Improved Statistical Pattern Analysis Monitoring for Complex Multivariate Processes Using Empirical Likelihood

Authors:

Jianwen Shao, Xin Zhang, Wenhua Chen, Xiaomin Shen

Date Submitted: 2021-07-19

Keywords: moving window, higher-order statistics, empirical likelihood, statistical pattern analysis

Abstract:

> This article developed an improved statistical pattern analysis (SPA) monitoring strategy for fault detection of complex multivariate processes using empirical likelihood. The technique based on statistical pattern analysis performs fault detection by inspecting change in the statistics of process variables (e.g., mean value, correlation coefficient, variance, kurtosis, etc.). It is capable of monitoring non-Gaussian or even nonlinear processes. However, the original SPA framework explicitly computes all the high-order statistics, which significantly increases the scale and dimensionality of the problem, especially in the case of complex multivariate processes. To alleviate this difficulty, we propose monitoring changes in the statistics with the same order using empirical likelihood, which is a widely used estimation method to construct confidence limits or regions for parameters with similar properties. As a result, changes in statistics of the same order can be translated into a single index; hence more information on the faulty conditions can be observed. Furthermore, by considering statistics of the same order, the scale of the problem is reduced significantly. The improved statistical pattern analysis monitoring strategy is suitable for monitoring complex multivariate processes. The performance of the improved method is illustrated by an application study to fault detection of the Tennessee Eastman (TE) process.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2021.0625
Citation (this specific file, latest version):	LAPSE:2021.0625-1
Citation (this specific file, this version):	LAPSE:2021.0625-1v1

DOI of Published Version: https://doi.org/10.3390/pr8121619 processes

MDPI

Article

Improved Statistical Pattern Analysis Monitoring for Complex Multivariate Processes Using Empirical Likelihood

Jianwen Shao [1,2], Xin Zhang [2], Wenhua Chen [1,*] and Xiaomin Shen [2]

[1] Zhejiang Province's Key Laboratory of Reliability Technology for Mechanical and Electronic Product, Zhejiang Sci-Tech University, Hangzhou 310018, China; jianwenshao@126.com

[2] Reasearch Division of Metrology in Transportation and Acoustics, Zhejiang Institute of Metrology, Hangzhou 310018, China; zhangxin900308@126.com (X.Z.); soominshim@163.com (X.S.)

* Correspondence: chenwh@zstu.edu.cn

Received: 4 November 2020; Accepted: 6 December 2020; Published: 9 December 2020

Abstract: This article developed an improved statistical pattern analysis (SPA) monitoring strategy for fault detection of complex multivariate processes using empirical likelihood. The technique based on statistical pattern analysis performs fault detection by inspecting change in the statistics of process variables (e.g., mean value, correlation coefficient, variance, kurtosis, etc.). It is capable of monitoring non-Gaussian or even nonlinear processes. However, the original SPA framework explicitly computes all the high-order statistics, which significantly increases the scale and dimensionality of the problem, especially in the case of complex multivariate processes. To alleviate this difficulty, we propose monitoring changes in the statistics with the same order using empirical likelihood, which is a widely used estimation method to construct confidence limits or regions for parameters with similar properties. As a result, changes in statistics of the same order can be translated into a single index; hence more information on the faulty conditions can be observed. Furthermore, by considering statistics of the same order, the scale of the problem is reduced significantly. The improved statistical pattern analysis monitoring strategy is suitable for monitoring complex multivariate processes. The performance of the improved method is illustrated by an application study to fault detection of the Tennessee Eastman (TE) process.

Keywords: statistical pattern analysis; empirical likelihood; higher-order statistics; moving window

1. Introduction

The last decades have witnessed great research progress in the field of fault detection and diagnosis using multivariate statistical process control techniques (MSPC). Among all MSPC techniques, traditional methods such as principal component analysis (PCA) [1–3], principal component regression (PCR) [4,5], and partial least squares (PLS) [6,7] are perhaps the most popular. One of the limitations of traditional methods is that they are designed for Gaussian processes, while it is commonly acknowledged that industrial systems are generally non-Gaussian or even nonlinear. To deal with non-Gaussian or nonlinear processes, different kinds of independent component analysis (ICA) [8–13] based approaches have been proposed. The basic idea of ICA based approaches is to decompose the process data into independent components so that non-Gaussian and Gaussian components can be separated. By considering statistical independence of the extracted components, ICA involves higher order statistics of process data implicitly.

More recently, He and Wang [14] suggested the use of statistical pattern analysis to monitor non-Gaussian batch processes and further extended it to the monitoring of continuous process. Unlike ICA based approaches, the SPA framework considers high order statistics of process data explicitly.

It examines change in the variance-covariance (e.g., mean, variance, cross-correlation, autocorrelation etc.) as well other higher order statistics of the process data using a window based approach. These statistics are called statistical patterns (SPs). It is assumed that the considered SPs followed a Gaussian distribution so that PCA could be applied to determine whether there was a faulty condition. The advantages of SPA are obvious: it does not employ a data preprocessing step and by using various statistics of process variables, it can capture different process characteristics. However, since the SPA framework uses the statistics of process variables explicitly, a large number of statistics may be required to fully capture the process characteristics; hence, significantly increasing the scale and dimensionality of the problem. As a result, it is not suitable for the monitoring of complex multivariate processes because the monitoring sensitivity will be reduced by considering a large number of statistics.

