Thresholded Multivariate Principal Component Analysis for Phase I Multichannel Profile Monitoring

Yuan Wang, Yajun Mei, Kamran Paynabar *
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

August 15, 2017

Abstract

Monitoring multichannel profiles has important applications in manufacturing systems improvement, but it is non-trivial to develop efficient statistical methods due to the facts that profiles are high-dimensional functional data with intrinsic inner-and inter-channel correlations, and that the change might only affect a few unknown features of multichannel profiles. To tackle these challenges, we propose a novel thresholded multivariate principal component analysis (PCA) method for multichannel profile monitoring. Our proposed method consists of two steps of dimension reduction: It first applies the functional PCA to extract a reasonable large number of features under the in-control state, and then uses the soft-thresholding techniques to further select significant features capturing profile information under the out-of-control state. The choice of tuning parameter for soft-thresholding is provided based on asymptotic analysis, and extensive numerical studies are conducted to illustrate the efficacy of our proposed thresholded PCA methodology.

Keywords: Change-point, Multichannel Profiles, Principal Component Analysis, Shrinkage Estimation, Statistical Process Control.

*Y. Wang is PhD student (Email: sxtywangyuan@gmail.com), Y. Mei is Associate Professor (Email: ymei@isye.gatech.edu), K. Paynabar is Assistant Professor (Email: kamran.paynabar@isye.gatech.edu), H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 756 Ferst Drive NW, Atlanta, GA 30332, USA. This is part of Y. Wang’s PhD dissertation.
1 Introduction

Profile monitoring plays an important role in manufacturing systems improvement (Noorossana et al. (2011), Qiu (2013)), and a typical setup is to monitor a sequence of profiles (e.g. curves or functions) over time to check whether the underlying functional structure of the profiles changes or not. Extensive research has been done for monitoring univariate profiles or real-valued functions in the area of statistical process control (SPC) in the past decades, and standard approaches are to reduce the univariate profiles in the infinite-dimensional or high-dimensional functional spaces to a low-dimensional set of features, e.g., shape, magnitude, frequency, regression coefficients, etc. See, for instance, work by Jin and Shi (2000), Ding et al. (2006), Jeong et al. (2006), Jensen et al. (2008), Berkes et al. (2009), Chicken et al. (2009), Qiu et al. (2010), Abdel-Salam et al. (2013). Meanwhile, nowadays manufacturing systems are often equipped with multiple sensors capable of collecting data simultaneously, and thus one often faces the problem of monitoring multichannel or multivariate profiles. While various methods have been developed for univariate profile monitoring, they often cannot easily be extended to multichannel profiles, and research on monitoring multivariate/multichannel nonlinear profiles is very limited. For some exceptions, see Jeong et al. (2007), Paynabar et al. (2013), Grasso et al. (2014), and Paynabar et al. (2016).

The primary goal of this paper is to present a Phase I method for multichannel profile data where multiple sensors measure the same/similar process variables or quality characteristics and where historical data consists of both in-control (IC) and out-of-control (OOC) observations. Similar to other Phase I methods, our proposed approach not only is capable of identifying and removing OOC observations, but also can estimate the baseline model parameters namely, mean and contrivance functions based on IC data that will be further used for Phase II (online) monitoring. A concrete motivating example of this paper is from a forging process, shown in Figure 1 and 2, in which multichannel load profiles measure exerted forces in each column of the forging machine. The data set includes 151 normal profiles and 69 default profiles, and each profile or observation is a four-dimensional vector function or four curves/channels/components. This inspires us to investigate Phase I data that are corrupted with OOC observations at a single unknown change point, and focus on how to better detect or estimate the change-point. It should be noted that although the Phase I
Figure 1: *Left:* A forging machine with 4 tonnage sensors. *Right:* A single run sample of four-dimensional functional data.

Figure 2: *Left:* Shape of workpieces at each operation. *Right:* Tonnage profile for normal and missing operations.

data may include both IC and OOC observations, their label is unknown (i.e. unsupervised setting). Our proposed methodology integrates multivariate functional Principal Component Analysis (PCA), shrinkage and change-point detection approaches for analyzing Phase I data corrupted with OCC observations. Here PCA and shrinkage are used for dimensionality reduction and extracting monitoring features with largest sensitivity to the process changes reflected in OOC data, whereas the change-point detection method combines these features to determine the time of the change.

On the one hand, our proposed methodology is closely related to several existing PCA-based SPC methods that successfully applied the functional PCA to SPC to deal with intrinsic inner- and inter-channel correlations of profiles, see Paynabar et al. (2013), Grasso et al. (2014), and Paynabar et al. (2016). On the other hand, our research is significantly different from these existing methods from the dimension reduction viewpoint. The existing PCA-based SPC methods often follow the standard PCA literature to select a few principal
components (PCs) that contain a large amount of variation or information under the IC state. While this is reasonable in the estimation or curve fitting/smoothing context under the IC state, it may not be effective in the SPC and change-point detection context, especially when the number of initially retained PCs are pretty large, since it does not take into account any information under the OOC state. To illustrate the important impact of the OOC state on PC selection, we provide a simple two-dimensional example in Figure 3. As can be seen in this figure, in a given Phase I data, the mean of the IC distribution is shifted along the second PC and thus, the second PC with a smaller variance is a better feature for monitoring. However, if the variation contribution rule is used, the first PC is selected and consequently the change becomes undetectable. To address the issue, our proposed methodology does not completely depend on the IC state to conduct an aggressive dimension reduction, and it allows some flexibility to capture the information about the direction of the change in OOC observations. This will help identify optimal PCs that can preserve the change in the PC domain, hence improve the control chart performance.

We should acknowledge that the importance of shrinkage is well-known in modern statistics for dimension reduction and feature selection of high-dimensional data, including the profile monitoring literature. Jeong et al. (2006) incorporated the hard thresholding into the Hotelling $T^2$ statistics in the context of online monitoring of single profiles, and Jeong et al.
(2007) proposed a hard thresholding method to obtain projection information by optimizing “overall relative reconstruction error”. Zou et al. (2012) applied LASSO shrinkage in linear model coefficients for online monitoring of linear profiles. However, these existing methods conduct one-shot dimension reduction for single-channel profiles, and it is unclear how to extend them directly in the complicated context of monitoring multi-channel profiles. Here we propose a novel way to apply shrinkage/thresholding to multi-channel profile monitoring: our proposed methodology splits the dimension reduction process into two steps using two different methods: PCA to extract and select important PC features, and soft-thresholding to select PCs that are significantly affected by the change.

The remainder of this paper is organized as follows. In Section 2, we state the mathematical formulation of multichannel profile monitoring. In Section 3, we propose our thresholded PCA method, and provide a guideline on how to select the corresponding tuning parameters. In Sections 4 and 5, we use the real forging process data and simulations to illustrate the usefulness of our proposed thresholded PCA method. Concluding remarks and future research directions are presented in Section 6.

2 Problem Formulation and Background

Suppose that a random sample of $m$ multichannel profiles, each with $p$ channels, is collected from a production process. Mathematically, each of the $m$ multichannel profile observations is a $p$-dimensional curve denoted by $X_i(t) = (X_i^{(1)}(t), ..., X_i^{(p)}(t))^T$ over the domain $t \in [0, 1]$, for $i = 1, \cdots, m$, where the superscript $T$ denotes the transpose. We assume that the process is initially IC and the $X_i(t)$’s have mean $\mu_1(t)$. At some unknown time, the process may become OOC in the sense that the mean of the profiles $X_i(t)$’s changes to $\mu_2(t)$. Specifically, we assume that the data are from the change-point additive noise model:

$$X_i(t) = \left\{ \begin{array}{ll}
\mu_1(t) + Y_i(t), & \text{when } i = 1, ..., \tau, \\
\mu_2(t) + Y_i(t), & \text{when } i = \tau + 1, ..., m,
\end{array} \right. \quad \text{for } 0 \leq t \leq 1,$$

(1)

for some unknown change-point $\tau; 1 \leq \tau < m$. For each $i = 1, \cdots, m$, the noise curve $Y_i(t) = (Y_i^{(1)}(t), \cdots, Y_i^{(p)}(t))^T$ is a $p$-dimensional vector function with mean $0$. We assume that the $m$ noise curves $Y_i(t)$’s are independent and identically distributed (i.i.d.) across different $i$’s, e.g., $Y_i(t)$ is independent of $Y_j(t')$ for all $i \neq j$ and for all $t$ and $t'$. On the other
hand, for any given $i = 1, \cdots, m$, there are spatial/temporal dependence within the noise curves $Y_i(t)$ across different components and across different $t$, e.g., the $p$ channels of $Y_i(t)$ are cross-correlated, and within each channel, $Y_i(t)$ is correlated with $Y_i(t')$ for $t \neq t'$.

