# Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes

YU DING

*Department of Industrial and Systems Engineering*
*Texas A&M University, College Station, TX 77843-3131*

LI ZENG and SHIYU ZHOU

*Department of Industrial and Systems Engineering*
*The University of Wisconsin-Madison, Madison, WI 53706*

Phase I analysis of nonlinear profiles aims at identifying the data from an in-control process as accurately as possible so that quality engineers can have a good reference to establish the control charts for a future process. Unlike linear profiles, which can be represented by a linear regression model with its model parameters used for monitoring and detection, nonlinear profiles are often sampled into high-dimensional data vectors and analyzed by nonparametric methods. Meanwhile, automatic in-process data-collection devices generate huge historical data sets, which must be analyzed for the presence of observations from out-of-control process conditions. The high dimensionality and data contamination present a challenge to the Phase I analysis of nonlinear profiles. This paper presents a strategy that consists of two major components: a data-reduction component that projects the original data into a lower dimension subspace while preserving the data-clustering structure and a data-separation technique that can detect single and multiple shifts as well as outliers in the data. Simulated data sets as well as nonlinear profile signals from a forging process are used to illustrate the effectiveness of the proposed strategy.

Key Words: Change-Point Detection; Clustering; Control Chart; Multivariate Analysis.

THIS PAPER investigates a strategy for Phase I analysis of nonlinear profile data. Phase I analysis, also called retrospective analysis in statistical process control (SPC), is applied to the set of historical process data, which is often a combination of data from the in-control condition (loosely called in-control data) and out-of-control conditions (loosely called out-of-control data). Phase I analysis attempts to identify the data from the in-control condition as accurately as possible so that quality engineers can have a good reference to establish the monitoring charts for a future process. For more discussions regarding the difference between analyses for Phase I and Phase II in SPC, please refer to Mahmoud and Woodall (2004) and Sullivan (2002).

As pointed out in Mahmoud and Woodall (2004), the collection of profile data for process monitoring appears to be increasingly common in industry practices. The focus of Mahmoud and Woodall (2004) is on the set of linear profiles, which can be represented by a model like those used in linear regression analysis. Based on the linear structured model, one usually uses a $T^2$ control chart monitoring the regression parameters in the model for the purpose of monitoring and detecting the changes in linear profiles. Section 1 of Mahmoud and Woodall (2004) presents a comprehensive account of state-of-the-art treatments for

Dr. Ding is an Assistant Professor in the Department of Industrial and Systems Engineering, Texas A&M University. His email address is yuding@iemail.tamu.edu.

Ms. Zeng is a Graduate Student in the Department of Industrial and Systems Engineering, The University of Wisconsin at Madison. Her email address is lzeng1@wisc.edu.

Dr. Zhou is an Assistant Professor in the Department of Industrial and Systems Engineering, The University of Wisconsin at Madison. His email address is szhou@engr.wisc. edu.

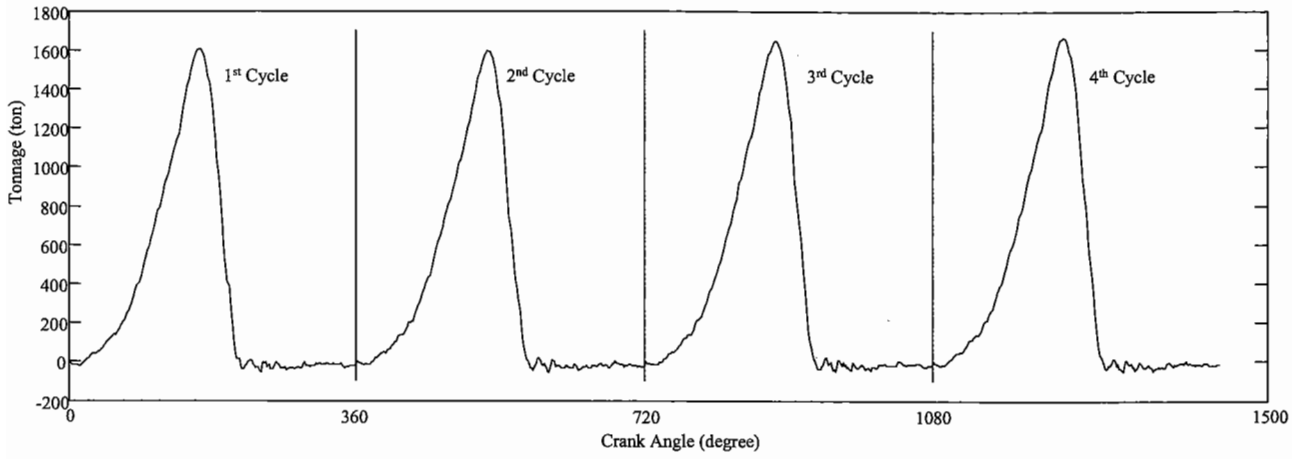


FIGURE 1. A Typical Nonlinear Profile Signal: the Forging Tonnage Signal.

linear profile data; interested readers please refer to the references therein for more details.

In this paper, we are concerned with a different class of profile data that cannot be adequately represented by a linear structured model and are generally labeled as nonlinear profiles. Figure 1 shows an example of such a nonlinear profile, which is the tonnage (i.e., forming force) signal of a forging process. The tonnage signal is obtained by strain sensors mounted on the supporting pillars of a forging press. Figure 1 shows a total of four cycles of tonnage signals from a crankshaft forging process, where the vertical axis is the forming force measured in tons, and the horizontal axis is the crank angle of the press. (The crank rotates 360° in every cycle.) A few other examples include a forming-force profile in a stamping process, a spatial profile constituted by the dimensional deviations measured at different locations on an automobile body, and a multistage quality profile constituted by the surface-finish measurements of a machined part that are tracked over a series of stages and operations. The profiles of Figure 1 are fairly smooth, but other nonlinear profiles, such as the spatial profile or the multistage profile, may not be inherently smooth. In this paper, we are concerned with the general category of nonlinear profiles, and the proposed method is not limited to smooth profiles.

Effective monitoring of nonlinear profiles is generally challenging. One immediate difficulty would be how to characterize a nonlinear profile. In practice, people have used some simple descriptive statistics to characterize the profiles, such as the maximum magnitude, the average value, and the separation between the 10% and 90% times on the waveform rise as well as on the fall (Knussmann and Rose (1993), Barnett et al. (1998)). When simple statistics are used, a great deal of information, especially those related to local characteristics and fine features in the original profile, may be ignored. For this reason, a monitoring system based merely on the simple statistics often suffers from a high false-alarm rate and/or a high miss-detection rate.

Nowdays, due to the fast development of computerized data-acquisition systems, a nonlinear profile can be sampled at very high frequency into a high-dimensional data vector. One could use this data vector for the monitoring purpose because it represents the profile well and captures necessary subtlety and fine features. As such, monitoring nonlinear profiles can be considered a particular application of multivariate process-control problems.

As for the Phase I analysis of nonlinear profiles, there are primarily two challenges to be addressed. The first challenge is the high data dimensionality resulting from the discretization of nonlinear profiles. In most cases, the profile signal could contain as high as several hundred data points, e.g., each cycle of the profile in Figure 1 contains 224 data points. Because of the curse of dimensionality, it will be ineffective to analyze such high-dimensional multivariate data directly. Even though linear profiles are also sampled into high-dimensionality data vectors, when it comes to analyzing the signals, the parameters of a linear model instead of the sampled data are used. In contrast, methods for analyzing nonlinear profiles are

almost exclusively nonparametric, and thus the data dimensionality involved in nonlinear profiles analysis is far greater than the parameter dimensionality in the case of linear profiles.

The second challenge is how to recognize the presence of data from out-of-control conditions and extract the data from the in-control condition. Including out-of-control Phase I data can lead to a biased estimation of the process mean and/or an inflated estimation of process variability, which will in turn affect the control limits, and thus eventually result in more false alarms and/or more missed detections in future monitoring.

This paper presents a strategy of Phase I analysis for nonlinear profiles monitoring. In light of the above discussion, the proposed strategy is naturally an integration of a data-reduction component and a data-separation/clustering component.

The most popular method employed to reduce the dimensionality of multivariate data is the principal components analysis (PCA) (Jackson (1991), Carreira-Perpinan (1997)). Basically, PCA will transform the original data and project them onto a lower dimensional space that preserves the majority of the variability in the original data. However, one limitation of PCA is that it does not guarantee to project the original data into a subspace that can maximize the separation of any clustering structure that may exist in the original data. In other words, even if the original data possess a distinct structure separating the in-control data from the out-of-control ones, the structure could become blurred if too few principal components are retained. Then, no matter how powerful the subsequent data-clustering algorithm is, the in-control data may not be able to be separated from the out-of-control ones. To bridge this gap, we will investigate an alternative data-reduction technique: independent component analysis (ICA) (Hyvarinen et al., 2001), with the objective of transforming the data into a subspace where the distinction of any existing structures in the data will be maximized in the resulting independent components (ICs). In the subsequent sections, we will discuss the conditions under which ICA can outperform PCA during the data reduction.

In analyzing Phase I data, people will investigate the data points outside a control limit initially established from the whole Phase I data set to see whether they correspond to the out-of-control conditions; if this seems likely, then the point is removed from the Phase I data set. As for data clustering and separation, traditional SPC also suggests applying this two-step procedure recursively to the rest of the data until no out-of-control data are detected. The final set of data points is treated as the representatives of the in-control process condition and is used to set up the control limits for future monitoring (Montgomery 2004). It will be shown later that this simple recursive procedure is mainly effective for the cases when scatter outliers constitute the out-of-control data points. But it is not effective with sustained shifts. This paper recommends a more sophisticated change-point detection algorithm, recently developed by Sullivan (2002), as the data-clustering tool, which is able to effectively detect single and multiple shifts as well as outliers.

