A Mixture Probabilistic PCA Model for Multivariate Processes Monitoring

Feng Zhang

Department of Industrial Engineering

Texas A&M University

College Station, TX 77843-3131

***Abstract* – A mixture probabilistic Principal Component Analysis (PCA) model is proposed as a multivariate process monitoring tool in this paper. High dimensional measurement data could be aggregated into some clusters based on the mixture distribution model, where the number of these clusters is automatically determined by the maximum likelihood estimation procedure. The multivariate statistical process monitoring mechanism is developed first with the learning of a finite mixture model for describing the local statistical patterns in each cluster, followed by the construction of the statistical process confidence intervals for the identified regions or nodes from T2 and Q charts. The abnormal input measurement would fall out of the acceptance region set by the confidence control limits and probabilistic PCA model. The experimental studies have illustrated that the mixture probabilistic PCA model conforms to the multivariate data well in the experiments involving Gaussian mixtures, and helps identify the underlying root causes of variation patterns in complicated multivariate manufacturing processes.**

I. INTRODUCTION

Due to the significant advances in in-process sensing technologies, it is not uncommon to collect vast quantities of information-rich measurement data for the potential use of monitoring and controlling. Automobile body assembly is a typical example of such complex manufacturing processes, where laser-optical dimensional stations for measuring automobile characteristic details are widely implemented at various assembly stages. An effective methodology is required to fully accommodate and utilize the automated in-process sensing technology to extract all quality-related diagnostic information from the inspection data. Fig. 1 illustrates the layout of 10 measurement points on the rear quarter panel subassembly of an automobile taken at the assembly process. The subassembly consists of a quarter panel, joined to a D-pillar reinforcement. The *y*/*z*-plane dimensional coordinates of five separate features are measured on each subassembly, providing a total of ten measured variables.

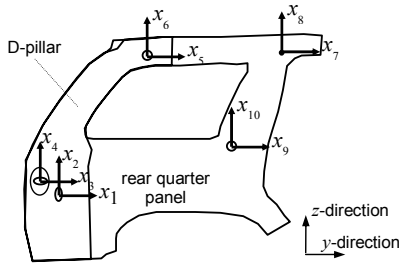

Fig. 1. Measurement layout on the quarter panel subassembly.

The assembly process involves hundreds of different locating elements. When a tooling element malfunctions, it will generally result some nonlinear variation pattern in the dimensional data.

When analyzing the in-process measurement data, one of the primary objectives is to identify the characteristics of any variation patterns that are present, which can provide insight into the major root causes of product-to-product dimensional variability. Constructing and interpreting the large array of 2D and/or 3D scatterplots that would be required for high-dimensional data is generally inconvenient. A more systematic approach to identifying a variation pattern is to represent and interpret the underlying data density structure. The incorporation of a suitable model describing the root causes that are clustered into some finite number of groups. The finite mixture of probabilistic PCA model is a good alternative to describe the measurement data under the above process descriptions.

Finite mixtures are a flexible and reasonable probabilistic modeling approach for the complicated multivariate data. Inferring the parameters of this mixture density model and determining the clusters from which the data are produced leads to the partitioning of observation data and identifying the manufacturing process.

The general method for fitting finite mixture models to observation data is the expectation-maximization (EM) algorithm [Dempster 1977], which converges to a maximum likelihood estimate of parameters. With too many components, the mixture model may over-fit the observation data, while data would be poorly and inflexibly expressed by a simpler model with too few components.

Thus the trade off between model complexity and data precise representation arises as an important issue in mixture modeling.

In this paper, we proposed a criterion for choosing the optimal number of clusters without intensive computations. The mixture of probabilistic PCA models is regarded as a good choice for finite mixture model paper with tractable and easier parameters estimation procedures. Once we explore the possibility of using finite complementary linear subspaces, clustering would be a straightforward strategy for separating and monitoring the manufacturing process variations. Finite mixture models permits the use of relatively simple models for each local data structure by offering excellent interpretation for observations, as well as the advantages of analytical and computational simplification.

