Fault Detection and Isolation of Non-Gaussian and Nonlinear Processes Based on Statistics Pattern Analysis and the $k$-Nearest Neighbor Method

Zhe Zhou, Jian Wang, Chunjie Yang, Chenglin Wen, and Zuxin Li*

ABSTRACT: Only low-order information of process data (i.e., mean, variance, and covariance) was considered in the principal component analysis (PCA)-based process monitoring method. Consequently, it cannot deal with continuous processes with strong dynamics, nonlinearity, and non-Gaussianity. To this aim, the statistics pattern analysis (SPA)-based process monitoring method achieves better monitoring results by extracting higher-order statistics (HOS) of the process variables. However, the extracted statistics do not strictly follow a Gaussian distribution, making the estimated control limits in Hotelling-$T^2$ and squared prediction error (SPE) charts inaccurate, resulting in unsatisfactory monitoring performance. In order to solve this problem, this paper presents a novel process monitoring method using SPA and the $k$-nearest neighbor algorithm. In the proposed method, first, the statistics of process variables are calculated through SPA. Then, the $k$-nearest neighbor (kNN) method is used to monitor the extracted statistics. The kNN method only uses the paired distance of samples to perform fault detection. It has no strict requirements for data distribution. Hence, the proposed method can overcome the problems caused by the non-Gaussianity and nonlinearity of statistics. In addition, the potential of the proposed method in early fault detection or safety alarm and fault isolation is explored. The proposed method can isolate which variable or its statistic is faulty. Finally, the numerical examples and Tennessee Eastman benchmark process illustrate the effectiveness of the proposed method.

INTRODUCTION

With the rapid development of industrial processes and the increase in process complexity, a large amount of process operation data has been generated. Fully mining and using the valuable information contained in the process data will be beneficial to early fault detection, thereby minimizing downtime and improving the safety of process operation. In this context, data-driven multivariate statistical process monitoring (MSPM) methods have been developed by leaps and bounds, where principal component analysis (PCA) methods are the most widely used. However, only low-order information of process data (i.e., mean, variance, and covariance) was considered, and high-order statistics (HOS) were ignored in the PCA-based process monitoring method. In addition, the threshold of Hotelling’s $T^2$ and SPE are calculated based on the premise that process variables satisfy a Gaussian distribution. Due to nonlinearity, non-Gaussianity, dynamic, and multimodality in industrial processes, it is difficult to satisfy this assumption in practice. Therefore, the traditional PCA-based process monitoring method has poor monitoring performance when facing the above problems.

Recently, there have been many studies on nonlinear process monitoring problems. To capture the nonlinear correlation structure, Yu et al. proposed a robust, nonlinear and sparse PCA method. Aiming at the excessive modeling redundancy problem caused by infinite-order mapping, a monitoring method based on constructing polynomial mapping is proposed in ref, which significantly improves the fault detection performance of nonlinear processes. Jiang and Yan proposed a parallel PCA-KPCA monitoring method for a process with linearly correlated and nonlinearly related variables. In order to detect the incipient faults of nonlinear industrial processes effectively, an enhanced KPCA method is proposed in ref 11. Mansouri et al. introduced a generalized likelihood ratio test into the KPCA method for fault detection of nonlinear processes. However,
these methods above do not consider the high-order statistics of the process data, and the calculation is complicated.

In addition, there is a lot of work to solve the problem of fault diagnosis of complex industrial processes using hybrid methods including Bayesian and data-driven approaches. Yu et al.\textsuperscript{13} proposed a two-stage fault diagnosis method that combines independent component analysis (ICA) with the Bayesian network (BN) to solve the problem in that conventional MSPM methods cannot isolate fault from unmonitored process variables. In order to reduce the cost of monitoring and alarm flooding, Wang et al.\textsuperscript{14} proposed a fault diagnosis technique that combines semi-parametric PCA and BN, which also achieves location of the root cause of faults in process variables. A PCA-BN hybrid method with multiple likelihood evidence for fault diagnosis is proposed in ref 15. This method enables BN to update more information about faults and improves process monitoring performance. Reference 16 uses KPCA and BN for fault diagnosis, which not only diagnoses the root cause of the fault but also shows the propagation path of the fault.

He and Wang\textsuperscript{17} proposed a statistics pattern analysis (SPA) monitoring framework to overcome these problems. This method extracts HOS from process data, such as skewness and kurtosis. Compared with the traditional MSPM methods that only use low-order statistics information, it has advantages in characterizing the non-Gaussianity and nonlinearity of the process data. They used the SPA to monitor the batch process because it calculates the statistics of batch process data, and data preprocessing is avoided. In ref 2, the sliding window is used to design the SPA to monitor the continuous process and obtain better monitoring performance on the Tennessee Eastman (TE) platform. In addition, the SPA is identified as a novel generation of the SPM method and can handle the 4 V challenges of big data.\textsuperscript{18–21}

Although the SPA framework has many advantages in monitoring processes of nonlinearity or multimodality,\textsuperscript{22–25} it is worth noting that the statistics extracted by the SPA is not a Gaussian distribution, in addition, there is a nonlinear relationship between the extracted statistics. Therefore, it is not suitable to use PCA combined with $T^2$ and SPE charts to monitor the extracted statistics. Ma et al.\textsuperscript{26} considered the nonlinearity of statistics: they used the advantages of kernel PCA in processing nonlinear data and proposed a process monitoring method based on statistics KPCA (SKPCA). However, SKPCA only considers the nonlinear problem of statistics, and the non-Gaussianity of statistics is ignored.

