Proceedings of the
46th IEEE Conference on Decision and Control
New Orleans, LA, USA, Dec. 12-14, 2007

Monitoring Non-normal Data with Principal Component Analysis and Adaptive Density Estimation

Gregory A. Cherry and S. Joe Qin

Abstract— **The issue of monitoring non-normally distributed data with principal component analysis (PCA) is addressed through the application of density estimation for evaluating the quality of the principal component scores. Although kernel density estimation has been previously cited as a method for monitoring such data, mixture models are proposed here in order to reduce model complexity and computational effort. Furthermore, several adaptation strategies for the density estimators are developed and suggestions are provided on their use. A rapid thermal anneal case study demonstrates how the estimators outperform the traditional Hotelling's T^2 statistic due to the presence of a first wafer effect.**

I. INTRODUCTION

Principal component analysis (PCA) is an algorithm that can be used for multivariate fault detection and diagnosis across a wide variety of applications within the semiconductor manufacturing environment, and its benefits are best realized for cases in which a large number of correlated variables need to be monitored simultaneously to indicate the quality of either the semiconductor product or the processing equipment [1, 14, 22, 23, 26, 28, 29]. But despite its flexibility and ever-expanding role within the industry, one frequently noted impediment of PCA is that it can be difficult to reach a balance between the numbers of type I errors (false alarms) and type II errors (missed faults) [20, 25].

The inability of conventional MSPC to correctly characterize faults in the presence of non-normal data has been well-documented, and several alternative approaches for dealing with this obstacle have been proposed. The method of mixture PCA relies on the same linear model as conventional PCA, but supplements it by grouping the transformed data set into a distinct set of clusters [3, 5]. Each cluster then defines a population with a given mean and covariance that can used to identify abnormal behavior. Alternatively, the method of nonlinear PCA projects the data set onto a nonlinear set of curves or surfaces, which can be extracted using a neural network approach [8, 9]. A method that uses support vector machines (SVM) for classification of multi-mode data has also been proposed, although its effectiveness is limited to those cases where the process conditions for the various faults have been observed in the training data sets [6].

While the nonlinear PCA, mixture PCA, and SVM methods have been shown to be effective for dealing with certain types of non-normal data, they still rely on assumptions that the data can be respectively described by curves or clusters, or that all fault classes have been observed in the past. In contrast, the method of kernel density estimation (KDE) has been proposed for score monitoring in conjunction with either conventional PCA [4, 16] or nonlinear PCA [13] in order to establish confidence bounds that are not constrained by such assumptions.

This paper considers the implications of making incorrect assumptions about the distributions of data sets when applying multivariate statistical process control. Building on previous work by Martin and Morris [16] and Chen *et al.* [4], the kernel estimator will be applied in order to characterize non-normal distributions of principal component scores. As an alternative to the kernel approach, adaptive mixture models for monitoring principal component scores will also be introduced. Although not previously applied in the context of MSPC, Priebe and Marchette [19] have shown that mixture models can provide similar estimates as the kernel estimators, but with a substantially reduced computational load. Additionally, several density estimator adaptation strategies will be presented and evaluated in terms of their ability to take into account process drift while maintaining sensitivity to faults.

The organization of the paper is as follows. Section II reviews the PCA algorithm, including the calculation of multivariate monitoring indices and their confidence limits. Sections III and IV describe the kernel density estimator and adaptive mixture model and how they can be used to estimate the distributions of the principal component scores in a non-parametric fashion. The methods are then applied to a rapid thermal annealing (RTA) process in Section V, which also considers several methods for adapting the estimators. Conclusions are provided in Section VI.

II. PCA BACKGROUND AND MULTIVARIATE STATISTICAL PROCESS CONTROL

Suppose that n samples of data have been collected from a manufacturing process in which m variables are observed for each sample. PCA is a method that allows the matrix of correlated, multidimensional data, $\mathbf{X} \in \mathbb{R}^{n\times m}$, to be decomposed into a much smaller set of uncorrelated variables called scores, $\mathbf{T} \in \mathbb{R}^{n\times l}$, according to

$$\mathbf{X} = \hat{\mathbf{X}} + \tilde{\mathbf{X}} = \mathbf{T}\mathbf{P}^T + \tilde{\mathbf{X}}. \qquad (1)$$

Prior to the analysis, $\mathbf{X}$ is typically scaled such that each column has zero mean and unit variance. The loadings matrix, $\mathbf{P} \in \mathbb{R}^{m\times l}$, is assembled by taking the l principal

G. A. Cherry is with Advanced Micro Devices, Inc., Austin, TX 78741, USA gregory.cherry@amd.com

S. J. Qin is with the Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, CA 90089-1211, USA sqin@usc.edu

1-4244-1498-9/07/$25.00 ©2007 IEEE.

eigenvectors of the correlation matrix, which is approximated by

$$\mathbf{R} \approx \frac{1}{n-1}\mathbf{X}^T\mathbf{X}. \tag{2}$$

The result is that $\mathbf{X}$ is decomposed into a modeled part $\hat{\mathbf{X}}$, and a residual part $\tilde{\mathbf{X}}$. Thus, for the purpose of multivariate statistical process control (MSPC), there are two performance indices that should be monitored. First, Hotelling's T^2 should be used to identify excursions within the principal component space (PCS), according to

$$T^2 = \mathbf{x}^T\mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^T\mathbf{x}. \tag{3}$$

Meanwhile, the squared prediction error (SPE) will identify excursions in the residual space as

$$SPE = \tilde{\mathbf{x}}^T\tilde{\mathbf{x}} = \mathbf{x}^T\tilde{\mathbf{P}}\tilde{\mathbf{P}}^T\mathbf{x}. \tag{4}$$

While several alternatives indices have been proposed which combine the T^2 and SPE into a single index[12, 15, 28], the separate monitoring of these quantities can give insight into the type of disturbance that has occurred. By monitoring T^2, disruptions can be attributed to an increased variation with variable relationships that are maintained as expected. In contrast, abnormal SPE behavior will indicate when the expected correlation no longer holds.

