Uneven Batch Data Alignment with Application to the Control of Batch End-Product Quality

Abstract

Batch processes are commonly characterized by uneven trajectories due to the existence of batch-to-batch variations. The batch end-product quality is usually measured at the end of these uneven trajectories. It is necessary to align the time differences for both the measured trajectories and the batch end-product quality in order to implement statistical process monitoring and control schemes. Apart from synchronizing trajectories with variable lengths using an indicator variable or dynamic time warping, this paper proposes a novel approach to align uneven batch data by identifying short-window PCA&PLS models at first and then applying these identified models to extend shorter trajectories and predict future batch end-product quality. Furthermore, uneven batch data can also be aligned to be a specified batch length using moving window estimation. The proposed approach and its application to the control of batch end-product quality are demonstrated with a simulated example of fed-batch fermentation for penicillin production.

Keywords: Variable batch lengths; Alignment; Partial least squares; Principal component analysis; End-product quality control.

1. Introduction

Batch/semibatch processing plays a significant role in the production of low-volume, high value-added products such as specialty polymers, pharmaceuticals and fine chemicals. This is mainly due to its flexibility and easy scale-up from laboratory procedures [1]. In order to obtain consistent and desirable batch end-product quality, it is vital to effectively monitor and control these processes during every stage of their operations. For example, early detection of abnormal conditions of a process not only saves energy and raw materials, but also makes it possible to find the cause and compensate for it as well with an appropriate control strategy [2].

Process monitoring and control techniques have been extensively addressed in the literature [3]. Due to the complexity of batch processes, it is usually difficult to develop mechanistic models based on physicochemical principles. However, a multitude of process variables such as temperatures, pressures

and flow rates are often routinely measured in modern manufacturing plants. Therefore, multivariate statistical process control (MSPC) methods, which are based on process history data, become favorite or even standard approaches for process monitoring and control. Among them, multi-way principal components analysis (PCA) and multi-way projection to latent structures (PLS), which are extensions of PCA and PLS to handle three-dimensional matrices, are most widely used [4, 5, 6, 7, 8].

The basic assumption underlying the methods of multi-way PCA and multi-way PLS for process monitoring and control is that all batch durations are the same and batch trajectories are properly synchronized to make sure that similar events happen at the same time points [9]. However, in industrial practice, the batch lengths of many processes are seldom fixed as the criterion for ending a batch is usually to meet some quality requirements rather than the time. Disturbances and changes in operating conditions and raw materials can easily lead to uneven batch lengths so as to meet the specified quality requirements at the end of batch runs. Thus it is necessary to align these uneven batch data so as to use them for process monitoring and control. Potential patterns of unsynchronized batch trajectories and possible approaches to synchronize them were discussed in [10]. In the simplest cases where the trajectories with different lengths overlap in the common time part, batch trajectories can either be cut to be the minimum length of all trajectories [11] or be expanded to be the longest length of all trajectories by estimating the absent parts of shorter trajectories using missing data algorithms [12]. The later option needs enough long batches for identifying the model to estimate missing data. Another approach is to find a proper indicator variable that can be used to replace the time dimension for synchronizing batch trajectories with variable lengths [13]. Several successful applications of this approach can be found in [14, 15, 16]. However, there may not exist such an indicator variable that must be strictly monotonic and has the same starting and ending values.

Batch trajectories with variable lengths can also be synchronized through appropriately translating, expanding and contracting localized segments within them. Typical methodologies for such kind of synchronization include Correlation Optimized Warping (COW) and Dynamic Time Warping (DTW) [9]. Their ability to align chromatographic and spectroscopic profiles were further compared in [17, 18]. COW approach was originally designed to correct peak shifts in chromatographic profiles [19] and it has some limitations for baseline corrections although the correlation coefficient offers a better similarity measurement. The application of COW in on-line batch process monitoring can be found in [20]. DTW approach, which was extensively developed in speech recognition to match similar events between signals, has a wider application in the synchronization of batch trajectories. DTW

uses the principle of dynamic programming to minimize a dissimilarity measurement or a distance between two trajectories and it may shift some feature vectors in time, compress some and/or expand others so that a minimum distance is achieved [21]. DTW was first introduced into chemometrics as a tool for supervision and fault detection with particular reference to bioprocess applications [22]. Then Kassidas et al. [2] adapted the DTW algorithm to batch statistical process control as a tool to synchronize batches with different lengths and/or different local times. Thereafter various modified versions of DTW were proposed in the literature for batch trajectory synchronization [23, 24, 9]. For example, a robust DTW algorithm was proposed in [23] through combining a moving window least squares procedure with derivative DTW so as to avoid singularity points and reduce the bias of alignment results towards the reference trajectory. Most recently, an adaptation from Kassidas et al.'s approach was introduced in [9] to achieve on-line synchronization of batch trajectories.

