A GLRGMM-Based Online Monitoring Strategy for Batch Processes Postprint

Authors: Xiaoqiang Zhao, Zhou Wenwei

Date: 2018-11-29T00:00:00+00:00

Abstract

To address the nonlinearity and dynamic characteristics of batch processes, a Global-Local Regularized Gaussian Mixture Model (GLRGMM) algorithm is proposed. First, a neighborhood preserving embedding algorithm is introduced to extract local manifold structure, which seeks a low-dimensional projection to preserve the global structure of the nonlinear process while maximizing the retention of local manifold features; then, regularization terms are introduced into the Gaussian mixture model to monitor and update the Gaussian models online, thereby capturing the manifold structure of nonlinear data and resolving the issue of data dynamics; finally, global-local monitoring metrics are integrated to achieve online monitoring. The effectiveness is validated through a penicillin fermentation process, and the results demonstrate that the proposed algorithm yields superior online monitoring performance compared to DPCA and GLNPE.

Full Text

An Online Monitoring Strategy for Batch Processes Based on GLRGMM

Zhao Xiaoqianga,b,c, Zhou Wenweia,b

aCollege of Electrical & Information Engineering, bKey Laboratory of Gansu Advanced Control for Industrial Processes, cNational Experimental Teaching Center of Electrical & Control Engineering, Lanzhou University of Technology, Lanzhou 730050, China

Abstract

To address the nonlinear and dynamic characteristics of batch processes, this paper proposes a Global-Local Regularization Gaussian Mixture Model (GLRGMM) algorithm. First, the Neighborhood Preserving Embedding algorithm is introduced to extract local manifold structure, seeking a low-dimensional projection that preserves the global structure of nonlinear processes while simultaneously retaining local manifold features to the maximum extent. Then, by introducing a regularization term into the Gaussian Mixture Model for online monitoring and updating of Gaussian models, the algorithm obtains the manifold structure of nonlinear data and solves the data dynamicity problem. Finally, integrated global-local monitoring indices are used to achieve online monitoring. Verification through the penicillin fermentation process demonstrates that the proposed algorithm achieves better online monitoring performance than DPCA and GLNPE.

Keywords: batch process; online monitoring; dynamic characteristic; NPE; GMM

0 Introduction

Batch processes have become an important industrial production method due to their strong operability and high flexibility, finding widespread application in the production and preparation of small-batch, high value-added products such as fine chemicals, biopharmaceuticals, and food and beverages [1,2]. Strong nonlinearity and dynamicity are inherent characteristics of batch processes, posing challenges in actual production and online monitoring. Seeking an effective solution is of great significance for reducing economic losses and avoiding casualties [3].

With the development of industrial process control systems, large amounts of process data have been acquired and effectively utilized. Multivariate statistical process monitoring (MSPM) methods, represented primarily by PCA, ICA, and PLS, have been rapidly developed and widely applied in industrial production. To address the nonlinearity and dynamicity in batch processes, methods such as KPCA [4], DPCA, and DPLS [5] have emerged and been successfully applied. Lee et al. [6] proposed multi-way kernel PCA (MKPCA); Peng et al. [7] proposed total kernel partial least squares (T-KPLS) based on T-PLS. These algorithms can effectively address nonlinear characteristics between process variables to some extent, but still suffer from high computational complexity and blind selection of kernel parameters, while ignoring temporal correlation and dynamic characteristics, making it difficult to guarantee process monitoring accuracy.

Fan et al. [8] proposed the Kernel Dynamic Independent Component Analysis (KDICA) algorithm based on kernel functions; Lu et al. [9] proposed a two-dimensional dynamic principal component analysis method. These methods eliminate temporal correlation by extending current data with historical data, but cannot effectively eliminate process data dynamicity. To simultaneously address nonlinearity and dynamicity in process data, Wang et al. [10] proposed dynamic kernel PCA (DKPCA); Zhang et al. [11] proposed recursive kernel PCA; Jia et al. [12] proposed batch dynamic kernel PCA. Although these methods can simultaneously monitor nonlinear and dynamic characteristics of process data, they do not consider global and local structure information of process data, still suffering from problems such as excessive number of principal components and high false alarm rates.

