Final quality prediction for multi-phase batch process based on phase cumulative product quality model

Xiaochu Tang, Yuan Li, Jinyu Guo and Zhi Xie

Transactions of the Institute of Measurement and Control published online 26 February 2014
DOI: 10.1177/0142331213501688

Article

TRANSACTIONS OF THE INSTITUTE OF MEASUREMENT & CONTROL

Transactions of the Institute of Measurement and Control
1–13

DOI: 10.1177/0142331213501688
tim.sagepub.com
SAGE

Final quality prediction for multi-phase batch process based on phase cumulative product quality model

Xiaochu Tang[1], Yuan Li[2], Jinyu Guo[2] and Zhi Xie[1]

Abstract

A novel online final product quality prediction scheme is proposed in this paper for the improvement of quality prediction in multi-phase batch processes. Phase cumulative product quality (PCPQ), which is quality cumulated from the beginning of the phase to the end, is introduced for quality prediction, and final product quality prediction offline is achieved by cumulating all the predicted PCPQ based on the corresponding PCPQ model. In this way, the quality prediction approach proposed not only explores the different effects of process variables in different phases on final product quality, but also takes the common effects of process variables in different phases into account. The PCPQ model and remained phase cumulative product quality (RPCPQ) model are combined to improve online prediction precision, without the missing observations estimation. The proposed approach is applied to a simulated penicillin fermentation process and the results of simulation demonstrate effectiveness and superiority.

Keywords

Common effect, final product quality prediction, phase cumulative product quality (PCPQ), remained phase cumulative product quality (RPCPQ)

Introduction

Batch or semi-batch processes have been widely applied to chemical, biological, pharmaceutical, injection moulding and semiconductor industries for manufacturing high value-added and low volume products (Kosanovich et al., 1996; Undey et al., 1999, 2000; Westerhuis and Coenegracht, 1997; Wise et al., 1999). In order to meet larger-scale commercial production, developing related investigations for batch processes has received increasing attention. Characterized by a prescribed processing of raw materials for a finite duration to convert them to products, a successful batch process means tracking this prescribed recipe with a high degree of reproducibility to produce consistent quality products. Due to raw material purities, variations in initial conditions and disturbances, batch processes suffer from a lack of reproducibility from batch-to-batch variations. Final product quality is usually available after completing the whole batch, which does not enable operators to take immediate and corrective actions to prevent degraded product quality. Therefore, it is necessary to develop an effective and accurate online quality prediction approach for improving product quality (Facco et al., 2009; Gunther et al., 2009; Lu and Gao, 2005; Undey et al., 2004).

Multivariate statistical process control (MSPC), such as principal component analysis (PCA) and partial least squares (PLS), has been regarded as an effective tool for statistical process monitoring (SPM), and fault detection and diagnosis (FDD) (Chen et al., 1996; Chiang et al., 2000; Geladi and Kowalshi, 1986; Jackson, 1991; Kano et al., 2001; Zhou et al., 2011). Multi-way principal component analysis (MPCA) and multi-way partial least squares (MPLS) modelling methods, pioneered by Nomikos and MacGregor, have been developed and successfully applied to batch processes for process monitoring and quality prediction (Nomikos and MacGregor, 1994, 1995). MPLS explains variations of process variables about their average trajectory at each point of time, and relates the quality variables with the process variables by extracting information on process variables that are most predictive on product quality. Although MPLS is regarded as an effective quality prediction approach, a major limitation of this method is that the future portions of process variable trajectories with respect to the current time for a new batch must be estimated using a filling-in-method (Ramaker et al., 2005). PARAFAC and Tucker models partitioned the total run time of a batch with respect to some scheduling points, which overcome data estimation as an another alternative approach (Louwerse and Smilde, 2000).

In industries, some batch processes include a single step, whereas many others are carried out in a sequence of steps, which are called multi-phase or multi-stage batch processes (Undey and Cinar, 2002). Conventional global MPLS modelling can be inevitably compromised for quality prediction of

[1]College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, China
[2]College of Information Engineering, Shenyang University of Chemical Technology, Shenyang, Liaoning Province, China

Corresponding author:
Yuan Li, College of Information Engineering, Shenyang University of Chemical Technology, Shenyang, Liaoning Province, 110042, China.
Email: li-yuan@mail.tsinghua.edu.cn.

multi-phase batch processes. This is because that global MPLS uses the whole process trajectory with final product quality to build a single model; it does not consider the multiplicity of phases in a multi-phase batch process. So, a stage-based process analysis and quality prediction strategy has been developed by Lu and Gao (2005) and Zhao et al. (2008). In their methods, the critical-to-quality phases are identified and quality prediction models are developed focusing on final quality and the critical-to-quality phases. Although their methods consider the multiplicity of phase in batch processes and the different effects of process variables of different phases on final product quality, it is difficult to reflect on the common effects of different phases. It just makes use of a weight value to measure the different effects of different phase on final product quality, which overlooks the common effects of different phases. Therefore, it is necessary to develop a more effective method for multi-phase quality prediction, which not only considers the different effects of different phases on final product quality, but also the common effects of different phases.

In this paper, a new quality prediction method approach is proposed for a multi-phase batch process. The basic idea is that final product quality can be regarded as the cumulative sum of phase cumulative product quality (PCPQ). This idea is based on the fact that final product quality is the cumulative result of changeable quality value at each time instance. Moreover, from the modelling viewpoint, final product quality depends on the whole process trajectory, i.e. cumulative effects of process behaviours on quality. That is to say, in a multi-phase batch, the final quality should depend on the common effects of the process trajectory in each phase. The conventional stage-based quality prediction method uses phase process trajectories with final product quality for modelling to reflect on different effects of each phase on quality. However, it is difficult to identify process trajectories affecting quality in a specific period with final quality alone (Duchesne and MacGregor, 2000). Therefore, the PCPQ model, which relates phase process trajectories to corresponding PCPQ, is developed to achieve PCPQ prediction. Such a model helps isolate the local and different effect of phase process trajectories on final product quality. Then, all PCPQ is cumulated to achieve final product prediction. In this way, final product quality based on the PCPQ model considers the different effect of phase process trajectories on final product quality. At the same time, by cumulating all the PCPQ, the common effect of all phases on the final product quality is also considered. The proposed prediction method can overcome the shortcomings of the conventional prediction method, which only considers the common effect or different effect of process trajectories on final product quality, and it has the following advantage: by introducing PCPQ, it allows final product quality prediction considering both common and different effects of process trajectories on quality. Meanwhile, the PCPQ model can well reflect on the inherent relation of process trajectories with quality. Furthermore, incorporating the PCPQ model with the remained phase cumulative product quality (RPCPQ) model achieves online prediction without data estimation. Therefore, it is more valuable for final product quality prediction for a multi-phase batch process.

The remainder of the article is organized as follows. The next section deduces the conventional MPLS for batch process data. Then, the modelling PCPQ and modelling RPCPQ method are given. The online quality prediction method proposed is given and application of the proposed method to penicillin fermentation demonstrates its effectiveness. Finally, the paper is concluded.

