A Nonstationary Process Monitoring Based on Mutual Information among Process Variables

Cheng Ji, Jingde Wang*, Wei Sun*

Beijing University of Chemical Technology, North Third Ring

Road 15, Chaoyang District, Beijing, 100029, China

Abstract: In practical chemical industrial processes, the feed valve is automatically adjusted in response to the control systems and the production load will be adjusted with market situation and administrative regulations. Therefore, process data display nonstationary statistics in practical operation condition and cannot satisfy the ideal assumptions of traditional multivariate statistical methods that process is assumed operating around one preset steady state. Under normal operating conditions, fluctuations or adjustments will only affect the mean and standard deviation of process variables, but the correlation among process variables should follow its inherent mechanism model, whose feature can be statistically captured within certain range. In this paper, a nonstationary process monitoring based on mutual information among process variables is proposed. The Euclidean distance (ED) of eigenvalues of the mutual information matrix under normal operation conditions is calculated to obtain a statistic. Once a fault occurs, the changes in correlation among process variables will be reflected in the mutual information matrix and corresponding ED will exceed the threshold, by which process monitoring can be implemented. A numerical simulation example and a practical cracking process are applied as case studies. The results show a better performance on monitoring nonstationary process than traditional principle component analysis method.

Keywords: Mutual information, Correlation Analysis, Euclidean Distance, Fault detection, Cracking Furnace

1. INTRODUCTION

A good alarm system is essential in chemical industrial processes because abnormal process deviation may lead to significant loss in economic effectiveness and process safety (Qin S., 2003). At present, most process monitoring methods in practical industrial processes still rely on process mechanism or human experience (Tian W., et al., 2020). But the increasing scale and complexity of chemical equipment make it difficult to obtain sufficient knowledge for establishing an efficient process monitoring. With the rapid development of computer technology and artificial intelligence, data-driven methods, which apply multivariate statistics or machine learning methods to detect process deviation in industrial process, have attracted great attention (Qin S., 2012).

Most abnormal process deviation can be detected by multivariate statistics methods with historical data under a given steady operating condition. Among them, Principal Component Analysis (PCA) is the most commonly used method for its advantages in dimension reduction and feature extraction, and has been successfully applied to practical processes monitoring. However, in process industry, the production load fluctuates in response to the control systems and is adjusted with market situation and administrative regulations. Process data display different distribution characteristics, and therefore cannot satisfy the ideal assumptions in traditional multivariate statistics methods that process is assumed operating around one preset steady state (Cinar., et al., 2007), which make it quite difficult to identify abnormal process deviation.

In order to solve this problem, Han et al, proposed a condition recognizer and combined it with PCA method (Han X., et al., 2018). She found that the PCA projection orientations of different data sets under normal operation conditions are very close, and the same PCA loading matrix can be used in different operational conditions by adjusting the normalization center. However, the real-time identification of current data condition will lead to a high computational load, and a time delay will exist in the condition recognizer using moving windows, which results in false alarms in transitional states.

The false alarms can be avoided if the common nonstationary feature under normal operating conditions can be extracted. The fluctuations and adjustments of product load will only result in change of mean and standard deviation of data, but the mutual information among process variables displays a unique feature. When a process deviation occurs, the changes in correlation among process variables will reflect in the statistical feature of mutual information matrix. Based on this idea, a process monitoring method for nonstationary process is proposed. Mutual information matrix in each window of process data is calculated in this work to characterize the correlation among process variables. To implement the fault detection, a statistic is selected as the Euclidean distance (ED) between the eigenvalues of mutual information matrix from historical data under normal operation conditions and that from real-time data.

The following parts of this paper are arranged as follows, the preliminaries of this work are briefly introduced in Section 2. The main idea and the implement procedure of the proposed method are explained in Section 3. In Section 4, the proposed method is applied to a numerical simulation example and a practical cracking process. The process monitoring results are shown and discussed. In Section 5, the paper is concluded.

2. METHODOLOGY

2.1 Mutual Information

Mutual information is a concept of information theory derived from Shannon entropy, which is a measurement of the uncertainty of a random variable proposed by Shannon (Shannon C E, 1948). Considering two-dimensional random variables X, Y, the mutual information between X and Y can be calculated as follows,

$$I(X,Y)=\sum_{xy} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ is the joint probability of variables X and Y. According to this formula, if X and Y are independent of each other, the joint probability of variables X and Y is equals to the product of the probability of variables X and the probability of variables Y, which means $I\ (x,y)$ is equal to 0, otherwise $I(x,y)$ is positive. Because the calculation of mutual information is based on the estimation of probability density, the nonlinear correlation can be also extracted, making it more suitable to be applied to nonlinear systems.

