Fault Detection Using Principal Components-Based Gaussian Mixture Model for Semiconductor Manufacturing Processes

Jianbo Yu

***Abstract*—Fault detection has been recognized in the semiconductor industry as an effective component of advanced process control framework in increasing yield and product quality. Recently, conventional process monitoring-based principal component analysis (PCA) has been applied to semiconductor manufacturing by quickly detecting when the process abnormalities have occurred. However, the unique characteristics of the semiconductor processes, nonlinearity in most batch processes, multimodal batch trajectories due to multiple operating conditions, significantly limit the applicability of PCA to the fault detection of semiconductor manufacturing. To explicitly address these unique issues in semiconductor processes, a principal components (PCs)-based Gaussian mixture model (GMM) (named PCGMM) has been proposed in this paper. GMM is capable to handle complicated data with nonlinearity or multimodal features by a mixture of multiple Gaussian components, which is very suitable to describe observations from semiconductor processes. Furthermore, two quantification indexes (negative Log likelihood probability and Mahalanobis distance) are proposed for assessing process states, and a Bayesian inference-based calculation method is further used to provide the process failure probability. The validity and effectiveness of PCGMM-based fault detection model are illustrated through two simulation processes and a semiconductor manufacturing process. The experimental results demonstrated that the proposed model is superior to PCA-based monitoring models and can achieve accurate and early detection of various types of faults in complicated manufacturing processes.**

***Index Terms*—Fault detection, Gaussian mixture model, multivariate statistical process control, principal component analysis, semiconductor manufacturing process.**

I. Introduction

THE SEMICONDUCTOR industry has obtained a big development over the past few decades. To obtain high benefits, semiconductor industry must keep cutting cost by focusing on factors related to manufacturing technology, reducing feature size, increasing wafer size, and improving yield [1], [2], which need to keep high reliability and effectiveness of semiconductor manufacturing processes. Semiconductor manufacturing is characterized by a dynamic, varying environment, compared to conventional manufacturing in the chemical industries [1], [3]. Thus, recent efforts have focused on characterizing and controlling variability critical manufacturing processes through adopting advanced process control (APC) and fault detection and classification [4]–[7] to reach desired process goals in operating individual process steps. This paper attempts to provide a Gaussian mixture model (GMM)-based fault detection model that enables semiconductor industry to achieve the following goals: 1) reducing variation of process quality; 2) decrease equipment downtime; 3) improve the quality of wafer; and 4) reducing the usage of test wafer.

Nowadays, some advanced automatic data collection and inspection techniques are now widely adopted in semiconductor manufacturing. The effective and efficient utilization of the large amount of data to characterize the process has become of increasing importance to a wide range of manufacturing industries. The online monitoring of the process state is essential for ensuring process safety and the delivery of high quality consistent products. Under multivariate processes, there are a number of quality characteristics that need to be monitored and controlled simultaneously. An appropriate approach is needed to monitor all these characteristics simultaneously. The usual practice has been to maintain a separate chart (i.e., univariate control charts) like Shewhart, CUSUM, and EWMA, for each quality characteristic [7]. However, this could result in many fault alarms when the characteristics are highly correlated.

To address this issue, multivariate statistical process control based on principal component analysis (PCA), partial least squares, and others [8], has been successfully applied to online continuous and batch process monitoring, in particular chemical processes and semiconductor manufacturing processes [9], [10]. As an effective multivariate statistical projection technique, PCA generates the first new principal components (PCs) from high correlated process data to relieve their correlations and capture the key process information. Typically, monitoring charts based on T^2 and squared prediction error (SPE) are often used to monitor out-of-control signals with the contribution plot, which shows the contribution of each measured variable to the monitoring statistic, being adopted to assist in the identification of an assignable cause of a fault. Early work using PCA for fault detection of semiconductor processes can be found in Wise and Gallagher [11].

Manuscript received August 26, 2009; revised January 6, 2011; accepted April 4, 2011. Date of publication May 12, 2011; date of current version August 3, 2011. This work was supported by the National Science Foundation of China, under Grant 71001060, and the Research Fund for the Doctoral Program of Higher Education of China, under Grant 20103108120010.

The author is with the Department of Mechanical Automation Engineering, Shanghai University, Shanghai 200072, China (e-mail: jianboyu@shu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2011.2154850

PCA-based monitoring methods have been tremendously successful in process control, and its extensive applications in batch processes have been widely studied recently across many different industries [12], [13]. In the field of semiconductor manufacturing, it was used to unfold optical emission spectra data for monitoring plasma etchers [13].

However, some unique characteristics of semiconductor manufacturing processes have posed difficulties to regular PCA-based fault detection models [2]. The unique characteristics of semiconductor processes are nonlinearity in most batch processes, multimodal batch trajectories product mix, and process steps with variable durations. These typical characteristics make modeling of regular PCA-based T^2 and SPE to be difficult, which results in the decreasing of performance of these monitoring models, significantly. To make PCA to be effective for multimodal data distribution, the "local" modeling which calculates T^2 for each local group, can address the multimodal problem [14]. For such monitoring models, priori process knowledge is required to manually segment the historical operating data into multiple local subgroups that correspond to different operating models.

First, the assumption that the PCs are Gaussian distribution when calculating the confidence bounds may be an invalid assumption particularly when the data are collected from complex manufacturing processes, e.g., processes with multiple steady state conditions or multiple operating states. To solve this important issue, some models have been proposed in opened literature recently. For example, nonlinear projection techniques [15], kernel density estimation [16], k-nearest neighbor model [2], the resulting will typically not be multivariate Gaussian. GMM [17], [18] has been used for describing the multimodal data distributions by a (limited) set of Gaussian components that can be parametrized differently. However, for semiconductor manufacturing processes, few effective models are proposed to handle and analyze process data with nonlinearity and multimodal features.

A second important issue is that two separate control charts, T^2 and SPE, are required to monitor the performance of a process. In practice, a heuristic approach is adapted, whereby the process state is observed to have changed from normal state if either of T^2 or SPE triggers an out-of-control alarm. In practice, a single monitoring chart will reduce the workload and operation cost of the operators as they will be exposed to only one monitoring chart. This is crucial for the wider acceptance of statistical process monitoring techniques in industry [18].

This paper proposes the application of GMM for the estimation of the probability density function (PDF) of semiconductor manufacturing process observations. GMM is based on unsupervised learning to organize itself according to the nature of the input data with complicated distribution, which means that the training is entirely data-driven. GMM is capable to handle complicated data with nonlinearity or multimodal features by using a mixture of multiple Gaussian model components, which is very suitable to describe data distribution from semiconductor processes. The use of GMM that monitors process changes without a prior knowledge of abnormal patterns is appealing in real-world manufacturing. Actually, only knowledge of the normal operation data sets of the process is required for training GMM. In this paper, two types of quantification indexes, i.e., negative log likelihood probability (NLLP), and Mahalanobis distance (MD) are proposed to evaluate the process states. To further address the issue of failure probability of semiconductor manufacturing processes, the Bayesian-inference-based failure probability calculation approach proposed by Yu and Qin [19] is combined into the GMM-based fault detection model. To avoid inputs with high dimension and sparsity, PCs generated by PCA are used as inputs of GMM in this paper. Thus, we call the proposed model as PCGMM. By using these comprehensible and quantitative assessment indexes, the PCGMM-based monitoring model provides improved fault detection performance for semiconductor manufacturing processes.

The rest of this paper is organized as follows. Section II presents PCA-based fault detection model. Section III presents GMM model. A PCGMM-based fault detection model is developed in Section IV. Simulation processes and a semiconductor process are used to illustrate the fault detection performance of PCGMM in Section V. Section VI gives conclusions and extensive discussions.

II. PCA-Based Process Monitoring

PCA in many ways forms the basis of the multivariate data analysis [20] and has been widely used in real-world process control. PCA has the capability to project the high-dimension data onto a lower-dimension space that contains the most variance of the original data and accounts for correlation among variables. However, PCA is a second-order method, which means it considers only mean and variance-covariance of given data, but lacks the ability to give high-order representations for non-Gaussian data, which is often the case of the real-world industrial cases.