Multi-way partial least squares (MPLS) of batch process data

In a batch process, assume that historical dataset consists of $i = 1, 2 \cdots, I$ batch runs where each of them has $j = 1, 2 \cdots, J$ variables measured over $k = 1, 2, \cdots, K$ time instances. The dataset collected from the batch process can be organized into a three-way array $\tilde{X}(I \times J \times K)$ and batch-wise unfolding, which creates the matrix $X(I \times JK)$ that is widely used to analyse batch process data including K time slice $\tilde{X}_k(I \times J)$. The corresponding final product quality variables $(m = 1, 2, \cdots, M)$ are arranged in a $Y(I \times M)$ matrix, as shown in Figure 1. MPLS is an extension of PLS to handle data in three-dimensional arrays. Application of MPLS to batch processes is equivalent to performing PLS on 2-D unfolded matrix. In order to eliminate the influence of nonlinearity and different measuring scale, the matrix $X(I \times JK)$ and $Y(I \times M)$ are mean centred and scaled to unit variance. PLS can be performed as follows

$$X = TP^T + E \qquad Y = TQ^T + F \tag{1}$$

For batch data, PLS decomposes the $X(I \times JK)$ and $Y(I \times M)$ into score matrix $T(I \times R)$ and loading matrix $P(JK \times R)$, $Q(M \times R)$, plus residual matrices $E(I \times JK)$, $F(I \times M)$, where T is given by:

$$T = XW\left(P^T W\right)^{-1} \tag{2}$$

This decomposition summarizes and compresses high-dimensional process trajectories into low-dimensional space, which is most relevant to the final product quality. Matrix T

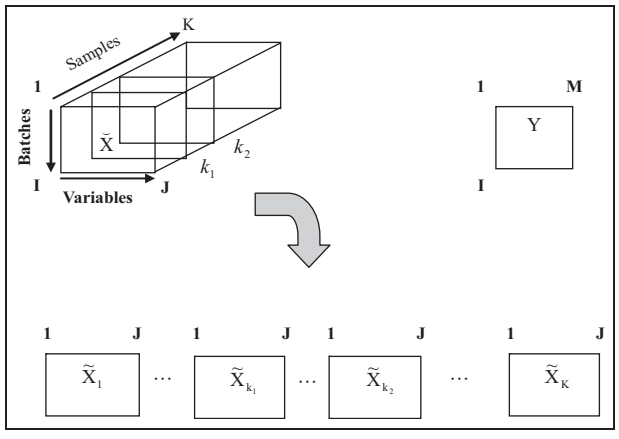

Figure 1. Unfolding the batch process data for the multi-way partial least squares (MPLS) modelling method.

represents the overall variability of each batch with respect to the other batches. Matrices P and W summarize the time variation of process variables about their average trajectories. Matrix Q relates the variability of the process variables to the final product quality.

Conventional MPLS use the whole process trajectories as input to pick up process variations of process variables to the final product quality. This indicates an idea that final product quality is the cumulative result of process trajectories over the entire batch on quality. However, for a multi-phase batch process, the whole process trajectories may be divided into several phases according to their different inherent natures. Therefore, a single model based on conventional MPLS applied to a multi-phase batch process for final quality prediction is not suitable. Moreover, developing a method to overcome the shortcomings of MPLS, which ignores the multi-characteristic and learning from the advantages, which considers the global-characteristic, is desirable for multi-phase batch process quality prediction.

Modelling phase cumulative product quality (PCPQ)

To build a model, the three-way data $\tilde{X}(I \times J \times K)$ is batch-wise unfolded and forms a two-way matrix, as described in Figure 1. For a multi-phase batch process, the two-way matrix can be divided into several different data blocks including different time slices $\tilde{X}_k(I \times J)$, which represent the different phase process data $X_c(I \times K_c J)(c = 1, 2, \cdots, C)$, as shown in Figure 2, where K_c is the number of samplings belonging to the phase c, C is the number of phase during a batch. In this article, we assume that the phase division has been achieved and the length of each batch is equal.

Final product quality may be regarded as the cumulative sum of quality, which is generated at each time instance. That is to say, in a multi-phase batch process, we can consider final quality the cumulative sum of quality, which is generated during each phase. PCPQ ΔY_c is introduced, which represents quality accumulated from the beginning of phase to the end, in order to achieve the final product quality prediction. Furthermore, the final product quality depends on the whole process trajectory. So, process trajectories including significant information about product quality can be used to estimate the final product quality. Conventional final quality prediction for multi-phase batch process only related final quality with each phase process trajectories for building a sub-PLS model, as described in Figure 2. However, phase process trajectories affect product quality only over a period, not from the beginning of a batch to the end, and final product quality depends on the whole process trajectory, rather than specific phase process trajectories. Therefore, it is difficult to identify phase process trajectories with final product quality. PCPQ is just the result of phase process trajectories run, and the PCPQ model can help isolate the local effects of process trajectories on product quality. Moreover, final product quality prediction based on the cumulative sum of PCPQ incorporates the information about the common effects of process trajectories on product quality. This is because, by introducing PCPQ, final product quality prediction is performed in the condition that all phase process trajectories effect on final product quality in common. It is unlike the conventional method in which the sub-model is built with a separate or independent effect of the phase process trajectories on final quality.

A PCPQ model is built by relating each phase process data $X_c(I \times K_c J)$ to the corresponding PCPQ $\Delta Y_c(I \times M)$, as described in Figure 3. Here, the phase average trajectory of process variables $\bar{X}_c(I \times J)$ is used as modelling input instead of fat unfolded 2-D array $X_c(I \times K_c J)$, which can be obtained as follows

$$\bar{X}_c(I \times J) = \frac{1}{K_c} \sum_{k \in c} \tilde{X}_k(I \times J) \tag{3}$$

Then the PLS algorithm is performed on $\bar{X}_c(I \times J)$ and ΔY_c at each phase. It can be summarized as follows

$$\bar{X}_c(I \times J) = T_c P_c^T + E_c \quad \Delta Y_c(I \times M) = T_c Q_c^T + F_c \tag{4}$$

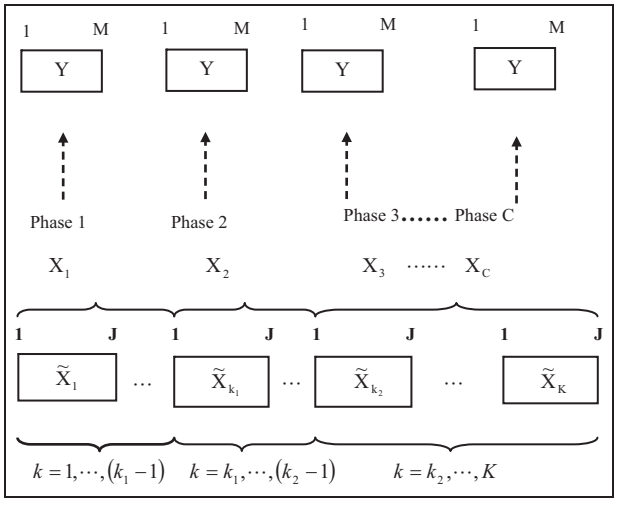

Figure 2. Conventional multi-phase batch process modelling method.

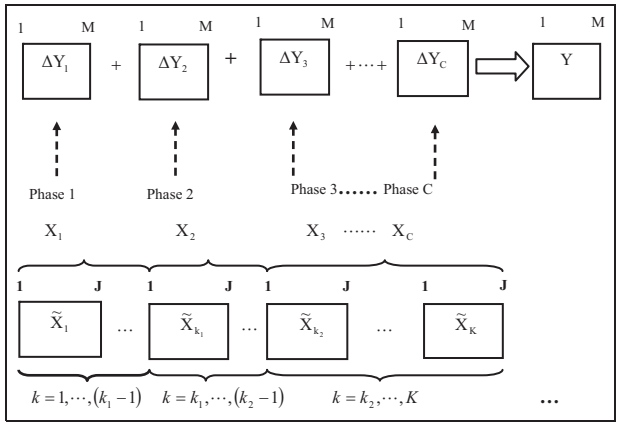

Figure 3. Phase cumulative product quality (PCPQ) modelling method for multi-phase batch process.

where $P_c(J \times A_c)$ and $Q_c(M \times A_c)$ are the loading matrices for $\bar{X}_c(I \times J)$ and $\Delta Y_c(I \times M)$, respectively. A_c is the number of latent variables. E_c and F_c are the residual matrices for $\bar{X}_c(I \times J)$ and $\Delta Y_c(I \times M)$, respectively. $T_c(I \times A_c)$ is the score matrix, which can be computed by the equation

$$T_c(I \times A_c) = \bar{X}_c W_c \left(P_c^T W_c\right)^{-1} \tag{5}$$

The predicted PCPQ $\Delta \hat{Y}_c(I \times M)$ for modelling data can be deduced as

$$\Delta \hat{Y}_c = T_c Q_c^T = \bar{X}_c W_c \left(P_c^T W_c\right)^{-1} Q_c^T \tag{6}$$

The final product quality prediction offline, according to the proposed method, can be performed as follows

$$\hat{Y} = \sum_{c=1}^{C} \Delta \hat{Y}_c \tag{7}$$

where $\hat{Y}$ is predicted final product quality. In conclusion, the final product quality prediction offline based on the proposed method can be concluded two steps: first, the new PCPQ is predicted by conducting PLS on phase process trajectories with PCPQ. Then, the final product quality is achieved by obtaining the sum of each PCPQ.

