Adaptive monitoring for transition process using dynamic mutual information similarity analysis

Yuchen He[1], Zhiqiang Ge[1], Zhihuan Song[1]

1. College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China
E-mail: 11232030@zju.edu.cn

Abstract: Due to the non-stationarity in a transition process, it is impossible to implement the monitoring using conventional statistical algorithms. In this paper, a novel identification and monitoring method for transitions using the mutual information similarity analysis (MISA) is proposed to cope with the problem. In the MISA method, the difference between different stable modes as well as transitions is identified. Therefore, the whole process can be divided into several small sub-segments. Considering the dynamic information contained in the process, the classical DPLS model is utilized for online identification and monitoring. Finally, the proposed algorithm is tested using the TE benchmark.

Key Words: Multimode process, Transition process, Mutual information similarity, Transition identification and Transition monitoring

1 INTRODUCTION

Considering the safety and efficiency of process producing, much attention have been paid to the process monitoring methods in last few decades. Traditional multivariate statistical process control (MSPC) algorithms such as principal component analysis (PCA) and partial least squares (PLS) have been widely used for process monitoring purpose[1-4]. Usually, the basic assumption of these methods is that the system status maintains stable and the process behavior can be described using one stable model. However, such assumption is often invalid in practice due to fluctuations in raw materials, aging of equipment and seasoning effects. Therefore, it is hard to model and analyze the process behavior using conventional MSPC algorithms[5]. In recent years, a couple of methods have been proposed to solve this problem [6-8]. In these methods, the process is usually divided into several stable modes and conventional MSPC methods are then applied on each stable segments. Actually, it is impossible that one stable mode is switched to the other one immediately. Commonly, there are always different kinds of transient process between these two stable modes. A transition is usually time-varying and changes from one stable mode to the other one. According to the characters of the transitions, several researches have been done in order to cope with the transition identification and monitoring problems. The Euclidean distance based clustering methods such as k-means are initially used for transition identification [9,10]. However, the result deteriorates when the characters of the transition are similar to the adjacent stable modes. As a result, Beaver et al. proposed the k-PCA models based algorithm where the whole process is considered to consist of a series of overlapping moving window and each window is then projected onto a specific model to obtain the current clustering results. The iteration will repeat until the condition of convergence is met [11]. Note that the transition is considered as a non-Gaussian process, and this characteristic should be taken into consideration. Therefore, Zhu et al.[12] proposed an unsupervised model based modeling and monitoring algorithm to cope with non-Gaussian process problem. However, ICs are usually difficult to choose and the algorithm is still carried out based on iteration, which makes the algorithm difficult to converge rapidly. Besides, the performance of the algorithm highly depends on the selection of initial model parameters. Moreover, dynamics are always found in transient process and this significant characteristic was not discussed in the above papers.

In this paper, a novel transition identification and online monitoring method is proposed to tackle the problem mentioned above. As an effective way to analyze the relationship between data, mutual information (MI) based similarity is utilized to identify which possible transitions the system status is transferred to. After that, the DPLS monitoring method is adopted to confirm the type of transition. If the statistics do not satisfy the corresponding control limits, it is considered as a fault and an alarm occurs. The proposed algorithm does not require any prior knowledge about stable mode to stable mode trajectory. It indicates that one stable mode can be transferred to any stable mode according to production requirements, which seem to give reality to the common view.

The rest of this paper is organized as follows. The MI algorithm is introduced firstly and MI based similarity index is derived in details in Section 2. After that, a MI-DPLS transition identification and monitoring scheme is proposed in Section 3. Then, the proposed algorithm is evaluated by the TE benchmark. In the final section, some conclusions are given.

