Fault diagnosis Based on An Improved Statistics Principal Analysis Similarity Factor

Hongxu Yin, Yuguang Wu, Zhijun Mu, Hongbin Geng, Xiaobo Li
Dezhou Power Supply Company, State Grid Shandong Electric Power Company, Dezhou shandong

Abstract—To settle the issue of conventional principal component analysis similarity factor cannot take advantage of higher-order statistics information of process variables, we presented an improved statistics principal analysis similarity factor (SPASF) to identify fault patterns in our work. In the improved SPASF approach, process data is first converted from original space into a new statistics space by the means of statistics pattern analysis (SPA) technology, and principal component analysis (PCA) is then employed to extract principal components in statistics space. At last, the pattern of snapshot dataset is recognized by measuring the similarity of principal components derived from statistics snapshot dataset and statistics historical fault dataset. The effectiveness of suggested SPASF based approach is verified through a case study on continuous stirring tank reactor (CSTR) by the means of recognizing fault patterns of snapshot datasets.

1. INTRODUCTION
With the wide application of computer control system, modern industrial production process tends to be much larger and more complicated. How to use the process fault diagnosis technology to ensure the safe and reliable operation of industry processes has become an important topic in the field of industrial process control [1,2].Principal component analysis (PCA) [3-5], as a data-driven based fault diagnosis method, has been widely studied and successfully applied to the study of fault detection in industry processes at present [6-8].

Although PCA based fault detection methods have achieved great success, identifying correct fault source for fault diagnosis has always been a problem in fault diagnosis research. The contribution plot [9-11] is simple and easy to use, but the variables in actual process often suffer from strong mutual correlation. Therefore, the variable with the largest contribution is not necessarily the real fault variable [12]. Compared with traditional contribution plot method, PCA similarity factor (PCASF) is a more effective fault identification approach. PCASF method was first used in literature [13], which measures the similarity of two datasets by comparing the angles between principal components of two PCA models. In consideration of the different data information described by each principal component direction, the corresponding eigenvalue is used to carry out weighting when calculating the angles between principal components in literature [14]. In order to make it easier for distinguishing different types of fault, a further improvement was made in PCASF method to calculate only the similarity between corresponding principal components in reference [15].

The existing PCASF method only uses low order statistics such as the mean and variance of non-Gaussian process data, and fails to use high-order statistics of process data to fully extract the feature information. Statistics pattern analysis (SPA) has been successfully applied to detect industry process faults [16-17] by using the high order statistics information of process data. SPA has the
capability of extracting process characteristic information more fully, which provides a new way to solve the problem of fault identification.

In this paper, SPA technology is adopted to improve the traditional PCASF method, and a fault diagnosis method based on an improved SPA similarity factor (SPASF) is proposed. The improved SPASF approach makes use of the high order statistics information of process data to extract the process features more fully, which enhances the fault identification effect. The fault diagnosis effectiveness of the improved SPASF is verified through a case study on continuous stirring tank reactor (CSTR) process.

2. THE BASIC PRINCIPLE OF SPA

Different from PCA, SPA monitors various statistics of process variables, which significantly reduces the influence of process nonlinearity and is more sensitive to the incipient faults [11].

2.1. Construct the statistics pattern vector

Let the matrix \( X \in \mathbb{R}^{n \times m} \) represent normal operating dataset in original space, where \( n \) is the number of samples and \( m \) indicates the number of variables. A data subset \( X_k \) is selected from the dataset \( X \) with the aid of a window whose width is set to be \( w \), where \( k \) is the index of sample. The form of data subset \( X_k \) is defined as:

\[
X_k = \begin{bmatrix}
    x_1(k) & x_1(k+w) & \cdots & x_1(k+w+l) \\
    x_1(k) & x_1(k) & \cdots & x_1(k) \\
    \vdots & \vdots & \ddots & \vdots \\
    x_1(k) & x_1(k) & \cdots & x_1(k)
\end{bmatrix}
\]

A statistical pattern (SP) is composed of various statistics of each original process variable, which is constructed in Eq. (2).

\[
S = [\mu | \Sigma | \Xi]
\]  

The \( \mu \) in Eq. (2) represents the first order statistics, which is set up by computing the mean value \( u_i \) of the \( i \)-th original process variable.

\[
u_i = \frac{1}{w} \sum_{l=0}^{w-1} x_i(k-l)
\]