PCA decomposes the data matrix $X \in R^{n\times d}$ (where n is the number of the samples and d is the number of variables) as the product of scores and loadings through using either the nonlinear iterative partial least squares [20] or a singular values decomposition algorithm as follows:

$$X = TP^T + E \tag{1}$$

where E is the residual matrix, and $T \in R^{n\times a}$ and $P \in R^{d\times a}$ are the score and loading matrices, respectively. Thus, the projection into PC space reduces the original set d of variables to a latent variables. The columns of P are actually eigenvectors of the covariance or correlation matrix of the variables associated with the a largest eigenvalues. Given a new sample vector x, the PCA score, prediction, and residual vectors are given as follows:

$$\text{Score} : t = P^T x \tag{2}$$

$$\text{Prediction} : x^{'} = PP^T x \tag{3}$$

$$\text{Residual} : e = (I - PP^T)x. \tag{4}$$

For abnormal detection in a new sample vector x, T^2 and SPE are often used. T^2 is a measure of the variation in PC space and the sum of the normalized squared scores, defines as

$$T^2 = t^T D^{-1} t = x^T P D^{-1} P^T x \tag{5}$$

where D is the diagonal matrix of the eigenvalues associated with the retained PCs. The 100(1-α)% upper control limit for T^2 can be obtained using the F-distribution [8] as follows:

$$T^2_{\lim} = \frac{a(N-1)}{N-a} F(a, N-1; \alpha) \tag{6}$$

where $F(a, N-1; \alpha)$ is an F-distribution with degrees of freedom $N-1$, and with level significance α [21].

On the contrary, a measure of variation not captured by the PCA model can be monitored using the SPE. The SPE statistics indicates how well each sample conforms to the model, measured by the projection of the sample vector on the residual space. The SPE is defined as the sum of square of residual e as follows:

$$SPE = e^T e = x^T (I - PP^T) x. \tag{7}$$

The upper control limit for the SPE can be computed from its approximate distribution [22] as follows:

$$Q_{\lim} = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{8}$$

where c_α is the standard normal deviate corresponding to the upper $1-\alpha$ percentile, $h_0 = 1 - 2\theta_1\theta_3/(3\theta_2^2)$, $\theta_i = \sum_{j=a+1}^{d} \lambda_j^i$ for $i = 1, 2, 3$, and λ is the eigenvalue associated with the jth loading vector of the data covariance.

III. Gaussian Mixture Model

Let $X = [X_1, \ldots, X_d]$ be a d-dimensional random variable, with $x = [x_1, \ldots, x_d]$ representing one particular outcome of X. It is said that X follows a finite mixture distribution when its PDF $p(x)$ can be written as a finite weighted sum of known densities. In cases where each component is the Gaussian density, X follows a Gaussian mixture. For a Gaussian mixture model with $M > 1$ components, the unknown mixture can be written as follows:

$$p(x) = \sum_{m=1}^{M} \pi_m p(x|m) \tag{9}$$

where $\pi_m \in (0, 1) (\forall m = 1, 2, \ldots, M)$ are the mixing proportions subject to $\sum_{m=1}^{M} \pi_m = 1$. For the Gaussian mixtures, each component density $p(x|m)$ is a normal probability distribution with mean μ_m and covariance matrix S_m

$$p(x|\theta_m) = \frac{1}{(2\pi)^{n/2} (S_m)^{-1/2}} \times \exp[0.5(x-\mu_m)^T S_m^{-1} (x-\mu_m)] \tag{10}$$

parametrized on the mean μ_m and the covariance matrix S_m, collectively denoted by the parameter vector $\theta_m = (\mu_m, S_m)$. Here, we encapsulate these parameters into a parameter vector to get $\phi = (\pi_1, \ldots, \pi_M; \theta_1, \ldots, \theta_M)$. Then (9) can be rewritten as

$$p(x|\phi) = \sum_{m=1}^{M} \pi_m p(x|\theta_m). \tag{11}$$

For the estimation problem, we assume a training set $X = \{x^{(1)}, \ldots, x^{(n)}\}$ with n independent and identically distributed samples of the random variable x. Learning aims at finding the number of components M and the optimum vector $\phi^* = (\pi_1^*, \ldots, \pi_M^*; \theta_1^*, \ldots, \theta_M^*)$, that maximizes the likelihood function

$$\phi^* = \arg\max_{\theta} L(\theta), \; L(\theta) = \prod_{i-1}^{n} p(x^{(i)}|\theta). \tag{12}$$

The usual choice for obtaining the optimum vector ϕ^* of the mixture parameters is the expectation-maximization (EM) algorithm [23]. EM is a powerful statistical tool for finding maximum likelihood solutions to problems involving observed and hidden variables. EM is an iterative procedure to find local maxima of $\log(p(X|\phi))$, where an E-step and M-step is implemented iteratively [24], [25]. For the case of Gaussian mixtures, the convergence behavior of EM is well studied [23].

Standard EM for mixture learning shows weakness which also affects the EM algorithm introduced above; it requires knowledge of M (i.e., the number of components) for reaching good local optimum. To overcome this difficulty, many deterministic criteria are proposed to estimate the appropriate number of components in GMM. Most of the practical model selection techniques are based on maximizing the type of criteria [24] as follows:

$$J(M, \phi(M)) = \log p(X; \phi(M)) - P(M) \tag{13}$$

where, $\log p(X; \phi(M))$ is the log-likelihood for the given data set X. This part can be maximized using the EM algorithm. However, introducing more components always increases the log-likelihood. The balance is achieved by introducing $P(M)$ that penalizes the complex solutions. Some examples of such model selection criterion are the Akaike information criterion (AIC) [26], the minimum description length (MDL) [27], the minimum message length [28], the Bayesian inference criterion (BIC) [29], and others. In this paper, we use the FJ criterion proposed by Figueiredo and Jain [30] for our model as follows:

$$L(\theta, X) = \log P(X|\theta) - \frac{N}{2} \sum_{m=1}^{M} \log\left(\frac{n\pi_m}{12}\right) - \frac{m}{2} \log \frac{n}{12} - \frac{m(N+1)}{2} \tag{14}$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta, X) \tag{15}$$

where N is the number of parameters specifying each component and n is the total number of training samples. Figueiredo and Jain's experimental results showed that this criterion is better than some other criterion such as MDL, BIC, Laplace-empirical criterion, and others. Generally, it is difficult to select one good model selection standard for GMM. An effective approach is to implement prior experimental to select one from

AIC, BIC, or FJ. Our prior experimental results show that FJ is similar with or better than other criterion (i.e., AIC and BIC) and thus is used as the model selection criterion in this paper.

IV. PCGMM-Based Fault Detection Model

A. Quantification Index for Process State Assessment

Once a GMM is created by using the normal data set, it is then used to monitor process states. The quantification criterion is needed to evaluate whether a new input is normal or abnormal, where a threshold is setup to ensure the required Type-I error. Abnormal detection can be implemented based on quantization index (QI). In this paper, two types of GMM-based QIs are proposed for quantifying the deviation degree of current process state with normal process state space. Furthermore, Bayesian-inference-based failure probability calculation approach proposed by [19] is adopted for the proposed fault detection system.

1) *QI 1: Negative Log Likelihood Probability:* For each new input, GMM provides $P(x)$ (9), the unconditional probability density (log likelihood), which indicates how the input follow the probability distribution of the GMM trained by normal data set. A GMM output corresponding to a novel data should be enough smaller than outputs of the GMM for normal data, namely, should be below a threshold. On the other hand, an input from the same region in input space as the training data should result in a $P(x)$ value that will be equal or greater than the threshold. Thus, log likelihood can be a good QI for evaluating the manufacturing process states. In general, since the log likelihood values are smaller than zero, to improve its intelligibility, NLLP is used as the QI in this paper as follows:

$$NLLP = -\log P(x|\phi). \tag{16}$$

2) *QI 2: Mahalanobis Distance:* MD is generally effective to measure the similarity between two data feature spaces and has been applied in semiconductor manufacturing process monitoring [31]. Based on the recognized best match component (BMC) θ_{BMC} in GMM for a new input x, the MD between x and θ_{BMC} can be calculated as follows:

$$D_{maha} = (x - \mu_{BMC})(S_{BMC})^{-1}(x - \mu_{BMC})^T \tag{17}$$

where BMC is the component in GMM that is closest to the input vector x (i.e., the minimal MD distance), and μ_{BMC} and S_{BMC} are the mean and covariance of the BMC, respectively. From the abnormal detection of view, the D_{maha} between input vector x and ϕ_{BMC} actually indicates how far away the input vector is deriving from the normal process state. Extremely high D_{maha} value, i.e., exceeding the threshold value, means that the input vector is an outlier or belongs to an abnormal class. D_{maha} can be used for quantifying the deviation degree of current process with normal process state space. Therefore, the process changes can be quantified and visualized by following the trends of D_{maha}. In this paper, GMM-based MD is different to the MD proposed in [31] because it measures distance between the input and the BMC of the GMM. Such quantifying index similar to minimum quantifying error proposed in the authors' previous work [32].