Given any two trajectories with different lengths, DTW can be applied to synchronize them as there is always an optimal solution to minimize the distance between these two trajectories according to certain criteria. Such a synchronization mechanism considers little about the process dynamics. As a result, DTW may distort the inherent process dynamics during the adjustment of localized segments and thus the ability for fault detection and diagnosis using the synchronized data can be reduced in some occasions [25]. Except for the synchronization of process variable trajectories, the batch end-product quality measured at the end of batch runs needs to be aligned as well due to the change of batch durations for shorter batches after the synchronization. For the indicator variable approach or the DTW approach, the future batch end-product quality at the synchronized batch duration time for those shorter batches can simply be assumed to be the same values at their original endpoints. However, this may not be true due to the change of batch durations for those shorter batches, especially in case of a large change of batch durations. An alternative approach is to predict the future batch end-product quality values at the synchronized batch duration time for shorter batches using a pre-determined model. Such a model can be identified using the method proposed in [26], where a series of created pseudo batches are synchronized to their batch endpoints and a PLS model is identified from the synchronized pseudo batch data. As the synchronized pseudo batch data come from various windows of batch runs, the identified PLS model is essentially a moving-window PLS model. It can be used for predicting future batch end-product quality as long as the process dynamics does not change for the time period that the moving windows cover.

Using the model identification method in [26], this paper proposes a two-step method for aligning

uneven batch data. The first step is to identify short-window PCA&PLS models by aligning batches with uneven lengths to their endpoints and the second step is to apply the identified PCA model to estimate missing trajectories and to apply the identified PLS model to predict future batch end-product quality for those shorter batches. The paper is organized as follows. The PCA&PLS models used in this paper are briefly introduced in Section 2. The proposed method is then presented and discussed in Section 3. Case study of applying the proposed method to align uneven batch data and control batch end-product quality with options of variable batch lengths for the fed-batch fermentation of penicillin is given in Section 4 and some conclusions are drawn in Section 5.

2. Preliminaries

For multi-way PLS, process variables are divided into two groups: one stands for predictor values such as measured process variable trajectories and the other stands for response values such as measured batch end-product quality variables. Each group of data is originally a three-dimensional matrix of size $I \times J \times K$, where I is the number of batches for which data are available, J is the number of variables that are measured and K is the number of samples collected during the batch run. Although there are various possibilities to unfold the data, the batchwise unfolding approach is the most logical way to model the difference among batches [8]. The unfolded data are further mean-centered and scaled to be unit variance and performing PLS on the obtained data results in a latent variable model of the form:

$$\mathbf{X} = \mathbf{T}_1\mathbf{P}_1^T + \mathbf{E}_1, \tag{1}$$

$$\mathbf{Y} = \mathbf{U}_1\mathbf{Q}_1^T + \mathbf{F}_1, \tag{2}$$

where $\mathbf{X}$ is a matrix of $I \times J_xK_x$ for predictor variables, $\mathbf{Y}$ is a matrix of $I \times J_yK_y$ for response variables, $\mathbf{P}_1$ of $J_xK_x \times A$ and $\mathbf{Q}_1$ of $J_yK_y \times A$ are the loading matrices, respectively. Here A is the number of latent variables. The scores $\mathbf{T}_1$ and $\mathbf{U}_1$ are related by a diagonal matrix $\mathbf{B}_1$ with $\mathbf{U}_1 = \mathbf{T}_1\mathbf{B}_1$. $\mathbf{T}_1 = \mathbf{X}\mathbf{W}_1(\mathbf{P}_1^T\mathbf{W}_1)^{-1}$, where $\mathbf{W}_1$ is the weight matrix [27]. Finally, $\mathbf{E}_1$ and $\mathbf{F}_1$ are residual matrices. The identified multi-way PLS model can be applied to control batch end-product quality through trajectory manipulation [5, 28]. Specifically, if there is no measurement for product quality variables or the response variables are not considered, multi-way PCA can be applied instead to model the correlation structure of all measured process variables:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{3}$$

where $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{E}$ are the corresponding matrices for the PCA model.

3. Uneven Batch Data Alignment Method

The critical assumption made when performing batchwise unfolding is that all the batches are of the same length. This assumption ensures that the number of columns in the unfolded data matrix is the same for each row, which allows PCA and PLS techniques to be readily applied. For uneven batch trajectories, it is necessary to align uneven batch data at first and then to identify PCA&PLS models for process monitoring and control.

The principle of the proposed uneven batch data alignment method is demonstrated using an example shown in Fig. 1(a) that depicts three batches with different lengths. The variables for batch end-product quality are only measured at the end of each batch run, i.e., y_1, y_2 and y_3 for these three batches in the example. Similar to the synchronization of pseudo batches in [26], these three batches are firstly aligned toward their endpoints and a modeling window is then selected to identify the multi-way PCA and PLS models for the selected window, as shown in Fig. 1(b). The length of the selected window is denoted as Wm and the identified short-window PCA&PLS models are referred to as Wm-PCA and Wm-PLS, respectively. It is worthy to note that the data structure of the Wm-PCA and Wm-PLS models for predictor variables can be regarded as a special case of variable-wise with time-lag unfolding approach [7, 29], where only the last row of unfolded data for each batch is selected to identify the PCA&PLS models and thus the identified PCA&PLS models are anticipated to be closely related to the process dynamics at the ending stage of batch runs. The identified PCA&PLS models are essentially moving-window models as the data used for identifying the models come from different windows of all available batch data. Such a model identification approach is valid as long as the process dynamics does not change for the time period that the moving windows cover. This is the case for most single-phase processes that can be modeled by a single PCA/PLS model.