In this paper, a new process monitoring method, which combines the superiority of SPA in extracting HOS information with kNN in processing the non-Gaussian and nonlinearity of data samples, is proposed to deal with the defects of those methods above. The fault detection method using the kNN rule (FD-kNN) only uses the distance between neighbors to perform fault detection; there is no restriction on the data distribution.\textsuperscript{27,28} The qualitative and quantitative analysis for the non-Gaussianity and nonlinearity of statistics are also presented. Meanwhile, the effect of window width on the non-Gaussianity of statistics is investigated. In addition, we explore the potential of the proposed method in fault diagnosis. Specifically, the variable contribution by kNN (VC$k$NN) is used to determine the contribution of different statistics to the detection index. kNN is a distribution-free method. The proposed method uses the kNN method to monitor process statistics to achieve fault detection and fault isolation. In terms of fault detection, compared with the conventional statistical techniques, it can overcome the non-Gaussian and nonlinear problems of process variables (statistics); in terms of fault isolation, compared to traditional contribution analysis methods, the proposed kNN-based variable contribution method is not affected by fault smearing. The experiments on numerical examples and TE processes illustrate the effectiveness of the proposed process monitoring method.

The contributions of the proposed method are as follows:

- The qualitative and quantitative analysis for the extracted statistics of process data is conducted and verifies that they are non-Gaussian and nonlinear.
- The statistics extracted from the process data are monitored by using the $k$-nearest neighbor method, which avoids the problems caused by non-linear, non-Gaussian characteristics of statistics.
- The employed variable contribution by kNN can indicate the statistic of the faulty variable, which may be useful for root cause analysis and has potential to perform incipient fault detection.

\section*{RELATED WORK}

The core foundation of the proposed method is related to the SPA and FD-kNN. This section will briefly review these two methods.

\textbf{SPA.} We use $Y_d$ to denote the data in a window of the original samples:

$$Y_d = [y_1 \ y_2 \ \cdots \ y_m]$$

$$= \begin{bmatrix}
\gamma_1(d-l+1) & \gamma_2(d-l+1) & \cdots & \gamma_m(d-l+1) \\
\gamma_1(d-l+2) & \gamma_2(d-l+2) & \cdots & \gamma_m(d-l+2) \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_1(d) & \gamma_2(d) & \cdots & \gamma_m(d)
\end{bmatrix}$$

where $l$ is the window width, $d$ is the last sampling time, and $m$ represents the number of variables.

The sampling data of each window can be used to calculate a statistics pattern (SP), which is composed of first-order statistics (FOS), second-order statistics (SOS), and HOS. As shown in eq 2

$$S \equiv [\mu^T \Sigma^T \Xi^T]$$

where $\mu$ denotes the FOS, i.e., variable means ($\mu_i$); $\Sigma$ denotes the SOS, which include variance ($\sigma_i$), correlation ($r_{ij}$), autocorrelation ($r_{ij}^2$), and cross-correlation ($r_{ij}^2$); $\Xi$ denotes the HOS, which is composed of skewness ($\gamma_i$) and kurtosis ($\kappa_i$). These statistics can be calculated as follows

$$\mu_i = \frac{1}{l} \sum_{t=0}^{l-1} \gamma_i(d-t)$$

$$\sigma_i = \frac{1}{l} \sum_{t=0}^{l-1} [\gamma_i(d-t) - \mu_i]^2$$

$$r_{ij} = \frac{1}{l} \sum_{t=0}^{l-1} [\gamma_i(d-t) - \mu_i][\gamma_j(d-t) - \mu_j]$$

$$\sqrt{\sigma_j}$$

18624

https://doi.org/10.1021/acs.omega.2c01279

ACS Omega 2022, 7, 18623–18637
and training samples are similar, but fault samples and training data are stacked into a matrix, denoted as SPs. The number of each window of process data, and all the SP vectors of training samples are obtained at current and the previous \( l \) sampling times. In ref 2, the SP matrix is then decomposed by PCA, and \( T^2 \) and SPE charts are established. For online monitoring, statistics are also extracted based on the data obtained at current and the previous \( (l-1) \) sampling times. After that, \( T^2 \) and SPE of extracted statistics are calculated to perform fault detection.

**FD-kNN.** The kNN rule was initially widely used in pattern classification. In December 2006, it was selected as one of the top ten classic algorithms in data mining. FD-kNN was first proposed by He and Wang.\(^{29}\) The main idea is to measure the difference between samples by distance; that is, normal samples and training samples are similar, but fault samples and training samples are significantly different.

- **Training phase (determine the detection control limit):**
  1. Use the Euclidean distance to get the kNNs of each training sample.
   \[
   d_{p,q} = \sqrt{(x_p - x_q)^2}, \quad p = 1, \ldots, n, \ q \neq p \tag{10}
   \]
  2. Calculate the distance statistic \( D_p^2 \).
   \[
   D_p^2 = \frac{1}{k} \sum_{q=1}^{k} d_{p,q}^2 \tag{11}
   \]
   where \( D_p^2 \) represents the average squared distance between the \( p \)th sample and its \( k \) neighbors, \( d_{p,q}^2 \) denotes the squared Euclidean distance between the \( p \)th sample and its \( q \)th nearest neighbor.
  3. Establish the control limit \( D_q^2 \) for fault detection. There are many ways to estimate \( D_q^2 \), such as the estimation using a noncentral chi-square distribution,\(^{30}\) kernel density estimation (KDE). The \((1 - \alpha)\)-empirical quartile \(^{30}\) of \( D_q^2 \) is used as the threshold in our proposed method.
   \[
   D_q^2 = D^2_{\lfloor n(1-\alpha) \rfloor} \tag{12}
   \]
   where \( D_q^2, i = 1, \ldots, n \) is the result of \( D_p^2 \) obtained by eq 12 in ascending order. \( \lfloor n(1-\alpha) \rfloor \) means discard the decimal part of \( n(1-\alpha) \) and keep the integer part.

- **Detection phase:**
  1. For a sample \( x \) to be tested, find its kNNs from the training set.
  2. Calculate \( D_x^2 \) between \( x \) and its \( k \) neighbors using eq 10.
  3. Compare \( D_x^2 \) with the threshold \( D_q^2 \). If \( D_x^2 > D_q^2 \), the process is considered abnormal. Otherwise, it is normal.

