Journal of Process Control 47 (2016) 136–149

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

# Key principal components with recursive local outlier factor for multimode chemical process monitoring

Bing Song, Shuai Tan, Hongbo Shi*

*Key Laboratory of Advanced Control and Optimization for Chemical Processes of the Ministry of Education, East China University of Science and Technology, Shanghai 200237, PR China*

ARTICLE INFO

*Article history:*
Received 11 November 2015
Received in revised form 28 June 2016
Accepted 11 September 2016

*Keywords:*
Multimode process monitoring
Principal component analysis
Recursive local outlier factor
Cumulative percent expression
Key principal components

ABSTRACT

Owing to various manufacturing strategies and demands of markets, chemical processes often involve multiple operating modes. How to identify mode from multimode process data collected under both stable and transitional modes is an important issue. This paper proposes a novel mode identification algorithm-recursive local outlier factor (RLOF) based on the sequential information in the time scale and the density information in the spatial scale. In this algorithm, not only the number of modes does not need to be determined in advance, but also details of mode switching can be acquired. In addition, the principal components (PCs) chosen by the variance of overall dataset in principal component analysis (PCA) cannot guarantee that all variables express information as completely as possible. Using the defined cumulative percent expression (CPE), this study chooses key PCs (KPCs) according to each variable. Moreover, fault diagnosis is realized via the contribution of every variable to key PCs. Finally, the monitoring performance is evaluated under the Tennessee Eastman (TE) benchmark and the continuous stirred tank reactor (CSTR) process.



## 1. Introduction

Modern chemical processes become larger and more complex with the development of technology. If a fault in a modern chemical process occurs, it will threaten the safety of operating persons and cause huge economic loss. Owing to the widely used distributed control system (DCS), multivariate statistical process monitoring (MSPM) methods have attracted more and more interests and become popular [1–6]. In traditional MSPM methods, there is an important assumption that the process only has a single operating mode. Complex chemical processes, however, often involve multiple operating modes because of various manufacturing strategies and demands of markets. Identifying mode from the multimode process dataset, which contains both stable and transitional modes, is significant. Principal component analysis (PCA) is regarded as one of the most essential algorithms among all the MSPM methods. Selecting principal components (PCs) in PCA is subjective since it only selects those PCs with larger variance of normal dataset. Thus, the chosen PCs may include incomplete variable information when some variables express information on those PCs with smaller variance of normal dataset. We need to study a novel approach to select key PCs (KPCs) to include entire variable information as completely as possible.

Various multimode process monitoring methods, which can be divided into the single model method and the multiple models method, have been intensively studied. In the single model method, how to make a single model describe different modes or contain different modes information is the key issue. Ma et al. [7] proposed a novel local neighborhood standardization strategy for scaling multimode data to follow single distribution. Ge et al. [8] developed a local model approach which employs the just-in-time learning scheme to solve the multiple modes problem. Song et al. [9] proposed a novel improved dynamic neighborhood preserving embedding algorithm, where only one model was established for multimode processes with dynamic behaviors of within-mode process data. Given that the data obtained from the same mode can be characterized by the local Gaussian distribution, the Gaussian mixture model (GMM) has been employed to handle the multimode problem successfully [10,11]. To construct a global model containing all the local models information, a novel framework called neighborhood based global coordination was developed [12]. Rashid and Yu [13] presented a hidden Markov model (HMM) to characterize shifting modes in multimode processes. Through designing an update termination rule and a switch scheme, a moving window local outlier factor approach was developed to handle operating modes change [14]. To deal with the multimode problem,

* Corresponding author.
*E-mail address:* hbshi@ecust.edu.cn (H. Shi).

http://dx.doi.org/10.1016/j.jprocont.2016.09.006
0959-1524/

Zhu et al. [15] established a mixture model using the Expectation-Maximization algorithm.

In the multiple models method, the key issues are how to identify mode from the multimode training dataset accurately in the offline modeling stage and how to choose the most suitable model or integrate results of all models in the online monitoring stage. For the purpose of identifying mode from multimode data, some basic clustering algorithms were employed [16,17]. Tong et al. [18] adopted the aggregated k-means clustering with the locality preserving indexing to cluster the multimode process dataset. The computational load is decreased and the number of clusters can be estimated. Zhu et al. [19] proposed the k-ICA–PCA models clustering algorithm which includes an ensemble moving window with an ensemble clustering solutions. Considering that multiple modes can be described by a mixture of Gaussian components, the GMM is constructed to identify the multimodality of hybrid systems [20]. Zhang et al. [21] employed the improved k-means clustering algorithm which can remove degeneracy to cluster the data. Recently, based on the serial correlation between data obtained from the same mode, Song et al. [22] proposed a novel clustering method which does not require the specification of the mode number in advance. Meanwhile, it does not consider the transitional mode between two stable modes. On the basis of the mode transformation probability, a novel model selection scheme was proposed by Wang et al. [23]. Using the resemblance of data characteristics, Tan et al. [24] developed a mode identification method to choose the proper model for the current data. Zhao et al. [25] designed a multiple PCA models approach which selects the right model for testing data according to the minimum squared prediction error. In order to integrate results of all models, the Bayesian fusion scheme was applied by Ge et al. [1,26,27]. Given that multiple models method rarely analyze the between-mode relationship, Zhao et al. [28] proposed a between-mode relative analysis algorithm where multiple modes have been predefined and every mode is divided into the increased part, the decreased part, the unchanged part, and the residual part. To monitor multimode multiphase batch processes, a concurrent phase partition and between-mode statistical modeling was developed, where the confidence limit of squared prediction errors in time-slice PCA models and time-segment PCA models are compared in order to partition phases [29].