Given that the data follow a multivariate-normal distribution and the number of sample points allows the mean and covariance to be accurately estimated, the χ^2 distribution with l degrees of freedom will adequately describe the T^2 index [28]. Conversely, if the residuals are normally distributed, the SPE should also follow a χ^2 distribution [17]. If the normality assumptions hold, normal operation, given control limit α and eigenvalues λ_j, can be characterized by

$$T^2 \leq \chi^2_\alpha(l) \equiv \tau^2 \tag{5}$$

and

$$SPE \leq g\chi^2_{h,\alpha} \equiv \delta^2, \tag{6}$$

with

$$\theta_i = \sum_{j=l+1}^{m} \lambda_j^i, \qquad g = \theta_2/\theta_1, \qquad \text{and} \qquad h = \theta_1^2/\theta_2. \tag{7}$$

While the control limits defined above will work reasonably well when the normality assumption is satisfied, actual processing data are often far from normal. Because the major correlations among variables are captured in the first few principal components, deviation from multivariate normality will strongly impact the frequency of type I and type II errors observed when monitoring Hotelling's T^2. The impact on the SPE index and its limits will be less pronounced because the residual space will mostly be made up of process noise and random fluctuations. Therefore, the strategy proposed in this work is to continue to monitor the SPE as usual with the control limit defined in (6), but to monitor the results in the principal component space through the use of density estimation techniques, which will be described in the next two sections.

III. KERNEL DENSITY ESTIMATION

In order to overcome the issue of non-normality in the raw data, Martin and Morris [16] proposed that non-parametric techniques be used to approximate the distributions of the PCA scores. Specifically, rather than assume a normal distribution for the scores, the kernel density estimator is used for approximating the distribution empirically. As applied for PCA monitoring, the kernel estimator takes the form

$$\hat{f}(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n} K\left(\mathbf{H}^{-1/2}\left(\mathbf{t} - \mathbf{T}_j\right)\right). \tag{8}$$

Thus, the probability density for a new observation, $\hat{f}$, is estimated by taking a weighted summation of n kernels, each centered at an observation, $\mathbf{T}_j$. The kernel function, K, provides the shape of the underlying distributions that are summed in order to give the overall density estimate. Most commonly, the kernel function is selected to be multivariate normal, yielding

$$\hat{f}(\mathbf{t}) = \frac{1}{n(2\pi)^{l/2}}|\mathbf{H}|^{-1/2}\sum_{j=1}^{n} e^{-1/2(\mathbf{t}-\mathbf{T}_j)^T H^{-1}(\mathbf{t}-\mathbf{T}_j)}. \tag{9}$$

The bandwidth, $\mathbf{H}$, defines the smoothness of the underlying distributions, which has a direct impact on the breadth of the final density estimate. Chen *et al.* compare and contrast three alternative bandwidth selection procedures, each based on a variant of leave-one-out cross validation [4]. For the purpose of the simulations that follow, the approach based least squares cross-validation is used to select a scalar bandwidth parameter [2]. Some of the earliest work on kernel density estimation, then referred to as the Parzen window method, can be found in [18].

One significant difference between this approach and the traditional Hotelling's T^2 approach is the flexibility in establishing confidence limits. Consider if there are 200 samples in the reference data set that is used both for building the PCA model and for assembling the kernel estimator. In this case, a 99% confidence limit for the data could be extracted by taking the third lowest value of $\hat{f}$. Alternatively, one could work under the assumption that all data within the reference data set are good, and set a 100% confidence limit to be equal to the maximum value of $\hat{f}$. While one may argue that confidence limits could be set for the Hotelling's T^2 in a similar manner, the presence of non-normal behavior would cause the limits to be overly extended, thus increasing the likelihood of type II errors.

IV. MIXTURE MODELS

A. General Mixture Models

While the methods described in section III have been proven effective in estimating the probability densities of non-normal distributions, they come with a rather substantial computational cost. Based on its implementation, every sample of historical data available should theoretically be added as a term in the overall density estimation model. However, due to the overlap of samples, the same distribution can be adequately described by a mixture model, which uses fewer terms, but weights them in order to generate a density estimate with the required characteristics [21]. The general mixture model, as applied for estimating the density of the principal component scores, takes the following form:

$$\hat{f}(\mathbf{t}) = \sum_{j=1}^{m} w_j \phi_j \left(\mathbf{t}|\mathbf{\Theta_j}\right). \tag{10}$$

As with the kernel equation, we are estimating the probability density, $\hat{f}$, for a given score vector, $\mathbf{t}$. However, rather than use all n samples of the reference data set for the estimation, a smaller number of m terms are used. Each term follows a preselected density distribution, ϕ_j, and is weighted by w_j, such that $\sum_{j=1}^{m} w_j = 1$. The distribution for each term is also parameterized by $\mathbf{\Theta}_j$, which for the case of a Gaussian mixture model includes a mean and covariance $(\mu_j, \mathbf{\Sigma}_j)$, yielding

$$\hat{f}(\mathbf{t}) = \frac{1}{(2\pi)^{l/2}} \sum_{j=1}^{m} w_j |\mathbf{\Sigma}_j|^{-1/2} e^{-1/2(\mathbf{t}-\mu_j)^T \mathbf{\Sigma}_j{}^{-1}(\mathbf{t}-\mu_j)}. \tag{11}$$

