Process Systems Engineering

Gaussian Process Regression and Bayesian Inference Based Operating Performance Assessment for Multiphase Batch Processes

Yan Liu, Xiaojun Wang, Fuli Wang, and Furong Gao

Ind. Eng. Chem. Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.iecr.8b00234 • Publication Date (Web): 03 May 2018

Gaussian Process Regression and Bayesian Inference Based Operating Performance Assessment for Multiphase Batch Processes

Yan Liu [a, b, c*], Xiaojun Wang [d], Fuli Wang [a, b], Furong Gao [c]

(a. College of Information Science & Engineering, Northeastern University, Shenyang, Liaoning, 110819, China;

b. State Key Laboratory of Synthetical Automation for Process Industries (Northeastern University), Shenyang, Liaoning, 110819, China;

c. Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong;

d. Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian, 116622, China)

Abstract: Batch processes have been playing a significant role in modern industrial processes. However, even if the operating conditions are normal, the process operating performance may still deteriorate away from optimal level, and this may reduce the benefits of production, so it is crucial to develop an effective operating performance assessment method for batch processes. In this study, a novel operating performance assessment method of batch processes is proposed based on both Gaussian process regression (GPR) and Bayesian inference. It is committed to solving the challenges of multiphase, process dynamics and batch-to-batch uncertainty that contain in most of batch processes. To characterize different dynamic relationships within each individual phase, multiple localized GPR-based assessment models are built firstly. Furthermore, the phase attribution of each new sample is determined, and two different identification results are obtained, i.e., a certain interval and a fuzzy interval between two adjacent phases. Then different online assessment strategies are designed correspondingly. When the operating performance is nonoptimal, cause

* Corresponding author at: College of Information Science & Engineering, Northeastern University, 3 Lane 11, Wenhua Road, Heping District, Shenyang, Liaoning, 110819, China.

Tel.: +86 024 83680351; Fax: +86 024 23890912.

Email：xiaoyan_105@163.com

variables are identified by variable contributions. Finally, the effectiveness of the proposed method is demonstrated by the fed-batch penicillin fermentation process.

Keywords: Batch processes, operating performance assessment, Gaussian process regression, Bayesian inference, nonoptimal cause identification

1. Introduction

Batch processes are commonly used in chemical, materials, pharmaceutical, biotechnology and semiconductor industries for the production of low-volume while high-value-added commodities, and they play a more and more important role in modern industrial production.[1, 2] However, as time goes on, process operating performance may deteriorate away from optimal state due to process disturbances, noise, and other uncertainties, and this may cancel the benefits of preliminary designs for process optimization and result in a degraded operating behavior. Therefore, it is very necessary to develop an effective operating performance assessment method for batch processes.

In the past several decades, many process and quality monitoring approaches have been proposed for batch processes.[3-7] Among them, multiway principal component analysis (MPCA) and multiway partial least squares (MPLS) are most widely used.[6-9] They extend the applications of principal component analysis (PCA) and partial least squares (PLS) techniques from continuous processes to batch processes, and allow process variable trajectory information to be projected into low-dimensional latent variable spaces. Therefore, batch process performance and product quality can be easily analyzed and monitored in the reduced space. As the study moving on, extensions to taking into account various factors are available, such as dynamic characteristics,[10, 11] nonlinearity,[12, 13] non-Gaussian distributions,[14, 15] and multiscale [16] of batch processes. However, it has recently been explored that the traditional MPCA/MPLS model cannot efficiently reveal the multiphase data behavior, even though it is very common in many batch processes. The phases defined by Undey and Cinar [17] is that ‘‘Steps occurring in a single processing unit as succession of events caused by operational or phenomenological (chemical reactions, microbial activities,

etc.) regimes''. Furthermore, Lu et al. [18] defined phases from the perspective of process monitoring, where phases are determined according to the changes in the underlying process correlations. In the present work, we apply the phase definition provided by Lu et al.[18] Being different from stages that lay their focuses on describing the physical operation units of a batch process, phases put the emphasis on the expression of process characteristics, such as variable relationships, data distributions, dynamic trajectories, amplitudes of measurements and so on. Compared with the stages, although the phases are more abstract, it helps learn more about the intrinsic process characteristics and enhance process understanding. Therefore, a batch process is desired to be divided into several phases to improve the modeling performance. In recent years, different phase division methods have been proposed, [19-21] and different modeling methods have been developed that take the phase effects into consideration. [11, 22, 23] The pioneer work has provided abundant theoretical bases for our following work.

In actual processes, the main task of process monitoring is to maintain the production process under normal operating conditions, but it can't satisfy the quest by enterprises for profits any longer. For most of plants, the production goal is to profit, and an effective way is to ensure that the process operates on optimal level throughout the batch production. By this point, the operating performance assessment of industrial processes came into being. [24] The purpose of process operating performance assessment is to get a measure on how far the current operating condition is from the optimum (or how optimal the current operating state is), and the potential assumption is that the operating conditions are normal. According to the process characteristics and plant personnel's attitudes of the operating performance, the performance levels can be divided into several grades, such as optimal, suboptimal, general, and poor. Through operating performance assessment, operators and managers can make a deeper understanding and mastering with the process operating performance, and propose reference suggestions on the operating adjustment and performance improvement. Despite a rich body of literatures in process monitoring, [25-27] studies in operating performance assessment of industrial processes are still in its infancy. In our

previous work, some methods in respect to continuous and multimode continuous processes have been developed on this issue. [28-30]

This paper is devoted to the operating performance assessment for multiphase batch processes with the considerations of process dynamics and batch-to-batch uncertainty. It is generally believed that the process operating performance has a close relationship with the comprehensive economic index, such as cost, profit, total revenue, product quality or the weighted integration of several production indices. If the comprehensive economic index approaches to or reaches the history optimal level, the process operating performance is usually considered as optimality. Thus, the comprehensive economic index is applicable to evaluate the process operating performance. However, it is hard to get online and usually obtained at the end of a batch production, which seriously affects the timeliness of online assessment. As an alternative way, the predicted value of the comprehensive economic index can be used in operating performance assessment. Many quality prediction methods of batch processes have been proposed so far. [31, 32] Considering the multiphase characteristics of batch processes, the quality prediction methods depend on the phase division strategies to some extent. For instance, the multivariate statistics based quality predictions are often applied to the batch processes whose phases are divided according to the variable relationships or covariance structures. [33] Since batch processes are usually characterized by non-stationary, batch-to-batch uncertainty and process variables frequently comprise deterministic trends, the process data of the whole batch often present non-Gaussian distribution. When the process variables have different change rates or amplitudes in two adjacent phases and the change rates are not very large in each individual phase, the local process data will show different distributions in a cycle of the batch production. In this sense, the batch processes can be divided into multiple phases based on the distribution characteristics, and each individual phase is actually represented by an approximate multivariate Gaussian distribution of the unfolded process data. This is quite consistent with the methodology of Gaussian mixture model (GMM), and some offline phase division algorithms based on GMM have been established, such as MPCA-GMM, [34] multiway Gaussian mixture model

(MGMM) [35] and GMM-based phase successive division (GMM-PSD). [36] Under the framework of probability, Gaussian process regression (GPR) [37] is regarded as an appropriate quality prediction method matching GMM. Compared with the deterministic modeling methods, GPR is more suitable for characterizing the complex relationships between the process and quality variables caused by process stochastic feature and system uncertainty.

In the present study, with the consideration of process dynamics and batch-to-batch uncertainty, a new operating performance assessment method based on GPR and Bayesian inference is proposed for multiphase batch processes. Considering the advantages of GMM-PSD in dealing with the uneven-length batch processes, it is used in offline phase division firstly. Then multiple local GPR-based assessment models are developed to characterize different dynamic relationships of the identified multiple phases, and the predictive distribution of the comprehensive economic index is estimated by averaging and weighting all the regression model parameter values with their posterior probabilities. This kind of assessment models can not only effectively handle the stochastic feature caused by process dynamics and batch-to-batch uncertainty but also ensure the accuracy of the assessment result. In online assessment, based on the result of offline phase division, the phase type can be identified as a certain interval or a fuzzy interval between two adjacent phases. In a certain interval, the phase type of the new sample can be uniquely identified, and only the corresponding assessment model is invoked for online assessment. While, if the new sample falls into a fuzzy internal between two adjacent phases, the assessment result is the weighted sum of those from two adjacent phases with the Bayesian inference-based posterior probabilities as the adaptive weights. It effectively reduces the error rate caused by misclassification and is very challenging for traditional assessment methods. When the process operating performance degenerates, the possible cause variables can be identified based on variable contributions, which helps managers and operators take appropriate operating adjustment strategy on production improvement.

The contributions of the proposed method are summarized as follows: (i) to ensure

the accuracy of the assessment result, the GPR-based assessment models are developed to deal with process dynamics and batch-to-batch uncertainty; (ii) for the new sample whose phase type is unknown, its posterior probabilities with respect to the assessment models of two adjacent phases are set as the adaptive weights, and then integrate the corresponding local assessment results for online operating performance assessment, which avoids the error evaluation caused by misclassification; (iii) the cause variables responsible for the nonoptimal operating performance can be determined by variable contributions.

The remainder of this article is organized as follows. Section 2 briefly reviews the methodologies of GMM and GPR. Then the proposed operating performance assessment method is introduced in Section 3. Section 4 demonstrates the effectiveness of the proposed assessment approach by the fed-batch penicillin fermentation process. Finally, some conclusions are drawn in Section 5.