In this article, we investigate the problem of Phase I profile monitoring with OOC data, in which all $m$ profiles $X_i(t)$’s are observed, and $\mu_1(t)$ and $\mu_2(t)$ are two unknown $p$-dimensional mean functions. The task is to utilize the observed $X_i(t)$’s to test the null hypothesis $H_0: \mu_1(t) = \mu_2(t)$ (i.e., no change) against the alternative hypothesis $H_a: \mu_1(t) \neq \mu_2(t)$ (i.e., a change occurs at some unknown time $1 \leq \tau < m$). In addition, we also impose the classical Type I error probability constraint:

$$P_{H_0}(\text{reject } H_0: \mu_1(t) = \mu_2(t)) \leq \alpha, \quad (2)$$

for some pre-specified constant $\alpha$, e.g., $\alpha = 5\%$.

To test the null hypothesis $H_0: \mu_1(t) = \mu_2(t)$ under the change-point additive noise model (1), it is intuitive to consider the two-sample test statistic for testing that the means of the first $\ell$ observations and the last $m - \ell$ are equal for each potential change-point $\ell = 1, 2, \ldots, m - 1$. This suggests us to define

$$\Delta_\ell(t) = \sqrt{\frac{\ell(m-\ell)}{m}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} X_i(t) - \frac{1}{m-\ell} \sum_{i=\ell+1}^{m} X_i(t) \right\}, \quad (3)$$

where the term $\sqrt{\ell(m-\ell)/m}$ standardizes the variance of profile difference. Then, one can utilize $\Delta_\ell(t)$ in (3) to search over $\ell$ for the most plausible place of the potential change-point.

When the original data are real-valued Gaussian random variables, the change-point additive noise model (1) was studied in James et al. (1987), which showed that the statistic $\Delta_\ell$ in (3) is actually the generalized log-likelihood ratio statistic when the alternative hypothesis is $\tau = \ell$. A challenge is that the problem can be ill-posed if the true change-point $\tau$ occurs near the boundary of $[1, m)$, which is often referred to as the edge effect. Fortunately, this can easily be circumvented by an additional assumption that $\rho_1 m \leq \tau \leq \rho_2 m$ for two constants $0 < \rho_1 < \rho_2 < 1$. Indeed James et al. (1987) proposed to reject the null hypothesis if $\max_{\rho_1 m \leq \ell \leq \rho_2 m} \Delta_\ell$ is large, and then investigated the theoretical properties of the corresponding Type I and Type II error probabilities. Moreover, the change-point additive noise model (1) has also been studied in other contexts such as a parameter change of exponential
distribution (Haccou and Meelis (1987)) or binomial distribution (Horvath (1989)), or the slope change of simple linear regression (Kim and Siegmund (1989)).

Our context of Phase I multichannel profile monitoring with corrupted data motivated us to investigate the change-point additive noise model (1) under the new context when the data are \( p \)-dimensional vector functions. Hence, we face a much more complicated problem of testing whether the \( p \)-dimensional vector function \( \Delta_\ell(t) \) in (3) is a zero vector function or not. Our main contributions are (i) to propose an efficient method to test whether or not \( \Delta_\ell(t) \) is a zero function by combining shrinkage with PCA, and (ii) provide the suggestions on the choices of tuning parameters through the investigation of the corresponding Type I and/or Type II error probability properties. To highlight our main ideas, we will not discuss the edge effect explicitly here: we develop our proposed statistical method under the assumption when the change-point \( 1 \leq \tau < m \) is unknown, but will avoid the edge effect by investigating its error probability properties when both \( \tau \) and \( m - \tau \) are moderately large, e.g., \( \tau = \rho m \) for some \( 0 < \rho < 1 \).

3 Our Proposed Thresholded PCA Methodology

In this section, we develop a thresholded multivariate functional PCA methodology for Phase I monitoring of multichannel profiles with OOC data. Recall that we re-formulate Phase I multichannel profile monitoring as the problem of testing whether or not the \( p \)-dimensional vector function \( \Delta_\ell(t) \) in (3) is a zero vector function for some candidate change-point \( \ell = 1, 2, \ldots, m - 1 \). In general, the functional hypothesis testing problem is highly non-trivial, and one useful approach is to utilize the PCA decomposition to project the \( \Delta_\ell(t) \)'s to the PCs, thereby reducing the functional hypothesis testing problem to the problem of testing whether all those \( p \)-dimensional coefficient variables are zero or not. For instance, this approach is adopted by Paynabar et al. (2016) which considers the scenario where all PC coefficients simultaneously become non-zero under the alternative change-point hypothesis. Here we consider a different scenario where only a few unknown PC coefficients become non-zero under the alternative hypothesis, especially when the number of PCs are large.

For the purpose of easy understanding, this section is divided into three subsections. In Subsection 3.1, we review the multivariate functional PCA method that estimates the basis
and the covariance matrices. This allows us to reduce the hypothesis testing problem from the functional space to the space of the PCA coefficients. In Subsection 3.2, our Phase I profile monitoring method is developed as a hypothesis testing of PCA coefficients augmented by soft-thresholding technique that has a semi-Bayesian interpretation of the generalized likelihood ratio test. In Subsection 3.3, based on asymptotic analysis, we provide a guidance on the choice of tuning parameters in our proposed methodology.

3.1 PCA Basis and Covariance Estimation

To have a better understanding of the basis and covariance matrix estimation of PCA under the change-point model in (1), we first consider the unrealistic case when the noise $p$-dimensional curves $Y_i(t)$’s in (5) were observable.

In our motivating application of multichannel profiles described in Section 1 (see Figure 1), all sensors are measuring the same variables in different locations and thus all components of multichannel profiles exhibit similar pattern. Hence we assume that all components of the $p$-dimensional functions $Y_i(t)$’s have similar covariance structures over time, and can be decomposed into the same real-valued basis functions $v_k(t)$’s. Under this assumption, it is sufficient to investigate the inner-channel covariance structure over time, and since the $p$-dimensional functions $Y_i(t)$’s have mean 0, we propose to evaluate it by the following total covariance function:

$$c(t, s) = E[Y_i(t)^T Y_i(s)] = \sum_{j=1}^{p} E[(Y_{i}^{(j)}(t))Y_{i}^{(j)}(s)] \quad \text{for } 0 \leq t, s \leq 1.$$  \hspace{1cm} (4)

Then the basis $v_k(t)$’s can be computed as the eigenfunctions of $c(t, s)$ as in the standard PCA. Note that for any given $j$-th channel or the curve $Y_i^{(j)}(t)$, the quantity $E[(Y_{i}^{(j)}(t))Y_{i}^{(j)}(s)]$ provides the corresponding channel-specific covariance function over time, and $c(t, s)$ in (4) essentially pools all components together to provide a single, overall estimate of the covariance function between different time steps $t$ and $s$. This approach is very efficient when different channels indeed have similar covariance structures, but the price we pay is that it might not be robust in other scenarios. Also this is often called separable PCA, which is different from the standard multivariate FPCA (see Section 8.5 in Ramsay and Silverman (2005)).
With the covariance function $c(t, s)$ in (4) and the corresponding first $d$ PCA eigenfunction bases $v_k(t)$’s, the projection of the $p$-dimensional noise function $Y_i(t)$ on the $k$-th PC can be presented as the $p$-dimensional random vector

$$\xi_{ik} = \langle Y_i(t), v_k(t) \rangle = \int_0^1 Y_i(t)v_k(t)dt,$$

for $i = 1, \ldots, m,$ and $k = 1, \ldots, d$. We assume that the coefficient vectors $\xi_{ik}$’s are independent multivariate normal distributed over different bases $k$’s and different observations $i$’s, and have different $p \times p$ covariance matrix $\Sigma_k$ for different PCs:

$$\Sigma_k = \mathbb{E}(\xi_{ik}\xi_{ik}^T) = \mathbb{E}\{\int_0^1 Y_i(t)v_k(t)dt \int_0^1 Y_i(t)^Tv_k(t)dt\},$$

(6)

Next, we discuss how to conduct the above PCA-related analysis based upon the observable profiles $X_i(t)$ under the change-point additive noise model in (1) when the noise curves $Y_i(t)$’s are unobservable. Note that the differences $Y_{i+1}(t) - Y_i(t) = X_{i+1}(t) - X_i(t)$ are observable for all $1 \leq i \leq m-1$ except $i = \tau$ if the change-point $\tau$ exists. Thus the covariance function $c(t, s)$ in (4) can be estimated by $\bar{Y}_{i+1}(t) - Y_i(t)$, which yields the approximation:

$$\hat{c}(t, s) = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (X_{i+1}(t) - X_i(t))^T (X_{i+1}(s) - X_i(s)).$$

(7)

Here the denominator is $2(m-1)$ since we estimate $c(t, s)$ in (4) from $Y_{i+1}(t) - Y_i(t)$ and the $Y_i(t)$’s are i.i.d. over $i = 1, \ldots, m$.

