Control, Automation and Instrumentation An integrated process monitoring approach combining dynamic independent component analysis and local outlier factor

Elham Tavasolipour[1], Mohammad Taghi Hamidi Beheshti[2*], and Amin Ramezani[3]

Received: 2015/6/7 Accepted:2015/7/11

Abstract

In this paper a novel process monitoring scheme for reducing the type I and type II error rates in the monitoring phase is proposed. First, the proposed approach uses an augmented data matrix to implement the process dynamic. Then, we apply independent component analysis (ICA) transformation to the augmented data matrix, and eliminate the outliers using the local outlier factor (LOF) algorithm. Finally, the control limit based on the LOF value of the cleaned data are obtained. In the monitoring phase, if the LOF value of each sample exceeds the control limit, fault has occurred; otherwise, data is normal. The proposed method is applied to fault detection in both a simple multivariate dynamic process and the Tennessee Eastman process. In both processes, type I and type II error rates are witnessed to reduce by considering the process dynamic and performing the LOF algorithm. Results clearly indicate better performance of the proposed scheme compared to the alternative methods.

Keywords: Local Outlier Factor; Independent Component Analysis; Tennessee Eastman process; Fault detection.

I. 1. INTRODUCTION

Recently, principal component analysis (PCA) has come to a wide use for monitoring multivariate processes. PCA is a dimension reduction technique that transforms the source signals into principal components (PCs) and uses the statistics such as T^2 and SPE to monitor the processes. In some applications, PCA is integrated into other methods aiming at a better performance. For example, PCA and partial least square (PLS) have been extended for fault detection in different applications [1–5]. Recently, kernel PCA (KPCA) has emerged as a nonlinear process monitoring technique for fault detection and identification that does not include nonlinear optimization [6,7]. PCA considers the Gaussian distribution for latent variables, although authors in [8] showed that PCA-extracted components rarely conform to a multivariate Gaussian distribution in many real industrial processes.

More recently, independent component analysis (ICA) has been introduced which can be considered as an extension of PCA. ICA takes the non-Gaussian distribution for latent variables and reconstructs the source signals into independent signals [9]. To enhance the fault detection performance, ICA and support vector machine (SVM) are integrated [10,11]. Kernel independent component analysis (Kernel ICA), on the other hand, has been proposed by Wang and Shi [12]. This approach uses kernel ICA to elicit the independent components accurately. Both PCA and ICA have the common limitation of considering a special distribution, Gaussian and non-Gaussian distribution respectively, for latent variables, whereas the variables in real industrial processes have mixture distribution [13]. In a few studies both PCA and ICA have been applied in two step: in one study, Kernel PCA and Kernel ICA have been used for fault detection by taking both Gaussian and non-Gaussian distribution in real industrial processes into account. In the second step, SVM is used to diagnose faults [14].

In addition, both PCA and ICA assume that the observations at one time are statistically independent from those in the past. This

1.Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.
2. Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran. .mbehesht@modares.ac.ir
3. Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.

assumption does not seem to be valid, because the data in real industrial processes have dynamic characteristics. Ku et al. [15] proposed dynamic PCA (DPCA) that utilizes an augmented matrix with time-lagged variables. Lee et al. [16] further extended DPCA to ICA and arrived at a new approach called dynamic independent component analysis (DICA). In this approach, first the data matrix is augmented with time-lagged variables and then ICA is applied to them. Results show that DICA might have better performance than ICA. In another study, Monroy et al. [17] combined DICA with SVM to improve the performance of fault diagnosis.

Another problem to be addressed in PCA and ICA is the effect of outliers, which has to be eliminated. The outliers can increase the type II error rate through enlarging the control limit. Recently, a novel process scheme, called Adjusted Outlier (AO), is proposed for ICA based on rectangular type measure, rather than elliptical type measure, to monitor processes [18]. The proposed scheme applies the AO algorithm to eliminate the outliers and to calculate the control limit. However, in this work, the correlation between variables at different times is ignored and it is assumed that variables do not have any dynamic characteristics. On the contrary, variables are dynamically related in industrial processes. To compensate for this limitation, the process dynamic is augmented to the ICA(AO) [19], in which both process dynamic and effect of outliers are considered. In the AO algorithm, the limitation of rectangle type measure does not seem to allow an accurate estimation of the nonlinear feature space boundary of normal operating condition (NOC). In [20], authors proposed a new process monitoring scheme by integrating ICA and local outlier factor (LOF). In this approach, the decision boundary of NOC can be determined more accurately by applying the LOF algorithm which also eliminated the outliers. Nevertheless, in the aforementioned study, the process dynamic has been ignored. It is evident that all these methods are carried out with a trade-off between the type I and type II error rates. In other words, the type I and type II error rates cannot be decreased concurrently.