II. FINITE MIXTURE MODEL

Principal component analysis (PCA) is a popular statistical tool for dimensionality reduction in multivariate processes analysis. For a set of $\{\boldsymbol{x}_n,\}$ with mean $\boldsymbol{\mu}$, the p principal components $\boldsymbol{e}_j$ are those orthonormal axes onto which the retained variance under projection is maximized.

The d-dimensional random vector $\boldsymbol{x}$ follows an M-component finite mixture distribution if its probability density function can be written

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^{M} \alpha_k \, p(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \tag{1}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_M$ are the mixing probabilities, and each $\boldsymbol{\theta}_k$ is the set of parameters defining the k^{th} component. The complete set of parameters $\boldsymbol{\theta} = \{\alpha_1, \ldots, \alpha_M, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ will be estimated to specify the finite mixture model with $\sum \alpha_k = 1$.

Given a set of N independent sample $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ the log-likelihood of $\boldsymbol{X}$ with respect to model (1) is

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}). \tag{2}$$

The optimal estimate of $\boldsymbol{\theta}$ is the ML estimate, $\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg\max_{\theta} \{\log p(\boldsymbol{X} \mid \boldsymbol{\theta})\}$, which cannot generally found analytically and depends on the specific density function forms $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$. The usual way for obtaining the ML estimates of mixture parameters is the EM algorithm [Figueiredo 2002, McLachlan 2000], which is an iterative procedure and converges to local maxima of log $p(\boldsymbol{X}|\boldsymbol{\theta})$. In EM algorithm, the observations $\boldsymbol{X}$ is interpreted as incomplete data, and $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$ is assumed to be complete data with missing data $\boldsymbol{Z}$. In model (1), $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$ is the binary label variables associated with each observation $\boldsymbol{x}_n$, indicating by which component the sample is produced. Thus, the updated log-likelihood is

$$\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{k=1}^{M} z_{nk} \log[\alpha_k \, p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)]. \tag{3}$$

The EM algorithm gives a sequence of estimate $\hat{\boldsymbol{\theta}}(t)$ (t = 0, 1, …) by alternating between expectation (E) step and maximization (M) step. Given the current estimate $\hat{\boldsymbol{\theta}}(t)$, the E step computes the conditional expectation

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t)) \equiv \mathrm{E}\left[\log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}(t)\right]. \tag{4}$$

The conditional expectation for z_{nk} is

$$w_{nk} = \mathrm{E}[z_{nk}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}(t)] = Pr\{z_{nk} = 1|\boldsymbol{X}, \hat{\boldsymbol{\theta}}(t)\}$$

$$= \frac{\hat{\alpha}_k(t) p(\boldsymbol{x}_n \mid \hat{\boldsymbol{\theta}}_k(t))}{\sum_{j=1}^{M} \hat{\alpha}_k(t) p(\boldsymbol{x}_n \mid \hat{\boldsymbol{\theta}}_j(t))}. \tag{5}$$

The M step updates the current $\hat{\boldsymbol{\theta}}(t)$ such that $\hat{\boldsymbol{\theta}}(t+1) = \arg\max_{\theta} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t))$. The convergence of the EM algorithm is theoretically guaranteed [Dempster 1977].

A complex nonlinear data density may be constructed by a collection of local linear models, with the advantage that each local model is easier to fit for data. The mixture of probabilistic PCA (MPPCA) model is appropriate for data modeling problem and usually requires two stages: first, a partition of the overall data into regions; second, fitting models by estimating the principal component within each region.