The relationship of the proposed Phase I analysis to the entire process-monitoring procedure is illustrated in Figure 2. Because our goal is the Phase I analysis, we choose not to specify the type of moni-

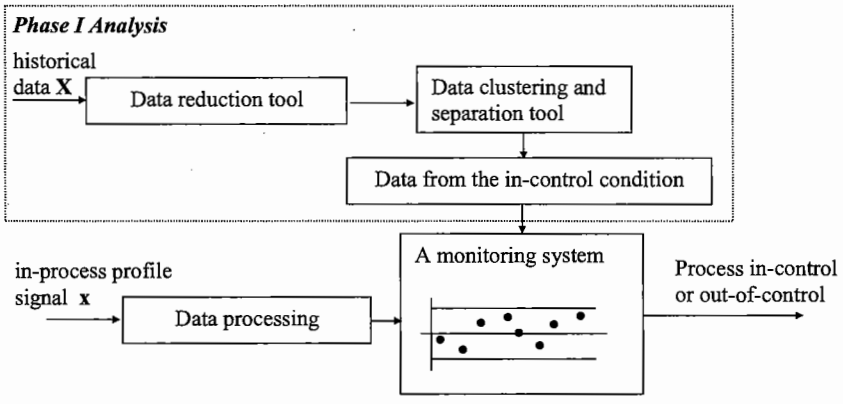


FIGURE 2. An Overview of Phase I Analysis.

toring chart to be used when a newly observed signal $\mathbf{x}$ becomes available. That is why a nonspecific "data processing" is used for Phase II in Figure 2.

The remainder of this article is organized as follows. In the subsequent sections, we first define the notation and assumptions used in the paper. Then we discuss the techniques used as the data-reduction and data-separation components, respectively. Afterward, we present two numerical examples: the first one uses a simulated data set to demonstrate the effectiveness of the recommended procedure versus other available methods; the second one is to apply the Phase I analysis to the profile signals obtained from a forging process. Finally, we will summarize the paper and include some concluding remarks.

## Notation and Assumptions

We denote an individual data point of the nonlinear profile by $x_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where $i$ is the cycle index, $j$ is the index for the data point within a cycle, $n$ is the total number of cycles, and $p$ is the dimension of the data vector for each cycle.

We can arrange the entire historical data set in a matrix $\mathbf{X}$ of dimension $n \times p$, each row of which is the data vector associated with one cycle. Meanwhile, we denote by $\mathbf{x}_i$ the transpose of the $i$th row of $\mathbf{X}$, of dimension $p \times 1$ and corresponding to the $i$th cycle of a sampled profile. For the tonnage signal in the forging process, the historical data set has $n = 530$ and $p = 224$.

Denote by $\mathbf{S}$ the sample covariance matrix of data matrix $\mathbf{X}$, i.e.,

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}},$$

where $\bar{\mathbf{x}} = (1/n) \sum_{i=1}^{n} \mathbf{x}_i$ and we assume $\mathbf{S}$ positive definite. Denote by $\{\lambda_k, \mathbf{e}_k\}$ the $k$th eigenvalue–eigenvector pair of $\mathbf{S}$. Without loss of generality, we will generally arrange the eigenvalues in descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. We refer to $\mathbf{e}_k$ as the $k$th eigenvector of $\mathbf{S}$, meaning that it is the eigenvector associated with the $k$th eigenvalue of $\mathbf{S}$.

As we mentioned in the introduction, the $n$ historical samples of a nonlinear profile may be a combination of both in-control and out-of-control conditions. In this paper, we make the following assumptions: (A1) A process switches between the in-control and out-of-control conditions infrequently. (A2) The data samples associated with the out-of-control condition will be the minority among all the historical samples. (A3) The measurements obtained under a specific process condition (the in-control condition or a specific out-of-control condition) are assumed to follow a multivariate normal distribution. We feel that the first two assumptions on the occurrence of out-of-control conditions are not restrictive. Actually, they quite reasonably reflect our observations of the behavior of a discrete-part manufacturing process, which usually has slow between-sample dynamics. The last assumption is not a surprise, either, because the multivariate normal distribution is arguably the most commonly used distribution in the practice of SPC (Montgomery 2004).

## Data Reduction: PCA Versus ICA

### Difference between PCA and ICA

The objectives of PCA do not include separating the existing data structure or clusters; PCA merely looks for the projection direction corresponding to large variations. A classical example that can be commonly found in the literature of ICA (e.g., Hyvarinen 2001) is shown in Figure 3. Suppose that there are two clusters of data in a two-dimensional space, representing the in-control data and the out-of-control one, each of which is depicted by an ellipse representing a distribution contour of a bivariate normal distribution. When PCA is performed on the whole data set, it will project the data into axis $e_1$, which is parallel to the direction of the largest variability in the data. After such a projection, the distinct structure of the original data clusters will not be preserved. ICA uses a different criterion, loosely defined as *interestingness*, when it chooses the projection direction. ICA will actually project the two-dimensional data into axis $e_2$, where the distinction of the structures in the original data is maximized.

The question is what defines the *interestingness*.

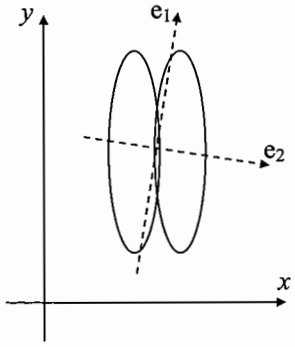


FIGURE 3. Difference Between PCA and ICA.

The general consensus (please refer to Huber (1985) and Jones and Sibson (1987)) is that the Gaussian distribution is the least interesting one, or that the interesting one should be non-Gaussian. The entropy definition measuring the non-Gaussianity for any continuous random variable $Y$ with density $f_Y(y)$ is

$$H(Y) = -\int f_Y(y) \log f_Y(y) dy = -E_Y[\log(f_Y(y))]. \tag{1}$$

One can optimize the entropy by varying the probability density function $f$, and the entropy is maximized when $f$ is Gaussian density and strictly smaller otherwise.

The above definition of interestingness appears to be well aligned with our objective of separating in-control data from the out-of-control ones. From assumption A3, the process outputs under a specific process condition follow a Gaussian distribution. Then, for combined data from both in-control and out-of-control conditions, the resulting distribution will be non-Gaussian. The projection directions for which the interesting features (or non-Gaussianity) of the data are in fact equivalent to the directions where the distinction of data clusters is obvious, e.g., the $e_2$ axis in Figure 3.

Of course, the projection for non-Gaussianity and that for the largest variability could coincide. When that happens, the resulting projection subspace from PCA and ICA will also coincide, which implies that using PCA for data reduction may be able to preserve the data-clustering structure as well in the reduced data. As illustrated in Figure 3, the clustering structures in the combined data manifest along the direction associated with the mean difference between the data samples—this direction is represented by $\mathbf{v}$, which will be defined later. An ideal projection method should project the original data onto the direction associated with the mean difference. The question of whether PCA will be able to preserve the data-clustering structure depends on whether the subspace spanned by its first several eigenvectors includes the direction of the mean difference.

In the sequel, we try to provide a general understanding of the aforementioned question by considering combined data from two Gaussian distributions: $\mathbf{X}_a$ has $n_a$ data samples with sample mean $\bar{\mathbf{x}}_a$ and sample covariance matrix $\mathbf{S}_a$, and $\mathbf{X}_b$ has $n_b$ data samples with sample mean $\bar{\mathbf{x}}_b$ and sample covariance matrix $\mathbf{S}_b$. The total number of data samples is $n = n_a + n_b$. Without loss of generality, we assume $\mathbf{X}_a$ is the in-control data, and hence $n_a > n_b$. Our purpose here is to facilitate conceptual understanding so that we consider a simple case that allows us to see the results clearly. For this reason, we assume that $\mathbf{S}_a = \lambda^a \mathbf{I}$ ($\mathbf{S}_a$'s eigenvalues are all equal) and $\mathbf{S}_b$ is diagonal with $\{\lambda_i^b\}_{i=1}^p$ as its eigenvalues. Denote by $\mathbf{v}$ the mean difference between the two data samples, i.e., $\mathbf{v} = \bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b$ and $\mathbf{S}$ as the sample covariance matrix of the combined data. Then, the $i$th eigenvalue $\lambda_i$ of $\mathbf{S}$ is

$$\lambda_i = \frac{(n_a - 1)}{n - 1}\lambda^a + \frac{(n_b - 1)}{n - 1}\lambda_i^b + m_i \frac{n_a n_b}{(n-1)n}\|\mathbf{v}\|^2, \tag{2}$$

where $\|\cdot\|$ is a 2-norm and $m_i$ is a constant, satisfying $0 \le m_i \le 1$ and $\sum_{i=1}^p m_i = 1$. (The derivation is included in Appendix I.) The procedure for determining $m_i$ is given in Wilkinson (1965, pp. 97–98), but it is algebraically involved.

When the elements in $\mathbf{v}$ are roughly the same, corresponding to a mean difference between data samples along a general direction in the original data space, we call it a whole-space mean difference. Under this circumstance, the subspace spanned by the major eigenvectors corresponding to the largest eigenvalues will include a substantial amount of the mean difference, meaning that the resulting PCs will be able to preserve the clustering data structure in the projected subspace. According to Equation (2), the order of the resulting $\lambda_i$ will be primarily decided by the order of the $\lambda_i^b$.

When the mean difference is more prominent in certain direction, i.e., a few elements in $\mathbf{v}$ are nonzero while the others are zero, we call this $\mathbf{v}$ a subspace mean difference. Suppose that we have $r$ nonzero elements in $\mathbf{v}$ and $p - r$ zero elements. Then the eigenvalues of $\mathbf{S}$ consist of two groups: the first group is associated with the $(p-r)$-dimensional subspace, where there is no mean difference and the eigenvalues correspond to $m_i = 0$ in Equation (2); the second group is associated with the $r$-dimensional subspace, where there is a mean difference, and the eigenvalues correspond to a nonzero $m_i$, and all the nonzero $m_i$ will still sum to unity, i.e., $\sum_{i=1}^r m_i = 1$.