### Analysis for the Non-Gaussianity and Nonlinearity of Statistics

Although the PCA fault detection method based on SPA (SPCA) improves the monitoring performance by using HOS information to describe the complex characteristics of process data, the statistics extracted by SPA is not satisfactory by a Gaussian distribution. Also, the relationship between different statistics is nonlinear. Faced with the above problems, the SPCA is theoretically flawed.

In this section, the qualitative and quantitative analysis for the non-Gaussianity and nonlinearity of statistics are presented. In addition, the effect of window width on the non-Gaussianity of statistics is investigated.

#### Non-Gaussianity of Statistics.

1. **Qualitative analysis:** The central limit theorem (CLT) says that under rather general circumstances, the sum of independent normalized random variables is a normal distribution at the limit.\(^{31}\) It needs to be emphasized that CLT assumes that the samples are drawn independently and the number of samples (in a window) is sufficiently large. In practice, however, the wider the window’s width, the greater the delay in giving the monitoring results. Moreover, the quantity of samples is limited by the specific process. In addition, if the overlap rate of two adjacent windows is high, the basic assumption of independent samples will be violated. Therefore, the extracted statistics are difficult to meet the assumption of CLT, which makes them non-Gaussian.

2. **Quantitative analysis:** The non-Gaussianity of each variable can be evaluated by negative entropy (NE).\(^{31}\) A kind of robust estimation of NE\(^{\alpha}\) is employed to calculate the non-Gaussianity of statistics. NE is non-negative, and it is 0 only when a variable is the standard normal distribution. The larger the NE of a random variable, the more it deviates from the standard normal distribution. For variable \( x \), its NE can be expressed as follows:
   \[
   NE_x = G(v) - G(v^*) \tag{13}
   \]
   where \( G(v) \) is the entropy of the standard normal distributed variable \( v \), \( G(v^*) \) is the entropy of \( x^* \), while \( x^* \) is the normalized result of \( x \). The \( G(v) \) can be calculated as a constant:
   \[
   G(v) = \log \sqrt{2\pi} + \int_{-\infty}^{\infty} \frac{v^2}{2\sqrt{2\pi}} e^{-v^2/2} dv = \log \sqrt{2\pi} + \frac{1}{2} \tag{14}
   \]
   As for \( G(x^*) \), it can be calculated as follows:\(^{32}\)
where $x^{*}(1), x^{*}(2), \ldots, x^{*}(M)$ denote $M$ measurements of $x^*$ arranged in the ascending order. Substituting eqs 14 and 15 into eq 13 yields the NE of $x$. It is hard to achieve zero for NE$_x$ even if $x$ has a Gaussian distribution because NE$_x$ is affected by the measurements of $x$ and the number of measurements. According to ref 34, the value $10^{-3}$ is taken as the threshold.

The Nonlinearity of Statistics.

1) Qualitative analysis: If process variables have a nonlinear relationship, the statistics of these variables cannot eliminate the nonlinear relationship of original variables. In addition, it can be seen from the calculation formula of the statistics that there is also a nonlinear relationship between the statistics of different orders.

2) Quantitative analysis: We use the surrogate data (SD) method to detect the nonlinearity of statistics. This method first designates a given linear process as the null hypothesis, then generates SD sets consistent with the null hypothesis, and finally calculates a discriminating statistic for the original data and each SD set. If the value calculated for the original data is significantly different from the set of values calculated from the SD, then the null hypothesis is rejected, and nonlinearity is determined.

(1) The null hypothesis $H_0$: the data come from a linear Gaussian process.

(2) Generate surrogate data sets for original data using Fourier transform (FT).

a) Fourier transform: For variable $x$ with $N$ samples, its FT can be expressed as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{2\pi kn}{N}\right)$$  \hspace{1cm} (16)

b) Randomize the phase:

$$X_r(k) = X(k) \exp(i\phi_k)$$  \hspace{1cm} (17)
\[
\begin{align*}
q_0 &= q_{N+1/2} = 0 \\
q_k &= -q_{N-k}, \quad k = 1, 2, ..., \frac{N-1}{2}
\end{align*}
\]  
(18)

where \(q_k\) is a random variable and obeys a uniform distribution in the range of \([0, 2\pi]\).

c) Inverse Fourier transform: After the inverse Fourier transform, the substitute data \(x_{SD}(n)\) of the original data can be obtained.

\[
x_{SD}(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \exp \left( -\frac{2\pi i kn}{N} \right)
\]  
(19)

(3) Calculate the statistic \(T_k\) of a significance test.

\[
T_k = \frac{|R_k - \bar{R}|}{\sigma_k}
\]  
(20)

where \(R_k\) denotes the redundancy of the original data and \(\bar{R}\) and \(\sigma_k\) denote the mean and standard deviation of redundancy of the surrogate data set, respectively. Redundancy can be calculated using eq 21.38

\[
R(x, y) = \frac{1}{2} \sum_{i=1}^{2} \log c_{ij} - \frac{1}{2} \sum_{i=1}^{2} \log \lambda_i
\]  
(21)

The covariance matrix of variables \(x\) and \(y\) is denoted as \(C\). \(c_{ij}\) and \(\lambda_i\) are diagonal elements and eigenvalues of the covariance matrix \(C\), respectively.

(4) Determine whether the null hypothesis is true. Under the condition that the significance level is 0.05, the threshold of significance testing is 1.96. If \(T_k \geq 1.96\), the null hypothesis \(H_0\) is rejected and original data is nonlinear; otherwise, accept the null hypothesis and original data is linear.