Therefore, rather than capture all samples in the reference data set and define a bandwidth matrix as in the kernel density approach, the mixture model requires that the number of terms, weights, and Gaussian parameters be defined at model build-time. The expectation-maximization (EM) algorithm has proven effective for fitting the weights and parameters for mixture models, with the only requirement being the a-priori knowledge of the number of terms in the mixture [7]. But because knowledge of the number of terms is often not readily known, a variant of Szewczyk's the time-evolving adaptive mixtures (TEAM) method will be implemented in this work [24].

B. Time-Evolving Adaptive Mixtures

The TEAM algorithm builds the mixture model by integrating the method of adaptive mixtures proposed by Priebe and Marchette [19] with the Dirichlet Process Priors approach of Escobar and West [10] and the iterative pairwise replacement algorithm (IPRA) of Scott and Szewczyk [21]. Adaptive mixtures is a method in which a mixture model is seeded with a single term centered at the value for the initial sample, and then iterates through the remainder of the sample population, creating new terms and updating existing terms along the way.

Given a precalculated population of scores, $\mathbf{T}$, the first step in the method is to center the initial term according to the first score vector, such that $\mu_1 = \mathbf{t}_1$. With only a single term at this point, that term takes all of the weight, $w_1 = 1$. The covariance for the first term must also be initialized. Given that we are dealing with a distribution of scores, the overall sample covariance can be used such that $\mathbf{\Sigma}_1 = \mathbf{\Lambda}$. Iteration through the remainder of the reference population is then performed on a sample-by-sample basis in order to build the mixture model.

As the iteration progresses for each remaining score in the reference population, $\mathbf{t}_k$, Dirichlet Process Priors are used to decide whether or not to add a new term. As originally formulated by Escobar and West [10] and demonstrated by Szewczyk [24], weights are calculated for all existing terms using

$$q_j^{(k)} = \phi\left(\mathbf{t_k}|\mu_j, \mathbf{\Sigma}_j\right) \tag{12}$$

These weights are then compared to a weight that is calculated from a new term generated from the current sample,

$$q_0^{(k)} = \gamma t_s\left(\mathbf{t}_k; \eta, M\right), \tag{13}$$

where $t_s(\mathbf{t}_k; \eta, M)$ is the t-density calculated at $\mathbf{t}_k$ with s degrees of freedom, mode η, and scale factor $M = (1+\tau)S/s$. A new term will be added if $q_0^{(k)} > q_i^{(k)}$, for all $i = 1, \ldots, m^{(k)}$. While the positive scalar, γ, has the largest influence on whether or not a term should be added, Szewczyk provides some additional guidelines for the selection of all parameters in order to maintain appropriate estimation while reducing the likelihood of adding redundant terms to the mixture [24].

If a new term is to be added, it is weighted according to an effective sample size, n, such that $w_{m^{(k)}+1}^{(k+1)} = 1/\min(k, n)$, centered at $\mu_{m^{(k)}+1}^{(k+1)} = \mathbf{t}_k$, and initialized with a covariance set to a preselected value. All other weights must be modified to maintain $\sum_{j=1}^{m^{(k)}+1} w_j^{(k+1)} = 1$. Although not applied in this work, the Markov Chain Monte Carlo (MCMC) method of Gibb's sampling can be used to set the parameters required for the new term [11].

Conversely, if a new term is not added, the following updating equations are used for the existing terms:

$$\rho_j^{(k+1)} = w_j^{(k)} \frac{\phi_j^{(k)}}{\hat{f}^{(k)}(\mathbf{t}_k)}, \tag{14}$$

$$w_j^{(k+1)} = w_j^{(k)} + \frac{1}{\min(k, n)}\left(\rho_j^{(k+1)} - w_j^{(k)}\right), \tag{15}$$

$$\mu_j^{(k+1)} = \mu_j^{(k)} + \frac{\rho_j^{(k+1)}}{w_j^{(k+1)}\min(k, n)}\left(\mathbf{t}_k - \mu_j^{(k)}\right), \tag{16}$$

$$\mathbf{\Sigma}_j^{(k+1)} = \mathbf{\Sigma}_j^{(k)} + \frac{\rho_j^{(k+1)}}{w_j^{(k+1)}\min(k, n)}\left(\mathbf{t}_k - \mu_j^{(k)}\right) \left(\mathbf{t}_k - \mu_j^{(k)}\right)^T. \tag{17}$$

The variable $\rho_j^{(k+1)}$ scales according to how closely the current observation matches term j, and it has a direct impact on all of the updating equations that follow. Terms that are near the current observation will cause $\rho_j^{(k+1)} > w_j^{(k)}$, which will increase the weight for that term based on (15), while also causing more significant updates to the mean and covariance as shown in (16) and (17). Conversely, terms that are further away from the new observation will have $\rho_j^{(k+1)} < w_j^{(k)}$, which will lead to a reduction in their weights and smaller changes on the parameters that characterize their distributions. Additionally, the presence of the $1/\min(k, n)$ term in the updating equations averages the results until n samples have been observed, and then continues to adapt the mixture model based on a moving window of the last n observations.