Now suppose that the elements in $\mathbf{v}$ along the direction of originally large variability are zero, say $v_1 = 0$. We are interested in knowing when an eigenvalue of $\mathbf{S}$ in the second group will be larger than the eigenvalues in the first group because of the mean difference existing in that subspace. Denote by $\lambda_1$ the largest eigenvalue in the first group and by $\lambda_i$ an eigenvalue in the second group. If $\lambda_1 > \lambda_i$, it implies that the direction of the largest PC will not align

with the direction of $\mathbf{v}$. For $\lambda_1 > \lambda_i$, we have

$$\frac{n_a-1}{n-1}\lambda^a + \frac{n_b-1}{n-1}\lambda_1^b > \frac{n_a-1}{n-1}\lambda^a + \frac{n_b-1}{n-1}\lambda_i^b + m_i\frac{n_a n_b}{(n-1)n}\|\mathbf{v}\|^2,$$

$$\Leftrightarrow$$

$$\frac{n_b-1}{n-1}(\lambda_1^b - \lambda_i^b) > m_i\frac{n_a n_b}{(n-1)n}\|\mathbf{v}\|^2$$

$$\Leftrightarrow$$

$$\|\mathbf{v}\| < \sqrt{\frac{(n_b-1)n}{m_i n_a n_b}(\lambda_1^b - \lambda_i^b)}. \quad (3)$$

If both $n_a$ and $n_b$ are much greater than 1, and let $n_b = \kappa n$, then $0 < \kappa < 0.5$ because we postulate $n_a > n_b$. As such, Equation (3) becomes

$$\|\mathbf{v}\| < \sqrt{\frac{1}{m_i(1-\kappa)}(\lambda_1^b - \lambda_i^b)}, \quad (4)$$

suggesting that when the mean difference is smaller than the right-hand side (RHS) amount in the above inequality, the projection output to the largest PC will miss the mean difference. Equation (4) can be further simplified for the case when there is only one nonzero element in $\mathbf{v}$, say $v_i \neq 0$. Then $m_i = 1$, so that Equation (4) becomes $v_i < \sqrt{[1/(1-\kappa)](\lambda_1^b - \lambda_i^b)}$. Because many PCA users likely keep a few largest PCs, this can be extended to more than one PC by replacing $\lambda_1^b$ with the smallest eigenvalue retained.

According to Equation (4), when the combined historical data has nearly the same amount of observations from out-of-control conditions as the in-control ones, i.e., $n_a \approx n_b$, the mean difference can go as large as $\sqrt{2(\lambda_1^b - \lambda_i^b)/m_i}$ without being projected onto the major PC's direction. The mean difference allowing $\lambda_1 > \lambda_i$ decreases when the amount of out-of-control data decreases; the other extreme approaches $\sqrt{(\lambda_1^b - \lambda_i^b)/m_i}$ when $n_a \gg n_b$.

Equation (4) holds for $n_b$ much greater than one. When $n_b$ is close to one, it corresponds to the scenario of outliers, and then Equation (3) is more appropriately approximated as

$$\|\mathbf{v}\| < \sqrt{\frac{(n_b-1)}{m_i n_b}(\lambda_1^b - \lambda_i^b)}, \quad (5)$$

under which circumstance, the allowed mean difference for $\lambda_1 > \lambda_i$ will be smaller than $\sqrt{(\lambda_1^b - \lambda_i^b)/m_i}$. When $n_b$ is 1 (a single outlier), the projected subspace associated with the major PCs will be able to capture the outlier, almost regardless of the magnitude of $\mathbf{v}$.

Combining the above points, for this simple case, we can conclude that PCA upon retaining a few largest eigenvalues will likely miss the data clustering structure when there is a subspace mean difference of mild magnitude. On the other hand, PCA is likely to preserve the data-clustering structure in its projection subspace for a whole-space mean difference, or a mean difference of large magnitude, or when the out-of-control data are merely some outliers. PCA will be worse off when the out-of-control data are nearly as numerous as the in-control data and be better off when there are many more in-control observations. Later, with the help of numerical examples, we show that these generalizations are still reasonable with more complicated covariance matrix structures.

The fact that the subspace created by PCA may not be sensitive to a mean difference has been noted in prior research. Runger (1996) proposed a $U^2$ chart for detecting a subspace mean shift (equivalent to the subspace mean difference in our case). In his paper, Runger assumed that the subspace within which a mean shift is likely to occur is known. A $U^2$ chart is simply a multivariate $\chi^2$ for the projected data in a predefined subspace that may be different from the PC's subspace. Following a similar idea, Runger et al. (2005) later presents a POBREP framework for monitoring multivariate fault patterns appearing only in a subspace; the $U^2$ chart can be considered as a special case of the more general POBREP approach. Again, the subspace within which the fault pattern appears is known. In our case, we do not know the subspace in which a mean difference lies; an ICA algorithm is an attempt to automatically find such a subspace by optimizing a linear transformation.

## Numerical Routine for ICA

Similar to PCA, ICA is also based on a linear transformation of the original data, but its computation procedure is far more complicated than that for PCA. Fortunately, ICA algorithms are available in commercial software language such as the MATLAB and *R*. We use the *fastICA* function in MATLAB to perform the data transformation in our implementation. One thing to note is that, in most efficient algorithms for finding ICs, an approximation has been used to replace the difficult-to-compute entropy, which is

$$J(x_i) = [E\{g(x_i)\} - E\{g(z)\}]^2, \quad (6)$$

where $x_i$ is the $i$th component in $\mathbf{x}$, $z$ is a standardized Gaussian, and $g(\cdot)$ is a smooth, nonlinear function. The objective function is solved through numerical optimization routines.

Even though ICA theoretically should do no worse than PCA in terms of finding the subspace of the data with clustering structures, the numerical routine of ICA may not be able to find the global optimum and thus could perform worse than PCA. This suggests the necessity of further research for better objective functions and more powerful, robust solution procedures to realize ICA, which is indeed an on-going pursuit. Among four versions of $g(\cdot)$ and two optimization procedures provided by the current MATLAB *fastICA* function, our empirical experiences indicate that using the Gaussian density function as $g(\cdot)$ together with a parallel (instead of a sequential) procedure to estimate ICs appears to be a better combination than the other options.

### Determining the Number of ICs and PCs

To utilize ICA or PCA for dimensionality reduction, we need also to decide the number of significant ICs or PCs to retain. In other words, we need to identify the number of significant variation directions in the data set. Different methods to solve this problem have been studied and compared through simulation by Apley and Shi (2001). This problem can be formulated as a hypothesis-testing problem: to determine if $k$ major variation directions exist in the system, we can test the hypothesis testing of $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \sigma^2 = \lambda_{k+1} = \cdots = \lambda_p$, where $\lambda_i$, $i = 1, \ldots, p$ are the eigenvalues of the covariance matrix of the profile data, and $\sigma^2$ represents the background noise level. Several asymptotic testing procedures are available. Apley and Shi (2001) recommended using either the Akaike information criteria (*AIC*) or the minimum description length (*MDL*) information criteria to estimate the number of significant variation directions, equivalent to the number of significant ICs or PCs to retain. In this paper, we choose the *MDL* criterion, which is defined as

$$MDL(l) = n(p-l)\log(a_l/g_l) + l(2p-l)\log(n)/2, \quad (7)$$

where $a_l$ and $g_l$ are the arithmetic mean and the geometric mean of the $p-l$ smallest eigenvalues of the sample covariance matrix of the profile data, respectively. To use this criteria, $MDL(l)$ is evaluated for $l = 0, \ldots, p-1$. The number of significant ICs or PCs to retain is chosen as the $l$ that minimizes $MDL(l)$.

When the *MDL* test retains a relatively large number of PCs, which may in turn cause difficulty in subsequent analyses, a visual aid such as a scree plot (i.e., a pareto plot) suggested by Johnson and Wichern (2002, p. 441) can be used to facilitate in deciding a more appropriate number of eigenvalues to retain.

### Apply PCA or ICA to the Wavelet Coefficients of Original Data

For analyzing nonlinear profiles, another widely used method is to perform a wavelet transformation and then work on the wavelet coefficients that presumably represent the original data. However, the number of the resulting nonzero wavelet coefficients is usually too many to enable effective decision making. Hence, dimensionality reduction is still a necessary step.

Prior work has been reported in using different methods to select a subset of wavelet coefficients from the whole set (Jin and Shi (1999, 2001), Koh et al. (1999a, b), Zhou et al. (2004), Pittner and Kamarthi (1999), Lada et al. (2002)). But it comes as no surprise that PCA is still the most commonly used method if an aggressive dimension reduction is required (e.g., Kosanovich and Piovoso (1997), Bakshi (1998)). It is our belief that ICA can also be applied to the wavelet coefficients, just as PCA, to help reduce the data dimensionality because our previous discussion is based on generic multivariate data. We are not yet certain about when it is helpful to include a wavelet transform to preprocess the profile data. That issue is out of the scope of this article but certainly warrants some future research.

## Data Clustering and Separation

Because of the independence among the resulting ICs, the data-separation task can be applied to each IC individually. We will treat a data sample as the in-control data if all corresponding ICs so indicate.

There are a few available alternatives to fulfill the data-separation task for an individual, univariate IC. The most commonly used procedure is a recursive application of a control chart, which removes any out-of-control point iteratively, namely, a control chart will be established using the entire data set; second, the data points outside the control limits will be removed. Then, the control limits will be revised using the rest of the data and be applied to the new data set. The procedure will be repeated until no out-of-control data point is found.

Another idea is to apply a clustering method to

the original data set to separate the data into different clusters, each of which presumably corresponds to different conditions (a different mean and/or a different variance). Suppose that, under the in-control condition, the process follows $N(\mu_0, \sigma_0^2)$ and, under out-of-control conditions, the process follows $N(\mu_i, \sigma_i^2)$, where $\mu_i \neq \mu_0$ or $\sigma_i^2 \neq \sigma_0^2$, $i = 1, \cdots, K$, and $K$ is the cluster number. Then, the resulting data may be modeled by a mixture model. An EMMIX (expectation-maximization mixture) clustering method (McLachlan and Peel, 2000) is available for separating the clusters in mixture data. Once the original data set is clustered by the EMMIX method, it would not be difficult to extract out the in-control data by utilizing assumption A2 that the in-control data are more numerous.