2. Preliminaries

2.1 Gaussian mixture model

GMM is usually used to model a collection of random variables arising from a number of latent classes or states.[38] Consider a J dimensional random variable $\boldsymbol{x} \in R^{J\times 1}$, the GMM with M components is written as:

$$G(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \omega_m g(\boldsymbol{x}|\theta_m), \tag{1}$$

where $\omega_m, m = 1,2,...,M$ is the prior probabilities of the m th Gaussian component C_m and satisfies the conditions of $0 \le \omega_m \le 1$ and $\sum_{m=1}^{M} \omega_m = 1$; $\theta_m = \{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ and $\boldsymbol{\Theta} = \{\omega_1, \omega_2, ..., \omega_M, \theta_1, \theta_2, ..., \theta_M\}$ separately represent the local and global Gaussian model parameters; $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and covariance matrix of the m th Gaussian component, respectively; $g(\boldsymbol{x}|\theta_m)$ is the corresponding multivariate Gaussian density function and expressed as follows:

$$g(\boldsymbol{x}|\theta_m)=\frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Sigma}_m\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)\right], m=1,2,...,M. \quad (2)$$

In order to establish the GMM, the unknown model parameters $\boldsymbol{\Theta}$ need to be estimated. Some learning methods, such as maximum likelihood estimation (MLE), expectation maximization (EM), and Figueiredo–Jain (F–J) algorithm, are usually used in parameter estimation of mixture model. [39, 40] In view of the ability of F-J algorithm [40] in automatically optimizing the number of Gaussian model components and estimating their statistical distribution parameters, it is adopted for model parameters estimation in this study. With the initialized model parameters $\boldsymbol{\Theta}^{(0)}=\left\{\omega_1^{(0)},\omega_2^{(0)},...,\omega_M^{(0)},\theta_1^{(0)},\theta_2^{(0)},...,\theta_M^{(0)}\right\}$, the s th two-step iteration involved in the F-J algorithm is as follows:

E-step:

$$P^{(s)}(C_m|\boldsymbol{x}_n)=\frac{\omega_m^{(s)}g(\boldsymbol{x}_n|\boldsymbol{\mu}_m^{(s)},\boldsymbol{\Sigma}_m^{(s)})}{\sum_{m=1}^{M}\omega_m^{(s)}g(\boldsymbol{x}_n|\boldsymbol{\mu}_m^{(s)},\boldsymbol{\Sigma}_m^{(s)})}. \quad (3)$$

M-step:

$$\boldsymbol{\mu}_m^{(s+1)}=\frac{\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)\boldsymbol{x}_n}{\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)}, \quad (4)$$

$$\boldsymbol{\Sigma}_m^{(s+1)}=\frac{\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)(\boldsymbol{x}_n-\boldsymbol{\mu}_m^{(s+1)})(\boldsymbol{x}_n-\boldsymbol{\mu}_m^{(s+1)})^T}{\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)}, \quad (5)$$

$$\omega_m^{(s+1)}=\frac{\max\left\{0,\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)-V/2\right\}}{\sum_{m=1}^{M}\max\left\{0,\sum_{n=1}^{N}P^{(s)}(C_m|\boldsymbol{x}_n)-V/2\right\}}, \quad (6)$$

where $\boldsymbol{x}_n$ is the n th sample of the modeling data $\boldsymbol{X}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,..,\boldsymbol{x}_N\right]^T\in R^{N\times J}$; $V=J^2/2+3J/2$ is the number of scalar parameters specifying each Gaussian component; $\boldsymbol{\mu}_m^{(s+1)}$, $\boldsymbol{\Sigma}_m^{(s+1)}$, and $\omega_m^{(s+1)}$ are the mean vector, covariance matrix, and prior

probability of the m th Gaussian component at the $(s+1)$ th iteration, respectively. The F-J algorithm is implemented in an iterative way until all parameters converge to the optimal solution. [40]

For an arbitrary new sample $\boldsymbol{x}_{new}$, the posterior probability corresponding to the m th Gaussian component can be represented as follows:

$$P(C_m \mid \boldsymbol{x}_{new}) = \frac{\omega_m g(\boldsymbol{x}_{new} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{m=1}^{M} \omega_m g(\boldsymbol{x}_{new} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}. \quad (7)$$

For a class of multiphase batch processes, where process variables have different change rates or amplitudes in two adjacent phases and the change rates are not very large in each individual phase, the phases can be characterized by their distribution functions. Although some variables show particular trends throughout the batch production, these trends can be significantly weakened in each individual phase. Thus, it can be considered that the unfolded process data from each individual phase approximately follow a multivariate Gaussian distribution, and GMM is applicable to cluster the batch process data from different phases.

2.2 Gaussian process regression

In many industrial processes, process variables often show stochastic feature owing to process dynamics and batch-to-batch uncertainty. Thus, the deterministic modeling strategies become ill-suited for characterizing the complex relationships between process variables and comprehensive economic index. To deal with this issue, GPR [37] is proposed. GPR focus its interest in two aspects: (i) making inferences about the relationship between process inputs and outputs and (ii) obtaining the conditional distribution of the outputs given the inputs, rather than the multivariate distribution of process inputs.

Denote the comprehensive economic index data corresponding to $\boldsymbol{X}$ as $\boldsymbol{y} = [y_1, y_2, .., y_N]^T \in R^{N\times 1}$. Assuming that the relationships between process variables and comprehensive economic index are approximately linear, and then the regression model between $\boldsymbol{X}$ and $\boldsymbol{y}$ can be described as below:

$$y_n = f(\boldsymbol{x}_n) + \varepsilon = \boldsymbol{x}_n^T \boldsymbol{\alpha} + \varepsilon, \tag{8}$$

where $\boldsymbol{\alpha} \in R^{J\times 1}$ is the regression parameter vector and follows a Gaussian prior distribution with zero mean and covariance $\boldsymbol{\Sigma}_\alpha$; ε is the Gaussian noise with zero mean and standard deviation σ.[37]

Moreover, given the process inputs and regression model parameter, the conditional probability density function of the output is represented as follows [37]:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\alpha}) &= \prod_{n=1}^{N} g(y_n|\boldsymbol{x}_n,\boldsymbol{\alpha}) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_n - \boldsymbol{x}_n^T\boldsymbol{\alpha})^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi)^{N/2}\left|\sigma^2\boldsymbol{I}\right|^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\right\|^2\right], \\ &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\alpha}, \sigma^2\boldsymbol{I}) \end{aligned} \tag{9}$$

where $\boldsymbol{I} \in R^{N\times N}$ is a unit matrix.

According to the posterior distribution over the regression model parameter, the posterior probability density function can be estimated through Bayesian inference strategy as follows:

$$P(\boldsymbol{\alpha}|\boldsymbol{X},\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\alpha})P(\boldsymbol{\alpha})}{P(\boldsymbol{y}|\boldsymbol{X})} = \frac{P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\alpha})P(\boldsymbol{\alpha})}{\int P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\alpha})P(\boldsymbol{\alpha})d\boldsymbol{\alpha}}, \tag{10}$$

and the corresponding distribution can be expressed as

$$P(\boldsymbol{\alpha}|\boldsymbol{X},\boldsymbol{y}) \sim \mathcal{N}(\bar{\boldsymbol{\alpha}}, \boldsymbol{A}^{-1}), \tag{11}$$

where $\boldsymbol{A} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_\alpha^{-1}$; $\bar{\boldsymbol{\alpha}} = \sigma^{-2}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{A}^{-1}$ are the mean vector and covariance matrix, respectively.

In deterministic modeling strategies, a single parameter is typically chosen by some criteria to predict the output of the new sample $\boldsymbol{x}_{new}$. However, under the consideration of process stochastic feature, all possible regression model parameters are weighted with their posterior probabilities. Thus, the predictive distribution of output $\hat{y}_{new} = f(\boldsymbol{x}_{new})$ is deduced as follows:

$$P(\hat{y}_{new}|\boldsymbol{x}_{new},\boldsymbol{X},\boldsymbol{y}) = \int P(\hat{y}_{new}|\boldsymbol{x}_{new},\boldsymbol{\alpha})P(\boldsymbol{\alpha}|\boldsymbol{X},\boldsymbol{y})d\boldsymbol{\alpha} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{x}_{new}^T\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y},\ \boldsymbol{x}_{new}^T\boldsymbol{A}^{-1}\boldsymbol{x}_{new}). \quad (12)$$

As seen from Eq. (12), the predictive distribution is also Gaussian, and the mean vector and covariance matrix are $\sigma^{-2}\boldsymbol{x}_{new}^T\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{x}_{new}^T\boldsymbol{A}^{-1}\boldsymbol{x}_{new}$, respectively.

Naturally, the predicted output of $\boldsymbol{x}_{new}$ is given by

$$\hat{y}_{new} = \sigma^{-2}\boldsymbol{x}_{new}^T\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}. \quad (13)$$

3. Gaussian process regression and Bayesian inference based operating performance assessment of multiphase batch processes

Due to the dynamic characteristics and random uncertainty, [19, 20, 22, 41] the stochastic assessment modeling method is more suitable for multiphase batch processes. In this study, a novel operating performance assessment method for multiphase batch processes is developed by integrating GPR with probabilistic inference strategy. The GMM-PSD algorithm is first used to divide different phases of batch processes. Furthermore, the GPR-based assessment models are established to characterize different dynamic relationships of the identified multiple phases, meanwhile, reveal the influences of different underlying process information on the process operating performance. During online assessment, the phase type of the new sample is identified firstly. If the new sample falls into the certain interval of a phase, only the assessment model of the corresponding phase is invoked for online assessment. While, if it belongs to a fuzzy internal between two adjacent phases, the assessment result can be calculated by incorporating those from two adjacent phases, where the Bayesian inference-based posterior probabilities are used as the adaptive weights. In this way, the error evaluation owing to misclassification is avoided effectively. When the process operating performance degenerates, the possible cause variables responsible for the nonoptimality can be identify based on variable contributions.

3.1 Establishment of assessment models

For a multiphase batch process, I uneven-length training batches are collected from the historical data. The process data of the i th batch are expressed in the form

of $\boldsymbol{X}_i = \left[\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,K_i} \right]^T$, where K_i is the number of samples of the i th batch, $i = 1, 2, ..., I$. The corresponding comprehensive economic index measured at the end of the batch is denoted as y_i. In order to remove the process noise and dynamics to some extent, all training batches are unfolded into a two-dimensional block matrix $\underline{\boldsymbol{X}} = [\boldsymbol{X}_{(k=1)}^T, \boldsymbol{X}_{(k=2)}^T, \ldots, \boldsymbol{X}_{(k=K_{\min})}^T, \ldots, \boldsymbol{X}_{(k=K_{\max})}^T]^T \in R^{K \times J}$ via variable-wise unfolding as shown in Fig. 1, where $K_{\min} = \min_{1 \le i \le I}(K_i)$ and $K_{\max} = \max_{1 \le i \le I}(K_i)$ are the minimum and maximum numbers of samples of the training batches, $K = \sum_{i=1}^{I} K_i$. Then the normalization approach proposed in Refs. 42 and 43 is applied to $\underline{\boldsymbol{X}}$, i.e.,

$$\underline{\tilde{x}}_{k,i,j} = \frac{\underline{x}_{k,i,j} - \overline{x}_j}{s_j}, \tag{14}$$

where $\underline{x}_{k,i,j}$ is the j th variable of the k th sample in the i th batch, $\overline{x}_j$ and s_j are mean and standard deviation of the j th variable of $\underline{\boldsymbol{X}}$.

