Accelerated Online Batch Process Local Monitoring and Soft Sensing Based on Pre-Clustered Fuzzy-C-JITL Multiway Partial Least Squares

Xichang Wang, Pu Wang, Jie Zhang, Xuejin Gao, Peng Chang, and Zheng Li

***Abstract*—Traditional Multiway Partial Least Squares (MPLS) methods focusing on the monitoring and quality prediction of batch processes have the model mismatch problem when encountered with different conditions such as variations in operating conditions, weather, environment, and raw materials, however local Just In Time Learning (JITL) MPLS methods forced on the aforementioned problems have the problem that the online selection of historical modelling data suffers unneglectable computing loading costs, which is heavier if dealing with large amount of historical data or using insufficient hardware capability. Focused on addressing these problems, this paper proposes a novel local modeling MPLS method. First, a pre-clustering procedure is applied during offline stage on historical data, giving the corresponding fuzzy c-means memberships and cluster centers. Second, during online stage the collected local online sample's fuzzy membership is calculated. Then rough historical samples are selected under the threshold derived from offline fuzzy memberships and then detailed selected samples are selected under the guidance of JITL. Finally, a local MPLS model will be built for online monitoring and soft sensing. The proposed Pre-Clustered Fuzzy-C-JITL MPLS algorithm reduces and transfers main part of the computational pressure from online historical sample selection stage to offline stage, improving the efficiency of online monitoring and soft sensing, while keeping the level of prediction accuracy in the same extent. This improvement can be obvious when under the circumstance like high timeliness requirement, low processing capability, huge and complex historical data, or long cycle length. The proposed approach is demonstrated through applications on the penicillin benchmark and an E. coli fermentation process and has its effectiveness examined.**

***Index Terms*—Batch processes, machine learning, multiway partial least squares, soft sensing.**

I. Introduction

Modern batch process industries have been developing through the trend of producing small quantity, responsive, and high value-added products. There are more and more kinds of measurable variables monitored online due to the massive application of data collection system. Some of the variables, such as biomass concentration, product concentration, etc., have high correlation with the final product quality and can even affect its pass rate. However, online measurements of these variables are typically difficult to obtain. Comparing to other process variables that can be easily accessed, there can be delays when using traditional measurement methods to acquire these variables and monitor the whole condition of the process [1], [2], which can consequently deteriorate the process monitoring and operation performance.

Focused on addressing this problem, Multivariate Statistical Process Monitoring (MSPM) methods have been developed under the situation with plenty of acquired data but lack of information. These methods can address the complexity of industrial data and extract useful information by analyzing correlations among different variables and building models. These soft sensing or process monitoring methods, such as Principal Partial Least Squares (PLS), Principal Component Regression (PCR) and Principal Component Analysis (PCA) [3]-[6], have been researched and applied in industries for years.

Many industrial processes, especially general batch processes, are usually not in static state, such as aging phenomenon of product facilities can be found along with time, multiphase problem can be detected due to the growth of microorganism and the switch of working conditions, as well as the replacement of operators, diversity between raw batches, and even changing temperature and humidity, seasonal variation in large time scale can also affect the final product quality. Hence, even though an accurate model is established based on the current condition, it can have model-plant mismatch problem due to quick or slow characteristic changes of the production process after a certain amount of period. Focused on addressing this problem, researchers have proposed recursive algorithms and multiphase algorithms [7]-[9]. The former use iterative calculation to introduce new data samples while gradually replacing old ones in order to renew the model with a good slow change process adaptability and are applicable to continues processes. The latter build models on each stage of a process, paying more attention on identifying the differences among process stages and transition periods. Recently, a local modeling method, Just In Time Learning (JITL) algorithm, have been drawn more and more attention in the field of MSPM [10]-[13]. JITL algorithm, compared to traditional algorithms, is easy to apply and have a relatively high compatibility on unaligned data, which is currently applied on continues processes and some batch processes [14]-[16].

Many related researches paid much attention to improving prediction accuracy or the timeliness of fault detection and

Manuscript received April 9, 2017; revised June 21, 2017.

X. Wang, P. Wang, X. Gao, P. Chang, and Z. Li are with the Faculty of Information Technology, Engineering Research Center of Digital Community (Ministry of Education), Beijing Laboratory of Urban Rail Transit, and Beijing Laboratory of Computational Intelligence System, Beijing University of Technology, Beijing, 100124 China (e-mail: wind-k@hotmail.com).

J. Zhang is with the School of Chemical Engineering and Advance Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU UK.

DOI: 10.7763/IJMO.2017.V7.583

diagnosis. However, JITL MPLS algorithm moves modeling procedure to online stage as well as the selection of historical data in addition, as a result the timeliness can be negatively affected by its local modeling nature. This paper presents the Pre-Clustered Fuzzy-C-JITL MPLS method to address this problem. First, Fuzzy C-means (FCM) algorithm is utilized to analysis the normalized historical data, giving their fuzzy memberships and calculating the cluster centers which are more representative over normal historical data. During online prediction, the relationship between the acquired data and the cluster centers are calculated and a rough selection throughout the historic data is made, then JITL algorithm is utilized to make a detailed selection among the roughly selected historical data. Last, a local model is built by using PLS algorithm with the selected data and then used for local online process monitoring and soft.

The paper is organized as follows. Section II briefly introduces several preliminaries about the related methods. The proposed method and the detailed procedure is given in Section III. Two case studies are detailed in Section IV and Section V draws some concluding remarks.