The third alternative is to use a change-point detection algorithm. A change-point detection algorithm detects the instances in an ordered sequence of observations when the distribution (characterized by its first two moments) undergoes a change. If all the change points in the original data set can be perfectly detected, the in-control data can then be easily separated. Sullivan (2002) recently developed a change-point detection algorithm, which can detect single/multiple change points as well as outliers very effectively and thus meets our needs here; a procedure for implementing Sullivan's method is outlined in Appendix II.

The question is which procedure will be the most effective in terms of data separation. Conceptually, the latter two procedures are more sophisticated than the recursive use of a control chart. The problem with the recursive use of a control chart is that the control limits used to remove the out-of-control data are contaminated by the out-of-control data themselves. Unless there exist only scattered outliers, the recursive use of control charts is unlikely to be effective.

Between the EMMIX clustering method and the change-point detection algorithm, the latter is likely to be more effective for the manufacturing processes of concern, of which we have assumed that the switches between the in-control and out-of-control conditions happen infrequently. It means that the in-control and out-of-control data will form reasonably long time-sequence segments, and outliers will not occur often. For example, given 10 data points, it is unlikely for the process to be that $\{1, 3, 5, 7, 9, 10\}$ are the in-control data points and $\{2, 4, 6, 8\}$ are the out-of-control data points. This type of data is similar to the mixture data but also different in the sense that, once there is a change to a distribution, the successive observations are likely to be from the same distribution. Obviously, the combined data from the manufacturing process are better modeled by a change-point model than a traditional clustering model that does not consider this ordered sequence information. As a result, the change-point detection utilizing the ordered sequence information can more effectively detect infrequent process changes. By contrast, the EMMIX clustering method does not consider the ordered sequence information at all and is more likely to assign a data point in the middle of a long segment to a different cluster.

Although the conceptual understanding points to the change-point detection algorithm for our application, the effectiveness of the three procedures has not been compared in the literature. In Sullivan (2002), the change-point detection algorithm was compared with the $X$-chart and the CUSUM chart. However, the comparison was performed only for the cases of a mean shift, while the variance was assumed unchanged. We perform a more comprehensive numerical comparison among the three procedures for a collection of 15 scenarios, including mean shifts, variance changes, and outliers.

In the following numerical comparison, we consider a univariate normally distributed sequence of 100 data points. For simplicity, we assume there is only one change point or one outlier, and the in-control observations are 60% of the data. Without loss of generality, the in-control distribution is taken to be zero mean ($\mu_0 = 0$) and unit variance ($\sigma_0 = 1$). At a change point, four types of mean shifts (small, medium, large, and very large) and three types of variance change (decreasing, none, and increasing) are considered, which give a total of 12 combinations. Together with three different magnitudes for an outlier, we compare the three algorithms for a total of 15 scenarios.

The detailed scenario information is summarized in Table 1. The $S_1$ to $S_{12}$ denote the 12 combinations of mean shifts and variance changes associated with a change point. Under each scenario, two mean shifts (denoted by $\delta$) are simulated and the average of their performance data will be used in the comparison. In the meanwhile, the variance after a change point, denoted by $\sigma_1$, is also indicated under each category. For example, under $S_1$, two mean shifts of $0.5\sigma_0$ and $1\sigma_0$, respectively, are used to represent the scenario with a small mean shift. Under $S_1$, the variance gets smaller ($\sigma_1 = 0.5\sigma_0$) after the change point.

TABLE 1. The Scenarios for Numerical Comparison

| | | | | |
|---|---|---|---|---|
| $S_1$ (small shift, decreased variance) | $S_2$ (medium shift, decreased variance) | $S_3$ (large shift, decreased variance) | $S_4$ (very large shift, decreased variance) | $S_5$ (small shift, same variance) |
| $\delta = 0.5\sigma_0$; $\delta = 1\sigma_0$; and $\sigma_1 = 0.5\sigma_0$ | $\delta = 1.5\sigma_0$; $\delta = 2\sigma_0$; and $\sigma_1 = 0.5\sigma_0$ | $\delta = 2.5\sigma_0$; $\delta = 3\sigma_0$; and $\sigma_1 = 0.5\sigma_0$ | $\delta = 3.5\sigma_0$; $\delta = 4\sigma_0$; and $\sigma_1 = 0.5\sigma_0$ | $\delta = 0.5\sigma_0$; $\delta = 1\sigma_0$; and $\sigma_1 = \sigma_0$ |
| $S_6$ (medium shift, same variance) | $S_7$ (large shift, same variance) | $S_8$ (very large shift, same variance) | $S_9$ (small shift, increased variance) | $S_{10}$ (medium shift, increased variance) |
| $\delta = 1.5\sigma_0$; $\delta = 2\sigma_0$; and $\sigma_1 = \sigma_0$ | $\delta = 2.5\sigma_0$; $\delta = 3\sigma_0$; and $\sigma_1 = \sigma_0$ | $\delta = 3.5\sigma_0$; $\delta = 4\sigma_0$; and $\sigma_1 = \sigma_0$ | $\delta = 0.5\sigma_0$; $\delta = 1\sigma_0$; and $\sigma_1 = 2\sigma_0$ | $\delta = 1.5\sigma_0$; $\delta = 2\sigma_0$; and $\sigma_1 = 2\sigma_0$ |
| $S_{11}$ (large shift, increased variance) | $S_{12}$ (very large shift, increased variance) | $O_1$ | $O_2$ | $O_3$ |
| $\delta = 2.5\sigma_0$; $\delta = 3\sigma_0$; and $\sigma_1 = 2\sigma_0$ | $\delta = 3.5\sigma_0$; $\delta = 4\sigma_0$; and $\sigma_1 = 2\sigma_0$ | magnitude $= 3\sigma_0$ | magnitude $= 6\sigma_0$ | magnitude $= 9\sigma_0$; |

Thus, $S_1$ corresponds to the case of small mean shift and a decreased variance. The other scenarios follow the same interpretation. For the three types of outliers, their magnitudes are indicated in Table 1 as a multiple of $\sigma_0$.

In order to compare the alternative procedures, we need to define a set of performance indices. Our thinking is as follows. In Phase I analysis, when the historical data is a combination of in-control and out-of-control data, the objective is to extract the in-control data as accurately as possible. The ideal case is where all in-control observations are correctly identified. We hence use it as a benchmarking reference. In reality, there are always misclassifications, either treating out-of-control as in-control or the other way around. Thus, a sensible performance measure should be able to evaluate the impact of misclassifications in Phase I analysis on the performance of a control chart used in Phase II.

In other words, when a data-separation algorithm is used to identify the in-control data, the output from each data-separation algorithm is actually a combination of some in-control data and some out-of-control data; we label the output as the contaminated data. The contaminated data are then used to set up the control limits to be used in Phase II. The performance of such a control chart in Phase II, characterized by its average run length ($ARL$), cannot be the same as the ideal case—they will generally do worse, either have a shorter $ARL_0$ (i.e., more frequent false alarms) or a larger $ARL_1$ (i.e., longer delayed detection), or both.

Obviously, the data-separation algorithm that achieves the smallest adverse change from the ideal case in both $ARL_0$ and $ARL_1$ measures is considered the best method. For this reason, we choose to use the changes in $ARL_0$ and $ARL_1$, as compared with the ideal case, to benchmark the performance of the three data-separation procedures. Please note that nonadverse changes, i.e., an increased $ARL_0$ or a decreased $ARL_1$ when using the contaminated data, are treated the same as no change.

First, consider the future monitoring on an actual in-control process in Phase II. We define the index for the change of $ARL_0$ as

$$\Delta ARL_0 = \frac{\max(ARL_{0,\text{ideal}} - ARL_{0,\text{contam}}, 0)}{ARL_{0,\text{ideal}}}, \quad (8)$$

where $ARL_{0,\text{ideal}}$ is the $ARL_0$ in the ideal case when the control limits are determined by the true in-control data, and $ARL_{0,\text{contam}}$ is the $ARL_0$ when the control limits are determined by the contaminated data. The numerator in Equation (8) is nonnegative, meaning that, when the $ARL_0$ is longer than the ideal $ARL_0$, the resulting numerator is zero.

Second, consider the future monitoring on a mean shift in Phase II. We define the index for the change of $ARL_1$ under specific mean-shift magnitudes. We adopt three $\Delta ARL_1$'s, corresponding to the cases of

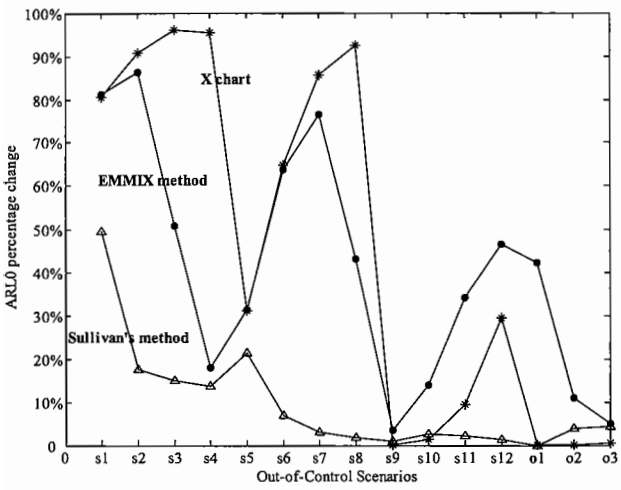


FIGURE 4. The Change of $ARL_0$ for the Three Separation Procedures.

when the mean shift of $\delta = 0.5\sigma_0$, $\sigma_0$, $2\sigma_0$, respectively, exists in the process during Phase II. So we define

$$\Delta ARL_1(\delta) = \frac{\max(ARL_{1,\text{contam}}(\delta) - ARL_{1,\text{ideal}}(\delta), 0)}{ARL_{1,\text{ideal}}(\delta)}, \tag{9}$$

where the notation generally follows that for $ARL_0$, except that $\delta$ indicates a specific mean shift. Note that because we are concerned with a longer $ARL_1$, the order of $ARL_{1,\text{ideal}}(\delta)$ and $ARL_{1,\text{contam}}(\delta)$ in the numerator of Equation (9) is flipped as compared with that in Equation (8).