Modelling remained phase cumulative product quality (RPCPQ)

In order to achieve online final product quality prediction, RPCPQ prediction is introduced to avoid future data estimation. Here, remained phase is defined as the phase that is running and does not run. So process data of the remained phase corresponding phase c is denoted as $\widehat{X}(I \times K_n J)$, which includes all the phase from phase c to the end. Here K_n is the number of time slice belonging to remained phase. Correspondingly, RPCPQ $\Delta \widehat{Y}_c(I \times M)$ is the quality accumulated from phase c to the end. The relation between X_c and $\widehat{X}_c$ can be described in Figure 4. The relation between $\Delta \widehat{Y}_c(I \times M)$ and $\Delta Y_c(I \times M)$ is given by

$$\Delta \widehat{Y}_c = Y - \sum_{c=1}^{c-1} \Delta Y_c \tag{8}$$

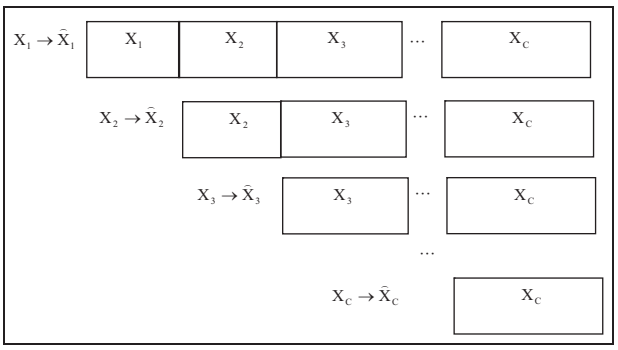

Figure 4. The corresponding relation between X_c and $\widehat{X}_c$.

Modelling RPCPQ is equal to performing PLS on remained phase process trajectories $\widehat{X}_c(I \times K_n J)$ and corresponding RPCPQ $\Delta \widehat{Y}_c(I \times M)$. During the course of a batch, the whole process trajectories of $\widehat{X}_c(I \times K_n J)$ are not available, and just the evolving process trajectories at running phase are available. In addition, with the development of the batch, the evolving process trajectories become fat so that the PLS algorithm is a difficult task.

In this paper, evolving average process trajectories of the remained phase, denoted as $\bar{\widehat{X}}_{cn}(I \times J)$, are used to predict RPCPQ. In this way, the above-mentioned shortcomings can be overcome. The $\bar{\widehat{X}}_{cn}$ can be given as follows

$$\bar{\widehat{X}}_{cn} = \frac{1}{n} \sum_{k=1}^{n} \tilde{X}_k \tag{9}$$

where n is the number of the evolving time slice, which belongs to the remained phase. When a phase is over, the value of n is equal to the K_c. In fact, PCPQ the prediction method is similar to multi-block quality prediction. That is to say, modelling PCPQ is finally conducted between the evolving process data of the running phase and the corresponding $\Delta \widehat{Y}_c(I \times M)$. This model assumes that all control moves in the remaining phase are nominal and follow those in the training data. The detailed modelling RPCPQ strategy is described in Figure 5. An important feature of this approach is that there is no need to estimate the missing data. Therefore, it reduces the uncertainty from data imputation. The PLS algorithm performed on $\left\{\bar{\widehat{X}}_{cn}(I \times J), \Delta \widehat{Y}_c(I \times M)\right\}$ is formulated as

$$\bar{\widehat{X}}_{cn}(I \times J) = \widehat{T}_c \widehat{P}_c^T + \widehat{E}_c \quad \Delta \widehat{Y}_c(I \times M) = \widehat{T}_c \widehat{Q}_c^T + \widehat{F}_c \tag{10}$$

where $\widehat{P}_c\left(J \times \widehat{A}_c\right)$ and $\widehat{Q}_c\left(M \times \widehat{A}_c\right)$ are the loading matrices for $\bar{\widehat{X}}_{cn}(I \times J)$ and $\Delta \widehat{Y}_c(I \times M)$, respectively. $\widehat{A}_c$ is the number of latent variables for RPCPQ model. $\widehat{T}_c\left(I \times \widehat{A}_c\right)$ is the score matrix, which can be computed by the equation

$$\widehat{T}_c\left(I \times \widehat{A}_c\right) = \bar{\widehat{X}}_{cn} \widehat{W}_c \left(\widehat{P}_c^T \widehat{W}_c\right)^{-1} \tag{11}$$

In the same way, the predicted RPCPQ $\Delta \hat{\widehat{Y}}_c(I \times M)$ for modelling the data can be deduced as

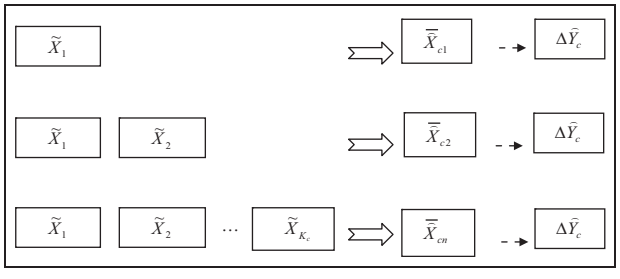

Figure 5. Remained phase cumulative product quality (RPCPQ) modelling method.

$$\Delta\hat{\hat{Y}}_c = \hat{\bar{T}}_c\hat{Q}_c^T = \hat{\bar{X}}_{cn}\hat{W}_c\left(\hat{P}_c^T\hat{W}_c\right)^{-1}\hat{Q}_c^T \quad (12)$$

where $\hat{W}_c$ is weighting matrix.

Online final product quality prediction

According to the idea of the proposed method, the final product quality is equal to the sum of all PCPQ. When online final product quality prediction is performed, the final product quality can be further regarded as the sum of PCPQ and RPCPQ. As above, by Equation (8), the predicted final quality can be deduced as

$$\hat{Y} = \sum_{c=1}^{c-1}\Delta\hat{Y}_c + \Delta\hat{\hat{Y}}_c \quad (13)$$

From Equation (13), we can see that the final product quality prediction include two parts. One part is PCPQ prediction and the other is RPCPQ prediction. For a new batch obtained at each time, x^{new} can be regarded as two parts of the process trajectories. One part is the process trajectories available of completed phases, and each one is denoted as x_c^{new}. The other part is the process trajectories available of uncompleted phases, and it is denoted as $\hat{x}_{cn}^{new}$. According to Equations (3) and (9), for a new batch, we can obtain phase average process trajectories, denoted as $\bar{x}_c^{new}(1\times J)$, and evolving average process trajectories of the remained phase, denoted as $\hat{\bar{x}}_{cn}^{new}(1\times J)$. The predicted PCPQ $\Delta\hat{Y}_c(I\times M)$ can be given by

$$\Delta\hat{Y}_c = \bar{x}_c^{new}W_c\left(P_c^TW_c\right)^{-1}Q_c^T \quad (14)$$

The predicted RPCPQ $\Delta\hat{\hat{Y}}_c(I\times M)$ can be obtained by

$$\Delta\hat{\hat{Y}}_c = \hat{\bar{x}}_{cn}^{new}\hat{W}_c\left(\hat{P}_c^T\hat{W}_c\right)^{-1}\hat{Q}_c^T \quad (15)$$

After predicted PCPQ and predicted RPCPQ are obtained, the final product quality can be obtained by Equation (10). During the course of a batch, online final product quality prediction is achieved by combing the PCPQ model with the RPCPQ model; the real-time prediction scheme is described in Figure 6.

