Data-driven based Fault Diagnosis using Principal Component Analysis

Shakir M. Shaikh[1], Imtiaz A. Halepoto[2], Nazar H. Phulpoto[3], Muhammad S. Memon[2], Ayaz Hussain[4], Asif A. Laghari[5]

[1]Department of Control Science and Engineering, HIT Harbin, China
[2]Department of Computer Systems Engineering, QUEST Nawabshah, Pakistan
[3]Department of Information Technology, QUEST Nawabshah, Pakistan
[4]Department of Electrical Engineering, BUETK, Khuzdar, Pakistan
[5]School of Computer Science & Technology, HIT Harbin, China

***Abstract*—Modern industrial systems are growing day by day and unlikely their complexity is also increasing. On the other hand, the design and operations have become a key focus of the researchers in order to improve the production system. To cope up with these chellenges, the data-driven technique like principal component analysis (PCA) is famous to assist the working systems. A data in bulk quanitity from the sensor measurements are often available in such industrial systems. Considering the modern industrial systems and their economic benifits, the fault diagnostic techniqes have been deeply studied. For example, the techniques that consider the process data as the key element. In this paper, the faults have been detected with the data-driven approach using PCA. In particular, the faults have been detected by using T^2 and Q statistics. In this process, PCA projects large data into smaller dimensions. Additionally it also preserves all the important information of process. In order to understand the impact of the technique, Tennessee Eastman chemical plant is considerd for the performance evaluation.**

Keywords*—Fault Diagnosis; Principal Component Analysis; Multivariate Statistical Approach; Tennessee Eastman Chemical Plant Introduction

I. Introduction

Industrial process managemen is one the key and emering issue in the small as well large industrial systems. Modern industrial services are in large-scale and they are extremely complex. In addition, the control of process management is carried out with a great number of parameters under the system. In the industries like manufacturing, there is a pressure to produce excellence in end-products, which is in bulk quantit. In parallel, it is also important to minimize the losses of rejection rates and to satisfy the ecological rules. To fulfill the demands, up-to-date industrial systems cover a huge number of parameters working under closed-loop controllers. For that, a data-driven design of system is one of the great interest both in research and academia. Engineering systems such as aircraft controllers, industrial processes, manufacturing systems, transportation systems, electric and electronic system are becoming more complicated to lead the failure. It will directly related to the system reliability, availability, safety and maintainability. One the other hand, such factor are very important for a good industrial system. Many of such systems rely on human efforts and the availablity. In order to improve the performance and industrial systems it is necessary to work on the automation. It reduces the human efforts and the cost so it effects the economic conditions. Nowadays, the demand for the automated systems is also increasing in the market. In this research area, there is need to study different operating constraint and applications of industrial automation that explore and elaborate the process of automation. In automated systems in it necesssary to implement techniques and policies for the fault diagnosis and repair. The fault diagnosis system tries to assure that the plant is safe by identifying unwanted events. It highlights the key issue that may degrade the overall performacne of the system. This information is necessary for plant engineer so that a quick action may be performed. So that an immidieate rescue could be performed for the safety of the industrial system. There are many techniques are available for the performance monitoring and control. PCA is one of the basic technique in the pool of famous techniques.

In Section II related work is discussed. Methology has been discussed in Section III in which implementation of techniques is done stepwise. In Section IV, PCA technique is applied in the industrial benchmark process. In Section V results have been discussed. Section VI concludes the work and provides the guidelines for the future.

II. Related Work

Multivariate statistical methods largely depend upon the huge quantity of past data to define the fluctuations in the process. Multivariate statistical process monitoring has the advantage of easy to design and make the analysis of process industries entire simple, due to this property it is most popular in industrial fault diagnosis systems while in detecting the abnormal operation from the process. The technique which has the capability to retain the major information and significant knowledge in a unique dataset that generates from the industrial process is PCA. There are many approaches are available for the fault diagnosis. These apporoaches use different parameters in order to detect the faults in the control systems. For example, the study in [1] highlited the fault diagnosis based on the neural netowrks. They combined such neural networks by an observer technique. Also, multivariate statistical approaches have been researched to deal with process monitoring [2]. PCA is first appeared in 1889 until now research is ongoing and applications are still in study.