In this work, we proposed monitoring the statistics of process variables with the same order using empirical likelihood [15] to increase sensitivity and reduce the dimensionality of the SP matrix. Empirical likelihood is a nonparametric method used to construct confidence regions for finite-dimensional parameters [16]. It is widely used to test, for example, symmetry about zero, change in distribution, independence, etc. [17]. By considering the change in statistics with the same order, it is possible to know which kind of characteristics have changed in the process variables and hence reveal more information about the potential process faults.

The rest of the paper is organized as follows. The next section presents the basic idea of statistical pattern analysis based monitoring strategy. In Section 3, empirical likelihood is introduced. Section 4 proposes the improved SPA monitoring strategy, followed by a demonstration of the application of the methodology in Section 5. Finally, a concluding summary is given in Section 6.

2. Statistical Pattern Analysis Based Monitoring Strategy

The rationale of SPA based monitoring strategy is that statistics of process variables under abnormal conditions will be different from those under normal operational conditions. Hence, process faults can be detected by considering the statistics of process variables instead of the process variables themselves. The SPA based monitoring strategy consists of two steps: construction of statistical patterns, and quantification of dissimilarity. A statistical pattern is a statistic calculated from a consecutive set of process measurements through a window based strategy. These statistics include the mean value, variance, and skewness, which capture the characteristics of individual variables; correlation coefficient that captures the interactions among variables as well as autocorrelation and cross-correlation that captures the process dynamics. In other words, statistics that capture different kinds of characteristics of complex multivariate processes can be considered.

Generation of statistical patterns can be achieved by a moving window approach. Denote the process measurement at time instance k by $\mathbf{x}(k)$, $\mathbf{x}(k) \in \mathbb{R}^m$,where m is the number of process variables, and a window of process measurements can be generated as

$$\mathbf{X}_k = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_m] = \begin{bmatrix} x_1(k-w+1) & x_2(k-w+1) & \cdots & x_m(k-w+1) \\ x_1(k-w+2) & x_2(k-w+2) & \cdots & x_m(k-w+2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(k) & x_2(k) & \dots & x_m(k) \end{bmatrix} \tag{1}$$

where w is the window length and $x_i(k)$, $i = 1, 2, \cdots, m$ is the most recent measurement of x_i. The statistical patterns are generated from Equation (1). The authors in [14,15] considered four sets of process statistics

$$\mathbf{S} \equiv [\boldsymbol{\mu}|\boldsymbol{\Sigma}|\Xi_1|\Xi_2] \tag{2}$$

In Equation (2), $\boldsymbol{\mu} \in \mathbb{R}^{M_1}$ relates to variable means, or first order statistics, which are calculated from Equation (1) as

$$\mu_i = \frac{1}{w}\sum_{l=0}^{w-1} x_i(k-l) \tag{3}$$

where M_1 is the number of first order statistics. On the other hand, $\boldsymbol{\Sigma} \in \mathbb{R}^{M_2}$ represents the second order statistics including variance v_i, correlation $r_{i,j}$, autocorrelation r_i^d, and cross correlation $r_{i,j}^d$. These statistics can be computed as

$$v_i = \frac{1}{w}\sum_{l=0}^{w-1} [x_i(k-l) - \mu_i]^2 \tag{4}$$

$$r_{i,j} = \frac{1}{w}\frac{\sum\limits_{l=0}^{w-1} [x_i(k-l) - \mu_i]\left[x_j(k-l) - \mu_j\right]}{\sqrt{v_i v_j}} \tag{5}$$

$$r_i^d = \frac{1}{w-d}\frac{\sum\limits_{l=d}^{w-1} [x_i(k-l) - \mu_i][x_i(k+d-l) - \mu_i]}{v_i} \tag{6}$$

$$r_{i,j}^d = \frac{1}{w-d}\frac{\sum\limits_{l=d}^{w-1} [x_i(k-l) - \mu_i]\left[x_j(k+d-l) - \mu_i\right]}{\sqrt{v_i v_j}} \tag{7}$$

where d is the time lag between variables and M_2 is the number of second order statistics.

Finally, $\Xi_1 \in \mathbb{R}^{M_3}$ represents the third order statistics like skewness γ_i and $\Xi_2 \in \mathbb{R}^{M_4}$ relates to the fourth order statistics like kurtosis κ_i, which can be calculated as

$$\gamma_i = \frac{\frac{1}{w}\sum\limits_{l=0}^{w-1} [x_i(k-l) - \mu_i]^3}{\left(\frac{1}{w}\sum\limits_{l=0}^{w-1} [x_i(k-l) - \mu_i]^2\right)^{3/2}} \tag{8}$$

$$\kappa_i = \frac{\frac{1}{w}\sum\limits_{l=0}^{w-1} [x_i(k-l) - \mu_i]^4}{\left(\frac{1}{w}\sum\limits_{l=0}^{w-1} [x_i(k-l) - \mu_i]^2\right)^{2}} - 3 \tag{9}$$

where M_3 and M_4 are the numbers of the third and fourth order statistics. Statistics higher than four orders can also be considered, however, for the sake of simplicity, only the above statistics were considered here. With different statistics calculated, a statistical pattern vector can be obtained by putting together all the statistical patterns in a row vector. The row vector reflects the statistical properties of process variables at the window from k-w + 1 to k, so that an SP matrix can be obtained as $\mathbf{S} \in \mathbb{R}^{M \times w}$, where M is the number of SPs and $M = M_1 + M_2 + M_3 + M_4$.