2.2 Kernel Density Estimation

The most important and difficult part in mutual information is the estimation of joint probability density. Histogram is the most commonly used method for probability density estimation. However, the results can be greatly affected by the width of the bins, resulting in a large estimation error (Bauer M., et al., 2006). In this work, kernel density estimation (KDE) method is applied. In KDE, a kernel function is used to fit the observed data points to estimate the true probability distribution. The calculation of KDE is as follows (Silverman B. W., 1986),

$$P(x)=\frac{1}{n}\sum_{i=1}^{n} K(x-x_i) \tag{2}$$

where $P(x)$ is the probability density of x, n is sample size, $K(x)$ is the kernel function. Generally, Gaussian function is selected as the kernel function. The joint probability density can be estimated in the same way,

$$P(x,y)=\frac{1}{n}\sum_{i=1}^{n} K\left(x-x_i, y-y_i\right) \tag{3}$$

where $P(x,y)$ is the joint probability density, n is sample size, $K(x,y)$ is the kernel function.

2.3 Euclidean Distance

ED is a commonly used definition of distance, and represents the true distance between two points in high-dimensional space. It can reflect the differences in absolute value between sequences (Gao X, et al., 2020), which is appropriate to measure the differences in the statistical feature of the mutual information matrix in this work.

Considering two sets of data, the mutual information matrix I_1, I_2 can be obtained using formula 1, the eigenvalues can be calculated by eigenvalue decomposition,

$$I_1 = P_1 \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} P_1^T \tag{4}$$

$$I_2 = P_2 \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} P_2^T \tag{5}$$

where P_1, P_2 are the eigenvectors and λ_i, σ_i are eigenvalues of the mutual information matrix.

The ED between the eigenvalues of the two matrix $Dis(\lambda, \sigma)$ can be calculated as follows,

$$Dis(\lambda,\sigma)=\sqrt{\sum_{i=1}^{n}\left(\lambda_i-\sigma_i\right)^2} \tag{6}$$

where n is the dimension of the data, $Dis(\lambda, \sigma)$ can represent the difference of correlation among process variables between the two data sets, which can be used to determine whether the two data sets are in a same operating condition.

3. THE PROPOSED PROCESS MONITORING METHOD

The purpose of this work is to detect the abnormal deviation in the nonstationary process with normal fluctuations or adjustments in production load. The main idea and implement procedure of the proposed method are presented in this section.

3.1 The Proposed Process Monitoring Method

As mentioned before, correlation among process variables should follow its inherent mechanism model under normal operation conditions. Therefore, the eigenvalues of mutual information matrix in each window of data under normal conditions displays a similar characteristic. The ED of the vector of the eigenvalues between two different windows should be a very small value.

To obtain a control limit, The ED of the vector between all pair of windows is calculated to obtain sufficient samples. The significance level is selected as a three-sigma threshold. When an abnormal process deviation occurs, the changes in correlation among process variables will be reflected in the mutual information matrix. ED between the current vector of the eigenvalues and that from normal operation conditions will increase and exceed its threshold, indicating that the abnormal deviation is detected.

3.2 The Procedure of the Proposed Method

The flowchart of the proposed process monitoring method is shown in Figure 1. The procedure can be divided to the following two parts:

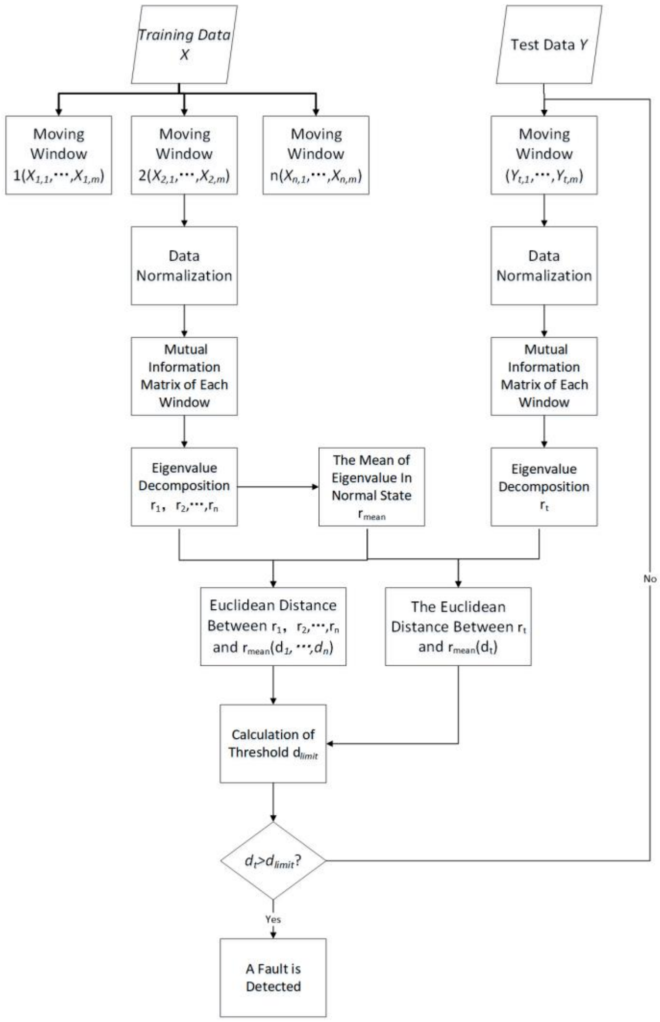

Fig. 1. The implement procedure of the proposed method.

Offline training:

The first step is to divide the training data into moving windows and normalize data in each window. Then correlation among process variables is calculated to obtain mutual information matrix of each window. The eigenvalues of each mutual information matrix are obtained by eigenvalue decomposition. Eigenvalues from each matrix are represented as a vector. The mean of all the vectors is calculated as the standard vector under normal operating conditions. The ED between the vector in each matrix and the standard vector is calculated to measure the difference of correlation among process variables between each window and the approximate standard window under normal operation conditions. A three-sigma threshold is chosen for the significance level.

Online monitoring:

For the online process monitoring, data in real-time moving window is normalized. The correlation among process variables under the operation condition of current window is calculated to obtain mutual information matrix. Eigenvalues of the matrix is obtained by eigenvalue decomposition. Calculate the ED between the vector of eigenvalues from real-time window and the standard one obtained from training data and compare the ED to the threshold.

If the ED is less than the threshold, the system is under a normal operation condition, otherwise an abnormal process deviation is identified if the ED exceeds the threshold.

4. CASE STUDIES

In this section, the proposed process monitoring method is applied to a numerical simulation example and an industrial cracking process.

4.1 Case Study on a Numerical Simulation

A simple two-dimension numerical simulation example is used to test the performance of the proposed process monitoring method. The training data is generated by the following equation,

$$y = \sqrt{x} + e_1 \tag{7}$$

where x is Gaussian distributed variable with a mean of 45 and a standard deviation of 0.2, y another variable that has a nonlinear relationship with x. e_1 is a zero-mean white noise with a standard deviation of 0.01.

In the simulation, 500 sampling points are generated as training data to obtain the mean of eigenvalues of mutual information matrix from data under normal operation conditions and the threshold of the ED. To illustrate the performance of the proposed monitoring on nonstationary process, test data are generated by the following two equations,

$$y = \sqrt{x} + e_2 \tag{8}$$

$$y = \log_2 x + e_2 \tag{9}$$

where e_1 is zero-mean white noises with a standard deviation of 0.01, e_2 is a zero-mean white noise with a standard deviation of 0.04.

The test sample includes 1500 data points. The first 500 samples are generated by Equation 7. In order to reflect the non-stationary characteristics of data, the standard deviation of the noise increase to 0.04 from the 500th sample point. A fault is introduced by completely changing the correlation between x and y in Equation 9 from the 1001st to the end. The test data are shown in Figure 2.

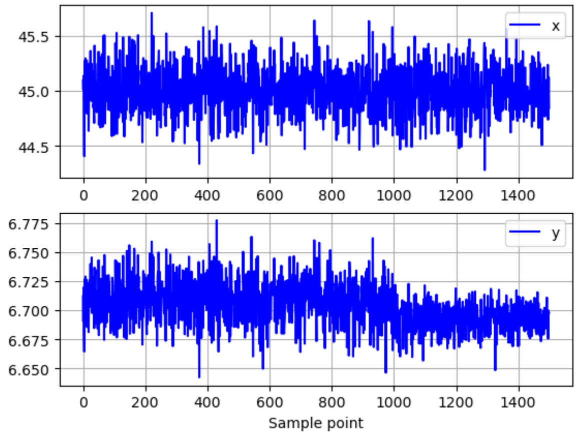

Fig. 2. Test data of the numerical simulation.