While the updating equations and initialization rules above are effective for maintaining a robust estimator with fewer terms than the kernel estimator, as time goes on there still exists a strong likelihood that redundant terms will be added to the model. Even though the terms may be distinctly separated in the initial stages of model development, there can be a tendency for terms to cluster around specific regions as time goes on. Weights may also slowly approach zero if the data drift substantially. Each additional term adds a certain amount of computational cost when performing density estimation, so to reduce this load, the iterative pairwise replacement algorithm (IPRA) is used [21].

After setting a limit on the number of terms allowed in the model, the IPRA algorithm will combine two terms into a single term in order to prevent that limit from being exceeded. In the case studies that follow, only a single pair of terms are combined when the threshold is reached. Alternatively, Szewczyk [24] applies IPRA several times over when the threshold is reached, which reduces the frequency of calls to IPRA.

As was the case for the kernel density estimation in Section III, standard rules do not exist for establishing the confidence limit, $\hat{f}_\alpha$, for adaptive mixtures. Rather, we must once again rely on historical observations to establish actionable thresholds for the density estimate. But as opposed to kernel density estimation, a set of historical observations is no longer maintained as part of the estimator. Recalling that an effect sample size, n, has previously been defined for the mixture model, it is now necessary to retain a vector of estimated densities for the last n observations. As before, confidence limits can be set by dividing this vector of observed estimates based on the desired quantiles (e.g. 95%, 99%, and 100%).

V. RAPID THERMAL ANNEAL CASE STUDY

To demonstrate how the aforementioned density estimation techniques can be applied with PCA for fault detection of non-normally distributed data, the rapid thermal annealing (RTA) process will be considered. RTA is a single-wafer process in which a series of halogen lamps are used to affect the electrical properties of the product. As opposed to furnace processes, RTA recipes are much shorter in duration, thus reducing the overall thermal impact on the wafers. While the single-wafer processing chambers are designed to improve uniformity across the wafer surface, variation from wafer-to-wafer can be a challenge. These wafer-to-wafer variability problems are particularly evident for the first few wafers processed after a significant idle time on the tool.

The data analyzed were gathered from a single chamber of an RTA tool. Mean values were calculated for 20 sensors across 5 recipe steps, for a total of 100 variables for each wafer. In addition to the chamber pressure, the temperatures and lamp voltages were also measured across multiple zones in the chamber. The PCA model and initial density estimators were trained with the first 498 wafers, while the remaining 3,608 wafers were used for testing. To enable the results of the study to be easily visualized, only two principal components are selected, even though seven components would have been chosen using the method of cross validation [27].

The principal component scores for the testing data set are provided in Fig. 1. The wafers in these plots are sorted chronologically, from earliest to latest. As shown, the scores experience some slow drifts and several shifts, which were the result of drifts and shifts in several of the individual variables. Such drifting behavior will be useful for demonstrating alternative adaptation strategies. Also, notice that the first wafers in these charts are denoted using the ● symbol. Chamber warm-up causes the principal component scores for the first wafers to be significantly different than the other wafers in the data set.

We can now compare the Hotelling's T^2 confidence limits with those for the estimators. The scores from the training data are displayed in both scatter plots in Fig. 2, along with the 95% and 99% normal confidence bounds based on Hotelling's T^2. The plots on the left and right then show the confidence bounds for the 498-term kernel estimator model and the 28-term mixture model, respectively. In the left plot for the kernel estimator, each of the training observations are shown as $\circ$'s, while all first wafers are further emphasized with $*$'s. The main sources of non-normality are thus the first wafers, and the samples taken from the single outlier lot shown in the bottom right corner of the plot.

It is readily observed that the elliptical Hotelling's T^2 limits do not adequately characterize the training data. The large regions within the ellipses where no observations are found could potentially lead to type I errors. Meanwhile, the presence of a large number of observations outside of the ellipses indicates that type II errors should be a concern, especially for first wafers. Now consider the limits based on the kernel estimation approach. These regions are much more successful at tightly conforming to the non-Gaussian behavior of the training data. Not only is the empty space within the regions substantially reduced, but it is done while retaining tolerance to the first wafers and the outlier lot.

The scatter plot on the right in Fig. 2 shows the limits from the estimator built from the training data using the TEAM approach. For the training model and the adaptation strategies described later in the section, the following settings were used: $s = 1$, $\eta = 0$, $\tau = 100$, $S = 5$, $\gamma = 1$, and $n = 60$. Again, notice how the confidence regions built using the density estimation approach more successfully enclose the training data than the ellipses based on Hotelling's T^2. When compared with kernel approach, the TEAM contours are much smoother around the edges, which is due both to the smaller number of terms used by the estimator and the tuning parameters that were chosen. While the kernel approach used all 498 observations in its estimation model, only 28 terms were captured using the TEAM algorithm. Each of these terms is centered at a $\circ$ in the figure. The distribution of the terms throughout the observed space was able to capture not only the first wafer effect, but also the many the small clusters of observations near the center.