To overcome these limitations, in the present paper, an integrated approach is proposed by combining dynamic ICA and LOF that is performed by first considering the process dynamic and then eliminating the effect of outliers. The elimination of outliers is performed based on the LOF algorithm. Here, the main advantage might be the fact that this algorithm does not consider a special distribution for variables. Therefore, this algorithm attributes the degree of being an outlier for both Gaussian and non-Gaussian distributions, thus conforming to the data in real industrial processes. Since the control limit, which is determined by the LOF algorithm, is a non-linear boundary compared to the NOC, it can be more accurate. This may, however, increase the type I error rate. In this paper, the type I and type II error rates were reduced by considering the process dynamic for type I, and by performing the LOF algorithm for type II, respectively. This, in turn, resulted in an enhancement of the process performance.

The remainder of this paper is organized as follows. In the next section, ICA algorithm is briefly introduced. After introducing the LOF algorithm in section 3, the proposed scheme is discussed in section 4. The experimental results are further presented in section 5. Finally in section 6, we present point out the concluding remarks.

II. 2. ICA-BASED PROCESS MONITORING

ICA is a new method that has recently been developed, in order to find a linear combination of independent data. This algorithm has enormous popularity among the dimension reduction techniques for its ability to extract the independent variables with a large amount of original data compared to the other methods. The algorithm converts the data into components which are statistically independent, or as independent as possible. In the ICA algorithm, it is assumed that the m variables can be measured by the linear combination of d unknown independent variables and may be presented as below:

$$x_1, x_2, \ldots, x_m \quad \Longrightarrow \quad s_1, s_2, \ldots, s_d \quad d \leq m \qquad (1)$$

Independent and measured variables have zero mean and the relationship between them is shown in the following equation:

$$\mathbf{X} = \mathbf{AS} \qquad (2)$$

Where **X**, **A**, **S** are the data matrix, the unknown composition matrix, and the independent

variables matrix respectively as follows:

$$\boldsymbol{X} = [\mathrm{x}(1), \mathrm{x}(2), \dots, \mathrm{x}(n)] \subset R^{m \times n} \quad (3)$$

$$\boldsymbol{S} = [\mathrm{s}(1), \mathrm{s}(2), \dots, \mathrm{s}(n)] \in R^{d \times n}$$

$$\boldsymbol{A} = [\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_d] \in R^{m \times d}$$

In the above equations, *m* is the number of variables and *n* is the number of measurements, and *d* is the number of independent components. The main problem in ICA is to find both matrix of **A** and the independent components matrix of **S** with only the data matrix **X** available. Also, the purpose of the ICA can be expressed as follows: The objective is to find a matrix **W** such that the $\widehat{\boldsymbol{S}}$ reconstruction matrix becomes as independent as possible according to the following equation:

$$\widehat{\boldsymbol{S}} = \boldsymbol{W}\boldsymbol{X} \quad (4)$$

In the above equation, the matrix of $\boldsymbol{W}$ is the inverse of the matrix $\boldsymbol{A}$. A useful algorithm for ICA is the FastICA that is demonstrated in [9]. The related software (FastICA toolbox) can be downloaded from (http://www.cis.hut.fi/projects/ica/fastica/).

III. 3. LOF ALGORITHM

In this section, we outline the details of the LOF algorithm [21]. Then the reason for its functionality for process monitoring is stated. This algorithm can be applied for eliminating the outliers and computing the control limit for the proposed monitoring scheme.

Definition 1 (*k-distance* of an object p): *k-distance* of an object *p* is denoted as *k-distance*(p) and is equal to Euclidian distance of *k*-th object in the neighbourhood of an object p. Where, *k* can be any positive integer.

Definition 2 (*k-distance* neighbourhood of an object p): *k-distance* neighbourhood of an object p, contains every object whose distance from p is not greater than *k-distance* itself and is denoted by $N_{k\text{-}distance}(\mathrm{p})$. (5)

$$N_{k\text{-}distance}(\mathrm{p}) = \{\{\mathrm{q} \in D \backslash \{\mathrm{p}\} \mid d(\mathrm{p},\mathrm{q}) \leq k\text{-}distance(\mathrm{p})\}\}$$

Henceforth, we use $N_k(\mathrm{p})$ instead of $N_{k\text{-}distance}(\mathrm{p})$ in our notation.