The major attraction of developing MMPCA model is that we can introduce a probabilistic data density for each component, called probabilistic PCA. We have assumed that the global data structure can be represented by finite linear component models. After the completion of clustering, each local subspace is then analyzed by PCA to learn the local statistical properties. Data in each cluster follow

$$\boldsymbol{x} = \mathbf{C}\,\boldsymbol{v} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{6}$$

where the latent variation source $\boldsymbol{v}$ are defined to be unit-variance independent Gaussians with loading matrix $\mathbf{C}$. The measurement error is a white noise, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. A key motivation of latent variable model is that, because of the isotropic structure of noise covariance matrix, the observed variable $\boldsymbol{x}$ are conditionally independently given the latent variables $\boldsymbol{v}$. And the dependencies between the measurement $\boldsymbol{x}$ are explained by a small number of latent variables. Thus, the probabilistic PCA model (6) effects a mapping from the lower latent space into the component subspace of the original high dimensional data

III. THE EM ALGORITHM FOR MPPCA

A. The EM Algorithm for MPPCA Model

Given the latent variable model (6) for each component, the log-likelihood takes

$$\log p(\boldsymbol{Y}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{k=1}^{M} z_{nk} \log[\alpha_k \, p(\boldsymbol{x}_n, \boldsymbol{v}_{nk}|\boldsymbol{\theta}_k)].$$

An iterative EM algorithm for optimizing the model parameters α_k, $\boldsymbol{\mu}_k$, σ_k^2 and $\mathbf{C}_k$ was developed [Tipping 1999]. Let $\pi_{nk} = \Pr\{z_{nk} = 1 \mid \boldsymbol{x}_n, \boldsymbol{\theta}(t)\}$.

$$\mathrm{E}\left[\log p(\boldsymbol{Y}|\boldsymbol{\theta})|\mathbf{X},\hat{\boldsymbol{\theta}}(t)\right]=\sum_{n=1}^{N}\sum_{k=1}^{M}\pi_{nk}\{\log\alpha_k$$
$$-\frac{d}{2}\log\sigma_k^2-\frac{1}{2}\mathrm{tr}(\mathrm{E}[\boldsymbol{v}_{nk}\boldsymbol{v}_{nk}^{\mathrm{T}}])-\frac{1}{2o_k^2}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2$$
$$+\frac{1}{\sigma_k^2}\mathrm{E}[\boldsymbol{v}_{nk}]^{\mathrm{T}}\mathbf{C}_k^{\mathrm{T}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)-\frac{1}{2\sigma_k^2}\mathrm{tr}(\mathbf{C}_k^{\mathrm{T}}\mathbf{C}_k\mathrm{E}[\boldsymbol{v}_{nk}\boldsymbol{v}_{nk}^{\mathrm{T}}])\}$$

The expectation and covariance of $\boldsymbol{v}_{nk}$ are $\mathrm{E}[\boldsymbol{v}_{nk}]=\mathbf{M}_k^{-1}\mathbf{C}_k^{\mathrm{T}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)$, and $\mathrm{E}[\boldsymbol{v}_{nk}\boldsymbol{v}_{nk}^{\mathrm{T}}]=\sigma_k^2\mathbf{M}_k^{-1}+\mathrm{E}[\boldsymbol{v}_{nk}]\mathrm{E}[\boldsymbol{v}_{nk}]^{\mathrm{T}}$, and $\mathbf{M}_k=\sigma_k^2\mathrm{I}+\mathbf{C}_k^{\mathrm{T}}\mathbf{C}_k$.

Then, we can get the following updates

$$\hat{\alpha}_k=\frac{1}{N}\sum_{n=1}^{N}\pi_{nk},\qquad \hat{\boldsymbol{\mu}}_k=\frac{\sum_{n=1}^{N}\pi_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N}\pi_{nk}},$$
$$\hat{\mathbf{C}}_k=\mathbf{S}_k\mathbf{C}_k(\sigma_k^2\mathbf{I}+\mathbf{M}_k^{-1}\mathbf{C}_K^{\mathrm{T}}\mathbf{S}_k\mathbf{C}_k)^{-1},$$
$$\hat{\sigma}_k^2=\frac{1}{d}\mathrm{tr}(\mathbf{S}_k-\mathbf{S}_k\mathbf{C}_k\mathbf{M}_k^{-1}\hat{\mathbf{C}}_k^{\mathrm{T}}) \qquad (7)$$

where $\mathbf{S}_k$ is the sample covariance matrix for the k^{th} component.

