IIE Transactions (2005) **37**, 569–584

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170590929036

Automatic feature selection for unsupervised clustering of cycle-based signals in manufacturing processes

SHIYU ZHOU[1,*] and JIONGHUA (JUDY) JIN[2]

[1] *Department of Industrial and Systems Engineering, The University of Wisconsin-Madison Madison, WI 53706, USA*
E-mail: szhou@engr.wisc.edu
[2] *Department of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA*
E-mail: jhjin@sie.arizona.edu

Received and accepted

Recent developments in sensing and computer technology have resulted in most manufacturing processes becoming a data-rich environment. A cycle-based signal refers to an analog or digital signal that is obtained during each repetition of an operation cycle in a manufacturing process. It is a very important class of in-process sensing signals for manufacturing processes because it contains extensive information on the process condition and product quality (e.g., the forming force signal in forging processes). In contrast with currently available supervised classification approaches that heavily depend on the training dataset or engineering field knowledge, this paper aims to develop an automatic feature selection method for the unsupervised clustering of cycle-base signals. First, principal component analysis is applied to the raw signals. Then a new method is proposed to select information containing principal components to allow clustering to be performed. The dimension of the problem can be significantly reduced through the use of these two steps. Finally, a model-based clustering method is applied to the selected principal components to find the clusters in the cycle-based signals. A numerical example and a real-world example of a forging process are used to illustrate the effectiveness of the proposed method. The proposed technique is an important data pre-processing technique for the monitoring and diagnostic system development using cycle-based signals for manufacturing processes.

1. Introduction

Due to recent developments in sensing and computer technology many process variables can now be measured online and automatically stored during production to allow manufacturing process monitoring and control. Among the currently available process variables, the *cycle-based* sensing signal is a very important class of signals in many manufacturing processes because it contains extensive information that is related to both product quality and process variables. As the name implies, a cycle-based signal is an analog or digital signal that is obtained using automatic sensing during each repetition of an operation cycle in a manufacturing process, for example, the tonnage signal (forming force) measured by strain sensors installed on a forging press machine. Figure 1(a) illustrates the tonnage signals of two production cycles that are sampled with respect to the crank angle of a forming press machine. This signal contains 224 data points in each cycle where the vertical axis is the forming force measured in tons, and the horizontal axis is the crank angle of the press. In Fig. 1(b), the two individual plots are superimposed to highlight the similarities and differences between these two cycles. In our case, the cycle-based signals are aligned naturally based on the crank angle. When the natural index is not available, a statistical method such as registration in functional data analysis can be used to align the cycle-based signals (Ramsay and Silverman, 1997).

In a forging process, the mean profile of the tonnage signals is determined by the physical process setups and working conditions i.e., material properties, workpiece geometry, press shut height, and press speed. Variations in these signals are inevitable due to natural process variations caused by the effects of inherent process disturbance factors, such as randomness in lubrication distribution and material uniformity, etc. For example, in Fig. 1(b), the observed difference in the peak of the tonnage signal could be due to a change in the process condition, such as insufficient lubrication, and differences at other parts of the tonnage profiles could reflect inherent natural process variations. Therefore, the tonnage signal contains extensive information on the forging process working condition and product quality.

These cycle-based signals exist not only in forging processes but also in many other kinds of manufacturing processes, e.g., the forming force in stamping processes, the

*Corresponding author

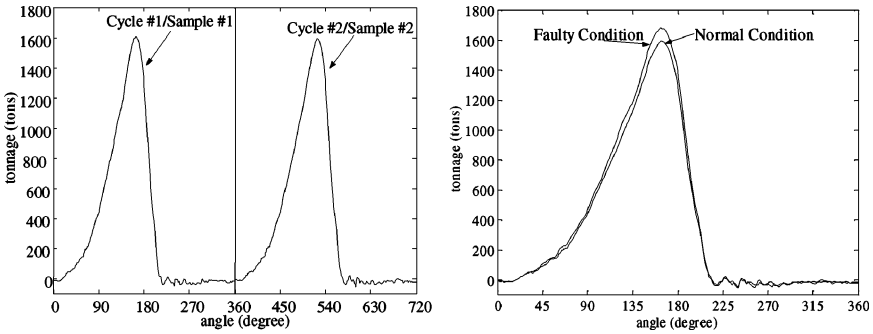

Fig. 1. (a) The forging tonnages of two cycles; and (b) superimposed plots for two cycles.

holding force and current signal in spot welding processes, the insertion force in engine assembly processes. Therefore, it is of considerable importance to develop a generic methodology for process monitoring and diagnosis that utilizes cycle-based signals.

Due to the complexity of analyzing high-dimensional signals, people often only use very simple statistics to characterize the signals and perform monitoring studies in industrial practice. For example, the maximum magnitude and the average value of the signal are the most commonly used statistics (Knussmann and Rose, 1993; Grogan, 2002). In these methods, a large portion of the information contained in the signals is not fully explored. Therefore, monitoring systems based on these simple statistics often suffer from high false-alarm rates and/or poor detection rates for various types of faulty conditions.

There are some reports in the literature on fully utilizing cycle-based signals for process monitoring and diagnosis purposes. Jin and Shi (1999) proposed a "feature-preserving data compression" procedure to reduce the data dimension after the use of a wavelet transformation on the tonnage signals of stamping processes. Their feature preserving criteria heavily depend on a pre-knowledge of the stamping process. Koh *et al.* (1999a, 1999b) introduced a uniformly most powerful test for individual coefficients in a Haar transformation of cycle-based signals. Based on this test, a monitoring system that is similar in nature to a Shewhart control chart is proposed to distinguish between normal and abnormal conditions in a process using cycle-based signals. This method is only effective if the correlations between the Haar coefficients are small and can therefore be neglected. If this is not the case then a large number of false alarms will be generated. For the root cause identification problem, Jin and Shi (2000) used a fractional factorial design of experiments approach to study the relationships between process variables and the variation patterns of the cycle-based signals. Pittner and Kamarthi (1999) proposed a wavelet-based procedure for feature extraction of signals. They transformed the signals into the wavelet domain and then selected wavelet coefficients based on the magnitude of the coefficients. This approach is different to the one we propose in which the differences between cycle-based signals are the most interesting characteristics. We wish to keep those coefficients that represent differences rather than simply selecting in terms of a large magnitude. Lada *et al.* (2002) have proposed a wavelet coefficient selection procedure that is not only based on the magnitude of the coefficients as in Pittner and Kamarthi (1999) but is also based on an additional term that penalizes the number of selected coefficients. Its purpose is to keep the number of wavelet coefficients small and hence simplifying subsequent analysis.

From the above review it is clear that currently available cycle-based signal analysis techniques either depend on engineering field knowledge, the availability of training samples, or the magnitude of the features in the signal. Little attention has been focused on the unsupervised clustering of cycle-based signals, which can automatically extract key features in the signals and group the signals into different clusters. Jin and Shi (2001) have proposed a method to develop a monitoring and diagnostic system that is able to achieve automatic learning and continuous improvement of the diagnostic performance. The proposed method includes automatic fault detection, optimal feature extraction, optimal feature subset selection, and diagnostic performance assessment. An exhaustive search method is used to conduct the feature selection. For a system with a large number of extracted features, the computational load of the exhaustive feature search becomes excessive.

In this article, we focus on feature selection for the unsupervised clustering analysis of cycle-based signals, which is very important in process monitoring and diagnostic system development. The clustering of information will lead to the following benefits: (i) the cluster information can help to discover changes in working conditions during continuous daily production runs; (ii) the production performance can be assessed according to the discovered clusters, thus enriching process knowledge and help to discover new optimal process conditions; (iii) the cluster information can be used as supervisory training sets in the monitoring and diagnostic system development. Since further analysis of the clusters and their associated working conditions can

reveal new knowledge for process monitoring, diagnosis, and improvement, the technique proposed in this article can also be considered as an important data pre-processing technique for data dimension reduction in the development of monitoring and diagnostic systems using cycle-based signals.

For the clustering of a high-dimensional dataset, the first step is often to reduce the data dimension. Several dimension reduction techniques have been developed in recent years including those of Carreira-Perpinan (1997) and also Tanaka and Mori (1997). However, some of these techniques, such as multi-dimensional scaling and projection pursuit techniques, are difficult to apply to the complex high-dimensional datasets encountered when attempting to find an optimal stress function or projection transformation. In other data dimension reduction methods, such as principle component analysis and the self-organizing map method, the feature selection techniques focus more on the faithful representation of the original data, instead of clustering (Duda *et al.*, 2001). The focus of this paper is on how to effectively select transformed features/variables so as to reduce the data dimension for the clustering. There is an extensive literature on variable selection in multiple regression and supervised classification (Draper and Smith, 1980; Guyon and Elissee, 2003). However, few results have been presented on feature selection in unsupervised clustering analysis. Fowlkes *et al.* (1987) have proposed a forward step-wise variable selection procedure for a hierarchical clustering technique but it cannot guarantee the global optimality, especially for a high-dimensional cycle-based signals with a complex correlation structure. Recently, Liu *et al.* (2003) developed an algorithm for the simultaneous feature selection and clustering under a Bayesian framework. The technique is quite effective but is somewhat complex and is computationally expensive.

In this paper, we propose a new technique for variable selection and clustering analysis of cycle-based signals. First, the cycle-based signals are modeled as a mixture of high-dimensional normal distributions. Each working condition or process fault corresponds to one component in this mixture model. Then, Principal Component Analysis (PCA) is used to find the major directions of the signal variations. A new strategy to pick out the principal components that contain the most information is proposed for the purpose of clustering. Based on this strategy, a subset of the principal components is selected for further analysis. Hence, the dimension of the problem is significantly reduced. Finally, a clustering algorithm is applied to the selected principal components to group the cycle-based signals. This technique can automatically find the clusters in a set of cycle-based signals in an unsupervised manner.

This article is organized as follows. In Section 2, the problem is formulated and an overview of the proposed method is presented. In Section 3, the key steps in the clustering, that is, PCA and the strategy of selecting of information containing principal components, are presented. Section 4 presents both a numerical simulation and a real-world example of a forging process to illustrate the effectiveness of this method. The conclusions are presented in Section 5.

