# Process Monitoring Based on Improved Principal Component Analysis

Xiao Yingwang[1, a]

[1] School of Automation, Guangdong Polytechnic Normal University, Guangzhou, China 510665

[a]ymy19701030@163.com

**Keywords:** Principal component analysis; Process monitoring; Principal-component-related variable; Double-effect evaporator process

**Abstract:** The information provided by $T^2$ and squared prediction error (SPE) test of principal component analysis (PCA) is not corresponding. An improved PCA is proposed which uses principal-component-related variable residual statistic and common variable residual statistic to replace SPE statistic. Then a simulated double-effect evaporator is monitored by using the proposed method and comparisons with the conventional PCA are made. The simulation result shows that the improved PCA can avoid the conservation of SPE statistical test and provide more explicit information about the process conditions. So the improved PCA has an enhanced fault diagnosing performance.

## Introduction

Principal component analysis (PCA) is one of multivariate statistical process monitoring methods which is currently studied widely. It has many reports in both theory and application [1, 2]. However, the tests of $T^2$ and $SPE$ have different meaning [3]. When $SPE$ statistic varies heavily (i.e. case (1) and case (3)), it indicates the relationship between normal working conditions which the PCA statistical model presented is destroyed and it caused by the fault of process or sensor. While $T^2$ statistic varies heavily but $SPE$ statistic varies slightly (i.e. the case (2)), it indicates the relationship between the variables still (almost) stable, whereas the process changes, it may be caused by the change of working conditions or caused by the fault which just breaks the relationship of variables a little. The approach of fault subspace proposed by Dunia et al and some consequent research based on this approach [4] discuss the problem of process monitoring for cases (1) and (3). However, the geometric approach they proposed for the case (2) cannot identify what caused $T^2$ statistic changed, whether by normal variation or by fault.

## Introduction to PCA Statistical Process Monitoring Model

Choose the data sets of normal production process $X_{n\times m}$ ($n$ samples, $m$ process variables) to establish the statistical model firstly. Data matrix need to be standardized, that is to transform the samples of every moment $x=[x_1, x_2, \mathbf{L}, x_m]^T$ in $X_{n\times m}$ as follows: according to the methods of centralization and standardization, we get $\bar{x}=D_s^{-1}[x-E(x)]$ and in which $E(x)=[m_1, m_2, \mathbf{L}, m_m]^T$ is the corresponding mean vector with regard to $x$ and the variance matrix $D_s=diag(s_1, s_2, \mathbf{L}, s_m)$, in which $s_j=\sqrt{E(x_j-m_j)^2}$ is the $j$-th $(j=1,2,\mathbf{L},m)$ standard deviation of process variables. The matrix after standardization is marked as $\bar{X}_{(n\times m)}$. The eigenvalue decomposition for $\bar{X}^T\bar{X}$ is:

$$\bar{X}^T\bar{X}=UD_lU^T \tag{1}$$

in which $D_l=diag(l_1, l_2, \mathbf{L}, l_m)$ where $l_i$ is the eigenvalues and $U_{m\times m}=[u_1, u_2, \mathbf{L}, u_m]$ where $u_i$ is the corresponding standard orthonormal eigenvectors of eigenvalue $l_i$.

The subspace consists of the front $k(k<m)$ dimensional linear independent vector

 

$P=[u_1,u_2,\mathbf{L},u_k]$ denotes the principal component subspace $\hat{S}$ and the back $m-k$ dimensional vector $\tilde{P}=[u_{k+1},u_{k+2},\mathbf{L},u_m]$ constitutes the residual subspace $\tilde{S}$. The principal component $k$ is selected according the percentage of the cumulative variance. The data vector $\bar{x}$ can be decomposed as:

$$\bar{x}=\hat{x}+\tilde{x}=\hat{C}\bar{x}+\tilde{C}\bar{x} \tag{2}$$

in which $\hat{x}$ and $\tilde{x}$ are the projections on $\hat{S}$ and $\tilde{S}$ respectively. The projection matrices are $\hat{C}=PP^T$ and $\tilde{C}=\tilde{P}\tilde{P}^T=I-\hat{C}$. Specifically, we establish two statistics, namely $T^2$ statistic and $SPE$ statistic. The statistic $T^2$ in $\hat{S}$ space is defined as:

$$T^2=\left\|D_{l_k}^{-1/2}t\right\|^2=\left\|D_{l_k}^{-1/2}P^T\bar{x}\right\|^2\le d_T^2 \tag{3}$$

in which $D_{l_k}^{-1/2}=diag(l_1^{-1/2},l_2^{-1/2},\mathbf{L},l_k^{-1/2})$. The $t=P^T\bar{x}$ is the principal component score vector and $d_T$ is the control limit for statistic $T^2$. The statistic $SPE$ which used for monitoring the residual in space $\tilde{S}$ is defined as:

$$SPE=\left\|\tilde{C}\bar{x}\right\|^2\le d_{SPE}^2 \tag{4}$$

in which $d_{SPE}$ is the control limit for $SPE$. The control limits $d_T$ and $d_{SPE}$ are determined by sampling distribution of the statistics $T^2$ and $SPE$. Thus we can establish the process statistical model through the analysis of the normal process data. We can judge whether the process change by taking the $T^2$ and $SPE$ tests for the sample at each moment based on statistical model.