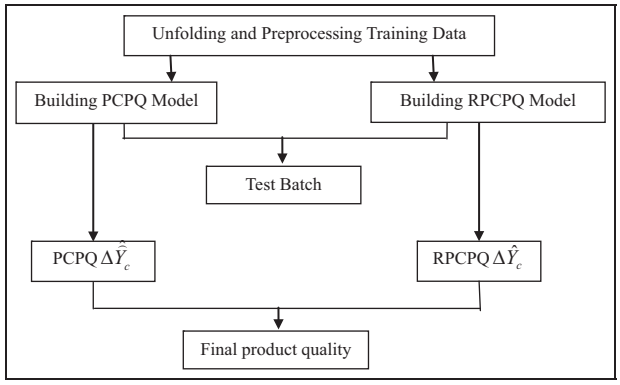

Figure 6. Flow diagram of the proposed online quality prediction method.

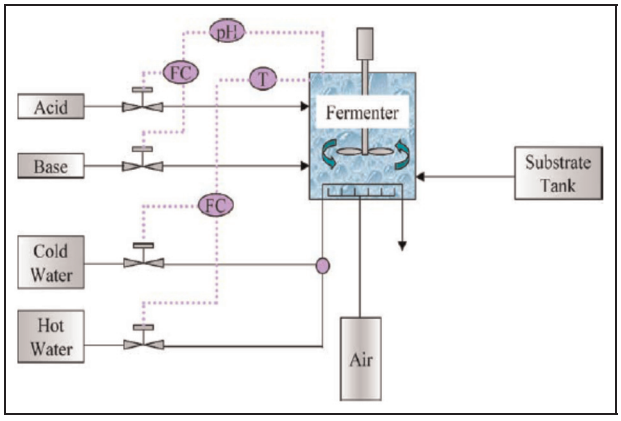

Figure 7. Flow diagram of the simulated penicillin fermentation process.

Simulation

Process description and data generation

A fed-batch penicillin fermentation process is used to demonstrate the validity and effectiveness of the proposed online final product quality prediction method. The process simulator has been developed by Birol et al. (2002) to provide a test platform for research and application of batch process monitoring and quality prediction methods (Birol et al., 2002; Lee et al., 2004a, b). The simulator is based on a mechanistic model proposed by Bajpai and Reuss (1980) and is available at: http://www.chbe.iit.edu/~cinar/. The flow diagram of penicillin fermentation process is shown in Figure 7. The process consists of two operational phases including a biomass growth phase and a penicillin production phase. To promote the amount of penicillin production, the first phase is carried out in a batch operation mode. After about 45 h, the process is switched to fed-batch operation mode with glucose being added continuously. The entire duration of each batch is 400 h with a sampling interval of 0.5 h. All batches are assumed to be of the same duration. The process variables and quality variables selected for modelling are provided in Table 1. During simulations, two proportional, integral and derivative (PID) controllers are adopted to control the pH and temperature by adjusting acid/base and cold/hot water flow rates. The substrate of glucose is continuously fed into the fermenter under an open-loop operation. In a real industrial setting, biomass and penicillin concentration are analysed offline by quality analysis experiments after a batch is completed. Online quality prediction provides a way of obtaining product quality during the process, making it possible to correct problems to ensure acceptable product quality. So, in this paper, biomass and penicillin concentration is selected as the quality variable for modelling and prediction.

In the simulation illustration, 48 normal batch runs are generated with slight variations under various operation conditions by the design of experiment (DOE) method. The

Table 1. Process variables and quality variables for penicillin modelling process.

No.	Process variables
1	Aeration rate
2	Agitator power
3	Substrate feed rate
4	Substrate feed temperature
5	Dissolved oxygen concentration
6	Culture volume
7	Carbon dioxide concentration
8	pH
9	Temperature
10	Generated heat
No.	**Quality variables**
1	Biomass concentration
2	Penicillin concentration

Table 2. Operation conditions for training data by design of experiment (DOE).

No.	pH	Temperature (K)
1	4.95	297
2	4.95	298
3	5.05	297
4	5.05	298
5	5	297
6	5	298

operating conditions are set as shown in Table 2. The corresponding PCPQ, which are biomass concentration and penicillin concentration, can be collected at 45 and 400 h. Then, 12 test batches with different normal stochastic changes in operation conditions from those training batches are generated for the validation of effectiveness of the proposed prediction method.

Illustration of simulation results

For multi-phase batch process modelling, phase division should be performed first. In this paper, the batch is divided into two phases according to process prior knowledge. One is phase from (0, 45 h], and the other is (45 h, 400 h]. With the training data set of 48 normal operation batches, PCPQ model is built for phase 1 and phase 2, respectively.

To illustrate the performance of offline prediction based on the proposed PCPQ modelling method, the final predicted product quality for training batches is shown in Figure 8. In Figures 8(a) and (b), a comparison of offline prediction for training data using the proposed method and Zhao's method is given. Moreover, to illustrate the offline prediction performance more clearly, plots of prediction error rate are given in Figure 8(c).

From Figures 8(a) and (b), we can see that the proposed PCPQ model shows the superiority for fitting training batches, where the predicted quality value fits well with the actual quality measurement, compared with Zhao's method. Moreover, from Figure 8(c), it can be seen that quality prediction based on the proposed method has a lower error rate, compared with Zhao's method. Table 3 summarized the prediction error rate in detail. It shows the proposed method has a lower average and maximum error rate. Therefore, offline quality prediction based on proposed method is superior.

As mentioned previously, the final product quality depends on all the critical process phases rather than each separate phase. PCPQ is introduced for modelling not only taking into account different effects of different phases on final quality but also considering the common cumulative effects of all the critical phases on final quality. This is just the reason that the proposed method for offline quality prediction obtained better prediction results. According to Zhao's method, the model is built between each phase and final quality, and final product quality prediction results are obtained by weighting each final quality at the end of each phase. It seems to be that the weight value is used to measure the percent ratio of effect of each phase. However, the weight value does not always reflect on the real effect very well. This could influence the prediction performance of the final product quality.

In addition, modelling based on the proposed method isolates the local effect of process variables on quality. The phase regression parameters plots for biomass concentration and penicillin concentration are given in Figures 9 and 10, respectively, compared with Zhao's method. In Figure 9, plots (a) and (b) represent regression parameters of phase 1 and 2 based on proposed PCPQ modelling method, and plots (c) and (d) represent the repression parameters based on Zhao's method. From Figures 9(a) and (b), we can see that, for phase 1, variables 6, 8, 9 and 10 have higher regression parameters, and for phase 2, variables 3, 6, 8, 9 and 10 have higher regression parameters, indicating that these variables are quality-correlated variables. However, in Figures 9(c) and (d), variable 8 does not show higher regression parameters in phase 1 and 2. In fact, variable 8 (pH) is the critical process variable affecting the product quality in the whole course of batch. Furthermore, from Figures 9(a) and (b), it is illustrated that variable 5, 6 and 7 are negatively related to the quality, and variables 8, 9 and 10 are positively related to the quality in phase 1. In phase 2, variable 3 is positively related to the quality and variables 6, 8, 9 and 10 are negatively related to the quality. This explores the different relations of process variables with the predicted quality. In phase 1, which is biomass growth phase, higher pH, higher temperature and higher generated heat can promote the biomass growth, so this results in higher biomass concentration. Variable 5 (dissolved oxygen concentration) and variable 7 (carbon dioxide concentration) in phase 1 will decrease due to greater need for biomass, and at the same time, biomass concentration will increase. Then, in phase 2, which is the penicillin production phase, when the pH, temperature and heat is up to a certain value, biomass concentration will decrease as these variables increased. Correlation of process variables with quality variables agrees well with the real physical process. However, due to only considering final product quality, regression parameters based on Zhao's method may not reflect on the correlation of process variable with quality accurately. Therefore, introducing the