Next, the estimates of basis functions $\hat{v}_k(t)$’s can be found as the eigenfunctions of $\hat{c}(t, s)$ in (7). As for the estimation of the covariance matrix $\hat{\Sigma}_k$ in (6) of coefficients $\xi_{ik}$, we again take advantage of the differences $Y_{i+1}(t) - Y_i(t)$ under the change-point additive noise model in (1), and approximate it by

$$\hat{\Sigma}_k = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \int_0^1 \{X_{i+1}(t) - X_i(t)\} \hat{v}_k(t)dt \int_0^1 \{X_{i+1}(t) - X_i(t)\}^T \hat{v}_k(t)dt.$$

(8)

Note that the estimated function $\hat{c}(t, s)$ in (7) and $\hat{\Sigma}_k$ in (8) is consistent under the reasonable regularity assumption of the alternative hypothesis, see Remark #2 in Paynabar et al. (2016).

3.2 Our Proposed Methodology

Recall that under our context of Phase I multichannel profile monitoring with corrupted data, we need to test whether the $p$-dimensional vector function $\Delta_r(t)$ in (3) is a zero vector
Given the estimated first $d$ real-valued PCs $\hat{v}_k(t)$'s, the $p$-dimensional vector function $\Delta(t)$ can be projected to these $d$ PCs, and denote the corresponding $p$-dimensional coefficients by

$$\eta_{tk} = \int_0^1 \Delta(t) \hat{v}_k(t) dt,$$

for $k = 1, 2, \ldots, d$. Note that for each given $t$, the $\eta_{tk}$'s are uncorrelated over different $k$'s. Then, we would like to test whether or not all $p$-dimensional coefficients $\eta_{tk}$ are zero vectors simultaneously for all candidate change-points $\ell = 1, \ldots, m - 1$ and for all possible PCs $k = 1, \ldots, d$.

A key idea is to consider the real-valued statistic

$$U_{\ell,k} = \eta_{tk}^T \hat{\Sigma}_k^{-1} \eta_{tk},$$

where $\hat{\Sigma}_k$ is defined in (8). Note that the statistics $U_{\ell,k}$'s in (10) are motivated from the scenario when the basis $v_k(t)$ and $\Sigma_k$ are known: if the estimates $\hat{v}_k(t)$ and $\hat{\Sigma}_k$ are replaced by their true values, it is straightforward from (1) to show that $\eta_{tk} \sim N(0, \Sigma_k)$ under the null hypothesis $H_0: \mu_1(t) = \mu_2(t)$ but $\eta_{tk} \sim N(\int_0^1 \Delta(t) \nu_k(t) dt, \Sigma_k)$ under the alternative hypothesis $H_a: \mu_1(t) \neq \mu_2(t)$. Hence, when the basis $\nu_k(t)$ and $\Sigma_k$ are known, the $U_{\ell,k}$'s in (10) are $\chi^2_p$-distributed under $H_0$, but should be stochastically larger than $\chi^2_p$ under $H_a$. When the estimates $\hat{v}_k(t)$ and $\hat{\Sigma}_k$ are used, we expect that similar conclusions also hold approximately, e.g., whether the value of $U_{\ell,k}$ in (10) is large or small indicates whether or not there is a change at time step $\ell$ along the PC $\hat{v}_k(t)$.

In Paynabar et al. (2016), the authors essentially assume that the change affects all PCs simultaneously, i.e., all $U_{\ell,k}$'s in (10) are non-zeros simultaneously for all $k = 1, \ldots, d$ when the true change-point $\tau = \ell$. Then the authors propose to rejects the null hypothesis $H_0: \mu_1(t) = \mu_2(t)$ if

$$\max_{1 \leq \ell < m} \sum_{k=1}^d U_{\ell,k} > L$$

for some pre-specified constant $L$ that is chosen to satisfy the Type I error constraint in (2).

In this paper, we consider a different scenario when the change only significantly affects some unknown PCs, also see Figure 3 in the Introduction for an illustrative example. In addition, this scenario might be more reasonable in practice when the number of the selected
PCs, $d$, is large. To develop efficient methods in this scenario, for a given candidate change-point $\ell$, we assume that all PCs are independent, and each has a prior probability $\pi$ getting affected by the changing event. In other words, for $k = 1, \ldots, d$, let $Z_k$ be the indicator whether the $k$-th PC is affected by the change in the OOC state or not, and then the changing indicators $Z_1, \ldots, Z_d$ are assumed to be i.i.d. with probability mass function $P(Z_k = 1) = \pi = 1 - P(Z_k = 0)$. When $Z_k = 1$, the $k$-th PC is affected, and $U_{\ell,k}$ in (10) represents the evidence of possible change in the log-likelihood-ratio scale. Treating $Z_k$’s as the hidden states, the joint log-likelihood ratio statistic of $Z_k$’s and $U_{\ell,k}$ when testing $H_0 : Z_1 = \ldots = Z_d = 0$ (no change) can be as

$$LLR(n) = \sum_{k=1}^{d} \{Z_k(\log \pi + U_{\ell,k}) + (1 - Z_k) \log(1 - \pi)\} - \sum_{k=1}^{d} \log(1 - \pi)$$

$$= \sum_{k=1}^{d} Z_k\{U_{\ell,k} - \log((1 - \pi)/\pi)\}.$$

Since the $Z_k$’s are unobservable, it is natural to maximize $LLR(n)$ over $Z_1, \ldots, Z_d \in \{0, 1\}$. Clearly $\max_{Z=0,1}(Zu) = \max(0, u)$ for any real-valued $u$. Hence, the generalized log-likelihood ratio becomes $\sum_{k=1}^{d} \max\{U_{\ell,k} - \log((1 - \pi)/\pi), 0\}$, which would be our proposed test statistics if we let $c = \log((1 - \pi)/\pi)$ be the tuning parameter.

Therefore, our proposed thresholded PCA methodology is defined as follows. We first define a test statistic

$$Q_m = \max_{1 \leq \ell < m} \sum_{k=1}^{d} \max(U_{\ell,k} - c, 0),$$

for some pre-specified “soft-thresholding” parameter $c \geq 0$. Then, we reject the null hypothesis $H_0 : \mu_1(t) = \mu_2(t)$ if and only if

$$Q_m > L$$

for some pre-specified $L$. Moreover, when relation (13) occurs, we can also estimate the change point by

$$\hat{\ell} = \arg \max_{1 \leq \ell < m} \sum_{k=1}^{d} (U_{\ell,k} - c)^+.$$  

The choices of the tuning parameters will be discussed in more details in the next subsection.
It is informative to compare our proposed thresholded PCA method in (12)-(14) with the existing method in (11) proposed by Paynabar et al. (2016). On one hand, from the pure mathematical viewpoint, the method in Paynabar et al. (2016) is a special case of our proposed method when the soft-thresholding parameter \( c = 0 \). On the other hand, from the statistical viewpoint, these two methods are designed for different scenarios: Paynabar et al. (2016) assumes that all PCs are affected by the change simultaneously, whereas we assume that only some PCs are significantly affected, especially when the number \( d \) of PCs is large. Hence, under our scenario, one has to smooth out those PCs that are not affected by the change, since they do not provide information about the change. To do so, we propose to develop our method by using the sum of the soft-thresholding transformation of the \( U_{\ell,k} \)'s in (10). While the use of thresholding is well-known in modern statistics for dimension reduction and feature selection of high-dimensional data, our application to PCA and multi-channel profile monitoring is new.

### 3.3 The Choices of Tuning Parameters

In our proposed thresholded PCA methodology in (12) and (13), there are three tuning parameters: (i) the number \( d \) of PCs in (12), (ii) the soft-thresholding parameter \( c \) in (12), and (iii) the control limit \( L \) in (13). Practically, one needs to determine \( d \) and \( c \) first before selecting the control limit \( L \), but below we will present the choice of \( L \) first for any given \( d \) and \( c \) since it is easier to understand from the statistical viewpoint.

In order to find the control limit \( L \) for our proposed methodology to satisfy the Type I error probability constraint in (2), assume, for now, that the number \( d \) of PCs and the soft-thresholding parameter \( c \) in the test statistic \( Q_m \) in (12) are given. Then the constraint in (2) becomes \( \Pr_{H_0}(Q_m > L) \leq \alpha \). Hence, the control limit \( L \) should be the upper \( \alpha \) quantile of the distribution of \( Q_m \) in (12) under \( H_0 \), and thus, it is sufficient to approximate or simulate the null distribution of \( Q_m \) under \( H_0 \).