Using the performance indices in Equations (8) and (9), we perform numerical simulations (with 5,000 replications) to compare the three data-separation procedures in the context of the 15 scenarios as defined in Table 1. The results of changes in $ARL_0$ and $ARL_1$ are displayed in Figures 4 and 5, respectively. In Figure 4, the vertical axis is the percentage of the $ARL_0$ change. The ideal is no change, and the smaller, the better the performance. For example, with Sullivan's algorithm for $S_1$, the $ARL_0$ is 50% shorter than the ideal case but it is worse (80% shorter) with the use of the EMMIX algorithm or the $X$-chart. In Figure 5(a)–(c), the vertical axis is the logarithm of the percentage changes in $ARL_1$, meaning that a zero on this axis actually corresponds to a 100% increase in $ARL_1$ instead of no change. In Figure 5(b), for example, with Sullivan's method for $S_3$, the $ARL_1$ will increase about $10^{-2} \times 100\% = 1\%$ (the reading from Figure 5(b) is about $-2$); with the EMMIX algorithm, the $ARL_1$ will increase about $10^{-1} \times 100\% = 10\%$ (the reading from Figure 5(b) is about $-1$), and with the $X$-chart, the $ARL_1$ will increase $10^0 \times 100\% = 100\%$ (the reading from Figure 5(b) is about 0), all as compared with the ideal case.

Figures 4 and 5 clearly demonstrate that the Sullivan's change-point detection algorithm outperforms (in terms of a smaller decrease in $ARL_0$ or a smaller increase in $ARL_1$) the other two methods for a vast majority of the scenarios. Only for outliers with a very large magnitude ($6\sigma_0$ and $9\sigma_0$) does the recursive use of an $X$-chart or the EMMIX method demonstrate some advantage. This verifies our conceptual understanding outlined before. As such, in our Phase I analysis, we choose Sullivan's change-point detection algorithm as the data separation tool.

## Numerical Examples

This section presents two examples. The first one is a simulated process, where we know when the process distribution changes. It will be used to compare the proposed Phase I analysis with other alternatives. The second example is a Phase I analysis on the tonnage signal from a forging process.

### Simulation Scenarios

The simulated data set consists of 1,000 samples of 20 variables, i.e., $n = 1{,}000$ and $p = 20$. The in-control distribution is a normal distribution of zero mean and variance around .5. Two process changes are injected into $x_1 - x_4$ and $x_5 - x_6$ at different time instances. The first change occurs to $x_1 - x_4$ at sample #101 and it has $n_{b1}$ samples. The first out-of-control observations follow a normal distribution with $\mu_1$ and

$$\Sigma_1 = \begin{bmatrix} 4 & 1.5 & 1.3 & .8 \\ 1.5 & 4 & 1.2 & .7 \\ 1.3 & 1.2 & 4 & .6 \\ .8 & .7 & .6 & \lambda_b \end{bmatrix},$$

where $n_{b1}$, $\lambda_b$, and $\mu_1$ are to be determined for six different scenarios. The second change occurs to $x_5 - x_6$ at sample #651. The out-of-control data follow a normal distribution of

$$\mu_2 = [-3 \quad 0]^{\mathrm{T}}, \quad \Sigma_2 = \begin{bmatrix} 9 & 1.9 \\ 1.9 & 2 \end{bmatrix},$$

and have $n_{b2}$ samples.

This simulation aims at verifying the general understanding of the difference among available methods presented in the previous sections. As such, six

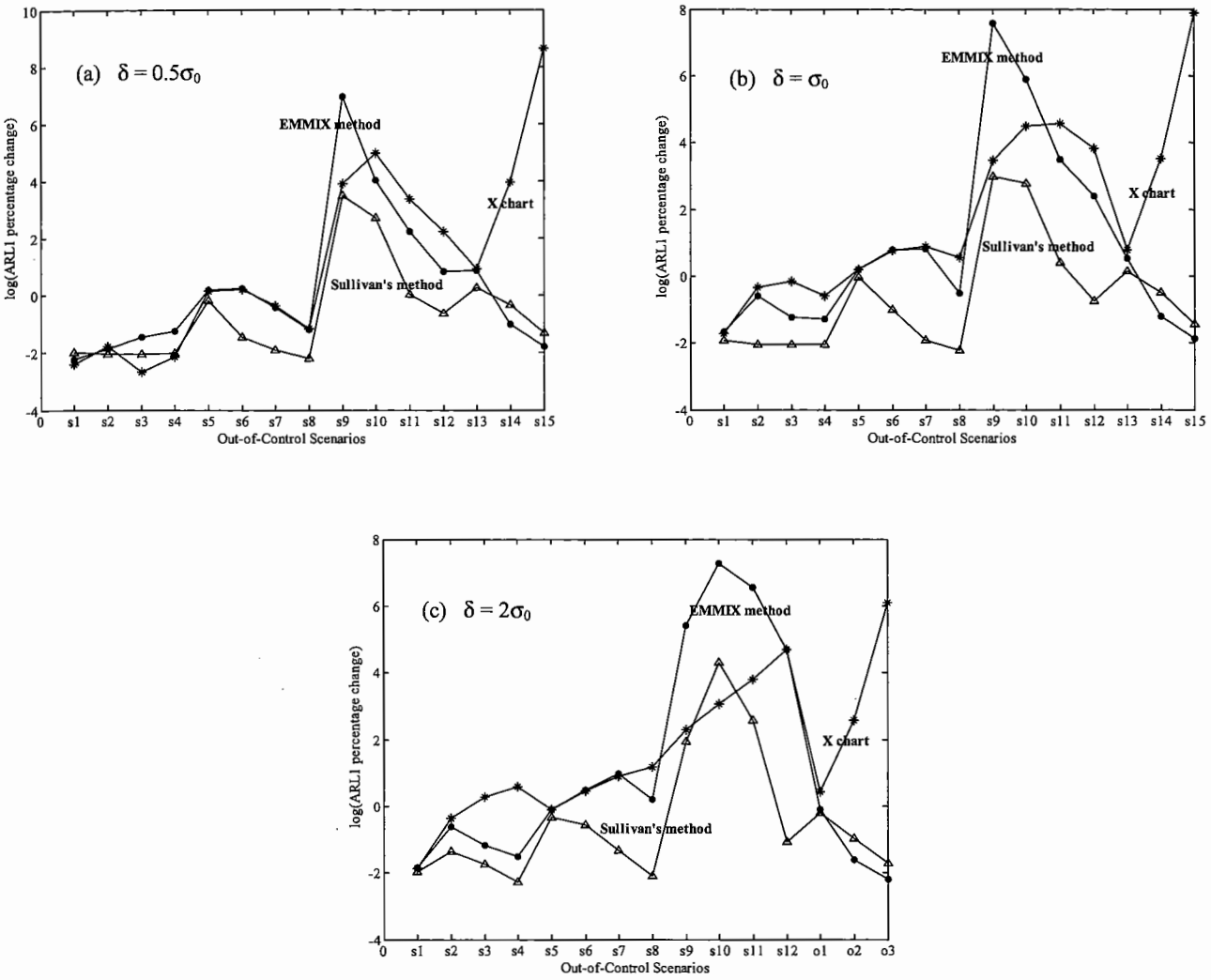


FIGURE 5. The Change of $ARL_1$ for the Three Separation Procedures.

simulation scenarios are created by changing the parameter values. The choices of parameters and the interpretations for the six scenarios are summarized in Table 2.

### Compared with Alternative Procedures

Our proposed Phase I analysis procedure integrates ICA and the change-point detection, and the combined procedure will be called ICA+CPD below. Two other alternatives are compared. One is to perform PCA to reduce the data dimension and then to apply the change-point detection algorithm directly to the resulting PCs, a procedure called PCA+CPD. Another alternative is the commonly used method that applies a multivariate Hotelling $T^2$ chart to a much reduced number of PCs. In other words, PCA will reduce the data dimension first, and then the Hotelling $T^2$ chart will be established for the resulting PCs to detect any points outside the control limit as out-of-control. Similar to other control charts, this procedure will be recursively applied to the data set until no out-of-control point occurs. This procedure is called PCA+$T^2$.

Toward our goal for Phase I analysis, which is to identify the in-control data accurately, we consider the following performance indexes. The first index is the percentage of in-control data correctly identified:

$$p_{ID} = \#\ of\ in\text{-}control\ data\ correctly\ identified/ \\ \#\ of\ total\ in\text{-}control\ data. \quad (11)$$

The second index indicates the rate of misclassifica-

TABLE 2. Summary for Six Simulation Scenarios

| Case | Parameters | Interpretation |
|---|---|---|
| 1 | $n_{b1} = 450$, $n_{b2} = 50$, $\lambda_b = .5$, $\boldsymbol{\mu}_1 = [-.2 \quad -.1 \quad -.1 \quad 2]^{\mathrm{T}}$ | Subspace mean difference of a mild magnitude; mean difference aligns with a small eigenvalue ($\lambda_b = .5$); a relatively large number of out-of-control data, $\kappa = .45$. |
| 2 | $n_{b1} = 5$, $n_{b2} = 1$, $\lambda_b = .5$, $\boldsymbol{\mu}_1 = [-.2 \quad -.1 \quad -.1 \quad 2]^{\mathrm{T}}$ | Outliers |
| 3 | $n_{b1} = 450$, $n_{b2} = 50$, $\lambda_b = 2.5$, $\boldsymbol{\mu}_1 = [-.2 \quad -.1 \quad -.1 \quad 2]^{\mathrm{T}}$ | Subspace mean difference of a mild magnitude; mean difference aligns with a large eigenvalue ($\lambda_b = 2.5$); a relatively large number of out-of-control data, $\kappa = .45$. |
| 4 | $n_{b1} = 150$, $n_{b2} = 50$, $\lambda_b = .5$, $\boldsymbol{\mu}_1 = [-.2 \quad -.1 \quad -.1 \quad 2]^{\mathrm{T}}$ | Subspace mean difference of a mild magnitude; mean difference aligns with a large eigenvalue ($\lambda_b = .5$); a relatively large number of out-of-control data, $\kappa = .15$. |
| 5 | $n_{b1} = 350$, $n_{b2} = 50$, $\lambda_b = .5$, $\boldsymbol{\mu}_1 = [-.2 \quad -.1 \quad -.1 \quad 4]^{\mathrm{T}}$ | Subspace mean difference of a large magnitude; mean difference aligns with a large eigenvalue ($\lambda_b = .5$); a moderate number of out-of-control data, $\kappa = .35$. |
| 6 | $n_{b1} = 250$, $n_{b2} = 150$, $\lambda_b = .5$, $\boldsymbol{\mu}_1 = [1.5 \quad 0 \quad 0 \quad -4]^{\mathrm{T}}$ | Whole-space mean difference of a small magnitude; a relatively large number of out-of-control data. |

tions:

$$p_{MIS} = \#\ \textit{of out-of-control data identified as in-control data/} \ \#\ \textit{of total in-control data.} \qquad (12)$$

Ideally, one would want $p_{ID} = 100\%$ and $p_{MIS} = 0$. Generally, a big value in $p_{ID}$ and a small value in $p_{MIS}$ are preferable. The above two indices will ultimately indicate how well the in-control mean vector and variance–covariance matrix can be estimated. The estimation of in-control mean vector and variance–covariance matrix will in turn affect the two types of errors for future monitoring.