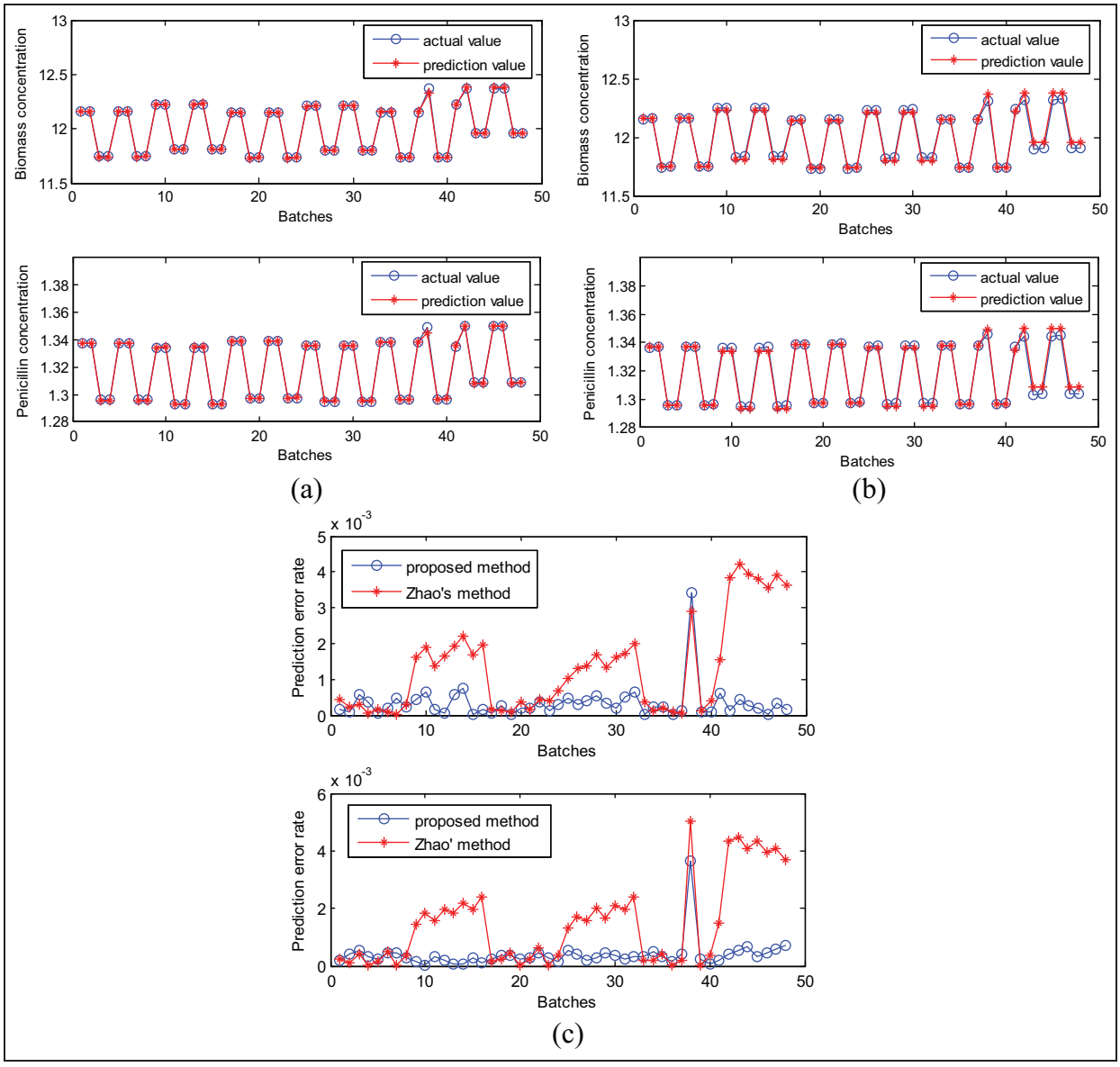

Figure 8. Offline quality prediction results for training data using: (a) the proposed method; (b) Zhao's method; (c) offline prediction error rate for biomass concentration and penicillin concentration, respectively.

Table 3. Average and maximum offline prediction error rate based on proposed method and Zhao's method.

Method	Offline prediction for biomass concentration		Offline prediction for penicillin concentration	
	Average error rate (%)	Maximum error rate (%)	Average error rate (%)	Maximum error rate (%)
Proposed method	0.04	0.36	0.03	0.34
Zhao's method	0.15	0.5	0.13	0.42

PCPQ for modelling quality prediction can reflect well on the correlation of process variable with quality – in this way improving the accuracy of quality prediction. In Figures 10(a) and (b), regression parameters for penicillin concentration based on the proposed method also give an agreeable result with the actual process. For penicillin concentration, depending on the biomass concentration, the correlation with process variables is the same as the biomass concentration. By the above analysis, we can see that the modelling based on the proposed method shows superiority.

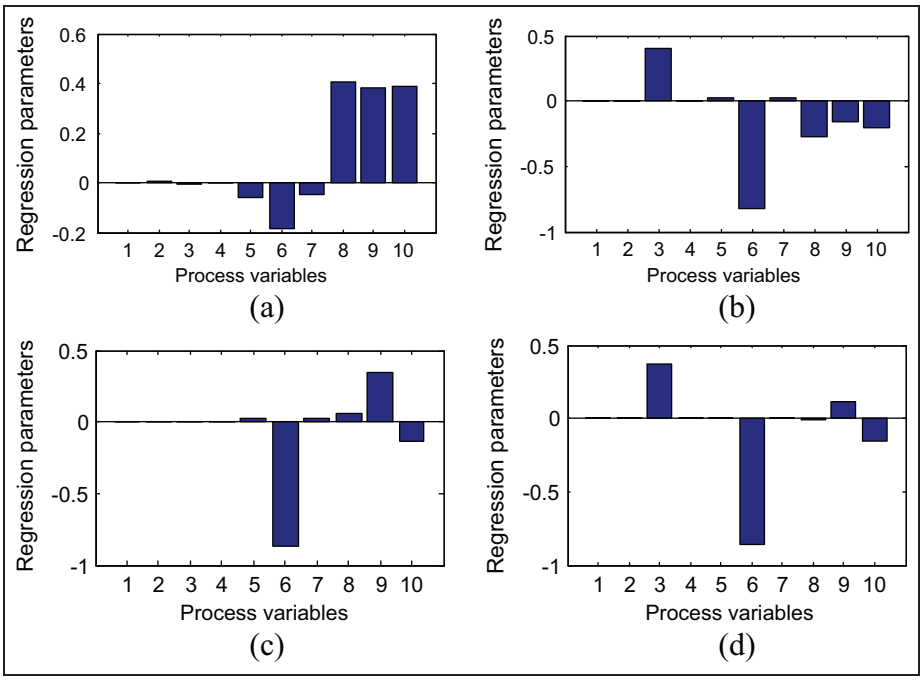

Figure 9. Regression parameters for biomass concentration based on the proposed method for phase 1 in (a) and phase 2 in (b); based on Zhao's method for phase 1 in (c) and for phase 2 in (d).