The \( \Sigma \) is the second order statistics, which contains variance \( v_i \), correlation coefficient \( r_{i,j} \), autocorrelation coefficient \( r_{i,i}^d \) and the interrelationship coefficient \( r_{i,j}^d \) between different process variables.

\[
v_i = \frac{1}{w} \sum_{l=0}^{w-1} [x_i(k-l) - u_i]^2
\]

\[
r_{i,j} = \frac{\sum_{l=0}^{w-1} [x_i(k-l) - u_i] [x_j(k-l) - u_j]}{\sqrt{v_i v_j}}
\]

\[
r_{i,i}^d = \frac{\sum_{l=0}^{w-d-1} [x_i(k-l) - u_i] [x_i(k+d-l) - u_i]}{v_i}
\]

\[
r_{i,j}^d = \frac{\sum_{l=0}^{w-d} [x_i(k-l) - u_i] [x_j(k+d-l) - u_j]}{\sqrt{v_i v_j}}
\]

where \( d \) is the time lag between two process variables.

The \( \Xi \) denotes the high order statistics, which is composed of the third order statistic skewness \( \gamma_i \).
and the fourth order statistic kurtosis $\kappa_i$.

$$y_i = \frac{1}{w} \sum_{l=0}^{w-1} [x_i(k-l) - u_i]^3 / \left( \frac{1}{w} \sum_{l=0}^{w-1} [x_i(k-l) - u_i]^2 \right)^{3/2}$$

(8)

$$\kappa_i = \frac{1}{w} \sum_{l=0}^{w-1} [x_i(k-l) - u_i]^4 / \left( \frac{1}{w} \sum_{l=0}^{w-1} [x_i(k-l) - u_i]^2 \right)^2 - 3$$

(9)

After different statistics are calculated, all obtained statistics are arranged in a row vector to construct a statistical pattern (SP). The SP in a new window is computed by moving the window to make a new data subset replace the old data subset.

2.2. The fault detection based on SPA

The SPA based fault detection method consists of two steps: (1) in the original data space, the moving window technology is used to calculate various statistics of process variables to transform original process data into a statistic space; (2) the PCA method is then applied in the statistic space to establish fault detection model.

A series of SPs is calculated from the training data to establish a matrix $X_s \in R^{n_s \times m_s}$, where $n_s$ is the number of SPs and $m_s$ is the number of statistics. The matrix $X_s$ is first normalized to zero mean and unit variance, then PCA model is built as shown in Eq. (10).

$$X_s = TP^T + \tilde{X}_s$$

(10)

$$T = X_sP$$

(11)

where $T \in R^{n_s \times l_s}$ indicates the score matrix and $P \in R^{m_s \times l_s}$ is the loading matrix. The $l_s$ denotes the number of principal components and $\tilde{X}_s$ is the residual matrix.

By moving the window forward one or more samples in the test dataset, a new SP $x_i \in R^{m_s}$ can be calculated. Its $D_r$ and $D_p$ monitoring indices in residual space and score space are computed respectively.

$$D_r = r^T r, \quad r = (I - PP^T)x_i$$

(12)

$$D_p = x^T P \Sigma_l P^T x_i$$

(13)

where $\Sigma_l$ is a diagonal matrix containing the $l_s$ largest singular values.

The control limits of $D_r$ and $D_p$ indices are estimated based on the training dataset $X_s$.

$$\sigma^2 = \theta \left( \frac{c_n \sqrt{2\theta h_0}}{\theta} + 1 + \frac{\theta h_0 (h_0 - 1)}{\theta^2} \right)^{-1}$$

(14)

$$\theta = \sum_{i=1}^{m} (\sigma)^{1/2} = 1, 2, 3 \quad h_0 = 1 - \frac{2\theta^2}{3\theta^2}$$

(15)

where $\alpha$ is the given significance level and $c_n$ is the standard deviation corresponding to the $(1-\alpha)$ point of Gaussian distribution.

$$T^2 = \frac{l_i(n_i^2 - 1)}{n_i(n_i - 1)} F_{\alpha}(l_1, n_s - l_s)$$

(16)
3. THE IMPROVED SPASF METHOD FOR FAULT IDENTIFICATION

The basic idea of the improved SPASF method is as follows: (1) the various statistics of snapshot dataset and historical fault datasets are calculated to acquire the statistics matrices in a new statistics space; (2) the modified PCA similarity factor (PCASF) method is then used to calculate the similarities between SPs in the statistics snapshot matrix and SPs in the statistics historical fault matrices to determine the fault pattern.