These methods have been successfully implemented in many industrial processes. MacGregor implemented PCA based process monitoring both in continuous and batch process and conclude PCA methods are capable of treating processes with a large correlated process data and can handle easily missing data [1]. Raich and Cinar [3] proposed a diagnosis method based on angle discriminant using PCA. Due to some difficulties, its applicability is not so large instead of that it fits in so many fields and successfully implemented. Though PCA could affect if it is applied in nonlinear problems because real systems are mostly nonlinear in nature, and this technique takes account linear combination due to its linear method. Kramer [4] has generalized PCA to the nonlinear case by using a neural network. These chemical engineering applications, are mostly nonlinear but the method is linear. Application of PCA in the real system have been applied at Dupont and other companies, published in many conferences and journals several types of researches have performed similar case work on data collected simulator of process [5][6]. For sake of easiness many dimensions of dataset proposed to get more from data in different views and plot in single dimension [7], on taking this step that will helps the operator to get information from more than multidimensional data [8]. In some cases multidimensional data acquire is quite difficult due to nonlinearities so an automation process proposed for process monitoring in [9]. The application of PCA in these type of problems motivated by three features. Number 1, PCA can develop a method which takes all the data in low dimension which helps out to get meaning from entire data from the training set by use of all dimensional data. Number 2 the data in structured format with help of PCA help to identifying the affected variables. Number 3, PCA can isolated the space which the variables contain useful information that variables have process information and rest in another subspace which contain noise. In this fault could occur in any subspace primarily [10], this step can increases the sensitivity of process monitoring to detect faults. As an effective data-driven process monitoring technique PCA can adapt complicated conditions according to rules of statistics. It is classical projection methods of multivariate statistical process monitoring which is then applied to train model beneath nominal conditions. Thus it detects online faults [11]. These techniques are highly demanded based on measurements [12]. In practical industrial outliers that is difficult to handle in spite of so many advantages and easy to design model [13]. Sensor failure, network transmission error, machine malfunction, database software, and data recording errors are mainly cause for irregularities produce data with noise [14]. Such cases outliers smoothed by mean and averaging [10].

III. Methodology

PCA is one of the famous and widely used technique. It has been effectively used in various areas including image processing, signal analysis, pattern recognition, data compression and process monitoring. This techniques is simple and efficient and have capability to process industrial data. It is familiar as influential tool for process monitoring. For this purpose, it is used in the process industry for process monitoring. PCA algorithm is a founding technique of automated process monitoring. These advanced PCA methods for example recursive, adaptive and kernel. These techniques extract the useful information from data in keeping view this it is a widely used area in fault diagnosis. It extracts orthogonal vectors in sets, known as loading vectors. It tells the amount of variance known to orthogonal vectors. Consider a process measurement matrix $X \in \mathbb{R}^{n\times m}$, where m is variables and n is observations in measurement matrix.

A. PCA based Fault Detection

Step 1: Pretreatment of data is done in this step. Normalize columns of X

Step 2: Obtain the Covariance of measurement matrix by

$$C = \frac{1}{n-1} X^T X \tag{1}$$

Step 3*:* In this step, the loading vector is extracted by obtaining the Singular Value Decomposition (SVD) from the above equation:

$$C = \frac{1}{n-1} X^T X = U\Sigma V^T \tag{2}$$

Where $\Lambda = diag(\lambda_1 \geq \cdots \geq \lambda_m \geq 0)$

Step 4: To calculate PCs (principal component) a divide V into the score and residual matrices.

$$\Lambda = \begin{bmatrix} \Lambda_{pc} & 0 \\ 0 & \Lambda_{res} \end{bmatrix}$$

$$\Lambda_{pc} = diag(\lambda_1, \dots \lambda_a) \quad \Lambda_{res} = diag(\lambda_{a+1}, \lambda_{a+2}, \dots \lambda_m$$

$$V = \begin{bmatrix} V_{pc} V_{res} \end{bmatrix} \quad V_{pc} \in \Re^{m\times a} \quad V_{res} \in \Re^{m\times(m-a)}$$

The value of L can be taken from V_{pc}:

$$L = V_{pc} \tag{3}$$

In the above equation V_{res} is the residual space.