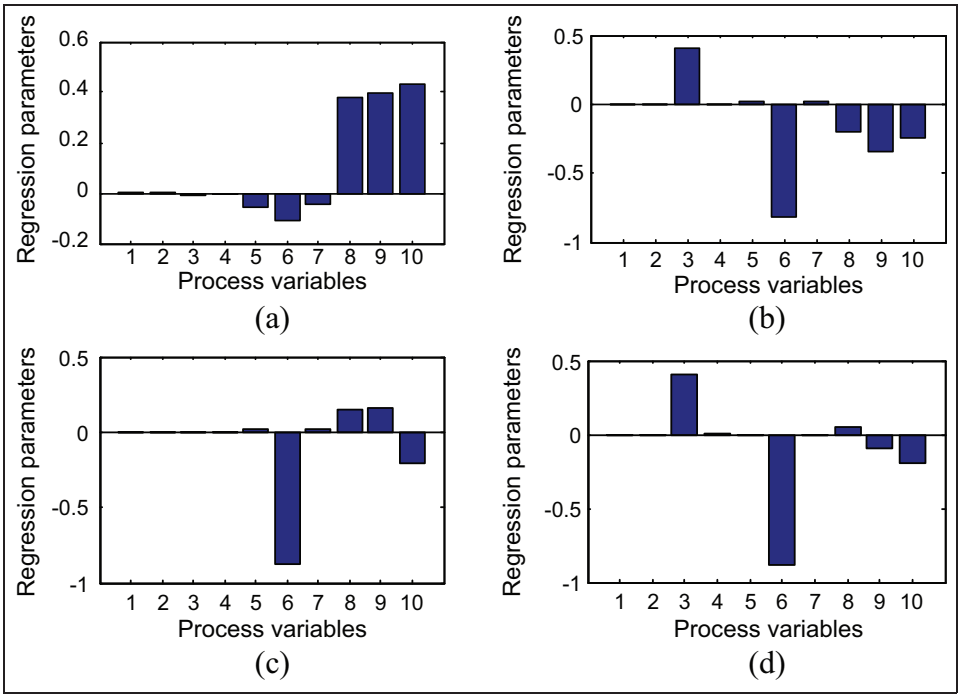

Figure 10. Regression parameters for penicillin concentration based on the proposed method in for phase 1 in (a) and for phase 2 in (b); based on Zhao's method for phase 1 in (c) and for phase 2 in (b).

To illustrate the performance of prediction, a comparison of offline prediction for 12 test batches, respectively, using proposed method and Zhao's method, is given in Figure 11. Although Zhao's method in Figures 11(c) and (d) obtains the desired prediction results, the proposed method in Figures 11(a) and (b) can show better performance of prediction in all the test batches. So, by comparison, the superiority of the proposed method for offline quality prediction is obvious for

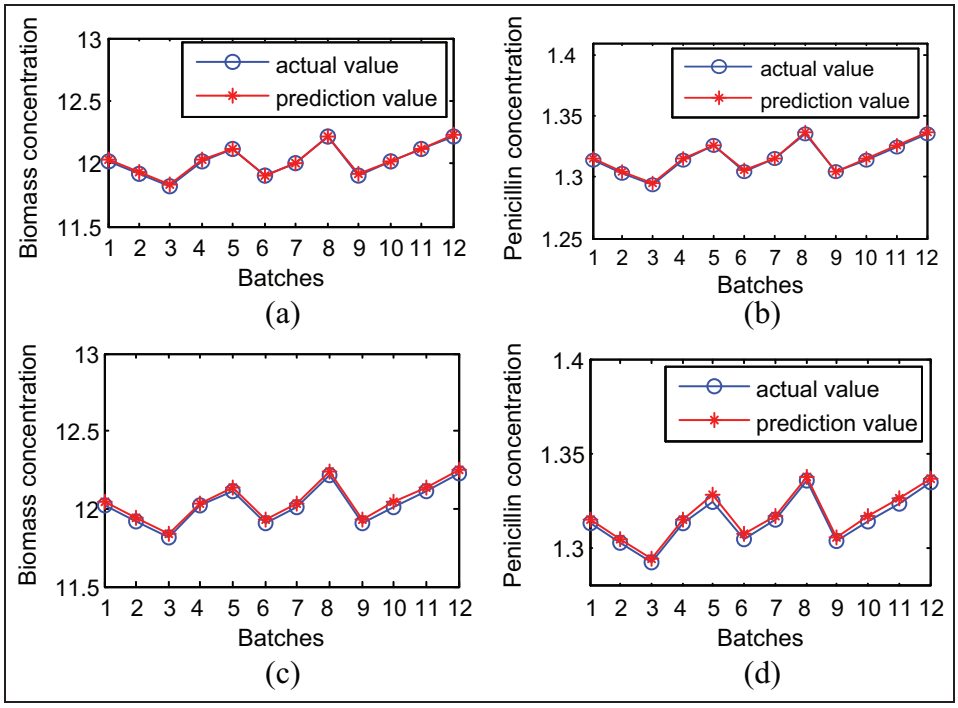

Figure 11. Offline quality prediction results for test data using the proposed method in (a) and (b); Zhao's method in (c) and in (d).

Table 4. Average and maximum online prediction error rate for biomass concentration and penicillin concentration.

Test batch	Online prediction for biomass concentration		Online prediction for penicillin concentration	
	Average error rate (%)	Maximum error rate (%)	Average error rate (%)	Maximum error rate (%)
1	0.18	0.38	0.21	0.40
2	0.22	0.77	0.23	1.16
3	0.24	1.62	0.25	1.93
4	0.17	0.37	0.19	0.39
5	0.08	0.46	0.12	0.41
6	0.11	0.87	0.15	1.02
7	0.06	0.37	0.11	0.34
8	0.11	0.40	0.13	0.37
9	0.14	0.82	0.16	1.09
10	0.11	0.44	0.15	0.38
11	0.12	0.34	0.13	0.31
12	0.16	0.42	0.17	0.37

both training and testing batches, which gives satisfying prediction results, demonstrating the model's fitness ability and prediction adaptability.

Online quality prediction is performed based on the proposed method by combining the PCPQ model and the RPCPQ model, without taking future data estimation into account. Figures 12 and 13 give online final quality prediction results of 12 test batches for biomass concentration and penicillin concentration, respectively. The average and maximum online prediction error rates for each test batch are listed in Table 4. From this table, we can see that the average prediction error rate of each test batch for biomass concentration is from 0.06% to 0.24%, and the maximum prediction error rate is less than 1.62%. For the penicillin concentration prediction, the average prediction error rate for each test batch is from 0.11% to 0.25%, and the maximum prediction error rate is less than 1.93%. The prediction precision is acceptable and desirable, which shows that the proposed online prediction method is effective and superior. Furthermore, in Figures 12 and 13, it is not difficult to find the online prediction result for all test batches is not the best at the beginning of a batch. This will be changeable with