There are a couple of numerical ways to do so by generating a large number of Monte Carlo simulations of \( Q_m \) under \( H_0 \). The first one is when there exists “retrospective profiles” dataset that are collected from an IC process performing under normal operating conditions. Then, in each Monte Carlo run, we can randomly select \( m \) profiles and compute the corresponding values of \( Q_m \). Alternatively, when “retrospective profiles” are not available, as suggested in
Paynabar et al. (2016), one can use the fact that $Q_m$ under $H_0$ has the same distribution as

$$G_m = \max_{1 \leq i < m} \sum_{1 \leq k \leq d} \left( \frac{(m-i)i}{m} \right) (\mathbf{z}_{k,1,i} - \mathbf{z}_{k,i+1,m})^T \mathbf{S}_{zk}^{-1} (\mathbf{z}_{k,1,i} - \mathbf{z}_{k,i+1,m}) - c$$

(15)

where $\{\mathbf{z}_{k,i}\}$ is a set of independent standard normal multivariate observations of dimension $p$, $\mathbf{z}_{k,\ell_1,\ell_2} = (\ell_2 - \ell_1 + 1)^{-1} \sum_{i=\ell_1}^{\ell_2} \mathbf{z}_{k,i}$, and $\mathbf{S}_{zk} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\mathbf{z}_{k,i+1} - \mathbf{z}_{k,i}) (\mathbf{z}_{k,i+1} - \mathbf{z}_{k,i})^T$. Then the control limit $L$ is chosen as the upper $\alpha$ quantile of simulated statistics $G_m$’s.

Next, let us discuss the choice of the number $d$ of PCs. In the PCA, a standard approach is to choose a small $d$ based on high percentage, say 95% or more, of total variation explained by the extracted PC-scores under the IC state. In some applications, the corresponding $d$ value indeed turns out to be small, e.g., $d = 4$ or 8 in the simulation studies of Paynabar et al. (2016). However, for the real data application of multichannel profiles in Section 1 (see Figure 1), the number of PCs required to explain 85%, 90%, 95%, 99% of total variation under the IC state would be 15, 21, 33, and 64, respectively. In view of the need of a small $d$ value in the standard PCA method, Paynabar et al. (2016) followed their intuition to choose $d = 15$ that explains 85% variation for the real data set, but their approach of lowering explained percentage is ad hoc.

Here we propose a different approach: we keep the standard choice of $d$ based on high explained percentage, but propose a new approach to circumvent the challenge when the choice of $d$ is large. Our rationale is to view the PCA as a dimension reduction tool that reduces raw data from the infinitely functional (or super-high-dimensional) space to an intermediate space of $R^d$ under the IC state. If $d$ is large and if a few PCs might be affected by the change, we can then conduct another dimension reduction to smooth out those PCs that do not provide information of the change under the OOC state. This new dimension reduction is based on the sum of the soft-thresholding transformation in (12) of the real-valued statistics $U_{\ell,k}$’s in (10).

Finally, we focus on the choice of soft-thresholding parameter $c$ in (12). Note that the baseline choice of $c$ is $c_0 = 0$, which yields the approach of Paynabar et al. (2016) for the scenario when the total number $d$ of PCs is small. Also recall that our proposed method is motivated from semi-Bayesian arguments with $c = \log((1 - \pi)/\pi)$ when each PC is affected by the change with probability $\pi$. Unfortunately, the prior probability $\pi$ is often unknown in practice. Below, we will discuss two alternative choices of the soft-thresholding parameter
c > 0 under the frequentist setting: one is only for the theoretical interest due to the optimality properties, and the other is our recommended choice in practice due to its simple but useful properties.

Intuitively, when the number $d$ of PCs is large, the parameter $c > 0$ in (12) should be chosen to be large enough to filter out those non-changing PC basis $\hat{v}_k(t)$’s, but cannot be too large to remove the PCs that are affected by the change. Hence, a suitable choice of $c$ will depend on the specific alternative hypothesis $H_a$ and its effects on the basis projections. Here we propose two different choices of the soft-thresholding parameter $c > 0$: one is based on the Central Limit Theorem (CLT) to derive the optimal choice of $c$ in (18) below under a specific $H_a$; and the other is based on the extreme theory that produces a simpler but practically useful choice of $c$ in (19). For that purpose, by (12), we have

$$P\left(\sum_{k=1}^{d}(U_{\ell,k} - c)^+ > L\right) \leq P\left(Q_m > L\right) \leq \sum_{\ell=1}^{m-1} P\left(\sum_{k=1}^{d}(U_{\ell,k} - c)^+ > L\right),$$  \hspace{1cm} (16)

which becomes $(m-1)P(\sum_{k=1}^{d}(U_{\ell,k} - c)^+ > L)$, as the data are i.i.d. over $\ell = 1, \cdots, m-1$. Hence, from the asymptotic viewpoint, $P(Q_m > L)$ and $P(\sum_{k=1}^{d}(U_{\ell,k} - c)^+ > L)$ go to 0 at the same rate when $m$ is fixed. In particular, when the Type I error constraint $\alpha$ goes to 0, the main probability of interest is to estimate

$$P_{H_0}(\sum_{k=1}^{d}(U_{\ell,k} - c)^+ > L_c),$$  \hspace{1cm} (17)

where $L_c$ is chosen so that this probability $\leq \alpha$.

Our proposed choices of $c$ correspond to two different methods to approximate the distribution of $\sum_{k=1}^{d}(U_{\ell,k} - c)^+$ under $H_0$ : one is CLT when $c$ is small, and the other is the extreme theorem when $c$ is large. Since these two methods yield different results on $c$, we present them separately in Proposition 1, which assumes that $\chi_p^2$ approximation applies to $U_{\ell,k}$’s.

**Proposition 1.** Assume that $U_{\ell,k} \sim \chi_p^2$ under $H_0$, for all $k = 1, \ldots, d$;

(a) *(The CLT approximation when $c$ is small).* Assume further that under $H_a$, exactly $d_0$ out of $d$ PCs are affected in the sense that $U_{\ell,k} \sim \chi_p^2(\delta^2p) = \epsilon_{tk}^T\epsilon_{tk}$ with $\epsilon_{tk} \sim N(\delta, I_p)$ for $k = 1, \ldots, d_0$, and $U_{\ell,k} \sim \chi_p^2$ for $k = d_0 + 1, \ldots, d$. Then when both $d_0$ and $d - d_0$ are
large, an approximate of the optimal choice of $c$ is

$$c_1 = \arg \min_{c \geq 0} \left\{ -\frac{(\mu_c^{(1)} - \mu_c)d_0}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d - d_0)(\sigma_c)^2}} + \frac{\sqrt{d\sigma_c}}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d - d_0)(\sigma_c)^2}} \right\}, \quad (18)$$

where $\mu_c = \mathbb{E}_0(U_{\ell,k} - c)^+$ and $(\sigma_c)^2 = \text{Var}_0(U_{\ell,k} - c)^+$ when $U_{\ell,k} \sim \chi^2_p$, $\mu_c^{(1)} = \mathbb{E}_1(U_{\ell,k} - c)^+$ and $(\sigma_c^{(1)})^2 = \text{Var}_1(U_{\ell,k} - c)^+$ when $U_{\ell,k} \sim \chi^2_p(\delta^2 p)$.

(b) (The extreme theory approximation when $c$ is large). For fixed $p$ channels, as $d \to \infty$, an upper bound of the optimal choice of the soft-thresholding parameter $c$ is

$$c_2 = p + 2\log(d). \quad (19)$$

The detailed proof of this Proposition will be given in the appendix. Roughly speaking, in Part (a), the $c_1$ value maximizes the power of the test under the specified alternative hypothesis $H_a$ in the proposition subject to the Type I error constraint $\alpha$ in (2), and the CLT is used to approximate these error probabilities. Despite its theoretical optimality properties, the $c_1$ value in (18) is often impractical, as it relies heavily on the specific assumption on the alternative hypothesis $H_a$, particularly the number of affected PCs, $d_0$. In our numerical simulation studies, we provide a data-driven approach to find $d_0$ value when the OOC models are assumed to be known, see subsection 4.2. Unfortunately, it is an open problem to choose a proper $d_0$ value under general conditions when little information about $H_a$ is available. Thus in our case study, we simply choose an ad hoc value $d_0 = d/3$.