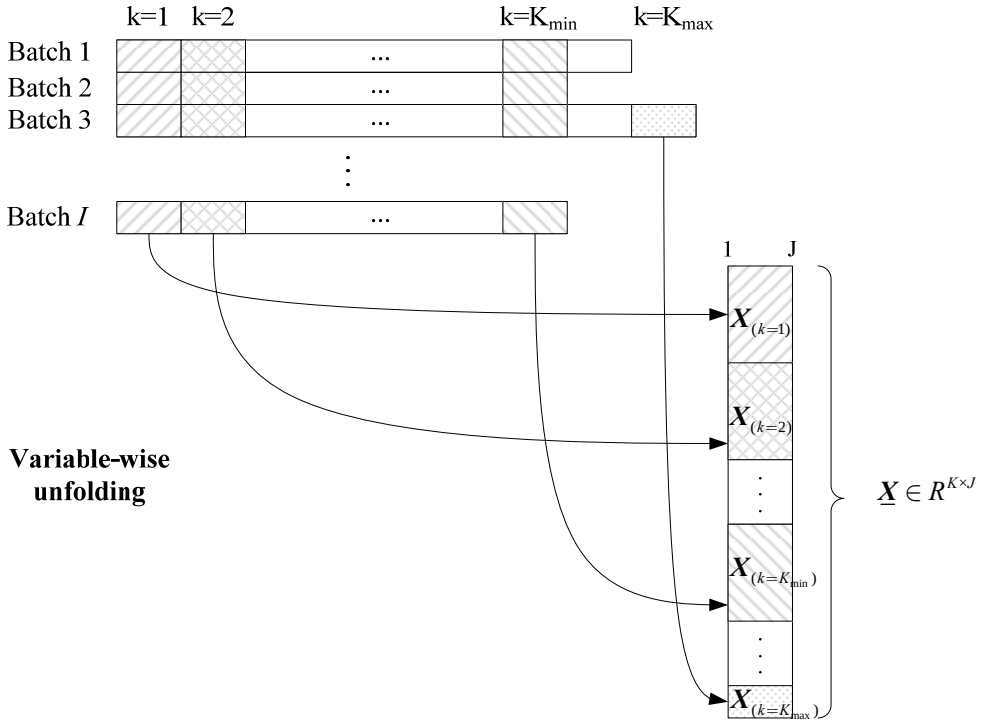

Fig.1. Illustration of the variable-wise unfolding of uneven-length training batches

Since the sample numbers of the comprehensive economic index and process

variables must be the same in modeling, the comprehensive economic index of each batch is extended to a vector by stacking it K_i times repeatedly, and then y_i is extended as $\boldsymbol{y}_i = [y_i, y_i, ..., y_i]^T \in R^{K_i \times 1}$. Furthermore, the comprehensive economic index data of all training batches are unfolded as $\underline{\boldsymbol{y}} = [\boldsymbol{y}_{(k=1)}^T, \boldsymbol{y}_{(k=2)}^T, \ldots, \boldsymbol{y}_{(k=K_{\min})}^T, \ldots, \boldsymbol{y}_{(k=K_{\max})}^T]^T \in R^{K \times 1}$. For simplicity, the normalized forms are still denoted as $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{y}}$, respectively.

In multiphase batch processes, if process variables have different change rates or amplitudes in two adjacent phases and the change rates are not very large in each individual phase, the unfolded process data approximately follow a multivariate Gaussian distribution in each individual phase. Hence, GMM-PSD strategy [36] is adopted for offline phase division. Firstly, with the unfolded data $\underline{\boldsymbol{X}}$, GMM is estimated through the F-J algorithm as formulated in Eqs. (3)~(6). Then the means of the posterior probabilities of the first h samples of each training batch are calculated with respect to each Gaussian component as follows:

$$P(C_m \mid \boldsymbol{x}_{i,h'}) = \frac{\omega_m g(\boldsymbol{x}_{i,h'} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{m=1}^{M} \omega_m g(\boldsymbol{x}_{i,h'} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}, \quad i = 1, 2, ..., I, m = 1, 2, ..., M, h' = 1, 2, ..., h, \tag{15}$$

$$\bar{P}_m = \frac{1}{Ih} \sum_{i=1}^{I} \sum_{h'=1}^{h} P(C_m \mid \boldsymbol{x}_{i,h'}), m = 1, 2, ..., M. \tag{16}$$

The target Gaussian component of phase 1 is determined by searching for the Gaussian component with the largest posterior probability, i.e., $m^* = \arg\max_m (\bar{P}_m)$. Thereafter, from the $h+1$ th sample of each training batch, only the posterior probability with respect to the target Gaussian component of phase 1, C_{m^*}, is calculated and used to determine whether the sample is still belonging to phase 1. Based on Bayesian inference, the posterior probabilities of the samples of phase 1 should be greater than those with respect to other Gaussian components. Hence, if the posterior probability is greater than a given threshold, it can be sure that the sample belongs to phase 1. In this way, the end moments of phase 1 of each batch are

determined finally. Because the phase may be uneven-length across different training batches, the end moments of phase 1 are different. As phase 1 is identified, the process data belonging to phase 1 are removed from each batch, and then the aforementioned procedures is applied to identify the second phase. This process is conducted iteratively until all phases are divided. Correspondingly, $\underline{\boldsymbol{y}}$ is divided into different data sets along with $\underline{\boldsymbol{X}}$. The detailed steps of offline phase division by GMM-PSD method can be referred in Ref. 36.

Assuming that C phases are identified from offline phase division, the process data and comprehensive economic index data of phase c are denoted as $\boldsymbol{X}^c = \left[\boldsymbol{x}_1^c, \boldsymbol{x}_2^c, ..., \boldsymbol{x}_{N_c}^c\right]^T \in R^{N_c \times J}$ and $\boldsymbol{y}^c = \left[y_1^c, y_2^c, ..., y_{N_c}^c\right]^T \in R^{N_c \times 1}, c = 1, 2, ..., C$, respectively. N_c is the number of samples of phase c. Then the GPR-based assessment model of phase c is established and formulated as follows:

$$\begin{aligned} P(\boldsymbol{y}^c \left| \boldsymbol{X}^c, \boldsymbol{\alpha}^c \right.) &= \prod_{n=1}^{N_c} g(y_n^c \left| \boldsymbol{x}_n^c, \boldsymbol{\alpha}^c \right.) \\ &= \prod_{n=1}^{N_c} \frac{1}{\sqrt{2\pi}(\sigma^c)^2} \exp\left[-\frac{(y_n^c - \boldsymbol{x}_n^{cT}\boldsymbol{\alpha}^c)^2}{2(\sigma^c)^2} \right] \\ &= \frac{1}{(2\pi)^{N_c/2}\left|(\sigma^c)^2 \boldsymbol{I}\right|^{1/2}} \exp\left[-\frac{1}{2(\sigma^c)^2} \left\| \boldsymbol{y}^c - \boldsymbol{X}^c\boldsymbol{\alpha}^c \right\|^2 \right] \\ &\sim \mathcal{N}(\boldsymbol{X}^c\boldsymbol{\alpha}^c, (\sigma^c)^2 \boldsymbol{I}), \end{aligned} \tag{17}$$

where $\boldsymbol{\alpha}^c$ is the regression parameter vector and follows a Gaussian prior distribution with zero mean and covariance $\boldsymbol{\Sigma}_\alpha^c$; σ^c is the standard deviation of the Gaussian noise.

For a new sample $\boldsymbol{x}_{new}$ that belongs to phase c, the predictive distribution of $\hat{y}_{new}^c = f(\boldsymbol{x}_{new})$ can be formulated as follows:

$$\begin{aligned} P(\hat{y}_{new}^c \left| \boldsymbol{x}_{new}, \boldsymbol{X}^c, \boldsymbol{y}^c \right.) &= \int P(\hat{y}_{new}^c \left| \boldsymbol{x}_{new}, \boldsymbol{\alpha}^c \right.) P(\boldsymbol{\alpha}^c \left| \boldsymbol{X}^c, \boldsymbol{y}^c \right.) d\boldsymbol{\alpha}^c \\ &\sim \mathcal{N}((\sigma^c)^{-2} \boldsymbol{x}_{new}^T (\boldsymbol{A}^c)^{-1} \boldsymbol{X}^{cT} \boldsymbol{y}^c,\ \boldsymbol{x}_{new}^T (\boldsymbol{A}^c)^{-1} \boldsymbol{x}_{new}), \end{aligned} \tag{18}$$

where $\boldsymbol{A}^c = (\sigma^c)^{-2} \boldsymbol{X}^{cT} \boldsymbol{X}^c + (\boldsymbol{\Sigma}_\alpha^c)^{-1}$.

In modeling, some computations require inversions of covariance matrices. However, it doesn't necessarily exist in real world scenarios for the collinearity of

process variables. In this case, the Moore–Penrose pseudo inverse [44] is calculated to solve the inverse of a singular matrix.

3.2 Online operating performance assessment

In online assessment, the basic premise is to choose the correct active assessment model, and this involves the issue of online phase identification for the new sample. As all phases have been identified during the offline phase division, the start and end moments of phase c of the i th batch are known and denoted as $k_{i,in}^{c}$ and $k_{i,out}^{c}$, respectively. Then the maximum interval of phase c is written as $[k_{in}^{c}, k_{out}^{c}]$ accordingly, where $k_{in}^{c} = \min_{1\le i\le I}(k_{i,in}^{c})$ and $k_{out}^{c} = \max_{1\le i\le I}(k_{i,out}^{c})$ are the earliest start and latest end moments of phase c. Since the uneven-length duration also presents in the single phase across different batches, k_{in}^{c} may be less than k_{out}^{c-1}, $c = 2,3,...,C-1$, and this further results in an overlapping interval between two adjacent phases, i.e., $\left[k_{in}^{c}, k_{out}^{c-1}\right]$. In the nonoverlapping interval, i.e., $(k_{out}^{c-1}, k_{in}^{c+1})$, the phase type of the new sample can be identified uniformly, thus it is called as certain interval; while the phase type is unknown if a sample falls into the overlapping interval of two phases, and this kind of interval is described as fuzzy interval. According to the intervals of each phase, the phase type of the new sample can be determined. Table 1 gives a demonstration of the intervals and the corresponding phase identification results.

Table 1 Intervals and online phase identification result

Interval	Phase type	Phase identification result
$[1, k_{in}^{2})$	certain interval	1
$[k_{in}^{2}, k_{out}^{1}]$	fuzzy interval	1 or 2
⋮	⋮	⋮
$[k_{in}^{c}, k_{out}^{c-1}]$	fuzzy interval	$c-1$ or c
$(k_{out}^{c-1}, k_{in}^{c+1})$	certain interval	c
⋮	⋮	⋮
⋮	⋮	⋮
$(k_{out}^{C-1}, k_{out}^{C}]$	certain interval	C

Supposing that the new sample belongs to phase c certainly, the assessment model of phase c is thus invoked, and the predicted output of $\boldsymbol{x}_{new}$ is given by

$$\hat{y}_{new}^{c} = (\sigma^{c})^{-2} \boldsymbol{x}_{new}^{T} (\boldsymbol{A}^{c})^{-1} \boldsymbol{X}^{cT} \boldsymbol{y}^{c}. \quad (19)$$

In order to simplify the assessment process, the predicted comprehensive economic index should be normalized, and the normalized form is defined as the assessment index and used for online operating performance assessment. Without loss of generality, assuming that the operating performance is optimal when comprehensive economic index is high, and the assessment index, γ_c, is calculated as follows:

$$\gamma_c = \begin{cases} 1, & if\ \hat{y}_{new}^{c} \geq y_{\max}^{c} \\ \dfrac{\hat{y}_{new}^{c} - y_{\min}^{c}}{y_{\max}^{c} - y_{\min}^{c}}, & if\ y_{\min}^{c} < \hat{y}_{new}^{c} < y_{\max}^{c}, \\ 0, & if\ \hat{y}_{new}^{c} \leq y_{\min}^{c} \end{cases} \quad (20)$$

where $y_{\max}^{c} = \max_{n}(y_n^c)$ and $y_{\min}^{c} = \min_{n}(y_n^c)$, $n = 1, 2, ..., N_c$, are the maximum and minimum values of the comprehensive economic index of phase c, respectively. As seen from Eq.(20), γ_c is between 0 and 1. When γ_c is close to 1, it means that the predicted value is close to the optimal value appeared in historical data, and the process operating performance is usually optimal; otherwise, if γ_c is nearly 0, the process operating performance is likely to be nonoptimal. In order to distinguish optimality from nonoptimality strictly, an assessment index threshold, $\eta (0.5 < \eta < 1)$, is introduced. If $\gamma_c \geq \eta$, it means that the process operating performance is optimal. On the contrary, we can say that the process is operating on the nonoptimal performance grade. The value of η can be determined through historical data and expert experience. According to the historical data, the maximum and minimum values of the comprehensive economic index can be counted, and then experts who are familiar with the production process can give a dividing value between optimality and nonoptimality. Furthermore, η is calculated as in Eq. (20) by replacing the predicted value of the comprehensive economic index with the dividing value.