Definition 3 (reachability distance of an object p): The reachability distance of an object p to object o in $N_k(\mathrm{p})$ is defined as follows: (6)

$$reach\text{-}dist_k(\mathrm{p},\mathrm{o}) = max\{k\text{-}distance(\mathrm{o}), d(\mathrm{p},\mathrm{o})\}$$

Definition 4 (local reachability density of an object p): The local reachability density of an object p is defined as follows:

$$lrd_k(\mathrm{p}) = 1/\left[\frac{\sum_{\mathrm{o} \in N_k(\mathrm{p})} reach\text{-}dist_k(\mathrm{p},\mathrm{o})}{N_k(\mathrm{p})}\right] \quad (7)$$

Definition 5 (local outlier factor (LOF) of an object p): The local outlier factor of an object p is defined as follows:

$$LOF(\mathrm{p}) = \frac{\sum_{\mathrm{o} \in N_k(\mathrm{p})} \frac{lrd_k(\mathrm{o})}{lrd_k(\mathrm{p})}}{N_k(\mathrm{p})} \quad (8)$$

If object p would be in the neighbouring of other objects in $N_k(\mathrm{p})$, the LOF(p) will become close to 1. because the ratio of the average density of objects in $N_k(\mathrm{p})$ is near the density of p. If p is an outlier, the LOF(p) becomes larger than 1. since the difference between the numerator and denominator in LOF(p) becomes very great.

The performance of elliptical type measurements like I^2 is useful when all of the independent variables conform to Gaussian variables. But if all of the independent variables conform to non-Gaussian distributions, the performance of rectangular measurements, like that of AO, will be employed rather than the elliptical-type measurements. This is despite the fact that the variables have both distributions (Gaussian and non-Gaussian distributions) in real industrial processes. Both in the elliptical type and the rectangular type measurements the fault detection boundaries are much larger than the actual NOC region. Since elliptical or rectangular distance type measure considers specific distribution for latent variables, their decision boundaries are hard to be located near the border of samples. This in turn results in an increase in type II errors, but by adopting LOF as the monitoring statistic, the decision boundary is

extracted along with the border of NOC regions [20], thus type II errors successfully decrease in comparison with elliptical or rectangular distance type measures. However, since only the normal samples around the border can be detected as faults when the decision boundary is too close to the border, LOF may cause some type I errors [20]. In this paper, the consideration of the process dynamic resulted in a reduction of type I error rate and the enhancement of process performance.

IV. 4. PROPOSED PROCESS MONITORING SCHEME

In this section, we introduce the proposed monitoring scheme that considers both process dynamic and the effect of outliers. Figure 1 illustrates the procedure for the proposed monitoring approach which contains two different phases: build-time and run-time modelling phases.

First, the original data matrix is augmented with time-lagged variables in order to take the process dynamic into account. Then ICA transformation is performed to reduce the dimension. In the next step, the LOF of the ICs are computed for eliminating the outliers and thus cleaning the data; afterwards the ICA and LOF algorithms are repeated to determine the control limit for online process monitoring. Throughout the run-time monitoring, first, the data at one time is augmented with time-lagged variables, and then the LOF computation is performed. If the LOF value exceeds the control limit, fault is detected; otherwise, data is normal, and the algorithm is repeated again.

Figure 1. The flow chart of dynamic ICA (LOF).

A. 4.1 Modeling phase

Step1: Obtain a training data set $\boldsymbol{X} \in R^{m \times n}$, where m and n are the number of variables and observations respectively.

Step2: Determine the time lag l and augment each observation vector with the previous observations and demonstrate the data matrix in the following form: (9)

$$\mathbf{X}(l) = \begin{bmatrix} \boldsymbol{x}_t^T & \boldsymbol{x}_{t-1}^T & \cdots & \boldsymbol{x}_{t-l}^T \\ \boldsymbol{x}_{t+1}^T & \boldsymbol{x}_t^T & & \boldsymbol{x}_{t-l+1}^T \\ \vdots & & \ddots & \vdots \\ \boldsymbol{x}_{t+n-1}^T & \boldsymbol{x}_{t+n-2}^T & \cdots & \boldsymbol{x}_{t+n-1-l}^T \end{bmatrix}$$

Where, $\boldsymbol{x}_t^T$ is the m-dimensional observation vector at time t and T is the transpose operator. Researchers have agreed that a value of l=1 or 2 is usually appropriate for lagged variables. In this method l=2 is chosen for process dynamic.

Step3: Perform FastICA algorithm, so that a de-mixing matrix **W** can be obtained. The estimated ICs can be shown as follows:

$$\hat{\mathbf{S}}_0 = \mathbf{W}_0\mathbf{X}(l) = [\hat{s}_0(1)\ \hat{s}_0(2) \ldots \hat{s}_0(n)] \quad (10)$$

Step4: Apply the LOF algorithm for each $\hat{s}_0(i)$ as follows: (11)