Once the Wm-PCA&Wm-PLS models for the selected modeling window are identified, they are to be used for extending shorter trajectories using missing data algorithms [30] and predicting future batch end-product quality for those shorter batches. Here the identified Wm-PCA model is used for estimating missing data while the identified Wm-PLS model is used for predicting future batch end-product quality. Such an arrangement aims to make full use of the identified Wm-PCA model for estimation and to make full use of the identified Wm-PLS model for prediction. As shown in Fig. 1(a), Batch 1 and Batch 2 are shorter batches compared to Batch 3 and thus the missing data in

Batch 1 and 2 can be estimated by placing the identified Wm-PCA model to a position that covers the missing part fully, which is shown in Fig. 1(c). Several missing data imputation methods have been proposed in the literature [30, 31]. The common idea of them is to make use of the underlying data pattern to deduce the missing part from the known part. Taking the missing data algorithm called Projection to the Plane as an example, its principle can be briefly stated as follows.

For the ith batch, the predictor values $\mathbf{X}_i$ can be partitioned as $\mathbf{X}_i = [\mathbf{X}_i^* \ \hat{\mathbf{X}}_i]$, where $\mathbf{X}_i^*$ denotes measured data while $\hat{\mathbf{X}}_i$ represents missing data. The loading matrix $\mathbf{P}$ from the Wm-PCA model can be partitioned accordingly as $\mathbf{P} = \begin{bmatrix} \mathbf{P}^* \\ \hat{\mathbf{P}} \end{bmatrix}$, where $\mathbf{P}^*$ and $\hat{\mathbf{P}}$ correspond to $\mathbf{X}_i^*$ and $\hat{\mathbf{X}}_i$, respectively. Originally, the score vector $\tau = \mathbf{X}_i\mathbf{P}$ is determined by both $\mathbf{X}_i^*$ and $\hat{\mathbf{X}}_i$. So the score vector τ connects the measured data and the missing data. It is to be optimized to best match the measured data and the optimized score vector $\hat{\tau}$ should also be the best available candidate to match the missing data. Therefore the missing data can be deduced from the optimal score vector $\hat{\tau}$, which is obtained from minimizing the following objective function:

$$J = \frac{1}{2}(\mathbf{X}_i^* - \tau\mathbf{P}^{*^T})(\mathbf{X}_i^* - \tau\mathbf{P}^{*^T})^T. \tag{4}$$

It can be seen that the objective function aims to match the known data using the score vector τ to be optimized. The optimal score vector can be obtained analytically by taking the derivative with respect to τ for the objective function and setting it to zero [30]:

$$\frac{dJ}{d\tau} = -\mathbf{P}^{*^T}(\mathbf{X}_i^* - \tau\mathbf{P}^{*^T})^T = 0, \tag{5}$$

$$\hat{\tau} = \mathbf{X}_i^*\mathbf{P}^*(\mathbf{P}^{*^T}\mathbf{P}^*)^{-1}. \tag{6}$$

When the matrix $\mathbf{P}^{*^T}\mathbf{P}^*$ is ill-conditioned, it is recommended to compute $\hat{\tau}$ through an optimization process instead [7]. Once the optimal score vector $\hat{\tau}$ is obtained, $\hat{\mathbf{X}}_i$ can be calculated as follows:

$$\hat{\mathbf{X}}_i = \hat{\tau}\hat{\mathbf{P}}^T. \tag{7}$$

The future batch end-product quality y for shorter batches can be predicted using the identified Wm-PLS model. The Wm-PLS model is placed to the same position as the Wm-PCA model and it uses the estimated missing data $\hat{\mathbf{X}}_i$ from the Wm-PCA model to predict future batch end-product quality:

$$\hat{y}_i = [\mathbf{X}_i^* \ \hat{\mathbf{X}}_i]\mathbf{W}_1(\mathbf{P}_1^T\mathbf{W}_1)^{-1}\mathbf{B}_1\mathbf{Q}_1^T. \tag{8}$$

In this way, the batch end-product quality can be updated for those shorter batches, just as the updated batch end-product quality $\hat{y}_1$ and $\hat{y}_2$ shown in Fig. 1(c).

The length of the modeling window for identifying the Wm-PCA&Wm-PLS models can be relatively small to reflect local dynamics of the process, then the estimation of missing data can be implemented in a recursive way as shown in Fig. 2. Although the modeling window may not cover the duration of the whole missing data part, it can be initially placed to cover only part of missing data and then move forward to estimate all missing data recursively. Such an approach can be regarded as a moving-window estimator for missing data. A benefit of such an approach is to make the most use of the known data to estimate missing data. In practice, the length of the modeling window can be selected heuristically with the consideration of local dynamics for accurate estimation and prediction using the identified Wm-PCA&Wm-PLS models.