By inspecting the statistical pattern vector at different windows using principal component analysis, the fault can be detected. The SPA monitoring framework considers the statistics quantifying the non-Gaussianity and nonlinearity of process data; it is capable of monitoring non-Gaussian and nonlinear processes. However, for complex multivariate processes, there is a large number of variables. Using the SPA framework may lead to consideration of too many statistics. For example, for a six variable process, considering the statistics listed from Equations (3)–(9) may involve six mean values, 21 variance and covariance components, six skewness values, and six kurtosis values. There was a total of 39 statistics, not to mention other terms like autocorrelation and cross correlation terms. To simplify the monitoring task, we used empirical likelihood to monitor changes in statistics with the same order, so that only four monitoring statistics were needed.

3. Empirical Likelihood

Empirical likelihood is a nonparametric approach used to define confidence regions for omnibus hypothesis testing. As a nonparametric approach, it is distribution free; the error of the confidence region obtained by empirical likelihood is Bartlett correctable [16].

As pointed out in Section 2, as there were too many statistics considered in the original SPA framework, we considered monitoring the change in statistics with the same order using empirical likelihood. Assume we have N normal data samples and the statistical pattern matrix $\bar{\mathbf{S}} = \left[\bar{\boldsymbol{\mu}}\ \bar{\boldsymbol{\Sigma}}\ \bar{\Xi}_1\ \bar{\Xi}_2 \right]$ has been calculated from the original data. As a new data sample $\mathbf{x}_{N+1}$ arrives, a moving window approach can be used to obtain the SP vector by considering the statistics of $\{\mathbf{x}(N-w+2), \cdots, \mathbf{x}(N+1)\}$, denoted as $\mathbf{S}_{N+1} = [\boldsymbol{\mu}_{N+1}\ \boldsymbol{\Sigma}_{N+1}\ \Xi_{1,N+1}\ \Xi_{2,N+1}]$. Take the first order statistic, for example, to monitor the change in first order statistics $\boldsymbol{\mu}_{N+1}$, another moving window approach should be considered (i.e., $\boldsymbol{\mu}_{N-s+2}, \boldsymbol{\mu}_{N-s+3}, \cdots, \boldsymbol{\mu}_{N+1}$). Denote the probability density function of $\boldsymbol{\mu}_{N-s+2}, \boldsymbol{\mu}_{N-s+3}, \cdots, \boldsymbol{\mu}_{N+1}$ as $p(\boldsymbol{\mu})$, and the following hypothesis test can be constructed

$$H_0 : \bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_0 \leftrightarrow H_1 : \bar{\boldsymbol{\mu}} \neq \boldsymbol{\mu}_0 \tag{10}$$

where $\boldsymbol{\mu}_0$ is the mean value of $\boldsymbol{\mu}_{N-s+2}, \boldsymbol{\mu}_{N-s+3}, \cdots, \boldsymbol{\mu}_{N+1}$ and s is the length of the sliding window. If the alternative hypothesis holds, then the new data sample $\mathbf{x}_{N+1}$ is a faulty sample with a fault in the first order statistics.

With the hypothesis test in Equation (10), it is possible to detect fault in the first order statistics. However, the hypothesis test cannot be used only when the confidence limit is available, which can be obtained by maximizing the following empirical likelihood function

$$L = \prod\nolimits_{i=N-s+2}^{N+1} p_i \tag{11}$$

subject to the following constraints

$$\sum_{i=N+s-2}^{N+1} p_i = 1, \quad \sum_{i=N+s-2}^{N+1} p_i \boldsymbol{\mu}_i = \bar{\boldsymbol{\mu}} \tag{12}$$

where p_i is the probability of $\boldsymbol{\mu}_i$. The maximum can be reached if and only if $p_i = \frac{1}{s}$ for all p_i, so that the null hypothesis $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_0$ holds. Otherwise, $\bar{\boldsymbol{\mu}} \neq \boldsymbol{\mu}_0$ holds and faults in the first order statistic can be observed.

To obtain the probability p_i, consider the following log-likelihood ratio

$$\widetilde{L} = -\log \frac{\prod_{i=N-s+2}^{N+1} p_i}{s^{-s}} = -\sum_{i=N-s+2}^{N+1} \log(sp_i) \tag{13}$$

subject to the constraints in Equation (12).

Using the Lagrange multiplier, we have

$$G = \sum_{i=N-s+2}^{N+1} \log(sp_i) - s\boldsymbol{\lambda}^T \sum_{i=N-s+2}^{N+1} \left(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}} \right) + \gamma \left(\sum_{i=N-s+2}^{N+1} p_i - 1 \right) \tag{14}$$

Differentiating Equation (14) and setting it to be zero, the probability p_i can be obtained as

$$p_i = \frac{1}{s}\frac{1}{1+\boldsymbol{\lambda}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})} \tag{15}$$

where $\boldsymbol{\lambda}$ is the solution of the following equation

$$J(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=N-s+2}^{N+1}\frac{\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}}{1+\boldsymbol{\lambda}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})} = 0 \tag{16}$$