**Illustrative Example.** A simple static nonlinear example is used to illustrate the non-Gaussianity and nonlinearity of statistics. The process model is as follows:

\[
\begin{align*}
x &= 2t + e \\
y &= x^2
\end{align*}
\]  
(22)

where \(t\) is a random variable that obeys a uniform distribution in the range of \([-0.5, 0.5]\) and \(e\) is the Gaussian noise with 0 mean and variance of 0.01. Ten thousand normal samples are generated by eq 22 for extracting statistics by SPA. The window width is 10 (i.e., \(l = 10\)) and the sliding step is 1. Each SP consists of 8 statistics, which include the FOS, SOS, and HOS of \(x\) and \(y\).

Figure 1 shows that the multivariate distribution of different statistics is non-Gaussian. The non-Gaussian of statistics is quantified by estimating the NE. It can be concluded from Table 1 that the NE of the statistics is almost always smaller than that of the original variable, but it is still greater than \(10^{-3}\). Therefore, the statistics extracted by SPA are indeed non-Gaussian.

The NE of the statistics can be affected by the window width and window shifting step. For a given number of statistics, the larger the window width, the smaller the NE of the statistics, that is, the weaker the non-Gaussianity, especially for HOS because the NE of the statistics is related to the number of the statistics determined by the window width and the window shifting step; therefore, as shown in Figure 2, there is an approximately monotonous decreasing trend between the NE (or non-Gaussianity) of the statistics and the window width. Figure 3a,b shows the nonlinearity detection results of the statistics (mean of \(x\) and mean of \(y\)); variance of \(x\) and kurtosis of \(x\) using the surrogate data method. It can be seen that \(T_k\) is always greater than 1.96. Therefore, the null hypothesis should be rejected with a probability of 95%: the statistics are nonlinear.

### PROCESS MONITORING METHOD USING THE SPA AND KNN ALGORITHM (SKNN)

**Fault Detection Based on SKNN.** The SKNN integrates the advantages of SPA in extracting HOS information and kNN in dealing with the non-Gaussian and nonlinearity of data samples.

- **Training step:**
  1. Extract the statistics matrix \(Y\)(SPs) from the training sets generated under the normal operation condition (NOC).
  2. Find kNNs for each sample (SP) in \(Y\) using eq 10.
  3. Calculate the distance statistic using eq 11.
  4. Determine the threshold \(D_{th}^2\) for fault detection using eq 12.

- **Online monitoring:**
  1. For a sample \(x\), a statistic sample \(s\) can be extracted by SPA.
  2. Find kNNs of \(s\) from \(Y\).
  3. Compare \(D_s^2\) with the threshold \(D_{th}^2\). If \(D_s^2 > D_{th}^2\), it is considered as a faulty sample. Otherwise, it is a normal sample.

**Fault Diagnosis Based on SKNN.** When a fault is detected, finding the root cause of the failure is the next more important thing. VCkNN\(^{27}\) is used to determine the contribution of different statistics to the detection index. The contribution of each statistic is not affected by other statistics; that is, there is no correlation. Therefore, the diagnosis result is not affected by the fault smearing.

- **Determine the contributions of statistics:**

\[
D_s^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} [c_{ij}^T (s - s_j)^2]
\]  
(23)

Equation 23 is the form of decomposing \(D_s^2\) into the sum of the contributions of \(m\) statistics.

\[
c_i^{knn} = \sum_{j=1}^{k} [c_{ij}^T (s - s_j)^2], \quad i = 1, ..., m
\]  
(24)

\[
\sum_{i=1}^{m} c_i^{knn} = D_s^2
\]  
(25)

where \(c_i^{knn}\) is the contribution from the \(i\)th statistic of \(s\) to \(D_s^2\).
Determine the faulty statistics:

\[ F = \{ i | \ell_{knn}^i > T_i \} \]  

where \( F \) contains all the faulty statistics diagnosed by SKNN and \( T_i \) represents the threshold for the \( i \)th statistic contribution. The threshold can be determined empirically; for example, the maximum contribution of the statistic in the training samples is used as the isolation threshold of the corresponding statistic. \(^{27}\)

**Remarks.**

1) The statistics can be flexibly selected for different purposes or process characteristics. For example, considering that the dynamics of the batch process are relatively weak, the only covariance is selected as the second-order statistics. \(^{17}\)

2) For the design of the window width and sliding step, the data set should be fully used. Otherwise, samples at the end of the data set will not be used. For example, if the size of the data set is 100 when the window width is 10 and the sliding step is 5, 19 SP are obtained exactly, and the data set is fully used; when the window width is 16 and the sliding step is 5, the last four samples of the data set will not be used resulting in the loss of key information of the process, making the model established in the training phase insufficient to describe the process accurately, thereby degrading the monitoring performance.

3) The number of nearest neighbors is selected, and \( k \) is determined using cross-validation.

### CASE STUDIES

In this section, we use numerical examples and TEP to explore the effectiveness of the proposed method SKNN in fault detection and diagnosis. The kernel-independent component analysis (KICA) \(^{39}\) method is also used to solve the non-
Gaussian and nonlinear problems of process data. Therefore, in addition to the SPCA and SKPCA methods, the monitoring performance of the fault detection method using SPA and KICA (SKICA) has also been verified.

**Numerical Simulations.** The NOC data are produced according to the following process model adopted from ref 4:

\[
x(k + 1) = \begin{bmatrix} 0.118 & -0.191 \\ 0.8470.264 \end{bmatrix} x(k) + \begin{bmatrix} 1.0 & 2.0 \\ 3.0 & -4.0 \end{bmatrix} o(k) + \begin{bmatrix} 0.193 \\ 0.320 \end{bmatrix} w(k - 1) + \begin{bmatrix} 0.689 \\ -0.749 \end{bmatrix} w(k - 1)
\]

(27)

\[y(k) = x(k) + o(k)\]

(28)

\[u^o(k) = \begin{bmatrix} 0.0811 & -0.226 \\ 0.4770.415 \end{bmatrix} u^o(k - 1) + \begin{bmatrix} 0.193 \\ -0.320 \end{bmatrix} w(k - 1)\]

(29)

\[u(k) = u^o(k) + v(k)\]

(30)

where \(w(k)\) is a noise with 0 mean and a standard deviation of 1, \(v(k)\) represents a noise with 0 mean and a variance of 0.1.