A common assumption for the use of the identified Wm-PCA model to estimate missing data and the use of the identified Wm-PLS model to predict future batch end-product quality is that the identified Wm-PCA&Wm-PLS models are valid for the time period that the modeling window covers. Under such an assumption, batch data can actually be aligned to be a specific batch length between the shortest and the longest batches using moving window estimation. As shown in Fig. 3, the modeling window is moved to a position between the shortest and the longest batches for aligning batch data to be a specific length between the shortest and the longest batches. Then missing data for Batch 1 and 2 can be estimated using the identified Wm-PCA model and batch end-product quality for all these three batches can be predicted using the identified Wm-PLS model. For batches without missing data such as Batch 3 in Fig. 3, the batch end-product quality at the assumed batch ending time can simply be predicted using all known data of predictor variables:

$$\hat{y}_i = \mathbf{X}_i^* \mathbf{W}_1 (\mathbf{P}_1^T \mathbf{W}_1)^{-1} \mathbf{B}_1 \mathbf{Q}_1^T. \tag{9}$$

Therefore, all batch data can be fully aligned to be the same length with the updated batch end-product quality at the assumed batch endpoint. For the example shown in Fig. 3, the updated batch end-product quality values are $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$. Using the aligned batch data with the specified batch length, new PCA&PLS models can be identified accordingly for process monitoring and control. The ability to specify a batch length for aligning uneven batch data can be beneficial for controlling batch end-product quality as flexible options for batch lengths exist and an optimal control strategy can be chosen from these options with the compromise of batch running lengths and available control efforts. Nevertheless, the batch lengths cannot be extended arbitrarily for some shorter runs as the operation

may already reach the critical moments such as the maximum yield of product. In such cases, the extended running can be detrimental for the production and the identified short-window PCA&PLS models become invalid for the extended time period due to the change of process dynamics. Therefore the extension of such shorter batches should be excluded from the proposed approach.

4. Case study

In order to assess and validate the above uneven batch data alignment approach, a benchmark simulation for the penicillin fed-batch fermentation process is used. The simulator, called Pensim, is based upon a series of detailed mechanistic models that describe the fermentation process [32]. The following process variables are collected hourly during the fermentation process: aeration rate, agitator power, substrate feed temperature, substrate concentration, dissolved oxygen concentration, culture volume, carbon dioxide concentration, pH, fermenter temperature, generated heat and substrate feed rate. The substrate feed rate is the manipulated process variable while the batch end-product quality is the biomass concentration measured at the end of batch runs. Except for varying substrate feed rate profiles for batch runs, the fed-batch fermentation process is also subject to disturbances to aeration rate, agitator power, substrate feed rate and substrate feed temperature. Furthermore, the solution concentration for the feeding substrate is oscillating around the constant value of $600g/l$ as a result of variations in the property of raw materials. All these disturbances or noise on operating conditions and raw materials contribute to varying batch lengths in practice. It is assumed that the target biomass concentration at the end of batch runs is $12g/l$ and samples are to be taken out for laboratory assay at 160th and 180th hour to see if the target is met. If the target has been met, the batch run is to be stopped immediately. Otherwise, the batch is to continue running up to the full length of 200 hours. So the batches are likely to have the length of 160, 180 and 200 hours, respectively. Other criteria for ending a batch run can also be performed to generate batch runs with variable lengths.

Taking 40 batches with variable lengths as an example for illustrating the uneven data alignment method, trajectories of biomass concentration for these 40 batches obtained from the simulator are plotted in Fig. 4. It can be seen that these batches are of variable batch lengths and the longest batch lasts 200 hours. Note that these biomass concentration trajectories are assumed to be unmeasured but are shown here for illustrative purposes.

According to the uneven data alignment method shown in Fig. 1, these 40 batches are first

aligned to their endpoints and a modeling window is selected for identifying the Wm-PCA&Wm-PLS models. The identified Wm-PCA model is further applied to estimate the missing trajectories in shorter batches and therefore all resulting batches are of the same length of 200 hours after feeding the missing data. The future substrate feed rate is assumed to be known in advance since it is the process input. Thus it is not estimated by missing data algorithms while all other process variables are to be estimated for shorter batches. Specifically, the missing data algorithm called projection to the plane is applied here. Other missing data algorithms such as trimmed score method can also be applied.

Taking the process variable of dissolved oxygen concentration as an example, the estimated future dissolved oxygen concentration trajectory against its actual trajectory for a batch with a running length of 160 hours is shown in Fig. 5, where three cases of Wm=80 hours, Wm=100 hours and Wm=120 hours are compared. It can be seen that the selected modeling window lengths do impact the accuracy of missing data estimation. Fig. 5 shows that in this particular application the selection of Wm=100 hours achieves the best performance in terms of missing data estimation. As a result, Wm=100 hours is selected in all the subsequent simulations. However, the selection of the modeling window length is quite heuristic and process dependent. Generally, larger modeling window lengths are favorable for missing data algorithms while shorter modeling window lengths are favorable for identifying accurate local PCA&PLS models. Therefore Wm should not be excessively large so as to ensure that the process dynamics along the modeling window does not change too much.

The alignment of the shorter batch using Wm=100 hours is further compared with the DTW approach in Fig. 6, where the reference trajectory for the DTW approach is selected to be the mean trajectory for all full-length batches with similar substrate feed rate profiles. It can be seen that the performance of the synchronization using the DTW approach is closely related to the selected reference trajectory as the shorter trajectory is mainly twisted to be close to the reference trajectory without considering the process dynamics. Therefore the synchronized trajectory by the DTW approach can stray away from the real trajectory while the proposed alignment method only estimates the future missing trajectory and the past known trajectory is kept intact.