Equation (16) can be solved using gradient search as follows

$$\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}(t) - \alpha\frac{\partial J(\boldsymbol{\lambda}(t))}{\partial\boldsymbol{\lambda}(t)} \tag{17}$$

where $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\lambda}(t+1)$ are the values of $\boldsymbol{\lambda}$ at the t-th and (t + 1)-th iteration of the gradient search; α is the learning rate; $\frac{\partial J(\boldsymbol{\lambda}(t))}{\partial\boldsymbol{\lambda}(t)}$ is the derivative of $J(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}(t)$ and can be obtained as

$$\frac{\partial J(\boldsymbol{\lambda}(t))}{\partial\boldsymbol{\lambda}(t)} = -\frac{1}{n}\sum_{i=N-s+2}^{N+1}\frac{\left(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}\right)^2}{\left(1+\boldsymbol{\lambda}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})\right)^2} \tag{18}$$

Combining Equations (17) and (18), the updating formula of $\boldsymbol{\lambda}$ in the gradient search can be obtained as follows.

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \frac{\alpha}{n}\sum_{i=N-s+2}^{N+1}\frac{\left(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}\right)^2}{\left(1+\boldsymbol{\lambda}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})\right)^2} \tag{19}$$

Once $\boldsymbol{\lambda}$ is determined, the probability p_i and hence the log-likelihood ratio can be computed from Equations (13) and (15). Similar procedures can be employed for second order statistics $\boldsymbol{\Sigma}$ and higher order statistics $\boldsymbol{\Xi}$.

4. Process Monitoring Strategy Based on Improved Statistical Pattern Analysis

With the statistical patterns obtained in Section 2, process monitoring can be performed by constructing monitoring plots using the empirical likelihood method in Section 3. Assume a statistical pattern matrix $\mathbf{S} = [\boldsymbol{\mu}\ \boldsymbol{\Sigma}\ \boldsymbol{\Xi}_1\ \boldsymbol{\Xi}_2]$ has already been obtained using the N training data samples by a moving window approach (with the window length of w). When a new data sample $\mathbf{x}(N+1)$ arrives, another statistical pattern vector $\mathbf{s}_{N+1} = [\boldsymbol{\mu}_{N+1}\ \boldsymbol{\Sigma}_{N+1}\ \boldsymbol{\Xi}_{1,N+1}\ \boldsymbol{\Xi}_{2,N+1}]$ can be obtained based on $\mathbf{x}(N+1)$ and its previous s-1 samples. The statistical pattern vector is then augmented with the previous w-1 vectors to form a new statistical pattern matrix $\mathbf{S}_{N+1}$. Empirical likelihood can then be performed between $\mathbf{S}_{N+1}$ and $\mathbf{S}$. Here, we performed a total of four empirical likelihood tests on the SP matrices, corresponding to the first, second, third, and fourth order statistical patterns. With the log-likelihood ratios estimated, fault detection can then be performed.

The process monitoring strategy can be divided into an offline training stage and an online monitoring stage, which are illustrated in Sections 4.1 and 4.2. Figure 1 shows the flowchart of the process monitoring strategy.

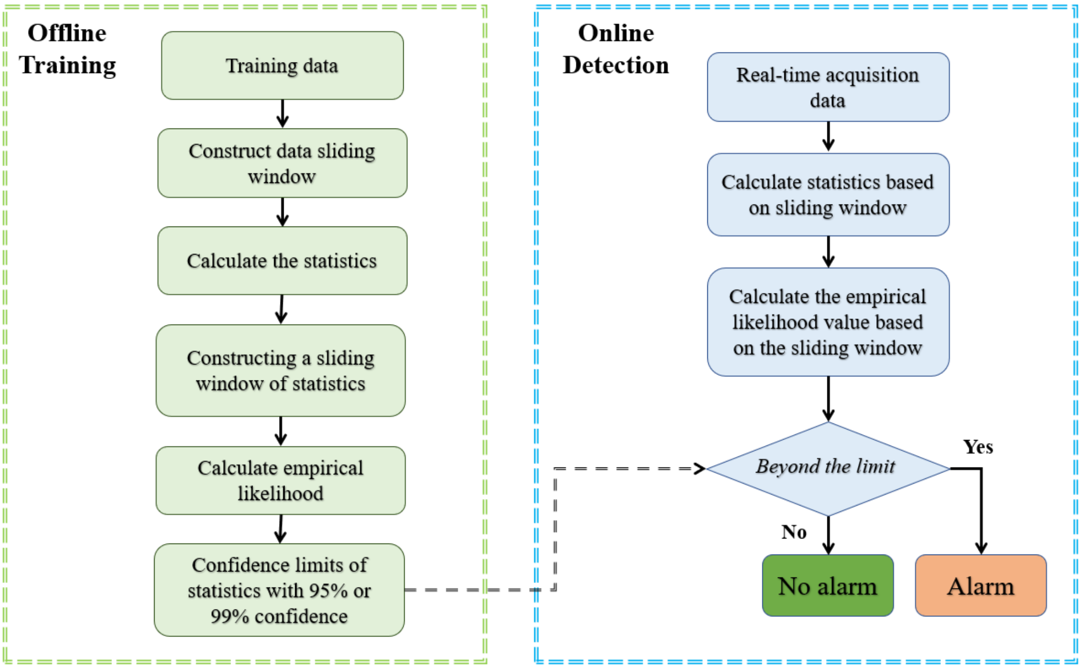
Offline Training
Training data
Construct data sliding window
Calculate the statistics
Constructing a sliding window of statistics
Calculate empirical likelihood
Confidence limits of statistics with 95% or 99% confidence
Online Detection
Real-time acquisition data
Calculate statistics based on sliding window
Calculate the empirical likelihood value based on the sliding window
Beyond the limit
Yes
No
No alarm
Alarm

Figure 1. Flowchart of offline training and online detection.