We also use an alternative index that may measure more directly the accuracy of the estimation of in-control mean and covariance matrix. To that end, we employ a Bayesian-type model-validation procedure as outlined in Gelman et al. (2003, p. 162) by treating the estimated mean vector and covariance matrix as the parameters of a multivariate normal distribution. Draw a large number of samples, say 10,000, from the above normal distribution. For each multivariate sample vector $\mathbf{x}_s$, compare it with the true distribution (the distribution parameters are the true mean vector and true covariance matrix) for a preset $\alpha$, say $\alpha = 1\%$. If $(\mathbf{x}_s - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} (\mathbf{x}_s - \boldsymbol{\mu}_0) < \chi^2_\alpha(p)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the true mean and true covariance, then the corresponding sample is not ruled out for being from the true distribution; otherwise, it is considered not from the true distribution. Computing the percentage of samples that do not satisfy the above test gives us an empirical $\alpha$ value. If the samples are indeed from the same distribution as the true distribution, the empirical $\alpha$ will match the preset $\alpha$ value. We can establish the significance of how the estimated distribution is different from the true distribution by calculating the $p$-value of the empirical $\alpha$.

We implement the three procedures to perform Phase I analysis on the simulated data set, respectively. The upper control limit (UCL) used in Sullivan's change-point detection algorithm is chosen to correspond to 0.0027 false-detection probability for a sequence of 1,000 observations. Given the decision rule for the first two procedures is that one observation is considered out-of-control if any univariate IC or PC is classified as out-of-control, the equivalent

TABLE 3. Comparison of the Alternative Phase I Analysis Procedures

| | ICA+CPD | | | | PCA+CPD | | | | PCA+$T^2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | $p_{ID}$ | $p_{MIS}$ | $\alpha$ (%) | $p$-value | $p_{ID}$ | $p_{MIS}$ | $\alpha$ (%) | $p$-value | $p_{ID}$ | $p_{MIS}$ | $\alpha$ (%) | $p$-value |
| 1 | .931 | .016 | 1.130 | .108 | .705 | .175 | 6.117 | .0 | .997 | .867 | 8.450 | .0 |
| 2 | 1.00 | .005 | 1.043 | .323 | 1.00 | .005 | 1.047 | .323 | .993 | .005 | .998 | .516 |
| 3 | .663 | .029 | 1.906 | .0 | .771 | .039 | 1.864 | .0 | .998 | .836 | 9.360 | .0 |
| 4 | .959 | .011 | 1.045 | .322 | .983 | .012 | 1.077 | .225 | .990 | .148 | 1.440 | .0 |
| 5 | .993 | .012 | 1.042 | .321 | .885 | .020 | 1.210 | .017 | .997 | .511 | 13.70 | .0 |
| 6 | .938 | .021 | 1.132 | .090 | .991 | .007 | 1.036 | .356 | .997 | .441 | 9.078 | .0 |

false-detection probability for the whole procedure is $1-(1-0.0027)^k$, where $k$ is the number of PCs or ICs retained. This equivalent false-detection probability is used to set the UCL for the $T^2$ chart in the PCA+$T^2$ procedure. In the Bayesian validation of the estimated distribution, $\alpha$ is chosen to be 1%. The distribution of $\alpha$ is established empirically via a 100,000-replicate simulation.

Table 3 summarizes the performance comparison results using the three competing methods. The values of $p_{ID}$, $p_{MIS}$, and empirical $\alpha$ are the average of 1,000 trials. The $p$-value is determined using the average empirical $\alpha$ and the distribution of $\alpha$ upon sampling.

From Table 3, we observe the following:

(1) Using the combination of PCA+$T^2$ chart for Phase I analysis is not appropriate unless there exist only a few outliers in the historical data set. Our experience indicates that the historical data set is much less clean than expected and usually contaminated by several groups of long runs of out-of-control data in a relatively large number. As such, the procedure of PCA+$T^2$ falls short of removing those out-of-control data and thus the estimation of the in-control distribution is significantly different from the true distribution, as indicated by a near-zero $p$-value. The other two procedures using the change-point detection will do equally well as the PCA+$T^2$ for the outlier case but much better when $n_{b1}$ and $n_{b2}$ are greater.

(2) According to our previous discussion about the difference between ICA and PCA, Case 1 is the scenario when ICA is supposed to do better than PCA, and it in fact does. Both $p_{ID}$ and $p_{MIS}$ from ICA are much better than those obtained by PCA. The resulting in-control data extracted by ICA will lead to an estimation of distribution closer to the true distribution ($p$-value = .108), while that obtained by PCA does not.

(3) For Cases 2–6, ICA does not always perform better than PCA. For Cases 2, 4, 5, both perform almost equally well. For some cases, e.g., Case 5, ICA performs better than PCA, as judged by the absolute value of $\alpha$ and the $p$-value. But, PCA apparently performs better than ICA for Case 6, which is the scenario of a whole-space mean difference. But ICA still manages to have a good enough identification of the in-control data so that the estimation of in-control distribution is not statistically different from the true value, as judged using a 5% significance level.

(4) Neither of the algorithms performs satisfactorily for Case 3, which has a subspace mean difference aligning with a relatively large eigenvalue. Both algorithms suffer from a relatively low identification rate ($p_{ID}$ of .663 and .771, respectively, much smaller than that for the other five cases), which causes inaccurate estimation of the in-control distribution. But in a relative sense, ICA performs similarly to PCA—the average empirical $\alpha$ is 1.906 versus 1.864.

Based on this comparison, we recommend using the procedure combining ICA and change-point detection for the Phase I analysis of nonlinear profiles. This combined procedure does best when expected and it also has a robust performance across many different scenarios.

## Application of the Phase I Analysis to Forging Process Data

The data set contains 530 profile signals, each of which is sampled into a 224-dimension vector, i.e.,

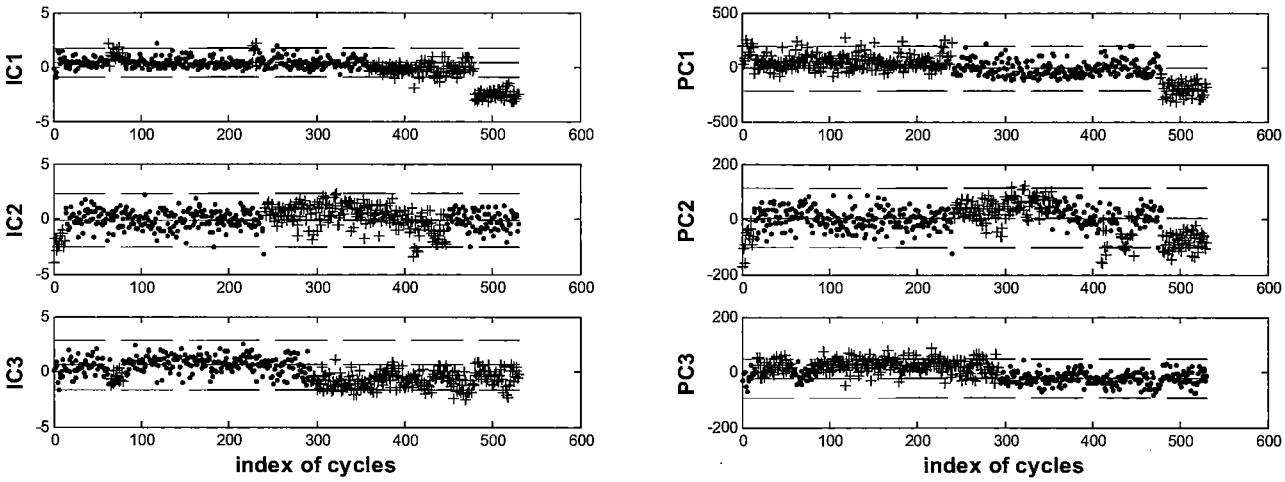


FIGURE 6. Phase I Analysis for the Profile Signals in the Forging Process; the Left Panel Is from the Procedure ICA+CPD, and the Right Panel Is from PCA+CPD.

$n = 530$ and $p = 224$. The same data set has been studied in Zhou and Jin (2004) for a different purpose. Interested readers may want to refer to Zhou and Jin (2004) for a more detailed background description of the forging process.

We suspect that this data set is contaminated by data from out-of-control conditions. We would like to see how the recommended procedure for Phase I analysis can help us find process change points rapidly.

The procedure combining ICA and change-point detection is executed the same as before. In this particular example, the *MDL* test retains up to 30 eigenvalues, which certainly makes the phase I analysis difficult. A scree plot indicates that the eigenvalues flat out from the fourth one onward. Hence, the first three eigenvalues and their eigenvectors are used. The UCL for Sullivan's algorithm is set in a way such that the false-detection probability for the whole procedure is 0.0027. The data separation is shown in Figure 6, where a "•" represents the in-control data point and a "+" indicates an out-of-control data point.