2. Problem formulation

2.1. *The modeling of cycle-based signals*

A cycle-based signal can be viewed as a high-dimensional multivariate vector whose dimensionality is equivalent to the number of data points within one cycle-based signal. For example, for the tonnage signals in Fig. 1(a and b) there are 224 data points per signal. Therefore, the tonnage signal can be viewed as a 224-dimensional vector. Clearly, this multi-dimensional vector is highly correlated over its 224 components. The collected samples of cycle-based signals under a fixed working condition are assumed to be independent and to follow an identical multivariate normal distribution. The covariance of the signal could be in an arbitrary structure, i.e., the different data points of a cycle-based signal can be correlated. In most cases, the multivariate sensing signal is affected by many process factors. For example, the tonnage signal in a forging process is affected by various factors such as the temperature of the workpiece, the shut height adjustment, lubrication, wear of the press, etc. Inevitably these factors may randomly vary around their individual nominal values. From the central limit theory, the distribution of the tonnage signals will tend to be a normal distribution. In this paper, for different working conditions which correspond to different process parameters, we assume they still follow a multi-dimensional normal distribution but with different parameters (means and variances).

If we denote $\mathbf{X}^{n\times p}$ as a set of samples of cycle-based signals, where n is the number of samples and p is the dimension of the signals, and denote $\mathbf{x}^T$ as a row vector of $\mathbf{X}^{n\times p}$ (one sample of cycle based signals), then at the kth working condition, the distribution of $\mathbf{x}$ is given as:

$$\phi_k(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}, \quad (1)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and the covariance of the measurements of the kth working condition, respectively.

Assume that there are q different working conditions existing in the collected set of signals $\mathbf{X}^{n\times p}$. Then, the distribution of $\mathbf{X}^{n\times p}$ can be represented as a finite mixture of normal distributions (McLachlan and Peel, 2000).

$$f_{\mathbf{x}}(\mathbf{X} \mid \boldsymbol{\mu}_1, \Sigma_1, \ldots, \boldsymbol{\mu}_q, \Sigma_q) = \prod_{i=1}^{n} \sum_{k=1}^{q} \tau_k \phi_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k) \quad (2)$$

where the $\tau_k, k = 1 \ldots q$, are positive numbers between zero and one that represent the occurrence probability of each working condition, thus, $\Sigma_{i=1}^{q/y} \tau_k = 1$.

2.2. *Proposed clustering method*

The objective of this paper is to provide an automatic feature extraction method for effective clustering of cycle-based signals under unknown working conditions. There is a huge body of literature on clustering techniques (Anon, 1989). They can be roughly divided into two categories: (i) non-parametric hierarchical clustering methods (Arabie *et al.*, 1998); and (ii) model-based clustering methods (McLachlan and Basford, 1988). In this paper, since the data is modeled as a mixture of normal distributions, the model-based clustering approach in which each component in the mixture model (ϕ_k in Equation (2)) represents a cluster in the data will be used.

The steps involved in clustering cycle-based signals are illustrated in Fig. 2. Since the signals have a high dimensionality and they will suffer the "curse of dimensionality" (i.e., the sample size needed to estimate the density function is proportional to the exponential of the number of dimensions, see the discussions is Carreira-Perpinan (1997) and Jimenez and Landgrebe (1998)), it is very difficult to apply directly clustering algorithms to the raw dataset.

In this paper PCA (Jackson, 1991) is used to reduce the dimension of the dataset as the first step in the data transformation. PCA linearly transforms the raw dataset into a new set of variables, called Principal Components (PCs). Given a dataset $\mathbf{X}^{n\times p}$ with p variables and n samples and $\mathbf{S}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, the ith principal component is given by:

$$y_i = \mathbf{e}_i^T \mathbf{x} = e_{i1}x_1 + e_{i2}x_2 + \cdots + e_{ip}x_p, i = 1 \ldots p, \quad (3)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $\mathbf{x}^T$ is a row vector of one sample of the p variables. Also, the sample variance of y_i is λ_j, $i = 1 \ldots p$, and the sample covariance between y_i and y_j is zero for $i \neq i$. In addition, the total sample variance is trace($\mathbf{S}$) that is equal to $\lambda_1 + \lambda_2 + \cdots + \lambda_p$, where trace($\mathbf{S}$) is the summation of the diagonal elements of $\mathbf{S}$. The sample variance explained by the ith principal component is given by λ_i/trace($\mathbf{S}$).

Since the PCs are linear combinations of the original variables, they will also follow a mixture of normal distributions. Another point worth mentioning is that PCA is an invertible transformation. All the variation information of the dataset is captured in the PCs and the eigenvectors of the covariance matrix of the dataset. In practice, after the PCA transformation, people often simply select the first few PCs with the largest eigenvalues and then apply clustering operations to these selected PCs. The rationale behind this practice is that the within-cluster variation is often smaller than the between-cluster variation. Therefore, it is expected that the clustering structure will show up in the first few PCs that have the largest eigenvalues. However, although it is usually true that PCs with a very small variation do not contain much information, and hence can be treated as noise, it is not always true that PCs with a large variation contain useful information for clustering (Chang, 1983; Yeung and Ruzzo, 2001). An example with two clusters is shown in Fig. 3(a–c).

In this figure, a two-dimensional dataset contains two distinguishable clusters. However, since the within-cluster variation of this dataset is quite large, the cluster structure is totally lost in the first PC as shown in the histogram of the first PC in Fig. 3(b). If the clustering is only based on the first PC, no distinguishable clusters can be found, or equivalently, we would say that there is no structure in the dataset. Therefore, the traditional method of selecting PCs *only* based on the eigenvalue magnitude is not appropriate for our clustering purpose. We wish to develop a new method that can effectively select PCs containing useful information from those PCs with large variations for future clustering studies. The method and criteria for the selection of information containing PCs are developed in Section 3.

After selecting the information containing PCs, the model-based clustering method is used to cluster the

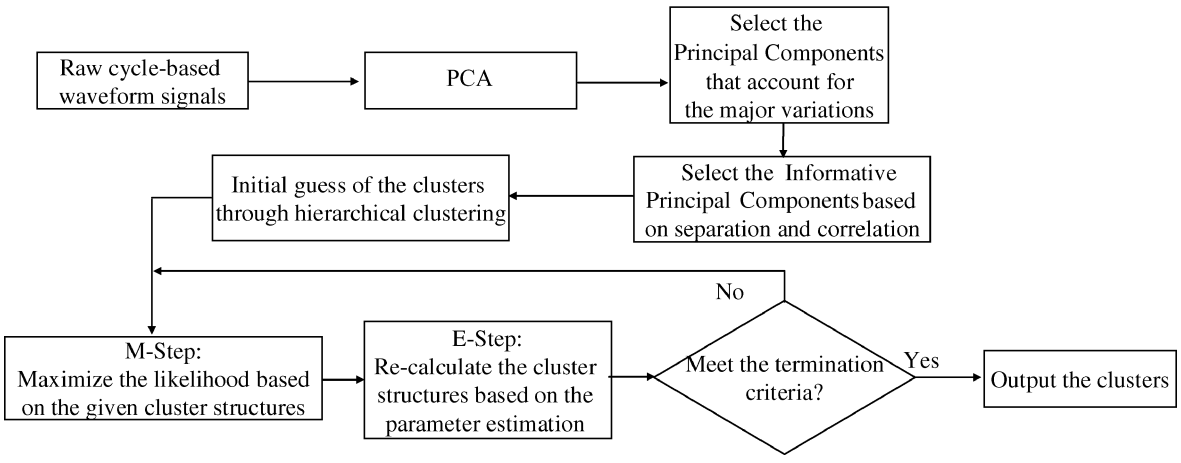

Fig. 2. Methodology overview.

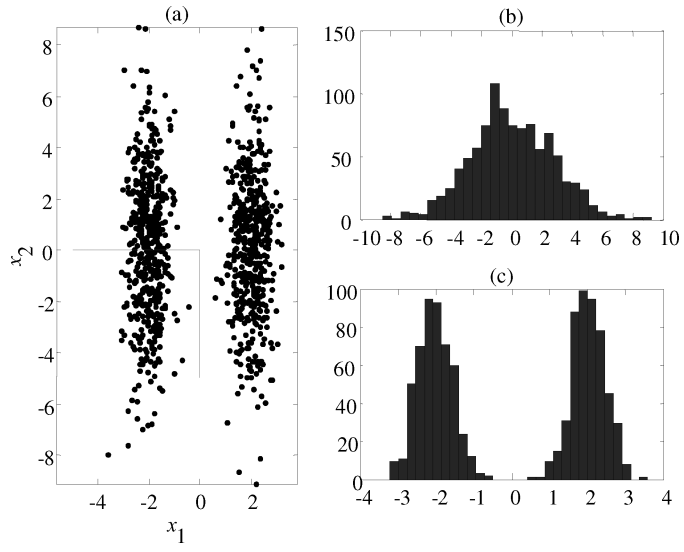

Fig. 3. (a) The original dataset; (b) a histogram of the first PC; and (c) a histogram of the second PC.

selected PCs. This can be done through the EM algorithm (Dempster *et al.*, 1977). The EM algorithm is a generic maximum-likelihood estimation technique for use on incomplete datasets. In the clustering of mixture models, the membership information of the data point is considered to be missing. In the maximization step, the model parameters are estimated using maximum likelihood estimation given the current guess of the incomplete data. In the expectation step, the guessed value of the incomplete data is updated based on the updated parameter estimates. In general, it is not an easy task to evaluate the clustering performance without knowing the true cluster structure in the dataset. However, the performance of model-based clustering can be evaluated through the Bayesian Information Criterion (BIC) (McLachlan and Peel, 2000) which is defined as:

$$\mathrm{BIC} \equiv 2l(\mathbf{x}, \hat{\boldsymbol{\theta}}) - m\log(n) \tag{4}$$

where $l(\mathbf{x}, \hat{\boldsymbol{\theta}})$ is the maximized log-likelihood of the mixture model, m is the number of independent parameters in the model, and n is the sample size. Based on the BIC, the number of clusters can be selected quantitatively.

It can be seen that the critical issue is to effectively select the information containing PCs for use in the proposed clustering method. The selection of the PCs should be based on the cluster structure rather than on the magnitude of their variations. A detailed discussion is presented in the following section.