Step 5: To obtain the matrix T, the following equation is used:

$$T = XL \tag{1-4}$$

B. PCA and the Faults

In the process, T^2 can be obtained bu the following equation:

$$T^2 = x^T L \Lambda_{pc}^{-1} L^T x \tag{4}$$

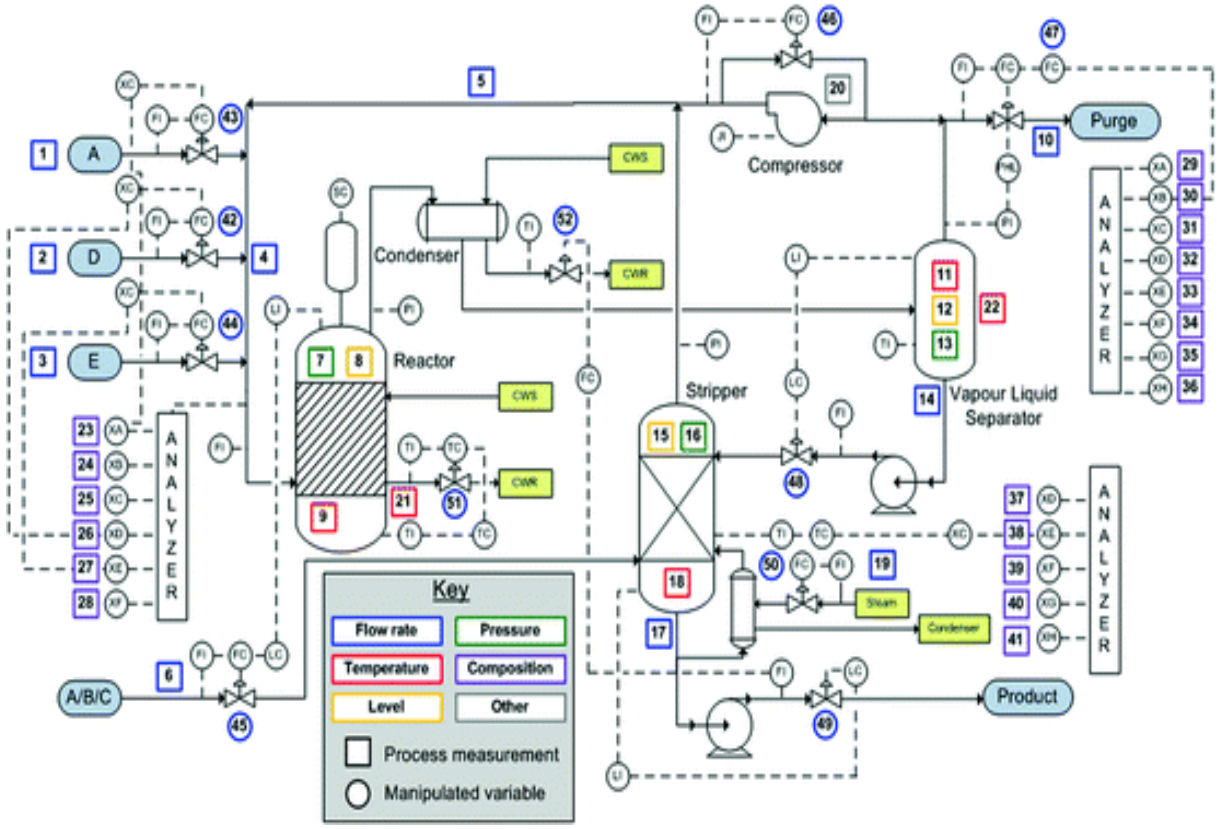

Fig. 1. Tennessee Eastman Process.