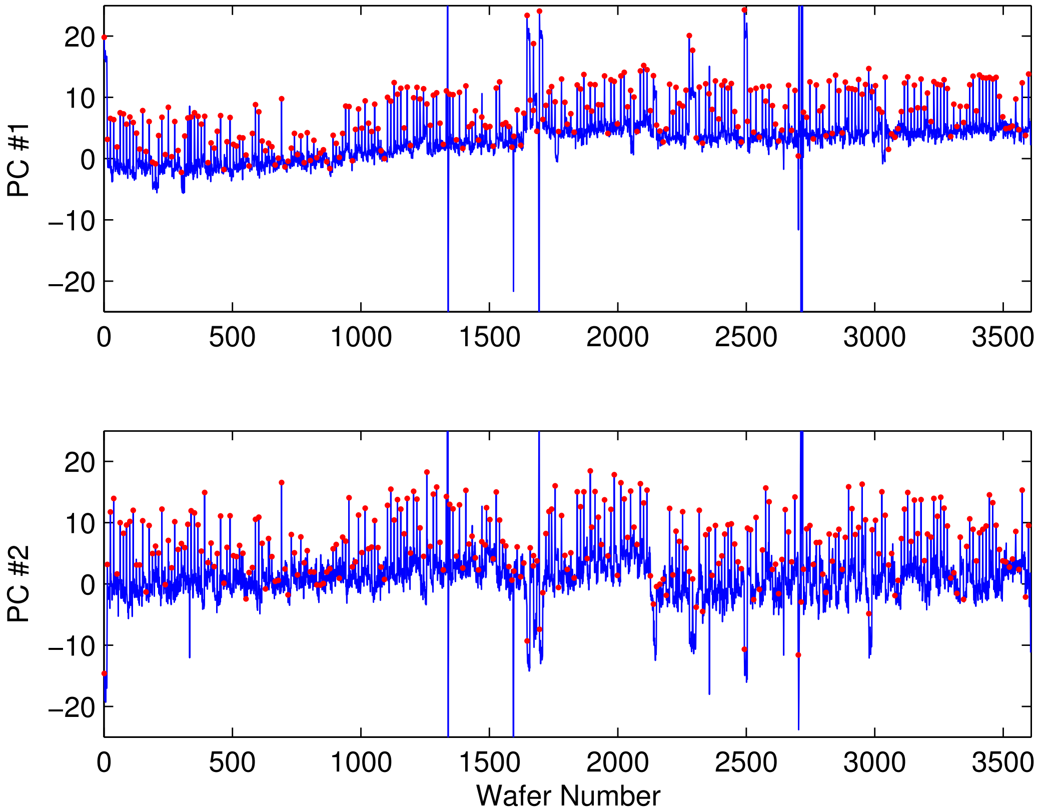

Fig. 1. Line charts of the drifting principal component scores for the test data. First wafers are denoted using the ● symbol.

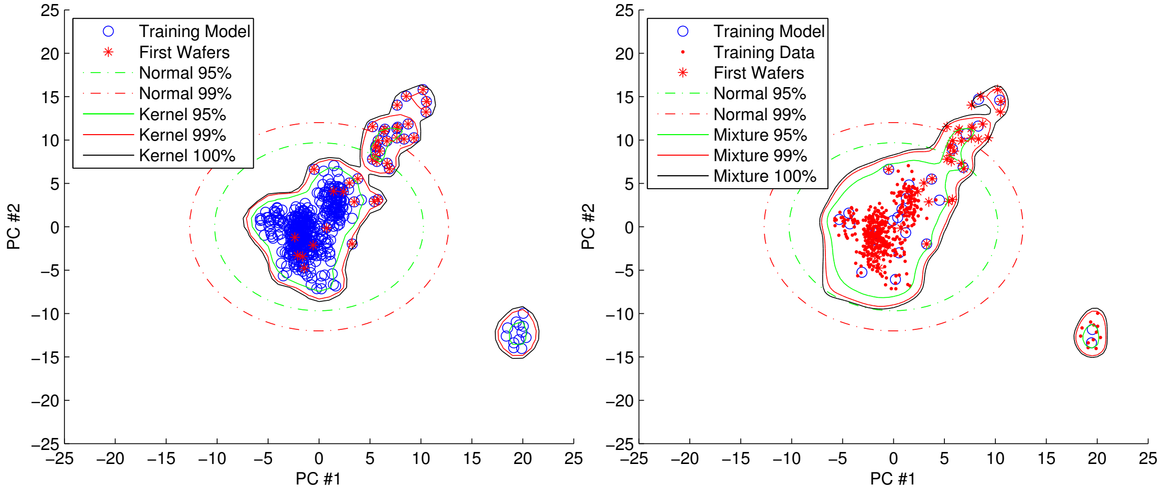

Fig. 2. Principal component scores for RTA training data, with confidence limits displayed based on Hotelling's T^2, the 498-term kernel estimator model (left plot), and the 28-term mixture model (right plot).

Now that it has been demonstrated that both KDE and the TEAM algorithms are both effective at describing the non-Gaussian behavior of the RTA training data, attention is now turned to the testing data, which have previously been shown to exhibit substantial drift with time. Several adaptation strategies will be demonstrated for the RTA case study.

A. Static Mode

The first adaptation strategy is one which is static, and thus does not adapt to any incoming samples. Such an approach is most appropriate if the process is expected to be stationary, and if the original reference data are representative of the entire allowable operating region of the process. In such a case, the static kernel and mixture estimators would detect any behavior that deviates from the region described by the reference data.

Fig. 3 provides snapshots of the last 498 samples of data superimposed on the static confidence regions for the kernel estimator and the TEAM estimator. By viewing the final snapshot figure it is clear that as time has progressed, the process has drifted steadily away from its initial region. Therefore, if such drifting behavior was undesired or unexpected, monitoring using the static estimation approach with the kernel or mixture method would have identified it. In contrast, observe that the confidence regions based on Hotelling's T^2, which did not closely match the initial population, would have been too wide to identify the drift in the main cluster of data.