4.1. Offline Training Stage

The offline training stage can be summarized as follows.

Step 1: Collect N normal samples under normal conditions, set the lengths of the two sliding windows as s and w.

Step 2: Perform the first moving window processing based on the normal samples, construct $N-s+1$ groups of measurement matrices using Equation (1). Calculate the first order, second order, third order, and fourth order statistics from the samples in each sub-window to form the statistical pattern matrix $\mathbf{S} = [\boldsymbol{\mu}\ \boldsymbol{\Sigma}\ \boldsymbol{\Xi}_1\ \boldsymbol{\Xi}_2]$ based on Equations (3)–(9).

Step 3: Divide the statistical patter matrix $\mathbf{S}$ into two parts with equal length, $\mathbf{S}_1$ and $\mathbf{S}_2$. Let $q = \frac{N-w+1}{2}$, use $\mathbf{S}_1$ as the base SP matrix, and divide $\mathbf{S}_2$ into a series of $q-s+1$ submatrices with the length of s using the moving window approach.

Step 4: Perform empirical likelihood tests between the four sets of statistics in $\mathbf{S}_1$ and those in the series of $q-s+1$ submatrices of $\mathbf{S}_2$. Since $\mathbf{S}_1$ and $\mathbf{S}_2$ contain SPs corresponding to the first-, second-, third-, and fourth-order statistics, a total of four sets of $q-s+1$ log-likelihood ratios, defined as l_m, l_s, l_v, l_k, can be obtained. For each set of log-likelihood ratios, determine its confidence limit using methods like kernel density estimation, or simply using the 95% or 99% quantiles as the confidence limits.

4.2. Online Monitoring Stage

Once the confidence limits for the four SPs have been obtained, it is now possible to perform online monitoring on new data samples. For a new sample $\mathbf{x}_{N+1}$, a new SP vector can be obtained as $\mathbf{s}_{N+1} = [\boldsymbol{\mu}_{N+1}\ \boldsymbol{\Sigma}_{N+1}\ \boldsymbol{\Xi}_{1,N+1}\ \boldsymbol{\Xi}_{2,N+1}]$ can be estimated. Based on $\mathbf{s}_{N+1}$ and its previous w-1 SP vectors, four SP matrices corresponding to the first-, second-, third-, and fourth-order statistics can be obtained, following the moving window approach discussed in Section 3.

Empirical likelihood tests are then performed between the SP matrices corresponding to the new data sample and those in $\mathbf{S}_1$ to get the log-likelihood ratios of l_m, l_s, l_v, l_k, which are used as monitoring statistics. If either of the four log-likelihood ratios exceed the confidence limit, a fault is detected and an alarm is triggered.

The advantages of the proposed empirical likelihood based monitoring strategy are double folded. In addition to reducing the size of the problem, it can provide a better understanding of the process changes. For example, if a change in first-order statistical patterns is detected, then the mean values of the variable is abnormal. If the second-order statistical patterns change, the fault causes an anomaly in variances or correlation structure. Third-order statistical patterns such as skewness are used to measure the degree of skewness of the data. If an anomaly occurs, it indicates that the process variables do not follow Gaussian distribution. Fourth-order statistical patterns such as kurtosis can be used to test the nonlinearity of the process data. If there is a violation, it indicates that the process becomes nonlinear.

5. Application Study

This section examines the performance of the improved statistical pattern analysis method on fault detection of the Tennessee Eastman (TE) process. The TE process is a simulation of a chemical process that has been widely used to test the performance of fault diagnosis and process control technologies. It was initially published by the Tennessee Eastman company [18] for academic research. The process contains 12 manipulated variables and 41 measured variables. There are four main operation units: reactor, condenser, compressor, and product separator. Four reactants labeled as Feed A, Feed C, Feed D, and Feed E were fed into the process and two products were produced. In addition, the process defines different constraints, process disturbances, and operating modes. Figure 2 presents a simplified diagram of the process.