L represent the set of loading vectors for large singular variance. The simplified equation will be:

$$j_{th,}T^2 = \frac{a(n-1)(n+1)}{n(n-1)} F_\alpha\left(a, n-a\right) \tag{5}$$

Where $F_\alpha(a, n-a)$ is the f-distribution. If the threshold value from equation no 1-5 exceeds the test statistics in equation no: 1-6 fault occurs. As T^2 statistics is demonstrated on the basis of loading vectors singular values, so it does have a problem to some inaccuracies [15] in the residual part values. The square prediction error is then used which utilizes the residual space [16].

$$Q = x^T V_{res} V_{res}^T x \tag{6}$$

Q −Statistics threshold is achieved by:

$$j_{th}, SPE = \left(\frac{\theta_1 \left(h_0 c_\alpha \sqrt{2\theta_2} \right)}{\theta_1} + 1 + \frac{\theta_2 h_0 \left(h_0 - 1 \right)}{\theta_1^2} \right)^{1/h_0}$$

Where

$$\theta_i = \sum_j^n = a+1 \lambda_j^{2i} \quad \text{and} \quad h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \tag{7}$$

Where c_α represents the standard deviation of the distribution parallel to the $(1-\alpha)$ percentile. Hence, the confidence level for the Q may be determined with the help of equation (7) in order to cope with abnormalities.

IV. The Benchmark Process

In this part to verify the algorithm like PCA, simulation is carried out in order to diagnosis the faults. It is computer-oriented simulator many types of research used it for comparison of different data-driven algorithms as well as model-based. It is like realistic simulator which mimic the original behavior of the typical chemical plant. TEP served as the desired benchmark to assessment algorithms for many techniques such as process observing control and fault diagnosis. TEP utilized to examine multivariate statistical process monitoring (MSPM) methods. It facilitates with different operating regimes. Figure 1 presents the flow of the process. A number of connected modules are present. These include a condenser, reactor, separator, stripper, and compressor. It consists with four in number of reactants or input and two products or output, along with by-product, and an inert by symbolically represented as A to H. There are 52 measurements by each process. Among them, 41 are the output variables and the rest are input variables. The output and input variables are depicted in Table 1[8]. Further, the process variables (from XMV(1) to XMV(11)) are used with the standard values (as in accordance with [8]). The researchers in [17] performed the simulation of TE process. There has been successfully worked done in these techniques

reported in. Training data set produced for working on this which includes both faulty and normal data set. The work in [18] proposed a control scheme which is thus applied in TEP. Flowchart of an industrial plant is demonstrated in figure 1. The gaseous components are presented as A, C, E, D and the inert B. It first feeds to the reactor. Formation of G and H component as a product from these inputs feed into a reactor. A simulator has been developed details can be found in [19]. Following equations are input-output relations of the process.

$$A(k)+C(k)+D(k)\rightarrow G(liq)$$
$$A(k)+C(k)+D(k)\rightarrow H(liq) \quad (8)$$
$$A(k)+C(k)+D(k)\rightarrow H(liq)$$
$$D(k)\rightarrow F(liq)$$

TABLE 1. LIST OF PROCESS VARIABLES AND MEASURED VARIABLES [8]

Tag	Description
XMV(1)	D feed flow
XMV(2)	E feed flow
XMV(3)	A feed flow
XMV(4)	A and C feed flow
XMV(6)	Purge Valve
XMV(7)	Separator pot liquid flow
XMV(8)	Stripper liquid product flow
XMV(10)	Reactor Cooling water flow
XMV(11)	Condenser Cooling water flow
XMEASV(1)	A feed (Stream 1)
XMEASV(2)	D feed (Stream 2)
XMEASV(3)	E feed (Stream 3)
XMEASV(4)	A and C feed
XMEASV(5)	Recycle flow
XMEASV(6)	Reactor feed rate
XMEASV(7)	Reactor Pressure
XMEASV(8)	Reactor level
XMEASV(9)	Reactor temperature
XMEASV(10)	Purge rate
XMEASV(11)	Separator temperature
XMEASV(12)	Separator level
XMEASV(13)	Separator pressure
XMEASV(14)	Separator underflow
XMEASV(15)	Stripper level
XMEASV(16)	Stripper pressure
XMEASV(17)	Stripper underflow
XMEASV(18)	Stripper temperature
XMEASV(19)	Stripper steam flow
XMEASV(20)	Compressor work
XMEASV(21)	Reactor water temperature
XMEASV(22)	Separator water temperature

In this equation the Component F is a by-product, that process is exothermic and cannot reversible. They are in first-order with concentrations higher temperature happened due to fastest reaction of component G over reaction of H component. Separator has vapors from reaction which is recycled again and again that is input to the compressor. The stream generated from the process keep for the use of by product and inert. Stripper "stream 10" is driven by a separator which is condensed.