While there are certainly many processes where tool drift would be a problem, such steady drift is not an issue for the RTA tool, which has a recipe that can be manipulated by run-to-run control and undergoes periodic maintenance to return the tool to its optimal processing conditions. For this reason, the errors that would have been triggered using the static approach would be considered false alarms in the manufacturing environment. Because the costs of investigating false alarms is high, and expectation that frequent false alarms will increase the likelihood of missed signals, clearly an alternative adaptation strategy would be preferred.

B. Selective Updating Mode

As a first attempt at an estimator that is able to adapt to the changing behavior of the process, a selective updating approach is proposed. To prevent the model from adapting to outliers, only samples that fall within the 100% limit are added to the model. For the kernel estimator, a moving window is used which updates with each new sample while maintaining a fixed number of terms. When each sample is added, the oldest term within the estimation model is removed to keep a total of 498 terms. Alternatively, the TEAM estimator adapts to new data based on its updating equations, initialization rules, and pairwise replacement algorithm, which were described above.

The results of applying the selective updating approach are shown in the final plot snapshots for KDE and TEAM in Fig. 4. As the estimators adapted to incoming data, two main observations can be made. First, the selective update strategy was very effective at capturing the gradual movement of the main cluster of the population for both KDE and TEAM. While the initial distributions were centered at the origin, by the time the final sample had arrived, they had shifted substantially to the right along the first principal component axis. The breadth of the confidence region for the TEAM method was somewhat greater than that for KDE, but this difference could have been reduced by modifying the tuning parameters for either method.

The second significant observation is that the estimators became very centralized quite quickly. This was due to the selective nature of the adaptation. Because not all samples were added to the model, and the majority of the samples fell in a region near the main cluster of data, the terms associated with first wafers were gradually neglected. For the KDE approach, the last of the first wafers was removed from the density estimation model around sample 1100. From that point forward, both estimators never again regained their tolerance to first wafers. Because the main purpose of applying density estimation techniques to the RTA data is to remain tolerant to first wafers, another adaption strategy must be considered.

C. Extreme Value Updating Mode

In the next adaptation strategy, only samples between the 95% and 100% limits are selected for adaptation. Again, the kernel estimator uses a moving window in which the oldest terms in the model are removed when new samples are added, while the TEAM model updates according to the algorithm defined above. Such a strategy is expected to be beneficial for processes that have a steady drift with time that must be tracked closely, but also have an infrequent sampling of non-normal behavior. By constraining updates to the most extreme values that are still within the allowable confidence region, it is expected that the tendency of the estimate to become excessively centralized will be reduced. As we saw earlier, a pure moving window with selective updating was ineffective at maintaining tolerance to such non-normal behavior.

As before, the results are presented with a plot of the final density estimate in Fig. 5. In contrast to the previous approach, these estimators were much more effective at maintaining tolerance to first wafers. At the same time, they were able to strike a good balance by also providing a reasonable adaption to the steady drift exhibited throughout the data series. In contrast to the static mode and the selective updating mode, this was substantially more effective at reducing false alarms.

D. Case Study Summary

The results of the case study are further summarized in Table I, which tabulates the number of alarms observed for each adaptation mode based on the 100% limit. Although no information was available to clearly indicate which alarms would actually be considered faults, it is reasonable to expect that the majority of alarms for the static mode can considered type I errors. Thus, by reducing the number of alarms with each adaptation strategy, it is expected that these false alarms would be reduced, thus providing a greater amount of confidence in the system's ability to detect truly faulty conditions.

	Overall		First Wafers	
Mode	KDE	TEAM	KDE	TEAM
Static	27%	14%	50%	45%
Selective Updating	11%	8%	65%	50%
Extreme Value Updating	5%	5%	30%	29%

TABLE I

ALARM FREQUENCY FOR RTA TESTING DATA BASED ON 100% CONFIDENCE LIMIT FOR ALL DENSITY ESTIMATOR ADAPTATION MODES.

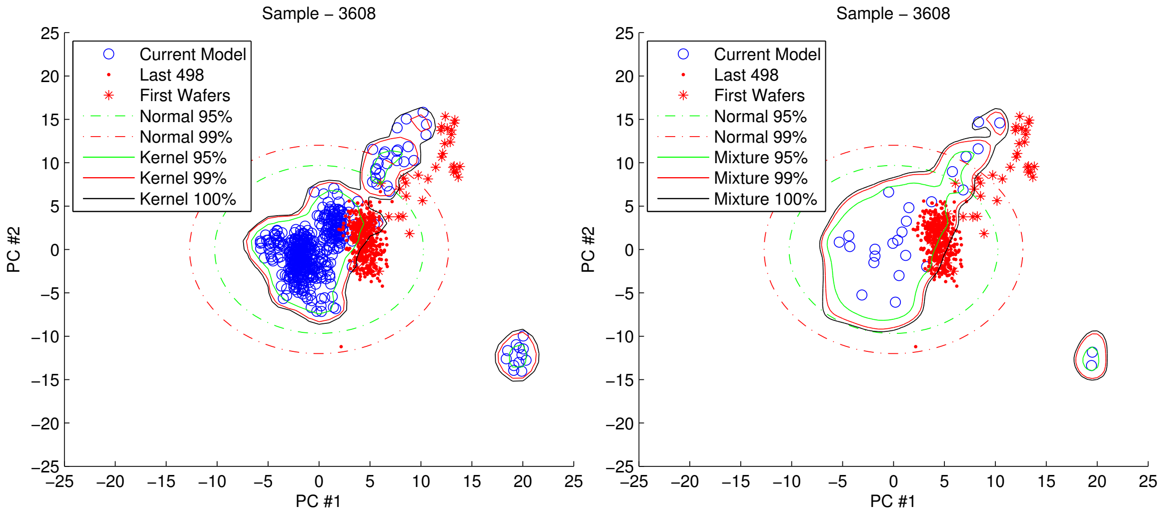

Fig. 3. Principal component scores for RTA testing data, with limits based on Hotelling's T^2 and the static kernel and mixture models.