B. Iterative Component Pruning

An important aspect in the proposed EM algorithm is to obtain the optimal number of components M, since ML criterion itself can not automatically determine M. Let $\mathcal{M}_k$ be the class of all possible k-component mixture with each component built from the probabilistic PCA model (6). The classes $\{\mathcal{M}_k\}$ are nested since $\mathcal{M}_k \subseteq \mathcal{M}_{k+1}$. Consequently, the log-likelihood function of $\boldsymbol{X}$ is a nondecreasing function of component size, and ML could be useless for estimating the number of M.

Some deterministic methods have been proposed for model selection by discarding a least probable component with few data points assigned to it [Figueiredo 2002]. However this criterion may not conform to the data structure when there are some components with much less data than the others. In this paper we introduced an iterative pruning method incorporating searches for the number of components. The measure $J(i, j, \boldsymbol{X})$ is actually an inner product of two possibility vectors, $J(i,j)=[\pi_{1i}\ \ \pi_{2i}\ldots\pi_{ni}][\pi_{1j}\ \ \pi_{2j}\ldots\pi_{nj}]^{\mathrm{T}}$, $1\le i<j\le M$. The larger J is, the more probable that i and j are produced by a single component. And for each iterative procedure, choose the components i and j with the largest J for component pruning.

After removing one component, the MMPCA model is re-trained until convergence by the EM algorithm. We need to take into account the decrease of log-likelihood resulting from the deletion of one component, since log-likelihood is nondecreasing on M. To accommodate this, a penalty term is introduced in the cost function which is a minimum message length based function penalizing larger value of component size [Wallace 1999].

IV. MULTIVARIATE SPC CHARTS

The EM algorithm combined with the iterative component pruning criterion allows us to identify each cluster of component in $\boldsymbol{x}$. Since each component is defined as a probabilistic PCA model, the multivariate processes inputs would fall into an accepted region if they are under the normal operation conditions. This forms the basis of applying multivariate statistical processes control (SPC) charts to monitoring the complex processes. The commonly used statistics T^2 and Q are two popular statistical tools for detecting faults.

Statistics T^2: The scalar measure as the sum of square of principal scores

$$T^2=(\boldsymbol{x}-\boldsymbol{\mu})\mathbf{U}_p\boldsymbol{\Lambda}_p^{-1}\mathbf{U}_p^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}, \qquad (8)$$

where $\boldsymbol{\Lambda}_p$ is the diagonal matrix containing the largest p eigenvalues.

Statistics Q: The scalar measure for the goodness of fit of sample data $\boldsymbol{x}$ to the probabilistic PCA model, defined as the sum of square of the residuals

$$Q=\boldsymbol{x}^{\mathrm{T}}(\mathbf{I}-\mathbf{U}_p\mathbf{U}_P^{\mathrm{T}})\boldsymbol{x}. \qquad (9)$$

In the probabilistic PCA model, the residuals are shown to be noise.

These two statistics T^2 and Q, can be used for monitoring multivariate processes. The $100(1-\alpha)\%$ confidence control limit Q_α is calculated directly from the assumption of probabilistic PCA component model, i.e., $Q_\alpha=\sigma^2\chi^2_{\alpha,d-p}$. The control limit T_α^2 for T^2 is :

$$T_\alpha^2=\frac{p(N-1)}{(N-p)}F_{\alpha,p,N-p}, \qquad (10)$$

where N is the sample size in each component and $F_{p,N-p}$ denotes the F distribution [Jackson 1991]. Usually the type I error probability α takes 0.99 or 0.95.

In view of process monitoring, the system is represented by a MMPCA model with M clusters of components. Each cluster performs a particular decision function on incoming process measurements with a set of estimated parameters pertaining to this component. Therefore, the final decision function outputs $\boldsymbol{O}_k(\boldsymbol{x})$ indicating the input variable $\boldsymbol{x}$ falls into accepted region, can be expressed as

$$O_k(\boldsymbol{x})=\begin{cases}\text{accept}, & \text{if } Q_k\le Q_{k\alpha} \text{ and } \mathrm{T}_k^2\le \mathrm{T}_{k\alpha}^2\\ \text{reject}, & \text{othereise}\end{cases},$$

where Q_k is the sum of squares of the residuals for cluster k and T_k^2 is the sum of squares of principal scores for k.