There are four variables (i.e., \(u_1, u_2, y_1, \) and \(y_2\)) used for monitoring. Since the process is dynamic, autocorrelation and cross-correlation are added to the second-order statistics. Therefore, a total of 54 statistics are extracted by SPA. For all four methods, the numbers of generated training samples and validation samples are 10,000 and 5000, respectively. There are 10,000 testing samples, of which the first 500 samples are normal, and the rest are faulty.

The fault setting is that the mean value of the first element \(w_1\) in \(w\) changes from 0 to 4.

The specific parameter settings of SPCA, SKPCA, SKICA, and SKNN are shown in Table 2. For SPCA, the number of principal components (PCs) is determined according to cumulative percent variance (CPV) more than 90%. For SKPCA, the size of the kernel parameter is selected as \(3m\) according to ref 5, and \(m\) is the number of variables. For SKICA, the number of independent components (ICs) is the same as that of the PCs. For SKNN, the number of nearest neighbors is 270. For SPCA, SKPCA, SKICA, and SKNN, the control limits of different indices are calculated at a confidence level of 99%.

Table 2. Parameter Settings of SPCA, SKPCA, SKICA, and SKNN

| methods  | statistics | PCs | ICs | window width | window sliding step | kernel parameter | \(k\) |
|----------|------------|-----|-----|--------------|---------------------|------------------|------|
| SPCA     | 54         | 13  |     | 25           | 15                  |                  |      |
| SKPCA    | 54         | 3   |     | 25           | 15                  | 270              |      |
| SKICA    | 54         | 4   | 4   | 25           | 15                  | 270              |      |
| SKNN     | 54         |     |     | 25           | 15                  | 3                |      |

Fault Detection:

- Fault Diagnosis:
  - The change of \(w\) has different effects on the four monitored variables. Figure 6 shows the difference between each variable of the training samples and the corresponding variable in the fault samples (only partial results are shown due to limited space). The fourth variable \(y_4\) has the maximum difference. Because the kernel method is difficult or even impossible to find the inverse mapping function from the feature space to the original space, the fault diagnosis of SKPCA and SKICA is not considered in this paper. Figure 7 is the comparison result of the diagnosis performance of SPCA and SKNN. SPCA cannot identify the essential ICs to be ignored.

- Fault Detection:
  - Figure 4 shows the NE of statistics, and we can observe that the NE of different statistics is greater than 10\(^{-3}\). Therefore, the statistics extracted from this dynamic process are non-Gaussian. The FDR and FAR of the four methods are shown in Table 3 (note that the FAR is obtained based on the validation samples). SKNN has the best detection performance among the three methods because it can overcome the problems caused by the non-linearity and non-Gaussianity of statistics. Figure 5 is the comparison result of the detection performance of the three methods. For SPCA, due to the non-Gaussian and nonlinearity of the statistics, the detection performance of SPCA is not good. The performance of SKICA is better than SPCA because the KPCA has the advantage in dealing with nonlinearity. Although the SKICA method can deal with non-Gaussian and nonlinear problems of process data, the fault detection performance of SKICA on experiments is not as good as that of SKNN. This may be due to the following reasons: to reduce the computational load of the SKICA method, ICA is performed after the kernel matrix is dimensionally reduced by PCA, which may cause the information used to extract essential ICs to be ignored.

![Figure 4](https://doi.org/10.1021/acsomega.2c01279/acsomega.2c01279_004)

Figure 4. Measures of non-Gaussianity of statistics for the numerical example.

Table 3. FDR (%) and FAR (%) of SPCA, SKPCA, SKICA, and SKNN for the Numerical Example

| index     | SPCA | SKPCA | SKICA | SKNN |
|-----------|------|-------|-------|------|
| \(T^2\)   |      |       |       |      |
| FDR       | 20.19| 10.57 | 13.56 | 57.10|
| FAR       | 0.9  | 0.9   | 1.5   | 0.0  |

ACS Omega 2022, 7, 18623–18637
Figure 5. Fault detection results of SPCA, SKPCA, SKICA, and SKNN for the numerical example. (a) SPCA; (b) SKPCA; (c) SKICA; (d) SKNN.

Figure 6. Difference between each variable of the training samples and the corresponding variable in the fault samples.

Figure 7. Fault diagnosis results of SPCA-T², SPCA-SPE and SKNN for the numerical example. (a) SPCA-T²; (b) SPCA-SPE; (c) SKNN.
have a smearing effect because VCkNN is defined in the original variable space (here means statistics space). In Figure 7c, the variable $y_i$ that contributes the most to the fault is successfully isolated. It is worth noting that the diagnosis result of SKNN indicates that the mean of $y_2$ has the maximum contribution as can be seen from Figure 8, the mean of $y_2$ increases significantly after the fault is introduced from the 33rd window. Therefore, it is effective to perform early fault detection or safety alarm by observing the mean of $y_2$.

**TE Benchmark Process.** When comparing the performance or effectiveness of process monitoring methods, the TEP is a benchmark choice. In ref 41, Downs and Vogel proposed the simulation platform. There are five major operating units in the TE process, namely, a reactor, a product condenser, a vapor–liquid separator, a recycle compressor, and a product stripper. The process has four kinds of reactants (A, C, D, E), two products (G, H), and contains a catalyst (B) and byproducts (F). The flowchart of the process is given in Figure 9. There are 11 manipulated variables (No.42–No.52), 22 process measurements (No.1–No.22), and 19 composition variables (No.23–No.41). For detailed information on the 52 monitoring variables and 21 fault patterns, see ref 42.