With the previous probability $P(\theta_i)$ and the conditional densities $p(x_t|\theta_i)$, the posterior probability $P(\theta_i|x_t)$ of an input vector x_t belonging to each Gaussian component in GMM can be computed by using Bayesian inference strategy.

Considering that each Gaussian component θ_i follows a unimodal Gaussian distribution, the MD of x from θ_i follows χ^2 distribution provided x belongs to θ_i. That is

$$D_{maha}((x_t, \theta_i)|x_t \in \theta_i)) = (x_t - \theta_i)S_i^{-1}(x_t - \theta_i)^T : \chi^2_m \tag{18}$$

where $D_{maha}((x_t, \theta_i)|x_t \in \theta_i))$ denotes the squared MD between x_t and the Gaussian component θ_i under the assumption $x_t \in \theta_i$, and χ^2_m has m degrees of freedom, which is the number of input vector dimension. Thus, a local MD-based probability relative each Gaussian component θ_i can be calculated as follows:

$$P_l^i(x_t) = \Pr\{((x, \theta_i)|x \in \theta_i)) \le ((x_t, \theta_i)|x_t \in \theta_i))\}. \tag{19}$$

$P_l^i(x_t)$ can be obtained by using the χ^2 PDF with appropriate degree of freedom. This index can serve as an indication of whether the input vector x_t belongs or not belongs to the corresponding Gaussian component. Due to the random characteristic that each input vector may come from multiple Gaussian components with the corresponding posterior probabilities, a global probability P_g is further defined to consider the local probability metrics across all Gaussian components as follows:

$$P_g = \sum_{i=1}^{M} P(\theta_i|x_t)P_l^i(x_t) \tag{20}$$

where the posterior probability $P(\theta_i|x_t)$ is used to incorporate the contribution of each local Gaussian component to the overall probabilistic index. Without determining the input vector x_t belongs to a single Gaussian component, such a probabilistic strategy can avoid the potential risk of false detection induced by misclassification. Since $0 \le P_l^i(x_t) \le 1$, we have

$$0 \le P_g \le \sum_{i=1}^{M} P(\theta_i|x_t) \le 1. \tag{21}$$

In this paper, the GMM-based MD and NLLP are used for monitoring process states, and are defined as PCGMM-NLLP and PCGMM-MD, respectively. Bayesian inference-based probability (BIP), i.e., P_g is used to provide process failure probability to offer more process state inform to users. It should be pointed out that, unlike PCA-based T^2 and SPE which are combined for process monitoring, PCGMM-MD and PCGMM-NLLP can be used solely for process monitoring.

Note that the latent variables in GMM do not follow a Gaussian distribution, thus, the control limit η_{NLLP} of PCGMM-NLLP cannot be determined directly from a particularly approximate distribution. An alternative approach to defining the nominal operating region of the PCGMM-NLLP is to use kernel density estimation (KDE) [16]. As for control limit of PCGMM-MD, the squared MD of an input x_t from the center of Gaussian follows χ^2 distribution with an appropriate freedom degree. Thus, the threshold η_{MD} of PCGMM-MD

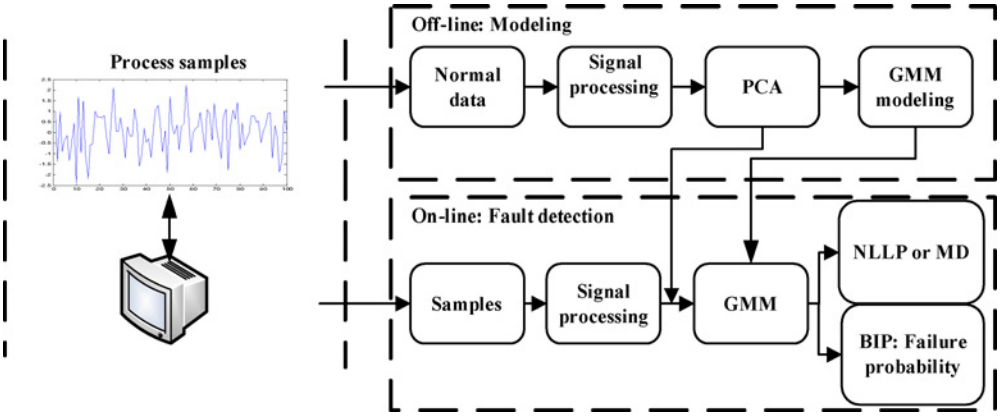

Fig. 1. Framework of PCGMM-based fault detection model.

with a significance α can be estimated using χ^2 statistics. In this paper, for PCGMM-MD and PCGMM-NLLP control charts, the confidence bound 99% (i.e., the Type I error is 1%) is used to setup the threshold η_{MD} and η_{NLLP}. Generally, if the type I error is high, the type II error will be low, while it will be high. Thus, the setting of the threshold of PCGMM-MD and PCGMM-NLLP is based on consideration of the balance between Type I and Type II errors according to the requirements of real-world applications.

B. Determination of Data Features for GMM

In general, GMM is difficult to describe probability density distribution of high dimension and sparse data. In addition, the number of parameters in a mixture model with k local Gaussian components increases by $2k(m + 1)$ when a new variable is added to the m observed variables. Furthermore, the number of parameters increases by $m + m^2$ when an additional Gaussian component is added to a mixture model. The reduction of data dimension by PCA alleviates these difficulties due to reduction of the number of parameters to be determined in GMM. Moreover, the using of PCA minimizes the information loss resulting from dimension reduction, due to PCA aims to extract information from the global variance. Thus, PCs that contains most key information of the data set are used as the inputs of GMM for fault detection.

It should be pointed out that PCA is applied to nonlinear data set. PCA will not be able to extract the nonlinear relations. However, PCA will still find the direction that corresponds to the largest variations of the data set. Thus, PCs extracted by PCA are good enough as the inputs of GMM in the nonlinear situation.

C. Algorithm Procedure of Using PCGMM

According to the above modeling scheme for PCGMM, a heuristic algorithm for applying PCGMM to monitor any abnormal deviation in the state of multivariate processes is proposed. PCGMM consists of two parts: offline modeling and online fault detection (see Fig. 1). The detailed procedure of PCGMM is as follows.

1) *Part I: Offline Modeling:* This part consists of five steps.

Step 1) Collect examples by recording/simulating time series measurements for each of the process variates when the process is in normal statistical condition. The pre-processing (e.g., normalization) of data set is implemented.

Step 2) PCA is first used to generate PCs.

Step 3) Start off the GMM modeling with the PCs using EM algorithm and FJ index, and estimate the model parameters based on the iterative steps of EM.

Step 4) Calculate the NLLP (or MD) using (16) [or (17)].

Step 5) Determine a threshold for NLLP (or MD) using KDE method (or χ^2 distribution estimation) with a confidence bound.

2) *Part II: Online Fault Detection:* For an incoming sample x, the online fault detection procedure consists of three steps.

Step 1) Input PCs generated from x to GMM to calculate NLLP (or MD) using (16) [or (17)].

Step 2) Compare NLLP (or MD) against the threshold η_{NLLP}(or η_{MD}). If $NLLP < \eta_{NLLP}$ (or $MD < \eta_{MD}$), it is classified as a normal sample. Otherwise, it is detected as a fault.

Step 3) Calculate the failure probability (P_g) of the process using (20) to obtain the risk probability of process failure.

V. Experiments and Results

To illustrate the performance of PCGMM, we applied it to the fault detection of manufacturing processes in different situations, including the presence of nonlinearity or multimodal distribution, and to compare it with the regular fault detection method PCA. Moreover, a real-world semiconductor process is used to further evaluate the performance of PCGMM. This real-world data set is publicly from Eigenvector Research, Inc. (http://software.eigenvector.com/Data/Etch/index.html).

