Use of Sparse Principal Component Analysis (SPCA) for Fault Detection

Shriram Gajjar*, Murat Kulahci **, Ahmet Palazoglu*

*University of California, Davis, CA 95616, USA
(Tel: 530-752-8774; e-mail: anpalazoglu@ucdavis.edu).
** Technical University of Denmark, Lyngby, Denmark and Luleå University of Technology, Luleå, Sweden
(e-mail: muku@dtu.dk)

Abstract: Principal component analysis (PCA) has been widely used for data dimension reduction and process fault detection. However, interpreting the principal components and the outcomes of PCA-based monitoring techniques is a challenging task since each principal component is a linear combination of the original variables which can be numerous in most modern applications. To address this challenge, we first propose the use of sparse principal component analysis (SPCA) where the loadings of some variables in principal components are restricted to zero. This paper then describes a technique to determine the number of non-zero loadings in each principal component. Furthermore, we compare the performance of PCA and SPCA in fault detection. The validity and potential of SPCA are demonstrated through simulated data and a comparative study with the benchmark Tennessee Eastman process.

Keywords: Multivariate quality control; Statistical Process Monitoring; Dimension Reduction

1. INTRODUCTION

Chemical process operations are typically subject to process or operational disturbances. Therefore, timely and effective detection and diagnosis of faults (monitoring) are critical to ensure safety and process stability, and to maintain optimal levels of operation. For process monitoring, techniques based on first principle models have been studied for more than two decades but their contribution to industrial practice has not been pervasive due to substantial cost and time required to develop a sufficiently accurate model for a complex chemical plant. On the other hand, in a large-scale unit, a Distributed Control System (DCS) collects data from sensor arrays distributed throughout the plant and stores the data at high sampling rates. This data contains information about the underlying process characteristics and can be used for process monitoring. For process monitoring, a plant operator monitors several DCS screens and using his/her experience and domain knowledge, focuses on critical process variables to anticipate and prevent abnormal or faulty process operations. In the absence of such experience or domain knowledge, however, more automated techniques are required to inform and advise plant operators. Moreover, abundance of data and multiple variables to monitor at once make the task of monitoring unfeasibly difficult.

Principal Component Analysis (PCA) based monitoring methods, which build statistical models from normal operation data and partition the measurements into a principal component subspace (PCS) and a residual subspace (RS), are among the most widely used multivariate statistical methods (Cinar et al., 2007). In these approaches, the dimensions of the PCA model, i.e., the number of principal components (PCs) retained, must be decided and this decision has an important role on the process monitoring performance. However, the approach to the determination of the number of PCs to be retained is not unique, especially due to the influence of sensor noise (Tamura and Tsujita, 2007). To tackle this challenge, a number of well-known techniques for selecting the number of PCs have been proposed. A simple approach is to choose the number of PCs for the explained variance to achieve a predetermined percentage, such as 85%, termed as cumulative percent variance (CPV) (Jackson, 2004). Other methods, including cross validation, average eigenvalue approach, variance of reconstruction error (VRE) criterion, and fault signal-to-noise ratio, have also been proposed (Tamura and Tsujita, 2007, Valle et al., 1999, Wold, 1978, Dunia and Qin, 1998). In addition, the fault detection ability has been shown to depend on the PCs retained in the PCA model (Kano et al., 2002). Togkalidou et al. (2001) also indicated that including components with smaller eigenvalues in the PCA model and excluding those with larger eigenvalues could improve the prediction quality. In most such approaches, rather than the magnitudes of component loadings, only the significant data variations are considered to extract PCA components as this does not require making any a priori assumptions about data structure. The downside however is that the resulting loadings of the extracted PCA components are difficult to interpret. Motivated by this perspective, the present work deals with several of the limitations inherently associated with the interpretation of loadings of retained PCs. Recently, Liu et al. (2015) demonstrated that their proposed adaptive sparse PCA method outperforms the PCA method. Xie et al. (2013) proposed shrinking principal component analysis (ShPCA) strategy for fault detection and isolation. Their works support the case that traditional PCA can be altered in such a way that the obtained loadings would have a clear interpretation without significant loss of information in each PC. Such an approach might help in the application of PCA as better understanding of the impact of PC loadings can clearly facilitate process monitoring. This paper attempts to use sparse principal component analysis (SPCA) proposed by
Zou et al. (2006) to illustrate its advantages with a synthetic example. Second, we compare fault detection rates using SPCA and PCA for data from the Tennessee Eastman benchmark process.