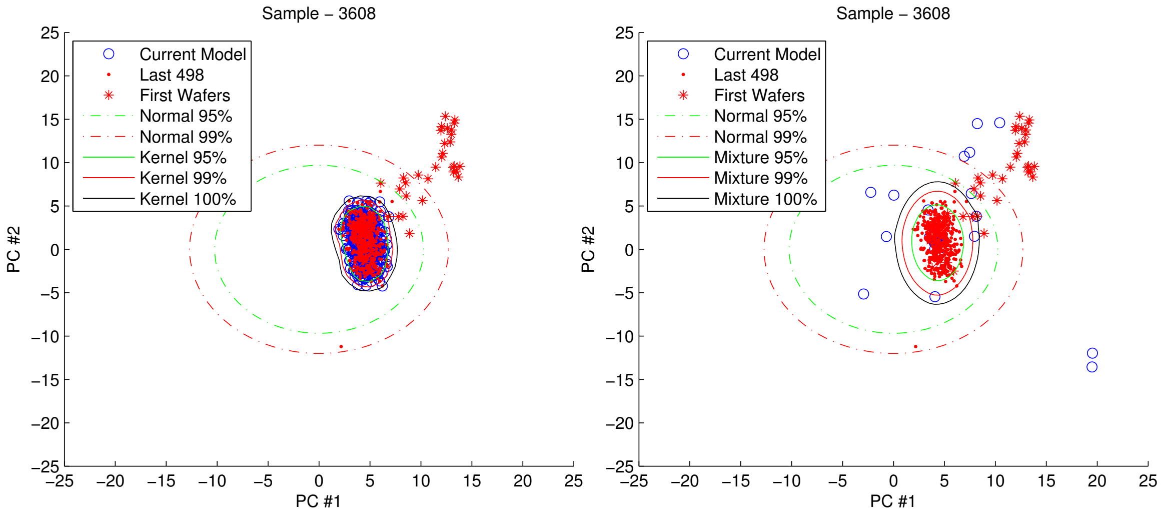

Fig. 4. Principal component scores for RTA testing data, with limits based on Hotelling's T^2 and the selectively updating kernel and mixture models.

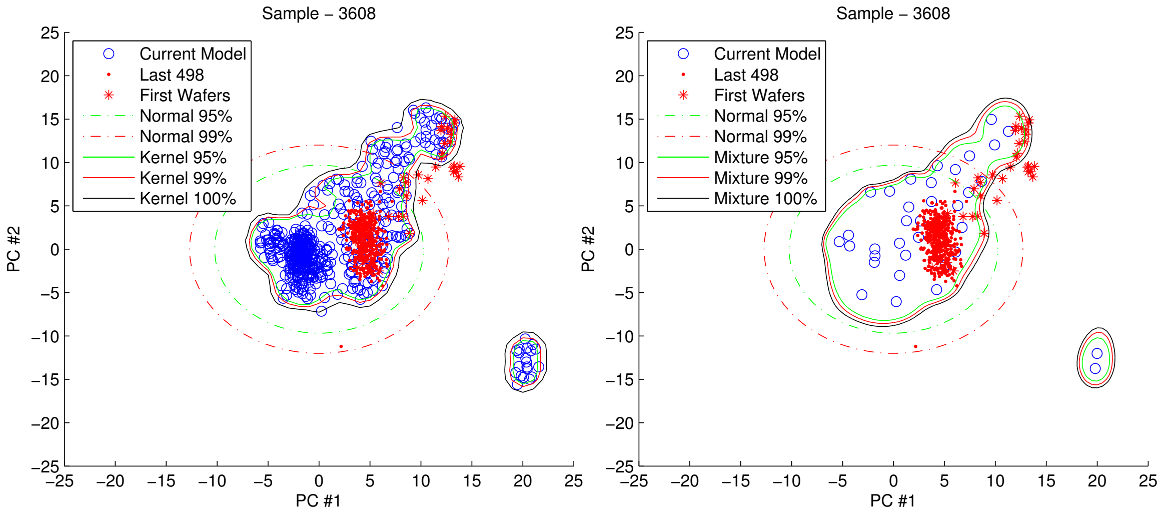

Fig. 5. Principal component scores for RTA testing data, with limits based on Hotelling's T^2 and the extreme value updating kernel and mixture models.

VI. CONCLUSIONS

In this paper, challenges were presented for monitoring semiconductor processing data that do not meet the typical MSPC assumption of multivariate normality. As originally presented by Martin and Morris [16], the kernel density estimator was used as an alternative to Hotelling's T^2 to provide control limits that more accurately reflect the distribution of the data. Additionally, the TEAM algorithm, which has been shown to be useful when estimating the densities of non-stationary data, was introduced as a more efficient method for monitoring the principal component scores.

As is the case in many semiconductor processes, another major difficulty is presented when the process being monitored exhibits a drift with time. In order for the system to be effective, it was emphasized that one must correctly characterize the nature of the drift and then decide on the adaptation strategy that reduces the number of type I and type II errors that are observed. Several adaptation strategies for both the kernel estimator and the mixture model were presented and applied to sensor data gathered from an RTA tool. The results of the study clearly illustrated the importance of not only the estimation algorithm itself, but also the rules that are used to decide which samples to include and which samples to ignore when performing the adaptation.