Particularly, if the new sample falls into the fuzzy interval $[k_{in}^{c}, k_{out}^{c-1}]$, both the assessment indices γ_{c-1} and γ_c are calculated, and Bayesian inference-based posterior

probabilities of $\boldsymbol{x}_{new}$ with respect to phase $c-1$ and c are set as the adaptive weights to integrate the corresponding local assessment results.

The posterior probabilities are calculated as follows:

$$P^{c-1}=\frac{\tilde{\omega}^{c-1}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c-1},\boldsymbol{\Sigma}^{c-1})}{\tilde{\omega}^{c-1}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c-1},\boldsymbol{\Sigma}^{c-1})+\tilde{\omega}^{c}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c},\boldsymbol{\Sigma}^{c})}, \tag{21}$$

$$P^{c}=\frac{\tilde{\omega}^{c}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c},\boldsymbol{\Sigma}^{c})}{\tilde{\omega}^{c-1}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c-1},\boldsymbol{\Sigma}^{c-1})+\tilde{\omega}^{c}g(\boldsymbol{x}_{new}|\boldsymbol{\mu}^{c},\boldsymbol{\Sigma}^{c})}, \tag{22}$$

where $\tilde{\omega}^{c-1}=\omega^{c-1}/(\omega^{c-1}+\omega^{c})$ and $\tilde{\omega}^{c}=\omega^{c}/(\omega^{c-1}+\omega^{c})$ represent the prior probabilities with respect to phase $c-1$ and c, respectively.

Furthermore, the assessment indices can be integrated as follows:

$$\gamma=P^{c-1}\gamma_{c-1}+P^{c}\gamma_{c}. \tag{23}$$

Then the process operating performance can be evaluated by comparing the assessment index γ with the threshold η.

3.3 Nonoptimal cause identification

Although the nonoptimal operating performance is not expected, it may occur due to manual operating error or process characteristics drift in production. Therefore, when the process operating performance is nonoptimal, it is much important to identify the possible cause for actual production processes. Being similar to the contribution plots-based fault diagnosis methods, [45, 46] the nonoptimal cause identification method is proposed based on variable contributions in this section. By approximately decomposing the assessment index into the weighted sum of different process variables, the variable contributions are constructed firstly. Then compare the contribution values with their statistical scopes obtained from the optimal level, and identify the responsible variables according to the given criterion. Since the contribution is a function of the process variable, the nonoptimal cause can be identified as long as the nonoptimality can be reflected through the measurable process variables. Therefore, the proposed method is applicable to nonoptimal conditions caused by sensors and process abnormalities.

As known from Eq.(20), the variable contribution to the assessment index γ_c is equivalent to the contribution to $\hat{y}_{new}^c$, so $\hat{y}_{new}^c$ is further decomposed into the following form:

$$\hat{y}_{new}^c = \boldsymbol{x}_{new}^T \boldsymbol{z}^c = \sum_{j=1}^{J} x_{new,j} z_j^c, \tag{24}$$

where $\boldsymbol{z}^c = (\sigma^c)^{-2}(\boldsymbol{A}^c)^{-1}\boldsymbol{X}^{cT}\boldsymbol{y}^c$; z_j^c is the j th element of $\boldsymbol{z}^c$, and $x_{new,j}$ is the j th variable of $\boldsymbol{x}_{new}$. Accordingly, the raw contribution of the j th process variable with respect to the assessment index γ_c is denoted as below:

$$contr_{raw,j}^c = x_{new,j} z_j^c, j = 1,2,...,J. \tag{25}$$

Intuitively, small $contr_{raw,j}^c$ causes small $\hat{y}_{new}^c$, and further leads to small γ_c. So it looks like that the process variables with smaller raw contributions should be considered as the responsible variables for nonoptimality. However, even though under the optimal operating performance, the raw contributions of different process variables are not the same, and some of them may be very small or negative. Therefore, it is more reasonable to compare each raw variable contribution with its average level under optimal operating performance, rather than only considering its absolute value. In view of this, $\tilde{I}$ batches under the optimal operating performance are selected from I training batches and used to calculate the mean of the raw contribution of each process variable. The reference data of each phase are denoted as $(\tilde{\boldsymbol{X}}^1, \tilde{\boldsymbol{y}}^1), (\tilde{\boldsymbol{X}}^2, \tilde{\boldsymbol{y}}^2), ..., (\tilde{\boldsymbol{X}}^C, \tilde{\boldsymbol{y}}^C)$, where $\tilde{\boldsymbol{X}}^c = [\tilde{\boldsymbol{x}}_1^c, \tilde{\boldsymbol{x}}_2^c, ..., \tilde{\boldsymbol{x}}_{\tilde{N}_c}^c]^T$ and $\tilde{\boldsymbol{y}}^c = [\tilde{y}_1^c, \tilde{y}_2^c, ...\tilde{y}_{\tilde{N}_c}^c]^T$, $c = 1,2,...,C$. $\tilde{\boldsymbol{A}}^c = (\tilde{\sigma}^c)^{-2}\tilde{\boldsymbol{X}}^{cT}\tilde{\boldsymbol{X}}^c + (\tilde{\boldsymbol{\Sigma}}_\alpha^c)^{-1}$ and $\tilde{\boldsymbol{z}}^c = (\tilde{\sigma}^c)^{-2}(\tilde{\boldsymbol{A}}^c)^{-1}\tilde{\boldsymbol{X}}^{cT}\tilde{\boldsymbol{y}}^c$ are the related parameters. The mean of the raw contribution of the j th process variable of phase c is calculated as follows:

$$contr_{mean,j}^c = \sum_{n=1}^{\tilde{N}_c} \tilde{x}_{n,j}^c \tilde{z}_j^c \Big/ \tilde{N}_c, \tag{26}$$

where $\tilde{z}_j^c$ is the j th element of $\tilde{\boldsymbol{z}}^c$, and $\tilde{x}_{n,j}^c$ is the j th variable of $\tilde{\boldsymbol{x}}_n^c, n = 1,2,...,\tilde{N}_c$.

Then we define the variable contribution as below:

$$contr_j^c = contr_{raw,j}^c - contr_{mean,j}^c, j = 1,2,...,J. \quad (27)$$

In theory, if $contr_j^c < 0$, it means that the j th variable is the responsible variable for the nonoptimal operating performance. Nevertheless, it is difficult to achieve an absolutely equal between $contr_{raw,j}^c$ and $contr_{mean,j}^c$ in actual processes. Hence, the upper and lower confidence limits of the variable contributions of the reference data are counted by kernel density estimation [47] for different phases, and they are denoted as CU_j^c and CL_j^c, respectively. If $CL_j^c \le contr_j^c \le CU_j^c$, it means that $contr_{raw,j}^c$ and $contr_{mean,j}^c$ are approximately equal statistically; otherwise, it can be concluded that they are obviously unequal. Furthermore, the process variables with contributions that are significantly less than CL_j^c are considered as the cause variables responsible for the nonoptimal operating performance. For the process variables whose contributions are greater than CU_j^c, they may play a role in improving the operating performance.

If $\boldsymbol{x}_{new}$ belongs to the fuzzy interval $[k_{in}^c, k_{out}^{c-1}]$, both $contr_j^{c-1}$ and $contr_j^c$ should be calculated, and the posterior probabilities are used as the adaptive weights to integrate them as the final variable contribution, i.e.,

$$contr_j = P^{c-1} contr_j^{c-1} + P^c contr_j^c, j = 1,2,...,J. \quad (28)$$

Correspondingly, the upper and lower confidence limits are constructed as follows:

$$\begin{aligned} CU_j &= P^{c-1} CU_j^{c-1} + P^c CU_j^c, \\ CL_j &= P^{c-1} CL_j^{c-1} + P^c CL_j^c, j = 1,2,...,J, \end{aligned} \quad (29)$$

The step-by-step procedure of GPR and Bayesian inference based operating performance assessment approach is summarized as below and the flow diagram is shown in Fig. 2.

(1) Collect the training batches for the establishment of assessment models.

(2) Unfold the three-dimension data into two-dimension matrix via variable-wise unfolding and perform the normalization on the unfolded matrix.

(3) Estimate GMM and identify all different phases based on GMM-PSD strategy.

(4) Develop GPR-based assessment models for different phases.

(5) Identify the phase type for new sample and invoke the corresponding model.

(6) Predict the comprehensive economic index of the new sample and construct the assessment index with it.

(7) Online assessment using the constructed assessment index.

(8) Identify the cause variables responsible for the nonoptimal operating performance based on variable contributions.