2. DESCRIPTION OF METHODS

2.1 Principal Component Analysis

PCA finds the strongest orthogonal directions of variation in a data, starting with the largest direction of variation. PCA includes the decomposition of data matrix $X \in \mathbb{R}^{n \times m}$ that contains $n$ regularly-sampled observations of $m$ process variables and is scaled to zero mean and unit variance, into a transformed subspace of reduced dimensions. The decomposition is expressed as follows:

$$X = TP^T = \hat{X} + E$$

(1)

where $T \in \mathbb{R}^{m \times m}$ and $P \in \mathbb{R}^{m \times m}$ are the score matrix and the loading matrix, respectively. The matrices $\hat{X}$ and $E$ represent the estimate of $X$ and the residual part, respectively. The principal component (PC) projection reduces the original set of variables to $l$ PCs. The decomposition assumes that the PC loadings are orthonormal and correspond to the squared prediction error from the modelling data, respectively. This approximating distribution is found to

$$Q_i = \text{e}_i^T \text{e}_i = x_i^T (I - P_l P_l^T) x_i$$

(3)

where, $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_m)$, $e_i$ is the error in $i^{th}$ observation vector, and $I \in \mathbb{R}^{m \times m}$ denotes the identity matrix. The status of the process can then be monitored by these two scalars $T^2$ and $Q$.

In ordinary PCA, the PCs are uncorrelated and their loadings are orthogonal. Since SPCA does not require the sparse PCs to be uncorrelated, the sample covariance is modified to account for this in the Hotelling's $T^2$ measure viz. SPCA $T^2$.

$$SPCA \ T^2 = t_i^T \gamma^{-1} t_i$$

(4)

where, $y$ is the covariance matrix of the PC scores.

The $T^2$ statistic measures the variation in latent variables subspace (Tong et al., 2013). The approximated control limits of $T^2$ and $Q$ statistics, with a certain confidence level (i.e., false alarm rate), can be determined from the normal operating data in several ways by applying the probability distribution assumptions (Kourti and MacGregor, 1995, Qin, 2003). The control limits can be calculated as follows:

$$T^2_a = \hat{T}^2 \text{S}^{-1} \hat{T} \leq \frac{(n-1)\alpha}{(n-1)} F_{\alpha}(l, l-1)$$

(5)

where, $S = T^T T/(n-1)$ is the covariance matrix of the PC scores of the training dataset. $F_{\alpha}(l, l-1)$ is the upper $\alpha$ percentile of the $F$-distribution with the degree of freedoms $l$ and $(n-1)$.

The $Q$ statistic is a measure of variation that breaks the normal process correlation, which is given as:

$$Q_i = e_i^T e_i = \sum_{j=1}^{n} e_{ij}^2 \leq Q_a$$

(6)

where, $e_{ij}$ is the error in $i^{th}$ variable at $j^{th}$ observation, $Q_a$ is the corresponding control limit, which can be calculated by using a weighted $F$-distribution (Nomikos and MacGregor, 1995):

$$Q_a = g \chi^2_{\alpha} \quad g = \frac{\nu}{2m-2h} = \frac{2m^2}{\nu}$$

(7)

where, $m$ and $\nu$ are the estimated mean and variance of the squared prediction error from the modelling data, respectively. This approximating distribution is found to

IFAC DYCOPS-CAB, 2016
June 6-8, 2016. NTNU, Trondheim, Norway

694
work well even in cases where the errors are not normally distributed (van Sprang et al., 2002).