In order to further demonstrate the capability of the identified Wm-PCA model for extending shorter trajectories, Fig. 7 shows the estimation of future carbon dioxide concentration trajectories for three shorter batches and compares them with their actual trajectories that would have been obtained if these batches were prolonged to 200 hours. The figure clearly shows that the identified

Wm-PCA using Wm=100 hours successfully estimates the future carbon dioxide trajectory for each of these three batches. Here the number of principal components for Wm-PCA is selected to be 16 accounting for over 90% data variability.

The future biomass concentrations for those shorter batches can also be predicted using the identified Wm-PLS model once all missing trajectories are available. Here the number of latent variables for Wm-PLS is selected to be 8, fitting over 70% of the variation in $\mathbf{X}$ space and over 98% of the variation in the $\mathbf{Y}$ space. The estimated biomass concentrations and their actual values for those shorter batches are shown in Fig. 8, where the actual biomass concentrations at 200 hours for those shorter batches are also obtained by prolonging the corresponding simulations. Fig. 8 shows the unaligned batch end-product quality for all shorter batches as well, i.e., the batch end-product quality is kept constant from their original endpoints for those shorter batches. It can be seen that the unaligned batch end-product quality values are consistently lower than their actual values. This is due to the fact that the shorter batches are aligned to be the length of 200 hours and accordingly the values for batch end-product quality can grow along with the prolonged batch duration time.

The uneven batch data are fully aligned once all missing trajectories are extended to 200 hours and the batch end-product quality is updated at 200th hours for those shorter batches. The aligned batch data can then be used to identify a new PLS model, which is referred to as full batch length PLS model (FBL-PLS). The accuracy of FBL-PLS can be further validated by testing batches with a unified batch length of 200 hours. Using the process variable trajectories from these testing batches, the biomass concentration at 200th hours can be predicted using FBL-PLS and the predicted values are to be compared with their actual values. The number of testing batches is selected to be 100 and the test results are shown in Fig. 9. The mean value and the standard deviation for the absolute prediction errors of the proposed approach is 0.0584 and 0.0454, respectively. The mean value and the standard deviation for the absolute prediction errors of the DTW approach is 0.1072 and 0.0760, respectively. It can be seen that the identified FBL-PLS model using the aligned batch data from the proposed approach provides better predictions for biomass concentration at 200th hour compared to the identified PLS model using the synchronized data from the DTW approach. Here the DTW approach only synchronizes batch trajectories of shorter batches while the batch end-product quality for those shorter batches is kept constant from their original endpoints.

The FBL-PLS model identified from the aligned batch data can be further applied to control batch end-product quality through trajectory manipulation [5, 28]. As illustrated in Fig. 3, uneven

batch data can be aligned to be a specific length between the shortest and the longest batches using moving window estimation. Here three options of batch lengths are demonstrated, i.e., the original uneven batch data are aligned to be 160, 180 and 200 hours, respectively. Then three FBL-PLS models can be identified based on these three sets of aligned batch data. At each control decision point for the control of batch end-product quality, three control strategies can be deduced from the identified FBL-PLS models and each control strategy has its specified batch running length. The future substrate feed rate trajectories and the resulting biomass concentration trajectories for these three options at the control decision point of 20th hour are plotted in Fig. 10 and 11, respectively. The target biomass concentration is set to be $13g/l$ for all these three options. It can be seen that an optimal control strategy is deduced from each option based on the identified FBL-PLS model with the specified batch length and all three options have successfully reached the desired target. However, a shorter batch running length for the same control target needs higher control efforts, as shown in Fig. 10. Furthermore, different batch running lengths result in different optimal control profiles and some control profiles tend to be more oscillatory than the others. Balancing the benefits of shorter batch running times and the sacrifice of extra control efforts, an overall optimal control strategy can be selected from these three options. The extra freedom to select batch running lengths at each control decision point can be extremely beneficial in practice, especially for saving time and energy.

5. Conclusions

This paper has studied an approach to align uneven batch trajectories and the corresponding batch end-product quality values. The principle of the proposed method is to identify short-window PCA&PLS models at first and then to apply the identified models to estimate missing trajectories for shorter batches and also to predict future batch end-product quality for those shorter batches. Thus all batches are to be the same length through feeding the missing data to shorter batches and updating the corresponding batch end-product quality. The proposed method can also align uneven batch data to be a specific batch length between the shortest and the longest batches. Thus extra flexibility exists for the control of batch-end product quality as the remaining batch running length is not fixed at each control decision point. The application of the proposed data alignment method to a benchmark simulation for penicillin fed-batch fermentation has demonstrated its effectiveness in estimating missing trajectories and predicting future batch end-product quality.

It should be emphasized that the proposed data alignment method is only applicable to those batch

processes that can be modeled by single PCA and PLS models. For batch processes with multiple phases or key events happening during the batch run that change the correlation characteristics, multiple local models should be employed to align data for each phase so as to ensure key events overlapping for all batches. Furthermore, for those processes that can hardly be modeled by a linear model such as PCA and PLS models, nonlinear-type modeling methods should be applied instead for uneven batch data alignment. The application of the proposed data alignment method to those complex processes with multiple phases and/or nonlinear process dynamics can be the future work.

Acknowledgement

The authors would like to acknowledge The Process Modeling, Monitoring, and Control Research Group at Illinois Institute of Technology who generously provided the source code for their Pensim simulator. Great gratitude is also owed to anonymous reviewers for their pertinent comments and suggestions to improve the paper. The project is funded by EPSRC with the grant number EP/G022445/1.