3. Selection of information containing PCs

3.1. *Assessment of the contributions of PCs to clustering*

To select information containing PCs, a quantitative evaluation method for the contribution of the PCs to the clustering should be first developed. Since the EM algorithm is used for the clustering, information containing PCs should be selected in such a way that the parameters of the components in the mixture model can be estimated with a high degree of accuracy. Now, the problem of how to select information containing PCs is changed to how to select useful variables (PCs) for the estimation of a mixture model, i.e., which variables (PCs) should be selected/added in a multivariate mixture model of normal distributions so that the accuracy of the estimation results can be improved.

In this article, we handle this problem using mathematical formulations by considering a two-dimensional mixture model with two components. Although it is a simple 2D case, some quantitative insights can be obtained from this study. Assume n samples of two-dimensional process variables are obtained as:

$$\mathbf{X}^{n\times 2} = \begin{bmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix},$$

and denote

$$\mathbf{X}_1^{n\times 1} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} \text{ and } \mathbf{X}_2^{n\times 1} = \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}.$$

(In our case, the clustering is on PCs. Therefore, $\mathbf{X}^{n\times 2}$ contains two PCs.) The mixture models for $\mathbf{X}^{n\times 2}$, $\mathbf{X}_1^{n\times 1}$, and $\mathbf{X}_2^{n\times 1}$ are given as follows:

$$f_{\mathbf{X}}\left(\mathbf{X}^{n\times 2} \mid \tau_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right) = \prod_{i=1}^{n}[\tau_1\phi_1(x_{i1}, x_{i2} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\tau_1)\phi_2(x_{i1}, x_{i2}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)], \tag{5a}$$

$$f_{\mathbf{X}_1}\left(\mathbf{X}_1^{n\times 1} \mid \tau_1, \mu_{11}, \sigma_{11}, \mu_{21}, \sigma_{21}\right) = \prod_{i=1}^{n}[\tau_1\phi_1(x_{i1}|\mu_{11}, \sigma_{11}) + (1-\tau_1)\phi_2(x_{i1} \mid \mu_{21}, \sigma_{21})], \tag{5b}$$

$$f_{\mathbf{X}_2}\left(\mathbf{X}_2^{n\times 1} \mid \tau_1, \mu_{12}, \sigma_{12}, \mu_{22}, \sigma_{22}\right) = \prod_{i=1}^{n}[\tau_1\phi_1(x_{i2}|\mu_{12}, \sigma_{12}) + (1-\tau_1)\phi_2(x_{i2} \mid \mu_{22}, \sigma_{22})], \tag{5c}$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_{112}\sigma_{11}\sigma_{12} \\ \rho_{112}\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{bmatrix},$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \sigma_{21}^2 & \rho_{212}\sigma_{21}\sigma_{22} \\ \rho_{212}\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{bmatrix},$$

and μ_{ij} is the mean of the jth variable of the ith component, σ_{ij}^2 is the variance of the jth variable of the ith component, ρ_{112} and ρ_{212} are the correlation coefficients between these two variables within each of the two components, respectively. Three estimation results can be obtained by using the model in Equations (5a), (5b) and (5c), respectively.

The variable (PC) selection problem is to determine which estimation result is the most accurate. Particularly, we need to determine if the estimation result using Equation (5a) is significantly more accurate than the estimation result using Equation (5b) or Equation (5c). If yes, both variables (x_1 and x_2) should be considered simultaneously. Otherwise, one variable can be deleted to reduce the data dimension in further clustering studies. In this way, the information containing PCs that should be included in the clustering variables can be selected.

It is known that the asymptotical accuracy of maximum likelihood estimation can be evaluated using the Fisher information matrix (Day, 1969). Define $L(\boldsymbol{\theta} \mid \mathbf{x})$ as the log-likelihood function of parameter set $\boldsymbol{\theta}$ (e.g., for the mixture model in Equation (5a), $L(\tau_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2 \mid \mathbf{x}) = \ln[\tau_1, \phi_1, (\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\tau_1)\phi_2(\mathbf{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)]$ where $\mathbf{x} = [x_1 \ x_2]^T$). The ijth element of the Fisher information matrix is given by:

$$\mathbf{F}_{ij}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 L(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right),$$

where $\boldsymbol{\theta}_i$ is the ith element of the parameter set. The asymptotical covariance of the estimation results of the maximum likelihood estimation is the inverse of $\mathbf{F}$. A "larger" information matrix will result in a smaller variation in the estimation, which means that the estimation is more efficient. Unfortunately, an analysis using the analytical method for the information matrix for a general maximum likelihood estimation of a multi-dimensional and multi-component mixture model is very difficult, if not impossible. However, insights can be gained through the analysis of a two-dimensional mixture model with two normal components. Based on these insights, we can obtain some guidelines on how to select useful variables for the estimation of a mixture model. In this paper, we will focus on the Fisher information on the critical parameter τ_1 in the mixture model. Comparing with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, τ_1 is a relatively important parameter in the clustering, which decides the size of each cluster. The purpose of the following section is to provide insights on which variables are needed to obtain an accurate estimation of τ_1.

3.1.1. *Assessment of the Fisher information of single variable case*

Denote $I(\tau_1)$ as the Fisher information regarding τ_1, then for a mixture model with two components ϕ_1 and ϕ_2, we can obtain (Hill, 1963):

$$I(\tau_1) = \frac{1}{\tau_1 \tau_2}\left[1 - \int_{-\infty}^{\infty} \frac{\phi_1(\mathbf{x})\phi_2(\mathbf{x})}{\tau_1\phi_1(\mathbf{x}) + \tau_2\phi_2(\mathbf{x})}\mathrm{d}\mathbf{x}\right], \quad (6)$$

where $\tau_2 = 1 - \tau_1$. Denote

$$S = \int_{-\infty}^{\infty} \frac{\phi_1(\mathbf{x})\phi_2(\mathbf{x})}{\tau_1\phi_1(\mathbf{x}) + \tau_2\phi_2(\mathbf{x})}\mathrm{d}\mathbf{x},$$

then for a single variable mixture of two normal distributions

$$S_1 = \int_{-\infty}^{\infty} \left(\frac{\exp\left\{-(1/2)(x-\mu_{11})^2/\sigma_{11}^2\right\}}{\left|2\pi\sigma_{11}^2\right|^{1/2}} \times \frac{\exp\left\{-(1/2)(x-\mu_{21})^2/\sigma_{21}^2\right\}}{\left|2\pi\sigma_{21}^2\right|^{1/2}}\right) \Bigg/ \left(\tau_1 \frac{\exp\left\{-(1/2)(x-\mu_{11})^2/\sigma_{11}^2\right\}}{\left|2\pi\sigma_{11}^2\right|^{1/2}} + \tau_2 \frac{\exp\left\{-(1/2)(x-\mu_{21})^2/\sigma_{21}^2\right\}}{\left|2\pi\sigma_{21}^2\right|^{1/2}}\right)\mathrm{d}\mathbf{x}, \quad (7)$$

where the mixture distribution of $\mathbf{x}$ with two components are $f(\mathbf{x}_i) = \tau_1\phi_1(\mathbf{x}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\tau_1)\phi_2(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Using variable substitution $z = (x - \mu_{11})/\sigma_{11}$, and denoting $\lambda_1 = \sigma_{11}/\sigma_{21}$, $d_1 = (\mu_{11} - \mu_{21})/\sigma_{21}$, we have:

$$S_1 = \int_{-\infty}^{\infty} \frac{\exp\{-(1/2)(\lambda_1 z + d_1)^2\}\lambda_1/|2\pi|^{1/2}}{\tau_1 + \tau_2 \exp\{-(1/2)[(\lambda_1 z + d_1)^2 - z^2]\}\lambda_1}\mathrm{d}z. \quad (8)$$

Assuming that τ_1 and τ_2 are constant then S_1 is a function of d_1 and λ_1. Figure 4(a and b) illustrates the relationships between S_1 and d_1 and λ_1 respectively. As would be intuitively expected S_1 reaches a maximum (which means that the Fisher information reaches a minimum and the estimation accuracy has its lowest value) at $d_1 = 0$ when $\lambda_1 = 1$. Under this condition, these two components totally overlap one another.

3.1.2. *Assessment of the Fisher information of two variables case*

In the two-variable case, a similar derivation can lead to the Fisher information regarding τ_1. Assume a two-dimension mixture distribution with two components to be:

$$f_{\mathbf{X}}(\mathbf{X}|\tau_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \tau_1\phi_1(x_1, x_2|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\tau_1)\phi_2(x_1, x_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad (9)$$

where the expressions for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are given in Equation (5) and $\mathbf{X}$ is $[x_1 \ x_2]^T$. The expression for Equation (6) will become very complicated if Equation (9) is directly applied. In order to keep the expression as simple as possible so as to be able to gain insights, a linear transformation of the variable is needed.