C. Main Process Variables

The process has 41 calculated variables and 11 input variables. They are listed in Table 1. Out of which 22 variables which sample every three minutes. There are 21 process faults as described in the work [17].

D. Simulated Faults in TEP

There are 21 faults in TEP process listed in Table 1. These faults affect mostly in chemical process parameters like process variables, kinetics, feed concentration and different types of actuators in the chemical process like pump valves. Data-driven approaches require online and offline training data, in this simulator there are 22 online test data sets, their duration for 48 hours plant operation and 960 samples of data are generated during simulations in [5], where faults are added after 160 data sample.

V. RESULTS AND DISCUSSION

There are a number of faults as depicted in the Table 2. All the values of Table 1 and 2 are directly taken from [8]. One of the first steps is to implement the PCA in order to detect such configured faults. The default dataset from the TE simulator is used for the experimentation and evaluation. The input and output variables are configured. Such as, XMEAS(1-22) and XMV(1-11). In real life the reason for the faults is unknown, similarly the TE also introduce faults at different subspaces.

According to Table 2, the fault in the condenser cooling water inlet temperature is represented by IDV2. It occurs after 160 data samples. In figure 2 PCA of IDV 2 is shown. IDV (6) involves in step type of fault, it is simulated by a sudden change in the reactor cooling water inlet temperature. The results of PCA diagnosis the affected variable from "A" feed loss in TE process are shown in figure 3. The algorithm for process monitoring is developed with the help of 960 samples taken from the ordinary process operations. Similarly, IDV (18) is unknown type of fault listed in Table 2, affect the process variable, it influence on the unknown variable. PCA-based statistics identifying the unknown variable shown in figure 4.

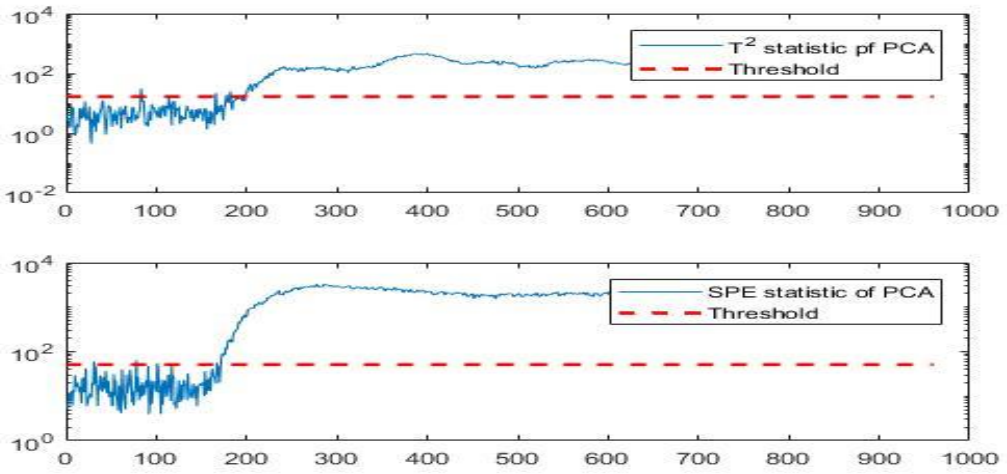

Fig. 2. Detection Result for Fault Scenario 2 PCA.