The statistics snapshot matrix \( S_s \in \mathbb{R}^{n_s \times m} \) of snapshot dataset is given as follows:

\[
S_s = \begin{bmatrix}
s_{11} & s_{12} & \cdots & s_{1m} \\
s_{21} & s_{22} & \cdots & s_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
s_{n_s,1} & s_{n_s,2} & \cdots & s_{n_s,m}
\end{bmatrix}
\]

After the matrix SS is scaled to zero mean and unit variance, the singular value decomposition is performed in Eq. (17).

\[
\frac{1}{\sqrt{n_s - 1}} S_s = U \Sigma V^T
\]

where the load vectors corresponding to the \( k_1 \) largest singular values in matrix V describes at least 95% changes of the dataset SS.

The statistics historical fault matrix HS is derived from historical fault dataset H and the matrix HS is also normalized. The \( k_2 \) principal components holding at least 95% variations of the dataset HS are also retained.

The parameter \( k \) is chosen as \( \max(k_1, k_2) \), then the first \( k \) principal components of datasets SS and HS can both describe at least 95% data variations. The corresponding score matrices of SS and HS are denoted as L and M, respectively.

The SPASF based method measures the similarity between the two score matrices L and M, which is defined as:

\[
S_{SPA} = \text{trace}(L^T M M^T L) / k
\]

The similarity of snapshot dataset S and historical fault dataset H can be evaluated by the similarity factor SSPA because the \( k \) important principal components of statistics datasets SS and HS are included in the score matrices L and M, respectively. The SPASF SSPA is within the interval of \([0,1]\). When SSPA is close to 1, the two datasets S and H are considered as the same fault pattern; when SSPA is close to 0, the two datasets S and H are deemed to be different fault patterns.

Considering that the data variation in each principal component direction is different, the cosine values of angles between the principal components are weighted by the square of the corresponding singular value. In order to make it much easier to distinguish different fault patterns, only the similarities between corresponding principal elements are calculated. Therefore, the improved SPASF \( S_{SPA}^M \) between statistics matrix SS of snapshot dataset S and statistics matrix HS of historical fault dataset H is formulated as:

\[
S_{SPA}^M = \frac{\sum_{i=1}^{k} (\lambda_i^L \lambda_i^M) m_i^L m_i^M}{\sum_{i=1}^{k} (\lambda_i^L \lambda_i^M)} = \frac{\sum_{i=1}^{k} (\lambda_i^L \lambda_i^M) \cos^2 \theta_i}{\sum_{i=1}^{k} (\lambda_i^L \lambda_i^M)}
\]

where vectors \( l_i \) and \( m_i \) denote the i-th principal component in score matrices L and M, respectively. \( \theta_i \) is the angle between the vectors \( l_i \) and \( m_i \). \( \lambda_i^L \) and \( \lambda_i^M \) indicate the square of the i-th singular value for score matrices L and M.
The improved SPASF $S_{SPAS}^M$ is indeed to calculate the mean of cosine values of the angles between corresponding principal components in statistics matrices SS and HS. Hence, the similarity factor $S_{SPAS}^M$ is also within the interval of [0,1]. For the same fault pattern, the value of $S_{SPAS}^M$ is close to 1; for different fault patterns, the value of $S_{SPAS}^M$ is close to 0.

4. A CASE STUDY ON THE CSTR PROCESS

In this paper, the simulation study on a continuous stirred tank reactor (CSTR) process [18] is performed to testify the effectiveness of the improved SPASF based fault diagnosis method. In the simulation of CSTR process, Gaussian measurement noise is added and 10,000 samples under normal operating conditions are collected as training dataset. Five fault patterns F1, F2, F3, F4, F5 listed in Fig. 1 are introduced to the CSTR process, and 10000 fault samples are collected for each fault pattern. In each sampling process, the fault is added at the 5000-th sample. The historical fault datasets F1H, F2H, F3H, F4H, F5H are constructed with the recorded five kinds of fault samples.

| Fault pattern | The description of fault |
|---------------|--------------------------|
| F1            | The activation energy slopes up |
| F2            | The heat transfer coefficient slopes down |
| F3            | The temperature sensor in reactor exists bias |
| F4            | The feed flow has a step change |
| F5            | The feed concentration slopes up or down |

During fault detection and identification, the number of retained principal components is determined according to the 95% variation cumulative contribution rate. The threshold values of Dp and Dr monitoring indices are computed by feat of 99% confidence upper limit. The time delay d between different variables is chosen to be 2 for calculating the autocorrelations and mutual correlations.