One useful result is that for two nonsingular covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, there exists a nonsingular matrix $\mathbf{C}$, such that $\mathbf{C}^T\boldsymbol{\Sigma}_1\mathbf{C} = \mathbf{I}$ and $\mathbf{C}^T\boldsymbol{\Sigma}_2\mathbf{C} = \boldsymbol{\Lambda}$, where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix (Scott, 1997). $\mathbf{C}$ is in the form of $\boldsymbol{\Sigma}_1^{-1/2}\mathbf{P}$ and $\mathbf{P}$ is an appropriate orthogonal rotation matrix that consists of the eigenvectors of $\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}$. Therefore, a variable transformation can be defined as $\mathbf{y} = \mathbf{C}^T(\mathbf{x} - \boldsymbol{\mu}_1)$. The distribution of these two components of the new variables are $N(\mathbf{0}, \mathbf{I})$

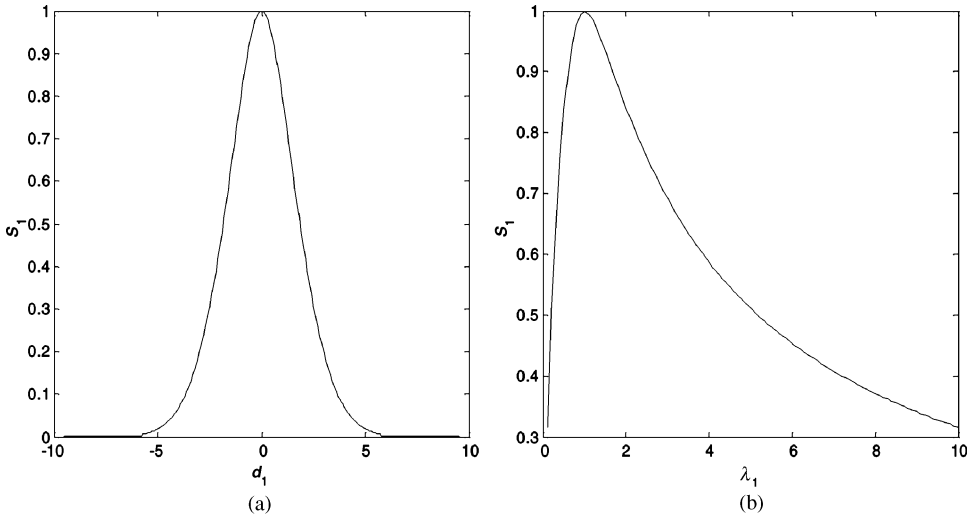

Fig. 4. (a) S_1 as a function of d_1 when $\lambda_1 = 1$; and (b) S_1 as a function of λ_1 when $d_1 = 0$.

and $N(-\mathbf{C}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \boldsymbol{\Lambda})$. The covariance structures of the transformed variables are much simpler than the original one. Moreover, it is known that MLE is invariant with respect to a linear transformation (Day, 1969). Therefore, the Fisher information can be studied based on **y** instead of **x**. Let $\mathbf{d}' = [d_1' \; d_2']^T = \mathbf{C}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \sigma_{21}'^2 & 0 \\ 0 & \sigma_{22}'^2 \end{bmatrix}, \tag{10}$$

and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \sigma_{21}^2/\sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2/\sigma_{12}^2 \end{bmatrix},$$

where $[D_1 \;\; D_2]^T$ is defined as the absolute value of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Clearly, these two variables contribute to the efficiency of the estimation individually. If the second variable is a normally distributed random variable, which means $\sigma_{12}^2 = \sigma_{22}^2$ and $D_2 = 0$, then the double integral is separated in Equation (11)

then by a straightforward derivation, we have

$$S_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\exp\left\{-(1/2)\left((y_1 + d_1')^2/\sigma_{21}'^2 + (y_2 + d_2')^2/\sigma_{22}'^2\right)\right\}/(2\pi)}{\tau_1\sigma_{21}'\sigma_{22}' + \tau_2 \exp\left\{-(1/2)\left((y_1 + d_1')^2/\sigma_{21}'^2 - y_1^2 + (y_2^2 + d_2')/\sigma_{22}'^2 - y_2^2\right)\right\}} \, \mathrm{d}y_1 \mathrm{d}y_2. \tag{11}$$

Clearly S_2 is a function of $\mathbf{d}'$, σ_{21}' and σ_{22}'. This expression can be further simplified under some special cases as discussed in the following section.

3.1.3. *Contributions of the second variable*

In order to understand when a variable is needed for the estimation of τ_1, its contribution to the estimation performance should be assessed. The following special cases will be discussed based on Equation (11).

Case 1: $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrices.
In this case:

$$\mathbf{C} = \boldsymbol{\Sigma}_1^{-\frac{1}{2}} = \begin{bmatrix} 1/\sigma_{11} & 0 \\ 0 & 1/\sigma_{12} \end{bmatrix}.$$

Hence, the absolute value of the components of $\mathbf{d}'$ is:

$$\begin{bmatrix} D_1/\sigma_{11} \\ D_2/\sigma_{12} \end{bmatrix},$$

and the integration with respect to y_2 can be integrated out. Therefore, y_2 will not contribute to the efficiency of the estimation. This result is not surprising because under this condition, x_1 is independent of x_2, and x_2 does not contain any structures. Therefore, x_2 will not contribute to the estimation.

Case 2: $\sigma_{12} = \sigma_{22}$, $\mu_{12} = \mu_{22}$, but $\rho_{112} \neq 0$ or $\rho_{212} \neq 0$.

In this case, there is no cluster structure in variable x_2, but there exists a correlation between variable 1 and variable 2. First, consider the relationship between $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Lambda}$. It is known that the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$. For a two-dimensional problem, we can obtain:

$$\sigma_{21}'^2 = \frac{(A - \sqrt{B})}{2(1 - \rho_{112}^2)\sigma_{11}^2} \quad \text{and} \quad \sigma_{22}'^2 = \frac{(A + \sqrt{B})}{2(1 - \rho_{112}^2)\sigma_{11}^2}, \tag{12}$$

where $A = \sigma_{11}^2 + \sigma_{21}^2 - 2\sigma_{11}\sigma_{21}\rho_{112}\rho_{212}$ and $B = \sigma_{21}^4 - 2\sigma_{21}^2\sigma_{11}^2 - 4\sigma_{21}^3\rho_{112}\rho_{212}\sigma_{11} + \sigma_{11}^4 - 4\sigma_{11}^3\rho_{112}\rho_{212}\sigma_{21} + 4\sigma_{11}^2\rho_{112}^2\sigma_{21}^2 +$

$4\sigma_{11}^2\rho_{212}^2\sigma_{21}^2$. Clearly, the variance of the second variable does not appear in the equation. Second, consider the distance $\mathbf{d}'$. To get the relationship between $\mathbf{d}'$ and $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, we need to know $\mathbf{C}^T$. However, even in this two-dimensional case, the general expression for $\mathbf{C}^T$ is too complicated to list. We can assume $\rho_{112} = 0$ to simplify the expression. Since ρ_{112} and ρ_{212} are symmetric in the problem setting, we can expect that the impacts of ρ_{112} and ρ_{212} on the Fisher information are the same. With $\rho_{112} = 0$, we can obtain:

$$\mathbf{C}^T \begin{bmatrix} D_1 \\ 0 \end{bmatrix} = \begin{bmatrix} d_1' \\ d_2' \end{bmatrix} = \begin{bmatrix} \dfrac{D_1 E}{\rho_{212}\sigma_{21}\sqrt{E^2/\left(\rho_{212}^2\sigma_{21}^2\sigma_{11}^2\right)+1}} \\ \dfrac{D_1 F}{\rho_{212}\sigma_{21}\sqrt{F^2/\left(\rho_{212}^2\sigma_{21}^2\sigma_{11}^2\right)+1}} \end{bmatrix}, \tag{13}$$

where

$$E = \frac{\sigma_{11}^2 + \sigma_{21}^2 + \sqrt{\sigma_{11}^4 - 2\sigma_{11}^2\sigma_{21}^2 + \sigma_{21}^4 + 4\sigma_{11}^2\sigma_{21}^2\rho_{212}^2}}{2\sigma_{11}^2} - 1,$$

$$F = \frac{\sigma_{11}^2 + \sigma_{21}^2 - \sqrt{\sigma_{11}^4 - 2\sigma_{11}^2\sigma_{21}^2 + \sigma_{21}^4 + 4\sigma_{11}^2\sigma_{21}^2\rho_{212}^2}}{2\sigma_{11}^2} - 1.$$

Clearly, the variance of the second variable also does not appear in the expression of $[d_1' \;\; d_2']^T$. Combining with Equation (12), it can be concluded that the variance of the second variable does not affect the asymptotical efficiency of the maximum likelihood estimation. Substituting Equations (12) and (13) into Equation (11), the effect of the correlation between x_1 and x_2 can be studied. Figure 5 illustrates the relationship between S_2 and ρ_{212}. Clearly, S_2 is a monotonically decreasing function with respect to ρ_{212}, which means that with a larger ρ_{212}, the Fisher information will become larger, and hence the variance of the estimation result will become smaller. Therefore, if the sample size is large, the second variable will still contribute to the estimation through the correlation among variables even though the second variable does not contain any cluster structures.

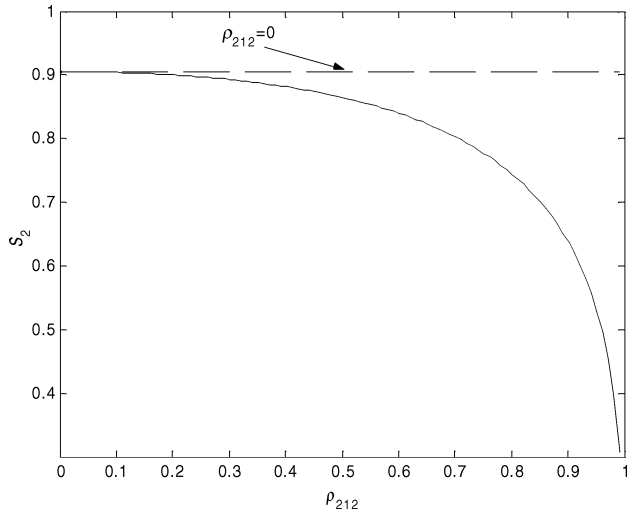

Fig. 5. S_2 as a function of ρ_{212} when $\sigma_{11} = 2$ and $\sigma_{21} = 3$.