II. Preliminaries

A. Partial Least Squares

Partial Least Squares (PLS) was proposed by Wold et al. [17] for developing regression models from correlated data sets, and has been applied in many fields such as economics, sociology and chemometrics, etc. [18], [19]

Given a set of data gathered from process operation, historical data matrix X represents the measured variables (independent variables) that can be directly monitored online, matrix Y represents the quality variables (dependent variables) gathered through some other kind of methods. When building the PLS model, the algorithm tries to maximize the interrelation between X and Y, decompose them into the following form.

$$X = TP^{\mathrm{T}} + E \tag{1}$$

$$Y = UQ^{\mathrm{T}} + F \tag{2}$$

where $X \in \mathbb{R}^{n\times m}$, and $Y \in \mathbb{R}^{n\times p}$. $T \in \mathbb{R}^{n\times R}$ and $U \in \mathbb{R}^{n\times R}$ are score matrices of X and Y respectively, $P \in \mathbb{R}^{n\times R}$ and $Q \in \mathbb{R}^{n\times R}$ are loading matrices of X and Y respectively. Superscript T represents the transpose of a matrix, $E \in \mathbb{R}^{n\times m}$ and $F \in \mathbb{R}^{n\times p}$ are the residuals of X and Y. The number of latent variables R can be decided by cross validation [20].

PLS is a biased regression method, its final regression model of dependent variable Y and independent variable X can be expressed as:

$$\hat{Y} = X\beta + E_{\hat{Y}} \tag{3}$$

where β is regression coefficient vector and $E_{\hat{Y}}$ is prediction error. The concept of the PLS method that used in this paper is as described in Table I, the detailed computation procedure can be found in reference [21].

TABLE I: Brief Principle of PLS

Let $X_0=X$, $Y_0=Y$;
Compute $S=X_0'Y_0$;
For $r=1, 2, \ldots, R$:
  If $r=1$: compute SVD of S;
  Else: compute SVD of $S-P(P'P)^{-1}P'S$;
  Weights: w=first left singular vector of SVD performed above;
  Compute scores: $t=X_0w$;
  Compute loadings: $p=X_0't/(t't)$;
  Construct w, t, and p to matrices W, T, and P respectively;
End
Compute regression coefficients $B_{\mathrm{PLS}}=WT-Y_0$;

B. Multiway Partial Least Squares

Batch processes are an important series of industrial production processes. MSPM methods and data gathered from batch process can be used to extract useful information and monitor the state of working condition, utilizing soft sensor algorithms and known real time sensor data (such as agitation power, temperature) to make predictions of quality data that cannot be timely measured online, have been one of the hot research fields of process data mining and production optimization.

Unlike continuous processes, data acquired from a batch process have not only the attribute of variables and samples (time), but also batches. To position a specific data one needs to know the number of variables, samples and batches, i.e., the data are three-way data. This characteristic existing in batch processes leads to the need of a multiway process before applying PLS algorithm, an unfolding method is needed to expand the three-way data into two-way data. In 1987, Wold et al. [22] proposed the Multiway PCA (MPCA) method with a discussion on the process data unfolding method. Nomikos and Macgregor proposed an unfolding method [23] (called NM method for short). The general idea of the method can be briefly described as follows. Given a set of three-way data $X(I\times J\times K)$ which represents normal working conditions of a batch process in I batches, sampled K times with J installed sensors (or plus offline measured quality variables e.g. COD5), the NM method expands the whole data matrix through batch direction, forming a two-way matrix $X(I\times JK)$. Wold *et al.* proposed another unfolding method [24] (WFKH method for short) in MPCA method. First, expand the data through variable direction, forming a new kind of two-way data $X(IK\times J)$. Then normalize the data along each column to a zero-mean and unit-variance data matrix. Finally construct the PLS or PCA model. This unfolding method is shown in Fig. 1.

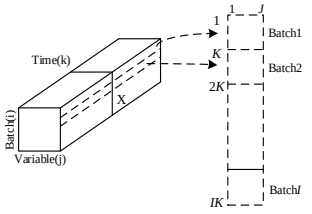

Fig. 1. Demonstration of WKFH expanding method.

The online application of NM unfolding method can have disadvantages. It needs to predict the data in the remaining period of the batch that have not yet been gathered, which can affect the accuracy of the model. In addition, if the gathered historic data have the problem of wrong alignment (i.e.

missing data or premature ending of the process), it will not be suitable using this method. WKFH unfolding method, which has avoided this issue, only have the problem when changing the quantity of variables during modeling or monitoring (yet NM method also suffers), however, this situation is much rarer comparing with the changing of the quantity of samples (time).

In terms of batch process modeling, MPLS can be divided into two parts, offline stage and online stage. In the offline part, the historical data is unfolded, normalized in order to acquire the corresponding score, loading and regression coefficients through MPLS. Statistical control limits for Hoteling T^2 statistic and squared prediction error (SPE) are also prepared for online monitoring, fault detection and diagnosis. During the online part, newly acquired data is normalized and is then used to calculate the T^2 and SPE indices and to see if their control limits are exceeded or not and quality variables are predicted as well.

When a fault occurs during online monitoring, the gathered data can depart from the scope of normal data. T^2 and SPE are usually used in MPLS or MPCA to judge if there is a fault or not [25].