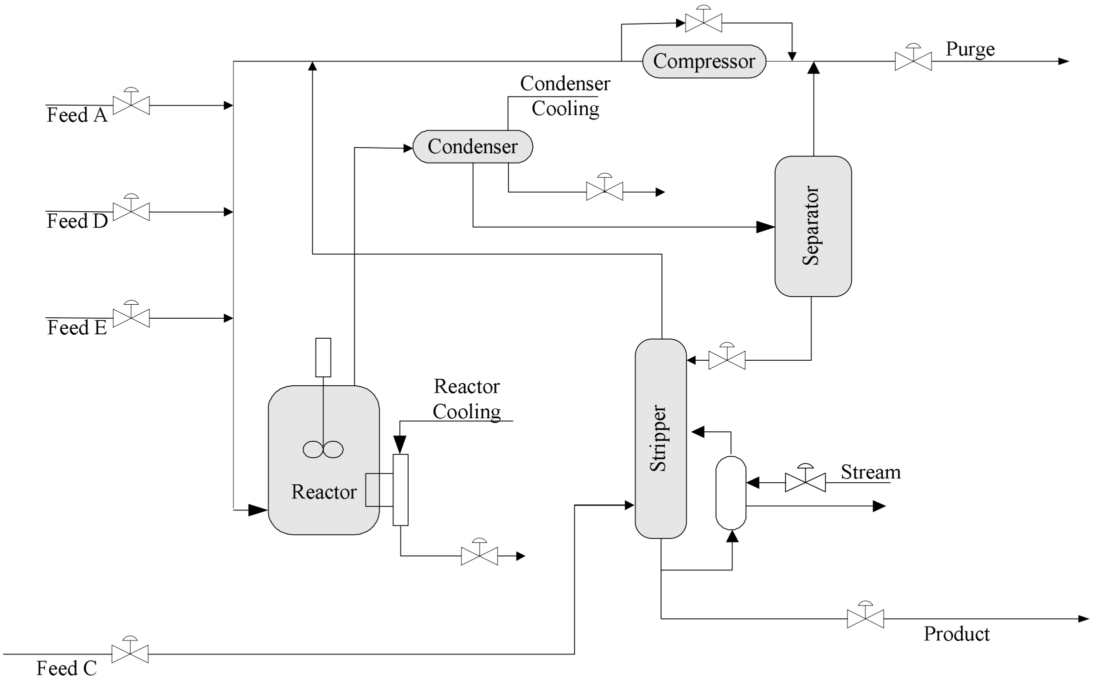

Figure 2. Simplified diagram of the Tennessee Eastman (TE) process.

Following the recommendation of [19], 22 measured variables and 11 manipulated variables were selected for process monitoring, as listed in Table 1.

For the purpose of fault detection, a normal dataset containing 500 samples was generated. A typical fault (i.e., fault 5) was considered and tested, which contained 960 samples. Fault 5 involves a step change in the condenser cooling water inlet temperature, which was introduced after the 160th sample. During the fault, the cooling ability was influenced and hence a change in the vapor–liquid ratio of the input flow separator was observed.

Table 1. Selected process variables for process monitoring. (RCW: reactor cooling water, CCW: condenser cooling water).

No.	Variable	No.	Variable	No.	Variable
x_1	A feed	x_{12}	Product separator level	x_{23}	D feed flow valve
x_2	D feed	x_{13}	Product separator pressure	x_{24}	E feed flow valve
x_3	E feed	x_{14}	Product separator underflow	x_{25}	A feed flow rate
x_4	Total feed	x_{15}	Stripper level	x_{26}	total feed flow valve
x_5	Recycle flow	x_{16}	Stripper pressure	x_{27}	compressor recycle valve
x_6	Reactor feed rate	x_{17}	Stripper underflow	x_{28}	purge valve
x_7	Reactor pressure	x_{18}	Stripper temp.	x_{29}	separator pot liquid flow valve
x_8	Reactor level	x_{19}	Stripper steam flow	x_{30}	stripper liquid product flow valve
x_9	Reactor temp.	x_{20}	Compressor work	x_{31}	stripper steam valve
x_{10}	Purge rate	x_{21}	RCW outlet temp.	x_{32}	RCW flow
x_{11}	Product separator temp.	x_{22}	CCW outlet temp.	x_{33}	CCW flow

For the purpose of fault detection, both window lengths of w, s were set as 50. Statistical pattern matrices involving mean value, variance, skewness, and kurtosis were constructed. Empirical likelihood was then used to construct the monitoring statistic for each of the pattern matrices and four statistics l_m, l_s, l_v, l_k were then constructed. The monitoring results are shown in Figure 3.

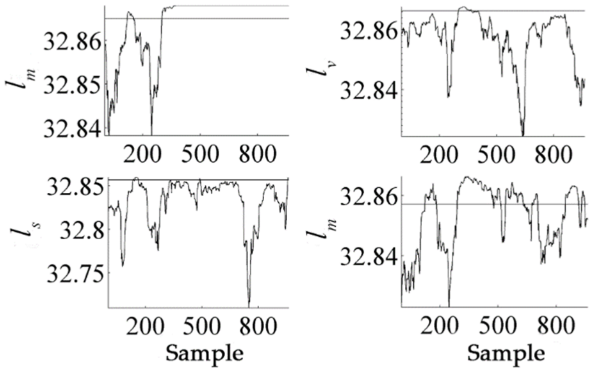

Figure 3. Monitoring results of the improved method for fault 5(the horizontal lines correspond to the confidence limits of the monitoring statistics).

In Figure 3, l_m is the monitoring statistic for the first order statistics (mean value); l_v is for the second order statistics (variance); l_s is for the third order statistics (skewness); and l_k is for the fourth order statistics (kurtosis). It should be noted that due to the introduction of two moving windows, the length of the monitoring statistics was reduced from 960 to 860, indicating that if correctly detected, the fault will cause violations in the statistics of Figure 3 after the 60th sample. From Figure 3, it can be seen that some violations could be observed for l_m and l_k after the 60th sample. At a later time, the fault becomes more severe and a significant number of violations were observed for l_m, especially after the 250th sample. Hence, the fault was successfully detected by the statistic l_m, which indicates that the fault influences both the mean values of process variables, resulting in a certain degree of nonlinearity in the process data. It should also be noted that the monitoring statistic l_m becomes a constant in the later stage. This was due to the fact that at the later stage, the impact of the fault on the mean values becomes so severe that the estimated log-likelihood values become very great. Hence, a cut-off line was introduced.