The parameter settings of SPCA, SKPCA, and SKNN are shown in Table 4. For all three methods, the number of training samples and the number of validation samples are 960 and 480, respectively. In addition, there are 960 testing samples where the fault is introduced from the 161st sample. The thresholds of different methods are all calculated at a confidence level of 99%. For second-order statistics, only the significant statistics will be selected. The specific selection method is the same as in ref 2. In order to facilitate the establishment of the model, for a large number of second-order statistics, we only select the numerically significant ones. The specific selection rules are as follows:

1. $r_{ij}$ is selected only if $|r_{ij}| > 0.5$ for more than 70% of the training SPs;
2. $r_i$ is selected only if $|r_i| > 0.5$ for more than 90% of the training SPs;
3. $r_{ij}$ is selected only if $|r_{ij}| > 0.5$ for more than 90% of the training SPs.

After statistics selection, a total of 223 statistics of the TE process are used for process monitoring.

![Figure 8. Man of $y_2$.](image)

![Figure 9. Flowchart of the Tennessee Eastman process.](image)
RESULTS AND DISCUSSION

This section analyzes and discusses the experimental results of the TE process.

- Fault Detection:
  - From Figure 10, the statistics extracted from the TE process are non-Gaussian because the NE of the statistics is greater than the threshold of $10^{-3}$. The FDR of the four methods are listed in Table 5. Note that faults 3, 9, and 15 are not considered here because they are difficult to be detected.\(^{1,43,44}\) For faults 5, 10, 11, 16, 18, 20, and 21, the FDR of the SKNN method is higher than those of SPCA, SKPCA, and SKICA. For faults 4, 5, 10, 16, 19, and 21, SPCA, SKPCA, and SKICA have difficulties detecting these faults (the FDR of $T^2$ is less than 30%). It can be concluded from the above experimental results that SKNN is superior to SPCA, SKPCA, and SKICA.

  The monitoring results of SPCA, SKPCA, SKICA, and SKNN for faults 10, 11, and 20 are shown in Figures 11–13, respectively. The reason why the performance SKNN is better than those of SPCA and SKPCA is that it can simultaneously overcome the nonlinearity and non-Gaussianity of statistics. For SPCA, the nonlinearity and non-Gaussianity of statistics destroy the premise of the threshold calculation in PCA. SKPCA only considers the nonlinear problem of statistics, and the non-Gaussianity of statistics is ignored. Although the SKICA method can deal with non-Gaussian and nonlinear problems of process variables (statistics), in the process of reducing the dimensionality of the kernel matrix and extracting ICs, some key information of the process may be lost, resulting in poor detection performance.

- Fault Diagnosis:
  - In the stage of fault diagnosis, fault 1, fault 7, and fault 10 are used for comparison. The faulty variables of fault 1 are the 1st variable, fourth

![Figure 10. Measures of non-Gaussianity of statistics for TEP.](https://doi.org/10.1021/acsomega.2c01279)

| fault | SPCA $T^2$ | SPCA SPE | SKPCA $T^2$ | SKPCA SPE | SKICA $T^2$ | SKICA SPE | SKNN |
|-------|------------|----------|-------------|-----------|-------------|----------|------|
| 1     | 99.25      | 100      | 99.25       | 99.25     | 99.25       | 44.03    | 100  |
| 2     | 96.27      | 98.51    | 96.27       | 97.76     | 97.01       | 96.27    | 98.51 |
| 4     | 1.49       | 100      | 0.75        | 91.79     | 0.75        | 8.21     | 100  |
| 5     | 21.64      | 50.75    | 23.88       | 30.60     | 24.63       | 22.39    | 55.97|
| 6     | 100        | 100      | 100         | 100       | 100         | 100      | 100  |
| 7     | 70.15      | 100      | 41.04       | 100       | 87.31       | 35.82    | 100  |
| 8     | 96.27      | 97.76    | 96.27       | 97.76     | 97.76       | 93.28    | 97.76|
| 10    | 16.42      | 64.93    | 26.87       | 50.75     | 28.36       | 18.66    | 72.39|
| 11    | 67.16      | 97.01    | 68.66       | 94.03     | 79.10       | 67.16    | 98.51|
| 12    | 98.51      | 99.25    | 99.25       | 99.25     | 99.25       | 99.25    | 99.25|
| 13    | 90.30      | 95.52    | 91.79       | 94.03     | 92.54       | 85.07    | 95.52|
| 14    | 100        | 100      | 100         | 100       | 100         | 100      | 100  |
| 16    | 22.39      | 73.13    | 25.37       | 54.48     | 30.60       | 14.18    | 76.87|
| 17    | 88.81      | 97.01    | 73.13       | 94.03     | 74.63       | 73.13    | 97.01|
| 18    | 88.81      | 90.30    | 18.66       | 89.55     | 18.66       | 18.66    | 91.04|
| 19    | 8.21       | 98.51    | 5.22        | 98.51     | 9.70        | 1.49     | 98.51|
| 20    | 50.00      | 77.61    | 49.25       | 70.15     | 55.97       | 49.25    | 80.60|
| 21    | 0          | 44.03    | 0           | 30.60     | 0           | 0        | 46.27|
| average | 61.98   | 88.02    | 56.43       | 82.92     | 60.86       | 51.49    | 89.35|

Table 5. FDR (%) of SPCA, SKPCA, SKICA, and SKNN for the TEP
variable, 18th variable, 19th variable, 38th variable, 44th variable, 45th variable, and 50th variable. In Figure 14, the SPCA-\(T^2\) and SPCA-SPE identify some nonfaulty statistics of fault 1 as fault variables due to the smearing effect. SKNN does not have a smearing effect. Therefore, the diagnosis result of SKNN is similar to the actual situation. The cause of fault 7 is the pressure loss of the C header. To maintain the reactor level, the controller adjusts the flow of stream 4 by changing the total feed flow rate (i.e., 45th variable). Therefore, it is the 45th variable that actually generates the exception. The cause of fault 10 is the random variation in the C feed temperature (stream 4). Specifically, it is the increase in variability of the 18th variable. Therefore, there is only one fault variable for fault 7 and
fault 10, which is the 45th variable and the 18th variable, respectively. The diagnosis results of fault 7 and fault 10 are shown in Figures 15 and 16, respectively. Both SPCA-SPE and SKNN correctly isolate the fault statistics. However, it is worth noting that the maximum contribution of the 45th
The statistic (i.e., the mean of the 45th variable) in Figure 15b,c is about 0.6 and 0.8, respectively; while the maximum contribution of the 70th statistic (i.e., the variance of the 18th variable) is about 0.7 and 0.9, respectively. In addition, the SPCA- \( T^2 \) identifies many nonfaulty statistics as faulty variables due to the smearing effect. In summary, the SKNN method is also superior to the SPCA method in fault diagnosis.