T^2 statistic reflects the extent that each sample deviate from the model, which subjects to F distribution:

$$T_k^2 = t_{new,k}\Lambda^{-1}t_{new,k}^{\mathrm{T}} \sim \frac{R(N^2-1)}{N(N-R)}F_{R,N-R,\alpha} \tag{4}$$

where $\boldsymbol{t}_{new,k}$ (k=1,2,…K) denotes the score of data collected at the k^{th} sample. Matrix Λ^{-1} represents the inverse covariance matrix of training data's score T. N denotes the number of rows in the unfolded training data, and α is the confidence limit.

SPE statistic, also called Q statistic, describes the residuals which have not been explained by the model, and is defined as

$$SPE_k = e_{new,k}e_{new,k}^{\mathrm{T}} \tag{5}$$

$$e_{new,k} = x_{new,k} - t_{new,k}P^{\mathrm{T}} \tag{6}$$

where $x_{new,k}$ (k=1,2,…,K) is the k^{th} new sample of the online process data. SPE statistic subjects to χ^2 distribution:

$$SPE_k \sim g_k\chi^2_{k,h_k,\alpha},\ g_k = \frac{v_k}{2m_{\mathrm{k}}}, h_k = \frac{2m_k^2}{v_k} \tag{7}$$

where g_k is a constant in one model, h_k is degree of freedom of the distribution, v_k and m_k are the mean and variance of the SEP of the k^{th} sample, and α is its confidence limit.

C. Just In Time Learning

Just In Time Learning (JITL) is a local modeling algorithm, having drawn more and more attention in the research of MSPM. According to the ideas of JITL, its processing procedure can be listed in three steps [12]. (1) Find the data from historic database which is the most relevant to the current data through a nearest neighbor criterion. (2) Build the local model with the selected relevant data. (3) Compute the output of the built model based on the current data, discard the current model after finished and wait for next sampling. When the next data is acquired, repeat the aforementioned three steps. A general data driven algorithm, e.g. PLS and PCR, is usually used in the modeling steps, i.e. steps (2) and (3). In other words, the main object of JITL algorithm is to select the most similar data corresponding to the current sample from historic database, while the process and modeling of the training data and its outputs are mainly responded by MSPM methods.

Given the newly acquired data $X_{\text{new},k}$, JITL algorithm calculates the similarity between $X_{\text{new},k}$ and historical data and chooses the similar data. In terms of data driven methods using JITL, the data preprocessing and model parameter estimation are done mainly during offline modeling phase, the modeling procedure which is done traditionally offline is transferred to online stage, so that it can improve the accuracy of the local model in real-time, however, with an increased computational burden which is not welcomed in real-time applications. To solve this problem, Chen et al. came up with a method which uses a forget factor to reduce the data amount for on line JITL data selection [16], Hu et al. [15], Kim et al. [14], and Yuan et al. [26] proposed a moving method and adaptive similarity method for JITL algorithm. In some condition, these methods can lose part of the historical data and have the problem such as the tuning of the forget factor or window width. These problems will have little problems in continuous processes. However, batch processes usually have high repeatability and periodicity and may have the potential useful information in the forgotten data. Focused on addressing this problem, a strategy is proposed in this paper to reduce the computational requirement while keeping the historical data intact.

Generally, v_k and m_k mentioned before are the statistics of all the historical batch data at the k^{th} sample, 'k^{th}' should correspond to current online sample, which may cause confusion if batch duration varies. While, in Just In Time Learning, v_k and m_k are the statistics of the current selected historical data during the online phase.

III. Pre-Clustered Fuzzy-C-JITL Multiway Partial Least Squares Algorithm

A. Preparation and Online Selection of Local Modeling Data

In this subsection, the Pre-Clustered Fuzzy-C-JITL MPLS strategy including rough selection and detailed selection is proposed in detail. The main idea of this strategy is to keep the model accuracy at the same level as normal JITL while reducing the computational requirement, reserving the full information of the historical data. During the application of the traditional JITL method, it is needed to calculate the similarity of all of the historical data each sample time, the calculation will conduct n times if there are n data samples in the historical data, notice that this calculation is done during online process monitoring and needs to be completed well within one sampling interval. The proposed strategy can be divided into three steps, i.e., offline preparation, online rough selection and online detailed selection, transferring and reducing the majority of the computational load from the online stage to the offline stage. First, an offline preparation step using fuzzy c-means is done during offline stage so that the fuzzy memberships of historical data and their

corresponding cluster centers can be presented for online stage. Second, during online rough selection step the fuzzy memberships are calculated between the newly sampled data vector and those cluster centers without, to be distinguish, dividing into groups. A set of related data is then chosen for online detailed selection, where a smaller amount of data is selected for PLS modeling by using JITL algorithm after the rough selection.

Fuzzy C-means clustering is an algorithm that can divide data into groups with cluster centers and membership. The method was introduced by Ruspini [27], developed by Dunn in 1973 [28] and generalized by Bezdek *et al.* [29] with a wide range of applications.

FCM can automatically select cluster centers through an iterative procedure after the number of clusters is set. During the iteration, the algorithm tries to constantly move every cluster center and calculate the membership of each data vector with cluster centers until convergence. The application of FCM offline has two advantages. First, a pre-analysis is done for the whole historical dataset and a similarity system with a wide cover range with a small parameter amount is built, which can reduce the computational burden and system IO requirement during online local modeling. Second, comparing to algorithms that directly uses FCM, this method trend to utilize its membership, without justifying the current process stage, group or alignment issue.