This work is supported by Fundamental Research Funds for the central Universities

978-1-4673-9714-8/16/$31.00 ©2016 IEEE

2 MI algorithm and MI based similarity index selection

When the system is operating at a specific system status, the statistical characteristic is unique. This uniqueness is often reflected using the Euclidean distance based indices. Actually, Euclidean distance based methods mainly concern about the correlation between samples while ignoring the relationship between variables. However, variable-wise relationship also plays an important role in system status uniqueness. Hence, important information have been lost in transition identification and the classification performance will deteriorate. Therefore, entropy based MI is introduced to help understand both variable-wise and sample-wise relationship between data. The reliability of MI methods in coping with non-Gaussian problem has been proved by previous works[13]. Both variable-wise and sample-wise relationships have been taken into consideration. A brief introduction about MI algorithm is given here. Before introducing MI algorithm, let us review the Shannon entropy of a single variable z , which can be expressed as follows:

$$H(z) = -\int dz\mu(z)log\mu(\mathrm{z}) \tag{1}$$

where "log" is the natural logarithm, and $\mu(z)$ represents the probability density of variable z . The mutual information between two data blocks can be expressed as[14]:

$$I(\mathrm{X,Y}) = \int\int dxdy\mu(\mathrm{x,y})log\frac{\mu(\mathrm{x,y})}{\mu_x(\mathrm{x})\mu_y(\mathrm{y})} \tag{1}$$

where $\mu_x(\mathrm{x}) = \int dy\mu(\mathrm{x,y})$ and $\mu_y(\mathrm{y}) = \int dx\mu(\mathrm{x,y})$ are the marginal densities of x and y , respectively. $\mu(\mathrm{x,y})$ represents the joint probability density. The above index can be further transferred into:

$$I(\mathrm{x,y}) = H(\mathrm{x}) + H(\mathrm{y}) - H(\mathrm{x,y}) \tag{2}$$

where $H(\mathrm{x})$ and $H(\mathrm{y})$ are the corresponding marginal entropies of variable x and y , respectively. $H(\mathrm{x,y})$ represents the corresponding joint entropy:

$$H(\mathrm{x,y}) = -\iint \mu(\mathrm{x,y})log\mu(\mathrm{x,y})dxdy \tag{3}$$

Considering the complexity in calculation of mutual information, the above integral polynomial can be transferred into a simple expression using k-Nearest Neighbor algorithm (kNN) as follows:

$$\begin{aligned} H(\mathrm{x,y}) = &\frac{dis_x + dis_y}{N}\sum_{i=1}^{N}\varepsilon(i) \\ &+\psi(N) - \psi(k) + \log(\mathrm{c}_{disx}\mathrm{c}_{disy}) \end{aligned} \tag{4}$$

where $c_{dis} = \pi^{d/2} / \Gamma(1 + d/2) / 2^d$ represents the volume of the d -dimensional unit cube for the Euclidean norm. dis_x and dis_y are the dimensions of variables x and y , respectively. Define $\varepsilon_x(i)/2$ and $\varepsilon_y(i)/2$ the distances between the same points projected into the x and y subspaces, respectively. $\psi(\cdot)$ represents digamma function. Besides, N represents the number of samples in the training data, $\psi(\mathrm{x}) = \Gamma(\mathrm{x})^{-1} d\Gamma(\mathrm{x}) / dx$ is the digamma function of variable x and $\Gamma(\mathrm{x}) = (\mathrm{x}-1)!$. In eq.(4), $\varepsilon(i) = \max\{\varepsilon_x(i), \varepsilon_y(i)\}$ represents the maximum distance between the i th sample point and its k th nearest neighbor. In Fig 1, black points represent the neighboring points and white points are the other points.