In recent years, manifold algorithms have been increasingly applied to process monitoring due to their superior capability in extracting local manifold structure features and handling high-dimensional data. Hu et al. applied Locality Preserving Projection (LPP) [13] and Neighborhood Preserving Embedding (NPE) [14] for effective monitoring of batch processes; Zhang et al. [15] proposed a Global-Local Structure Analysis (GLSA) algorithm that maintains global structure information while obtaining local manifold structure; Yu [16] proposed Local and Global PCA (LGPCA) based on GLSA; Tong et al. [17] proposed Multi-Manifold Projection (MMP) algorithm combining global and local variance information for process monitoring. Although these algorithms extract global and local information, they do not effectively solve data nonlinearity and dynamicity problems.

Since Gaussian Mixture Model (GMM) can better handle complex process data and has been used to estimate complex data distributions, it has been successfully applied to industrial processes with nonlinear and dynamic characteristics [18~20]. To address the nonlinear and dynamic characteristics of batch processes while effectively extracting global-local information from process data, this paper proposes the Global-Local Regularization Gaussian Mixture Model (GLRGMM) algorithm. As NPE can better obtain local manifold features of high-dimensional data, we improve NPE to obtain a new low-dimensional projection that captures high-dimensional global-local manifold structure, achieving dimensionality reduction and feature extraction for nonlinear process data. Then, dynamic manifold regularization is applied to GMM to automatically update and obtain current data. Finally, through Negative Log Likelihood Probability (NLLP) and Mahalanobis Distance (MD) statistics, global and local probability are integrated to generate a comprehensive monitoring index GLP (Global Local Probability) for online monitoring of batch processes.

---

1 Global-Local Neighborhood Preserving Embedding Algorithm

1.1 Global Structure Preservation

For the space sample set $\{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^D$, we seek a projection matrix A that maps the data through $Y = AX$ to a low-dimensional feature space, preserving the global structure information of original data according to the maximum variance direction. Its objective function is shown in equation (3):

$$J_{\text{global}} = \text{tr}(ACA^T)$$

where $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$ is the covariance matrix, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

1.2 Local Structure Preservation

Local structure preservation excavates local structural information by reconstructing each data point through its neighboring points. The objective function for preserving local structure is shown in equation (4):

$$J_{\text{local}} = \text{tr}(AXMX^TA^T)$$

where $M = (I - W)^T(I - W)$ is the weight matrix, and W is the weight coefficient matrix.

1.3 Global-Local Neighborhood Preserving Embedding (GLNPE)

NPE can excavate the low-dimensional manifold structure hidden in high-dimensional space and preserve the local manifold structure of original space. Its purpose is to map high-dimensional data in $\mathbb{R}^D$ space to a relatively low-dimensional feature space $\mathbb{R}^d$. Although NPE can well preserve local manifold structure and reveal nonlinearity hidden in high-dimensional data, it ignores global structure and cannot better preserve data structure information. Therefore, this paper first proposes the Global-Local Neighborhood Preserving Embedding (GLNPE) algorithm.

To more fully extract process data features, we need to find a projection that can simultaneously preserve global and local information in low-dimensional space, i.e., maximize equation (3) while minimizing equation (4). To solve this complex optimization problem, we propose a new objective function:

$$J = \eta J_{\text{global}} + (1 - \eta)J_{\text{local}} = \text{tr}(AXMX^TA^T) + \eta\text{tr}(ACA^T) - \eta\text{tr}(AXMX^TA^T)$$

where η is an adjustable balance parameter, and we set $\eta = 0.85$. The minimization problem in equation (5) can be solved using Laplacian method to obtain the projection matrix A:

$$(1 - \eta)M + \eta C = A^T\Lambda A$$

The low-dimensional feature space data can thus be obtained:

$$Y = AX$$

Through the projection mapping $Y = AX$, high-dimensional data is projected to low-dimensional feature space while preserving global-local structure information to the maximum extent, which is more helpful for extracting nonlinear data features.