The fault F4 is utilized to illustrate the fault detection performance of SPA based method. For SPA, the window width is set as 50 samples and the window moves forward 50 samples each time in training dataset while the window width is also selected as 50 samples and the window moves forward only one sample each time in test dataset. Fig. 1 shows the monitoring results of SPA based approach. It can be found that both SPA Dp and Dr indices detect fault F4 at the 4951-th window simultaneously, which confirms the fault detection effect of SPA based method.
After the fault F4 is detected by SPA, the improved SPASF based approach is adopted to recognize its fault pattern. The snapshot dataset F4S of fault F4 is constructed using 5000 fault samples from the 5001-th sample to the 10000-th sample in the test dataset. The window width is chosen as 150 samples and the window moves forward 50 samples each time in both the snapshot dataset F4S and the historical fault dataset F4H. The similarity factors $S_{PCA}$, $S_{SP4}$ and $S_{SP4}^M$ of PCASF, modified PCASF, SPASF and improved SPASF based approaches are calculated to identify the pattern of snapshot dataset F4S, which are listed in Table 2.

### Table 2. Four similarity factors for snapshot dataset F4S.

| Similarity factor | F1H | F2H | F3H | F4H | F5H |
|-------------------|-----|-----|-----|-----|-----|
| $S_{PCA}$         | 0.9311 | 0.8905 | 0.998 | 1 | 0.999 |
| $S_{SP4}$         | 0.6884 | 0.6099 | 0.6498 | 1 | 0.7196 |
| $S_{PCA}^M$       | 0.0334 | 0.2724 | 0.4808 | 1 | 0.0231 |
| $S_{SP4}^M$       | 0.0105 | 0.04 | 0.1551 | 1 | 0.0215 |

The values of similarity factors $S_{PCA}$, $S_{SP4}$, $S_{PCA}^M$ and $S_{SP4}^M$ for dataset F4S are represented by histograms in Fig. 2, Fig.3, Fig. 4 and Fig. 5 for an intuitionistic comparison.
As can be seen from Table 2, when fault identification is carried out by using similarity factors $S_{PCA}$ and $S_{SPA}$, the values of $S_{SPA}$ and $S_{SPA}$ between snapshot dataset F4S and historical fault dataset F4H are the largest. However, the values of $S_{SPA}$ and $S_{SPA}$ between dataset F4S and the other historical fault datasets F1H, F2H, F3H and F5H are also large, which leads to that the fault type cannot be accurately identified. Although the fault identification effect of $S_{MPCA}$ is improved compared with that of $S_{SPA}$ and $S_{PCA}$, it is still not as good as that of $S_{MSPA}$ and it needs to be further improved. It can be clearly seen from the histograms in Fig. 5, Fig. 6 and Fig. 7 that there is little difference in the values of $S_{PCA}$, $S_{SPA}$ and $S_{MPCA}$ for different fault patterns, which is not conducive to the accurate fault identification.
When the similarity factor $S^M_{SPAS}$ is utilized to recognize fault pattern of snapshot dataset $F4S$, the value of $S^M_{SPAS}$ between datasets $F4S$ and $F4H$ is the largest, which is computed as 1. Meanwhile, the values of $S^M_{SPAS}$ between dataset $F4S$ and the other four historical fault datasets are very small, which are all less than 0.16. In this way, the fault pattern of snapshot dataset $F4S$ can be clearly distinguished and the root of the fault can be located correctly. The histogram in Fig.8 clearly shows that the dataset $F4S$ has the greatest similarity with fault pattern $F4$ while the similarities with the other four fault patterns are very small.

In the simulation, the fault amplitude of fault $F4$ starts from 25% of the nominal value and increases by 5% in turn until the fault amplitude reaches 125% of the nominal value [9]. A total of 21 simulation experiments are conducted. The success rates of similarity factors $S^M_{PCA}$, $S^M_{SPAS}$, $S^M_{PCAS}$ and $S^M_{SPAS}$ for recognizing fault pattern of snapshot dataset $F4S$ are given in Table 3.