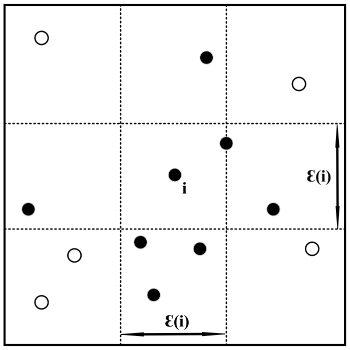


Fig 1. The schematics of MI neighboring sample selection

According to the kNN method, $H(\mathrm{x})$ and $H(\mathrm{y})$ can be calculated as:

$$\begin{aligned} H(\mathrm{x}) = -\frac{1}{N}\sum_{i=1}^{N}\psi(num_x(i)+1) + \psi(N) + \\ \log c_{disx} + \frac{dis_x}{N}\sum_{i=1}^{N}log\varepsilon(i) \end{aligned} \tag{5}$$

$$\begin{aligned} H(\mathrm{y}) = -\frac{1}{N}\sum_{i=1}^{N}\psi(num_y(i)+1) + \psi(N) + \\ \log c_{disy} + \frac{dis_y}{N}\sum_{i=1}^{N}log\varepsilon(i) \end{aligned} \tag{6}$$

where $num_x(i)$ or $num_y(i)$ are the numbers of points within vertical line $x = x_i = \pm\varepsilon(i)$ or horizontal line $y = y_i = \pm\varepsilon(i)$, respectively. Substitute eq.(4)(5)(6) into eq.(2), the MI index between x and y can be computed as:

$$I(\mathrm{X,Y}) = \psi(\mathrm{k}) - \langle\psi(num_x + 1) + \psi(num_y + 1)\rangle + \psi(\mathrm{N}) \tag{7}$$

where num_x and num_y indicate the number of points falling into the regime of k th sample in x direction and y direction, respectively. For further details, please refer to Kraskov's work[14]. Then the MI based similarity which can estimate the correlation between different data can be expressed as follows:

$$S_{MI} = I(\bar{D}_1, \bar{D}_2) \tag{8}$$

where $\bar{D}_1$ and $\bar{D}_2$ are two data blocks whose means and variances are normalized. Eq.(8) shows that the bigger the mutual information between two different data blocks, the larger the MI based similarity value. It is quite reasonable since the relationship is expressed in the form of entropy. When the two data blocks share more information, the entropy which is reflected in the similarity index gets larger. When one data block is dissimilar to the other one, the entropy between them is small and the similarity index decreases.

3 MI similarity index based offline and online identification

In this section, the transition offline identification is firstly carried out using MI based similarity and DPLS algorithm. Transition identification and monitoring mainly consist of two parts: offline identification, online identification & online monitoring. Explicit explanations for each part is shown in the next two sub-sections.

3.1 Transition offline identification

In the transition offline identification, the MI based similarity is utilized to confirm the similarity between the reference data block and the other data blocks. Classical moving window strategy is adopted here to obtain the corresponding data blocks. The reference data block is usually chosen from stable modes. Considering that the durations of stable modes are much bigger than transitions, it is easy to select corresponding reference data block. The schematic diagram of identification using MI based similarity is shown in Fig 2.

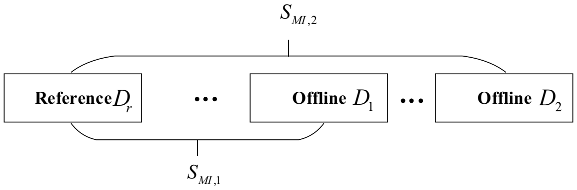


Fig 2. The schematic diagram of identification using MI based similarity

where $S_{MI,2}$ and $S_{MI,2}$ are the corresponding similarity between reference data block and two offline data blocks, respectively. The calculation of similarity will continue until all offline data blocks are tested. After a series of similarity indices are obtained, subtractive clustering algorithm is applied to divide the whole process into several small segments. Windows are classified into different clusters according to their similarity with the reference window. It should be noted that there exists several samples which is involved by two or more different windows according to moving window strategy. The cluster label for these samples can be determined using the following membership probability

$$\mathbf{P}_{ib} = \frac{n_{ib}}{LM^{-1}} \tag{9}$$

where L and M are the window length and moving step, respectively. n_{ib} is the number of times each sample i assigned to cluster b. Sample i belongs to the cluster with the biggest membership probability. After the whole process is separated into several segments, the location for each stable modes and transitions should be determined. Note that the durations of stable modes are much bigger than those of transitions. It is easy to recognize stable modes from transitions. Define δ as the duration of shortest stable mode. Segment whose length is bigger than δ is considered as a stable mode. Segment whose length is shorter than δ is considered as a sub-segment in a transition process.