2 Global-Local Regularization Gaussian Mixture Model (GLRGMM) Algorithm

Assume the original variable space data is $X \in \mathbb{R}^{n\times m}$, where n is the number of samples and m is the number of variables. The Gaussian mixture model for samples can be written as $p(X) = \sum_{k=1}^{K} \omega_k g(X|\mu_k, \Sigma_k)$, where K, ω_k, μ_k, and Σ_k represent the number of Gaussian components, and the weight, mean, and variance of the k-th Gaussian component, respectively.

Through GLNPE, two data points X_i and X_j in feature space are extracted, and their cosine similarity matrix is $S_{ij} = \cos\langle X_i, X_j\rangle$. Define the regularization term R_k:

$$R_k = \sum_{i=1}^{n}\sum_{j=1}^{n} p(C_k|X_i)p(C_k|X_j)S_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ik}f_{jk}S_{ij}$$

where I is the identity matrix, $f_{ik} = p(C_k|X_i)$ is the posterior probability that X_i belongs to the k-th Gaussian component, and $L = I - S$.

By minimizing R_k to determine the posterior probability, smooth transitions between different Gaussian components are achieved, removing Gaussian components with minimum posterior probability and updating current Gaussian components to realize smooth transition of dynamic processes and real-time data updating. Introduce a new objective function containing regularization term R_k:

$$J_{\text{new}}(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \omega_k g(X_i|\mu_k, \Sigma_k) - \lambda \sum_{k=1}^{K} R_k$$

where λ is the regularization parameter. Use the K-means algorithm to obtain initial parameters $\theta^0 = \{\omega_k^0, \mu_k^0, \Sigma_k^0\}_{k=1}^{K}$, and determine Gaussian component parameters $\theta = \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ through EM iteration.

E-step: Determine posterior probability by minimizing R_k:

$$p^{(s+1)}(C_k|X_i) = \frac{\omega_k^{(s)} g(X_i|\mu_k^{(s)}, \Sigma_k^{(s)})}{\sum_{j=1}^{K} \omega_j^{(s)} g(X_i|\mu_j^{(s)}, \Sigma_j^{(s)})}$$

M-step: Update parameters:

$$\omega_k^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} p^{(s+1)}(C_k|X_i)$$

$$\mu_k^{(s+1)} = \frac{\sum_{i=1}^{n} x_i p^{(s+1)}(C_k|X_i)}{\sum_{i=1}^{n} p^{(s+1)}(C_k|X_i)}$$

$$\Sigma_k^{(s+1)} = \frac{\sum_{i=1}^{n} p^{(s+1)}(C_k|X_i)(x_i - \mu_k^{(s+1)})(x_i - \mu_k^{(s+1)})^T}{\sum_{i=1}^{n} p^{(s+1)}(C_k|X_i)}$$

GLRGMM more fully extracts global-local structure features of nonlinear data through GLNPE, which better conforms to actual batch process data characteristics. The detailed algorithm flow is shown in [Figure 1: see original paper].

---

3 Online Monitoring Strategy for Batch Process

3.1 Batch Process Data Preprocessing

Batch process data obtained from real-time monitoring is three-dimensional data: $X(I \times J \times K)$, where I is the number of batches, J is the number of variables, and K is the sampling time. In actual production, due to different sampling times across batches and other reasons, it is impossible to completely repeat the same batch production, inevitably leading to unequal data length. This paper adopts the shortest length method [21], using the shortest batch data as the standard and hard-truncating the remaining batches to ensure all batch data have equal length.

To model three-dimensional batch process data, the equal-length data needs to be unfolded into two-dimensional data similar to continuous processes. This paper combines the advantages of batch unfolding (which can remove nonlinearity) and variable unfolding (which can eliminate data prediction problems). The unfolding scheme is shown in [Figure 2: see original paper].