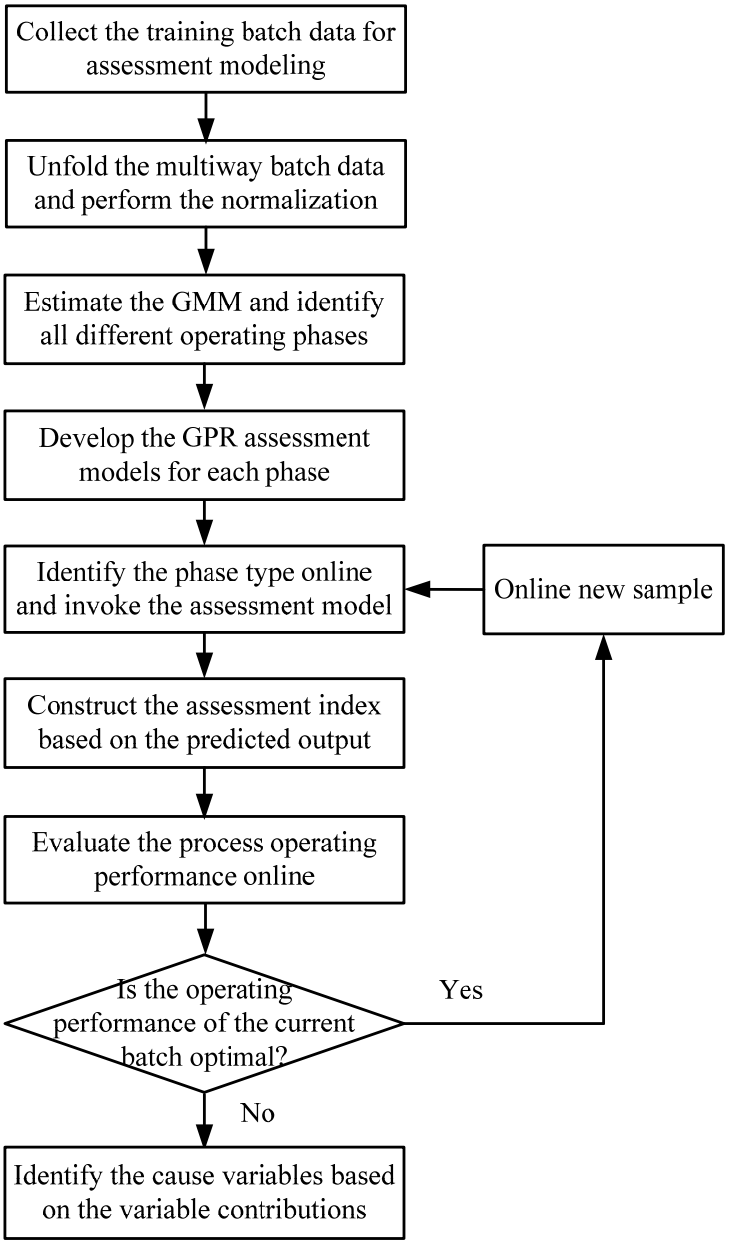

Fig.2. Diagram of GPR and Bayesian inference based operating performance assessment approach

4. Case study

4.1 Fed-batch penicillin fermentation process description

The fed-batch penicillin fermentation process [48, 49] is a typical multiphase batch process and has received wide attentions for its academic and industrial importance. There are two physical stages in this process: preculture and fed-batch. The preculture stage starts with small amounts of biomass and substrate. Most of the initially added substrate is consumed by the microorganisms after about 45 h, and the process is switched from preculture stage to fed-batch stage which usually continues about 355 h. In fed-batch stage, the penicillin increased exponentially until the stationary procedure. Because the biomass growth rate must be kept constant for the aim of optimizing penicillin production, the substrate is supplied continuously into the fermentor, rather than being added all at once in the beginning. Meanwhile, in order to maintain the constant temperature and pH values, two proportional-integral-derivative (PID) control loops are implemented in the fermentor for manipulating the acid/base and hot/cold water flow ratios, respectively. In 2002, a modular simulator (PenSim v2.0) for fed-batch fermentation was developed by Ali Çinar et al. from the Monitoring and Control Group of the Illinois Institute of Technology. It can simulate the concentrations of biomass, CO_2, hydrogen ion, penicillin, carbon source, oxygen, and heat generation under various operating conditions. The diagram of the penicillin fermentation process is shown in Fig. 3.

Known from the mechanism of penicillin fermentation process, the process operating performance depends on the final penicillin concentration. Under the normal operating conditions, the higher penicillin concentration corresponds to the better operating performance, while the poorer operating performance usually leads to lower penicillin concentration. Thus, the final penicillin concentration is used as the comprehensive economic index in this study. In addition, the process variables related to the penicillin concentration are selected for operating performance assessment and listed in Table 2. Table 3 gives the initial operating conditions of the production.

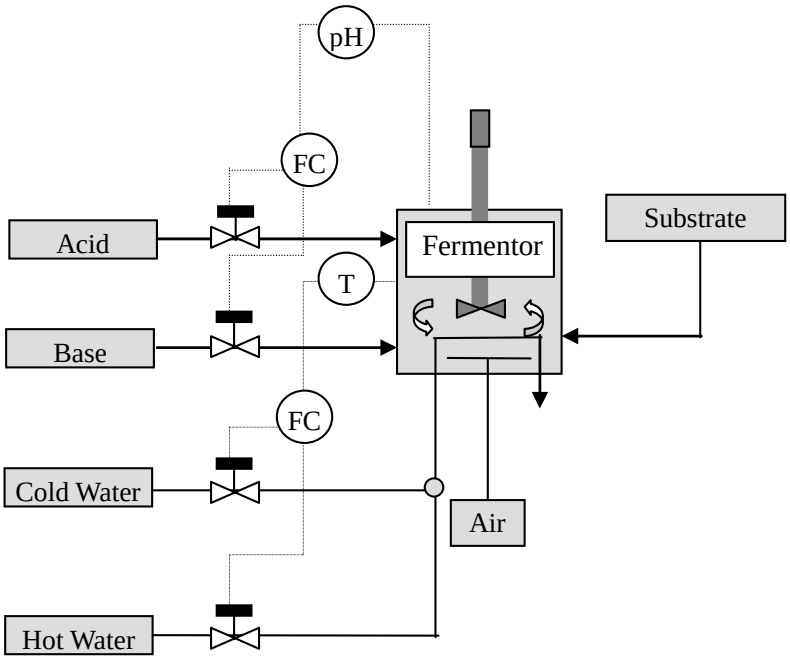

Fig. 3 Flow diagram of the penicillin fermentation process

Table 2 Process variables used for operating performance assessment

No.	Process variables	Normal operating ranges
1	Aeration rate (L/h)	3-10
2	Agitator power (W)	20-50
3	Substrate feed rate (L/h)	0.035-0.045
4	Substrate feed temperature (K)	296-298
5	Substrate concentration (g/L)	5-50
6	Dissolved oxygen concentration (%)	1.16
7	Biomass concentration (g/L)	0-0.2
8	Culture volume (L)	100-150
9	Carbon dioxide concentration (mole/L)	0.5-1
10	pH	4-6
11	Generated heat (kcal/h)	0

Table 3 Initial operating conditions of penicillin fermentation process

Process variables	Initial operating conditions
Substrate concentration(g/L)	15
Dissolved oxygen concentration (%)	1.16
Biomass concentration (g/L)	0.1
Penicillin concentration (g/L)	0
Culture volume (L)	100
Carbon dioxide concentration (mole/L)	0.5
pH	5
Fermentor temperature (K)	298
Generated heat (kcal/h)	0

To establish the assessment models, a total of 50 training batches are produced

under the normal operating conditions, and the durations are between 390h and 420h with the sampling interval of 1h. By using GMM-PSD method, five phases are identified from offline phase division. Table 4 shows the result of offline phase division, and the preculture stage and fed-batch stage are further divided into two and three phases, respectively. This can be attributed to the fact that a stage may contain multiple data distributions caused by the intrinsic biochemical reactions or the changes in external operating conditions. In preculture stage, most of the initially added substrates are consumed quickly by the microorganisms. To capture the regularity of data distribution, meanwhile, ensure that the data of each phase approximately follow a multivariate Gaussian distribution, the preculture stage is further divided into two phases, although its duration is only about 45 h. The fed-batch stage, whose duration is about 355 h and almost eight times length of preculture stage, is just divided into three phases. It is because that the variation trends of the variable trajectories noticeably slows down in this stage, and three phases are enough to describe the diversity of data distributions. In addition, the batch-to-batch variations cause the inconsistent phase durations across different batches, thus form the fuzzy intervals between phases.

Table 4 Result of offline phase division

Phase no.	Earliest start moment k_{in}^{c}	Latest end moment k_{out}^{c}	Phase division result
1	1	34	Certain interval of phase 1: $[1,30)$; Fuzzy interval between phase 1 and 2: $[30,34]$;
2	30	47	Certain interval of phase 2: $(34,43)$; Fuzzy interval between phase 2 and 3: $[43,47]$;
3	43	125	Certain interval of phase 3: $(47,106)$; Fuzzy interval between phase 3 and 4: $[106,125]$;
4	106	223	Certain interval of phase 4: $(125,195)$; Fuzzy interval between phase 4 and 5: $[195,223]$;
5	195	420	Certain interval of phase 5: $(223,420]$.

To test whether the process data of each phase follow Gaussian distributions, the multivariate normality test algorithm proposed in Ref. 50, i.e., F-Straight Method Based on Mahalanobis Distance (FSMD), is implemented before modeling. The basic idea of FSMD is that, when the process data approximately follow a multivariate

Gaussian distribution, the distribution function of the squared Mahalanobis distances of samples, $D_{(n)}$, is a specific F distribution, and it can be replaced by its empirical distribution function. Denote the quantile of F distribution as F_n, and the above argument is equivalent to test whether $D_{(n)} \approx F_n$. Let the linear regression equation between $D_{(n)}$ and F_n be $F_n = a + bD_{(n)}$, the significance test is performed on it based on the regression standard deviation, s, and the mean of the quantile, $\overline{F}$. If $s/\overline{F} \le 0.15$, the significance test is valid and the linear regression equation can reflect the real relationships between $D_{(n)}$ and F_n; otherwise, the significance test is invalid. Furthermore, if both of the conditions $|a| < \sigma$ and $|b-1| < \beta$ are satisfied, where σ and β are usually set as $\overline{F} \times 5\%$ and 0.2, it is usually considered that the fitting line, $F_n = a + bD_{(n)}$, passes through the original point with the slope of 1. Thus, the hypothesis test that the data follows a Gaussian distribution should not be rejected. Otherwise, reject the null hypothesis. In summary, when the conditions $s/\overline{F} \le 0.15$, $|a| < \sigma$ and $|b-1| < \beta$ are satisfied at the same time, the data follow a Gaussian distribution. The detailed procedures of FSMD are given in Appendix A. As shown in Table 5 and Fig. 4, the test results illustrate that, in each individual phase, the unfold process data approximately follow a multivariate Gaussian distribution.