The results of the case study were presented using two-dimensional plots that overlay sample data with the density estimate contours. Because only two principal components were calculated from the RTA data, such plots could be used to effectively demonstrate the ability of KDE and TEAM to capture the non-normal behavior. However, if three or more principal components are required to adequately describe the main correlation structure in the data, another method of visualizing the results would be required. Just as SPC charts that track Hotelling's T^2 can be used to identify excursions based on any number of principal components in a single chart, we could similarly chart the calculated density estimate for each sample in order to take advantage of our non-Gaussian monitoring techniques.

VII. ACKNOWLEDGMENTS

The authors thank the sponsors of the Texas-Wisconsin Modeling and Control Consortium (TWMCC) for their continued support. Also, the Advanced Process Control (APC) Technology group at Advanced Micro Devices, Inc., is acknowledged for making the research possible.

REFERENCES

[1] G.G. Barna. Procedures for implementing sensor-based Fault Detection and Classification (FDC) for Advanced Process Control (APC). *SEMATECH Technology Transfer Document 97013235A-XFR*, 1997.

[2] A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360, 1984.

[3] J. Chen and J. Liu. Mixture principal component analysis models for process monitoring. *Ind. Eng. Chem. Res.*, 38:1478–1488, 1999.

[4] Q. Chen, R.J. Wynne, P. Goulding, and D. Sandoz. The application of principal component analysis and kernel density estimation to enhance process monitoring. *Control Eng. Practice*, 8:531–543, 2000.

[5] S.W. Choi, J.H. Park, and I. Lee. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Comput. Chem. Eng.*, 28:1377–1387, 2004.

[6] Y.H. Chu, S.J. Qin, and C. Han. Fault detection and operation mode identification based on pattern classification with variable selection. *Ind. Eng. Chem. Res.*, 43:1701–1710, 2004.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statistical Society*, Ser. B, 39:1–38, 1977.

[8] D. Dong and T.J. McAvoy. Batch tracking via nonlinear principal component analysis. *AIChE J.*, 42(8):2199–2208, 1996.

[9] F. Doymaz, J. Chen, J.A. Romagnoli, and A. Palazoglu. A robust strategy for real-time process monitoring. *J. Proc. Cont.*, 11:343–359, 2001.

[10] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Amer. Statistical Assoc.*, 90:577–588, 1995.

[11] A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *J. Amer. Statistical Assoc.*, 85:398–409, 1990.

[12] D.M. Hawkins. The detection of errors in multivariate data using principal components. *J. Am. Statist. Assoc.*, 69:340–344, 1974.

[13] F. Jia, E.B. Martin, and A.J. Morris. Non-linear principal components analysis with application to process fault detection. *Int. J. Sys. Sci.*, 31(11):1473–1487, 2000.

[14] W. Li, H. Yue, S. Valle, and J. Qin. Recursive PCA for adaptive process monitoring. *J. Proc. Cont.*, 10:471–486, 2000.

[15] K.V. Mardia. Mahalanobis distances and angles. In *Multivariate Analysis - IV*, pages 495–511. North-Holland, 1977.

[16] E.B. Martin and A.J. Morris. Non-parametric confidence bounds for process performance monitoring charts. *J. Proc. Cont.*, 6(6):349–358, 1996.

[17] P. Nomikos and J.F. MacGregor. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37(1):41–59, February 1995.

[18] E. Parzen. On the estimation of a probability density function and the mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.

[19] C.E. Priebe and D.J. Marchette. Adaptive mixtures: Recursive nonparametric pattern recognition. *Pattern Recognition*, 26:771–785, 1991.

[20] H.J. Ramaker, E.N.M. van Sprang, J.A. Westerhuis, and A.K. Smilde. The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chem. Intell. Lab. Sys.*, 73:181–187, 2004.

[21] D.W. Scott and W.F. Szewczyk. From kernels to mixtures. *Technometrics*, 43(3):323–335, 2001.

[22] A. Skumanich, J. Yamartino, D. Mui, and D. Lymberopoulos. Advanced etch applications using tool-level data. *Solid State Tech.*, 47(6):47–52, 2004.

[23] J.A. Smith, K.C. Lin, M. Richter, and U. LevAmi. Practical, real-time multivariate FDC. *Semicond. Int.*, 27(13):51–56, 2004.

[24] W.F. Szewczyk. Time-evolving adaptive mixtures. Submitted for publication in *J. Comput. Graph. Stat.*, 2005.

[25] H.W. Tong and C.M. Crowe. Detection of gross errors in data reconciliation by principal component analysis. *AIChE J.*, 41:1712–1722, 1995.

[26] B.M. Wise, N.B. Gallagher, S.W. Butler, Jr. D.D. White, 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*, 13:379–396, 1999.

[27] S. Wold. Cross validatory estimation of the number of components in factor and principal component analysis. *Technometrics*, 20(4):397–406, November 1978.

[28] H.H. Yue and S.J. Qin. Reconstruction based fault detection using a combined index. *Ind. Eng. Chem. Res.*, 40:4403–4414, 2001.

[29] H.H. Yue, S.J. Qin, R.J. Markle, C. Nauert, and M. Gatto. Fault detection of plasma etchers using optical emission spectra. *IEEE Trans. Semicond. Manuf.*, 13(3):374–385, 2000.