A very interesting special case of case 2 is $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, i.e., $\sigma_{21} = \sigma_{11}$ and $\rho_{112} = \rho_{212}$. In this case, $\sigma_{21}' = 1$ and $\sigma_{22}' = 1$ from Equation (12). S_2 is determined only by $\mathbf{d}'$. Defining Δ as the Euclidean length of $\mathbf{d}'$ and using the variable transformation:

$$\mathbf{z} = \frac{1}{\Delta}\begin{bmatrix} d_1' & -d_2' \\ d_2' & d_1' \end{bmatrix}\mathbf{y},$$

we can transform Equation (11) into:

$$S_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\exp\{-(1/2)(\Delta^2 + z_1^2 + z_2^2 + 2\Delta z_1)\}/(2\pi)}{\tau_1 + \tau_2\exp\{-(1/2)(\Delta^2 + 2\Delta z_1)\}}\mathrm{d}z_1\mathrm{d}z_2. \tag{14}$$

Clearly, the integration of z_1 and z_2 are separated. Since:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}z_2^2\right)\mathrm{d}z_2 = 1,$$

therefore:

$$S_2 = \int_{-\infty}^{\infty} \frac{\exp\{-(1/2)(\Delta^2 + z_1^2 + 2\Delta z_1)\}/(\sqrt{2\pi})}{\tau_1 + \tau_2\exp\{-(1/2)(\Delta^2 + 2\Delta z_1)\}}\mathrm{d}z_1. \tag{15}$$

Comparing Equation (15) and Equation (8) with $\lambda_1 = 1$, they are the same except that Δ replaces d_1. Since Δ is the length of $\mathbf{d}'$, which is $\mathbf{C}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\mathbf{C} = \boldsymbol{\Sigma}_1^{-\frac{1}{2}}\mathbf{P}$, where $\mathbf{P}$ is an orthogonal rotation matrix, $\mathbf{P}$ will not change the length of a vector. Therefore Δ^2 is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. In a two-dimensional case, we can further obtain (Chang, 1976):

$$\Delta^2 = \left(D_1^2 + D_2^2 - 2\rho D_1 D_2\right)/(1 - \rho^2), \tag{16}$$

where $[D_1 \; D_2]^T$ is defined as the absolute value of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and ρ is ρ_{112} or ρ_{212}. It is straightforward to prove that Δ^2 is monotonically increasing with respect to ρ. However, since S_1 is monotonically decreasing with Δ^2 based on Fig. 4(a), even if D_2 is zero, the second variable will still contribute to the efficiency of the estimation through the correlation between these two variables.

Based on the above studies, it can be seen that an assessment of the contribution of the PCs to data clustering should consider two factors: (i) whether the PCs contain a multiple cluster structure; and (ii) whether the PCs have within-cluster with other PCs that contain a cluster correlations structure. A summary remark is given as follows:

- A variable will contribute to the model-based clustering if it contains a cluster structure (i.e., the variable itself is a mixture of different distributions) or it has within-cluster correlations with other variables that contain a cluster structure. It is worth mentioning that this point

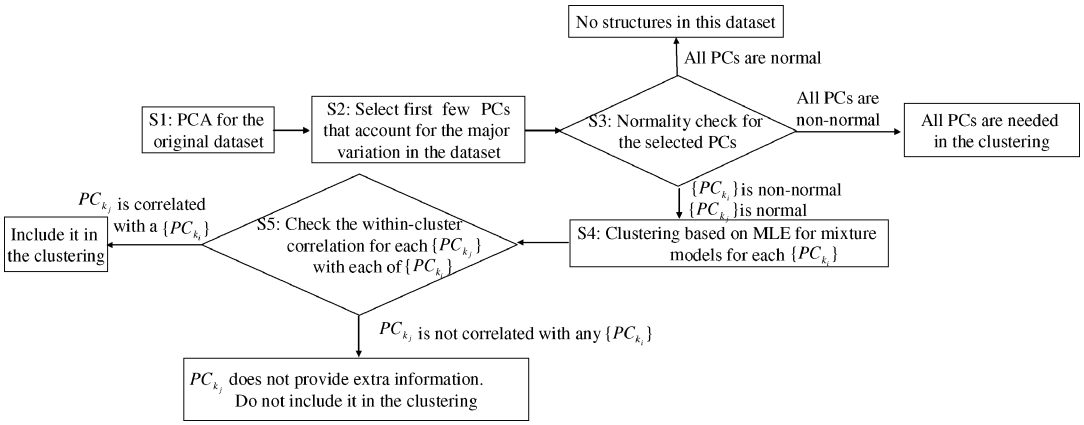

Fig. 6. The decision making process procedure for variable selection for subsequent clustering.

is intuitive and can be qualitatively illustrated by a simple example as follows. Consider a 2-D discrimination problem and assume that there are two clusters that only differ from one another in the mean of the first variable. These two variables are correlated with one another and hence the distributions of the clusters are tilted ellipses. Clearly in this case, the best linear discriminator will not be parallel to either axis of the coordinate system. The second variable plays a role in the discrimination. Furthermore, if the correlation between the two variables gets larger, the ellipse gets narrower. Hence, the clustering result will get better. Another point that needs to be emphasized is that the selected PCs are actually uncorrelated with one another overall although they could be highly within-cluster correlated. Therefore, there is no colinearity issue of the selected PCs. The above mathematical analysis provides us with some quantitative insights for checking the significance of variable correlations on the estimation accuracy.

- It should be pointed out that the property derived from the Fisher information is an asymptotic property, which means that it is true only for large sample sizes. If the sample size is small then sample uncertainty will play an important role. Unfortunately, it is very difficult to quantify the impact of the sample size. Numerical study is often needed to obtain the impacts for particular cases. Nevertheless, because of the automated sensing technology in modern manufacturing processes, a large sample set is assumed in the paper.

3.2. *Procedures to select information containing PCs for clustering*

3.2.1. *Overview of the selection procedures*

In the above analytical study, we identified two contributing factors (the cluster structure and the within-cluster correlation with other variables) of a variable to the estimation efficiency of a mixture model. An automatic analysis procedure for selecting information containing PCs is developed as shown in Fig. 6.

S1. PCA of the original dataset.

S2. Since the PCs that correspond to small variations can be considered as noise, only those PCs that correspond to large variations are kept. Accumulation of 90%–95% of the total variation is the commonly used threshold in PC selection in practice.

S3. To identify if the PCs contain cluster structures, we use a normality test. If a PC passes the normality test (i.e., the PC follows a normal distribution), it does not contain any cluster information. The detailed discussion on how to check the normality will be given in Section 3.2.2. Based on the normality test, if all the PCs are normal, then no structures can be found in the dataset. On the other hand, if all PCs are non-normal, then all the PCs will contribute to the clustering at least through the cluster structure within themselves. We need to keep all of them in the clustering. If some of the PCs are non-normal and some are normal, we will keep the non-normal distributed PCs for clustering, and further judge whether the normally distributed PCs should be included in the clustering in the following step.

S4. A normally distributed PC could still contribute to the model-based clustering through the within-cluster correlation. To check the within-cluster correlation, cluster information on the dataset is needed. A practical way to conduct the test is as follows. Assume PC_{k_j} is a PC that is normally distributed and $\{PC_{k_i}, i = 1 \ldots n_1\}$ is the set of PCs that are non-normally distributed, where n_1 is the total number of non-normally distributed PCs. Then for each i, $i = 1 \ldots n_1$, maximum likelihood estimation can be used to find the clusters in PC_{k_i}. Based on these cluster structures, we can check if a within-cluster correlation exists between PC_{k_i} and PC_{k_j}.

S5. Assume in step 4 that PC_{k_i} can be clustered into two clusters. Denote $PC^l_{k_i}$ as the lth sample of PC_{k_i} and $\{PC^l_{k_i}, l \in C_1\}$ and $\{PC^l_{k_i}, l \in C_2\}$ are the two

clusters. Then we can check the correlation between $\{PC^l_{k_i}, l \in C_1\}$ and $\{PC^l_{k_j}, \in C_1\}$ and the correlation between $\{PC^l_{k_i}, l \in C_2\}$ and $\{PC^l_{k_j}, l \in C_2\}$. If either cluster shows correlation behavior then PC_{k_j} will contribute to the clustering of PC_{k_i}. Hence, we shall include it in the final clustering. If PC_{k_j} is not within-cluster correlated with any PC_{k_i}, then it only contains non-structured normally distributed noise. It should be excluded in the final clustering. The existence of the correlation can be tested using a t-distribution (Montgomery and Runger, 1994). Similar to the normality test, a critical value needs to be selected for the correlation test. Since correlation only contributes to the clustering through another variable that does have a cluster structure, only a strong correlation will result in a significant impact on the clustering as shown in Fig. 5. Hence, the critical value often needs to be selected to be large. The critical value can be selected quantitatively following the procedure in Section 3.1. Our experience also shows that a critical value of 0.2–0.5 is a good choice in most cases.

After these five stages, the information containing PCs will be picked out from all the PCs. One point to be highlighted is that we have selected the variable based solely on its contribution to a two-variable clustering case. This means that the interaction of three or more variables within the cluster is ignored at the PC selection step, however, any higher-order correlation among more than two variables is considered in later stages through the model-based clustering of all selected PCs.

3.2.2. *Normality check to detect cluster structures*

There are many normality checking routines, such as the Kolmogorov-Smirnov test, Lilliefors test, the Jarque-Bera test, etc. (Mardia, 1980). In this study, we use the Jarque-Bera test. Other procedures are also used and it is found that the performance of other test procedures is similar. The statistics used in the Jarque-Bera test are:

$$Q = \frac{n}{6}\left[J^2 + \frac{(B-3)^2}{4}\right], \tag{17}$$

where J and B are the skewness and kurtosis of the random variable, respectively, and n is the sample size. It is known that Q asymptotically follows a chi-square distribution with two degrees of freedom. For a specific α-value and under large sample condition, if $Q < \chi^2_{\alpha,2}$, then the null hypothesis (the sample follows a normal distribution) cannot be rejected. Otherwise, it concludes that the sample does not follow a normal distribution, and hence contains a multiple cluster structure. To provide guidelines on the application of the Jarque-Bera test, Operation Characteristic (OC) curves are generated as shown in Fig. 7(a–d) through Monte Carlo simulation.