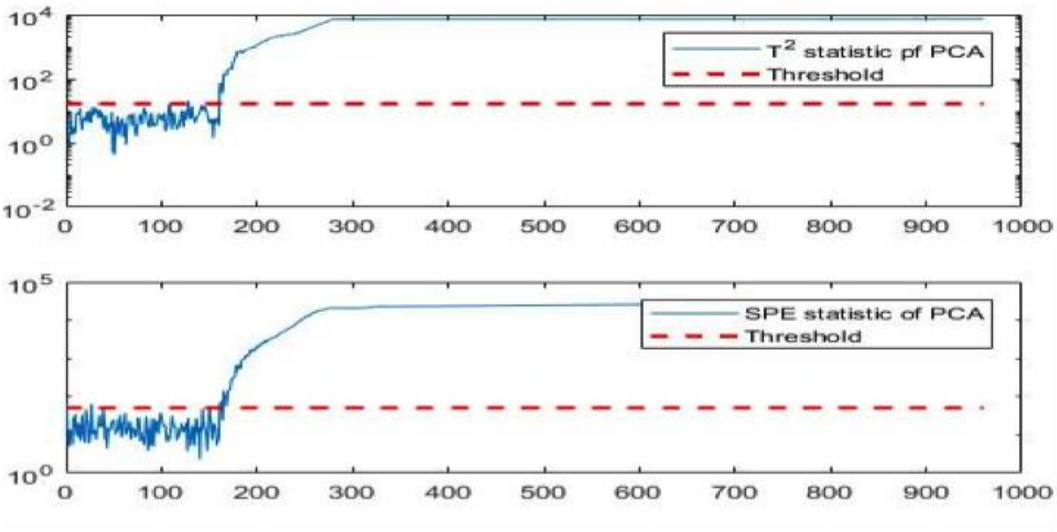

Fig. 3. Detection Result for Fault Scenario 6 PCA.

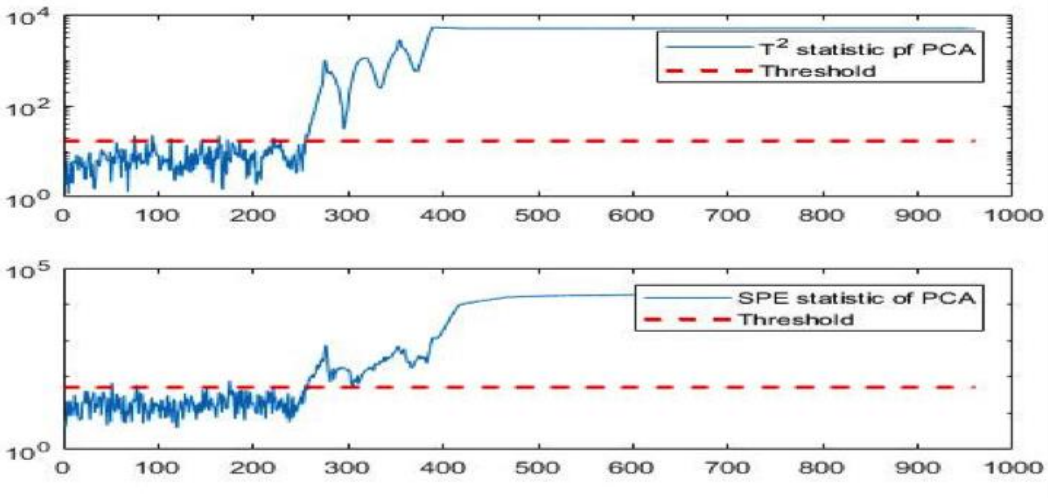

Fig. 4. Detection Result for Fault Scenario 18 PCA.

TABLE 2. DEFINITION OF FAULTS [8]

Fault number	Process Variable	Type
IDV(1)	A/C feed ratio, B composition constant	Step
IDV(2)	B composition, A/C ration constant	Step
IDV(3)	D feed temperature	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss	Step
IDV(7)	C header pressure loss-reduced availability	Step
IDV(8)	A, B C feed composition	Random variation
IDV(9)	D feed temperature	Random variation
IDV(10)	C feed temperature	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water valve	Random variation
IDV(13)	Reaction Kinetics	Slow Drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	--	--
IDV(17)	--	--
IDV(18)	--	--
IDV(19)	--	--
IDV(20)	--	--
IDV(21)	Steady state position	Fixed

VI. CONCLUSION

The fault diagnosis is very important for the optimized systems. Specifically, the data-driven techniques are getting famous. This work presents a detailed study on the data-driven technique. The design of a data-driven technique using the PCA is proposed. PCA is simple efficient and easy to design due to these properties it is frequently used in fault diagnosis techniques. The industrial benchmark, Tennessee Eastman process is used for the simulation and analysis of the data-driven technique, which successfully detects the faults.