The outputs of FCM are cluster centers and the corresponding membership. The detailed FCM method can be found in Bezdek et al.'s paper [29]. Through the comparison with the old membership, historical data with high relation can be preliminary isolated as rough selection dataset, which will then be detailed selected. To calculate the current membership, an equation from FCM needs to be used:

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right)^{-1}, \ 1 \le k \le N, 1 \le i \le c \qquad (8)$$

where N denotes the sample number, c denotes the cluster number, and parameter d can be expressed as:

$$d_{ik} = \|x_k - v_i\|_A = (x_k - v_i)^{\mathrm{T}} A (x_k - v_i) \qquad (9)$$

where matrix A can be taken as identity matrix I, i.e. Euclidean Norm, m is weight exponent which equals 2 here, v_i is the already acquired i^{th} cluster center, x_k is the new online acquired normalized data, and u_{ik} is the representation of the partition, i.e., membership of x_k.

The online rough selection can be done under the direction of membership by selecting relative data from the historical data, the amount of rough selection dataset can be defined by membership threshold or data proportion. In this paper the proportion threshold is chosen for case studies by selecting a certain percentage of historical data which have the minimum absolute difference with the current online sample. Those selected data are then prepared for the detailed selection section.

The online detailed selection step can be done by using JITL mentioned before. In this paper the often used Euclidean norm, i.e., $d(x, x_i) = \|x, x_i\|^2$ is used for simplicity in discussion, yet other distance measurements can also be used and even the similarity measurement of FCM can be altered correspondingly. The modeling database capacity of the detailed selection can also be limited by membership threshold or data proportion, while as the data amount in the detailed selection section can be very small, the quantity of sample for modeling is defined as the criterion in the case studies.

B. Offline Modeling and Online Monitoring of Batch Process

The proposed batch process monitoring method is as follow.

1) Acquire historical data $X_{raw}(I \times J \times K)$ and $Y_{raw}(I \times J_Y \times K)$, where J_Y is the number of variables in Y.
2) Unfold and normalize the historical data, getting $X(IK \times J)$ and $Y(IK \times J_Y)$, as well as their corresponding means and variances.
3) Apply FCM to the unfolded X, obtaining the corresponding cluster centers and membership matrix $U(IK \times C)$.
4) Online monitoring, acquire a new data vector $X_{new\_raw}(J \times 1)$, normalize the data with the mean and variance that already known offline. However, since there is only one vector sampled at a time, X_{new} is normalized through the equation shown below, where $X_{mean}(J \times 1)$ and $X_{var}(J \times 1)$ are means and variances of historical data.

$$X_{new}(i) = \frac{\left(X_{new\_raw}(i) - X_{mean}(i)\right)}{X_{var}(i)}, i = 1, 2, \dots, J \qquad (10)$$

5) Use (8) to calculate the corresponding membership $U_{new}(1 \times C)$ and use U to construct the rough selection dataset X_{rough}.
6) Calculate the similarity between X_{new} and X_{rough} through JITL, select the detailed selection dataset X_{sub} for modeling, with the corresponding quality data Y_{sub}.
7) Construct the PLS model of X_{sub} and Y_{sub}, as well as the regression coefficients B, 99% (or 95%) confidence control limits T^2_{limit} and SPE_{limit}.
8) Use (3) to make prediction Y_{pre}, which need to be reversely normalized into the quality data. Use equations 4 and 5 to calculate the T^2 and SPE statistics, alarm if inappropriate.
9) Jump to step (4) for the next measurement until the whole production process is done.

IV. Case Studies

A. Fed-Batch Penicillin Fermentation Process

Fed-batch penicillin fermentation process is one of typical batch processes. Its penicillin concentration has direct connection with the performance of the whole process as well as the final product of the process. Birol Gülnur et al. [30] from Department of Chemical and Environmental Engineering, Illinois Institute of Technology, Chicago, developed the Pensim 2.0 Benchmark simulation platform with complex reaction equations embedded. In this case study the platform is utilized to verify the proposed algorithm. A typical fermentation duration is 400 hours, in this study,

the sampling interval was 0.5 hour, the selected measured variables X and quality variables Y are shown in Table II. Totally 61 batches of normal operation data were collected, which were then added up to the three-way data (61 batches × 11 variables ×800 samples).

To verify the effectiveness of the proposed algorithm, the comparisons with traditional MPLS and JITL MPLS was done. Leave one out method was taken into account to evaluate the performances of the algorithms when modeling and predicting normal batches process and quality. In this case, totally 61 sets of experiments with cross validation were conducted. The number of latent variables in MPLS was 5, and that in JITL MPLS was 3. The number of selected data amount in JITL MPLS was 10 samples. The number of latent variables in the proposed method was 3, the number of rough selection cluster center was 3 and selected sample percentage was 10%, and the final detailed selection amount was 10 samples. The computer used in the case studies was equipped with an Intel® Core™ i5-3230M 2.60GHz CPU and had a memory of 8GB ROM. The time that the algorithms spent was logged as well as the prediction accuracy, i.e. Root Mean Square Error. Generally, Root Mean Square Error (RMSE) is used to assess the prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2} \qquad (11)$$

where n represents the total amount of the samples, $\hat{Y}_i$ denotes the prediction of the i^{th} sample, while Y_i denotes the actual value of the ith sample. RMSE was calculated after the new process was fully collected and monitored.