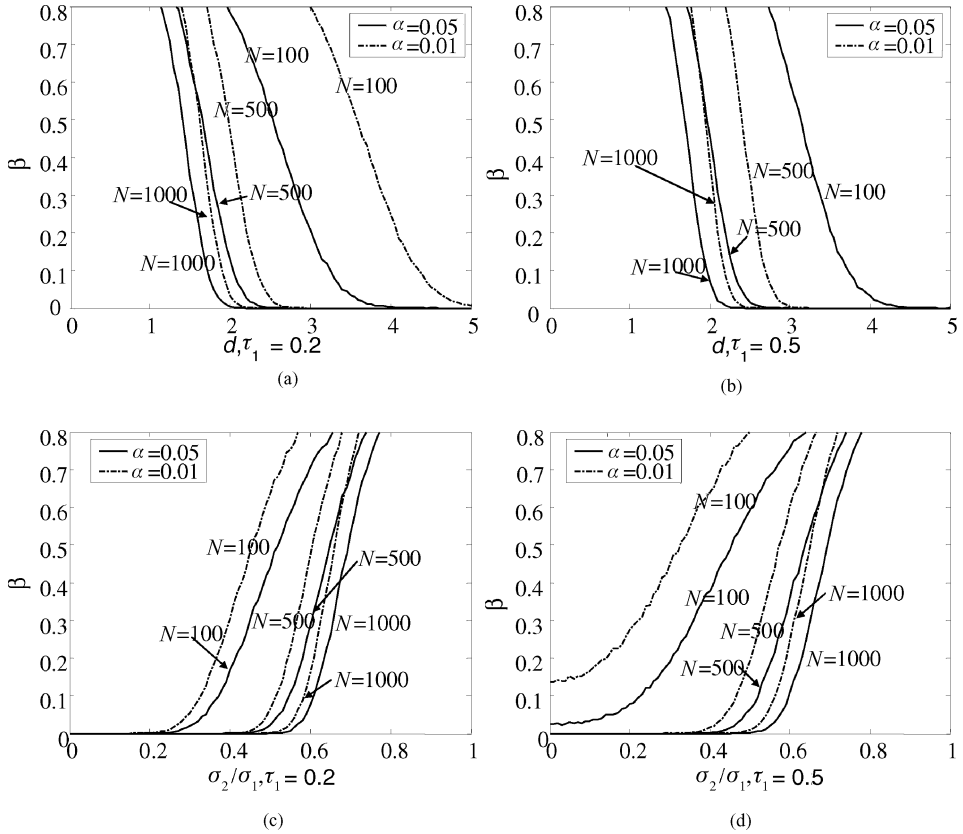

Fig. 7. OC curves for the Jarque-Bera test for typical mixture distributions; (a) $\sigma_1 = \sigma_2 = 1$, $\tau_1 = 0.2$, and $d = \mu_2 - \mu_1$; (b) $\sigma_1 = \sigma_2 = 1$, $\tau_1 = 0.5$, and $d = \mu_2 - \mu_1$; (c) $\mu_1 = \mu_2 = 0$, $\tau_1 = 0.2$, and $\sigma_1 = 1$; (d) $\mu_1 = \mu_2 = 0$, $\tau_1 = 0.5$, and $\sigma_1 = 1$.

The OC curves illustrate the miss detection probability (also called the β error) for a specific mixture distribution under a given specific false-alarm probability (also called the α error). For example, the solid curves in Fig. 7(a) are the β errors when the α error equals 0.05 and the mean difference of the two components in the mixture distribution varies from zero to five, and the proportion of one component is 0.2.

Monte Carlo simulation is used to obtain these OC curves. Since the Jarque-Bera test is an asymptotical test, we cannot use $\chi^2_{\alpha,2}$ as the critical value when the sample size is small. Therefore, the first step generating the OC curve is to obtain the critical values through simulation. First, 10,000 replicated samples with sample sizes of 100, 500 and 1000 of a univariate standard normal distribution $N(0,1)$ are generated. Then for each sample, the Q statistic is calculated. Finally, the 95% and 99% points of the Q statistic (please note that we have 10,000 Q statistics for each sample size), which are taken as the critical values, are obtained for each sample size. The 95% points for sample sizes of 100, 500 and 1000 are 5.08, 5.94 and 6.04, respectively. Similarly, the 99% points for sample sizes of 100, 500 and 1000 are 11.21, 11.13, and 10.24, respectively. It is interesting to note that $\chi^2_{0.05,2}$ and $\chi^2_{0.01,2}$ are 5.99 and 9.21, respectively. It is clear that when the sample size increases, the percentile values get closer to the asymptotical values.

After obtaining the critical values for a specific α value and sample size, we can further obtain the β error for a specific non-normal distribution. Assume a univariate variable x is a mixture of two normal distribution $N(0, 1)$ and $N(d, \sigma_2/\sigma_1)$ and the proportion of the first component is τ_1. Given these parameters, samples of this mixture distribution can be generated and the corresponding Jarque-Bera test statistics can be obtained. In the Monte Carlo simulation, 10,000 replicates are generated under specific conditions (sample size n, d and σ_2/σ_1). From these replicates, the miss-detection probability can be estimated as the ratio between the number of Q statistics less than the corresponding critical value and total number of replicates (10,000 in our case).

With this OC curve, we can select the proper α-value based on the desired separation (this is often obtained based on engineering judgment). Clearly when the α error gets smaller, the β error gets larger. When the separation (difference in means and/or variances of the components of the mixture) gets larger, the β error gets smaller for a given α error. Figure 7(a–d) clearly illustrates these properties. In practice, we can often determine the level in the separation of the components based on engineering judgment. For example, it is known that two normal distributions with a mean separation of two standard deviations cannot be separated with a small miss-detection rate and false-alarm rate. Therefore, the critical values in the normality test can be set in such a way that it will only detect the mixture that contains components separated by more than two standard deviations.

4. Case studies

4.1. *A numerical example*

To illustrate the effectiveness of the proposed method, a numerical study is conducted. A dataset that consists of 1000 samples of 20 variables is generated. Among these 20 variables, 15 variables ($x_6 - x_{20}$) only contain normal noise with a zero mean and a variance between $0 \sim 2$. $x_1 \sim x_3$ follow a three dimensional mixture of normal distributions with $\boldsymbol{\mu}_1^{1-3} = [6\ 2\ 0]^T$,

$$\boldsymbol{\Sigma}_1^{1\sim3} = \begin{bmatrix} 12 & 1.5 & 1.3 \\ 1.5 & 2 & 1.2 \\ 1.3 & 1.2 & 2 \end{bmatrix},$$

$$\boldsymbol{\mu}_2^{1\sim3} = [0\ \ 00]^T,\ \boldsymbol{\Sigma}_1^{1\sim3} = \mathbf{I},\ \text{ and }\ \tau_1 \sim 0.4.$$

Also x_4–x_5 follows a two-dimensional mixture of normal distribution with $\mu_1^{4\sim5} = [-3\ 0]$,

$$\Sigma_1^{4\sim5} = \begin{bmatrix} 9 & 1.9 \\ 1.9 & 2 \end{bmatrix},$$

$\boldsymbol{\mu}_2^{4\sim5} = [0 \quad 0]^T$, $\Sigma_2^{4\sim5} = \mathbf{I}$, and $\tau_1 = 0.35$. The first component of $x_1 \sim x_3$ happens during the sample number 101$\sim$500. The first component of $x_4 \sim x_5$ happens during the sample number 651$\sim$1000. Since they are not overlapping one another, there are three clusters in the generated data representing three working conditions: the first faulty condition (101$\sim$500) that is represented by the first component of $x_1 \sim x_3$, the second faulty condition (651$\sim$1000), and the normal working condition that is represented by the rest. Figure 8 shows the samples of $x_1 \sim x_6$. From the plot, it can be seen that faulty condition 1 is separated from the other working conditions. However, faulty condition 2 is not easily distinguished from the normal working condition.

PCA was performed on this dataset and Fig. 9(a and b) shows the results. Figure 9(a) is a Pareto plot of the

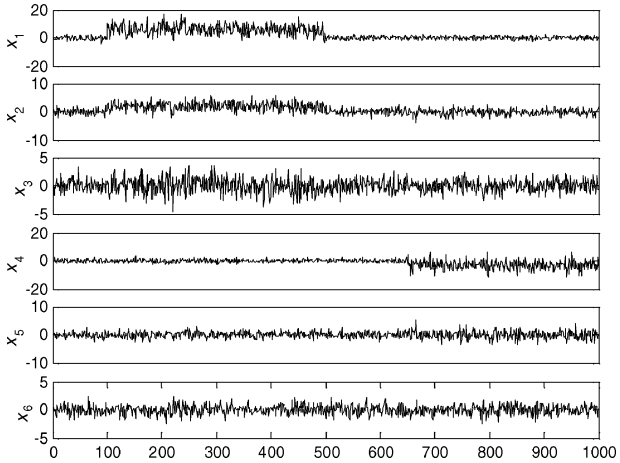

Fig. 8. The first six dimensions of the generated dataset.

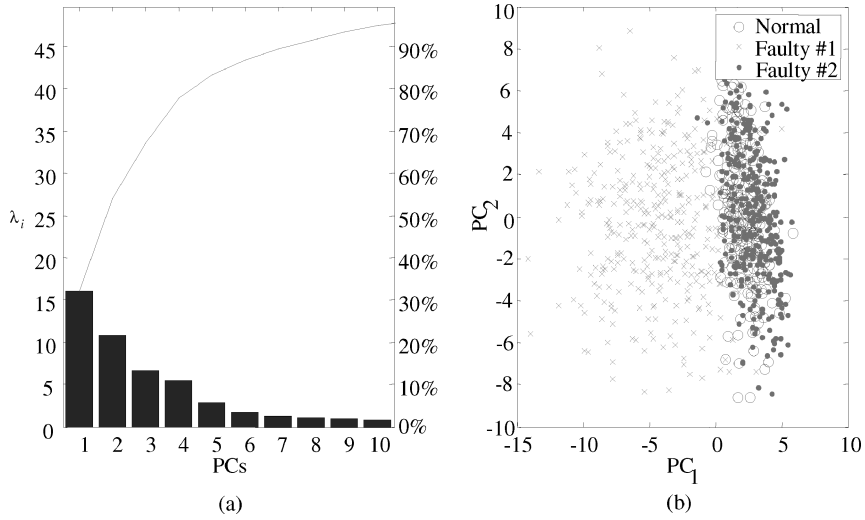

Fig. 9. The PCA results for the dataset: (a) a Pareto plot of the eigenvalues of the PCs: and (b) the first two PCs.

eigenvalues of the PCs. It is clear that the first 10 PCs account for more than 95% of the total variation. It is therefore safe to only consider the first 10 PCs in the clustering analysis. However, not all PCs contain information. Figure 9(b) shows the first two PCs which account for about 50% of the total variation. However, in these two PCs, the normal condition and faulty condition 2 completely overlap one another. Thus, we will not get a good result if only the first two PCs are used.