References

[1] S. A. Russell, P. Kesavan, J. H. Lee, and B. A. Ogunnaike. Recursive data-based prediction and control of batch product quality. *AIChE Journal*, 44:2442–2458, 1998.

[2] A. Kassidas, J. F. MacGregor, and P. A. Taylor. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44:864–875, 1998.

[3] E. B. Martin and A. J. Morris. An overview of multivariate statistical process control in continuous and batch process performance monitoring. *Transactions of the Institute of Measurement & Control*, 18:51–60, 1996.

[4] P. Nomikos and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40:1361–1375, 1994.

[5] J. Flores-Cerrillo and J. F. MacGregor. Control of batch product quality by trajectory manipulation using latent variable models. *Journal of Process Control*, 14:539–553, 2004.

[6] H. Zhang and B. Lennox. Integrated condition monitoring and control of fed-batch fermentation processes. *Journal of Process Control*, 14:41–50, 2004.

[7] J. Flores-Cerrillo and J. F. MacGregor. Latent variable MPC for trajectory tracking in batch processes. *Journal of Process Control*, 15:651–663, 2005.

[8] M. Golshan, J. F. MacGregor, M. J. Bruwer, and P. Mhaskar. Latent variable model predictive control (LV-MPC) for trajectory tracking in batch processes. *Journal of Process Control*, 20:538–550, 2010.

[9] J. M. Gonzalez-Martinez, A. Ferrer, and J. A. Westerhuis. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. *Chemometrics and Intelligent Laboratory Systems*, 105:195–206, 2011.

[10] T. Kourti. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. *Annual Reviews in Control*, 27(2):131–139, 2003.

[11] S. G. Rothwell, E. B. Martin, and A. J. Morris. Comparison of methods for dealing with uneven length batches. *Proceedings of the 7th International Conference on Computer Application in Biotechnology (CAB7)*, pages 387–392, 1998.

[12] T. Kourti. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics*, 17:93–109, 2003.

[13] P. Nomikos and J. F. MacGregor. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37(1):41–59, 1995.

[14] C. Duchesne, T. Kourti, and J. F. MacGregor. Multivariate SPC for startups and grade transitions. *AIChE Journal*, 48(12):2890–2901, 2002.

[15] C. Ündey, S. Ertunç, and A. Çinar. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. *Industrial & Engineering Chemistry Research*, 42(20):4645–4658, 2003.

[16] M. Zarzo and A. Ferrer. Batch process diagnosis: PLS with variable selection versus block-wise PCR. *Chemometrics and Intelligent Laboratory Systems*, 73(1):15–27, 2004.

[17] V. Pravdova, B. Walczak, and D. L. Massart. A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, 456(1):77–92, 2002.

[18] G. Tomasi, F. van den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5):231–241, 2004.

[19] N. Nielsen, J. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17–35, 1998.

[20] M. Fransson and S. Folestad. Real-time alignment of batch process data using COW for on-line process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 84(1-2):56–61, 2006.

[21] M. Nadler and E. P. Smith. *Pattern Recognition Engineering*. 1993.

[22] K. Gollmer and C. Posten. Supervision of bioprocesses using a dynamic time warping algorithm. *Control Engineering Practice*, 4(9):1287–1295, 1996.

[23] Y. Zhang and T. F. Edgar. A robust dynamic time warping algorithm for batch trajectory synchronization. In *American Control Conference, 2008*, pages 2864–2869, June 2008.

[24] H. J. Ramaker, E. N. M. van Sprang, J. A. Westerhuis, H. F. M. Boelens, and A. K. Smilde. Dynamic time warping of spectroscopic BATCH data. *Analytica Chimica Acta*, 498(1-2):133–153, 2003.

[25] Y. Yao and F. Gao. A survey on multistage/multiphase statistical modeling methods for batch processes. *Annual Reviews in Control*, 33:172–183, 2009.

[26] O. Marjanovic, B. Lennox, D. Sandoz, K. Smith, and M. Crofts. Real-time monitoring of an industrial batch process. *Computers and Chemical Engineering*, 30:1476–1481, 2006.

[27] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[28] J. Wan, O. Marjanovic, and B. Lennox. Disturbance rejection for the control of batch end-product quality using latent variable models. *Journal of Process Control*, 22(3):643–652, 2012.

[29] M. Golshan, J. F. MacGregor, M. J. Bruwer, and P. Mhaskar. Latent variable MPC for trajectory tracking in batch processes: role of the model structure. *Proceedings of the 2009 American Control Conference, St. Louis, MO, USA*, pages 4779–4784, 2009.

[30] P. R. C. Nelson, P. A. Taylor, and J. F. MacGregor. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, 35:45–65, 1996.

[31] F. Arteaga and A. Ferrer. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics*, 16:408–418, 2002.

[32] G. Birol, C. Ündey, and A. Çinar. A modular simulation package for fed-batch fermentation: penicillin production. *Computers and Chemical Engineering*, 26:1553–1565, 2002.

