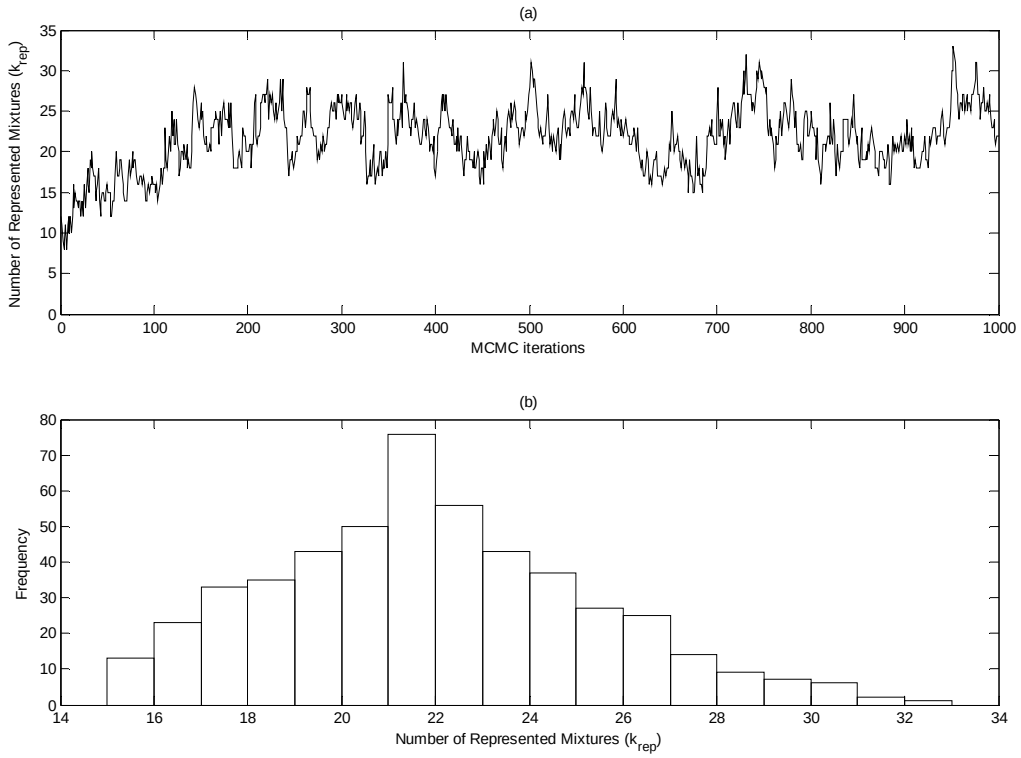

Figure 1: (a): Number of represented mixtures versus MCMC iterations; (b): Frequency of number of represented mixtures, after 500 iterations.

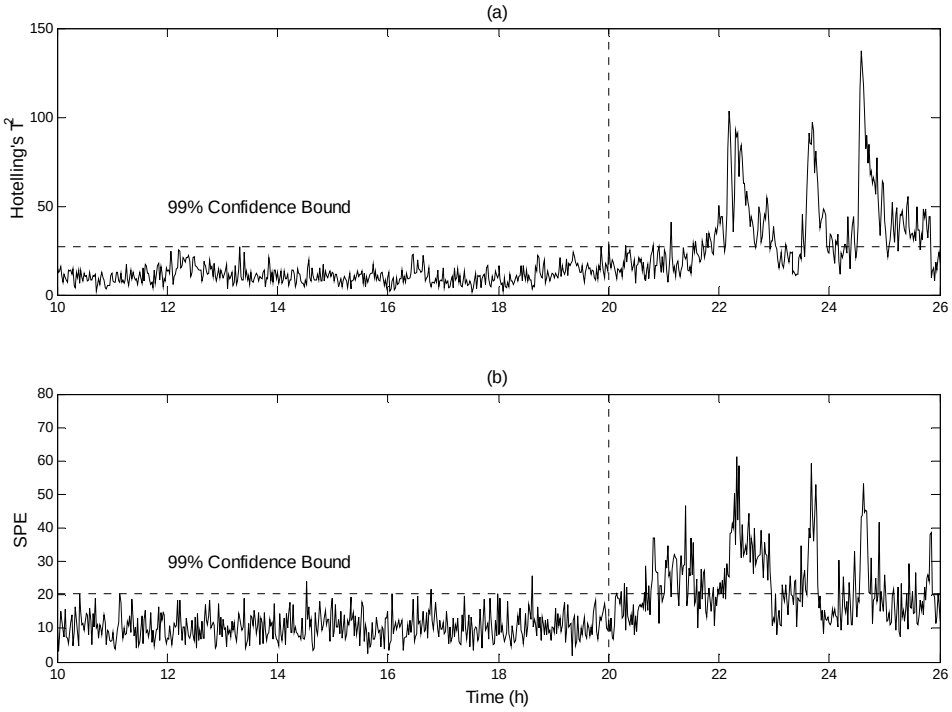

Figure 2: Process monitoring with (a): Hotelling's T^2 and (b): SPE. Fault was introduced at 20 *hrs*.