Following the scheme proposed in this paper, information containing PCs can be selected. In this case, PC_1, PC_4, PC_7 and PC_9 fail the normality test at a significance level at 0.1. Then EM clustering is conducted on each of these PCs. Based on the clustering results, within-cluster correlation is checked for other PCs with a critical value of 0.2. It turns out that PC_2 and PC_3 should also be included in the analysis. Figure 10 shows the first two selected PCs, the first PC and the fourth PC. It is clear that these PCs contain significant amounts of cluster information.

The EM clustering algorithm is applied to the original dataset, the first 10 PCs, and the six selected PCs, respectively. Since the true cluster structure is known in this case, we can use an adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985) to evaluate the performance of the clustering. An adjusted Rand index is a quantitative index that measures the agreements between two partitions of a set. The maximum value is one, which means that these two partitions are identical to one another. The expectation of the adjusted Rand index for two random partitions is zero. The adjusted Rand index of the clustering results based on the original dataset, the first 10 PCs, and the selected PCs are denoted as $Rand_{\mathrm{X}}$, $Rand_{10\mathrm{PCs}}$, and $Rand_{\mathrm{Selected\ PCs}}$, respectively. To compare the clustering performance, 50 numerical cases are conducted and $Rand_{\mathrm{X}}$, $Rand_{10\mathrm{PCs}}$, and $Rand_{\mathrm{Selected\ PCs}}$ are shown in Fig. 11.

Clearly, in most cases, the clustering results based on selected PCs have a higher adjusted Rand index than the results based on the original dataset and those based on

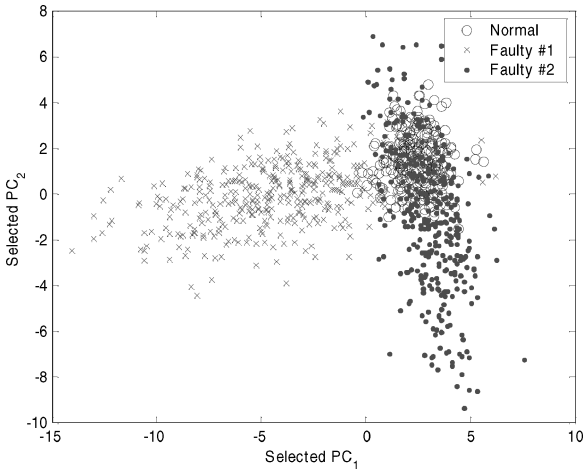

Fig. 10. The first two selected PCs.

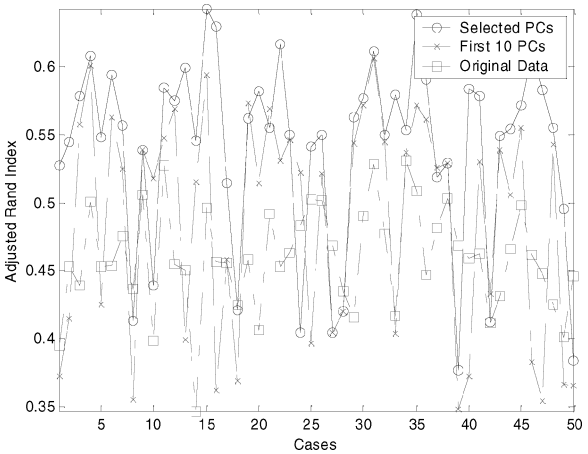

Fig. 11. The adjusted Rand indices for 50 cases.

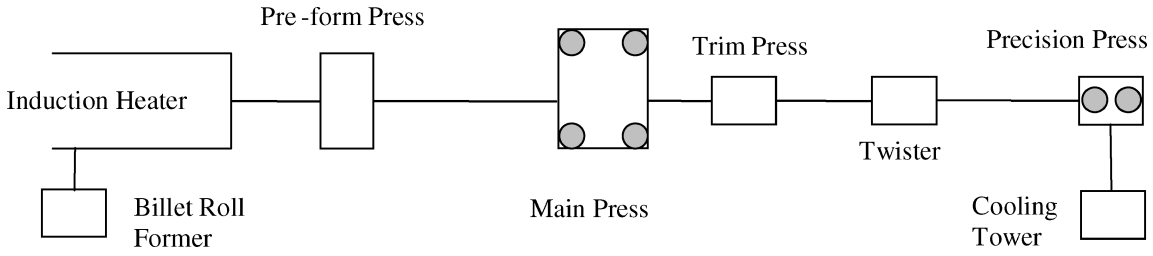

Fig. 12. A typical crankshaft forging process.

the first 10 PCs. Quantitatively, in about 90% of cases, $Rand_{\text{selectedPCs}}$ is higher than $Rand_{\text{X}}$; and in about 80% of cases, $Rand_{\text{selected PCs}}$ is higher than $Rand_{\text{10 PCs}}$. The evidence is clear that the proposed scheme is quite helpful to the clustering in this case study. In the next section, a real-world example is presented to show the effectiveness of the proposed technique.

4.2. *Case study for unsupervised clustering of forging tonnages*

A historic dataset of forging tonnage signals is used to evaluate the efficiency of the proposed method. A diagram of a typical crankshaft forging process is shown in Fig. 12.

In this process, the workpiece passes a heater, a pre-form press, the main press, the trim press, a twister, precision finishing press, and finally the cooling tower. The major deformation of the workpiece happens at the main press. Strain sensors are mounted on the columns of the main press to measure the tonnage signals. The dataset contains 906 tonnage signals from the strain sensor mounted on one of the columns of the main press. The sensor readings have been calibrated according to the force. Therefore, the output of the sensor is tonnage force, instead of strain.

These 906 tonnage signals are shown in Fig. 13(a). It can be seen that the signals are very similar to one another. It is necessary to first cluster the signals into different groups for further analysis. However, it is very difficult to cluster the signals into different groups based on the original time-domain signals as shown in Fig. 13(a) because of their high-dimensionality. Following the proposed scheme as shown in Fig. 2 in this article, PCA is performed. The Pareto plot of the variances of the PCs is shown in Fig. 13(b). The first 10 PCs account for more than 95% of the total variation. Therefore, only these PCs are considered in the clustering. For each PC, there are 906 samples. The shape of the eigenvector corresponding to the largest eigenvalue is shown in Fig. 13(c). From this shape, it can be seen that the variation pattern of the cycle-based signals is related to their profile mean value. The histograms of the first four PCs are shown in Fig. 14(a–d).

A normality test with a significance level of 1% is used to check the normality of the first 10 PCs. It is found that PC_1–PC_4, PC_6, PC_8 and PC_9 fail the test, which means that they contain certain separation information. Based on these PCs, a correlation test with a critical value of 0.3 is conducted and the result shows that no other PCs need to be included in the analysis.

The model-based clustering is conducted based on the selected PCs and the first 10 PCs, respectively. The results are shown in Fig. 15(a and b).

The 906 tonnage signals are all clustered into three clusters in these two cases. However, clusters 2 and 3 from the

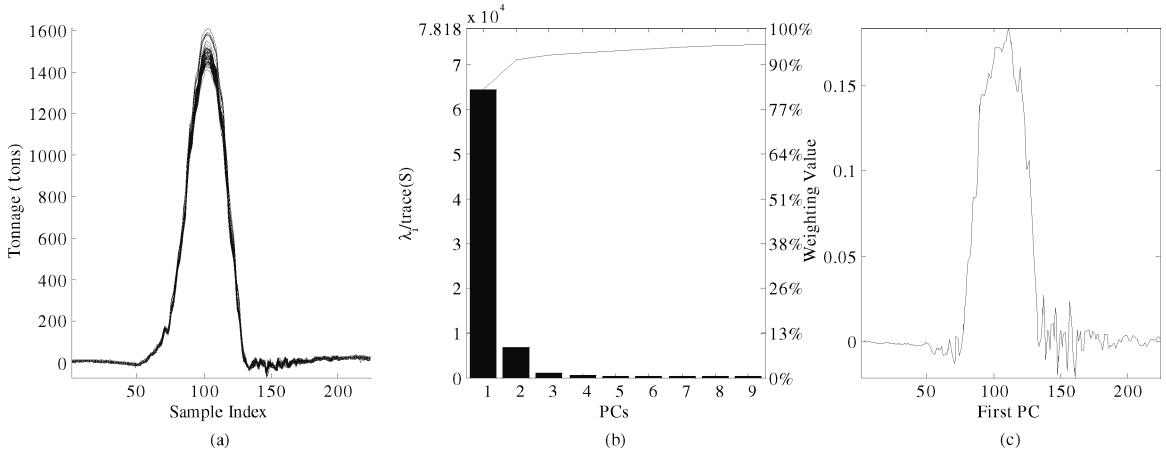

Fig. 13. (a) The 906 tonnage signals; (b) Pareto plot of the variances of the PCs; and (c) the shape of the eigenvector corresponding to the largest eigenvalue.

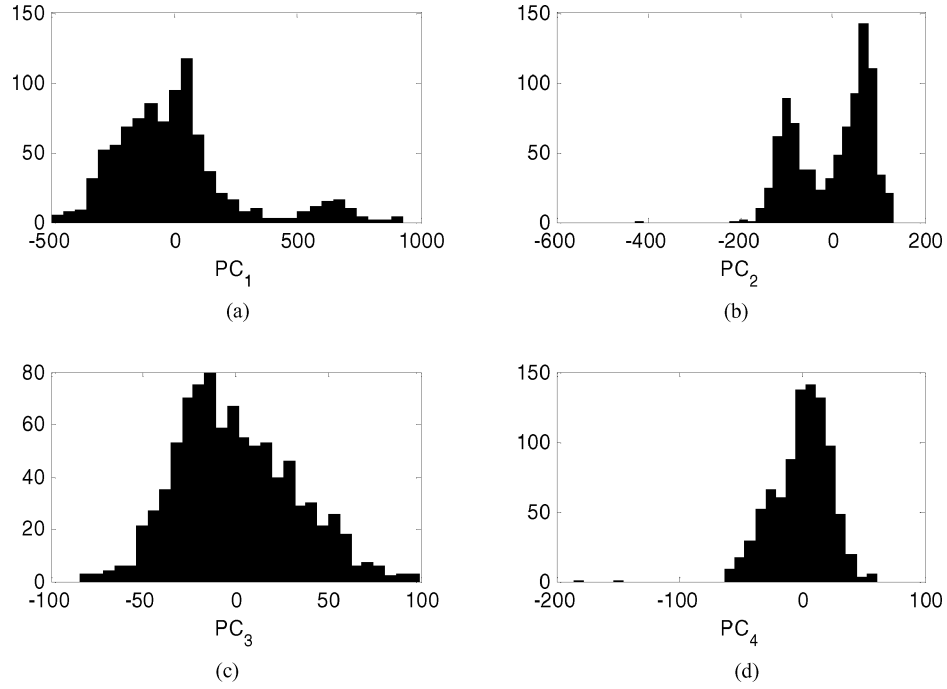

Fig. 14. Histograms of the first four PCs: (a) PC_1; (b) PC_2; (c) PC_3, and (d) PC_4.

first 10 PCs are mixed together in the first two PCs. They are more separated in the clustering results based on the selected PCs. The BIC values of these two clustering cases are also calculated. We obtain that $BIC_{\text{first 10 PCs}} = -83\,970$ and $BIC_{\text{selected PCs}} = -61\,010$, which means that the estimation based on the selected PCs is a better model. The process is further investigated based on the clustering results. It is shown that the data points of clusters 1 and 2 possibly come from two different shut height setups. The third cluster that has a larger variation may be a result of a variation in the lubrication. With this information at hand, further discrimination and online monitoring can be pursued.