List of Figures:

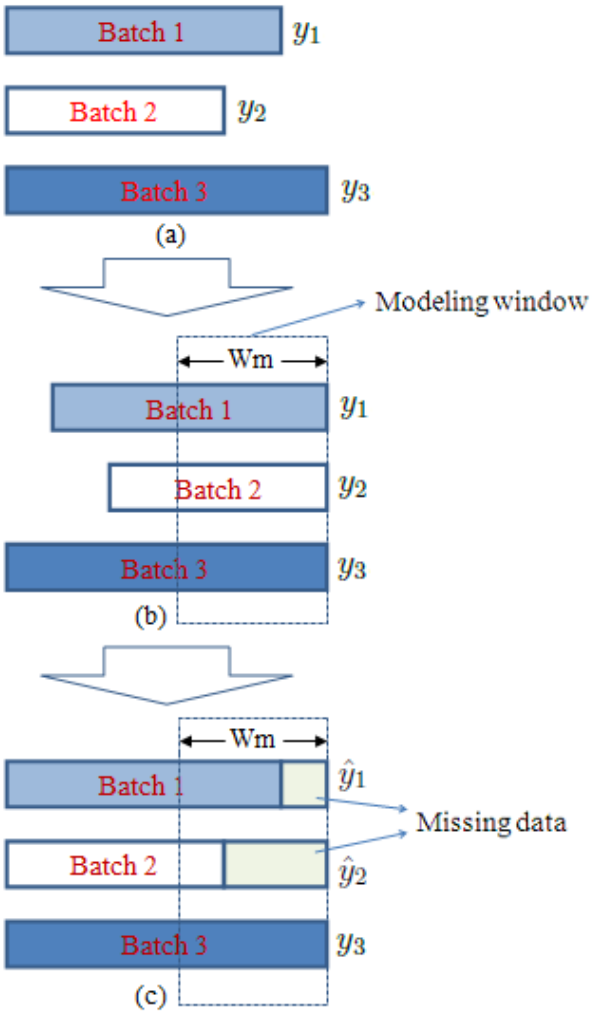

Figure 1: Alignment of uneven batch data

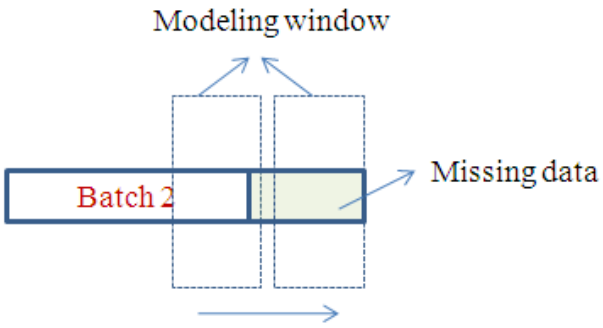

Figure 2: Estimate missing data recursively

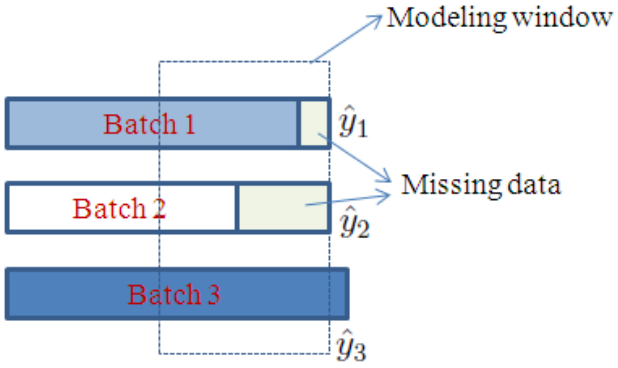

Figure 3: Align uneven batch data to be a specific length between the shortest and the longest batches