TABLE II: PROCESS VARIABLES AND QUALITY VARIABLES

Variable no.	Variable name
x_1	Aeration rate (L/h)
x_2	Agitator power (W)
x_3	Substrate feed flow rate (L/h)
x_4	Substrate feed temperature (W)
x_5	Substrate concentration (g/L)
x_6	DO (% saturation)
x_7	Culture volume (L)
x_8	CO2 concentration (mmole/L)
x_9	pH
x_{10}	Temperature (K)
y_1	Penicillin concentration (g/L)

TABLE III: STATISTICAL PERFORMANCES OF THE THREE METHODS

Statistics	MPLS	JITL MPLS	The proposed method
Offline time(s)	0.252	0.187	2.008
Online time(s)	1.375	45.565	8.312
Average of RMSE	0.1418	0.0295	0.0312
Variance of RMSE	0.0176	0.0026	0.0021

As can be seen from Table III, the offline modeling time of JITL MPLS was shorter than that of MPLS. The reason of this is that the modeling stage of JITL MPLS is not done offline, that is, the offline stage of JITL MPLS is finished after unfolding the three-way data into two-way data, while MPLS needs not only to unfold the data, but also to model these data. The proposed method needs to do the preparation step for the two-way data hence it also had a longer time than JITL MPLS. Comparing to JITL MPLS, the proposed method had a significant time reduction in online monitoring, this is because the reduction in calculation of the similarity parameter amount due to rough selection with less cluster centers, as well as that the detailed selection sample amount selected from rough selection is far less than traditional JITL MPLS method. The RMSE indexes of the latter two methods have an acceptable range in their mean and variable statistics, while the MPLS method which took global data into account worked worse than the other two methods in local prediction accuracy. Fig. 2 illustrates that the three methods had a same trend in batches while the traditional MPLS is with big prediction deviation, however, as for the batches with small prediction deviation, the local modeling methods, i.e. methods with JITL had a batter stability. In addition, the proposed method has a relatively same level and trend in prediction accuracy with that of JITL MPLS in each batch.

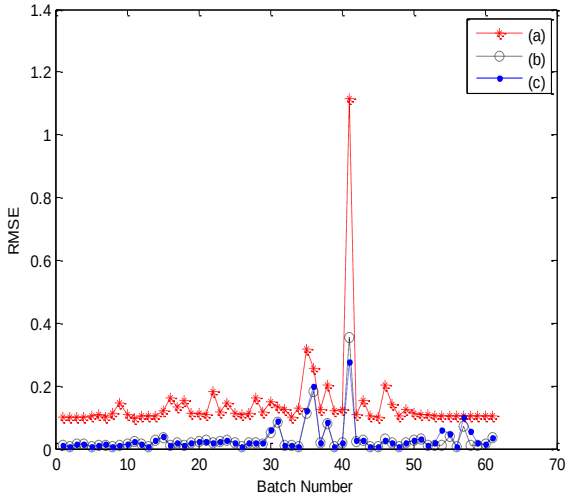

Fig. 2. Prediction performances of 61 tested batches: (a) MPLS, (b) JITL MPLS, (c) The proposed method.

Another main function of PLS in the application of industrial process is fault detection and diagnosis. In order to compare the fault detection ability, the aforementioned different algorithms were used to monitor two abnormal batches where faults were introduced from the 400th sample (199.5 hour) to the batch end. The first abnormal batch was with a ramp fault happened at aeration rate with a ratio of -1% per hour. In the second abnormal batch, a step fault on the aeration rate with a fault magnitude of -10% was introduced.

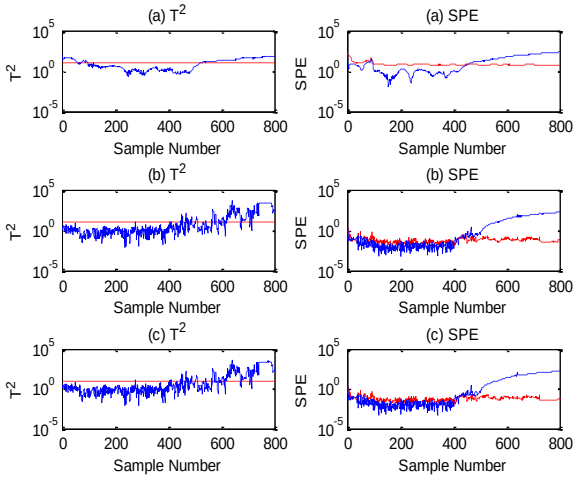

Fig. 3. Online monitoring charts of Pensim fault batch 1 of different algorithms: (a) MPLS, (b) JITL MPLS and (c) The proposed algorithm.

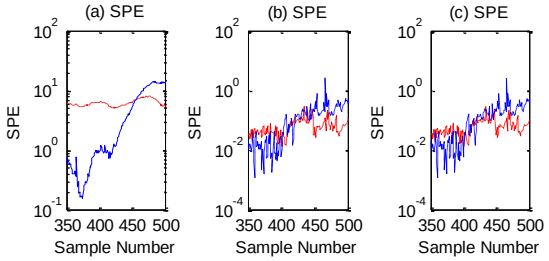

Fig. 4. Zoomed monitoring charts of Pensim fault batch 1 of different algorithms: (a) MPLS, (b) JITL MPLS and (c) The proposed algorithm.