A. Experimental Setup

The EM was run for certain cycles to train each GMM of each case and the final error which is equivalent the negative total log likelihood value was recorded as the evaluation standard of trained GMM. FJ index [30] is used to choose the GMM with the highest FJ value as the fault detection

model. Before implementing EM, the parameter initialization of GMM need to be done to lead to faster convergence of EM. K-means is used for initialization of GMM parameters, because it is by far the most convenient to use and hence the most efficient as it requires the least amount of effort and time to configure while providing good performance. The termination for EM implements the following termination criteria: 1) terminate when the total number of generations specified by the user is reached (1000), and 2) terminate when the log likelihood value of GMM cannot be improved in continuous two iterations. The termination criterion 2) is often met in the following experiments, which means the convergence of EM is enough fast not to implement 1000 iterations.

B. Nonlinear Process

The first simulation process is a modified three variable nonlinear process but only one factor, originally suggested by Dong and McAvoy [33] as follows:

$$\begin{aligned} x_1 &= t + e_1 \\ x_2 &= 1.5t^2 - 4t + e_2 \\ x_3 &= -1.5t^3 + 4t^2 + e_3 \end{aligned} \tag{22}$$

where e_1, e_2, and e_3 are independent noise variables, $N(0, 0.01)$ and $t \in [0.01, 2]$. Normal data consisting of 100 samples are generated according to these equations. The data are scaled to zero mean and unit variance. PCA is implemented and then the first two PCs are selected because they capture 99.25% variance information of the given data. One set of test data consisting of 300 samples are also generated. A disturbance is performed separately during generation of the test data set; a step change of x_2 by -0.4 is introduced at sample 101.

The scatters of the training and testing data points for the disturbance are presented in Fig. 2. The conventional PCA-based monitoring procedure is first carried out on the test data. Based on the first two PCs, the T^2 and SPE for PCA monitoring of the process under 99% confidence bound with the disturbance are presented in Fig. 3(a) and (b), respectively. It is obvious from these charts that PCA is not able to detect the disturbance under nonlinear situation; it only captures the dominant randomness.

GMM uses the first two PCs as its inputs. First, we need to find the optimal number of Gaussian components. As shown in Fig. 4, when the number of components increases from 1 to 8, the model selection criterion, FJ, increases at the beginning. As we discussed previously, larger the FJ, and better the GMM. Thus, we would prefer optimal number of components = 8, which are enough to describe the nonlinear data distribution of the process. Fig. 5 presents the data distribution and probability density description (i.e., the contour) of those Gaussian components in the constructed GMM. It can be easily seen from Fig. 5 that the established GMM precisely characterizes the nonlinear distribution of the operating data (i.e., the training data). It is worthwhile to emphasize that a prior knowledge on the number of components is not required prior to the GMM estimation.

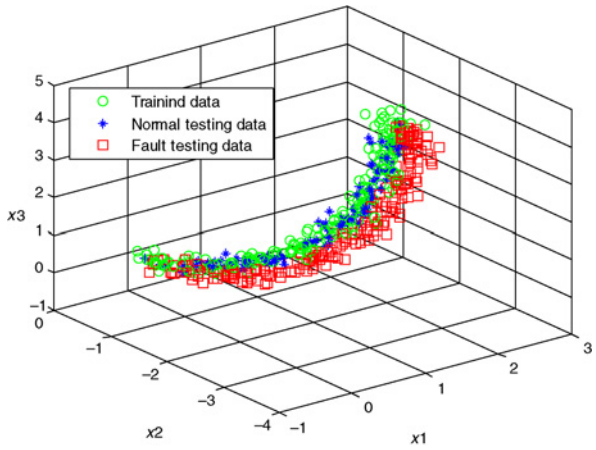

Fig. 2. Nonlinear process: scatter plot of training and testing data.

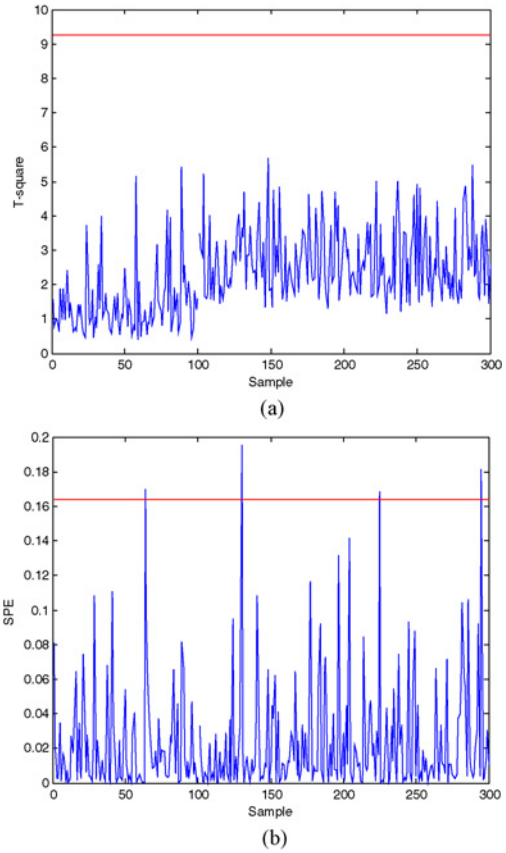

Fig. 3. Nonlinear process: PCA-based T^2 and SPE control charts. (a) T^2. (b) SPE.

Based on the trained GMM, PCGMM-MD, PCGMM-NLLP, and BIP charts under 99% confidence bound to the disturbance give the results presented in Fig. 6(a)–(c), respectively. It is obvious that the PCGMM-MD and PCGMM-NLLP detected the disturbance successfully, and BIP reported many failure probabilities with value 1 once it occurs in the sample 101, which proves that GMM-based monitoring models are very sensitive to the small faults occurred in nonlinear processes.

Due to nonlinear in the process data, the confidence bound of PCA covers a much wider rang than it does in the linear pro-

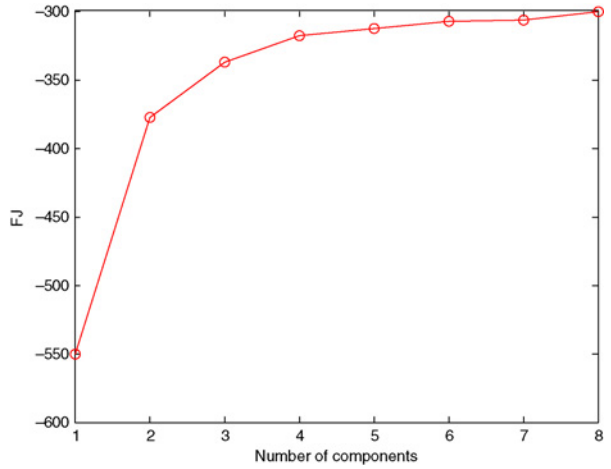

Fig. 4. Nonlinear process: FJ change curve for different number of components in GMM.

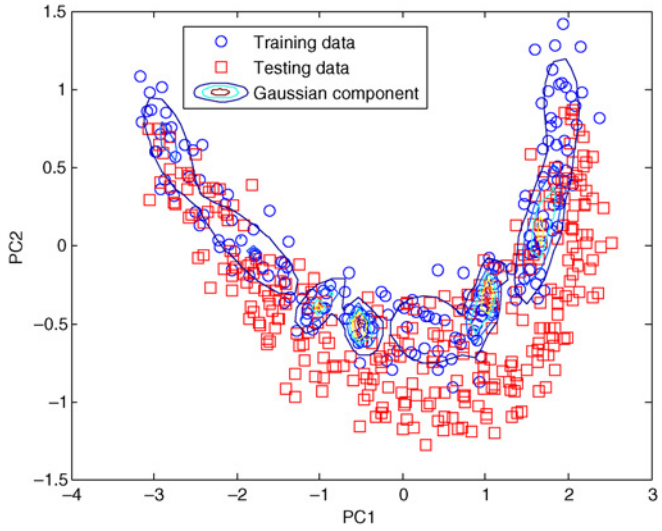

Fig. 5. Nonlinear process: scatter plot of two PCs of the training and testing data along with data distribution estimation of GMM.

cess (i.e., normalization data distribution). This process case illustrates how nonlinearity negatively affects effectiveness of PCA, and GMM extracts nonlinear features by using mixtures of multiple Gaussian components to describe complicated data distribution, and thus handles process nonlinearity very well. Meanwhile, PCGMM-MD, PCGMM-NLLP, and BIP charts look similar to the traditional control charts which can be visually inspected by operators in a similar manner.