The major objective of further investigation is to analysis of non-Gaussian process data, since in industries mostly system are non-linear in nature. In this work it is assumed that data which are under consideration is Gaussian. A framework should be established that will directly constructed from process data for construction of fault tolerant architecture.

REFERENCES

[1] ZHOU D, LI G, QIN S J. Total projection to latent structures for process monitoring[J]. AIChE Journal, Wiley Online Library, 2010, 56(1): 168–178.

[2] DE JONG S. SIMPLS: an alternative approach to partial least squares regression[J]. Chemometrics and intelligent laboratory systems, Elsevier, 1993, 18(3): 251–263.

[3] LI G, LIU B, QIN S J . Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: The dynamic T-PLS approach[J]. IEEE transactions on neural networks, IEEE, 2011, 22(12): 2262–2271.

[4] YIN S, ZHU X, KAYNAK O. Improved PLS focused on key-performance-indicator-related fault diagnosis[J]. IEEE Transactions on Industrial Electronics, IEEE, 2015, 62(3): 1651–1658.

[5] CHIANG L H, PELL R J, SEASHOLTZ M B. Exploring process data with the use of robust outlier detection algorithms[J]. Journal of Process Control, Elsevier, 2003, 13(5): 437–449.

[6] FRANK P M. Analytical and qualitative model-based fault diagnosis–a survey and some new results[J]. European Journal of control, Elsevier, 1996, 2(1): 6–28.

[7] YIN S, DING S X, HAGHANI A . A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process[J]. Journal of Process Control, Elsevier, 2012, 22(9): 1567–1581.

[8] STEVEN X. Data-driven Design of Fault Diagnosis and Fault-tolerant Control Systems [M]. SPRINGER, 2014.

[9] AGRAWAL V, PANIGRAHI B K, SUBBARAO P M V. Review of control and fault diagnosis methods applied to coal mills[J]. Journal of Process Control, Elsevier, 2015, 32: 138–153.

[10] TSCHUMITSCHEW K, KLAWONN F. Incremental quantile estimation[J]. Evolving Systems, Springer, 2010, 1(4): 253–264.

[11] TRACY N D, YOUNG J C, MASON R L. Multivariate control charts for individual observations[J]. Journal of quality technology, Taylor & Francis, 1992, 24(2): 88–95.

[12] BRITTO R da S. Detecção de falhas com PCA e PLS aplicados a uma planta didática[J]. Pós-Graduação em Engenharia Elétrica, 2014.

[13] BISHOP C M. Pattern recognition and machine learning (information science and statistics)[J]. Springer-Verlag New York, Inc., Secaucus, NJ, 2006.

[14] CHEN T, MORRIS J, MARTIN E. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), Wiley Online Library, 2006, 55(5): 699–715.

[15] CHEN Z, ZHANG K, DING S X . Improved canonical correlation analysis-based fault detection methods for industrial processes[J]. Journal of Process Control, Elsevier, 2016, 41: 26–34.

[16] CHEN Z, DING S X, ZHANG K . Canonical correlation analysis-based fault detection methods with application to alumina evaporation process[J]. Control Engineering Practice, Elsevier, 2016, 46: 51–58.

[17] WISE B M, GALLAGHER N B. The process chemometrics approach to process monitoring and fault detection[J]. Journal of Process Control, Elsevier, 1996, 6(6): 329–348.

[18] PIOVOSO M J, KOSANOVICH K A, PEARSON R K. Monitoring process performance in real-time[C]//American Control Conference, 1992. IEEE, 1992: 2359–2363.

[19] Chen, Zhiwen, et al. "Improved canonical correlation analysis-based fault detection methods for industrial processes." Journal of Process Control 41 (2016): 26-34.