First, three-dimensional data $X(I \times J \times K)$ is unfolded in the batch direction into two-dimensional data $X(I \times JK)$ and standardized. This unfolded data can eliminate nonlinearity and dynamic characteristics to some extent. Then the standardized two-dimensional data is reset to three-dimensional data $X(I \times J \times K)$, and unfolded in the variable direction into two-dimensional data $X(KI \times J)$ for standardization. This unfolding method can eliminate data prediction problems in online monitoring, improve real-time monitoring accuracy, and reduce false alarm rates.

3.2 GLP (Global Local Probability) Monitoring Index

To statistically evaluate the global probability index, we first introduce Negative Log Likelihood Probability (NLLP):

$$\mathrm{NLLP}(X_i) = -\log p(X_i|C_k)$$

The global probability index is obtained by comparing the NLLP of test data with that of training data:

$$P_{\text{global}} = \Pr\{X_{\text{test}} | \text{NLLP}(X_{\text{test}}) \le \text{NLLP}(X_{\text{train}})\}$$

Introduce Mahalanobis Distance (MD) that follows a χ^2 distribution:

$$\text{MD}(X_i) = (X_i - \mu_k)^T \Sigma_k^{-1} (X_i - \mu_k)$$

Define the local probability index:

$$P_{\text{local}} = \Pr\{X_{\text{test},k} \in C_k | \text{MD}(X_{\text{test},k}) \le \text{MD}(X_{\text{train},k}), X_{\text{test},k} \in C_k\}$$

The integrated global-local probability generates the GLP monitoring index:

$$\text{GLP} = \sum_{k=1}^{K} g_k P_{\text{global}}(C_k | X_{\text{test}}) P_{\text{local}}(X_{\text{test}} | C_k)$$

3.3 Offline Modeling

a) Collect several batches of data $X(I \times J \times K)$ under normal operating conditions as training samples, perform data preprocessing using the method in Section 3.1, and finally obtain standardized two-dimensional data Y.

b) Preserve global-local information through GLNPE, obtain low-dimensional feature space projection data Y from equations (6) and (7).

c) Regularize data Y using GLRGMM, and determine parameters θ using the EM algorithm.

d) Calculate $P_{\text{global}}(X_{\text{train}})$ and $P_{\text{local}}(X_{\text{train}})$ through equations (20) and (22), respectively.

3.4 Online Monitoring

a) Preprocess test data using the method in Section 3.1.

b) Obtain projection data Y_{test} for new test data through the GLNPE algorithm.

c) Regularize new projection data Y_{test} using GLRGMM, and determine parameters θ using the EM algorithm.

d) Calculate $P_{\text{global}}(X_{\text{test}})$ and $P_{\text{local}}(X_{\text{test}})$ through equations (20) and (22), respectively.

e) Calculate global and local probability indices through equations (21) and (23), respectively.

f) Calculate the GLP monitoring index through equation (24). If GLP > 0.99, a fault is detected; otherwise, repeat steps b)–f) until a fault alarm occurs or the production process ends.

---

4 Case Study

Penicillin, as an important antibiotic, is widely used clinically. Its production process is a typical batch process with strong nonlinearity and dynamic characteristics. This paper generates batch process data based on the Pensim 2.0 standard penicillin simulation platform [22] developed by the Illinois Institute of Technology. The platform can generate data for each variable at each moment during the penicillin fermentation process under different initial conditions and operating conditions for analysis.

The reaction time for each batch is set to 400 h, with a sampling time of 0.5 h. Under normal range settings with different initial conditions, 35 batches of normal operating data are generated without introducing faults. From the 18 generated variables, 10 process variables are selected as monitoring variables (), forming three-dimensional data $X(35 \times 10 \times 800)$ as training samples.

TABLE:1 Process variables - Aeration rate (L/h) - Agitator power (w) - Feed temperature (K) - Dissolved oxygen concentration (mmol/L) - Reactor volume (L) - Exhaust CO concentration (mmol/L) - Fermenter temperature (K) - Generated heat (K) - Cold water flow rate (L/h)

Pensim 2.0 provides three fault types (aeration rate, agitator power, substrate flow rate). This paper introduces fault type 1 (variable 1, aeration rate) and fault type 2 (variable 2, agitator power), adding 15% step signals during time 200–300 h (sampling points 400–600) as fault signals. The generated data serve as fault test data for online monitoring.