C. Multimodal Process

The second simulation process is a three-variable multimodal process, originally suggested by Liu and Chen [34]. The measured variables are generated using

$$x_j = \Xi_j s_j + e_j,\ j = 1, \ldots, 3 \tag{23}$$

where $s_j = [s_{1,j}, s_{2,j}]^T$ denotes Gaussian distribution data sources in which $s_{i,j} \sim N(a_{i,j}, \sigma_{i,j})$ and e_j is a normal distribution with zero-mean and standard deviation of 0.05. The transformation matrix $\Xi_j \in R^{3\times 2}$, in which the elements

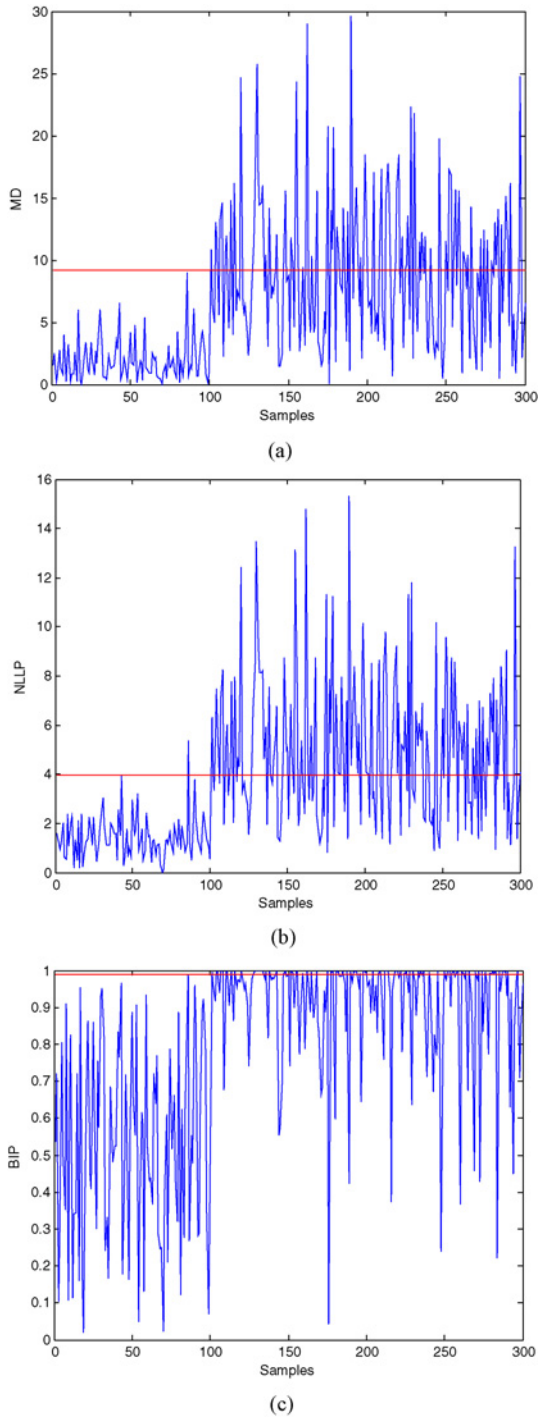

Fig. 6. Nonlinear process: GMM-based monitoring charts. (a) PCGMM-MD. (b) PCGMM-NLLP. (c) BIP.

are different for each mode in order to reflect the variable correlation changes among different modes is written as

$$\Xi_j = \begin{bmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \\ \varphi_j & 1-\varphi_j \end{bmatrix}. \tag{24}$$

Two-hundred samples are generated for each mode and thus total 600 samples are used as the normal training data set. Table I lists the parameter values for generating the training data set where M1, M2, and M3 represent three modes,

TABLE I
PARAMETERS FOR THE MULTIMODAL PROCESS CASE WITH THREE MODES

	M1	M2	M3
s_1	$N(0, 1)$	$N(10, 1)$	$N(10, 2)$
s_2	$N(0, 1)$	$N(10, 2)$	$N(-10, 1)$
θ	$\pi/3$	$\pi/4$	$\pi/6$
φ	0.5	0.6	0.4

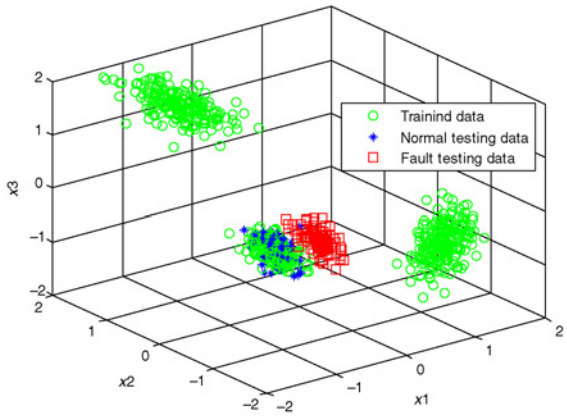

Fig. 7. Multimodal process: scatter plot of training and testing data.

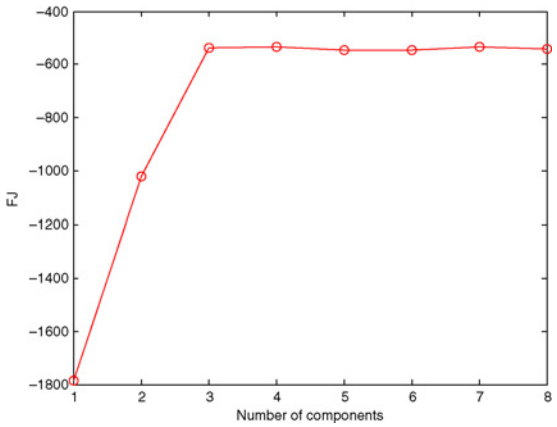

Fig. 8. Multimodal process: FJ change curve for different number of components in GMM.

respectively. Fig. 7 presents the data scatter of the training dataset.

PCA is first used to the training data, which is normalized to zero and unit variance, and the first two PCs are kept, which captured more than 85% of the total variance information. The score vectors are then used to train GMM, where model selection is implemented by using FJ (see Fig. 8). Since the GMM with four components shows the best performance, we choose it as our monitoring model. Fig. 9 shows the scatter plots of the projected data with the two PCs in the training and testing data set along with the data distribution estimation of GMM. It can be easily seen that the constructed GMM precisely characterizes the multimodal distribution of the operating data (i.e., the training data).

To illustrate the performance of the GMM-based monitoring models, a period test data are generated in the following

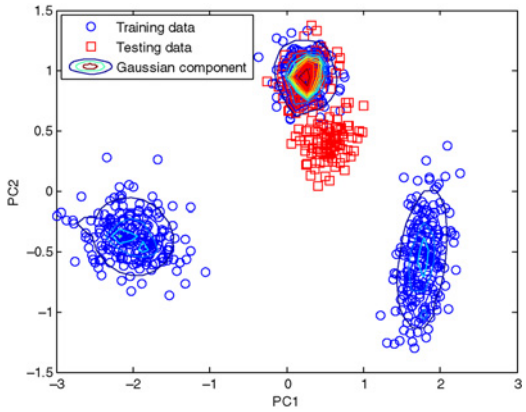

Fig. 9. Multimodal process: scatter plot of two PCs of the training and testing data along with data distribution estimation of GMM.

faulty disturbance scenario. The system is initially running at mode 1, and then a bias error of 4 is added to x_1 from the 101st through 200th samples. The test data points are plotted in Fig. 7, and the projected test data with two PCs are also exhibited in Fig. 9.

PCA-based T^2 and SPE are implemented on the testing data set from the three different modes. The T^2 and SPE of the process are presented in Fig. 10(a) and (b), respectively. It is obvious from the two charts that PCA-based monitoring models are not able to detect this faulty disturbance under multimodal situation.

The results indicate that the conventional PCA is not capable to handle process data from the multimodal data distribution because the prerequisite assumption of unimodal normal distribution becomes invalid in this situation. Because PCA can only catch linear structure from the multimodal data, the inflated global covariance makes both T^2 and SPE statistics insensitive to disturbances, the multimodal environment imposes limitations on PCA and makes it less effective.

In contrast, the GMM-based monitoring models are performed, and the PCGMM-MD, PCGMM-NLLP, and BIP chart for the disturbance are presented in Fig. 11(a)–(c), respectively. It is very obvious that PCGMM-MD and PCGMM-NLLP detect most faults from sample 101 to 200, while they do not trigger false alarms for the normal observation from sample 1 to 100 (only PCGMM-MD chart triggers one false alarm). BIP also reported many failure probabilities values with 1 for this disturbance. As opposed to the complete failure of T^2 and SPE statistic, PCGMM-MD and PCGMM-NLLP are capable of detecting such faulty disturbance without triggering too many fault alarms. The results of the disturbance in this process further demonstrate that the GMM-based monitoring models outperform the PCA-based T^2 and SPE control charts under multimodal operating conditions. GMM is capable to handle multimodal data distribution because it uses mixtures of Gaussian components to effectively describe the complicated data distribution (see Fig. 9).