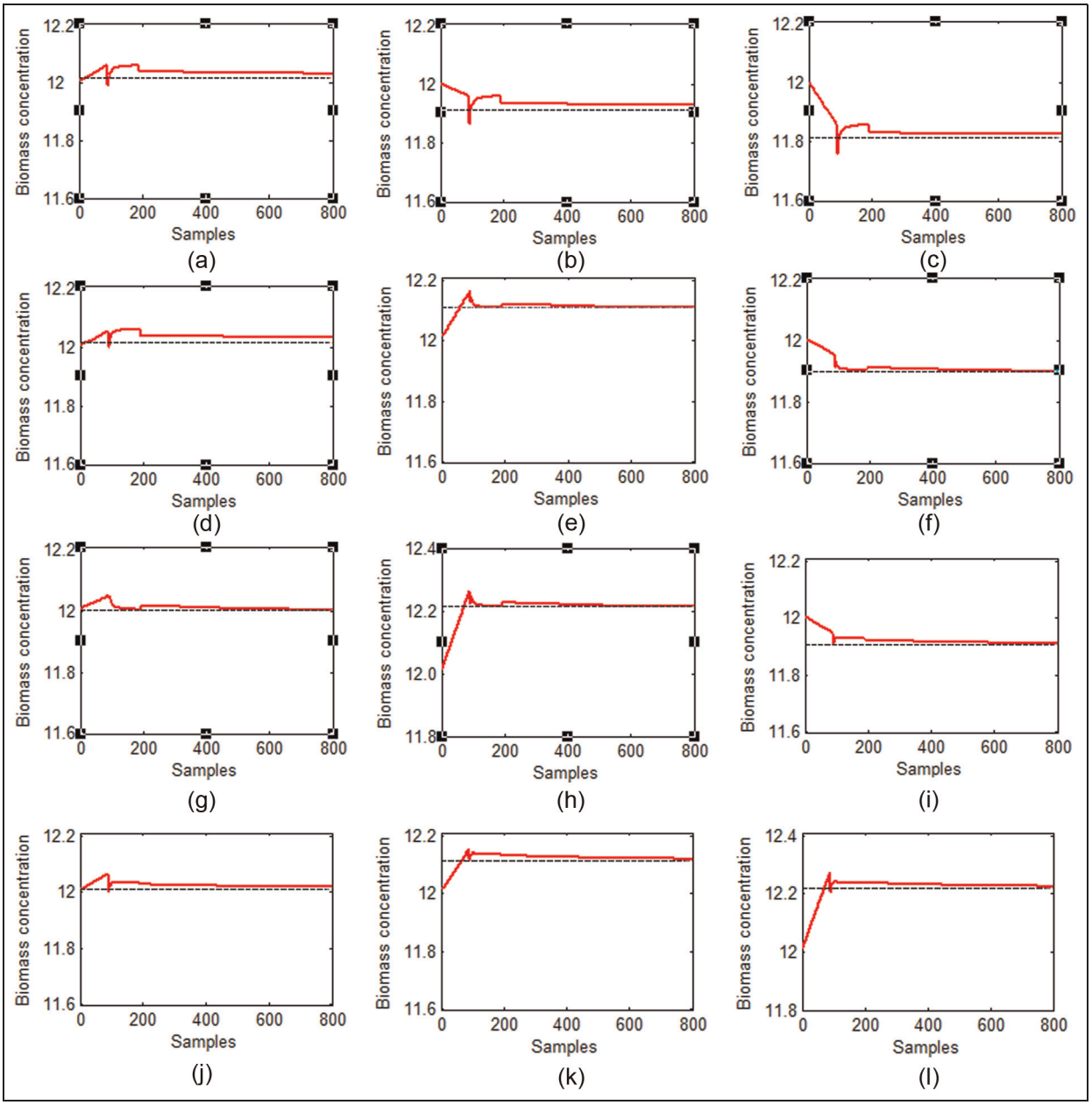

Figure 12. Online quality prediction of 12 test batches for biomass concentration from (a) to (l), respectively (sold line, online prediction; dash line, real measurement of final quality).

process development, and we can get a more accurate quality prediction.

Conclusions

In this paper, a PCPQ model-based quality prediction method has been proposed for improvement for quality prediction in a multi-phase batch process. Considering the change of correlation structures between process variables and the final product quality in different phases, and the common effect of process variables on final product quality, the PCPQ is introduced to develop a new quality prediction method, taking both critical factors into account at the same time. In the proposed method, a separated PCPQ model is built not only to recognize the different correlation characteristics but also to isolate the local effect of certain process variables on quality. Furthermore, the final product quality prediction is achieved by accumulating all the PCPQ in this way, and considering the common effect of all the different effect on quality. Online final product quality prediction is performed by predicting the PCPQ and RPCPQ according to the corresponding PCPQ and RPCPQ models. For a new batch, an integrated phase is used for PCPQ prediction, and the other

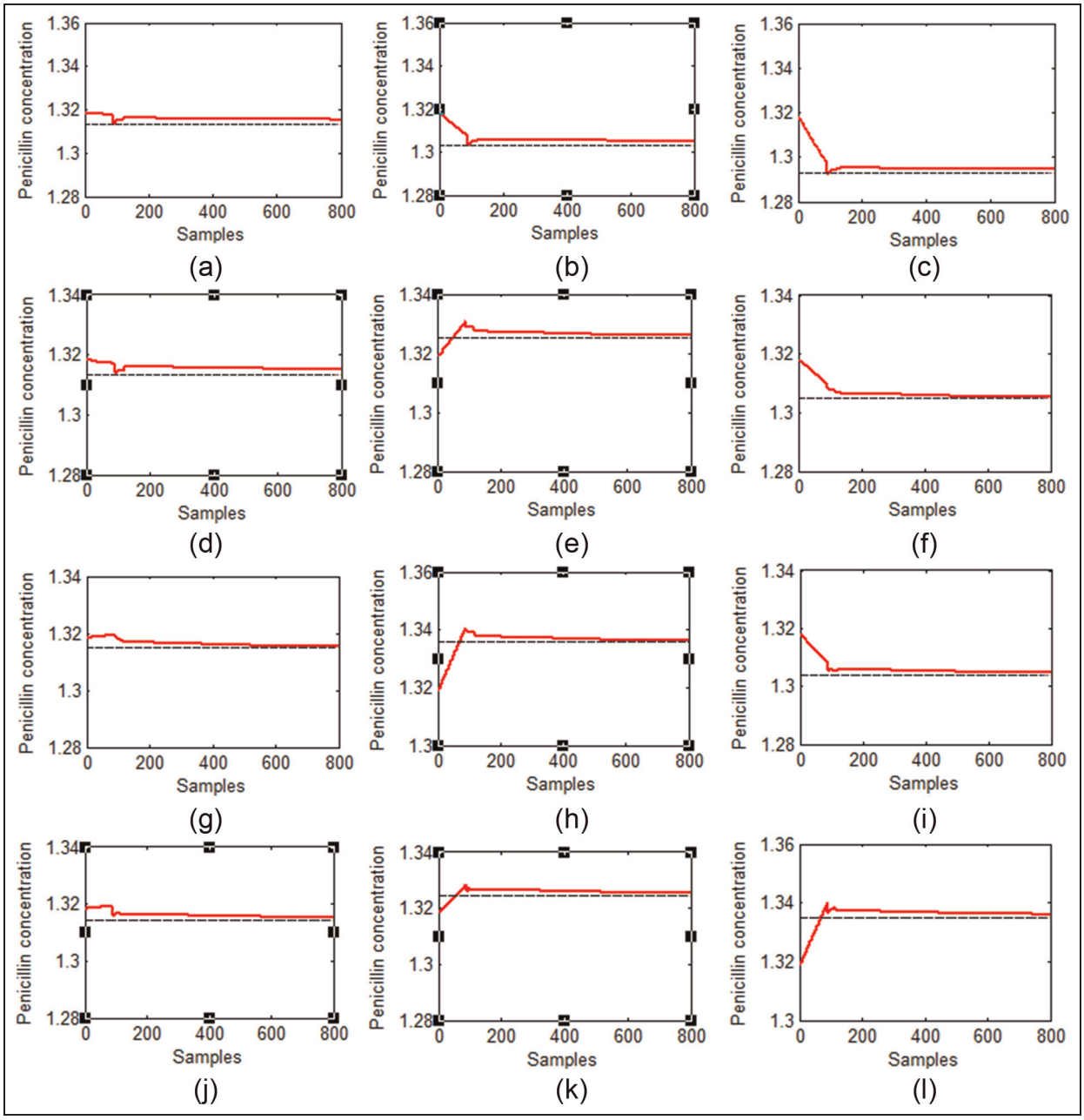

Figure 13. Online quality prediction of 12 test batches for penicillin concentration from (a) to (l), respectively (sold line, online prediction; dash line, real measurement of final quality).

process trajectories can be used for RPCPQ prediction. Finally, the final product quality prediction is realized. The application of a simulated fed-batch penicillin fermentation process shows the effectiveness and superiority, compared with conventional quality prediction in a multi-phase batch process, regardless of offline or online quality prediction.