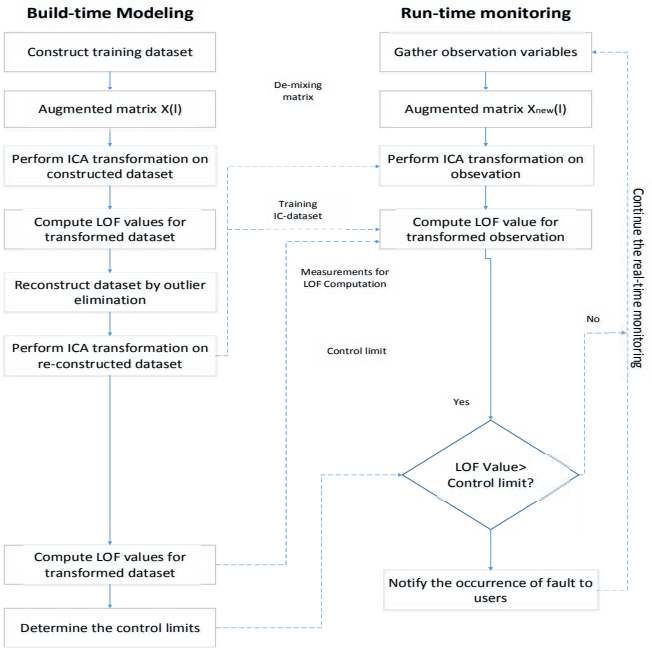

Figure 1. The flow chart of dynamic ICA (LOF).

$$LOF(\hat{\mathbf{S}}_0) = [LOF(\hat{\mathbf{s}}_0(1)) \;\; LOF(\hat{\mathbf{s}}_0(2)) \ldots LOF(\hat{\mathbf{s}}_0(n))]$$

Where (12)

$$LOF(\hat{\mathbf{s}}_0(i)) = \frac{\sum_{o \in N_k(\hat{s}_0(i))} \frac{lrd_k(o)}{lrd_k(p)}}{N_k(p)} \quad \text{for}$$

$i = 1,2,\ldots,n$

Step5: Eliminate the outliers by 99.3% limit determined by Kernel Density Estimation (KDE). A univariate kernel estimator with kernel k is defined by (13)

$$\hat{f}(x) = \frac{1}{nh}\sum_i k\left(\frac{x - x_i}{h}\right)$$

Where, x is the data point under consideration, x_i is the observation, h is the window width (also known as the smoothing parameter), n is the number of samples and k is the kernel function. The type of the kernel function is not of high significance and the Gaussian kernel is widely used for that matter [22]. In the present study, we utilize the Gaussian kernel. Some of the n samples are eliminated from the normal samples by following this step. Therefore, assume that the n' samples are remained, so that $\mathbf{X}' \in R^{m \times n'}$.

Step6: Reconstruct $\mathbf{X}'$ with time-lagged variables and denote as $\mathbf{X}'(l)$ and again, perform the FastICA algorithm, then obtain the $\mathbf{W}_1$ matrix for monitoring phase as follows: (14)

$$\hat{\mathbf{S}}_1 = \mathbf{W}_1 \mathbf{X}'(l) = [\hat{\mathbf{s}}_1(1) \ldots \ldots \ldots \hat{\mathbf{s}}_1(n')]$$

Step7: Apply the LOF algorithm for $\hat{\mathbf{S}}_1$ and determine 99% control limit by KDE method.

B. 4.2 Monitoring phase

Step1: Obtain a new data matrix, $\mathbf{X}_{new}$.

Step2: Augment the data matrix with lag l and denote as $\mathbf{X}_{new}(l)$.

Step3: Calculate ICs by $\hat{\mathbf{s}}(t) = \mathbf{W}_1 \mathbf{X}_{new}(l)$.

Step4: Compute the LOF value of $\hat{\mathbf{s}}(t)$ and denote as $LOF(t)$, (15)

$$LOF(t) = \frac{\sum_{o \in N_k(\hat{s}(t))} \frac{lrd_k(o)}{lrd_k(\hat{s}(t))}}{N_k(\hat{s}(t))}$$

In this phase, we applied an approximated LOF algorithm rather than the original algorithm for the purpose of reducing computational costs whose details are outlined in [19].

Step5: Compute whether the fault is occurred or not. If $LOF(t)$ exceeds the control limit, fault is detected; otherwise, the sample contains normal data.

V. 5. EXPERIMENTS

In this section, the efficiency of the proposed method with two process data sets is investigated. First, the proposed method is applied to a simple multivariate dynamic process. Then, the Tennessee Eastman (TE) process is tested. In this paper, three schemes are implemented to compare different performances of fault detection; schemes 1 and 2 are extracted from [17] and scheme 3 is extracted from [20].

Scheme 1 (ICA(I^2)): performs ICA transformation first, then eliminates the outliers with 99.3% limit determined by I^2 values with the KDE method, then uses I^2 monitoring statistic to determine the 99% control limit.

Scheme 2 (ICA(*AO*)): first, performs ICA transformation and then eliminates the outliers with AO rejection rule. Each sample in the modeling phase which exceeds 99.3% of the limit determined with the KDE method, is eliminated. Finally, the 99% control limit of AO values by the KDE method is determined.

Scheme 3 (ICA(LOF)): performs ICA transformation first and then eliminates the outliers with the LOF rejection rule. Each sample in the modeling phase which exceeds 99.3% of the limit determined with the KDE method, is eliminated. Finally, the 99% control limit of LOF values by the KDE method is determined.