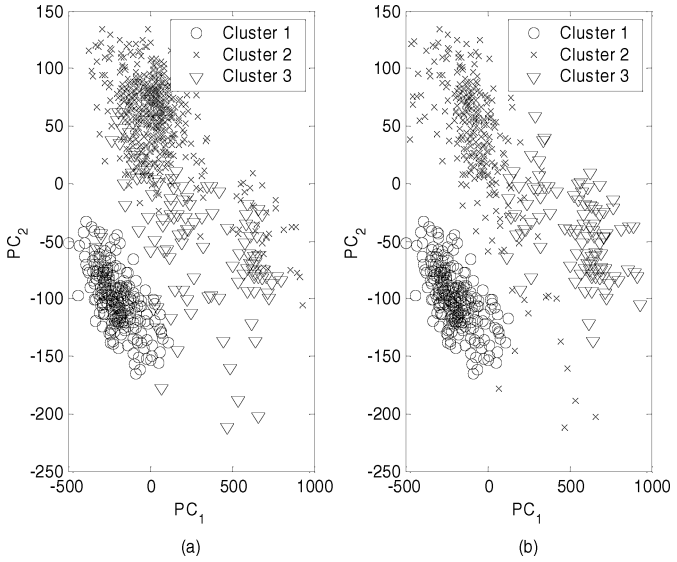

Fig. 15. The clusters obtained using: (a) the first 10 PCs; and (b) the selected PCs.

5. Conclusions

An automatic feature extraction method for dimension reduction and an analysis procedure for unsupervised clustering of cycle-based signals are proposed in this paper. First, the principal component analysis is used to linearly transform the signals into PCs. Then, the PC accounting for large variations and "information containing" PCs are selected among all the PCs. It is found that a cluster structure in the PCs and the within-cluster correlation between PCs will contribute to the model-based clustering. A simulation study and a real-world example of forging tonnage signal clustering illustrated the effectiveness of this method. Based on the initial clustering results for the forging process, physical insights into the process can be obtained.

The proposed technique has strong engineering relevance. Due to rapid developments in sensing technology, a huge amount of historical measurement data are now usually available. The proposed unsupervised technique can cluster these historical data into different preliminary groups to provide more information or a starting point for monitoring and diagnostic system design.

One point that needs to be mentioned is that PCA is applied to the raw signals in this paper. However, this method can also be applied to other transformations of the raw

signals without modification. For example, wavelet transformation can be used to transform the raw signal into the time-frequency domain. Then the PCA and variable selection can be applied to the wavelet coefficients. In this way, the method proposed in this paper can be integrated with engineering-knowledge based variable selection methods. Furthermore, we assumed that the cycle-based signal follows a normal distribution. This assumption will be more accurate for the wavelet coefficients since they can be viewed as a summation of the cycle-based signal at a certain interval. These problems are currently under investigation.

Acknowledgements

The authors would like to thank the Editors and reviewers for their insightful comments and suggestions, which have significantly improved the paper quality and readability. The authors also gratefully acknowledge the financial support of NSF grant 0330356 and a National Science Foundation (NSF) CAREER award DMI-0133942.

References

Anon (1989) Discriminant analysis and clustering. *Statistical Science*, **4**(1), 34–69.

Arabie, P., Hubert, L.J. and Soete, G.D. (eds). (1998) *Clustering and Classification*, World Scientific, River Edge, NJ.

Carreira-Perpinan, M. (1997) A review of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield, Sheffield, UK.

Chang, W.C. (1976) The effects of adding a variable in dissecting a mixture of two normal populations with a common covariance matrix. *Biometrika*, **63**(3), 676–678.

Chang, W.C. (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, **32**, 267–275.

Day, N.E. (1969) Estimating the components of a mixture of normal distribution. *Biometrika*, **56**(3), 463–474.

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

Draper, N.R. and Smith, H. (1980) *Applied Regression Analysis*, Wiley, New York, NY.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*, 2nd edn., Wiley, New York, NY.

Fowlkes, E.B., Gnanadesikan, G. and Kettenring, J.R. (1987) Variable selection in clustering and other contexts, in *Design, Data, and Analysis: By Some Friends of Cuthbert Daniel*. Wiley, New York, NY.

Grogan, R. (2002) High speed stamping process improvement thru force and displacement monitoring. Technical report, Helm Instrument Company, Maumee, OH.

Guyon I. and Elissee, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Hill, B.M. (1963) Information for estimation of the proportions in mixtures of exponential and normal distributions. *Journal of the American Statistical Association*, **58**, 918–932.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, **2**, 193–198.

Jackson, J.E. (1991) *A User's Guide to Principal Components*, Wiley, New York, NY.

Jimenez, L. and Landgrebe, D.A. (1998) Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, **28**(1), 39–54.

Jin, J. and Shi, J. (1999) Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics*, **41**(4), 327–339.

Jin, J. and Shi, J. (2000) Diagnostic feature extraction from stamping tonnage signals based on design of experiments. *ASME Transactions, Journal of Manufacturing Science and Engineering*, **122**(2), 360–369.

Jin, J. and Shi, J. (2001) Automatic feature extraction of signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing*, **12**, 257–268.

Knussmann, K.D. and Rose, C. (1993) Signature-based process control (SbPC^{TM}). Technical report, Signature Technologies, pp. 311–325.

Koh, C.K.H., Shi, J., Williams, W. and Ni, J. (1999a) Multiple fault detection and isolation using the Haar transform—part 1: theory. *ASME Transactions, Journal of Manufacturing Science and Engineering*, **121**(2) 290–294.

Koh, C.K.H., Shi, J., Williams, W. and Ni, J. (1999b) Multiple fault detection and isolation using the Haar transform—part 2: application to the stamping process. *ASME Transactions, Journal of Manufacturing Science and Engineering*. **121**(2), 295–299.

Lada, E.K., Lu, J.-C., and Wilson, J.R. (2002) A wavelet-based procedure for process fault detection. *IEEE Transactions on Semiconductor Manufacturing*, **15**(1), 79–90.

Liu, J.S., Zhang, J.L., Palumbo, M.J. and Lawrence, C.E. (2003). Bayesian clustering with variable and transformation selections. *Bayesian Statistics* **7**, 249–275.

Mardia, K.V. (1980) Tests of univariate and multivariate normality, in *Handbook of Statistics: Vol 1*, Krishnaiah, P. R. (eds.), North-Holland, Amsterdam, pp. 274–320.

McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models*, Dekker, New York, NY.

McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*, Wiley, New York, NY.

Montgomery, D. and Runger, G.C. (1994) *Applied Statistics and Probability for Engineers*, Wiley, New York, NY. pp. 519–520.

Pittner, S. and Kamarthi, S.V. (1999) Feature extraction from wavelet coefficients for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(1), 83–88.

Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis*, Springer-Verlag, New York, NY.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66** 441–458.

Scott, J.R. (1997) *Matrix Analysis for Statistics*, Wiley, New York, NY, pp. 154–157.

Tanaka, Y. and Mori, Y. (1997) Principal component analysis based on a subset of variables: variable selection and sensitivity analysis. *American Journal of Mathematical and Management*, **17**(1), 61–89.

Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**(9), 763–774.

Biographies

Shiyu Zhou is an Assistant Professor in the Department of Industrial and Systems Engineering at the University of Wisconsin-Madison. He received his B.S. and M.S. degrees in Mechanical Engineering at the University of Science and Technology of China in 1993 and 1996 respectively, and a Master's in Industrial Engineering and a Ph.D. in Mechanical Engineering at the University of Michigan in 2000. His research interests are focused on in-process quality and productivity improvement methodologies

incorporating statistics, system and control theory, and engineering knowledge. The objective is to achieve automatic process monitoring, diagnosis, compensation, and their implementation in various manufacturing processes. He is a member of IIE, INFORMS, ASME, and SME.

Jionghua (Judy) Jin received her B.S. and M.S. degrees in Mechanical Engineering, both from the Southeast University in 1984 and 1987 respectively, and her Ph.D. degree in Industrial and Operations Engineering at the University of Michigan in 1999. She is currently an Assistant Professor in the Department of Systems and Industrial Engineering at the University of Arizona. Her research focuses on developing a unified methodology for quality and reliability improvement through the fusion of statistics with engineering models. Her research expertise is in the areas of systematic process modeling for variation analysis, automatic feature extraction for monitoring and fault diagnosis, and optimal maintenance decision with the integration of the quality and reliability interaction. Her research is sponsored by National Science Foundation, the Air Force Office of Scientific Research, the US Department of Transportation, and Global Solar Energy Inc. She was a recipient of a CAREER Award from the National Science Foundation in 2002 and a PECASE award in 2004. She received the best paper award from the ASME, Manufacturing Engineering Division in 2000. She is a member of INFORMS, IIE, ASQC, ASME, and SME.

Contributed by the On-Line Quality Engineering Department