On the other hand, the second order statistic l_v and third order statistic l_s could hardly detect any violation, indicating no change happened in the second and third order statistical patterns. In addition, fault 5 also influenced the fourth order statistic l_k, indicating that the fault introduced some kinds of process nonlinearity. This can be explained as step change in the cooling water inlet temperature may influence the process setpoint and introduce process nonlinearity.

After the fault is detected, it is important to isolate which variable is affected by the fault. Since few violations are detected in the skewness and kurtosis in this monitoring model, they are not considered in fault isolation. The deviations of mean values and kurtoses between normal and faulty samples for the 33 variables are shown in Figure 4. For a clearer inspection, Table 2 presents the variables with the most significant deviations in mean values and kurtoses due to the fault.

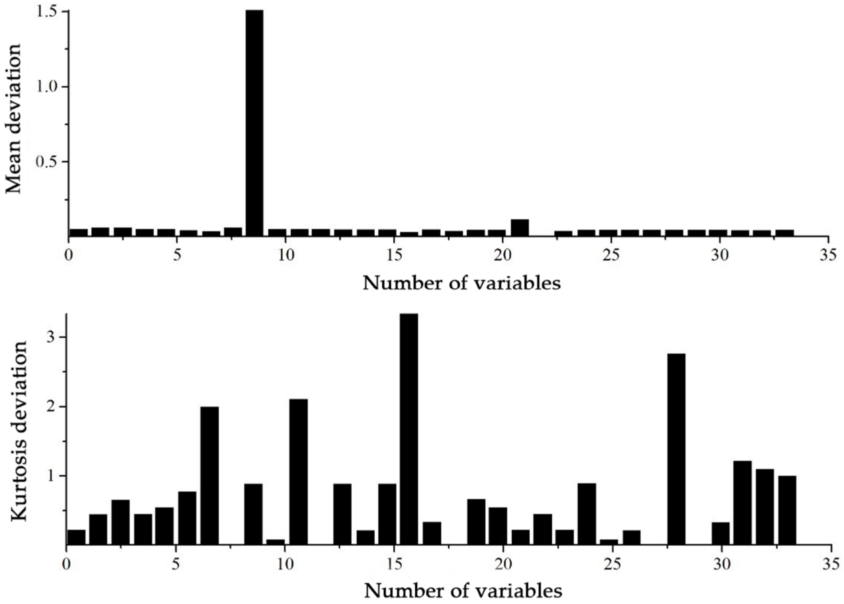

Figure 4. Monitoring results of statistics corresponding to a single variable in fault 5.

Table 2. Variables with the most significant deviations in mean values and kurtoses due to fault 5.

Statistics	Variables with the Most Significant Deviations in Means and Kurtoses					
Mean	x_9	x_{21}				
Kurtosis	x_{16}	x_{28}	x_{11}	x_7	x_{31}	x_{32}	x_{33}

It can be seen from Table 2 that the variables x_9, x_{21}, x_{16}, x_{11}, x_7, x_{31}, x_{32}, and x_{33} had significant deviations. This can be explained as fault 5 involves a step change in the inlet temperature of the condenser cooling water, so the reactor cooling water outlet temperature x_{21} of the cooling water of the separator will be affected. In addition, since the gas flow is cooled by the condenser and fed into the separator, it will cause the product separator temperature x_{11}, the separator cooling water flow x_{33}, and the separator pot liquid flow x_{29}.The separated steam is recycled into the reactor through the compressor, causing changes in the reactor temperature x_9, reactor cooling water flow rate x_{32}, and reactor cooling water outlet temperature x_{21}.

For comparison, the SPA monitoring method proposed in [14] was also tested. In SPA monitoring, four sets of statistical patterns (i.e., mean, variance, skewness and kurtosis) are considered and a SP matrix with 132 statistical patterns was obtained by setting the window length as 50. It was found that retaining eight principal components was sufficient to capture 95.3% variance. Hence, the number of principal components was set as eight. The monitoring results are shown in Figure 5. It can be seen from Figure 5 that the fault was successfully detected by the T^2 and Q statistics of SPA monitoring, however, it did not reveal any information on fault type. Furthermore, the SP matrix included a total of 132 patterns, which may cause difficulties in subsequent fault isolation.

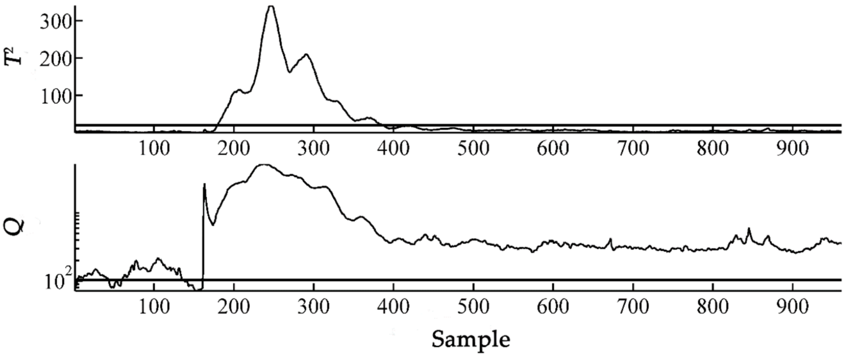

Figure 5. Monitoring results of the statistical pattern analysis (SPA) approach for fault 5.