In both scheme 3 and the proposed approach, the k parameter used for eliminating outliers and for computing the control limit in the LOF algorithm was determined as 20.

Note that the outliers must be removed from the training data set $\mathbf{X}$, before augmenting the time-lagged variables. Eliminating the outliers from $\mathbf{X}(l)$ may not allow for a complete rejection of the effect of outliers; moreover, this may derive incorrect results in the monitoring phase.

A. 5.1 Simple multivariate process

This simulation process has five variables with autocorrelation, which was developed by Ku et al. [15]. (16)

$$\boldsymbol{z}(k) = \begin{bmatrix} 0.118 & -0.191 & 0.287 \\ 0.847 & 0.264 & 0.943 \\ -0.333 & 0.514 & -0.217 \end{bmatrix} \boldsymbol{z}(k-1) + \begin{bmatrix} 1 & 2 \\ 3 & -4 \\ -2 & 1 \end{bmatrix} \boldsymbol{u}(k-1)$$

$$\boldsymbol{y}(k) = \boldsymbol{z}(k) + \boldsymbol{v}(k) \quad (17)$$

Where $\boldsymbol{u}$ is the correlated input with:

(18)

$$\boldsymbol{u}(k) = \begin{bmatrix} 0.811 & -0.226 \\ 0.477 & 0.415 \end{bmatrix} \boldsymbol{u}(k-1) + \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix} \boldsymbol{w}(k-1)$$

Where $\boldsymbol{w}$ is a random vector that is uniformly distributed over the interval (-2, 2). The output $\boldsymbol{y}$ is equal to $\boldsymbol{z}$ plus a random noise $\boldsymbol{v}$, which has been normally distributed with a zero mean and a variance of 0.1. Both $\boldsymbol{y}$ and $\boldsymbol{u}$ are the output of the process. In the modeling phase, we trained the monitoring scheme using normal data with 500 samples. Then we tested the performance of the proposed method using fault data with 800 samples. In this simulation, fault was generated by changing the mean shift of the first element $(\boldsymbol{w}_1)$ of $\boldsymbol{w}$, where a fault is introduced from sample 201 and continued to the end.

When the step change becomes larger, the effect of fault is more pronounced in the process. In this paper, we examine the variation of the step size from 0.5 to 3. The monitoring results of the above four schemes are summarized in Table 1. The type I error rate is the rate of the misclassified normal samples to the total normal samples from observation 1 to 200. The type II error rate is the rate of the misclassified fault samples to the total fault samples from observation 201 to 800. Among these schemes, ICA(I^2) and ICA(AO) displayed higher type II error rate. This is because they use the elliptical and the rectangular type measurements, respectively, which are not appropriate for non-Gaussian variables and cause increase in type II error rate. In the LOF algorithm, the control limit is a non-linear boundary toward normal samples, therefore it can be more accurate and may result in a decrease in the type II error rate; however, since the control limit is too close to the border and only the normal samples around the border can be detected as faults, the type I error rate may increase. Because of this limitation in ICA(LOF), type I error rate is higher than that of DICA(LOF). In DICA(LOF), adding the dynamic process reduces sensitivity of the control limit toward normal samples, and type I error rate decreases thus, enhancing the process performance.

Figure 2 illustrates the monitoring results for step size 0.5 using DICA(LOF). Fault is introduced from sample 201 and continued to the end. The 99% control limits are also shown as dotted lines in all of these figures. In Figure 2, since the step size is too small, fault detection seems to be difficult, but in Figure 3 by increasing the step size to 3, monitoring results can detect the fault from sample 201 to the end. The results of the type II error rate of DICA(LOF) seem to be most appealing among the other monitoring methods.

Table 1. Type I and type II error rates in simple ultivariate process (%).

Step size	ICA(I^2)		ICA(AO)		ICA(LOF)		DIAC(LOF)	
	Type	Type	Type	Type	Type	Type	Type	Type
0.5	1	95	1	93	3	81	2	25
1	0	88	2	86	3	69	2	19
1.5	0	83	1	77	4	50	0	9
2	0	45	1	14	1	14	2	5
2.5	1	29	2	7	1	1	1	1
3	0	14	2	1	3	0	0	0
Average	**0.3**	**59.2**	**1.2**	**46.2**	**2.3**	**35.8**	**1.2**	**9.8**

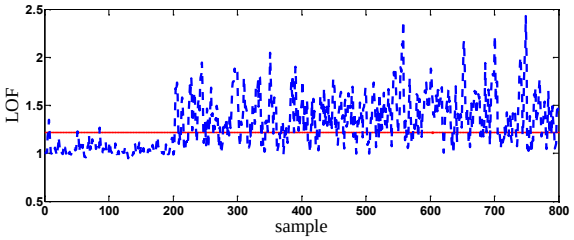

Figure 2. Monitoring result of simple multivariate process with DICA(LOF) in the step size 0.5.