Fig. 3 is the statistic monitoring charts for the first abnormal batch. Fig. 3 illustrates that MPLS statistic chart denoted a fault alarm at the beginning, while the other two did not. On the other hand, from the analysis for the SPE statistic which was better than T^2 statistic, the latter two local methods could detect the fault in an earlier stage (both at the 404th sample) which can be more clearly seen in zoomed Fig. 4 corresponding to Fig. 3, while the former one detected the fault at the 455th sample. From the Fig. 3 can be seen that the proposed method shows almost the same fault detection time with general JITL MPLS method. However, as a ramp fault was introduced, the two method could not alarm at the very beginning. The zoomed Fig. 3 is as shown in Fig. 4, with a sample ranged from 350 to 500.

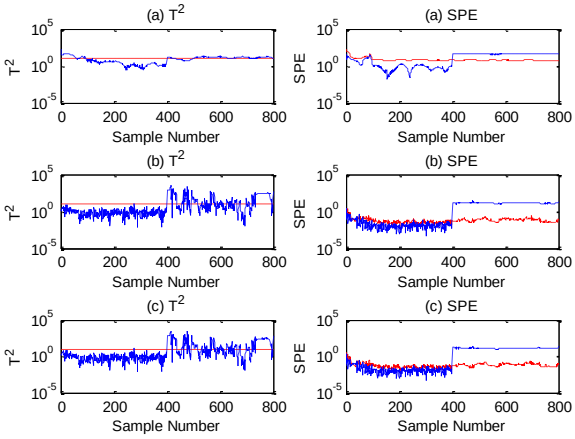

Fig. 5. Online monitoring charts of Pensim fault batch 2 of different algorithms: (a) MPLS, (b) JITL MPLS and (c) the proposed algorithm.

Fig. 5 is the statistic charts for the second abnormal batch. The fault was successfully alarmed by all of the three methods as shown in Fig. 5, while it can be noticed that the T^2 and SPE statistics of the two algorithms using JITL is better than that of MPLS, and still, the proposed method shows almost the same performance with regarding to the general JITL MPLS method.

B. E. Coli Fermentation Process

The Escherichia coli (E. coli) fermentation process is a typical batch process which utilizes the transgenic E. coli to produce some pharmaceutical proteins. An experiment of E. coli fermentation producing interleukin (Fig. 6) was conducted in a pharmaceutical factory in Beijing, China. The strain grows in a fermenter will experience some stages such as adaptation phase, exponential growth phase, and stationary phase. Usually one batch of this process lasts for 5.5 hours, and operators need to decide to take different actions, e.g., feeding, heating up, inducing, based on a variable called OD600, i.e., the broth absorbance at the wavelength of 600 nm. OD600 Not only can direct the operators, but also has strong correlation with the final product concentration. However, the measurement of OD600 in this factory is offline and with some delay, so there is a scope for the application of MSPM method.

Fig. 6. Illustration of the E. coli fermentation system.

In this experiment the chosen measured variables and quality variables are as shown in Table IV. The fermentation time was 5.5 hours and the sampling interval was half an hour. Totally 11 batches of normal operational data were collected forming a three-way matrix (11 batches × 8 variables × 12 samples).

To verify the effectiveness of the proposed method, the experimented methods were the same as the cases in the penicillin platform as well as the PC environment. The number of latent variables in MPLS was 5, and that in JITL MPLS was 5. The number of selected data amount in JITL MPLS was 29 samples. The number of latent variables in the proposed method was 3, the number of rough selection cluster centers was 5, the selected sample percentage was 30%, and the final detailed selection amount was 29 samples. Table V shows the performances of each method, and Fig. 7 shows the RMSE index of each leave one out batch.

TABLE IV: PROCESS VARIABLES AND QUALITY VARIABLES

Variable no.	Variable name
x_1	Time (s)
x_2	Temperature (K)
x_3	Agitator power (W)
x_4	Aeration rate (L/h)
x_5	Pressure (Pa)
x_6	DO (% saturation)
x_7	Inner pH
x_8	Measured pH
y_1	OD600

TABLE V: STATISTICAL PERFORMANCES OF THE THREE METHODS

Statistics	MPLS	JITL MPLS	The proposed method
Offline time(s)	0.006	0.003	0.023
Online time(s)	0.010	0.027	0.026
Average of RMSE	0.4831	0.4659	0.3931
Variance of RMSE	0.1590	0.3115	0.0601

As can be inferred from Table V, whether the offline computing time or the online computing time the three methods spent were all acceptable because of the amount of collected data were small. Although slight, the time consumption features among the three method were as the same as those in penicillin platform and the proposed method

had a batter stability.

In respect of E. coli process, a step fault happened in rotating speed with a relative magnitude of +15% was introduced into an abnormal batch. The corresponding T^2 and SPE statistics are shown in Fig. 8.

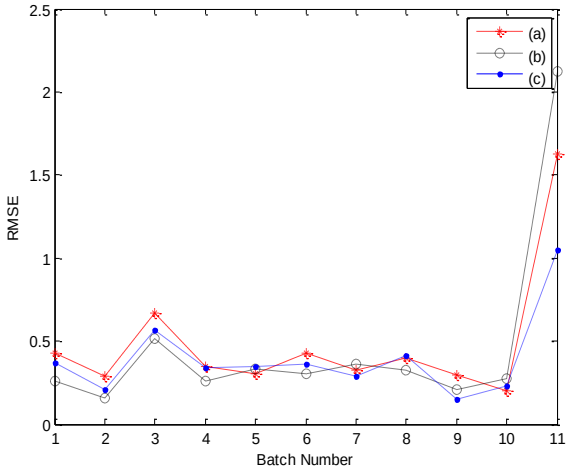

Fig. 7. Prediction performances of 11 tested batches.