D. Semiconductor Manufacturing Process

The manufacturing of semiconductor is introduced as an example of the monitoring of batch manufacturing processes.

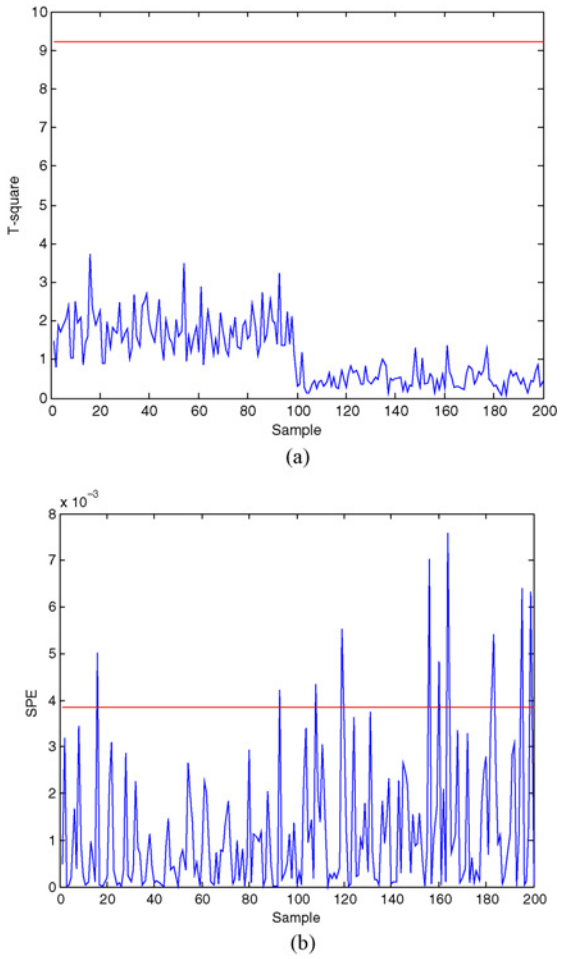

Fig. 10. Multimodal process: PCA-based T^2 and SPE control charts. (a) T^2. (b) SPE.

This industrial case is used to compare different fault detection methods. Although there are many steps in this manufacturing process of semiconductors, this paper focuses specially on an Al stack etch process performed on the commercially available Lam 9600 plasma etch tool [10]. The goal of this process is to etch the TiN/Al-0.5% Cu/TiN/oxide stack with an inductively coupled BCL_3/CL_2 plasma. The key parameters of interest are the linewidth of the etched A1 line, uniformity across the wafer, and the oxide loss. The detail information about the semiconductor manufacturing process refers to literature [10].

In the etch process, it would be ideal to have sensors which directly reflected the state of the wafers in the manufacturing process. However, with a few exceptions, wafer state sensors are typically unavailable in original equipment manufacturer processing tools [10]. Thus, the alternative is to setup more commonly available process state sensors to obtain the wafer state information. The process is equipped with several sensor systems including machine state variables, optical emission spectroscopy of the plasma, and a radio-frequency monitoring system. Thus, data from 12 process sensors, which are listed in Table II, were collected during the wafer processing stage, which was of duration 90 s. A sampling interval of 1 s was used in the analysis. Thus, for each batch, the data are of the order (90*12). A series of three experiments (numbers

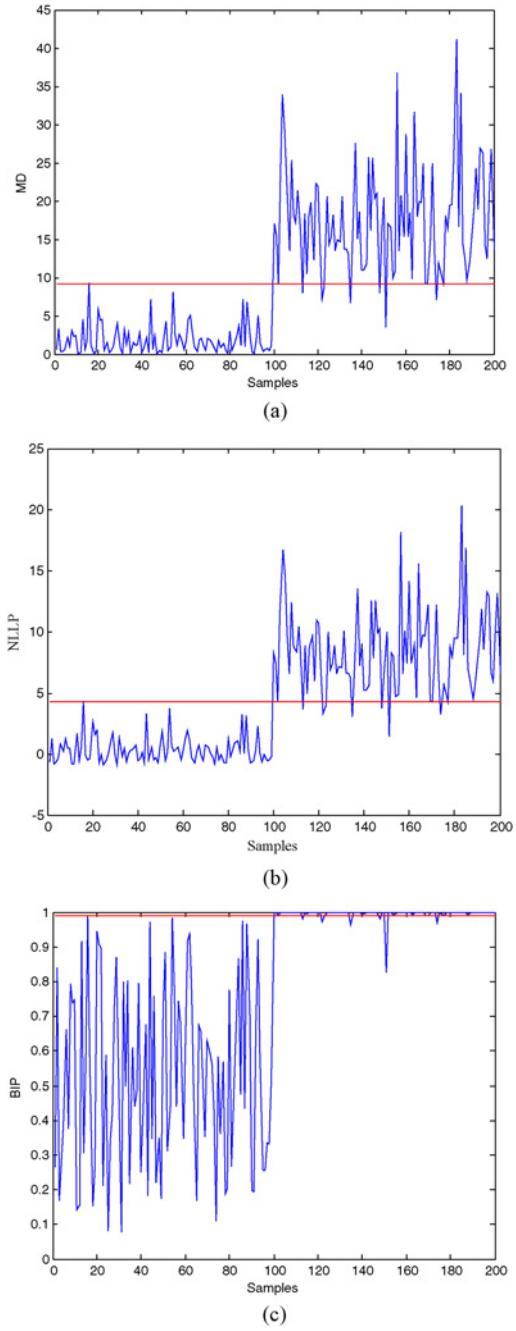

Fig. 11. Multimodal process: GMM-based monitoring charts. (a) PCGMM-MD. (b) PCGMM-NLLP. (c) BIP.

29, 31, and 33), resulting in three distinct data groups, were performed where faults were intentionally introduced by changing specific manipulated variables [transformer coupled plasma power, radio frequency (RF) power, pressure, plasma flow rate and helium chunk pressure]. The three experiments consist of total 108 normal wafers with 20 faulty batches induced from changing the transformer-coupled plasma power, RF power, pressure, CL_2 and BCL_3 flow rate, and He chuck pressure. The 56th wafer is excluded from 108 normal wafers because it has very few examples compared with other normal wafers. The experiments were run several weeks apart and data from different experiments have different mean and somewhat different covariance structure.

TABLE II
VARIABLES USED FOR MONITORING THE SEMICONDUCTOR PROCESS

Variable	Definition	Variable	Definition	Variable	Definition
1	End point A detector	5	RF phase error	9	Transformer-coupled plasma phase error
2	Helium pressure	6	RF power	10	Transformer-coupled plasma reflected power
3	RF tuner	7	RF impedance	11	Transformer-coupled plasma load
4	RF load	8	Transformer-coupled plasma tuner	12	VAT valve

In this paper, 21 batches, seven from each group (i.e., experiment 29, 31, and 33), are selected from the normal batches as the validation data set to investigate the fault detection performance of the proposed approach. The remaining of 107 nominal batches were used build nominal PCA representations.

The data from an experimental study take the form of three-way arrays. To understand the nature of the data available with which to monitor batch processes, consider a typical batch run in which $j = 1, 2, \ldots, J$ variables, are measured at $k = 1, 2, \ldots, K$ time intervals throughout the batch. Similar data will exist on a number of such batch runs $i = 1, 2, \ldots, I$. To analyze such 3-D data ($I \times J \times K$), "multiway" analysis methods have been proposed to "unfold" the 3-D array into a 2-D matrix and conventional PCA is then applied to the unfolded data matrix [12]. This paper unfolds the data array ($I \times J \times K$) into a large 2-D matrix ($I \times JK$). The unfolded data matrix is mean centered and scaled to attenuate the major nonlinear behavior of the process. This will avoid the important variable of small magnitudes from being taken over by less important but larger magnitude variables. Based on the unfolded data, multiway PCA is used to decompose the data ($I \times JK$) into a series of PCs consisting of score vectors (T), and loading matrices (P), plus a residual matrix (E). It was observed that the initial three PCs explain 24.4%, 12.1%, and 3.6% of the total variance, respectively, which supports the selection of only first two PCs. Furthermore, the prior experimental results show that using of more PCs cannot improve the performance of GMM-based monitoring models significantly. Thus, the first two PCs are used as input features of models.