The conventional PCA-based monitoring method is first carried out to detect the fault. The PCA model is built using the training data generated by Equation 7. The corresponding T^2 control charts under 99% confidence level in the test data are shown in Figure 3. It is obvious that the value of T^2 statistic is below the control limit line. The fault cannot be detected by PCA method because there are no significant changes in the mean and standard deviation of data, but the correlation of the variables is completely changed.

The results using the proposed monitoring method is performed in Figure 4. Considering both the calculation error and computational loads, the length of moving windows is selected as 100 in this work. It can be observed that the ED index value remains below the control limit line for all of the initial 1000 samples, which means that the normal fluctuation will not lead to false alarm in this method. After the 1000th sample point, the ED increases quickly to exceed the threshold and stays above the threshold until the 1500th sample point.

It can be concluded that the proposed monitoring methods are more sensitive to detect the slight variation of the correlation among variables, but not sensitive to the variation of the mean or standard deviation of data. Therefore, the proposed method can be applied to monitor nonstationary process with fluctuations and adjustments in production load.

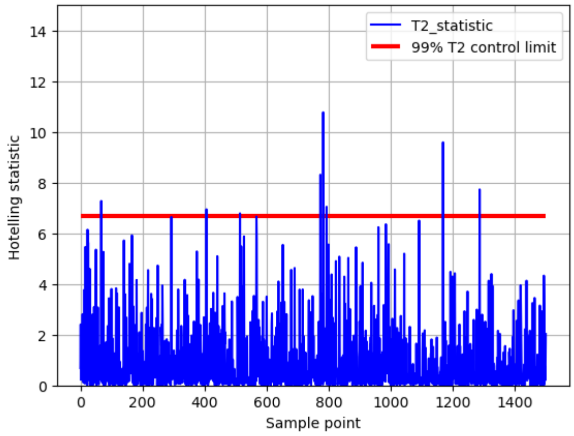

Fig. 3. Process monitoring result based on Principal component analysis.

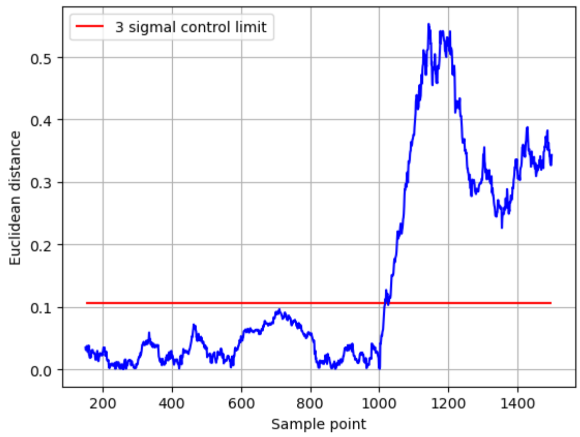

Fig. 4. Process monitoring result based on the proposed method.

4.2 Case Study on an industrial Cracking Process

In this section, an industrial naphtha cracking process is applied as a case study. Ethylene from naphtha cracking is a key product in chemical industry and plays an important role in the national economy. Naphtha cracking furnace is the key equipment in ethylene production process. The structure of the furnace is shown in Figure 5. In the furnace, coil outlet temperature (COT) must be controlled strictly because the product quality can be easily influenced by COT. Therefore, it is important to establish a process monitoring method on the cracking furnace, especially on the temperature variable of COT. The process data are nonstationary because the feed flow of naphtha is frequently adjusted with the market situation and administrative regulations, which provides huge challenges for the detection of process faults.