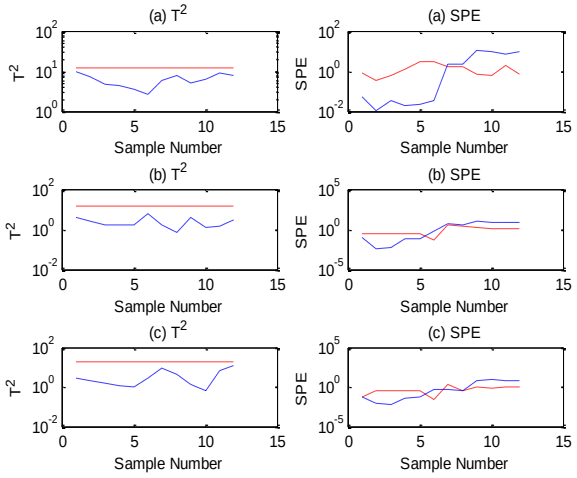

Fig. 8. Online monitoring charts of E. coli fault batch 1 of different algorithms: (a) MPLS, (b) JITL MPLS and (c) the proposed algorithm.

V. Discussion and Conclusions

This paper proposed a method named Pre-Clustered Fuzzy-C-JITL MPLS, focusing on the online computational pressure when applying general JITL MPLS and the periodicity and some other characteristics of batch processes with enormous historical data. The method has three steps in brief, i.e. offline preparation, online rough selection and online detailed selection, transferring the main computational pressure from online stage of process to offline stage, while keeping the same level of algorithm's online monitoring and soft sensing capability. The effectiveness of the proposed method is verified by both simulation and actual experiment.

The parameters such as quantity of rough selection sample, detailed selection sample and cluster center can all affect the performance of the method directly or indirectly, among them those that mainly affect the computational pressure are quantity of rough selection sample and cluster center, while those that mainly affect the model accuracy are quantity of detailed selection sample, etc. So there is a trade-off between computing load and model accuracy, yet this strategy can easily ensure the model accuracy in the same level. But it is noteworthy that the reduction of online computing time of Pensim case study which have a larger amount of historical data is 81.76% (which can be derived from Table III), and the reduction of online computing time of this case study which have a smaller amount of historical data is 3.70% (can be derived from Table V), thus comparing these two improvements, it is reasonable to have a good expectation on the potential that the proposed method can have a good efficiency when dealing with industrial processes with big data. In addition, by looking through the two case studies one can find an interesting problem. As can be seen form the test batch No. 41 in Fig. 2 and test batch No. 11 in Fig. 7, the proposed method performed the best among algorithms in both cases, which may be an accident, or demonstrate that the proposed method also has a potential in dealing with some outlier batches, an intensive analysis can be done in further research.

Acknowledgment

The project was supported in part by the National Natural Science Foundation of China (Grant nos. 61174109, 61364009 and 61640312) and the Beijing Natural Science Foundation (Grant no. 4172007).

References

[1] P. Nomikos and J. F. MacGregor, "Multi-way partial least squares in monitoring batch processes," *Chemometr Intell Lab*, vol. 30, pp. 97-108. 1995.

[2] Y. Zhang and Z. Hu, "On-line batch process monitoring using hierarchical kernel partial least squares," *Chemical Engineering Research and Design*, vol. 89, pp. 2078-2084, 2011.

[3] M. J. F and K. T, "Statistical process control of multivariate processes," *Control Eng Pract*, vol. 3, pp. 403-414, 1995.

[4] E. Vigneau, D. Bertrand, and E. M. Qannari, "Application of latent root regression for calibration in near-infrared spectroscopy," *Comparison with Principal Component Regression and Partial Least Squares*, pp. 231-238, 1996.

[5] J. F. MacGregor, C. Jaeckle, C. Kiparissides, and M. Koutoudi, "Process monitoring and diagnosis by multiblock PLS methods," *AICHE J*, vol. 40, pp. 826-838, 1994.

[6] T. Komulainen, M. Sourander, and S. Jämsä-Jounela, "An online application of dynamic PLS to a dearomatization process," *Comput Chem Eng*, vol. 28, 2004, pp. 2611-2619.

[7] Y. Yao and F. Gao, "A survey on multistage/multiphase statistical modeling methods for batch processes," *Annu Rev Control*, vol. 33, pp. 172-183, 2009.

[8] U. C. and C. A., "Statistical monitoring of multistage, multiphase batch processes," *IEEE Control Systems*, vol. 22, pp. 40-52, 2002.

[9] Y. Wang, D. Zhou, and F. Gao, "Iterative learning model predictive control for multi-phase batch processes,"*J Process Contr*, vol. 18, pp. 543-557, 2008.

[10] Z. Ge and Z. Song, "A comparative study of just-in-time-learning based methods for online soft sensor modeling," *Chemometr Intell Lab*, vol. 104, pp. 306-317, 2010.

[11] L. Xie, J. Zeng, and C. Gao, "Novel Just-In-Time Learning-Based Soft Sensor Utilizing Non-Gaussian Information," *IEEE T Contr Syst T*, vol. 22, pp. 360-368, 2014.