GLRGMM modeling on training data yields 3 Gaussian components, as shown in [Figure 3: see original paper]. DPCA, GLNPE, and GLRGMM algorithms are used for monitoring. For DPCA and GLNPE, monitoring charts for fault 1 are shown in [Figure 4: see original paper]–[Figure 7: see original paper]; detailed monitoring results are shown in and . GLRGMM monitoring results for fault 1 and fault 2 are shown in [Figure 8: see original paper] and [Figure 9: see original paper].

[Figure 3: see original paper] Modeling results of training data

[Figure 4: see original paper] T^2 monitoring diagram of DPCA for fault 1

[Figure 5: see original paper] SPE monitoring diagram of DPCA for fault 1

[**Figure 6: see original paper**] T^2 monitoring diagram of GLNPE for fault 1

[**Figure 7: see original paper**] SPE monitoring diagram of GLNPE for fault 1

[**Figure 8: see original paper**] Monitoring diagram of GLRGMM for fault 1

[**Figure 9: see original paper**] Monitoring diagram of GLRGMM for fault 2

TABLE:2 Monitoring statistics of different methods for fault 1 (%)

Method	T^2 (Statistic)	SPE (GLNPE)	GLP (GLRGMM)
DPCA	-	-	-
GLNPE	-	-	-
GLRGMM	-	-	-

TABLE:3 Monitoring statistics of different methods for fault 2 (%)

Method	T^2 (Statistic)	SPE (GLNPE)	GLP (GLRGMM)
DPCA	-	-	-
GLNPE	-	-	-
GLRGMM	-	-	-

As shown in [Figure 4: see original paper]-[Figure 7: see original paper], both DPCA and GLNPE produce false alarms and missed alarms for fault 1. Although DPCA can address dynamicity to some extent, shows that GLNPE has lower false alarm and missed alarm rates and higher detection rate than DPCA. This is because DPCA cannot well preserve global and local structures in nonlinear data processes, only handling dynamic characteristics of linear processes. GLNPE not only preserves global structure information but also effectively extracts low-dimensional manifold structure hidden in high-dimensional space. Although not the most effective solution for nonlinearity, by preserving global-local information of nonlinear data, its fault detection rate is significantly better than traditional methods that directly ignore nonlinearity.

shows monitoring statistics for fault 2, further confirming that GLNPE achieves better monitoring performance than DPCA, validating the effectiveness of preserving global-local information for nonlinear process monitoring.

[Figure 8: see original paper] and [Figure 9: see original paper] show GLRGMM monitoring for fault 1 and fault 2. GLRGMM outperforms DPCA and GLNPE, detecting all variables of fault 1 without missed alarms, with a false alarm rate of only 0.5% and fast detection speed. For fault 2, although there are missed alarms, no false alarms occur, with a detection rate of 91.04% and high fault sensitivity. This is because GLRGMM not only maximally preserves global-local structure of nonlinear data but also considers data dynamicity, demonstrating clear superiority for processes with nonlinear and dynamic characteristics.

---

5 Conclusion

Strong nonlinearity and dynamicity are inherent characteristics of batch processes. The monitoring strategy based on GLRGMM proposed in this paper not only preserves global-local structure of nonlinear data through improved NPE but also performs online updating of Gaussian models through introduced regularization terms, finally integrating global-local probability indices for online monitoring. Simulation results from the penicillin fermentation process demonstrate that the proposed method achieves better monitoring performance than DPCA and GLNPE, reducing false alarm and missed alarm rates while improving the speed and accuracy of fault monitoring.

References

[1] Guo Jinyu, Yuan Tangming, Li Yuan. Fault detection method for uneven-length multimode batch processes [J]. CIESC Journal, 2016, 67(7): 2916-2924.

[2] Wang Pu, Li Chunlei, Gao Xuejin, et al. Research on batch process monitoring method based on multi-way kernel entropy component analysis and angular structure statistic [J]. Chinese Journal of Scientific Instrument, 2017, 38(1): 174-180.