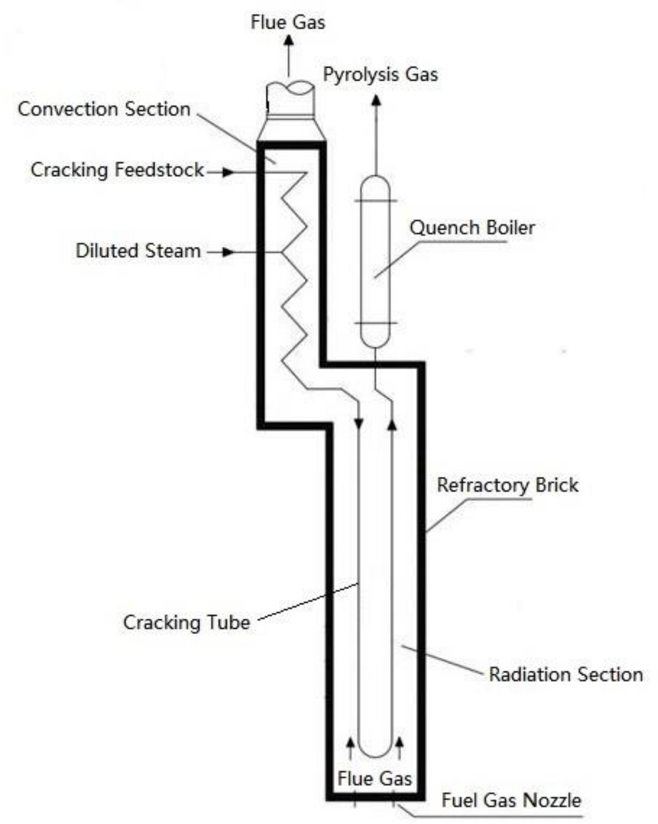

Fig. 5. Naphtha cracking furnace.

Next the proposed method is applied to the actual operation data of a cracking furnace. Totally 63 process variables are selected, which is shown in Table 1. Data with 1000 sample points under normal operation conditions are used as training data, and test data are shown in Figure 6. It can be noticed that the feed flow of naphtha is adjusted, and there is a step fault in COT at the 395th sample point. This deviation will lead to huge influence on product quality, therefore it is important to detect the deviation in time.

The monitoring results of PCA method and the proposed method are shown in Figure 7, Figure 8 and Figure 9. It should be noticed in Figure 9 that the result is shown from the 100th sample of the test data because of the moving window applied in proposed method. Although the process deviation has been identified in PCA method, both T^2 statistic and SPE statistic exceed the control limit line when the feed flow of naphtha is adjusted at the first 200 sample points, resulting in a high false alarm rate. In contrast, the adjustment of the operation conditions is identified as normal operation in the proposed method, and when COT steps, the ED increases quickly to exceed the threshold, indicating that the system is under an abnormal condition. By comparison, the proposed method has a better performance on monitoring nonstationary processes.

Table 1. Variables in cracking furnace

Variables	Description	Quantity
F1	Naphtha mass flow	7
F2	Diluted steam mass flow	7
QL	Total fuel gas calorific value	1
P1	Feed pressure	1
P2	Cross section pressure	6
P3	Outlet pressure	6
TA	Temperature in furnace A side	1
TB	Temperature in furnace B side	1
T1	Feed temperature	1
T2	Diluted steam temperature	1
T3	Cross section temperature	6
COT	Coil outlet temperature	25

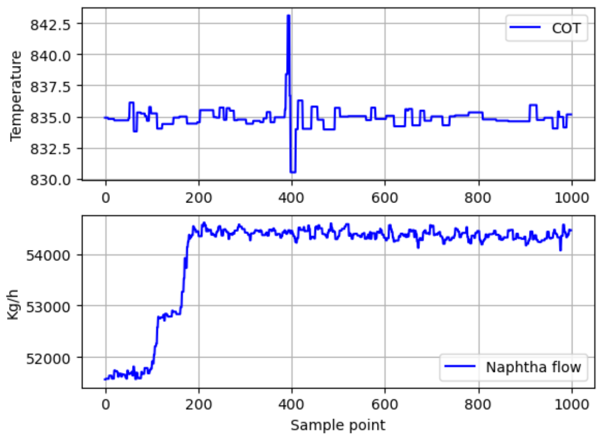

Fig. 6. Test data from naphtha cracking process.