The main steps of offline identification is shown as follows

1. Normalize the offline data;
2. Choose a reference data block from stable modes;
3. Calculate the similarity indices between reference data block and offline data blocks;
4. Separate the whole process into several segments by applying subtractive clustering;
5. Determine the stable modes and transitions according to their durations.

3.2 Transition online identification and monitoring

After transitions and stable modes are identified using offline data, the online identification is carried out for transitions and stable modes. First, the MI based similarity is still utilized for online identification. It is quite reasonable that the online identification begins with a stable mode since the durations of stable modes are much bigger than those of transitions. The MI based similarity indices between online data block and offline data blocks in stable modes are then calculated. Define β_i as the similarity limit of the i th stable mode. β_i can be calculated using the kernel density estimator[15]. Suppose that γ_i is the similarity index between online data block and the i th stable mode. If $\gamma_i > \beta_i$, online data block may belong to the i th stable mode. Otherwise, online data block may belong to the other stable modes or be a fault. Note that there are might be several stable modes which satisfy the above condition. The next step is to figure out which stable mode the online data block really belongs to. Hence, DPLS model is employed here. If the statistics of online data are under the corresponding control limits of possible stable mode, the online data are considered to belong to this stable mode. Otherwise, the online data block does not belong to this stable mode. After all possible stable mode is checked, the stable mode which satisfies all conditions is chosen. If no candidate stable mode satisfies the condition, the online data is a fault and an alarm occurs. After the beginning stable mode type is determined, subsequent online data are tested and monitored. If online data does not satisfy either similarity index or DPLS control limits, this online data may belong to a transition or just be a fault. Suppose that the process starts with stable mode j, all possible transitions starting with stable mode j are tested using the beginning data of each transition. Similarly, candidate transitions are chosen according to MI based similarity criterion. DPLS models are then adopted to find out whether the online data belongs to a transition or are faulty. The online identification continues until an alarm occurs or all online data are tested. Simultaneously, the online monitoring is also accomplished by using the DPLS modelling and monitoring. The Hotelling's T^2 and SPE statistics are employed for process monitoring:

$$T^2 = t\Lambda^{-1}t \tag{10}$$

$$SPE = \boldsymbol{e}\boldsymbol{e}^T \tag{11}$$

where t represents the corresponding score vector and

$\Lambda = \frac{1}{N}\mathbf{T}^T\mathbf{T}$ is the corresponding covariance of score matrix. $\boldsymbol{e}$ is the prediction errors in accordance to the DPLS model. The confidence limits for T^2 and SPE are defined as:

$$T^2 \sim \frac{A(N^2 - A)}{N(N-A)}F_{A,N-A,\alpha}$$
$$SPE \sim \frac{g}{2f}\chi^2_{\frac{2f^2}{g}} \tag{12}$$

where f and g are the corresponding estimated mean and variance of the modelled data.

The main steps of transition online identification and online monitoring is listed as follows

1. Normalize the online data;
2. Determine the type of the beginning stable mode using MI based similarity index and DPLS model;
3. Monitor the following online data until neither conditions are not satisfied;
4. Identify the online data using the beginning parts of transitions which starts with the stable mode. Choose candidate transitions according to their performance of MI based similarity;
5. Monitor the online data using the DPLS model based on the candidate transition in the offline data;
6. Determine whether the online data belong to a transition.

The schematic diagram of transition identification and online monitoring is shown in Fig 3.