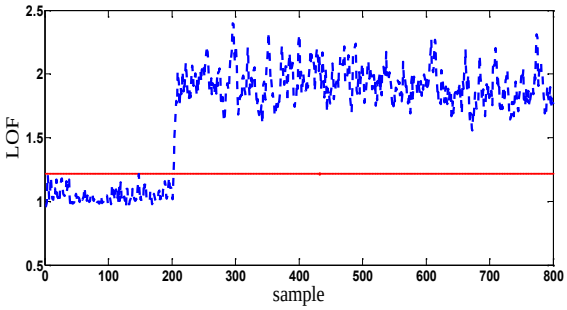

Figure 3. Monitoring result of simple multivariate process with DICA(LOF) in the step size 3.

B. 5.2 Tennessee Eastman process

In this section, the proposed method is applied to the Tennessee Eastman process simulation data and is further compared with other methods. The Tennessee Eastman process was created by the Eastman chemical company for the purpose of studying, developing, and evaluating process control technology [23]. In this paper, we used the same simulation data generated by Chiang et al. [24]. These data can be downloaded from http://brahms.scs.uiuc.edu. The process contains five major unit operations: a reactor, a product condenser, a compressor, a separator, and a striper. The process also involves 22 continuous measurements, 19 composition measurements, and 12 manipulated variables. In this paper, 33 monitored process variables, 22 continuous process measurements and 11 manipulated variables are selected. The simulated data are used for training and testing data sets, and are later collected in the sampling interval of 3 minutes. A set of programmed faults (fault 1-21) is listed in the first column of Table 2. The training and testing data set for each fault consist of 500 and 960 observations, respectively. All faults in the test data set were introduced from sample 161 and were continued to the end. The monitoring results of the four schemes are summarized in Table 2.

Type I error rate is the rate of misclassified normal samples to total normal samples from observation 1 to 160. Type II error rate is the rate of misclassified fault samples to total fault samples from observation 161 to 960. Faults 3, 9 and 15 are not computed in the average rate because they are too small and leave almost no effect. Among all these methods, the proposed method displays the lowest average type II error rate. This is due to the fact this method considers the effect of outliers by applying the LOF algorithm. Although ICA(AO) considers the effect of outliers too, in the AO algorithm, the limitation of the rectangle type measure does not allow accurate estimation of the nonlinear feature space boundary of NOC. In addition, LOF algorithm attributes the degree of outlierness to the data, regardless of their distribution, and conforms to data in real industrial processes. The proposed method displays the lowest average type I error rate. ICA(I^2) and ICA(AO) use elliptical and rectangular type measurements for calculating the control limit. Since these methods consider specific distributions for variables, difference between the actual control limit and the estimated control limit increases significantly. But in DICA(LOF) adding the dynamic process reduces sensitivity of the control limit towards normal samples, so type I error rate decreases and this enhances the process performance in turn. Results indicate that among all four methods, DICA (LOF) causes the lowest type I and type II error rates. This method is also the only method that could decrease the type I and type II error rates concurrently.

Table 3. shows type II error rate by adding the process dynamic to schemes 1 and 2. Among these methods, DICA(AO) displays better performance than DICA(I^2) in fault detection and the proposed method has the lowest type II error rate and shows the best performance.

Table 2. Type I and type II error rates in TE process (%).

Fault no. descriptions	**ICA(I^2)**		**ICA(AO)**		**ICA(LOF)**		**DICA(LOF)**	
	Type I	**Type II**	**Type**	**Type II**	**Type I**	**Type II**	**Type**	**Type II**
1. A/C feed ratio, B composition constant	3	0	0	0	2	0	0	0
2. B composition, A/C ratio constant	2	2	0	2	1	1	0	1
4. Reactor cooling water inlet temperature	5	39	3	16	1	0	1	0
5. condenser cooling water inlet temperature	3	0	1	0	1	0	1	0
6.A feed loss	6	0	0	0	1	0	1	0
7.C header pressure loss	0	1	1	0	2	0	0	0
8.A,B,C feed composition	0	3	0	3	1	1	1	2
10.C feed temperature	0	22	1	18	2	10	2	2
11.Reactor cooling water inlet temperature	3	48	4	30	1	26	1	1
12.condenser cooling water inlet temperature	0	1	5	0	2	0	1	0
13.Reaction kinetics	4	6	0	5	1	4	0	4
14.Reactor cooling water valve	1	0	1	0	2	0	1	0
16.Unknown	3	29	0	22	9	6	3	4
17.Unknown	3	7	2	6	2	4	1	2
18.Unknown	0	10	0	10	2	9	0	9
19.Unknown	1	31	0	20	0	13	0	0
20.Unknown	1	13	1	9	1	17	1	8
21.Valve position constant	4	55	18	38	11	46	3	42
Average	**2.2**	**14.8**	**2.0**	**9.9**	**2.4**	**7.6**	**0.94**	**4.2**