Although both, $T^2$ and $Q$ are used simultaneously for process monitoring, they address different situations of the process, and their roles in process monitoring are not symmetric but complementary. The $Q$ index measures variability that breaks the normal process correlation, which often indicates a faulty situation. The $T^2$ index measures the distance to the origin in the PCS. PCS contains information about deviations from the expected behaviour that represents signals while the noise information is mostly contained in the RS. The normal region defined by the $Q$ control limit includes residual components that are mainly noise. Overall, $T^2$ statistic measures the significant variability within the PCs, while $Q$ indicates how well the reduced dimensional PCA model can describe the significant process variation. In most cases, $Q$ is considered as an assistant statistic for $T^2$ considering that the value of $Q$ can be significantly affected by the process noise. The process is considered out-of-control if the statistics of a new observation exceeds the control limit.

Once either $T^2$ or $Q$ statistic detects a fault, a contribution plot is often used to identify the root cause. Contribution plots provide information about which variables contribute to the distance between the points in a $Q$ or $T^2$ chart and the sample mean of the data. The high contribution of a variable implies it could be the main source of out-of-control signal. The variable contribution to $Q$ is calculated as follows:

$$Con_Q = [(I - P_i^TP_i)\bar{x}_i]^T$$

(8)

The variable contribution to $T^2$ is calculated as follows:

$$Con_{T^2} = x_i^TP_i\Omega^{-1}P_i^T$$

(9)

where, $x_i$ is the $i^{th}$ observation vector.

2.4 Process monitoring procedure

A real-time monitoring scheme can be implemented efficiently using a two-step algorithm. First, a sufficiently large number of observations from a good process are collected to obtain the PCA model. The number of PCs to be retained is determined based on criteria mentioned in section 2.1. As the new observations become available, they are projected onto the loading matrix to obtain the score matrix. The monitoring statistics $T^2$, SPCA $T^2$ and $Q$ are calculated. The complete monitoring procedures are summarized as follows:

2.4.1 Off-line modelling:

1. Sufficient data is acquired when a process is operated under normal operating conditions. Each column (variables) of the data matrix is normalized, i.e., scaled to zero mean and unit variance.
2. PCA is applied to the data matrix, and the loading matrix $P \in \mathbb{R}^{n \times k}$ and the eigenvalue matrix $E = diag \{ \lambda_1, \lambda_2, ..., \lambda_n \}$ are obtained.
3. The threshold of cumulative percent variation, $\eta$ is specified.
4. The control limits are computed according to equations 5 and 7.

2.4.2 Real-time monitoring:

When the real-time observations become available, the monitoring task can be initiated.

1. For real-time monitoring, each new sample is scaled based on the mean and the variance obtained from offline model. Subsequently, it is projected onto loadings to obtain the corresponding scores.
2. The first $I$ PCs that capture the dominant variability are retained.
3. If the fault detection statistics are within the control limit, the process is judged as in control.

3. CASE STUDIES

3.1 Synthetic data

To demonstrate the methods and illustrate the advantages, we use a multivariate dataset for which the correlation structure ($\Omega$) is specifically defined. First, we generate data ($X$) from the random normal distribution $N(0, I)$. Second, Cholesky decomposition of the correlation matrix is multiplied by data generated and the resulting matrix ($D$) becomes the required synthetic data. $D$ will have the specified correlation $\Omega$.

$$D = [\text{Cholesky}(\Omega)][X]$$

(11)

3.2 Tennessee Eastman Process (TEP)

TEP is a well-known benchmark for testing the performance of various fault detection methods (Lyman and Georgakis, 1995). The process has 22 continuous process measurements, 12 manipulated variables, and 19 composition measurements sampled less frequently. Details on the process variables and their descriptions can be found in (Chiang et al., 2001). A total of 33 variables, excluding composition variables, are used for monitoring in this study.

The detailed description of two fault scenarios used in this study is given in Table 1. The data are generated at a sampling interval of 3 min and can be downloaded from http://brhaps.scs.uiuc.edu. 960 normal samples are used to build the model. Each fault data set contains 960 samples, with the fault introduced after sample 160. Along with the CPV criterion, a Scree test (Cattell, 1966) was also used to determine the number of PCs to retain. Based on this test, we retained 14 PCs to capture 85% variance from the data.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A/C feed ratio, B composition constant (stream 4)</td>
<td>Step</td>
</tr>
<tr>
<td>4</td>
<td>Reactor cooling water inlet temperature</td>
<td>Step</td>
</tr>
</tbody>
</table>
4. RESULTS AND DISCUSSION

4.1 Synthetic data

To first determine the NNZV to be retained for each PC in SPCA, we systematically increased the NNZV on the corresponding PC and obtained the total explained variance by SPCA. The NNZV for a sparse PC was selected when the variance captured was similar to the variance that was explained by PCA. As shown in Fig. 1, the NNZV to be retained on PC1, PC2 and PC3 were 5, 3 and 2, respectively.