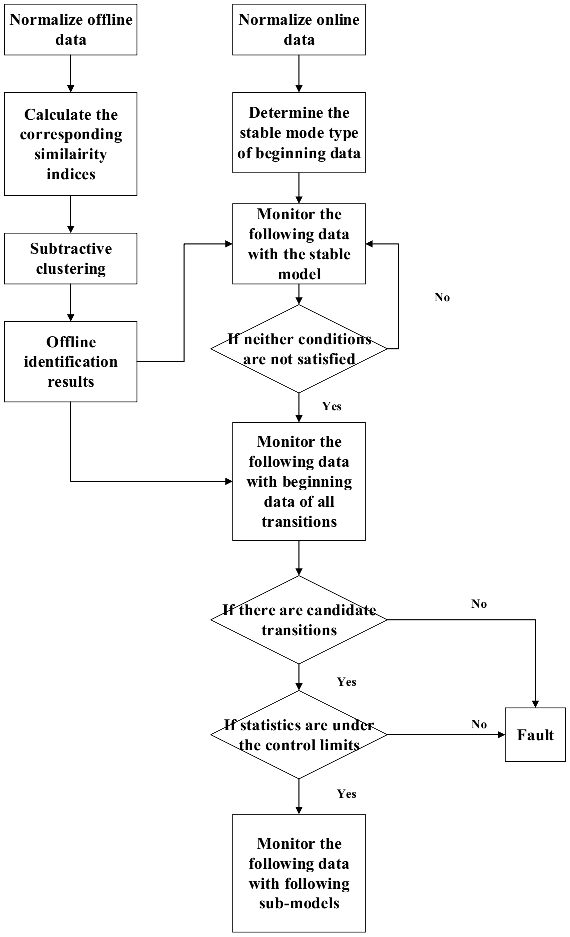

Normalize offline data
Calculate the corresponding similairity indices
Subtractive clustering
Offline identification results
Normalize online data
Determine the stable mode type of beginning data
Monitor the following data with the stable model
If neither conditions are not satisfied
No
Yes
Monitor the following data with beginning data of all transitions
If there are candidate transitions
No
Yes
If statistics are under the control limits
No
Fault
Yes
Monitor the following data with following sub-models

Fig 3. The schematic diagram of transition identification and online monitoring

4 Case study

In this section, the superiority of the proposed algorithm is evaluated on the TE benchmark. Tennessee Eastman (TE) Industrial benchmark was created by the Eastman Chemical Company. It has been widely adopted for algorithm evaluation. The whole process consists of five major unit operations: a product condenser, a recycle compressor, a product stripper, a vapor-liquid separator, and a reactor. For more details about the TE process, please refer to Chiang's work[16]. 41 measured variables and 12 manipulated variables are included in TE benchmark. In order to simulate the multimode operation, two adjacent stable operation modes along with transition between them are given in this case, which are shown in Table 1. 1500 samples are contained in the multimode process. Process status is switched at the 601th point by setting reactor level and temperature set-point to different values[17]. In this paper, 31 process variables and 19 quality variables are used for modeling, which is shown in Table 2 and Table 3.

Table 1. Two operation stable modes in TE benchmark

Mode	Description	Samples
1	Set reactor level at 75%, reactor pressure at 2705kPa and temperature set-point at 120.4°C	1-600
2	Set reactor level at 65%, reactor pressure at 2705kPa and temperature set-point at 125.4°C	601-1500

Table 2. Process variables in TE benchmark

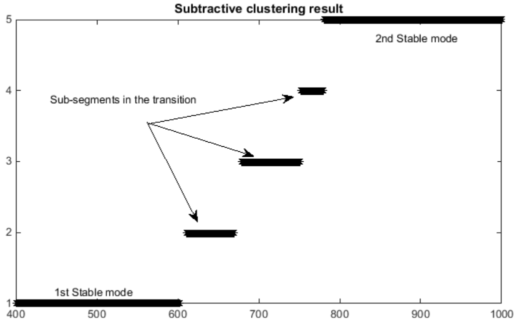
No.	Process variables	No.	Process variables
1	A feed(stream 1)	17	Stripper underflow(stream 11)
2	D feed(stream 2)	18	Stripper temperature
3	E feed(stream 3)	19	Stripper steam flow
4	A & C feed(stream 4)	20	Compressor work
5	Recycle flow(stream 8)	21	Reactor cooling water outlet temperature
6	Reactor feed rate(stream 6)	22	Separator cooling water outlet temperature
7	Reactor pressure	23	D feed flow(stream 2)
8	Reactor level	24	E feed flow(stream 3)
9	Reactor temperature	25	A feed flow(stream 1)