The rationale of Part (b) is completely different, as it is to find largest $c$ to satisfy the Type I error constraint $\alpha$ in (2). This is similar to use the following well-known fact to choose the soft-thresholding parameter of $\sqrt{2\log(d)}$ for $d$ i.i.d. $N(0, 1)$ random variables, see Fan (1996),

$$\lim_{d \to \infty} \frac{\max_{1 \leq k \leq d} |Z_k|}{\sqrt{2\log(d)}} = 1 \quad \text{almost surely}$$

when the $Z_k$’s are i.i.d. $N(0, 1)$. Here we extend the critical value from $\sqrt{2\log(d)}$ for the i.i.d. $N(0, 1)$-distributed $Z_k$’s to $c_2$ for the i.i.d. $\chi^2_p$-distributed $U_{\ell,k}$’s for fixed $p$ as $d \to \infty$, and these two critical values are asymptotically equivalent when $p = 1$. Due to its simple but useful properties, the $c_2$ value in (19) is our recommended choice for the soft-thresholding parameter $c$ in practice, as it only depends on the number of channels ($p$) and the number of PCs ($d$).
It is also useful to comment on the main assumption of Proposition 1, which is the \( \chi^2_p \)-approximation for \( U_{\ell,k} \)'s in (10). We acknowledge that this holds rigorously only when the bases \( \nu_k(t) \)'s and \( \Sigma_k \)'s are known. However, we shall emphasize that we only need the approximation of the distributions of \( U_{\ell,k} \)'s in (10) to derive an approximation choice of the soft-thresholding parameter \( c \). For this reason, the assumption of \( \chi^2_p \)-approximation for \( U_{\ell,k} \)'s in (10) is not bad when the PCA is used to estimate \( \nu_k(t) \)'s and \( \Sigma_k \)'s.

4 Simulation Studies

In this section, we conduct extensive simulation studies that are motivated from the real forging dataset with the focus of detecting different types of changes with different magnitudes. A high-level description of our simulation setting is as follows. In each Monte Carlo simulation run, we generate \( m = 200 \) “observable” profiles from the change-point model in (1). Under the null hypothesis \( H_0 \), all \( m = 200 \) profiles are simulated from the generative model under the IC state. Depending on different assumptions of the alternative hypothesis \( H_a \), we will generate \( m = 200 \) profiles from the change-point model in (1) with change-point \( \tau = 100 \), i.e., the first 100 profiles, \( X_1(t), ..., X_{100}(t) \), are simulated from the generative model under the IC state, and the last 100 profiles, \( X_{101}(t), ..., X_{200}(t) \) are simulated from the generative model under one of many different OOC states.

For the purpose of comparison, three methods are applied to each set of \( m = 200 \) simulated “observable” profiles to detect whether there is a change-point or not. The baseline method is the PCA method developed in Paynabar et al. (2016) which assumes all PCs are affected by the change simultaneously, and can be regarded as the special case of our proposed method with the soft-thresholding parameter of \( c = c_0 = 0 \). The other two methods are our proposed thresholded PCA methods with two specific non-zero soft-thresholding parameters: one is \( c_1 \) in (18) and the other is \( c_2 \) in (19). For each of these three specific methods, we choose the respective control limit \( L \) in (13) to satisfy Type I error constraint \( \alpha = 0.05 \) by using the profiles generated from the null hypothesis \( H_0 \). Then each method is applied to profiles generated under the alternative hypothesis to see whether it is able to correctly detect the change time \( \tau = 100 \) or not. This process is repeated for 200 times, and the average performances are reported and compared for three different parameters \( c_0, c_1, c_2 \).
It is important to emphasize that our proposed thresholded PCA methods do not use any information or knowledge of the profile generative models under the IC or OOC states below, which are only used to generate the $m = 200$ “observable” 4-dimensional vector profiles $X_i(t)$’s.

For better presentation, the remainder of this section is divided into three subsections. In subsection 4.1, we use the real profiles and B-splines to present the generative models of 4-channel profiles under the IC state as well as a total of $3 \times 7 = 21$ different OOC states. This allows us to generate observed profiles $X_i(t)$’s from the change-point additive noise model in (1). In subsection 4.2, our proposed thresholded PCA methods are applied to the generated profiles $X_i(t)$’s, and the performance of the values of $c_1$ and $c_2$ in Proposition 1 is then compared with that of the baseline value $c_0 = 0$. In subsection 4.3, we compare these three methods under a more subtle case when 2 out of 4 channels in the 4-channel profiles are affected by the change. While our proposed thresholded PCA methods are not designed for this subtle case, we run simulations to see their performance as compared to the baseline method without thresholds.

## 4.1 Profile Generative Models

The generative model for profiles under the IC state is built from the 207 normal profiles, $X_1(t), \ldots, X_{207}(t)$, in the real forging dataset, and we propose to do so by using B-splines. To be more specific, we generate an unevenly spaced set of 66 B-spline bases in $[0, 1]$, and after orthogonalization and normalization we obtain basis $B_1(t), \ldots, B_{66}(t)$ using the “orthogonal-splinebasis” Package in the free statistical software R 3.1.2. Based on our experiences, the choice of 66 bases yields the best tradeoff to balance the empirical fitting of real normal profiles and the computational simplicity, though one can easily extend our simulations to another number of bases. Then our proposed generative model for normal profiles is of the form

$$X(t) = \sum_{i=1}^{66} \tilde{\theta}_i B_i(t), \quad (20)$$

where the 4-dimensional vectors $\tilde{\theta}_i$’s are assumed to be multivariate normally distributed.

It remains to estimate the distribution of $\tilde{\theta}_i$’s in (20) under the IC state from the observed 207 normal profiles in the real forging dataset. Note that in the original forgoing dataset,
each normal profile $X_i(t)$ is observed at 1200 different $t$ points, i.e., $t = j/1200$ for $j = 1, \cdots, 1200$. To speed up our computation and to reduce the profile noises, we first apply a non-overlapping moving average function with the window size of 3 to each profile, resulting in 207 “smoothed” 4-dimensional normal profiles $X(t)$’s with $t \in \left\{ \frac{j}{100} ; j = 1, \ldots, 400 \right\}$. Next, we fit each of 207 normal profiles $X(t)$ with B-spline basis $B_1(t), \ldots, B_{66}(t)$ using the least square estimation method, i.e., $\min_{\theta_1, \ldots, \theta_{66}} \| X(t) - \sum_{i=1}^{66} \hat{\theta}_i B_i(t) \|^2$. Hence, for each given B-spline basis $i = 1, \cdots, 66$, we obtain 207 fitted values for the 4-dimensional vector $\hat{\theta}_i$, which allow us to compute the corresponding sample mean and sample covariance matrix, denoted by $\bar{\theta}_i$ and $\Sigma_{\theta_i}$, respectively. The simulated normal profiles are then generated from (20) under the assumption that the IC distribution of $\tilde{e}_i$ is $N(\bar{\theta}_i, \Sigma_{\theta_i})$.

For profiles under the OOC state, we assume that the generative OOC model is the same as (20) but the means of $\tilde{e}_i$’s will be different. We will consider a total of $3 \times 7 = 21$ different OOC cases, depending on the location and magnitude of changes. First, we consider three different cases, depending on which location or subinterval of $[0,1]$ changes at the original profile scale of $X_i(t)$ for $0 \leq t \leq 1$, or equivalently, which subset of the 66 different $\tilde{\theta}_i$ in the model (20) changes their means:

(I) a local change of $\tilde{\theta}_i$ for $30 \leq i \leq 37$, i.e., over the interval $\frac{200}{400} \leq t \leq \frac{300}{400}$;

(II) a local change of $\tilde{\theta}_i$ for $16 \leq i \leq 29$, i.e., over the interval $\frac{90}{400} \leq t \leq \frac{140}{400}$; and

(III) a global change of $\tilde{\theta}_i$ for all $1 \leq i \leq 66$, i.e., over the interval $0 \leq t \leq 1$.

Second, we consider seven different magnitudes of the change. Observe that given the same magnitude of the change, it is the most difficult to detect the local change of Case (I) (where the peak of the profile occurs), and it is the easiest to detect the global change of Case (III). Hence we assign different magnitudes so that the detection powers of these cases are comparable: we assume that the real-valued mean of the $i$-th component $\tilde{\theta}_i^{(j)}$ changes from the IC value $\theta_i$ to the OOC value $\theta_i + 0.005 + 0.005 \times \Delta$. Here we set $\Delta = h + 1$ for local change in Case (I), $\Delta = h$ for local change in Case (II), and $\Delta = (0.1)h$ for global change in case (III), and we consider seven different values of $h$: $h = 1, 2, \cdots, 7$. In summary, depending on the location and magnitude of changes, there are a total of $3 \times 7 = 21$ OOC cases, and the numerical values are inspired from the real forging dataset, particularly the
Table 1: The value of $d_0$ and soft-thresholding parameters $c$’s

<table>
<thead>
<tr>
<th></th>
<th>baseline $c_0$</th>
<th>$d_0$</th>
<th>$c_1$ in (18)</th>
<th>$c_2$ in (19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOC-case (I)</td>
<td>0</td>
<td>15</td>
<td>4.9</td>
<td>11.6</td>
</tr>
<tr>
<td>OOC-case (II)</td>
<td>0</td>
<td>9</td>
<td>7.0</td>
<td>11.6</td>
</tr>
<tr>
<td>OOC-case (III)</td>
<td>0</td>
<td>12</td>
<td>4.5</td>
<td>11.6</td>
</tr>
</tbody>
</table>

locations of changes. The smaller magnitudes of the change are artificial so as to have a better understanding of the detection power as a function of change magnitudes.