**Principle-component-related Variable Residuals Statistic and Common Variable Residuals Statistic**

If the process variables which are closely related with principal component from statistic $SPE$ are isolated, then coordinate with the statistic $T^2$, thus will make up for lack of the conventional PCA. Therefore, we propose an improved PCA method. In the improved PCA, those process variables closely related to principal component are isolated to constitute a new residual statistic (i.e. the $PVR$ statistic).

Theorem 1 [5]: Let $T$ be the score matrix of the former $k$ principal component of process data $X_{n\times m}$ ( $n$ samples, $m$ variables), then the squared multi-correlation coefficient of $x_j(j=1,2,\mathbf{L},m)$ and $T$ is :

$$g_j=r^2(x_j,T)=\sum_{i=1}^{k}l_i p_{j,i}^2 \tag{5}$$

in which $p_{j,i}$ is the each element of loading matrix $P$, $l_i$ is the eigenvalue of the covariance matrix after $X_{n\times m}$ standardization which can be calculated by the equation (1). Setting a threshold for multi-correlation coefficient $g$, those multi-correlation coefficients of process variable and principal component which are greater than the threshold are labeled as PVs (Principle-component-related Variables; PVs). The threshold of $g$ would be selected after determining the number of principal components.

Definition 1: Assuming there are $s(s<m)$ PVs in the set of process variable, we call the residual they constituted as the statistic $PVR$. Supporting the former $s$ variables of the process variable data set are PVs, then we call remaining $(m-s)$ detected variables as Common Variables

(Common Variables; CVs), we call the residual they constituted as the statistic $CVR$. In which they are similar to the equation (4):

$$PVR = \left\| D_{PV}\tilde{x} \right\|^2 = \left\| D_{PV}\tilde{C}x \right\|^2 \le d_{PV}^2 \;\; CVR = \left\| D_{CV}\tilde{x} \right\|^2 = \left\| D_{CV}\tilde{C}x \right\|^2 \le d_{CV}^2 \tag{6}$$

in which $D_{PV(m\times m)} = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix}$ and $D_{CV(m\times m)} = \begin{bmatrix} 0 & 0 \\ 0 & I_{(m-s)} \end{bmatrix}$, $d_{PV}$ and $d_{CV}$ are the control limits.

Similar to $SPE$ test, then we need to establish the limits $d_{PV}$ and $d_{CV}$ of the statistics $PVR$ and $CVR$ respectively. From the equation (6), it shows the relationship of $PVR$, $CVR$ and $SPE$ is:

$$SPE = PVR + CVR \tag{7}$$

The control limit of statistics $PVR$ and $CVR$ can directly be calculated by following equation:

$$SPE = PVR + CVR = w_{PVR} \times SPE + w_{CVR} \times SPE, \;\; w_{PVR} + w_{CVR} = 1 \tag{8}$$

Another significance of multi-correlation coefficient is its square which reflecting how much information of a variable which is summarized by the principal component [6], so we can take the value of the weights $w_{PVR}$ and $w_{CVR}$ in the equation (8) as:

$$w_{PVR} = 1 - \sum_{j\in PV} g_j \Big/ \sum_{j=1}^{m} g_j, \quad w_{CVR} = 1 - \sum_{j\in CV} g_j \Big/ \sum_{j=1}^{m} g_j = 1 - w_{PVR} \tag{9}$$

**The Simulation Study**

Double-effect evaporator is a kind of multi-effect evaporation. The detailed process description about the double-effect evaporator is shown in the paper [7].In this paper, all variables except $M_2$, $M_1$ and steady-state operating variable are monitored, i.e. the variables $\{W_F, W_{SO}, X_2, X_1, T_2, T_1, W_{S2}, W_{S1}, W_2, W_1\}$. Choose 400 sampling data of monitored variables under normal steady state, and the sampling interval is 3 seconds, that can get data matrix $X(400\times 10)$, then establish the PCA statistical model. Because the cumulative percent variance of the 4 former principal component is 87.6521, choose the number of principal component $k = 4$. According to the formulas of the statistics control limits, the control limits of corresponding statistics are: $d_T^2 = 11.0125(99\%)$, $d_T^2 = 9.1442(95\%)$, $d_{SPE}^2 = 5.9211(99\%)$ and $d_{SPE}^2 = 4.7057(95\%)$.