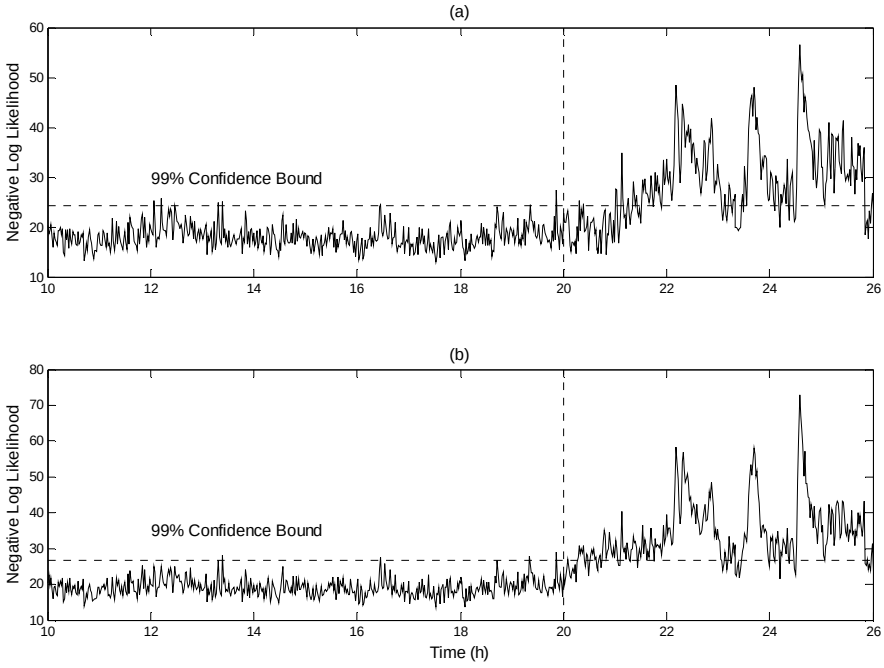

Figure 3: Process monitoring using IGMM. (a): pdf of the PCA scores is estimated to calculate the confidence bound; (b): estimation of the joint pdf of the PCA scores and log-SPE. Fault was introduced at 20 *hrs*.

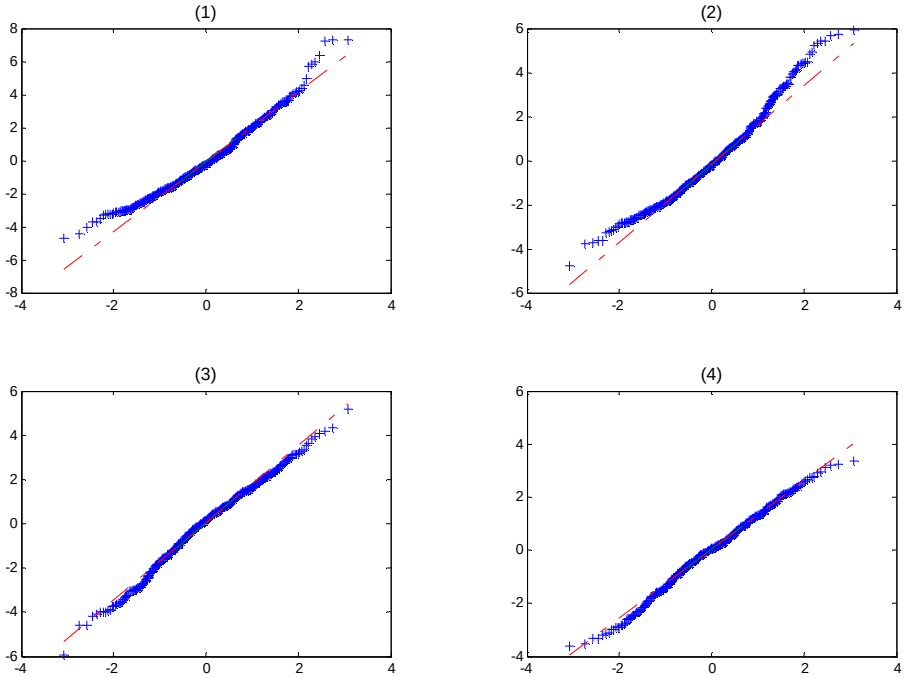

Figure 4: Quantile-quantile plots. The horizontal axes are the quantiles of a standard Gaussian distribution and the vertical axes are the quantiles of the PCA scores corresponding to the largest four eigenvalues (1-4).