4.2 Performance Comparison

In this subsection, we report the performance of our proposed thresholded PCA method with three different choices of the soft-thresholding parameter $c$, and our objective is to see whether the $c_1$ and $c_2$ values in Proposition 1 will yield a better performance as compared to the baseline $c_0 = 0$ in the sense of detecting those $2 \times 3 \times 7 = 42$ OOC cases.

In order to have a fair comparison, we fix the number of PCs as $d = 45$ for all three choices of soft-thresholding $c$ values, since that will explain more than 90% of the profiles variance for the 200 simulated forging dataset from (20). Here this $d$ value is different from those in the real dataset, likely due to different sample sizes or number of profiles in the dataset. The threshold $L = L_c$ in (13) for each $c$ value is chosen to satisfy the common Type I error constraint $\alpha = 0.05$. Table 1 lists the specific values of $c_0, c_1, c_2$ used in our study.

Note that $c_0$ and $c_2$ does not depend on the location or magnitudes of the changes, although $c_1$ does depend on the location of the change.

When computing the $c_1$ value in Proposition 1, we need to know the value of $d_0$, the number of affected PCs that are relevant to the change among the total of $d = 45$ selected PCs. The $d_0$ values reported in Table 1 are chosen by the following data-driven method: We first obtain $U_{\ell,k}^{H_0}\{k = 1,...,d\}$’s under $H_0$ using the simulated IC profiles and record the $A$ as the top 10% $U_{\ell,k}^{H_0}$ values. Then, we compute $U_{\ell,k}^{H_1}\{k = 1,...,d\}$’s under $H_1$ using simulated OOC profiles, and count how many $U_{\ell,k}^{H_1}$’s are greater than the threshold $A$. This gives an estimate of $d_0$ since it indicates the number of altered $U_{\ell,k}$’s if a specific fault occurs.

Figures 4-6 plot the detection power of our proposed methods with three different choices
Figure 4: Case (I) with a local change for $30 \leq i \leq 37$. Each curve represents our proposed method with a specific soft-thresholding $c$ values: Red line with circle (baseline $c_0 = 0$); blue line with square ($c_1$); and black line with star ($c_2$). The detection power of each method is plotted as the function of the 7 different change magnitudes based on 200 Monte Carlo runs. The standard errors of each detection power curve are between 0.01 and 0.04.
Figure 5: Case (II) with a local change for $16 \leq i \leq 29$. Each curve represents our proposed method with a specific soft-thresholding $c$ values: Red line with circle ($c_0 = 0$); blue line with square ($c_1$); and black line with star ($c_2$). The detection power of each method is plotted as the function of the 7 different change magnitudes based on 200 Monte Carlo runs. The standard errors of each detection power curve are between 0.01 and 0.04.
Figure 6: Case (III) with a global change for all $1 \leq i \leq 66$. Each curve represents our proposed method with a specific soft-thresholding $c$ values: Red line with circle (baseline $c_0 = 0$); blue line with square ($c_1$); and black line with star ($c_2$). The detection power of each method is plotted as the function of the 7 different change magnitudes based on 200 Monte Carlo runs. The standard errors of each detection power curve are between 0.01 and 0.03.
of soft-thresholding $c$ values as functions of change magnitudes. Figure 4 deals with the OOC case (I) where a local change affects the rise, peak, and fall segments of the profiles, and all three methods seem to have comparable detection powers, although $c_0 = 0$ is slightly worse. Figure 5 shows that under the OOC case (II), both $c_1$ and $c_2$ can greatly improve the detection power as compared with the baseline $c_0 = 0$, especially when the change magnitude is small (e.g., $h \leq 5$). For large change magnitudes, all three methods have detection power close to 1, implying that all reasonable methods should be able to detect large changes.

In Figure 6, we consider the OOC case (III) when a global change occurs over $[0, 1]$, and it is interesting to see that both $c_1$ and $c_2$ can yield a larger detection power than the baseline $c_0 = 0$, especially for small magnitude $h$. This might be counter-intuitive, as one might expect that thresholding does not help a global change. To gain a deep understanding, we further compare the distributions of the statistics $U_{\ell,k}$’s for all $d = 45$ bases under the IC case and the OOC case (III) state of a global change. Figure 7 plots the box plot of $U_{\ell=100,k}$ under the both IC and OOC-case(III) states for all $d = 45$ PCs. From the box plots, for the global change, it is surprising that almost half of $U_{\ell,k}$’s have a similar or smaller median value under OOC than IC, which implies that almost half of PCs have not been significantly affected by the change. We feel that this is the reason why soft-thresholding help improve the detection power in the global change case, as it can filter out those insignificant $U_{\ell,k}$’s or PCs.

We also evaluate the performance of our proposed method in terms of estimating the change-point $\tau$. When the true $\tau = 100$ is estimated as $\hat{\tau}$, we consider three different measures: $E(|\hat{\tau} - \tau|)$, $P(|\tau - \hat{\tau}| \leq 1)$ (denoted by P1) and $P(|\tau - \hat{\tau}| \leq 3)$ (denoted by P3). Table 2 reports the Monte Carlo simulation results under these three criteria based on 200 runs. In general all three values $c_0, c_1$ and $c_2$ yield comparable results in terms of estimating $\tau$, and it is interesting to note that the thresholding values $c_1$ and $c_2$ often have larger P1 and P3 than the baseline $c_0 = 0$ for the OOC case (II) with the local-mean shift cases. This suggests that thresholding might be able to locate the small, local change more precisely. One “strange” observation in Table 2 is that $E(|\hat{\tau} - \tau|)$ is not necessarily monotone as a function of the change magnitude $h$, and one possible explanation is because $|\hat{\tau} - \tau|$ takes on integer values, 0, 1, 2, · · · , 100, and $\hat{\tau}$ is a biased estimator.
Figure 7: Box plots of $U_{\ell,T=100, k}$ under the $H_0$ hypothesis and $H_1$ hypothesis for case (III) under all 4 channels affected scenario with $h = 4$ based on 1000 replications. The X-axis with $k = 1, \ldots, 45$ represents the projection on the $k$-th PC. This plot implies that even for the global change, the OOC distribution of the $U_{\ell, k}$’s is not necessarily stochastically larger than those IC distribution over all $k = 1, \ldots, 45$ PCs. This explains why soft-thresholding can improve the detection power in the global change case, as it can filter out those $U_{\ell, k}$’s that have smaller OOC values.
Table 2: Comparison of detection biases for each algorithms under 3 different OOC cases for all 4 channels affected scenario.