For fault 7, it can be seen from Figure 15 that the faulty variable isolated by SKNN is the mean of the 45th variable, not the variable itself. For fault 10, in Figure 16, the fault statistic isolated by SKNN is the variance of the 18th variable. As can be seen from Figure 17a,b, the mean of the 45th variable and the variance of the 18th variable increase significantly after the fault is introduced from the 26th window. Therefore, for fault 7 and fault 10, it is effective to perform early fault detection by observing the mean of the 45th variable and the variance of the 18th variable.

### CONCLUSIONS

The SPA method can extract different statistics to capture the complex characteristics of the process. Therefore, it is a feasible framework for complex industrial process monitoring. The non-Gaussianity and non-linearity of extracted statistics under the SPA framework cause the failure of the traditional MSPM methods. In order to overcome these problems, a new process monitoring method based on SPA and the kNN algorithm is proposed in this paper. In the premise of the kNN method, the data samples are not required to satisfy a certain distribution. Therefore, the proposed method can avoid the problems caused by the non-Gaussianity and nonlinearity of extracted statistics while inheriting the advantages of SPA in extracting HOS information. In addition, in terms of fault diagnosis, SKNN also has better performance because it is not affected by fault smearing. The experiments on numerical examples and TE processes verify the effectiveness of the proposed method.

The strength of the proposed method is that it can overcome the non-Gaussian and nonlinear problems in the SPA framework as well as alleviating the fault smearing phenomenon in the fault isolation of statistics. However, the proposed method still has limitations when dealing with multimode processes because the statistics extracted by SPA still retain the multimodality characteristics in the original sample space, which will make the calculation of detection thresholds seriously deviate from the normal level.

### AUTHOR INFORMATION

#### Corresponding Author

Zuxin Li — School of Science and Engineering, Huzhou College, Huzhou 313000, China; orcid.org/0000-0002-5816-9728; Email: lzx@zjhu.edu.cn

#### Authors

Zhe Zhou — School of Engineering, Huzhou University, Huzhou 313000, China; orcid.org/0000-0002-0445-2560

Jian Wang — School of Engineering, Huzhou University, Huzhou 313000, China

Chunjie Yang — State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China

Chenglin Wen — School of Automation, Guangdong University of Petrochemical Technology, Maoming 525000, China

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c01279

#### Notes

The authors declare no competing financial interest.
**ACKNOWLEDGMENTS**

Z.Z. received funding from the NSFC under Grant 61703158. Z.Z. received funding from Huzhou Municipal Natural Science Foundation of China under Grant 2021YZ03. Z.Z. received funding from the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2002. Z.L. received funding from the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2002. Z.Z. received funding from Huzhou Municipal Natural Science Foundation of China under Grant 2021YZ03. Z.Z. received funding from the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2002. Z.L. received funding from the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China under Grant ICT2002.

**REFERENCES**

(1) Chiang, L. H.; Russell, E. L.; Braatz, R. D. Fault Detection and Diagnosis in Industrial Systems; Springer London, 2001, DOI: 10.1007/978-1-4471-0347-9.

(2) Wang, J.; He, Q. P. Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis. Ind. Eng. Chem. Res. 2010, 49, 7858–7869.

(3) Nomikos, P.; Macgregor, J. F. Monitoring Batch Processes Using Multiway Principal Component Analysis. AIChE J. 1994, 40, 1361–1375.

(4) Ku, W.; Storer, R. H.; Georgakis, C. Disturbance detection and isolation by dynamic principal component analysis. Chemom. Intell. Lab. Syst. 1995, 30, 179–196.

(5) Lee, J. M.; Yoo, C. K.; Choi, S. W.; Vanrolleghem, P. A.; Lee, I. B. Nonlinear process monitoring using kernel principal component analysis. Chem. Eng. Sci. 2004, 59, 223–234.

(6) Lee, J. M.; Yoo, C. K.; Lee, I. B. Fault detection of batch processes using multiway kernel principal component analysis. Comput. Chem. Eng. 2004, 28, 1837–1847.

(7) Choi, S. W.; Lee, C.; Lee, J.-M.; Park, J. H.; Lee, I.-B. Fault detection and identification of nonlinear processes based on kernel PCA. Chemom. Intell. Lab. Syst. 2005, 75, 55–67.

(8) Yu, H.; Khan, F.; Garaniya, V. A sparse PCA for nonlinear fault diagnosis and robust feature discovery of industrial processes. AIChE J. 2016, 62, 1494–1513.

(9) Yu, H.; Khan, F. Improved latent variable models for nonlinear and dynamic process monitoring. Chem. Eng. Sci. 2017, 168, 325–338.

(10) Jiang, Q.; Yan, X. Parallel PCA–KPCA for nonlinear process monitoring. Control Eng. Pract. 2018, 80, 17–25.

(11) Cai, P.; Deng, X. Incipient fault detection for nonlinear processes based on dynamic multi-block probability related kernel principal component analysis. ISA Trans. 2020, 105, 210–220.

(12) Mansouri, M.; Nounou, M.; Nounou, H.; Karim, N. Kernel PCA-based GLRT for nonlinear fault detection of chemical processes. J. Loss Prev. Process Ind. 2016, 40, 334–347.