The input features influent the performance of recognizers, significantly. These features that are sensitive to faults of processes are able to improve the performance of monitoring models. Unlike the GMMs only using PCs of PCA as input features for previous simulation processes, we combine PCs, T^2, and SPE as the input features of GMM to improve its detection performance.

Before implementing fault detection, we first compare the modeling process of GMM and T^2 by using the first two PCs. Fig. 12 shows the PCA scores plot of the process data, the multimodal data distribution were described by GMM with four Gaussian components. From this figure, it can be observed that the multimodal property in this dataset invalidates the underlying Gaussian assumption with respect to the traditional confidence bound. Therefore, the common global T^2 fails to identify many of the non-conforming data points. In contrast, the GMM-based approach provides a more appropriate confidence bound that identifies the non-conforming batches and effectively recognizes the distinct cluster in the data.

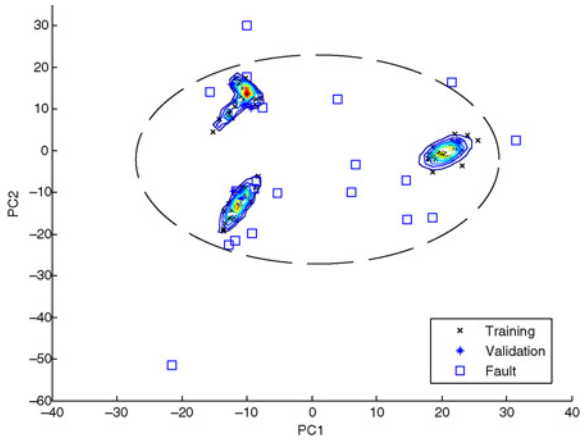

Fig. 12. Bivariate scores plot for PCs 1 (PC1) and 2 (PC2) with 99% confidence bounds defined by the PCs-based GMM (——) and T^2(——): training (×), validation (*), and fault(□).

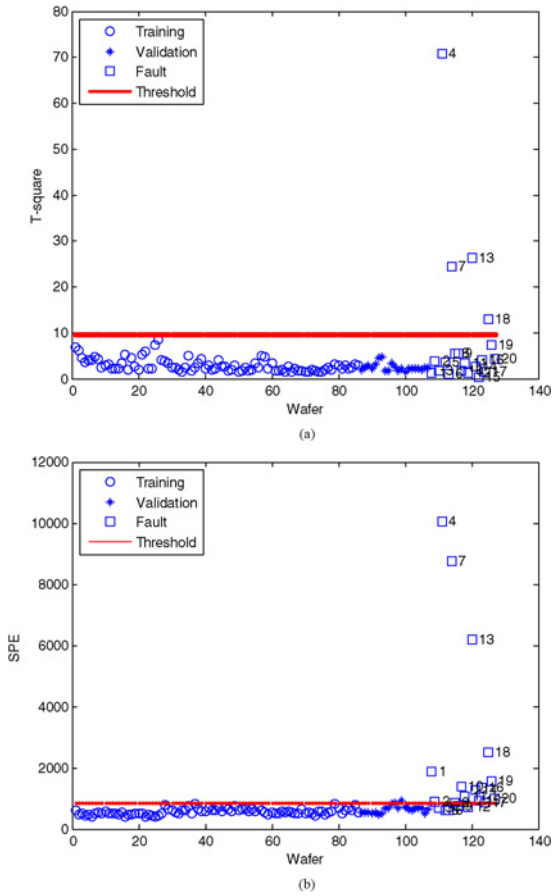

Fig. 13. Semiconductor manufacturing process: fault detection results using PCA. (a) T^2 chart. (b) SPE chart.

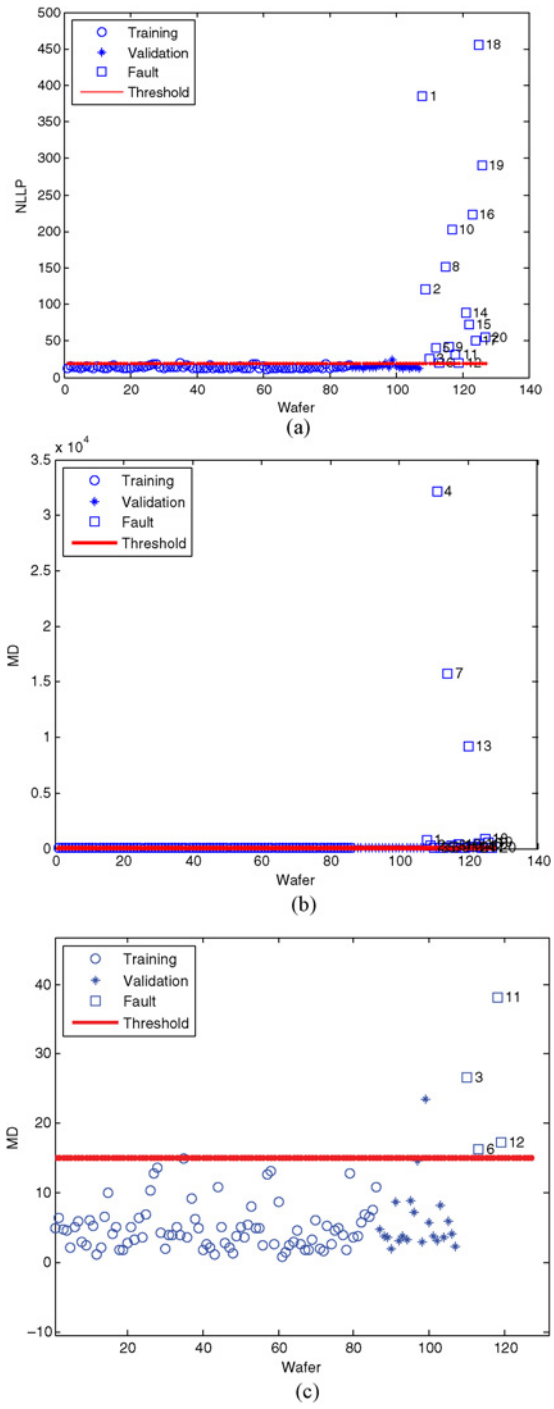

Fig. 14. Semiconductor manufacturing process: fault detection results using PCGMM. (a) PCGMM-NLLP. (b) PCGMM-MD. (c) Local enlargement of (b).

Fig. 13(a) and (b) shows the fault detection results based on the T^2 and SPE under 99% confidence bound, respectively. The T^2 chart detects only four faults and SPE chart misses six faults. Detailed fault detection results are listed in Table III. The less efficiency of PCA in this case can be explained by the characteristic multimodal distribution of the batch trajectories. As discussed earlier, the process data are from three experiments that were run several weeks apart. Due to process shifts or drifts during the whole experimental process,

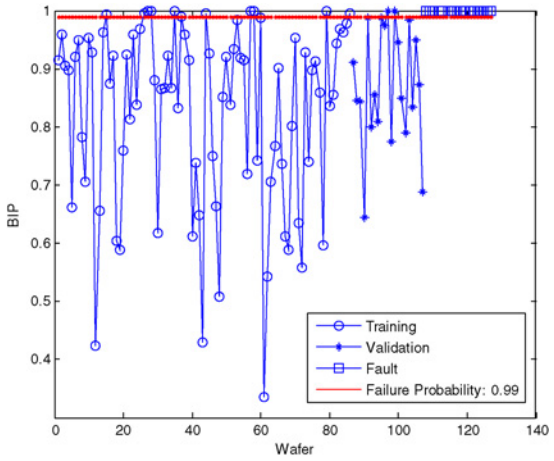

Fig. 15. Semiconductor manufacturing process: failure probability (i.e., BIP).

TABLE III
FAULT DETECTION RESULTS BY PCA-T^2, PCA-SPE, PCGMM-NLLP, AND PCGMM-MD

Fault no.	PCA-T^2	PCA-SPE	PCGMM-NLLP	PCGMM-MD
1		√	√	√
2		√	√	√
3			√	√
4	√	√	√	√
5			√	√
6				√
7	√	√	√	√
8		√	√	√
9			√	√
10		√	√	√
11		√	√	√
12			√	√
13	√	√	√	√
14		√	√	√
15		√	√	√
16		√	√	√
17			√	√
18	√	√	√	√
19		√	√	√
20		√	√	√

different experiments result in different data distributions. It should be pointed out that process shifts or drifts are inevitable in real-world applications due to various normal operations like preventive maintenance and part replacement [2].