### Table 3. The Success Rates of Four Similarity Factors for Identifying Snapshot Dataset $F4S$.

| parameter  | $S^M_{PCA}$ | $S^M_{SPAS}$ | $S^M_{PCAS}$ | $S^M_{SPAS}$ |
|------------|-------------|--------------|--------------|--------------|
| Success rate | 52.38% | 71.43% | 66.67% | 100% |

According to Table 3, by changing the fault amplitude of fault $F4$ to conduct multiple simulation experiments, the similarity factor $S^M_{SPAS}$ can accurately identify fault pattern of dataset $F4S$ in each simulation with the 100% success rate. The fault diagnosis effect of the improved SPASF $S^M_{SPAS}$ is obviously better than the other three similarity factors.

5. CONCLUSIONS

To solve the problem of PCASF based methods that they cannot utilize the higher order statistics information of process data, an improved SPASF approach is developed by combining the SPA technology with the modified PCASF method. In the improved SPASF method, the original process data is first processed using SPA technology to calculate various statistics process variables, and then the modified PCASF method is utilized to compute the similarities of statistics snapshot dataset and statistics historical fault datasets in the statistics space for the purpose of recognizing fault pattern of snapshot dataset. The simulation results on the CSTR process show that the improved SPASF method has better fault identification performance than the existing PCASF based methods.

**REFERENCES**

[1] Liang J, Qian J. Multivariate statistical process monitoring and control: recent developments and applications to chemical industry. Chin. J. Chem. Eng., 2003,11(2): 191-203.

[2] Ye L, Shi X, Liang J. A multi-level approach for complex fault isolation based on structured residuals. Chin. J. Chem. Eng., 2011,19 (3): 462-472.

[3] Li H, Xiao D Y. Survey on data driven fault diagnosis methods. Control and Decision, 2011, 26(1): 1-9.

[4] Shen Y, Ding S, Haghani A. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process. Journal of Process Control, 2012, 22: 1567-1581.

[5] Qin S J. Survey on data-driven industrial process monitoring and diagnosis. Annual Reviews in Control, 2012, 36(2): 220-234.

[6] BinShams M A, Budman H M, Duever T A. Fault detection, identification and diagnosis using CUSUM based PCA. Chemical Engineering Science, 2011, 66(20): 4488-4498.
[7] Yu J B. Local and global principal component analysis for process monitoring. Journal of Process Control, 2012, 22: 1358-1373.
[8] Li Yuan, Tang Xiao chu. Improved performance of fault detection based on selection of the optimal number of principal components. Acta Automatic Sinica, 2009, 35(12): 1550-1557.
[9] Macgregor J F, Jacckle C. Kipanissides C, Koutodi M. Process monitoring and diagnosis by multiblock PLS methods. AIChE Journal, 1994, 40: 826-838.
[10] Miller P, Swanson R E. Contribution plots: a missing link in multivariate quality control. Appl. Math. Comput. Sci., 1998, 8: 775-792.
[11] Kourtii T, MacGregor J F. Multivariate SPC methods for process and product monitoring. Qual. Tech., 1996, 28: 409-428.
[12] Qin S J. Statistical process monitoring: basics and beyond. J. Chemom., 2003, 17(8): 480-502.
[13] Krzanowski W J. Between-groups comparisons of principal components. Journal of the American Statistical Association, 1979, 74(367): 703-707.
[14] Singhal A, Seborg D E. Matching patterns from historical data using PCA and distance similarity factors[C]. Proceedings of American Control Conference, Arlington, 2001: 1759-1763.
[15] Deng X G, Tian X M. Fault diagnosis method based on PCA similarity factor. Journal of Jilin University (Engineering and Technology Edition), 2004, 34: 207-210.
[16] Wang J, He Q P. Multivariate statistical process monitoring based on statistics pattern analysis. Ind. Eng. Res., 2010, 49(17): 7858-7869.
[17] He Q P, Wang J. Statistics pattern analysis: a new process monitoring framework and its application to semiconductor batch processes. AIChE Journal, 2010, 49(17): 7858-7869.
[18] Deng X G, Tian X M. Multivariate statistical process monitoring using multi-scale kernel principal component analysis[C]. Fault Detection, Supervision and Safety of Technical Processes, Beijing, China, 2006: 108-113.