| Case   | $E(|\hat{\tau} - \tau|)$ | $P(|\tau - \hat{\tau}| \leq 1)$ | $P(|\tau - \hat{\tau}| \leq 3)$ |
|--------|--------------------------|----------------------------------|----------------------------------|
|        | $h$ | $c_0$ | $c_1$ | $c_2$ | $c_0$ | $c_1$ | $c_2$ | $c_0$ | $c_1$ | $c_2$ |
| (I)    |     |       |       |       |       |       |       |       |       |       |
| 1      | 5.18 ± 1.71 | 1.14 ± 1.89 | 0.86 ± 2.09 | 0.18 | 0.15 | 0.19 | 0.40 | 0.44 | 0.36 |
| 2      | 1.57 ± 1.33 | 1.89 ± 1.35 | 2.65 ± 1.73 | 0.22 | 0.22 | 0.22 | 0.51 | 0.50 | 0.39 |
| 3      | 0.95 ± 1.21 | 1.51 ± 1.27 | 0.59 ± 1.38 | 0.27 | 0.26 | 0.25 | 0.54 | 0.54 | 0.47 |
| 4      | 0.81 ± 1.10 | 1.03 ± 1.02 | 0.59 ± 1.26 | 0.31 | 0.36 | 0.33 | 0.57 | 0.63 | 0.54 |
| 5      | 0.28 ± 0.86 | 0.30 ± 0.88 | 0.09 ± 0.77 | 0.38 | 0.36 | 0.42 | 0.63 | 0.63 | 0.63 |
| 6      | 0.13 ± 0.73 | 0.13 ± 0.79 | 0.58 ± 0.47 | 0.41 | 0.42 | 0.46 | 0.65 | 0.68 | 0.66 |
| 7      | 0.14 ± 0.54 | 0.63 ± 0.46 | 0.29 ± 0.49 | 0.47 | 0.46 | 0.49 | 0.70 | 0.72 | 0.70 |
| (II)   |     |       |       |       |       |       |       |       |       |       |
| 1      | 2.15 ± 2.37 | 0.16 ± 2.45 | 2.30 ± 2.30 | 0.09 | 0.22 | 0.22 | 0.24 | 0.36 | 0.36 |
| 2      | 1.98 ± 1.64 | 0.78 ± 1.57 | 0.18 ± 1.50 | 0.24 | 0.35 | 0.35 | 0.48 | 0.51 | 0.53 |
| 3      | 1.12 ± 0.88 | 0.76 ± 1.08 | 1.19 ± 1.23 | 0.39 | 0.40 | 0.42 | 0.60 | 0.61 | 0.63 |
| 4      | 0.11 ± 0.70 | 0.67 ± 0.78 | 0.43 ± 0.71 | 0.48 | 0.53 | 0.56 | 0.72 | 0.76 | 0.76 |
| 5      | 0.51 ± 0.62 | 0.02 ± 0.54 | 0.24 ± 0.57 | 0.58 | 0.67 | 0.65 | 0.78 | 0.83 | 0.86 |
| 6      | 0.22 ± 0.53 | 0.51 ± 0.49 | 0.49 ± 0.49 | 0.70 | 0.76 | 0.73 | 0.86 | 0.91 | 0.90 |
| 7      | 0.50 ± 0.48 | 0.02 ± 0.14 | 0.07 ± 0.16 | 0.77 | 0.81 | 0.80 | 0.90 | 0.95 | 0.95 |
| (III)  |     |       |       |       |       |       |       |       |       |       |
| 1      | 0.07 ± 1.13 | 0.16 ± 1.07 | 1.18 ± 1.25 | 0.35 | 0.34 | 0.31 | 0.57 | 0.57 | 0.50 |
| 2      | 0.58 ± 1.11 | 0.37 ± 1.01 | 0.52 ± 1.07 | 0.39 | 0.35 | 0.34 | 0.60 | 0.61 | 0.54 |
| 3      | 0.85 ± 1.05 | 0.67 ± 0.94 | 0.51 ± 0.84 | 0.43 | 0.39 | 0.38 | 0.64 | 0.61 | 0.56 |
| 4      | 0.15 ± 0.90 | 0.11 ± 0.73 | 0.43 ± 0.82 | 0.45 | 0.41 | 0.40 | 0.67 | 0.64 | 0.61 |
| 5      | 0.13 ± 0.84 | 0.11 ± 0.66 | 0.27 ± 0.55 | 0.47 | 0.45 | 0.45 | 0.69 | 0.68 | 0.66 |
| 6      | 0.44 ± 0.79 | 0.04 ± 0.58 | 0.03 ± 0.52 | 0.49 | 0.48 | 0.46 | 0.70 | 0.71 | 0.67 |
| 7      | 0.39 ± 0.63 | 0.15 ± 0.54 | 0.01 ± 0.53 | 0.51 | 0.49 | 0.47 | 0.72 | 0.72 | 0.67 |
4.3 A more subtle case when 2 out of 4 channels are affected

Recall that for profiles under the OOC state, we assume that the generative OOC model is the same as (20) but the 4-dimensional random coefficients \( \tilde{\theta}_i = (\tilde{\theta}_i^{(1)}, \tilde{\theta}_i^{(2)}, \tilde{\theta}_i^{(3)}, \tilde{\theta}_i^{(4)}) \) are non-zero. A subtlety is how to define a non-zero 4-dimensional vector. In our simulations so far, we assume that all 4 components of \( \tilde{\theta}_i \) are non-zero simultaneously, i.e., all 4 channels are affected by the change simultaneously. In this subsection and only in this subsection, we consider a new OOC state when only 2 out of 4 channels are affected by the change. To be more specific, under the new OOC state, only the first 2 components/channels, \( \tilde{\theta}_i^{(1)} \) and \( \tilde{\theta}_i^{(2)} \), have OOC mean, and the last 2 components/channels, \( \tilde{\theta}_i^{(3)} \) and \( \tilde{\theta}_i^{(4)} \), keep the IC mean. All other OOC settings are the same as those in the previous subsection. While our proposed methods are not designed for this scenario, we run simulation to see their performance nevertheless.

Figure 8 plots the detection power of our proposed methods when only 2 out of 4 channels/components are affected. It is clear from the top and middle panels of Figure 8 that the \( c_1 \) and \( c_2 \) values greatly outperforms the baseline \( c_0 = 0 \) value for almost all shift magnitudes in the OOC case of local changes. In the bottom panel for the OOC case (III) of the global change, the detection power improvement is significant for \( c_2 \) as compared to the baseline \( c_0 = 0 \). We feel this might be due to the new spatial sparsity where the profile means of only two channels have shifted. While our proposed thresholded PCA method is not designed specifically for the spatial sparsity, the thresholding can actually take care of spatial sparsity to yield better detection power. In addition, if we compare Figure 8 with Figures 3-5, it is evident that the detection powers in the case of only 2 out of 4 affected channels are less than those in the case of all 4 channels affected by the change. In other words, it is more difficult to detect if only 2 out of 4 channels are affected by the change, which is consistent with our intuition.

5 Analysis of Real Forging Dataset

In this section, we apply our proposed thresholded PCA methodology to the real forging manufacturing process dataset in Paynabar et al. (2016) that includes 151 normal profiles and 69 fault profiles.
Figure 8: When only 2 out of 4 channels/components are affected. The three plots correspond to three OOC cases, depending on which subset of the 66 different $\tilde{\theta}_i$ in the model (20) changes their means. **Upper:** case (I) with a local change for $30 \leq i \leq 37$; **Medium:** case (II) with a local change for $16 \leq i \leq 29$ and **Bottom:** case (III) with a global change for all $1 \leq i \leq 66$. In each figure, each curve represents our proposed method with a specific soft-thresholding $c$ values: Red line with circle ($c_0$); blue line with square ($c_1$); and black line with star ($c_2$). The detection power of each method is plotted as the function of the 7 different change magnitudes. The standard errors of each detection power curve are between 0.01 and 0.04.
First, we apply our methods to the specific case study setup in Paynabar et al. (2016) where the 151 normal profiles are followed by the 69 fault profiles, i.e., the change-point $\tau = 151$ for the change-point model in (1). As mentioned before, Paynabar et al. (2016) intuitively chose the first $d = 15$ PCs that explain 85% of variations under the IC state, as they felt the baseline method of $c_0 = 0$ is suitable when the number $d$ of PCs should not be too large. With $d = 15$, the baseline of $c_0 = 0$ can correctly detect the true change-point $\tau = 151$. We also apply our proposed methods with $c_1$ and $c_2$ to the case of $d = 15$, and they can also correctly detect the true change-point $\tau = 151$. In the case study, $c_1$ in (18) was computed under a specific alternative hypothesis $H_a$ where OOC magnitude $\delta = 1$ and the number of affected PCs is $d_0 = d/3$ out of the $d$ selected PCs. Thus, it should be noted that this $c_1$ value might not be an optimal choice for other alternative hypotheses.

Next, we re-run the above analysis using $d = 21$ and $d = 33$ PCs that explain 90% or 95% of variations. It turns out that all three choices of $c$'s have the same performance on this particular full data set, and all correctly detect the true change-point $\tau = 151$. In other words, we do not have to worry about the choice of $d$ for this particular full real data set.

Moreover, instead of a single full dataset with $m = 151 + 69 = 220$ profiles, we use random sampling to get a better understanding of the robustness of our proposed methods in many smaller datasets. Note that we do not bootstrap the full dataset, since identical profiles/observations might be problematic or misleading in the PCA context. Instead we randomly generate many subsets of profiles that are half size of the full dataset. To be more specific, for each subset of profiles, we randomly selected 75 out of 151 normal real forgoing profiles and 35 out of 69 fault profiles. Then each subset of profiles consists of the 75 normal profiles followed by 35 fault profiles with the change-point $\nu = 75$. Next, we selected the first $d$ PCs that explain 95% of variations for this subset of real profiles, and our proposed methods with three different tuning parameters $c_0, c_1, c_2$ were applied to see whether or not we can detect the true change-point $\tau = 75$ in this subset of real forgoing data. This process was repeated 1000 times.