The left panel of Figure 6 is the result from ICA+CPD, indicating that there are five change points in the data set, partitioning the data set into six segments. The first suspected process change happens at sample #14, the second one at sample #65, the third one at sample #80, the fourth one at sample #241, and the fifth one at #480. It is reasonable that the first segment of roughly 13 samples is different from the rest of the process because each machine will usually have a transition period when it starts working. Careful reviews of the process history also confirm the authenticity of the fifth change point, where the last data segment of 50 samples has obviously undergone a mean shift from the previous data. It turns out that the coolant for the forging press was changed at that instance. There is no clear reason to explain the second to the fifth change points. After consulting the engineers, we believe that a collection of uncontrollable or hard-to-control factors, such as temperature and humidity, may have contributed to these changes. Since during an actual forging fabrication process, the process is considered to be in-control under such conditions, we will combine the second to the fifth data segments to form the in-control data set. That gives us a total of 467 observations for the in-control process condition, which should be used for setting up the monitoring scheme for future observations.

The right panel of Figure 6 is the result from PCA+CPD. Apparently, the data segments in the PCs are more than those in ICs, and the last segment of the data is not seen as clearly shifting away from the main stream in the PCs as it is seen in the ICs. Had the original shift been of a little smaller magnitude, using the first three PCs could have missed the important process change. Also, using the PCs, the majority of the data will be classified as out-of-control and the common ground from the in-control process condition is not so easy to be established.

In Figure 7, we plot the functional responses of

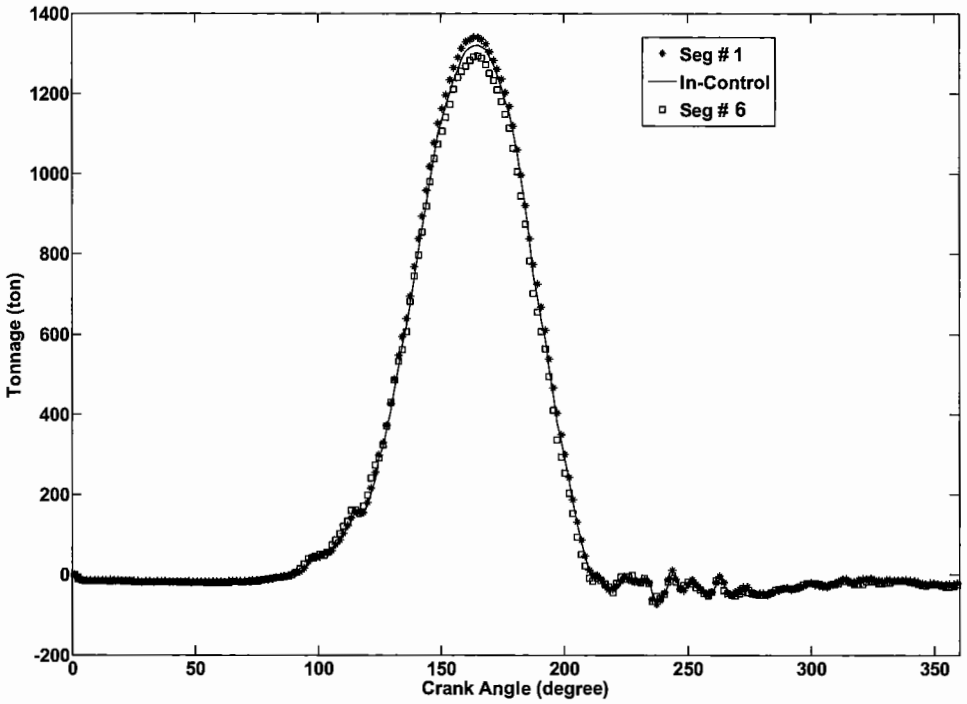


FIGURE 7. Functional Responses of the Segments in Figure 6.

three data segments detected in Figure 6: the first segment, the in-control data that combine the second to the fifth segment, and the sixth segment. The curves in Figure 7 are the average of the profiles within each one of the three segments. Apparently, the three segments demonstrate a noticeable deviation from each other, primarily around the peak area. This seems consistent with our understanding from Figure 6.

It is also worth noting that the recommended procedure for Phase I analysis can be recursively applied to the historical data, especially when there exist some outliers with large magnitudes that may overshadow other process changes. It would not be difficult to apply the recommended procedure to remove the outliers first and then to detect other process changes so as to extract the in-control data as accurately as possible.

It turns out that the change-point analysis splits the data sequence into subsequences, some of which apparently do not seem to be interestingly different (such as segments #2 to #5 in the left panel of Figure 6). This may be due to the fact that real process data rarely satisfy the model assumptions perfectly. The model of processes gives successive readings that, while all is well, are independently and identically distributed. This of course is seldom actually true, and real data sets always contain some apparently statistically real structure that we may not care to notice.

## Concluding Remarks

This paper investigates a strategy for performing Phase I analysis for high-dimensional nonlinear profiles. The presented Phase I analysis procedure consists of two major components: a data-reduction task, realized by the method of independent components analysis, and a data-separation task, realized by the change-point detection algorithm described in Sullivan (2002). Inclusion of the ICA plays a central role in the recommended procedure. The ICA finds the subspace in which the distinction of any existing structures in the data is maximized. This helps the latter algorithm to separate the in-control data from out-of-control observations. Moreover, the change-point detection algorithm can detect multiple change points effectively. Our study leads to the general conclusion that Sullivan's change-point detection algorithm is more effective than a clustering method for an ordered sequence. This conclusion seems to be applicable to broader application domains.

In this paper, we assume that the in-control data

represent the preponderance of the historical data set when developing the strategy for Phase I analysis. The question could be how one might know if there are too many out-of-control observations. From our experience, we believe the fraction of nonconforming of products (Montgomery, 2004) can be used as a rule of thumb to check if there are too many out-of-control cases. When the process is in control, the product it produces will often be conforming (not always true, though). Thus, a low fraction of nonconforming generally indicates a dominance of in-control data.

Many nonlinear profiles may possess a smoothness property, which motivates the research in the area of functional data analysis (Ramsey and Silverman 2002). For those nonlinear profiles, using a functional data analysis to preprocess the data and to find a more concise and robust representation could help the latter analysis, especially when the measurements themselves are noisy. Our current study is intended for a general category of nonlinear profiles that may or may not be inherently smooth. We have not explored how to combine the functional data analysis with the subsequent data reduction and clustering procedures. It is certainly a valuable aspect that is worth future research efforts.

One referee also points out that, to perform clustering on mixture data, a computationally much faster algorithm than the EMMIX algorithm we used is the hard allocation approach proposed in Fraley and Raftery (2002). We would like to share this valuable reference with the readers because it should definitely help in the cases of large-size data sets when computation becomes crucial.

## Acknowledgments

The authors gratefully acknowledge financial support from the NSF under grants DMI-0330356, CMS-0427878, and DMI-0348150. The authors also appreciate the editor and the referees for their valuable comments and suggestions.

## Appendix I
## Derivation of Equation (2)

Given the set of data as the combination of $\mathbf{X}_a$ and $\mathbf{X}_b$, the sample mean of the combined data is

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n_a} \mathbf{x}_i^a + \sum_{i=1}^{n_b} \mathbf{x}_i^b}{n_a + n_b} = \frac{n_a \bar{\mathbf{x}}_a + n_b \bar{\mathbf{x}}_b}{n},$$

where $\mathbf{x}_i^a$ and $\mathbf{x}_i^b$ are the values in samples $a$ and $b$, respectively. The sample covariance matrix of $\mathbf{x}$ is

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n-1}\left\{\sum_{i=1}^{n_a+n_b}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n_a}(\mathbf{x}_i^a - \bar{\mathbf{x}})(\mathbf{x}_i^a - \bar{\mathbf{x}})^{\mathrm{T}} \right. \\
&\qquad \left. + \sum_{i=1}^{n_b}(\mathbf{x}_i^b - \bar{\mathbf{x}})(\mathbf{x}_i^b - \bar{\mathbf{x}})^{\mathrm{T}}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n_a}\left(\mathbf{x}_i^a - \frac{n_a}{n}\bar{\mathbf{x}}_a - \frac{n_b}{n}\bar{\mathbf{x}}_b\right)\right. \\
&\qquad \times \left(\mathbf{x}_i^a - \frac{n_a}{n}\bar{\mathbf{x}}_a - \frac{n_b}{n}\bar{\mathbf{x}}_b\right)^{\mathrm{T}} \\
&\qquad + \sum_{i=1}^{n_b}\left(\mathbf{x}_i^b - \frac{n_a}{n}\bar{\mathbf{x}}_a - \frac{n_b}{n}\bar{\mathbf{x}}_b\right) \\
&\qquad \left. \times \left(\mathbf{x}_i^b - \frac{n_a}{n}\bar{\mathbf{x}}_a - \frac{n_b}{n}\bar{\mathbf{x}}_b\right)^{\mathrm{T}}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n_a}\left(\mathbf{x}_i^a - \bar{\mathbf{x}}_a + \frac{n_b}{n}(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)\right)\right. \\
&\qquad \times \left(\mathbf{x}_i^a - \bar{\mathbf{x}}_a + \frac{n_b}{n}(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)\right)^{\mathrm{T}} \\
&\qquad + \sum_{i=1}^{n_b}\left(\mathbf{x}_i^b - \bar{\mathbf{x}}_b + \frac{n_a}{n}(\bar{\mathbf{x}}_b - \bar{\mathbf{x}}_a)\right) \\
&\qquad \left. \times \left(\mathbf{x}_i^b - \bar{\mathbf{x}}_b + \frac{n_a}{n}(\bar{\mathbf{x}}_b - \bar{\mathbf{x}}_a)\right)^{\mathrm{T}}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n_a}(\mathbf{x}_i^a - \bar{\mathbf{x}}_a)(\mathbf{x}_i^a - \bar{\mathbf{x}}_a)^{\mathrm{T}}\right. \\
&\qquad + \sum_{i=1}^{n_a}\frac{n_b^2}{n^2}(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)^{\mathrm{T}} \\
&\qquad + \sum_{i=1}^{n_a}\frac{2n_b}{n}(\mathbf{x}_i^a - \bar{\mathbf{x}}_a)(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)^{\mathrm{T}} \\
&\qquad + \sum_{i=1}^{n_b}(\mathbf{x}_i^b - \bar{\mathbf{x}}_b)(\mathbf{x}_i^b - \bar{\mathbf{x}}_b)^{\mathrm{T}} \\
&\qquad + \sum_{i=1}^{n_b}\frac{n_a^2}{n^2}(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)^{\mathrm{T}} \\
&\qquad \left. + \sum_{i=1}^{n_b}\frac{2n_a}{n}(\mathbf{x}_i^b - \bar{\mathbf{x}}_b)(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)^{\mathrm{T}}\right\} \\
&= \frac{1}{n-1}\left\{(n_a-1)\mathbf{S}_a + (n_b-1)\mathbf{S}_b \right. \\
&\qquad \left. + \frac{n_a n_b}{n}(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)(\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_b)^{\mathrm{T}}\right\} \\
&= \frac{(n_a-1)}{n-1}\mathbf{S}_a + \frac{(n_b-1)}{n-1}\mathbf{S}_b + \frac{n_a n_b}{(n-1)n}\mathbf{v}\mathbf{v}^{\mathrm{T}}.
\end{aligned}$$