It can be seen from Figures 3 and 5 that the improved statistical pattern analysis based on empirical likelihood not only successfully detected the fault, but could also learn the type of fault and the specific information of faulty variables while reducing the scale of the monitoring model, leading to a simple and parsimonious model. Although the SPA monitoring method in [14] also successfully detected the step fault, it could not provide any further fault information. Therefore, the improved statistical pattern analysis based on empirical likelihood is more effective and comprehensive than the SPA monitoring method. One additional issue about the proposed method is that it involves a solution of an optimization problem for each online sample, resulting in greater computation load. In this case study, the average computation time for the empirical likelihood method was 0.0325s on a personal computer with the configuration of an Intel (R) Core (TM) i7-6700 CPU @ 3.40 GHz, RAM: 8.0 GB, which is acceptable for online application. Hence, the computation load does not pose a serious problem for our method.

6. Conclusions

This article proposed an improved statistical pattern analysis monitoring method based on empirical likelihood. The basic idea is to monitor the statistical patterns with the same order independently using empirical likelihood. As a result, changes in statistical patterns with specific orders can be detected independently and faults occurring in each order can be isolated.

Compared to the original SPA framework, the improved method reduced the scale of the monitoring problem and provides more information on the faulty conditions. A case study on the TE process demonstrated that the improved statistical pattern analysis monitoring strategy detected the different effects of faults on various statistics, provided more fault information, and is suitable for the monitoring of complex multivariate processes with the change in statistics of each order.

Author Contributions: Conceptualization, J.S. and W.C.; Methodology, J.S.; Software, J.S.; Validation, J.S., X.Z., and W.C.; Formal analysis, J.S.; Investigation, J.S. and X.Z.; Resources, J.S. and W.C.; Data curation, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NQI Scientific Research Project of Zhejiang Province Market Supervision Administration (grant No. 20190106), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (grand No. U1709210), National Natural Science Foundation of China (grand No. 51875529, Key Research and Development Project of Zhejiang Province (grand No. 2019C03108).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S.N. A review of process fault detection and diagnosis, part II: Qualitative models and search strategies. *Comput. Chem. Eng.* **2003**, *27*, 313–326. [CrossRef]
2. Liu, Y.; Zeng, J.; Xie, L.; Luo, S.; Su, H. Structured joint sparse principal component analysis for fault detection and isolation. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2721–2731. [CrossRef]

3. Ouyang, Y. Evaluation of river water quality monitoring stations by principal component analysis. *Water Res.* **2005**, *12*, 2621–2635. [CrossRef] [PubMed]
4. Jolliffe, I.T. A note on the use of principal components in regression. *Appl. Stat.* **1982**, *31*, 300–303. [CrossRef]
5. Gemperline, P.J.; Long, J.R.; Gregoriou, V.G. NonIine multivariate calibration using principle components regression and artificial neural networks. *Anal. Chem.* **1991**, *63*, 313–323. [CrossRef]
6. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]
7. Zhu, E.Y. A kind of PLS method suitable to deal with the fingerprinting data of Chinese medicine. *Comput. Appl. Chem.* **2005**, *22*, 639.
8. Chen, K.X.; Shen, J.Z. *Modern Digital Theory*; Zhejiang University Press: Hangzhou, China, 2001.
9. Lee, J.M.; Yoo, C.K.; Lee, I.B. Statistical process monitoring with independent component analysis. *J. Process Control* **2004**, *14*, 467–485. [CrossRef]
10. Hyvarinen, A.; Oja, E. Independent component analysis: Algorithms and application. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
11. Yang, J.; Gao, X.; Zhang, D.; Yang, J.Y. Kernel ICA: An alternative formulation and its application for face recognition. *Pattern Recognit.* **2005**, *38*, 1784–1787. [CrossRef]
12. Kano, M.; Tanaka, S.; Hasebe, S.; Hashimoto, L.; Ohno, H. Monitoring independent components for fault detection. *AIChE J.* **2003**, *49*, 969–976. [CrossRef]
13. Chiang, L.H.; Russeil, E.L.; Braatz, R.D. *Fault Detection and Diagnosis in Industrial Systems*; Springer: London, UK, 2001.
14. Wang, J.Q.; Peter, H. Multivariate statistical process monitoring based on statistics pattern analysis. *Ind. Eng. Chem. Res.* **2010**, *49*, 7858–7869. [CrossRef]
15. Wang, Q. Development summary of empirical likelihood inference approach. *Adv. Math.* **2004**, *33*, 141–151.
16. Liu, Y.Q.; Zou, C.L.; Zhang, R.C. Empirical likelihood for the two-sample mean problem. *Stat. Probab. Lett.* **2008**, *78*, 548–556. [CrossRef]
17. Xu, L.; Ding, X.W.; Lin, J.G. Statistical diagnostics for partially linear models based on empirical likelihood. *Chin. J. Appl. Probab. Stat.* **2011**, *27*, 91–102.
18. Downs, J.J.; Vogel, E. A plant-wide industrial process control problem. *Comput. Chem. Eng.* **1993**, *17*, 245–255. [CrossRef]
19. Xie, L.; Lin, X.; Zeng, J. Shrinking principal component analysis for enhanced process monitoring and fault isolation. *Ind. Eng. Chem. Res.* **2013**, *52*, 17475–17486. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.