To make PCA to be effective in such industrial case with multimodals, the "local" modeling of Wise *et al.* [10], which calculates T^2 for each local group, can address the multimodal problem. However, the determination of the number of clusters is still a problem before building "local" models.

Four Gaussian components are used for GMM according to the FJ calculation. The PCGMM-NLLP and PCGMM-MD charts under 99% confidence bound are presented in Fig. 14(a) and (b), respectively. We can see that both the PCGMM-MD and PCGMM-NLLP charts successfully detect all the 20 faults except for fault 6 with one false alarm, and their failure probabilities (i.e., BIP) are all 100% as shown in Fig. 15.

It should be pointed out that we plotted a failure probability line with 0.99 in BIP chart, which is used to let users know the process failure probability levels. The detailed detection results can be found in Table III. The results of faults in this semiconductor process further demonstrate that PCGMM-NLLP and PCGMM-MD outperform PCA-based T^2 and SPE control charts in detecting various types of faults under multimodal operating conditions. Meanwhile, both the PCGMM-NLLP and PCGMM-MD control charts looks similar to the traditional control charts which can be visually inspected by practitioners in a similar manner.

VI. Conclusion

In this paper, a new fault detection model using PCGMM was proposed to explicitly solve some important issues in most semiconductor processes: nonlinearity and multimodal trajectories. The extracted PCs by PCA were used to improve the performance of GMM with dimension reduction and global variance information preservation. Because the proposed PCGMM model does not need assumption about the linearity of the process and it detects faults by using a mixture of multiple Gaussian components, PCGMM naturally handles process nonlinearity and multimodal situations. Based on the complicated data describing capability of GMM, two process state quantifying indexes (i.e., NLLP and MD) with failure probability (i.e., BIP) were proposed to assess the process states for fault detection. The goal of such a scheme is to reduce variations in the semiconductor manufacturing process by quickly and accurately detecting faults, thus leading to the production of a more consistent and valuable products. The experimental results on simulation and real-world cases indicate that PCGMM outperforms PCA-based fault detection methods in nonlinearity and multimodal cases. The intension of this paper is to provide an alternative fault detection model for the semiconductor research community so that utility of statistical learning models like GMM can be examined in some complicated process cases in which the regular PCA-based models do not work well. An interesting issue would be to develop PCGMM-based fault diagnosis model for semiconductor manufacturing processes and our work in this direction is underway.

Acknowledgment

The authors would like to express sincere appreciation to the anonymous referees for their detailed and helpful comments to improve the quality of this paper.

References

[1] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, Nov. 2000.

[2] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 20, no. 4, pp. 345–354, Nov. 2007.

[3] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*. New York: Wiley, 2006.

[4] S. J. Qin, G. Cherry, R. Good, J. Wang, and C. A. Harrison, "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *J. Process Control*, vol. 16, no. 3, pp. 179–191, Mar. 2006.

[5] Q. P. He and J. Wang, "Statistical fault detection of batch processes in semiconductor manufacturing," in *Proc. AIChE Annu. Conf.*, Nov. 2006.

[6] G. A. Cherry and S. J. Qin, "Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 2, pp. 159–172, May 2006.

[7] C. J. Spanos, "Statistical process control in semiconductor manufacturing," *Proc. IEEE*, vol. 80, no. 6, pp. 819–830, Jun. 1992.

[8] J. E. Jackson, *A User's Guide to Principal Components*. New York: Wiley, 1991.

[9] G. Cherry, R. Good, and S. J. Qin, "Semiconductor process monitoring and fault detection with recursive multiway PCA based on a combined index," in *Proc. AEC/APC Symp. XIV*, Sep. 2002, pp. 130–139.

[10] B. M. Wise, N. B. Gallagher, S. W. Butler, D. D. White, Jr., and G. G. Barna, "A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process," *J. Chemometrics*, vol. 13, nos. 3–4, pp. 379–396, Jul. 1999.

[11] B. M. Wise and N. B. Gallagher, "The process chemometrics approach to process monitoring and fault detection," *J. Process. Contr.*, vol. 6, no. 6, pp. 329–348, Dec. 1996.

[12] P. Nomikos and J. MacGregor, "Multivariate SPC charts for monitoring batch processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, Feb. 1995.

[13] H. Yue, S. Qin, R. Markle, C. Nauert, and M. Gatto, "Fault detection of plasma etchers using optical emission spectra," *IEEE Trans. Semicond. Manuf.*, vol. 13, no. 3, pp. 374–385, Aug. 2000.

[14] S. J. Zhao, J. Zhang, and Y. M. Xu, "Monitoring of processes with multiple operating modes through multiple principal component analysis models," *Ind. Eng. Chem. Res.*, vol. 43, no. 22, pp. 7025–7035, Sep. 2004.

[15] R. Shao, F. Jia, E. B. Martin, and A. J. Morris, "Wavelets and nonlinear principal components analysis for process monitoring," *Contr. Eng. Practice*, vol. 7, no. 7, pp. 865–879, Jul. 1999.

[16] E. B. Martin and A. J. Morris, "Non-parametric confidence bounds for process performance monitoring charts," *J. Process Contr.*, vol. 6, no. 6, pp. 349–358, Dec. 1996.

[17] J. Chen and J. Liu, "Mixture principal component analysis models for process monitoring," *Industr. Eng. Chem. Res.*, vol. 38, no. 4, pp. 1478–1488, Feb. 2004.

[18] T. Chen, J. Morris, and E. Martin, "Probability density estimation via an infinite Gaussian mixture model: Application to statistical process monitoring," *Appl. Statist.*, vol. 55, no. 5, pp. 699–715, Oct. 2006.

[19] J. Yu and S. J. Qin, "Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models," *AICHE J.*, vol. 54, no. 7, pp. 1811–1829, May 2008.

[20] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987.

[21] R. A. Johnson and D. W. Wichen, *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[22] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, Aug. 1979.

[23] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Royal Statistic Soc.*, vol. 30, no. 1, pp. 1–38, 1977.

[24] Z. Zivkovic and F. V. D. Heijen, "Recursive unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 26, no. 5, pp. 651–656, May 2004.

[25] B. Zhang, C. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," *Pattern Recog.*, vol. 37, no. 1, pp. 131–144, Jan. 2004.

[26] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[27] J. Rissansen, "Stochastic complexity," *J. Royal Statistic. Soc. Series B (Methodological)*, vol. 49, no. 3, pp. 223–239, 1987.

[28] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Royal Statistic. Soc. Series B (Methodological)*, vol. 49, no. 3, pp. 240–265, 1987.

[29] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[30] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[31] K. Sakamoto, M. Obata, and N. Koyama, "Semiconductor manufacturing apparatus, method of detecting abnormality, identifying cause of abnormality, or predicting abnormality in the semiconductor manufacturing apparatus, and storage medium storing computer program for performing the method," U.S. Patent 7751921, 2006.

[32] J. B. Yu and L. F. Xi, "Using an MQE chart based on a self-organizing map NN to monitor out-of-control signals in manufacturing processes," *Int. J. Product. Res.*, vol. 46, no. 21, pp. 5907–5933, Nov. 2008.

[33] D. Dong and T. J. McAvoy, "Nonlinear principal component analysis, based on principal curves and neural networks," *Comput. Chem. Eng.*, vol. 20, no. 1, pp. 65–78, Jan. 1996.

[34] J. Liu and D. S. Chen, "Operational performance assessment and fault isolation for multimode processes," *Ind. Eng. Chem. Res.*, vol. 49, no. 8, pp. 3700–3714, Mar. 2010.

Jianbo Yu received the B.E. degree from the Department of Industrial Engineering, Zhejiang University of Technology, Zhejiang, China, in 2002, the M.E. degree from the Department of Mechanical Automation Engineering, Shanghai University, Shanghai, China, in 2005, and the Ph.D. degree from the Department of Industrial Engineering and Management, Shanghai Jiaotong University, Shanghai, in 2009.

In 2008, he was a Visiting Scholar with the Center for Intelligent Maintenance System, University of Cincinnati, Cincinnati, OH. Since 2009, he has been working as a Research Fellow with the Department of Mechanical Automation Engineering, Shanghai University. His current research interests include manufacturing process control, intelligent condition-based maintenance, machine learning, and statistical analysis.