Because both $\mathbf{S}_a$ and $\mathbf{S}_b$ are diagonal matrices, the

eigenvalues of

$$\frac{(n_a-1)}{n-1}\mathbf{S}_a+\frac{(n_b-1)}{n-1}\mathbf{S}_b$$

are simply

$$\left\{\frac{(n_a-1)}{n-1}\lambda^a+\frac{(n_b-1)}{n-1}\lambda_i^b\right\}_{i=1}^{p}.$$

Also notice that $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ is a matrix of unit rank. According to Theorem 8.1.8 in Golub and Van Loan (1996, p. 397), the $i$th eigenvalue of $\mathbf{S}$ is

$$\lambda_i=\frac{(n_a-1)}{n-1}\lambda^a+\frac{(n_b-1)}{n-1}\lambda_i^b+m_i\frac{n_an_b}{(n-1)n}\|\mathbf{v}\|^2,$$

where $\|\cdot\|$ is a 2-norm and $m_i$ is a constant, satisfying that $0\le m_i\le 1$ and $\sum_{i=1}^{p}m_i=1$.

## Appendix II
## Sullivan's Algorithm for Change-Point Detection

Suppose there are $m$ independent observations, $x_1,x_2,\ldots,x_m$, from one or more univariate normal distributions. There are $R$ shifts in the mean, and the shift locations are $T_r$, $r=1,\ldots,R$, such that $0<T_1<\cdots<T_R<m$. Sullivan's algorithm consists of three steps:

**Step 1:** This step includes $m-1$ substeps. At each substep of $K=1,\ldots,m-1$, the observations are separated into $m-K+1$ clusters with $m-K$ boundaries indexed by $k$, $k=1,2,\ldots,m-K$. Associated with each boundary is a location $l_k$, the last observation in the cluster, and a distance $d_k$, which measures the dissimilarity of the means of its adjacent clusters by

$$d_k=\frac{|\bar{x}_k-\bar{x}_{k+1}|}{s\sqrt{\dfrac{m_k+m_{k+1}}{m_km_{k+1}}}}$$

where $m_k$ and $m_{k+1}$ are the numbers of observations in the adjacent clusters, $\bar{x}_k$ and $\bar{x}_{k+1}$ are the sample means, and $s$ is an estimate of the common standard deviation of all clusters. Without loss of generality, the value $s=1$ is used. Remove $k^*$ that corresponds to the smallest distance and save its location and distance as $l^*_{m-K}=l_{k^*}$ and $d^*_{m-K}=d_{k^*}$. Meanwhile, update the remaining distances for the next substep. Finally, we can get two $m-1$-dimensional sequences, $\{l_i^*\}$ and $\{d_i^*\}$.

**Step 2:** Calculate the robust estimator at the point where $0.2m$ boundaries remain by

$$s_r^2\frac{1}{m-K-1}\sum_{k=1}^{K+1}\sum_{i=1+T_{k-1}}^{T_k}(x_i-\bar{x}_k)^2,$$

where $K=0.2m$, rounded to the nearest integer, and the mean of cluster $k$ is

$$\bar{x}_k=\frac{1}{T_k-T_{k-1}}\sum_{i=1+T_{k-1}}^{T_k}x_i,\qquad 1\le k\le K+1.$$

Then calculate the updated distances

$$\{d_i^{*\prime}\}=\{d_i^*/s_r\}.$$

**Step 3:** Build the control chart of $\{d_i^{*\prime}\}$. If no out-of-control point exists, the process is determined to be in control. Otherwise, the index of the last out-of-control point, $n$, denotes the number of shifts or outliers and the first $n$ elements in $\{l_i^*\}$, correspondingly, indicate their locations. The upper control limit of the control chart can be estimated in advance with numerical simulations as follows. In the $j$th, $j=1,\ldots,M$, simulation: (1) generate a set of $m$ random observations following standard normal distribution. (2) Apply Steps 1 and 2 on that set and save $d_j=\max\{d_1^{*\prime},d_2^{*\prime}\}$. If the false-detection probability is set to be $\gamma$, then UCL equals the $100(1-\gamma)$th percentile of $\{d_j\}_{j=1}^{M}$.

## References


APLEY, D. W. and SHI, J. (2001). "A Factor-Analysis Method for Diagnosing Variability in Multivariate Manufacturing Processes". *Technometrics* 43, pp. 84–95.

BAKSHI, B. R. (1998). "Multiscale PCA with Application to Multivariate Statistical Process Monitoring". *AIChE Journal* 44(7), pp. 1596–1610.

BARNETT, K.; DUGGIRALA, R.; HITZ, D.; and SPIEWAK, S. (1998). "Monitoring and Control of Equipment and Process in Cold Extrusion". *SAE 1998 Transactions. Journal of Materials and Manufacturing* 107, pp. 991–1001.

CARREIRA-PERPINAN, M. (1997). "A Review of Dimension Reduction Techniques". Technical Report CS-96-09, Dept. of Computer Science, University of Sheffield.

FRALEY, C. and RAFTERY, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation". *Journal of the American Statistical Association* 97, pp. 611–631.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; and RUBIN, D. B. (2003). *Bayesian Data Analysis, 2nd edition.* Boca Raton, FL, Chapman and Hall/CRC.

GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations, 3rd edition.* Baltimore, MD, John Hopkins University Press.

HUBER, P. J. (1985) "Projection Pursuit". *Annals of Statistics* 13, pp. 435–475.

HYVARINEN, A.; HARHUNEN, J.; and OJA, E. (2001). *Independent Component Analysis.* Wiley, New York, NY.

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

JIN, J. and SHI, J. (1999). "Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelets". *Technometrics* 41, pp. 327–339.

JIN, J. and SHI, J. (2001). "Automatic Feature Extraction of Waveform Signals for In-Process Diagnostic Performance Improvement". *Journal of Intelligent Manufacturing* 12, pp. 267–268.

JOHNSON, R. A. and WICHERN, D. W. (2002). *Applied Multivariate Statistical Analysis, 5th edition.* Upper Saddle River, NJ, Prentice Hall.

JONES, M. C. and SIBSON, R. (1987). "What Is Projection Pursuit?" *Journal of the Royal Statistical Society, Series A* 150, pp. 1–36.

KNUSSMANN, K. D. and ROSE, C. (1993). "Signature-based Process Control (SbPC$^{TM}$)". Technical Report, Signature Technologies, Inc., Oak Point, Texas.

KOH, C. K. H.; SHI, J.; WILLIAMS, W.; and NI, J. (1999a). "Multiple Fault Detection and Isolation Using the Haar Transform—Part 1: Application to the Stamping Process". *ASME Transactions, Journal of Manufacturing Science and Engineering* 121, pp. 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: Theory". *ASME Transactions, Journal of Manufacturing Science and Engineering* 121, pp. 295–299.

KOSANOVICH, K. A. and PIOVOSO, M. J. (1997). "PCA of Wavelet Transformed Process Data for Monitoring". *Intelligent Data Analysis* 1, pp. 85–99.

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, pp. 79–90.

MAHMOUD, M. A. and WOODALL, W. H. (2004). "Phase I Analysis of Linear Profiles with Calibration Applications". *Technometrics* 46, pp. 380–391.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models.* New York, NY, John Wiley & Sons.

MONTGOMERY, D. C. (2004). *Introduction to Statistical Quality Control, 5th edition.* Wiley, New York, NY.

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, pp. 83–88.

RAMSEY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* New York, NY, Springer-Verlag.

RUNGER, G. C. (1996). "Projections and the $U^2$ Multivariate Control Chart". *Journal of Quality Technology* 28, pp. 313–319.

RUNGER, G. C.; BARTON, R. R.; CASTILLO, E. D.; and WOODALL, W. H. (2005). "Optimal Monitoring of Multivariate Data for Fault Patterns". *Journal Quality Technology*, to appear.

SULLIVAN, J. H. (2002). "Detection of Multiple Change Points from Clustering Individual Observations". *Journal of Quality Technology* 34, pp. 371–383.

WILKINSON, J. H. (1965). *The Algebraic Eigenvalue Problem.* Oxford, England, Clarendon Press.

ZHOU, S. and JIN, J. (2005) "Automatic Feature Selection for Unsupervised Clustering of Cycle-Based Signals in Manufacturing Processes". *IIE Transactions* 37, pp. 569–584.

ZHOU, S.; SUN, B.; and SHI, J. (2004). "An SPC Monitoring System for Cycle-Based Waveform Signals Using Haar Wavelet Transform". *IEEE Transactions on Automation Science and Engineering*, accepted.