To build a good monitoring framework, it is the first step to establish a consistent and accurate mixture models with the historical data sets collected under normal process operations. The cluster decision function output based on these two statistics is accepted when the corresponding system input pattern fall in to the ellipsoid enveloped by $(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{S}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)=c$. On the other hand, input pattern

which is far away from the ellipsoidal center yield a rejection output signal.

V. Experimental Studies

A. Mixture of Probabilistic PCA

A 3-D synthetic data set was generated from 3 components with 200 data points in each cluster. We investigated the proposed EM algorithm on the synthetic data set, where each clustering is equipped with $p = 1$ principal axis. The clustering probabilities are $\pi_1 = \pi_2 = \pi_3 = 1/3$, the means are $[2\ 1\ .25]^T$, $[1\ 1.5\ .25]^T$, $[.25\ 2\ .25]^T$, and equal covariance matrices with noise $\sigma = 0.1$. The estimates for parameters from the EM algorithm ere $\hat{\pi}_1 = .312, \hat{\pi}_2 = .351, \hat{\pi}_3 = .337$, and the respective noise estimates in the three clusters are .096, .115 and .095.

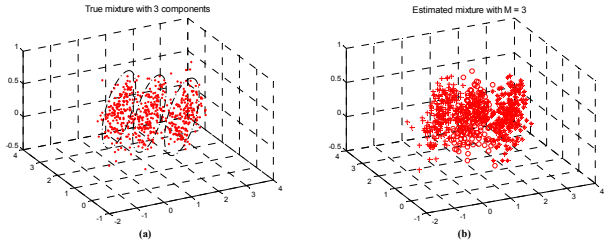

Fig. 2. Fitting a mixture probabilistic PCA data set in 3-D space with $p = 1$. (a) The true mixture with dashed ellipses indicating the components. (b) The final estimated clusters with $M = 3$, the data points in each components are denoted by "+","o" and "*" respectively.

The experiment verified that the cost function (9) approached its minimum at $M = 3$, which is the successful estimate. We arbitrarily chose a large enough initial guess for M, and then using iterative component pruning to search for the optimal component size. $L(\boldsymbol{\theta}(M), \boldsymbol{y})$ as a function of components size M was plotted in Fig. 3, showing that data mixtures with $M > 3$ and $M = 1, 2$ have higher values and were discarded consequently. Therefore, the proposed EM algorithm can automatically select the number of components given the observed sample data.

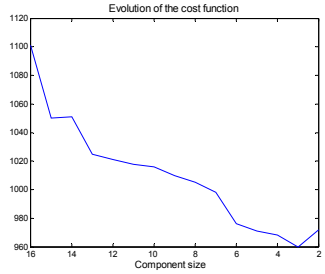

Fig. 3. *Evolution* of the cost function $L(\boldsymbol{\theta}(M), \boldsymbol{y})$.

B. Autobody Assembly Process Example

The automobile assembly example introduced in Section 1, provided a prime application of multivariate data density modeling and on-line process monitoring.

In Fig. 1, the z-direction denotes the up/down direction, and the y-direction denotes the left/right direction. The measurements are obtained after the subassembly is joined to the bodyside and represent the deviation of the coordinates from their nominal positions. A failure in any tooling element may be a root cause of product dimensional variability. And several different groups of tooling faults may be present in different subassembly stations in autobody assembly line. One possibly root cause was found to be a loose tooling element in the fixture that is used to locate the quarter panel subassembly. The loose tooling element allowed the subassemblies to rotate by small amounts about the x_9/x_{10} feature after being placed into the fixture. When a subassembly was subsequently clamped into place and welded to the bodyside, it retained the incorrect, rotated position. The subassemblies would rotate clockwise on some automobiles and counter-clockwise on others. Another variation pattern may be caused by a z-direction translation of the D-pillar with respect to the quarter panel. A design flaw in the fixture used to locate the D-pillar when it is joined to the quarter panel allowed the D-pillar to translate freely by small amounts in the z-direction.