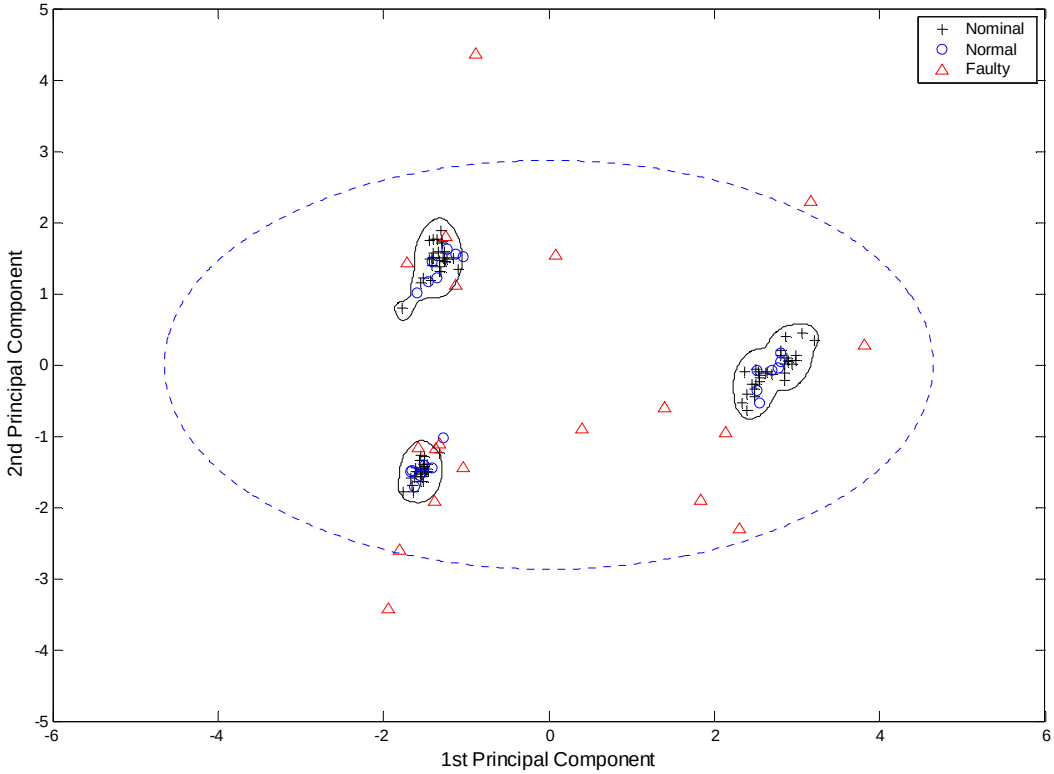

Figure 5: Bivariate scores plot for principal component 1 and 2 with 99% confidence bounds defined by the infinite GMM (solid line) and Hotelling's T^2 (dotted line).

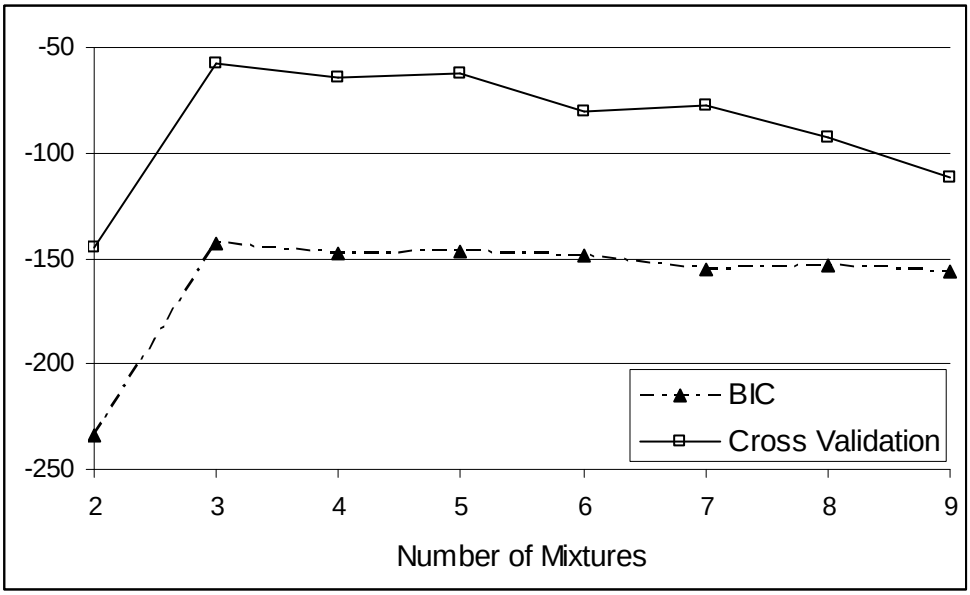

Figure 6: Selection of the number of mixtures in GMM. The vertical axis represents the BIC value, and the log-likelihood for 5-fold cross validation, respectively.

Table 1: Process faults

Case	Disturbance
IDV(1)	A/C feed ratio (step change)
IDV(10)	C Feed Temperature (random variation)
IDV(14)	Reactor cooling water valve (sticking)
IDV(12+15)	Condenser cooling water inlet temperature (random variation) and valve (sticking)

Table 2: Error rates (%) for false alarm and missing error, under different process faults.

Model	False Alarm	Missing Error			
		IDV(1)	IDV(10)	IDV(14)	IDV(12+15)
Hotelling's T^2 & SPE	1.2	1.3	9.7	3.0	26.3
IGMM	1.4	0.3	6.0	0.7	14.3

Table 3: Variables used for monitoring of semiconductor process.

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

Table 4: Summary of errors. The number of mixtures in GMM was selected to be 3, based on both BIC and cross validation.

Model	False Alarm	Missing Error
GMM	3	5
IGMM	1	2