10	Purge rate(stream 9)	26	A & C feed flow(stream 4)
11	Product separator temperature	27	Purge valve(stream 9)
12	Product separator level	28	Separator pot liquid flow(stream 10)
13	Product separator pressure	29	Stripper liquid product flow(stream 11)
14	Product separator underflow(stream 10)	30	Stripper steam valve
15	Stripper level	31	Reactor cooling water flow
16	Stripper pressure		

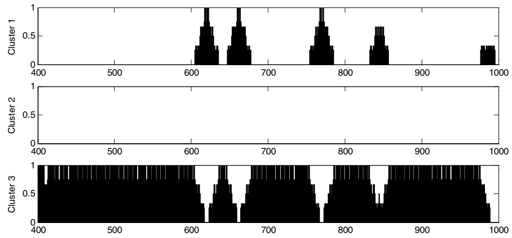
Table 3. Quality variables in the TE benchmark

No.	Quality variables	No.	Quality variables
1	A constituent(stream 6)	11	E constituent(stream 9)
2	B constituent(stream 6)	12	F constituent(stream 9)
3	C constituent(stream 6)	13	G constituent(stream 9)
4	D constituent(stream 6)	14	H constituent(stream 9)
5	E constituent(stream 6)	15	D constituent(stream 11)
6	F constituent(stream 6)	16	E constituent(stream 11)
7	A constituent(stream 9)	17	F constituent(stream 11)
8	B constituent(stream 9)	18	G constituent(stream 11)
9	C constituent(stream 9)	19	H constituent(stream 11)
10	D constituent(stream 9)		

Firstly, MI based similarity is tested for identification performance on TE benchmark. The MI based method and k-ICA-PCA algorithm are performed in the same data and the testing result is shown in Fig 4. It can be clearly seen that the transition can be identified completed using MI base similarity identification method. On the contrary, the identification result using k-ICA-PCA method does not seem very well. Part of the transition is misclassified into stable 1 or stable 2. There are no sample classified into the transition which is very different from the actual situations.



(a)



(b)

Fig 4. Transition identification results using (a) the proposed method (b) the k-ICA-PCA

Another 700 online samples are introduced in the online identification validation. In these 700 online samples, two stable modes and one transition are included. A fault is introduced by changing the reactor pressure from 2705kpa to 2725kpa at the 450th sample. The monitoring results using DPLS are shown in Fig 5 and Fig 6. The fault can be easily detected as long as it occurs. Meanwhile, the missing alarm rate and the false alarm rates are both very low.