[3] Zhao Xiaoqiang, Wang Tao. Batch process fault diagnosis based on TGNPE algorithm [J]. CIESC Journal, 2016, 67(3): 1055-1062.

[4] Ben I K, Liama M, Weihs C. Variable window adaptive kernel principal component analysis for nonlinear nonstationary process monitoring [J]. Computers & Industrial Engineering, 2011, 61(3): 437-466.

[5] Chen J, Liu K C. On-line batch process monitoring using dynamic PCA and dynamic PLS models [J]. Chemical Engineering Science, 2002, 57(1): 63-75.

[6] Lee J M, Changkyoo Y, In-Beum L. Fault detection of batch processes using multi-way kernel principal component analysis [J]. Computers & Chemical Engineering, 2004, 28(9): 1837-1847.

[7] Peng Kaixiang, Zhang Kai, Li Gang, et al. Contribution rate plot for nonlinear quality related fault diagnosis with application to the hot strip mill process [J]. Control Engineering Practice, 2013, 21(4): 360-369.

[8] Fan Jicong, Wang Youqing. Fault detection and diagnosis of non-linear non-Gaussian dynamic processes using kernel dynamic independent component analysis [J]. Information Sciences, 2014, 259(3): 369-379.

[9] Lu Ningyun, Yao Yuan, Gao Furong, et al. Two-dimensional dynamic PCA for batch process monitoring [J]. AIChE Journal: Chemical Engineering Research and Design, 2005, 51(12): 3300-3304.

[10] Wang Ting, Wang Xiaogang, Zhang Yingwei, et al. Fault detection of nonlinear dynamic processes using dynamic kernel principal component analysis [C]// Proc of the 7th World Congress on Intelligent Control and Automation. 2008: 3009-3014.

[11] Zhang Yingwei, Li Shuai, Teng Yongdong. Dynamic processes monitoring using recursive kernel principal component analysis [J]. Chemical Engineering Science, 2012, 72(16): 78-86.

[12] Jia Mingxing, Chu Fei, Wang Fuli, et al. On-line batch process monitoring using batch dynamic kernel principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 2010, 101(2): 110-122.

[13] Hu Kunlun, Yuan Jingqi. Multivariate statistical process control based on multiway locality preserving projections [J]. Journal of Process Control, 2008, 18(7-8): 797-807.

[14] Hu Kunlun, Yuan Jingqi. Statistical monitoring of fed-batch process using dynamic multiway neighborhood preserving embedding [J]. Chemometrics & Intelligent Laboratory Systems, 2008, 90(2): 195-203.

[15] Zhang Muguang, Ge Zhiqiang, Song Zhihuan, et al. Global-local structure analysis model and its application for fault detection and identification [J]. Industrial & Engineering Chemistry Research, 2011, 50(11): 6837-6848.

[16] Yu Jianbo. Local and global principal component analysis for process monitoring [J]. Journal of Process Control, 2012, 22(7): 1358-1373.

[17] Tong Chudong, Yan Xuefeng. Statistical process monitoring based on a multi-manifold projection algorithm [J]. Chemometrics & Intelligent Laboratory Systems, 2014, 130(15): 20–28.

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

[19] Chen Tao, Zhang Jie. On-line multivariate statistical monitoring of batch processes using Gaussian mixture model [J]. Computers & Chemical Engineering, 2010, 34(4): 500-507.

[20] Yu Jianbo. Semiconductor manufacturing process monitoring using Gaussian mixture model and Bayesian method with local and nonlocal information [J]. IEEE Trans on Semiconduct, Manufact, 2012, 25(3): 408-422.

[21] Kourti T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, startups and grade transitions [J]. Journal of Chemometrics, 2003, 17(1): 93-109.

[22] Birol G, Ündey C, Cinar A. A modular simulation package for fed-batch fermentation: penicillin production [J]. Computers and Chemical Engineering, 2002, 26(11): 1553-1565.

Note: Figure translations are in progress. See original paper for figures.