Table 3. Type II error rates in TE process

Fault no. descriptions	**DICA(I^2)**	**DICA(AO)**	**DICA(LOF)**
1. A/C feed ratio, B composition constant	0	0	0
2. B composition, A/C ratio constant	1	1	1
4. Reactor cooling water inlet temperature	3	0	0
5. condenser cooling water inlet temperature	0	0	0
6.A feed loss	0	0	0
7.C header pressure loss	0	0	0
8.A,B,C feed composition	2	2	2
10.C feed temperature	18	10	2
11.Reactor cooling water inlet temperature	46	17	1
12.condenser cooling water inlet temperature	0	0	0
13.Reaction kinetics	5	4	4
14.Reactor cooling water valve	0	0	0
16.Unknown	18	9	4
17.Unknown	10	4	2
18.Unknown	10	10	9
19.Unknown	19	5	0
20.Unknown	12	8	8
21.Valve position constant	54	38	42
Average	**11.0**	**6.0**	**4.2**

Figure 4. to Figure 7. illustrate the monitoring results of faults 4, 7, 16 and 21 using DICA(LOF). Fault is introduced from sample 161 and continued to the end. In Figure 4 and Figure 5, by occurring the fault, statistic increases from control limit and fault can be successfully detected. But in the case of fault 16 (Figure 6), statistic decreases from control limit in some samples. Therefore this fault has 4% of type II

error rate. All these methods display the highest type II error in the case of fault 21. Figure 7 reveals a shortcoming of the proposed method in fault detection: this method cannot detect the fault until sample 600; therefore type II error rate increase.

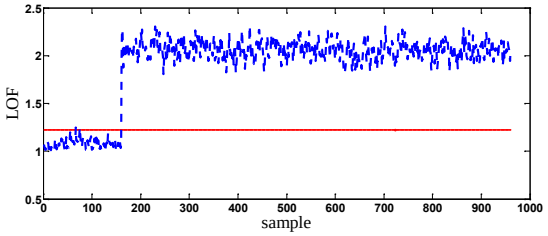

Figure 4. Monitoring result of TE process with DICA(LOF) in the case of fault 4.

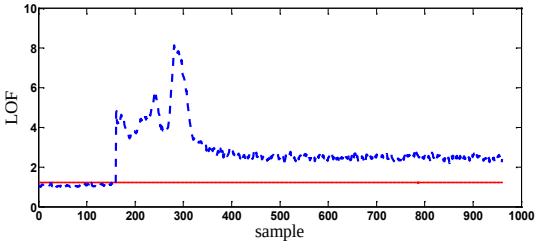

Figure 5. Monitoring result of TE process with DICA(LOF) in the case of fault 7.

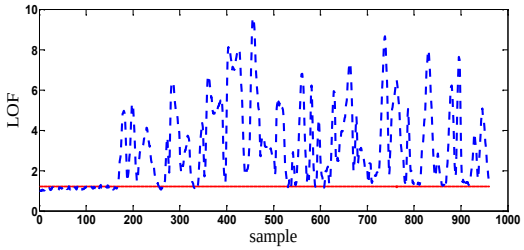

Figure 6. Monitoring result of TE process with DICA(LOF) in the case of fault 16.

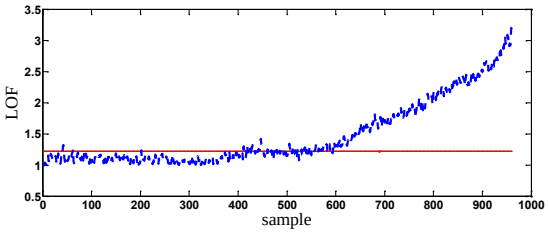

Figure 7. Monitoring result of TE process with DICA(LOF) in the case of fault 21.

I. 6. CONCLUSION

In this paper, we introduce a novel process monitoring scheme integrating DICA and LOF. The advantage of this method is first taking into consideration the process dynamic and then eliminating the effect of outliers by the LOF algorithm. Moreover, this algorithm does not consider any specific distribution for variables which result in conforming the variables to the data in real industrial processes. Additionally, we use LOF monitoring scheme for determining the control limit. The proposed method was applied both to a simple multivariate dynamic process and to the TE process. In both processes, type I and type II error rates are witnessed to reduce by considering the process dynamic and performing the LOF algorithm, respectively. The proposed method has shown superior performance as compared to the alternative methods.

References

[1] B.R. Bakshi. "Multiscale PCA with application to multivariate statistical process monitoring" AIChE J, 44, 1998, 1596–1610.

[2] D. Dong, T.J. McAvoy. "Nonlinear principal component analysis-Based on principal curves and neural networks" Comput. Chem. Eng, 20, 1996, 65–78.

[3] J.V. Kresta, J.F. Macgregor, T.E. Marlin. "Multivariate statistical monitoring of process operating performance" Can. J. Chem. Eng, 69, 1991, 35–47.