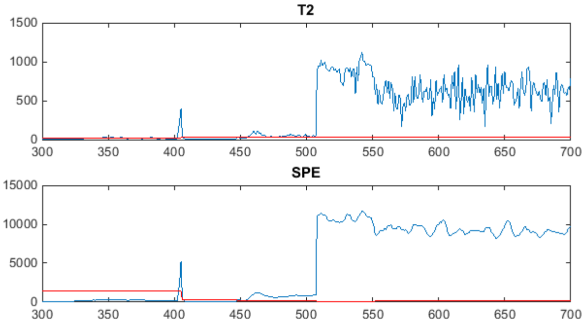


Fig 5. Fault detection for fault 1 using DPLS

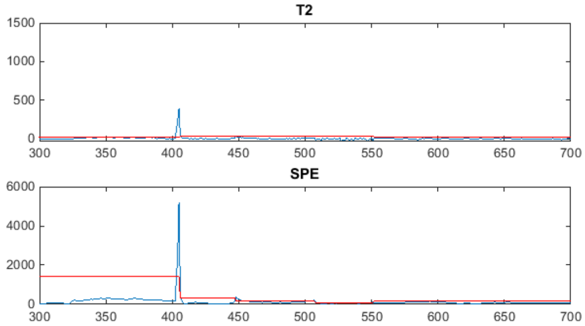


Fig 6. Fault detection for normal data using DPLS

5 Conclusions

In this paper, a novel MI based similarity combined with DPLS monitoring models algorithm is proposed for transition identification, modeling and process monitoring. Both variable-wise and sample-wise correlations are taken fully consideration in a transition. In the offline identification, the MI similarity between the reference data block and other offline data blocks is calculated. After the offline identification is finished, the MI based similarity index is also used in the transition online identification. Combined with DPLS monitoring, possible candidate transitions are then tested. Simultaneously, the online monitoring is also accomplished. Finally, the TE benchmark are employed for algorithm performance validation. The result proves the superiority of the proposed algorithm in transition identification and monitoring.

REFERENCES

[1] Dayal, B.S. and J.F. MacGregor, *Improved PLS algorithms.* Journal of Chemometrics, 1997. **11**(1): p. 73-85.

[2] Qin, S.J. and R. Dunia, *Determining the number of principal components for best reconstruction.* Journal of Process Control, 2000. **10**(2–3): p. 245-250.

[3] Höskuldsson, A., *PLS regression methods.* Journal of Chemometrics, 1988. **2**(3): p. 211-228.

[4] Lindgren, F., P. Geladi, and S. Wold, *The kernel algorithm for PLS.* Journal of Chemometrics, 1993. **7**(1): p. 45-59.

[5] Ge, Z., Z. Song, and F. Gao, *Review of Recent Research on Data-Based Process Monitoring.* Industrial & Engineering Chemistry Research, 2013. **52**(10): p. 3543-3562.

[6] Yu, J. and S.J. Qin, *Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models.* AIChE Journal, 2008. **54**(7): p. 1811-1829.

[7] Tan, S., et al., *Multimode Process Monitoring Based on Mode Identification.* Industrial & Engineering Chemistry Research, 2011. **51**(1): p. 374-388.

[8] Zhang, Y., et al., *Modeling and monitoring for handling nonlinear dynamic processes.* Information Sciences, 2013. **235**(0): p. 97-105.

[9] Hwang, D.-H. and C. Han, *Real-time monitoring for a process with multiple operating modes.* Control Engineering Practice, 1999. **7**(7): p. 891-902.

[10] Ge, Z., et al., *Utilizing transition information in online quality prediction of multiphase batch processes.* Journal of Process Control, 2012. **22**(3): p. 599-611.

[11] Beaver, S., A. Palazoglu, and J.A. Romagnoli, *Cluster Analysis for Autocorrelated and Cyclic Chemical Process Data.* Industrial & Engineering Chemistry Research, 2007. **46**(11): p. 3610-3622.

[12] Zhu, Z., Z. Song, and A. Palazoglu, *Process pattern construction and multi-mode monitoring.* Journal of Process Control, 2012. **22**(1): p. 247-262.

[13] Rashid, M.M. and J. Yu, *A new dissimilarity method integrating multidimensional mutual information and independent component analysis for non-Gaussian dynamic process monitoring.* Chemometrics and Intelligent Laboratory Systems, 2012. **115**(0): p. 44-58.

[14] Kraskov, A., H. Stögbauer, and P. Grassberger, *Estimating mutual information.* Physical Review E, 2004. **69**(6): p. 066138.

[15] Duong, T., et al., *Feature significance for multivariate kernel density estimation.* Computational Statistics & Data Analysis, 2008. **52**(9): p. 4225-4242.

[16] Chiang, L.H., *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes.* 2000.

[17] Rashid, M.M. and J. Yu, *Hidden Markov Model Based Adaptive Independent Component Analysis Approach for Complex Chemical Process Monitoring and Fault Detection.* Industrial & Engineering Chemistry Research, 2012. **51**(15): p. 5506-5514.