[12] C. Cheng and M. Chiu, "A new data-based methodology for nonlinear process modeling," *CHEM ENG SCI*, vol. 59, pp. 2801-2810, 2004.

[13] C. Cheng and M. Chiu, "Nonlinear process monitoring using JITL-PCA," *Chemometr Intell Lab*, vol. 76, pp. 1-13, 2005.

[14] S. Kim, R. Okajima, M. Kano, and S. Hasebe, "Development of soft-sensor using locally weighted PLS with adaptive similarity measure," *Chemometr Intell Lab*, vol. 124, pp. 43-49, 2013.

[15] Y. Hu, H. Ma, and H. Shi, "Enhanced batch process monitoring using just-in-time-learning based kernel partial least squares," *Chemometr Intell Lab*, vol. 123, 2013, pp. 15-27, doi:10.1016/j.chemolab.2013.02.004.

[16] M. Chen, S. Khare and B. Huang, "A unified recursive just-in-time approach with industrial near infrared spectroscopy application," *CHEMOMETR INTELL LAB*, vol. 135, pp. 133-140, 2014.
[17] S. Wold, A. Ruhe, H. Wold, and I. W. Dunn, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, pp. 735-743, 1984.
[18] L. Kaufmann and J. Gaeckler, "A structured review of partial least squares in supply chain management research," *Journal of Purchasing and Supply Management*, vol. 21, pp. 259-272, 2015.
[19] J. Hulland and R. I. S. O. Business, "Use of partial least squares (PLS) in strategic management research: A review of four recent studies," *Strategic management journal*, vol. 20, pp. 195-204, 1999.
[20] G. Li *et al.*, "Total PLS based contribution plots for fault diagnosis," *Acta Automatica Sinica*, vol. 35, pp. 759-765, 2009.
[21] S. De Jong. "SIMPLS: An alternative approach to partial least squares regression," pp. 251-263, 1993.
[22] S. Wold, P. Geladi, K. Esbensen, and J. Öhman, "Multi-way principal components-and PLS-analysis," *J Chemometr*, vol. 1, pp. 41-56, 1987.
[23] P. Nomikos and J. F. MacGregor, "Monitoring batch processes using multiway principal component analysis," *Aiche J*, vol. 40, pp. 1361-1375, 1994.
[24] S. Wold, N. Kettaneh, H. K. Fridén, and A. Holmberg, "Modelling and diagnostics of batch processes and analogous kinetic experiments," *Chemometr Intell Lab*, vol. 44, pp. 331-340, 1998.
[25] V. Venkatasubramanian, *et al.*, "A review of process fault detection and diagnosis: Part III: Process history based methods," *Comput Chem Eng*, vol. 27, pp. 327-346, 2003.
[26] X. Yuan, Z. Ge, and Z. Song, "Locally weighted kernel principal component regression model for soft sensing of nonlinear time-variant processes," *Ind Eng Chem Res*, vol. 53, pp. 13736-13749, 2014.
[27] E. H. Ruspini, "Numerical methods for fuzzy clustering," *Inform Sciences*, vol. 2, pp. 319-350, 1970.
[28] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," vol. 1973.
[29] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput Geosci-Uk*, vol. 10, pp. 191-203, 1984.
[30] G. Birol, C. Ündey, and A. Cinar, "A modular simulation package for fed-batch fermentation: Penicillin production," *Comput Chem Eng*, vol. 26, pp. 1553-1565, 2002.

Xichang Wang was born at Shandong Province of China. He completed bachelor degree in the field of measurement and control technology and instrument, Northeastern University at Qinhuangdao, China, in 2011.

He is now pursuing his doctoral degree in Beijing University of Technology, Beijing, China.

Mr. Wang's research interests include machine learning, soft sensing and quality prediction.

Pu Wang was born at Anhui Province of china. He received Ph.D. degree in the field of control and optimization of industry process from China University of Mining and Technology in 1988.

He is currently a professor and Ph.D. supervisor in Beijing University of Technology, Beijing, China.

Prof. Wang's main research fields include control and optimization of industry process, complex system control and computer control system.

Jie Zhang received his BSc degree in control engineering from Hebei University of Technology, Tianjin, China, in 1986 and his Ph.D degree in control engineering from City University, London, in 1991.

He is a senior lecturer in the School of Chemical Engineering & Advanced Materials, University of Newcastle, England.

Dr. Zhang's research interests are in the general areas of process system engineering including process modelling, batch process control, process monitoring, and computational intelligence. He has published over 250 papers in international journals, books, and conferences. He is a senior member of IEEE, a member of the IEEE Control Systems Society, and IEEE Computational Intelligence Society. He is on the editorial boards of a number of journals including neurocomputing published by Elsevier.

Xuejin Gao received his Ph.D. degree from Beijing University of Technology, China, in 2006.

He is now a professor in Beijing University of Technology.

Prof. Gao's main research interests include multivariate statistical process monitoring, fault diagnosis.

Peng Chang received his M.Sc. degree from Beijing University of Technology, China, in 2006.

He is now a lecturer in Beijing University of Technology.

Dr. Chang's main research interests include multivariate statistical process monitoring, fault diagnosis and optimization control.

Zheng Li was born in the Inner Mongolia, China. She obtained master's degree in the field of control engineering at Beijing University of technology in 2015.

She is now a doctoral candidate at Beijing University of Technology, China.

Her research area includes soft sensing and fault prediction.