Funding

The authors would like to acknowledge the National Natural Science Foundation of China under grant numbers 61174119, 61034006, 60774070, Liaoning Province Foundation under grant number 2009R47, the education department research project of Liaoning Province L2012139 and Liaoning Province doctoral start funds 20131089.

References

Bajpai RK and Reuss M (1980) A mechanistic model for penicillin production. *Journal of Chemical Technology and Biotechnology* 30: 332–344.

Birol G, Undey C and Cinar A (2002) A modular simulation package for fed-batch fermentation: penicillin production. *Computer and Chemical Engineering* 26: 1553–1565.

Chen J, Bandoni A and Romagnoli JA (1996) Robust statistical process monitoring. *Computers & Chemical Engineering* 20: S497–S502.

Chiang LH, Russell EL and Braatz RD (2000) Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 50: 243–252.

Duchesne C and MacGregor JF (2000) Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics and Intelligent Laboratory Systems* 51: 125–137.

Facco P, Doplicher F, Bezzo F, et al. (2009) Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process. *Journal of Process Control* 19: 520–529.

Geladi P and Kowalshi B (1986) Partial least squares regression: a tutorial. *Analytica Chimica Acta* 185: 1–17.

Gunther JC, Conner JS and Seborg DE (2009) Process monitoring and quality variable prediction utilizing PLS in industrial fed-batch cell culture. *Journal of Process Control* 19: 914–921.

Jackson JE (1991) *A User's Guide to Principal Components*. New York: John Wiley & Sons, Inc.

Kano M, Hasebe S, Hashimoto I, et al. (2001) A new multivariate statistical process monitoring method using principal component analysis. *Computers and Chemical Engineering* 25: 1103–1113.

Kosanovich KA, Dahl KS and Piovoso MJ (1996) Improved process understanding using multiway principal component analysis. *Industrial & Engineering Chemistry Research* 35: 138–146.

Lee JM, Yoo CK and Lee IB (2004a) Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis. *Journal of Biotechnology* 110: 119–136.

Lee JM, Yoo CK and Lee IB (2004b) Fault detection of batch processes using multiway kernel principal component analysis. *Computers and Chemical Engineering* 28: 1837–1847.

Louwerse DJ and Smilde AK (2000) Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science* 55: 1225–1235.

Lu N and Gao F (2005) Stage-based and quality prediction for batch processes. *Industrial & Engineering Chemistry Research* 44: 3547–3555.

Nomikos P and MacGregor JF (1994) Monitoring batch processes using multi-way principal component analysis. *AIChE Journal* 40: 1361–1375.

Nomikos P and MacGregor JF (1995) Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* 30: 97–108.

Ramaker H-J, Sprang EN van, Westerhuis JA, et al. (2005) Fault detection properties of global, local and time evolving models for batch process monitoring. *Journal of Process Control* 15: 799–805.

Undey C and Cinar A (2002) Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Systems Magazine* 22: 40–52.

Undey C, Boz I, Oztemel E, et al. (1999) Statistical monitoring of multistage batch processes. *Proceedings of the AIChE Annual Meeting*.

Undey C, Tatara E and Cinar A (2004) Intelligent real-time performance monitoring and quality prediction for batch/ fed-batch cultivations. *Journal of Biotechnology* 108: 61–77.

Undey C, Tatara E, Williams BA, et al. (2000) A hybrid supervisory knowledge-based system for monitoring penicillin fermentation. *Proceedings of the American Control Conference 2000*, pp. 3944–3948.

Westerhuis JA and Coenegracht PMJ (1997) Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics* 11: 379–392.

Wise BM, Gallagher NB, Butler SW, et al. (1999) A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics* 13: 379–396.

Zhao C, Wang F, Mao Z, et al. (2008) Quality prediction based on phase-specific average trajectory for batch processes. *AIChE Journal* 54: 693–705.

Zhou D, Li G and Li Y (2011) *Data-driven Based Process Fault Detection and Diagnosis Technology*. Beijing: Science Press.

Appendix: a PLS algorithm

Assuming input matrix $X(n\times p)$ scaled to $E_0(n\times p)=\left(E_{01},\cdots,E_{0p}\right)$ and output $Y(n\times q)$ scaled to $F_0(n\times q)=\left(F_{01},\cdots,F_{0q}\right)$. t_1 is the first score of E_0, $t_1=E_0w_1$, w_1 is the first axe of E_0. The first axe stands for the maximum direction of data variations, and w_1 is a unit vector, i.e. $\|w_1\|=1$. u_1 is the first score of F_0, $u_1=F_0c_1$. c_1 is the first axe of F_0, i.e. $\|c_1\|=1$. A PLS algorithm requires that t_1 and u_1 can stand for the variations of X and Y. This can be mathematically expressed as:

$$Var(t_1)\rightarrow\max \tag{A1}$$

$$Var(u_1)\rightarrow\max \tag{A2}$$

Meanwhile, due to modelling need, it is required that t_1 has the maximum explanations, i.e. t_1 and u_1 has the maximum correlation relationship. Therefore, there exists an expression as follows:

$$r(t_1,u_1)\rightarrow\max \tag{A3}$$

To summarize the above analysis, the PLS algorithm is equivalent to requiring the covariance between t_1 and u_1 to be the maximum value. It can be expressed as follows:

$$Cov(t_1,u_1)=\sqrt{Var(t_1)Var(u_1)}r(t_1,u_1)\rightarrow\max \tag{A4}$$

This can be mathematically expressed as a constrained optimization problem:

$$\max_{w_1,c_1}\langle E_0w_1,F_0c_1\rangle \quad s.t.\begin{cases}w_1{}'w_1=1\\ c_1{}'c_1=1\end{cases} \tag{A5}$$

Using a Lagrange operator, the optimization problem finally leads to a simple analytic solution

$$E_0{}'F_0F_0{}'E_0w_1=\theta_1^2w_1 \tag{A6}$$

$$F_0{}'E_0E_0{}'F_0c_1=\theta_1^2c_1 \tag{A7}$$

Where $\theta_1=2\lambda_1=2\lambda_2=w_1{}'E_0{}'F_0c_1=c_1{}'F_0{}'E_0w_1$, and θ_1 is the maximal objective function value, i.e. θ_1^2 is the largest Eigen-value of matrix $E_0{}'F_0F_0{}'E_0$. So, w_1 is the unit eigenvector of matrix $E_0{}'F_0F_0{}'E_0$ corresponding to maximum eigenvalue θ_1^2 and c_1 is the unit eigenvector of matrix $F_0{}'E_0E_0{}'F_0$ corresponding to maximum eigenvalue θ_1^2.

Then, regression equations, which are E_0 and F_0 related to t_1 and u_1, are obtained as:

$$E_0 = t_1 {p_1}' + E_1 \tag{A8}$$

$$F_0 = u_1 {r_1}' + F_1^* \tag{A9}$$

$$F_0 = t_1 {q_1}' + F_1 \tag{A10}$$

where E_1, F_1^* and F_1 are the residual of three regression equations, respectively. The regression coefficient vectors are calculated by

$$p_1 = \frac{{E_0}' t_1}{\|t_1\|^2} \tag{A11}$$

$$r_1 = \frac{{F_0}' u_1}{\|u_1\|^2} \tag{A12}$$

$$q_1 = \frac{{F_0}' t_1}{\|t_1\|^2} \tag{A13}$$

After the regression coefficient is available, the parameter update is performed. Substitute residual matrices E_1 and F_1 for E_0 and F_0. Then, the second axe w_2 and c_2 are obtained. Stop the parameter updating until the retained the number of principal components are available.