One of the keys to diagnosing faults is the incorporation of a suitable model describing the root causes that are clustered into a finite number of groups. The MPPCA model appears to represent the high dimensional data effectively. For the clustered data resulting from distinct root causes, they are assumed to obey model (6).

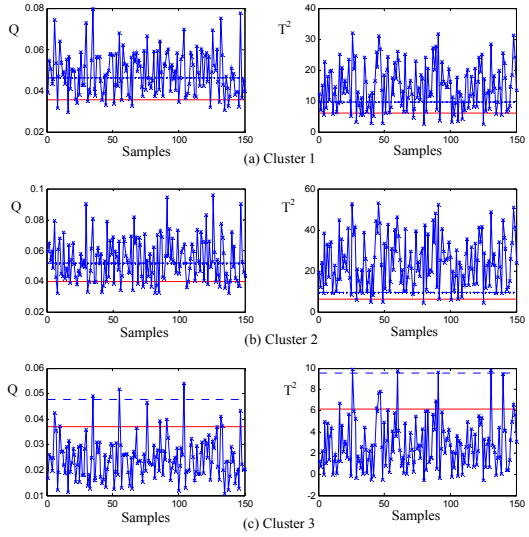

Fig. 4. The multivariate monitoring Q and T2 charts for 3 distinct clusters labeled (a)-(c) respectively, each contain 95% (solid line) and 99% (dotted line) confidence limits.

Using the proposed EM algorithm on the sample with $d = 10$ and $N = 460$, it estimated that there are three distinct clusters existing in $\boldsymbol{x}$ with $p = 2$ for each component. The control limits, $\boldsymbol{T}_\alpha^2$ and Q_α of each component are calculated respectively as shown in Fig. 4.

Applying the Q and T^2 multivariate SPC charts on 150 new process sample data detected some out-of-control signals, which were identified to be generated from the third root cause. The reason is obvious by observing the three in-process monitoring charts shown in Fig. 4. Since

for cluster 1 and 2, most Q and T^2 values of these on-line measurement data exceeded the corresponding control limits greatly, however, by inspecting the control charts for cluster 3, only several measurement data points yielded fault signals.

The results in this real problem example indicated that the MPPCA model has the capability of detecting the variability faults in the autobody assembly processes involving high-dimensional measurement data, and identifying the underlying root causes by classifying them into distinct components. Although we illustrated with a real example from automobile assembly process, similar variability analysis and process monitoring scenarios exist in many other manufacturing applications.

VI. Conclusion

The proposed mixture probabilistic PCA model takes the advantages of PCA and finite mixture models, contributing a lot to compressing noise and extracting the complicated process features to form a set of simpler and smaller informative clusters for analysis. So it offers a suitable alternative for multivariate processes monitoring and fault detection.

The developed method for learning finite mixture models is able to determine the proper number of components without much intensive computational searching. It can also avoid the drawback of standard EM algorithm which is sensitive to initialization. The derived EM algorithm adjusted each component's eigenvectors, noise and dimension to the local data structure. A real example from the automobile assembly process has demonstrated the potential usage of this data modeling method for multivariate manufacturing process diagnosing and monitoring.

VII. References

[1] C. Archer and T. K. Leen, "Building Better Clusters with Adaptive PCA," submitted to *Neural Computation*, 2002.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc. (B)*, 39, 1-38., 1977.

[3] Mario A. T. Figueiredo and A. K. Jain "Unsupervised Learning of Finite Mixture Models," *IEEE Trans. on P.A.M.I.*, 24(3), 381-396, 2002.

[4] C. E. McCulloch, "Maximum Likelihood Algorithm for Generalized Linear Mixed Models," *J.A.S.A.*, 92,162-170, 1997.

[5] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley & Sons, 2000.

[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290, 2323-2326, 2000.

[7] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation*, 11(2), 443-482, 1999.

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