As the most basic approach among many MSPM methods, PCA has been extended to dynamic PCA, multi-way PCA, kernel PCA, multi-scale PCA, probabilistic PCA, and so on [30–34]. Although PCA has been applied in process monitoring successfully, how to reasonably select PCs is still an open question. Up to now, various algorithms have been proposed for choosing PCs. The cumulative percent variance (CPV) method, which is the most widely used selects those PCs with larger variance of training dataset. Once the CPV value exceeds a threshold, the corresponding PCs are obtained. In the variance of reconstruction error (VRE) scheme, the optimal PCs can be determined when the fault reconstruction error reach minimum according to the best reconstruction of the variables [35]. In the cross-validation method, a part of training data are used to construct the model first and then the prediction of the established model is compared with the other training data. The PC would be added if the prediction residual sum of squares is smaller than the residual sum of squares [36]. The signal-to-noise ratio strategy is used to choose the number of PCs on the basis of the sensitivity of fault detection [37]. In summary, the PCs with larger variance of normal dataset are chosen in these classical methods. On the other hand, since the fault data are different from the normal data, those PCs with smaller variance of normal dataset would capture the largest variations of the fault data. Togkalidou et al. [38] pointed out that those PCs with larger variance may not include enough information to predict. Jiang et al. [39] proposed a new method which selects the PCs online considering that the fault data cannot be guaranteed by the offline-selected PCs. Given that the CPV method would lose important variable information, from the perspective of original variables, Song et al. [22] selected the PCs according to the relevance between the variable and the PC. It only selected the PC with the highest relevance, which is unreasonable when the variable equally expresses information on each PC.

This paper proposes a multimode process monitoring scheme named key principal components with recursive local outlier factor (KPCs-RLOF). Firstly, considering that traditional clustering algorithms ignore the information in the time scale, the clustering result may disaccord with the actual situation and the details of mode switching cannot be obtained. In this article, the training dataset obtained from both stable and transitional modes is identified to every mode by applying both the sequential information in the time scale and the density information in the spatial scale. Secondly, given that the fault data may not lead to a significant variation in those PCs with larger variance of training dataset, selecting PCs according to the variance of normal dataset may lose important information. Motivated by the cumulative percent method, this paper develops a novel PCs selection approach called cumulative percent expression (CPE) on the basis of the expression degree of each variable on PCs. Considering that the fault occurs in the original variable, it needs to be emphasized that the proposed CPE scheme chooses key PCs from the perspective of each variable. Thus, the key PCs space would contain full variable information. Thirdly, to determine the right model for the testing data in the online monitoring stage, the LOF value of the testing data for each mode is calculated and compared. Once a fault has been detected, a contribution plot which applies the selected key PCs is designed to find the fault original variables. Finally, the effectiveness of KPCs-RLOF is proved under the multimode TE benchmark and the multimode CSTR process.

There are three contributions in this article: (1) A novel multimode process monitoring method (KPCs-RLOF) is developed. (2) A novel mode identification approach is proposed on the basis of the sequential information in the time scale and the density information in the spatial scale. (3) A new strategy named cumulative percent expression is developed to select key PCs.

The remaining content is organized as follows: First, the PCA and LOF algorithms are depicted briefly. Second, the detailed mode identification approach and the key PCs selection strategy are presented. Third, the details of how to choose the right model and find the original fault variables are shown. Then, the proposed method is studied under the multimode TE benchmark and the multimode CSTR process. Finally, some conclusions of this work are given.

## 2. Preliminaries

This section reviews the conventional PCA and LOF briefly.

### 2.1. Principal component analysis

The objective of PCA is to acquire several uncorrelated PCs for representing all the original variables. Let $\boldsymbol{X} \in R^{n\times m}$ denote the training dataset with $n$ data and $m$ variables. Then, $\boldsymbol{X}$ can be expressed using the singular value decomposition as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

$$\hat{X} - TP^T \tag{2}$$

where $\boldsymbol{T}$ is the score matrix, $\boldsymbol{P}$ is the loading matrix representing the relationship of the original variable to PCs and $\boldsymbol{E}$ is the residual matrix [40]. Specifically, the number of retained PCs can be determined using the CPV method. The $T^2$ statistic based on retained PCs with larger variance is constructed to monitor the feature space and