It turns out that all three methods correctly rejected the null hypothesis of no change in each of these 1000 repetitions at the significant level $\alpha = 5\%$, e.g., all methods have observed empirical detection power of 1. In addition, the Monte Carlo estimates of $\text{E}(|\hat{\tau} - \tau|)$ based on 1000 repetitions are 0.056, 0.049, 0.039 for $c = c_0, c_1, c_2$, respectively. The corresponding
estimates of $\Pr(|\tau - \hat{\tau}| \leq 1)$ is 0.991, 0.995, and 0.998, respectively, whereas all three methods have observed empirical probabilities of 1 for $\Pr(|\tau - \hat{\tau}| \leq 3)$. Our interpretation is that the change of real profiles seems to be significantly large, and thus all reasonable profile monitoring algorithms, including our proposed methods with any of the three $c$ values in (12), are able to detect the change correctly. However, our proposed thresholded PCA methodologies with $c_1$ and $c_2$ were more accurate to estimate the true change-point than the baseline method of $c_0 = 0$.

6 Conclusions and Future Work

In this paper, we proposed a thresholded multivariate PCA for multichannel profile monitoring. The novelty of our proposed method is to conduct dimension reduction in two steps: We first apply multivariate PCA to reduce high dimensional multichannel profiles to a reasonable number of features (e.g., PCs) under the IC state, and then use soft-thresholding techniques to further select informative features under the OOC state. These two steps allow us to include only those PCs that are informative to the change and smooth out the noisy ones, thereby yielding efficient monitoring. We also gave a couple of suggestions on how to select tuning parameters based on asymptotic analysis. Moreover, we used real forging process dataset and B-splines to build generative methods for multichannel profiles under the IC state and $3 \times 7 = 21$ different OOC states, depending on the location, and magnitude of the changes. Our numerical studies demonstrated that soft-thresholding can significantly improve the performance of PCA.

There are a number of interesting problems that have not been addressed here. From the theoretical point of view, it will be useful to investigate the efficiency of our proposed methods, and to find an optimal value of soft-thresholding parameter $c$ that can adaptively adjust for different OOC states. Another direction is to investigate how to extend our proposed method to Phase II online profile monitoring. That will be more challenging, partly because it is more difficult to select informative PCs due to fewer OOC profiles since one observes profiles one at a time. Therefore, our research should be interpreted as a starting point for further investigation. Finally, in this work we assumed that signals are temporally independent. Relaxing this assumption and extending the proposed methodology
for temporally correlated multichannel signals is another topic for future research.

Acknowledgement

The authors would like to thank the Editor, Associate Editor and two anonymous reviewers for their detailed and constructive comments that greatly improved the presentation of the article. The authors also gratefully acknowledge the support of the National Science Foundation under grant NSF CMMI-1362876 and CMMI-1451088.

Appendix: Proof of Proposition 1

Here we will provide the detailed proof of Proposition 1. Let us first prove part (a) when the central limit theorem (CLT) is applicable to $\sum_{k=1}^{d}(U_{\ell,k} - c)^+$. This can occur when the soft-thresholding parameter $c$ is small and the number $d$ of basis is large. By the notation in part (a), under $H_0$, the terms $(U_{\ell,k} - c)^+$ are i.i.d. with mean $\mu_c$ and variance $(\sigma_c)^2$, and thus $\sum_{k=1}^{d}(U_{\ell,k} - c)^+ \approx N(\mu_c d, \sigma_c^2 d)$ for any given $\ell$. Hence, the probability in (17) can be approximated by

$$P_{H_0} \left( N(0, 1) > \frac{L_c - \mu_c d}{\sigma_c \sqrt{d}} \right). \tag{21}$$

To satisfy the Type I error constraint (2) with small $\alpha$, the threshold $L = L_c$ can be approximated by $L_c \approx \mu_c d + \sigma_c \sqrt{d} z_u$, where $z_u = z$ such that $P(N(0, 1) > z) = u$.

Likewise, we can also derive the relationship of power function of the proposed test. Under the alternative hypothesis $H_a$ with the change time $\tau$, the term $(U_{\tau,k} - c)^+$ has mean $\mu_c$ and variance $(\sigma_c)^2$ if the $k$-th component is unaffected, and has mean $\mu_c^{(1)}$ and variance $(\sigma_c^{(1)})^2$ if affected. Recall that there are $d_0$ components are affected. When both $d_0$ and $d-d_0$ are relatively large, the CLT is applicable to both $\sum_{k=1}^{d_0}(U_{\tau,k} - c)^+$ and $\sum_{k=d_0+1}^{d}(U_{\tau,k} - c)^+$.

Hence, the power function of the proposed test is of the order of

$$P_{H_1} \left( \sum_{k=1}^{d}(U_{\tau,k} - c)^+ > L_c \right) = P_{H_1} \left( \sum_{k=1}^{d_0}(U_{\tau,k} - c)^+ + \sum_{k=d_0+1}^{d}(U_{\tau,k} - c)^+ > L_c \right)$$

$$\approx P_{H_1} \left( d_0 \mu_c^{(1)} + \sqrt{d_0 \sigma_c^{(1)}} Z_1 + \mu_c (d - d_0) + \sqrt{d - d_0} \sigma_c Z_2 > \mu_c d + \sigma_c z_u \sqrt{d} \right)$$

$$= P_{H_1} \left( N(0, 1) > -\frac{(\mu_c^{(1)} - \mu_c) d_0}{\sqrt{d_0}(\sigma_c^{(1)})^2 + (d - d_0)(\sigma_c)^2} + \frac{\sqrt{d_0} \sigma_c}{\sqrt{d_0}(\sigma_c^{(1)})^2 + (d - d_0)(\sigma_c)^2} z_u \right),$$

29
where $Z_1$ and $Z_2$ are independent $N(0, 1)$ random variables, and the last equation is from the fact that $aZ_1 + bZ_2 \sim N(0, a^2 + b^2)$. To maximize the power function under $H_a$, a natural choice of $c$ is the one that maximizes the above expression, and this leads to the $c_1$ value in (18), and thus part (a) holds.

Now let us prove part (b) by using the extreme theory approximation when $c$ is large. In this case, instead of the CLT, we will explore the following facts:

$$P_{H_0}(\sum_{k=1}^{d} (U_{\ell, k} - c) > L_c) < P_{H_0}\left(\max_{1 \leq k \leq d} U_{\ell, k} > c\right) < \sum_{k=1}^{d} P_{H_0}(U_{\ell, k} > c) = dP(\chi_p^2 > c).$$

Here the first equality follows from the simple fact that $(U_{\ell, k} - c)^+ > 0$ for some $1 \leq k \leq d$ when $\sum_{k=1}^{d} U_{\ell, k} - c > L_c > 0$, and the last equality uses the main assumption of the proposition that $U_{\ell, k} \sim \chi_p^2$ under $H_0$. To satisfy Type I error constraint in (2), it suffices to find $c$ such that $P(\chi_p^2 > c) \approx \frac{2}{d}$ for fixed $p$ and $\alpha$ as $d \to \infty$, i.e., $\log P(\chi_p^2 > c) \approx -\log(d)$.

When $p = 1$, for large $c > 0$, we have

$$P(\chi_1^2 > c) = 2P(N(0, 1) > \sqrt{c}) \approx 2\frac{\phi(\sqrt{c})}{\sqrt{c}} = \frac{2}{\sqrt{\sqrt{\pi}c}} \exp\left(-\frac{c}{2}\right),$$

where we use the well-known fact that $\frac{1}{u+1/u} \phi(u) \leq P(N(0, 1) > u) \leq \frac{1}{u} \phi(u)$ for all $u > 0$. Taking logarithm both sides, we have $c \approx 2\log(d)$ to satisfy Type I error constraint in (2) for fixed $\alpha$ as $d$ goes to $\infty$. This is consistent with the well-known fact that $\sqrt{2\log(d)}$ is the critical soft-thresholding value for the $d$ i.i.d. $N(0, 1)$ random variables, see Fan (1996).

Now we need to extend the above arguments from $p = 1$ to any $p > 1$. The crucial step is to approximate $P(\chi_p^2 > c)$ for large $c > 0$. By Lemma 1 of Inglot and Ledwina (2006), we have

$$\frac{1}{2}E(c) \leq P(\chi_p^2 > c) \leq \frac{1}{\sqrt{\pi}(c - p + 2)E(c)},$$

for $p \geq 2$ and $c > p - 2$ and

$$E(c) = \exp\left\{-\frac{1}{2}\left(c - p - (p - 2)\log(c/p) + \log p\right)\right\}.$$

This implies that $\log P(\chi_p^2 > c)$ is asymptotically equivalent to $\log E(c) \approx -(c - p)/2$ for fixed $p$ as $c \to \infty$, also see Theorems 4.1 and 5.1 in Inglot (2010). Thus in order to satisfy Type I error constraint in (2), we have $c \approx p + 2\log(d)$ if we ignore all non-essential constants when $\alpha$ and $p$ are fixed, and $d$ goes to $\infty$. Hence, part (b) is proved.
References