[4] W. Li, H.H. Yue, S. Valle-Cervantes, S.J. Qin. "Recursive PCA for adaptive process monitoring" J. Process Control, 10 , 2000, 471–486.

[5] B.M. Wise, N.B. Gallagher. "The process chemometrics approach to process monitoring and fault detection" J. Process Control, 6 , 1996, 329–348.

[6] J.-M. Lee, C. Yoo, S.W. Choi, P.A. Vanrolleghem, I.-B. Lee. "Nonlinear process monitoring using kernel principal component analysis" Chem. Eng. Sci, 59, 2004, 223–234.

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

[8] E.B. Martin, A.J. Morris. "Non-parametric confidence bounds for process performance monitoring charts" Journal of Process Control, 6, 1996, 349–358.

[9] A. Hyvärinen, E. Oja. "Independent component analysis: algorithms and applications" Neural Netw, 13, 2000, 411–430.

[10] C.-C. Hsu, M.-C. Chen, L.-S. Chen. "Intelligent ICA–SVM fault detector for non-Gaussian multivariate process

monitoring" Expert Syst. Appl, 37, 2010, 3264–3273.

[11] Hsu C-C, Chen M-C, Chen L-S. "Integrating independent component analysis and support vector machine for multivariate process monitoring" Computers & Industrial Engineering, 59, 2010, 145–156.

[12] L. Wang, H. Shi. "Multivariate statistical process monitoring using an improved independent component analysis" Chem. Eng. Res. Des, 88, 2010, 403–414.

[13] Z. Ge, Z. Song. "Multimode process monitoring based on Bayesian method" J. Chemometr, 23, 2009, 636–650.

[14] Y. Zhang. "Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM" Chem. Eng. Sci, 64, 2009, 801–811.

[15] W. Ku, R.H. Storer, C. Georgakis. "Disturbance detection and isolation by dynamic principal component analysis" Chemom. Intell. Lab. Syst, 30, 1995, 179–196.

[16] J.-M. Lee, C. Yoo, I.-B. Lee. "Statistical monitoring of dynamic processes based on dynamic independent component analysis" Chem. Eng. Sci, 59, 2004, 2995–3006.

[17] I. Monroy, R. Benitez, G. Escudero, M. Graells. "DICA enhanced SVM classification approach to fault diagnosis for chemical processes, in: Jacek Jeżowski and Jan Thullie (Ed.)" Comput. Aided Chem. Eng., Elsevier, 2009, 267–272.

[18] C.-C. Hsu, L.-S. Chen, C.-H. Liu. "A process monitoring scheme based on independent component analysis and adjusted outliers" Int. J. Prod. Res, 48, 2010, 1727–1743.

[19] C.-C. Hsu, M.-C. Chen, L.-S. Chen. "A novel process monitoring approach with dynamic independent component analysis" Control Eng. Pr., 18, 2010, 242–253.

[20] J. Lee, B. Kang, S.-H. Kang. "Integrating independent component analysis and local outlier factor for plant-wide process monitoring" J. Process Control, 21, 2011, 1011–1021.

[21] M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander. "LOF: Identifying Density-Based Local Outliers" in: Proc. 2000 ACM SIGMOD Int. Conf. Manag. DATA, ACM, 2000, 93–104.

[22] B.W. Silverman. "Density estimation for statistics and data analysis" CRC press, 1986.

[23] J.J. Downs, E.F. Vogel. "A plant-wide industrial process control problem" Comput. Chem. Eng, 17, 1993, 245–255.

[24] L.H. Chiang, R.D. Braatz, E. Russell. "Fault Detection and Diagnosis in Industrial Systems" Springer, 2001.

Elham Tavasolipour received the B.S. degree in Electrical Engineering from Shahrood University of Technology of Iran in 2011 and the M.S degree in Electrical Engineering from Tarbiat Modares University in 2013. Her areas of interest include fault detection, fault tolerance control, and system identification.

Mohammad Taghi Hamidi Beheshti received the B.S. degree in Electrical Engineering from University of Nebraska, Lincoln, Nebraska of USA in 1984 and the M.S. degree in Electrical Engineering from Wichita State University, Wichita, Kansas of USA in 1987 and the PhD degree in Electrical Engineering from Wichita State University, Wichita, Kansas of USA in 1992 .His areas of interest include optimal control, adaptive control, and network control systems.

Amin Ramezani received the B. S. Degree in electronic engineering from Shahid Beheshti University of Iran, in 2001 and the M.S. degree in electrical engineering from Sharif University of Technology of Iran, in 2003 and the Ph.D. degree in Control engineering at University of Tehran of Iran, in 2011. His areas of interest include Process Control and Automation, advanced Instrumentation, Fault Tolerant Control Systems, Discrete Event Systems and Tele-Operation, Intelligent Transportation Systems.