(13) Yu, H.; Khan, F.; Garaniya, V. Modified independent component analysis and Bayesian network-based two-stage fault diagnosis of process operations. Ind. Eng. Chem. Res. 2015, 54, 2724–2742.

(14) Wang, Y.; Liu, Y.; Khan, F.; Imtiaz, S. Semiparametric PCA and bayesian network based process fault diagnosis technique. Can. J. Chem. Eng. 2017, 95, 1800–1816.

(15) Amin, M. T.; Imtiaz, S.; Khan, F. Process system Fault Detection and Diagnosis Using a Hybrid Technique. Fifth International Conference on Chemical Engineering; 2017.

(16) Gharaibegheri, H.; Imtiaz, S. A.; Khan, F. Root Cause Diagnosis of Process Fault Using KPCA and Bayesian Network. Ind. Eng. Chem. Res. 2017, 50, 3445–3454.

(17) He, Q. P.; Wang, J. Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. AIChE J. 2011, 57, 107–121.

(18) He, Q. P.; Wang, J.; Shah, D.; Vahdat, N. Statistical process monitoring for IoT-Enabled cybermanufacturing: opportunities and challenges. IFAC-PapersOnLine 2017, 50, 14946–14951.

(19) He, Q. P.; Wang, J. Statistical process monitoring in the era of smart manufacturing: American Control Conference, 2017.

(20) He, Q. P. Statistics Pattern Analysis: A Statistical Process Monitoring Tool for Smart Manufacturing. Comput.-Aided Chem. Eng. 2017, 44, 2071–2076.

(21) He, Q. P.; Wang, J.; Shah, D. Feature Space Monitoring for Smart Manufacturing via Statistics Pattern Analysis. Comput. Chem. Eng. 2019, 126, 321–331.

(22) Deng, X.; Tian, X. Nonlinear process fault pattern recognition using statistics kernel PCA similarity factor. Neurocomputing 2013, 121, 298–308.

(23) Zhou, B.; Gu, X. Multi-block statistics local kernel principal component analysis algorithm and its application in nonlinear process fault detection. Neurocomputing 2020, 376, 222–231.

(24) Ning, C.; Chen, M.; Zhou, D. Hidden Markov Model-Based Statistics Pattern Analysis for Multimode Process Monitoring: An Index-Switching Scheme. Ind. Eng. Chem. Res. 2014, 53, 11084–11095.

(25) Shah, D.; He, Q. P.; Wang, J. A spectroscopic chemometric modeling approach based on statistics pattern analysis. IFAC-PapersOnLine 2018, 51, 369–374.

(26) Ma, H. H.; Hu, Y.; Shi, H. B. Statistics kernel principal component analysis for nonlinear process fault detection. 2011 9th World Congress on Intelligent Control and Automation; 2011.

(27) Zhou, Z.; Wen, C.; Yang, C. Fault Isolation Based on k-Nearest Neighbor Rule for Industrial Processes. IEEE Trans. Ind. Electron. 2016, 63, 2578–2586.

(28) Zhou, Z.; Li, Z.; Cai, Z.; Wang, P. Fault Identification Using Fast k-Nearest Neighbor Reconstruction. Processes 2019, 7, 340.

(29) He, Q. P.; Wang, J. Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. IEEE Trans. Semicond. Manuf. 2007, 20, 345–354.

(30) Verdier, G.; Ferreira, A. Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing. IEEE Trans. Semicond. Manuf. 2011, 24, 59–68.

(31) Hyvärinen, A.; Oja, E. Independent Component Analysis: Algorithms and Applications. Neural Networks 2000, 13, 411–430.

(32) Cai, L.; Thornhill, N. F.; Pal, B. C. Multivariate Detection of Power System Disturbances Based on Fourth Order Moment and Singular Value Decomposition. IEEE Trans. Power Syst. 2017, 4298–4297.

(33) Erik, G. M.; John, W. F. ICA Using Spacings Estimates of Entropy. J. Mach. Learn. Res. 2003, 4, 1271–1295.

(34) Choudhury, M. A. A. S.; Shah, S. L.; Thornhill, N. F. Diagnosis of poor control-loop performance using higher-order statistics. Automat. 2004, 40, 1719–1728.

(35) Theiler, J.; Eubank, S.; Longtin, A.; Galdrikian, B.; Farmer, J. D. Testing for nonlinearity in time series: The method of surrogate data. Phys. D 1992, 58, 77–94.

(36) Engbert, R. Testing for nonlinearity: the role of surrogate data. Chaos, Solitons Fractals 2002, 13, 79–84.

(37) Schreiber, T.; Schmitz, A. Improved Surrogate Data for Nonlinearity Tests. Phys. Rev. Lett. 1996, 77, 655–658.

(38) Paluš, M.; Albrecht, V.; Dvořák, I. Information theoretic test for nonlinearity in time series. Phys. Lett. A 1993, 175, 203–209.

(39) Lee, J.-M.; Qin, S. J.; Lee, I.-B. Fault Detection of Non-Linear Processes Using Kernel Independent Component Analysis. Can. J. Chem. Eng. 2007, 85, 526–536.

(40) Valle, S.; Li, W.; Qin, S. J. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. Ind. Eng. Chem. Res. 1999, 38, 4389–4401.

(41) Downs, J. J.; Vogel, E. F. A plant-wide industrial process control problem. Comput. Chem. Eng. 1993, 17, 245–255.

(42) Ricker, N. L. Optimal steady-state operation of the Tennessee Eastman challenge process. Comput. Chem. Eng. 1995, 19, 949–959.

(43) Lee, J. M.; Qin, S. J.; Lee, I. B. Fault detection and diagnosis based on modified independent component analysis. AIChE J. 2006, 52, 3501–3514.
Zhang, Y. Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM. *Chem. Eng. Sci.* 2009, 64, 801–811.