Table 5 Results of multivariate normality test of each phase

Phase no.	$s/\overline{F}$ (<0.15)	$\lvert a \rvert$	$\overline{F} \times 5\%$	$\lvert b-1 \rvert$ (<0.2)	Distribution characteristics
1	0.0339	1.3768	1.8321	0.1238	Gaussian
2	0.0410	1.4559	1.8965	0.1267	Gaussian
3	0.0312	1.1879	1.8133	0.1031	Gaussian
4	0.0194	0.9076	1.7448	0.0873	Gaussian
5	0.0239	1.3765	1.6544	0.0993	Gaussian

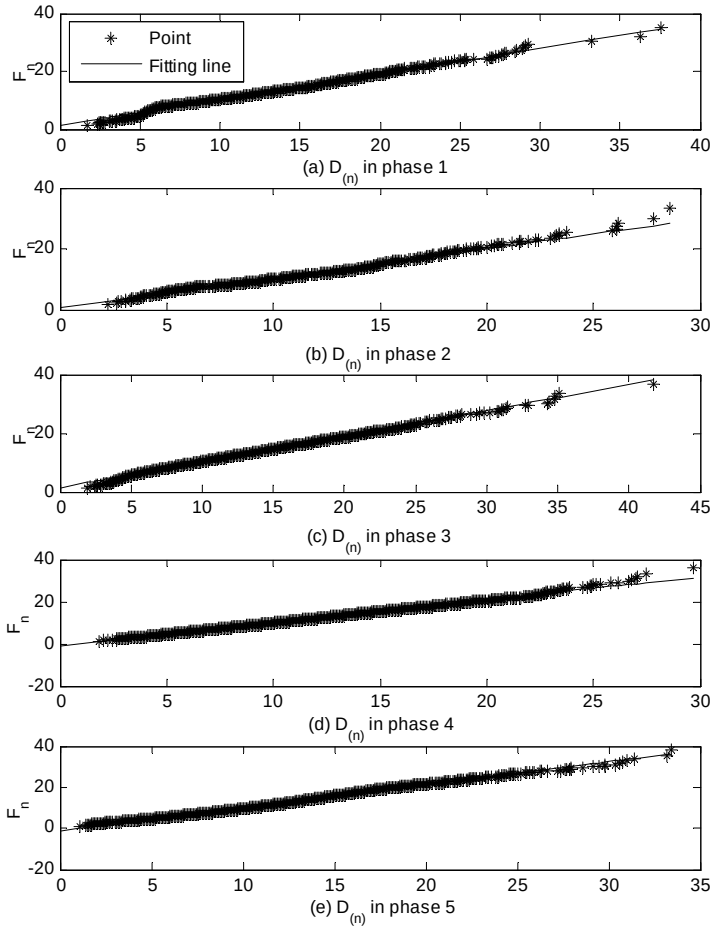

Fig. 4 Multivariate normality test of the process data of each phase ('*' denotes the point $(D_{(n)}, F_n)$, '—' is the fitting line $F_n = a + bD_{(n)}$)

Then GPR-based assessment models are developed for each phase. Additionally, the root mean square error (RMSE) and maximum relative error (MRE) are used to evaluate the quality and reliability of the assessment models:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n_t=1}^{N_t} (y_{n_t} - \hat{y}_{n_t})^2}{N_t}}, \tag{30}$$

$$\mathrm{MRE} = \max_{n_t} \left\{ \left| \frac{y_{n_t} - \hat{y}_{n_t}}{y_{n_t}} \right| \right\}, n_t = 1, 2, ..., N_t, \tag{31}$$

where N_t is the number of samples, y_{n_t} and $\hat{y}_{n_t}$ are the actual and predicted values of the comprehensive economic index of the n_t th sample, respectively. The RMSE and MRE of the modeling data are summarized in Table 6. The values of RMSE and MRE

are all very small, so the predictive abilities of the assessment models are satisfactory for the modeling data.

Table 6 RMSE and MRE of the modeling data

Phase no.	RMSE	MRE
1	0.0257	0.0435
2	0.0298	0.0339
3	0.0543	0.0309
4	0.0337	0.0859
5	0.0207	0.0537

4.2 Operating performance assessment and nonoptimal cause identification

In online assessment, two test batches under optimal and nonoptimal operating performances are generated separately. In optimal test batch, the aeration rate, agitator power, substrate feed rate, pH, and culture volume are in their optimal ranges as shown in Table 7. Its duration is 412 h. According to the process mechanism [48], the fed-batch process operation causes a volume change in the fermentor, and the loss in volume due to evaporation is very significant in penicillin production. Some important raw materials, such as biomass and substrate, will be lost along with evaporation, and it could further affect the final penicillin concentration. Thus, the nonoptimal operating performance can be simulated by adjusting the culture volume to 150L, which is within the normal operating range but outside the optimal range. Temperature and culture volume are two important factors associated with the evaporative loss. When the temperature is kept as a constant, the evaporative loss just depends on the culture volume. The larger the culture volume is, the more the evaporative loss is, which eventually results in the decline of the penicillin concentration. The duration of the nonoptimal test batch is 392h. Fig. 5 shows the variable trajectories of optimal and nonoptimal test batches, as well as the division lines between intervals at sampling instant 30, 34, 43, 47, 106, 125, 195, and 223 given by Table 4, respectively. It is clearly seen from Fig. 5 that the trajectories of culture volumes in two test batches are significant differences throughout the batch runs, and the trajectories of other process variables also present different change trends after the production going into the fed-batch stage. Thus, it leads to a great

difference in the final penicillin concentrations under optimal and nonoptimal operating performances.

Table 7 Ranges of optimal operating performance

Process variables	Optimal operating ranges
Aeration rate (L/h)	8.5-10
Agitator power (W)	29-31
Substrate feed rate (L/h)	0.042-0.045
pH	4.9-5.1
Culture volume (L)	100-115

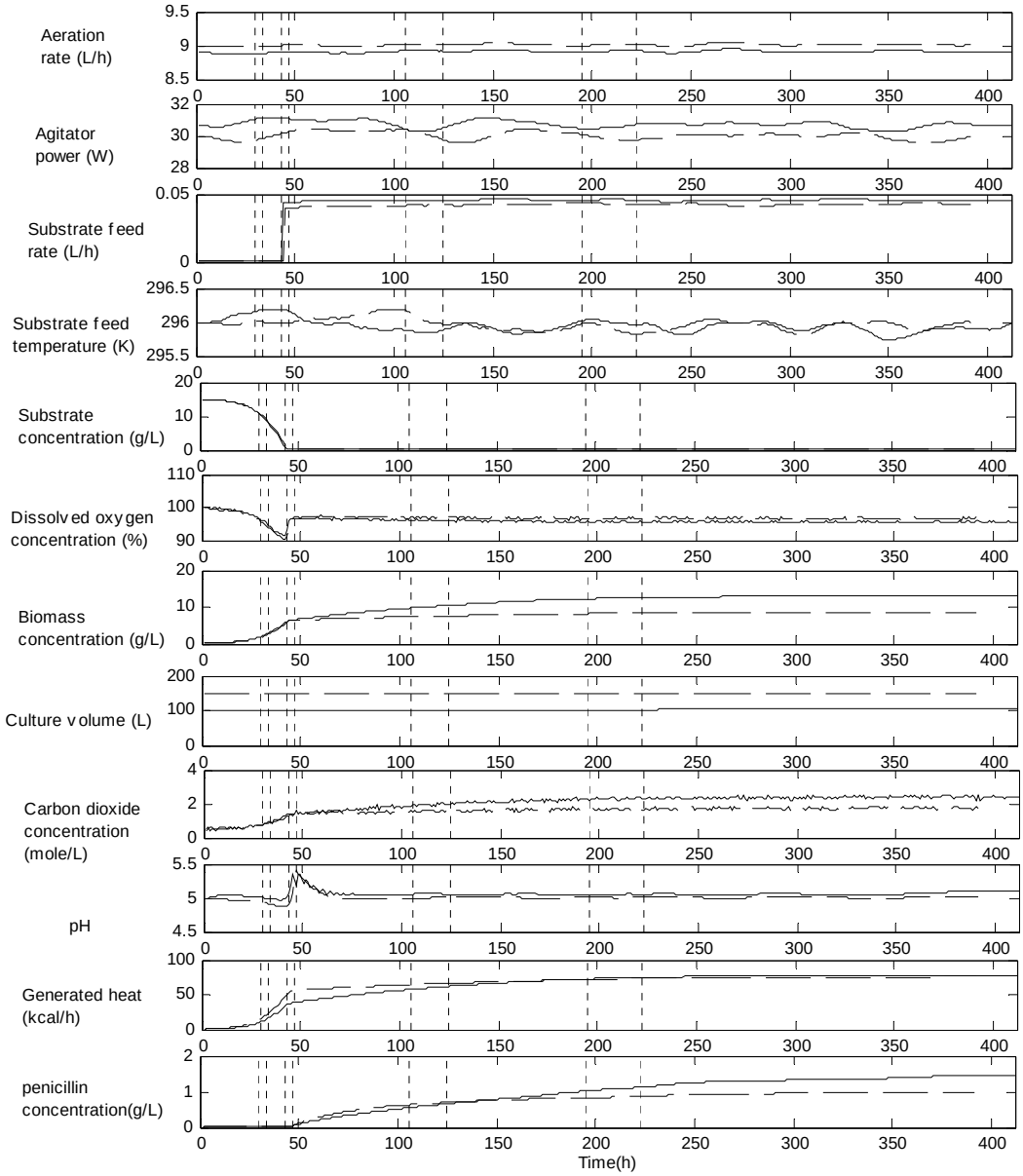

Fig. 5 Trajectories of process variable and penicillin concentration of optimal and nonoptimal test batches ('—'denotes optimal test batch, '--'represents nonoptimal test batch, and ':' is the division line between certain and fuzzy intervals)

The proposed assessment method is first applied to the optimal test batch. The predicted penicillin concentration and the actual value are shown in Fig. 6(a), as well as the relative errors calculated by $\mathrm{RE}_{n_t} = (y_{n_t} - \hat{y}_{n_t}) / y_{n_t}$ given in Fig. 6(b). As can be seen from Fig. 6, the predicted values match well the trend of the actual values. Furthermore, both RMSE and MRE are small enough as shown in Table 8. It indicates that GPR-based assessment models have satisfactory predictive performance, and it is applicable to predict the final penicillin concentration for online evaluation. The assessment index threshold η is set as 0.8. Fig. 7 displays the online assessment result for the optimal test batch. The assessment indices are larger than 0.8 throughout the batch production, and the operating performance of the test batch is optimal, so the online assessment result is highly consistent with the actual situation.