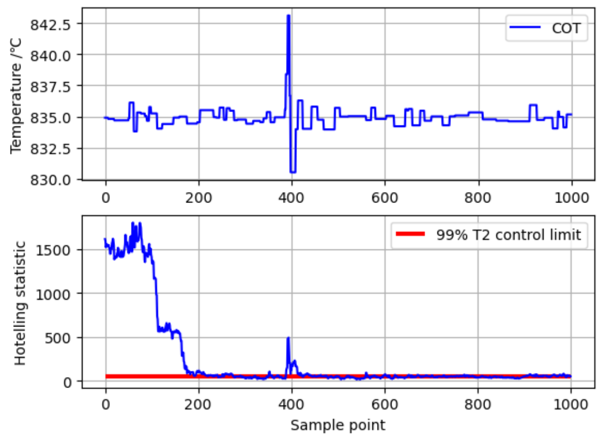

Fig. 7. Process monitoring result based on Principal component analysis (T^2 statistic).

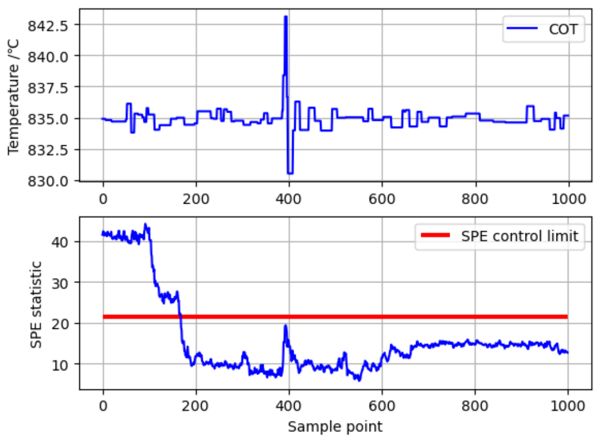

Fig. 8. Process monitoring result based on Principal component analysis (SPE statistic).

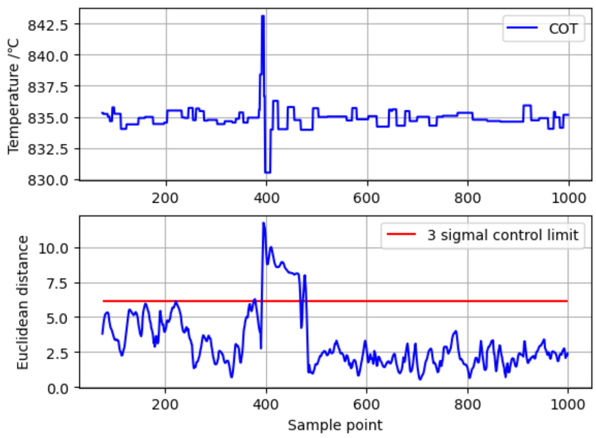

Fig. 9. Process monitoring result based on the proposed method.

5. CONCLUSIONS

In this work, a process monitoring method for a nonstationary process is proposed based on mutual information among process variables. It can be concluded that the correlation among process variables will not be significantly affected by normal fluctuations or adjustments in production load. By the proposed method, process monitoring on a numerical simulation and an industrial naphtha cracking process is implemented. The results show that the abnormal process deviations can be effectively identified, which is consistent with the process operation record. The method provides a brand-new way to detect the abnormal process deviation in specific nonstationary processes with fluctuations or adjustments in production load.

REFERENCES

Bauer, M., Cox, J. W., Caveness, M. H., Downs, J. J., and Thornhill, N. F. (2006). Finding the direction of disturbance propagation in a chemical process using transfer entropy. *IEEE transactions on control systems technology*, 15(1), 12-21.

Cinar, A., Palazoglu, A., and Kayihan, F. (2007). *Chemical process performance evaluation*. CRC press.

Gao, X., Yang, F., and Feng, E. (2020). A process fault diagnosis method using multi-time scale dynamic feature extraction based on convolutional neural network. *The Canadian Journal of Chemical Engineering*, 98(6), 1280-1292.

Han, X., Tian, S., Romagnoli, J. A., Li, H., and Sun, W. (2018). PCA-SDG based process monitoring and fault diagnosis: application to an industrial pyrolysis furnace. *IFAC-PapersOnLine*, 51(18), 482-487.

Joe Qin, S. (2003). Statistical process monitoring: basics and beyond. Journal of Chemometrics: A Journal of the Chemometrics Society, 17(8-9), 480-502.

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *ANNUAL REVIEWS IN CONTROL*, 36(2), 220-234.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379-423.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Tian, W., Liu, Z., Li, L., Zhang, S., and Li, C. (2020). Identification of abnormal conditions in high-dimensional chemical process based on feature selection and deep learning. *Chinese Journal of Chemical Engineering*, 28(7), 1875-1883.