Choose the data changed to analyze for the first example. The data is picking 100 samples when the feed-in concentration of solute $X_F$ is changing from 0.02 kg solute / kg solution to 0.025 kg solute / kg solution. Use the conventional PCA first and the Fig. 1 shows the monitoring results of $T^2$ and $SPE$. It indicates process changed from $T^2$ plot while no change in $SPE$ plot, that matches with the fact. However, it can not determine the causes of change from Fig. 1. Because as pointed out in the introduction, the monitoring results in Fig. 1 may also appear when the fault of process occurred does not significantly change the relationship between the monitoring variables. Therefore, the point of view of process changing that is the monitoring results corresponding to case (2) which viewed by conventional PCA is unreliable.

The detection results by using improved PCA is shown in Fig.2. In Fig.2, it shows that neither $PVR$ nor $CVR$ is detected significant change. Because both the $PVR$ and $T^2$ plots reflect the information of $PV$ variable, the monitoring of $PVR$ plot further determines there's no change for residual of $PV$ variable. While the monitoring of $CVR$ plot indicates there's no change of $CV$ variable, so the change of $T^2$ plot can be considered as caused by the change of normal process and there is no fault during the process.

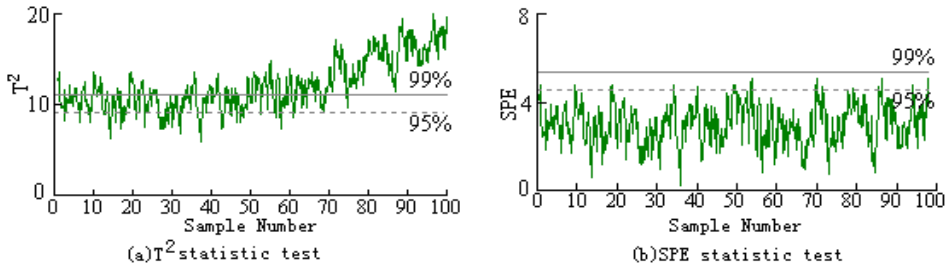


Fig. 1 $T^2$ and $SPE$ plots for case 1

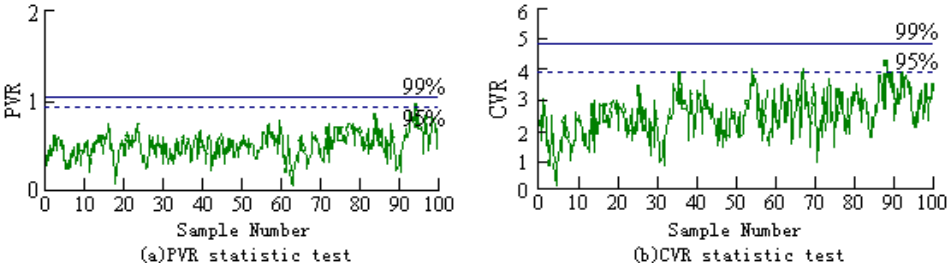


Fig. 2 $PVR$ and $CVR$ plots for case 1

## Conclusions

The improved PCA use the two new statistics $PVR$ and $CVR$. The process change and the fault can be distinguished by overall consideration of the change in $T^2$ and $PVR$ statistics due to $T^2$ and $PVR$ reflect all information of the process variables those closely related to principal component.

Compared with the conventional PCA, the effectiveness of improved PCA can be proved by the simulation monitoring for double-effect evaporation.

## Acknowledgements

The work is supported by the National Natural Science Fund Program of China (61174123).

## References

[1] Yajun Wang, Mingxing Jia and Zhizhong Mao: Journal of Industrial and Engineering Chemistry Vol. 21(1) (2015), p. 328-337

[2] Junichi Mori, Jie Yu: Journal of Process Control Vol. 24(1) (2014), p. 57-71

[3] M. A. Bin Shams, H. M. Budman and T. A. Duever: Chemical Engineering Science Vol. 66(20) (2011), p. 4488-4498

[4] Chunhui Zhao, Youxian Sun: Control Engineering Practice Vol. 21(10) (2013), p. 1396-1409

[5] Gertler J., W. Li, Y. Huang and T. McAvoy: AIChE Journal Vol. 45 (1999), p. 323-334

[6] Yves Roggo, Pascal Chalus, Lene Maurer, Carmen Lema-Martinez, Aurélie Edmond, Nadine Jent: Journal of Pharmaceutical and Biomedical Analysis Vol. 44(3) (2007), p. 683-700

[7] Ku W., Storer R. H. and Georgakis C: Chemometrics and Intelligent Laboratory Systems Vol. 30 (2003), p. 179-196