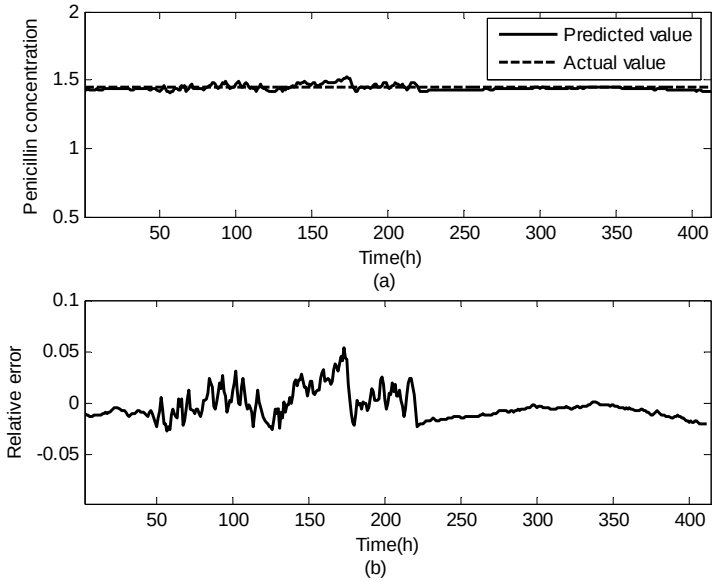

Fig. 6 Predicted final penicillin concentration and relative errors of optimal test batch

Table 8 RMSE and MRE of optimal test batch

Phase no.	RMSE	MRE
1	0.0100	0.0132
2	0.0098	0.0130
3	0.0134	0.0308
4	0.0194	0.0546
5	0.0111	0.0231

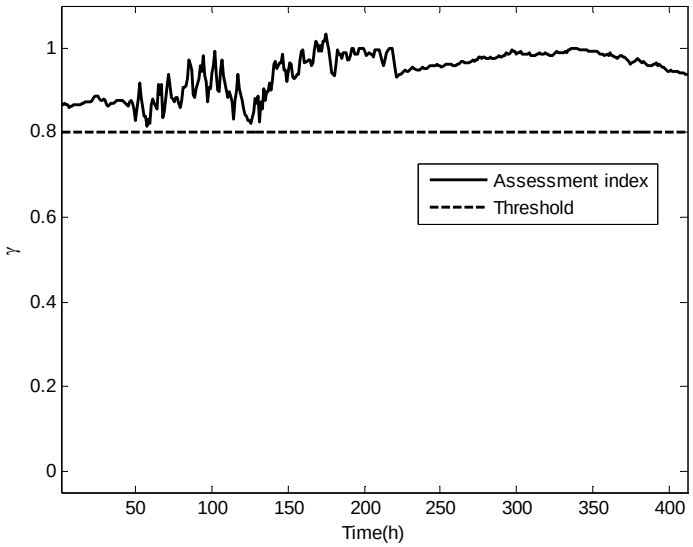

Fig. 7 Online assessment of the optimal test batch

For nonoptimal test batch, the predicted penicillin concentration and the actual values are shown in Fig. 8(a), and the relative errors between them are given in Fig. 8(b). Being similar to the optimal case, most of the samples have been well estimated with very small relative errors. Table 9 summarizes the RMSE and MRE, and both of them are small enough to verify the accuracy of the prediction results. Fig. 9 shows the online assessment result of the nonoptimal test batch. Since the assessment indices are much lower than the threshold, it means that the nonoptimal operating performance has been accompanied by this batch until the end. In order to further identify the reasons for the nonoptimal operating performance, the variable contributions are calculated at the initial moment of nonoptimality, as shown in Fig. 10. Only the contribution of variable 8 (culture volume) is beyond the corresponding lower confidence limit. Because the culture volume is just the responsible variable for the nonoptimality, the cause identification result is consistent with the actual situation.

Table 9 RMSE and MRE of nonoptimal test batch

Phase no.	RMSE	MRE
1	0.0244	0.0294
2	0.0218	0.0312
3	0.0282	0.0751
4	0.0267	0.0700
5	0.0120	0.0302

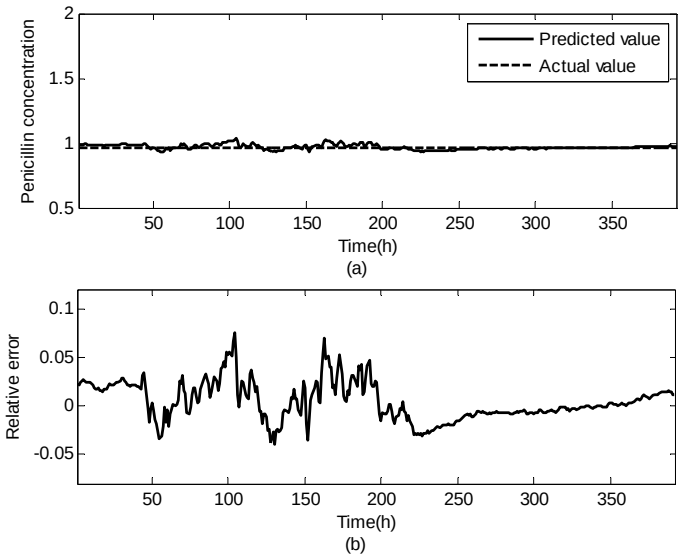

Fig. 8 Predicted final penicillin concentration and relative errors of nonoptimal test batch

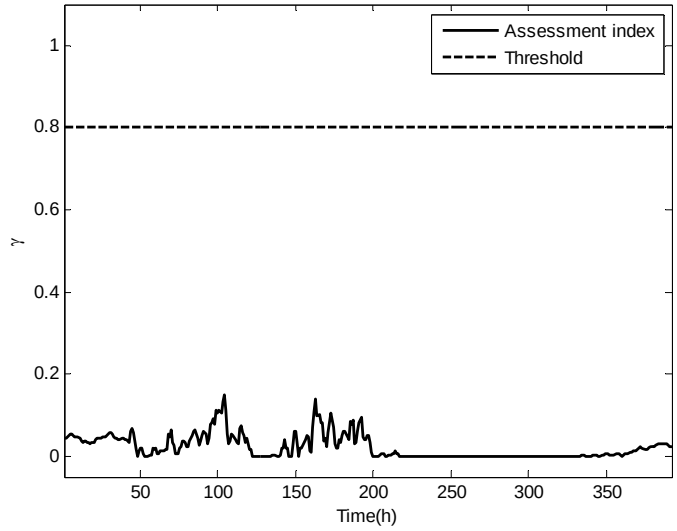

Fig. 9 Online assessment of nonoptimal test batch

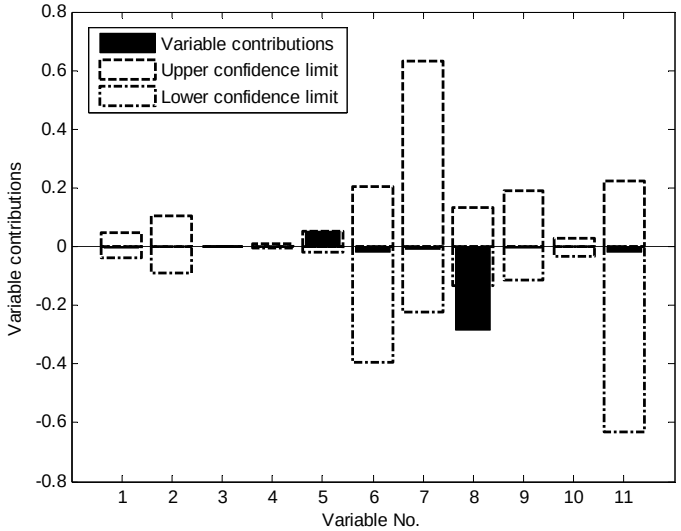

Fig. 10 Variable contributions under nonoptimal operating performance

In summary, no matter under optimal or nonoptimal operating performance, the process operating performance can be evaluated timely and accurately by the proposed method, and the responsible process variables can be identified to reflect the actual situation.

5. Conclusions

This paper proposes a novel operating performance assessment method for multiphase batch processes based on GPR and Bayesian inference. To take the process dynamics and batch-to-batch uncertainty into account, GPR-based assessment models are established for different phases. Furthermore, from the view point of data distribution, the local influences of production operation on the operating performance of the whole batch are described for each phase. Since the uneven-length characteristic embodies in phases across different batches, the online phase identification results can be a certain interval or a fuzzy interval between two adjacent phases. In the case of a certain interval, only the corresponding assessment model is invoked to evaluate the operating performance. While, in the case of a fuzzy interval, both of the assessment models corresponding to the adjacent phases are used. The assessment results are adaptively integrated with the Bayesian inference-based posterior probabilities as the weights and used to evaluate the operating performance of the batch process, which is very challenging for traditional assessment methods. For the nonoptimal operating performance, the variable contributions are constructed and compared with the upper and lower confidence limits to determine whether it is responsible for the nonoptimality. Case study on the fed-batch penicillin fermentation process is used to demonstrate the effectiveness of the proposed method. As seen from the simulation results, the proposed online assessment framework is able to give an efficient operating performance assessment for multiphase batch processes. Furthermore, the responsible process variable identified by the variable contributions is consistent with the actual situation, which provides valuable reference for operators and managers to make further production adjustment.

Acknowledgments

We appreciate the financial support from National Natural Science Foundation of China (Nos. 61703078, 61673198, 61533007 and 61702070), China Postdoctoral Science Foundation (No.2017M611247), Fundamental Research Funds for the Central Universities (No.N160403002), Postdoctoral Science Foundation of Northeastern University (No.20170309), Hong Kong Research Grant Council on Synchronized Process Control of Hot-Runner Temperatures (No.670525433), the Funds for Creative Research Groups of China (No. 61621004), Stat Key Laboratory of Synthetical Automation for Process Industries Fundamental Research Funds (No. 2013ZCX02-04).

Appendix A: F-straight method based on Mahalanobis distance [50].

For an random matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T \in R^{N\times J}$, if the samples follow a Gaussian distribution $\boldsymbol{X} \sim \mathcal{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the population mean and covariance matrix respectively, the squared Mahalanobis distances of $\boldsymbol{x}_n, n = 1, 2, , ..., N$ follow χ^2 distribution, i.e.,

$$D(\boldsymbol{x}_n, \boldsymbol{\mu}) = (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \sim \chi^2(J). \quad \text{(A 1)}$$

When the population mean and population covariance matrix are unknown, the estimated values are usually used as alternative. In this case, the squared Mahalanobis distance follows the F distribution

$$D(\boldsymbol{x}_n, \boldsymbol{\mu}) = (\boldsymbol{x}_n - \bar{\boldsymbol{\mu}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}) \sim \frac{J(N^2-1)}{N(N-J)} F(J, N-J), \quad \text{(A 2)}$$

where $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ are the estimated values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. When $\boldsymbol{S}$ is singular and $rank(\boldsymbol{S}) = p < J$, the pseudo inverse of $\boldsymbol{S}$ is used by Mardia [44] and Eq. (A 2) becomes

$$D(\boldsymbol{x}_n, \boldsymbol{\mu}) = (\boldsymbol{x}_n - \bar{\boldsymbol{\mu}})^T \boldsymbol{S}^{+} (\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}) \sim \frac{p(N^2-1)}{N(N-p)} F(p, N-p), \quad \text{(A 3)}$$

where $\boldsymbol{S}^{+}$ is the Moore-Penrose pseudo inverse of the covariance matrix $\boldsymbol{S}$. It has been further proven that D is independent to the form of the pseudo inverse [44].