4.1.1 Loadings obtained from PCA and SPCA

The cumulative variance captured by the first 3 PCs in PCA and SPCA is 87% but the SPCA provides a much sparser representation. In this synthetic example, SPCA correctly identifies the important variables. Moreover the associated variables on the 3PCs do not overlap as expected from the correlation structure in D. This example demonstrates that SPCA not only reduces the data dimension but also makes it easy for the user to interpret the associated variable loadings.

![Figure 1: Variance captured by each PC in SPCA while the NNZV on each PC is systematically increased.](image)

For fault detection, each data sample is first projected onto PC subspace, then Hotelling’s $T^2$ and $Q$ statistic are obtained. If the sample violates the control limit for the statistics, then one has to backtrack what variable in the sample was the root cause. As seen in Table 2, SPCA PCs are relatively sparse. The advantage of sparse loadings is that it makes the diagnosis easier. Also, Average Run Length (ARL), determines the average number of samples that will be taken before an out-of-control condition is detected. We have simulated datasets to compare ARL from SPCA and PCA wherein we introduced shift in mean for a few variables and also changed the underlying correlation structure. Table 3 shows some of the ARL results from a mean shift in variable 1 and a change in the correlation between variable 1 through 5 from 0.9 to 0.4. It can be observed that the results for both SPCA and PCA are comparable.

4.2 Tennessee Eastman process (TEP)

In this study, we retained 14 PCs to achieve a predetermined CPV of 85% by PCA. Then for SPCA, we used the same method as described for synthetic example to determine the NNZV in each 14 PCs in addition to limiting the total CPV captured to min. of 80%. Thus, the sparsity in the PC loadings is achieved via the variance captured trade-off. The NNZV determined for PC1 to PC14 were 16, 15, 8, 3, 2, 2, 2, 2, 2, 2, 1, 1 and 1 respectively. This is quite remarkable as 11 out of 14 retained PCs are now only linear functions of 3 or less variables as opposed to being a linear combination of 33 variables as in the conventional PCA.

### Table 2. PC loadings comparison for PCA and SPCA

<table>
<thead>
<tr>
<th>Var 1</th>
<th>PC1</th>
<th>SPCA</th>
<th>Var 2</th>
<th>PC2</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.440</td>
<td>0.441</td>
<td>-0.040</td>
<td>0.445</td>
<td>-0.029</td>
<td>0.020</td>
</tr>
<tr>
<td>0.450</td>
<td>0.446</td>
<td>-0.033</td>
<td>0.443</td>
<td>-0.029</td>
<td>0.020</td>
</tr>
<tr>
<td>0.450</td>
<td>0.446</td>
<td>-0.033</td>
<td>0.443</td>
<td>-0.029</td>
<td>0.020</td>
</tr>
</tbody>
</table>

### Table 3. ARL based on 1000 simulations, Phase I from 500 samples, Phase II from 2000 samples

<table>
<thead>
<tr>
<th>Fault</th>
<th>SPCA</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean shift</td>
<td>SPCA</td>
<td>$T^2$</td>
</tr>
<tr>
<td>Var 1</td>
<td>mean = 0.5</td>
<td>10</td>
</tr>
<tr>
<td>Var 2</td>
<td>mean = 1</td>
<td>74</td>
</tr>
<tr>
<td>Var 3</td>
<td>mean = 2</td>
<td>67.5</td>
</tr>
</tbody>
</table>

Correlation changed (off-diagonals for Var 1-5 in Ω = 0.4)

<table>
<thead>
<tr>
<th>Fault</th>
<th>SPCA</th>
<th>$T^2$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>Corr = 0.4 from 0.9</td>
<td>181.2</td>
<td>2.4</td>
</tr>
</tbody>
</table>

4.2.1 Loading plots

As it can be seen in Fig. 2, a PC1, the first PC, has non-zero loadings for all variables whereas 16 variables have non-zero loadings for sparse PC1as shown in Fig. 2b. Thus, after a fault is detected on PC1, one has to identify a root cause only from 16 out of 33 variables when using SPCA. The same phenomenon can be observed in Fig. 2 for other PC pairs. The sparsity of loadings should facilitate the fault diagnosis which will be further investigated.