Since the distribution function can be replaced by its empirical distribution functions, the multivariate normality test can be transformed into the problem that whether the empirical distribution function of statistic D is equivalent to a specific F distribution. Firstly, the order statistics of the squared Mahalanobis distance $D_n(n=1,2,...,N)$ are defined as

$$D_n : D_{(1)} \le D_{(2)} \le ... \le D_{(N)}. \tag{A 4}$$

Then the empirical distribution function of D_n is described as

$$F_N(D_{(n)}) = \frac{t-0.5}{N} = r_n (n=1,2,...,N). \tag{A 5}$$

In addition, the quantile of F distribution corresponding to probability r_n is given by

$$F_n = \begin{cases} \dfrac{J(N^2-1)}{N(N-J)} F_{r_n}(J, N-J), rank(\boldsymbol{S}) = J \\ \dfrac{p(N^2-1)}{N(N-p)} F_{r_n}(p, N-p), rank(\boldsymbol{S}) = p \end{cases}. \tag{A 6}$$

If hypothesis test H_0 is valid, $D_{(n)} \approx F_n$. Thus, the plot of $(D_{(n)}, F_n)$ should scatter on a line passing through the original point with the slope of 1. Let the linear regression equation between $D_{(n)}$ and F_n be $F_n = a + bD_{(n)}$, where $a = \overline{F} - b\overline{D}$, $b = \sum_{n=1}^{N}(D_{(n)} - \overline{D})(F_n - \overline{F}) \Big/ \sum_{n=1}^{N}(D_{(n)} - \overline{D})^2$, $\overline{D} = \sum_{n=1}^{N} D_{(n)} \Big/ N$ and $\overline{F} = \sum_{n=1}^{N} F_n \Big/ N$. To test whether the linear regression equation reflects the linear relation between $D_{(n)}$ and F_n, perform the significance test on it, and the regression standard deviation s, $s = \sqrt{\sum_{n=1}^{N}(F_n - (a + bD_{(n)}))^2 \Big/ (N-2)}$, is used. If $s/\overline{F} > 0.15$, the significance test is invalid, and reject the null hypothesis. Otherwise, the significance test is valid and the linear regression equation can reflect the real relations between $D_{(n)}$ and F_n. Furthermore, compare the intercept a with 0 and regression coefficient b with 1. It is usually considered that the regression line passes through the original point with the

slope of 1 as long as the following condition is satisfied

$$\begin{cases} |a| < \sigma \\ |b-1| < \beta \end{cases}, \quad \text{(A 7)}$$

where σ and β are the thresholds and usually set as $\bar{F} \times 5\%$ and 0.2, respectively. If the condition in Eq. (A 7) is satisfied, the hypothesis test that the data follows a Gaussian distribution should not be rejected. Otherwise, reject the null hypothesis.

References

(1) Kourti, T.; MacGregor, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 3.

(2) Rendall, R. R.; Lu, B.; Castillo, I.; Chin, S. T.; Chiang, L. H.; Reis, M. S. A unifying and integrated framework for feature oriented analysis of batch processes. *Ind. Eng. Chem. Res.* **2017**, *56*, 8590.

(3) Louwerse, D. J.; Smilde, A. K. Multivariate statistical process control of batch processes based on three-way models. *Chem. Eng. Sci.* **2000**, *55*, 1225.

(4) Meng, X.; Morris, A. J.; Martin, E. B. On-line monitoring of batch processes using a PARAFAC representation. *J. Chemom.* **2010**, *17*, 65.

(5) He, Q. P.; Wang, J. Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. *AIChE J.* **2015**, *57*, 107.

(6) Nomikos, P.; MacGregor, J. F. Monitoring batch processes using multiway principal component analysis. *AIChE J.* **1994**, *40*, 1361.

(7) Nomikos, P.; MacGregor, J. F. Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97.

(8) Nomikos, P.; MacGregor, J. F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41.

(9) Kourti, T.; Nomikos, P.; MacGregor, J. F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control* **1995**, *5*, 277.

(10) Chen, J.; Liu, K. C. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chem. Eng. Sci.* **2002**, *57*, 63.

(11) Sang, W. C.; Morris, J.; Lee, I. B. Dynamic model-based batch process monitoring. *Chem. Eng. Sci.* **2008**, *63*, 622.

(12) Lee, J. M.; Yoo, C.; Lee, I. B. Fault detection of batch processes using multiway kernel principal component analysis. *Comput. Chem. Eng*. **2004**, *28*, 1837.

(13) Jia, M.; Chu, F.; Wang, F.; Wang, W. On-line batch process monitoring using batch dynamic kernel principal component analysis. *Chemom. Intell. Lab. Syst.* **2010**, *101*, 110.

(14) Yoo, C. K.; Lee, J. M.; Vanrolleghem, P. A.; Lee, I. B. On-line monitoring of batch processes using multiway independent component analysis. *Chemom.*

Intell. Lab. Syst. **2004**, *71*, 151.

(15) Rashid, M. M.; Jie, Y. Nonlinear and non-Gaussian dynamic batch process monitoring using a new multiway kernel independent component analysis and multidimensional mutual information based dissimilarity approach. *Ind. Eng. Chem. Res.* **2012**, *51*, 10910.

(16) Rato, T. J.; Blue, J.; Pinaton, J.; Reis, M. S. Translation-invariant multiscale energy-based PCA for monitoring batch processes in semiconductor manufacturing. *IEEE Trans. Autom. Sci. Eng.* **2017**, *14*, 894.

(17) Undey, C.; Cinar, A. Statistical monitoring of multistage, multiphase batch processes. *IEEE Contr. Syst. Mag*. **2002**, *22*, 40.

(18) Lu, N.; Gao, F.; Wang, F. Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J.* **2004**, *50*, 255.

(19) Zhang, S.; Zhao, C.; Gao, F. Two-directional concurrent strategy of mode identification and sequential phase division for multimode and multiphase batch process monitoring with uneven lengths. *Chem. Eng. Sci.* **2018**, *178*, 104.

(20) Zhao, C.; Sun, Y. Step-wise sequential phase partition (SSPP) algorithm based statistical modeling and online process monitoring. *Chemom. Intell. Lab. Syst.* **2013**, *125*, 109.

(21) Zhao, C. H.; Wang, F. L.; Lu, N. Y.; Jia, M. X. Stage-based soft-transition multiple PCA modeling and on-line monitoring strategy for batch processes. *J. Process Control* **2007**, *17*, 728.

(22) Zhao, C. Phase analysis and statistical modeling with limited batches for multimode and multiphase process monitoring. *J. Process Control* **2014**, *24*, 856.

(23) Ng, Y. S.; Srinivasan, R. An adjoined multi-model approach for monitoring batch and transient operations. *Comput. Chem. Eng.* **2009**, *33*, 887.

(24) Ye, L. B.; Liu, Y. M.; Fei, Z. S.; Liang, J. Online probabilistic assessment of operating performance based on safety and optimality indices for multimode industrial processes. *Ind. Eng. Chem. Res*. **2009**, *48*, 10912.

(25) Wang, J.; He, Q. P. Multivariate statistical process monitoring based on statistics pattern analysis. *Ind. Eng. Chem. Res.* **2010**, *49*, 7858.

(26) Zhao, C.; Huang, B. A full-condition monitoring method for nonstationary dynamic chemical processes with cointegration and slow feature analysis. *AIChE J.* **2017**. doi.org/10.1002/aic.16048

(27) Wang, Y.; Zhao, C. Probabilistic fault diagnosis method based on the combination of nest-loop fisher discriminant analysis and analysis of relative changes. *Control Eng. Pract*. **2017**, *68*, 32.

(28) Liu, Y.; Wang, F. L.; Chang, Y. Q. Online fuzzy assessment of operating performance and cause identification of non-optimal grades for industrial processes. *Ind. Eng. Chem. Res.* **2013**, *52*, 18022.

(29) Liu, Y.; Chang, Y. Q.; Wang, F. L. Online process operating performance assessment and nonoptimal cause identification for industrial processes. *J. Process Control* **2014**, *24*, 1548.

(30) Liu, Y.; Wang, F. L.; Chang, Y. Q.; Ma, R. C. Comprehensive economic index prediction based operating optimality assessment and nonoptimal cause identification for multimode processes. *Chem. Eng. Res. Des.* **2015**, *97*, 77.
(31) Qin, Y.; Zhao, C. H.; Wang, X. Z.; Gao, F. R. Subspace decomposition and critical phase selection based cumulative quality analysis for multiphase batch processes. *Chem. Eng. Sci.* **2017**, *166*, 130.
(32) Corbett, B. ; Mhaskar, P. Subspace identification for data-driven modeling and quality control of batch processes. *AIChE J.* **2016**, *62*, 1581.
(33) Lu, N.; Gao, F. Stage-based process analysis and quality prediction for batch processes. *Ind. Eng. Chem. Res.* **2005**, *44*, 3547.
(34) Yoo, C. K.; Villez, K.; Lee, I. B.; Rosén, C.; Vanrolleghem, P. A. Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnol. Bioeng*. **2007**, *96*, 687.
(35) Yu, J.; Qin, S. J. Multiway Gaussian mixture model based multiphase batch process monitoring. *Ind. Eng. Chem. Res.* **2009**, *48*, 8585.
(36) Liu, Y.; Wang, F. L.; Chang, Y. Q.; Ma, R. C.; Zhang, S. M. Multiple hypotheses testing-based operating optimality assessment and nonoptimal cause identification for multiphase uneven-length batch processes. *Ind. Eng. Chem. Res.* **2016**, *55*, 6133.
(37) Williams, C. K.; Rasmussen, C. E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, 2006.
(38) Young, D. S.; An overview of mixture models. *Stat. Survey* **2008**, *0*, 1.
(39) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, 1995.
(40) Figueiredo, M. A. T.; Jain, A. K. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2002**, *24*, 381.
(41) Zhang, S.; Zhao, C. Stationarity test and Bayesian monitoring strategy for fault detection in nonlinear multimode processes. *Chemom. Intell. Lab. Syst.* **2017**, *168*, 45.
(42) Wold, S.; Kettaneh, N.; Friden, H.; Holmberg, A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331.
(43) Lu, N. Y.; Gao, F. R.; Yang, Y. H.; Wang, F. PCA-based modeling and on-line monitoring strategy for uneven-length batch processes. *Ind. Eng. Chem. Res.* **2004**, *43*, 3343.
(44) Mardia, K. V. Mahalanobis distances and angles. *Multivariate Anal. IV.* **1977**, 495.
(45) Qin, S. J.; Valle, S.; Piovoso, M. J. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemom.* **2001**, *15*, 715.
(46) Xuan, J.; Xu, Z.; Sun, Y. Incipient sensor fault diagnosis based on average residual-difference reconstruction contribution plot. *Ind. Eng. Chem. Res*. **2014**, *53*, 7706.
(47) Khosrow, D. Density estimation for statistics and data analysis. *Technometrics* **1992**, *29*, 495.

(48) Birol, G.; Ündey, C.; Çinar, A. A modular simulation package for fed-batch fermentation: Penicillin production. *Comput. Chem. Eng*. **2002**, *26*, 1553.

(49) Birol, G.; Ü ndey, C.; Parulekar, S. J.; Çinar, A. A morphologically structured model for penicillin production. *Biotechnol. Bioeng*. **2002**, *77*, 538.

(50) Zhang, S.; Wang, F.; Zhao, L.; Wang, S.; Chang, Y. A novel strategy of the data characteristics test for selecting a process monitoring method automatically. *Ind. Eng. Chem. Res.* **2016**, *55*, 1642.

Abstract graphic