![Figure 2: Comparison of variable loadings on PC1 and PC2 vs. PC2 and PC3 obtained from PCA (a, c) and SPCA (b, d) for TEP.](image)
4.2.2 Fault detection, Type I and Type II errors
For the PCA-based monitoring method, the number of PCs greatly affects the ability of fault detection (Tamura and Tsujita, 2007). Thus the total variance captured will impact the fault detection. In the case of SPCA because of the sparsity-variance trade-off, about 5% less variance is captured. However, our preliminary results as shown in Table 4 illustrate that SPCA based fault detection is comparable to that of conventional PCA.
To quantifying the fault detection performance, in this work, the missed detection rate is expressed as the ratio of faulty data samples that are not detected as faulty to the total faulty data samples specific to a fault.

### Table 4. Fault detection comparison of PCA and SPCA

<table>
<thead>
<tr>
<th>Fault</th>
<th>PCA $\tau^2$ (Q (SPE)) Missed detection rate (%)</th>
<th>SPCA $\tau^2$ (Q (SPE)) Missed detection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fault #1</td>
<td>0.00  0.00</td>
<td>0.00  0.00</td>
</tr>
<tr>
<td>Fault #2</td>
<td>0.00  0.00</td>
<td>0.00  0.00</td>
</tr>
<tr>
<td>Fault #3</td>
<td>0.00  0.00</td>
<td>0.00  0.00</td>
</tr>
<tr>
<td>Fault #4</td>
<td>0.00  0.00</td>
<td>0.00  0.00</td>
</tr>
</tbody>
</table>

4.2.3 Fault detection visualization
Faults #1 and #4 were selected for this paper since Fault #1 is expected to be detected by Hotelling’s $T^2$ whereas Fault #4 is expected to be detected by the $Q$ statistic. Figure 3 shows fault detection results for PCA and SPCA with corresponding control limits. Although 5% less variance is explained by SPCA, it can be observed from Fig. 3 that the fault detection capability is comparable to PCA.

4.2.4 Fault diagnosis

Figure 4 is a comparative plot of the variable contributions to $Q$ statistic for PCA and SPCA for Fault #4 associated with a step change in the reactor cooling water inlet temperature. A variable that is likely to cause a fault should have a higher and more persistent contribution to this statistic so that it can be isolated quickly and reliably to rectify the fault. As illustrated in Fig. 4, SPCA correctly identifies variables 21 and 32 (the reactor cooling water outlet temperature and reactor cooling water flow) as the likely root cause whereas PCA identifies variables 6 and 9 (reactor feed rate and reactor temperature) which reflects a secondary effect and may confound the diagnosis leading to a potential wild goose chase in pursuit of fault diagnosis. Moreover, there may be several causes that could affect variables 6 and 9 whereas variables 21 and 32 are directly related to fault #4. This is the essential benefit of SPCA as it reduces superfluous influences of all variables on a PC, thereby capturing causal relationship more effectively.

5. CONCLUSIONS

In this work we focus on the use of SPCA and discuss its potential for fault detection and diagnosis. The feasibility and validity of the proposed technique are demonstrated through a synthetic example and a benchmark process. We showed that in the benchmark process comparison while SPCA captured 5% less variance as compared to PCA, the fault detection capability is not compromised. Moreover, we showed that SPCA correctly identified the variables that were likely to cause a fault in question.

An important aspect of the proposed technique that requires further investigation is the analysis of the PC scores to
discern features during faulty operating conditions. Such characteristics will guide fault isolation as sampled data is acquired in real-time. Furthermore, such knowledge can be used for fault prediction (Skittides and Früh, 2014) and taking pre-emptive action for fault prevention.

REFERENCES