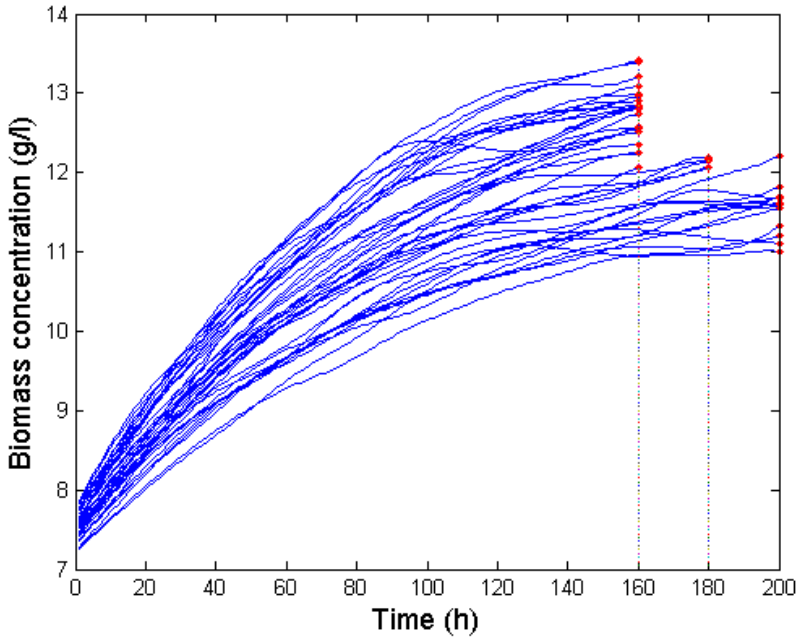

Figure 4: Batches with variable lengths

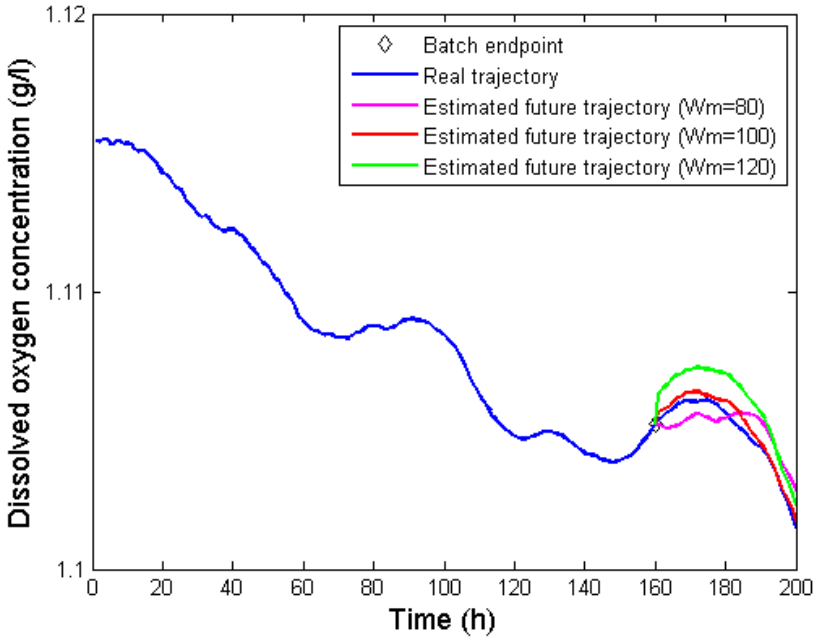

Figure 5: The impact of modeling window lengths on missing data estimation

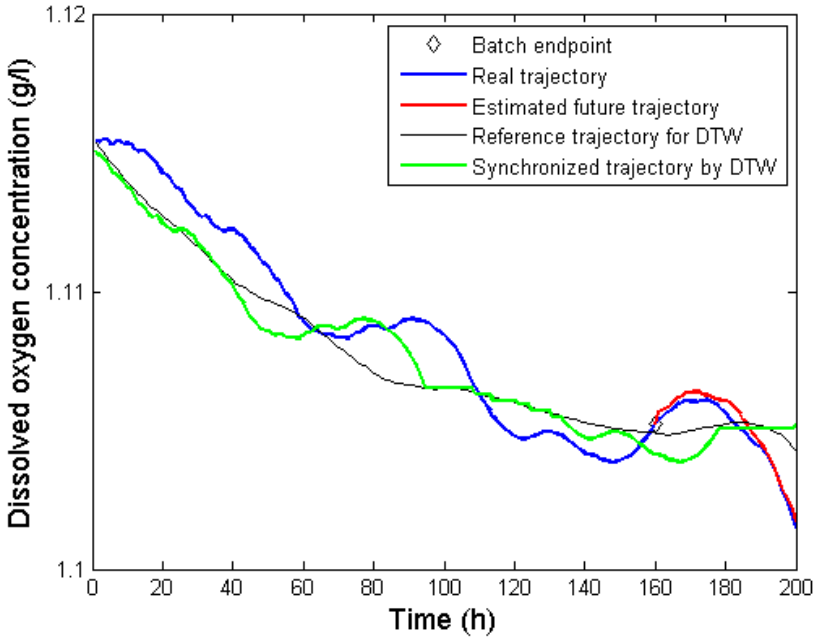

Figure 6: The comparison of uneven batch data alignment methods

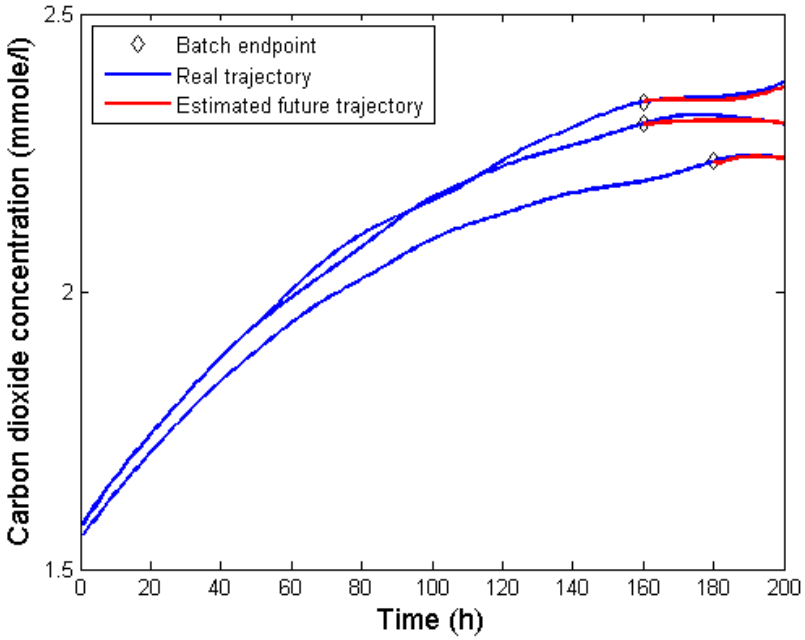

Figure 7: The estimation of missing data for three shorter batches

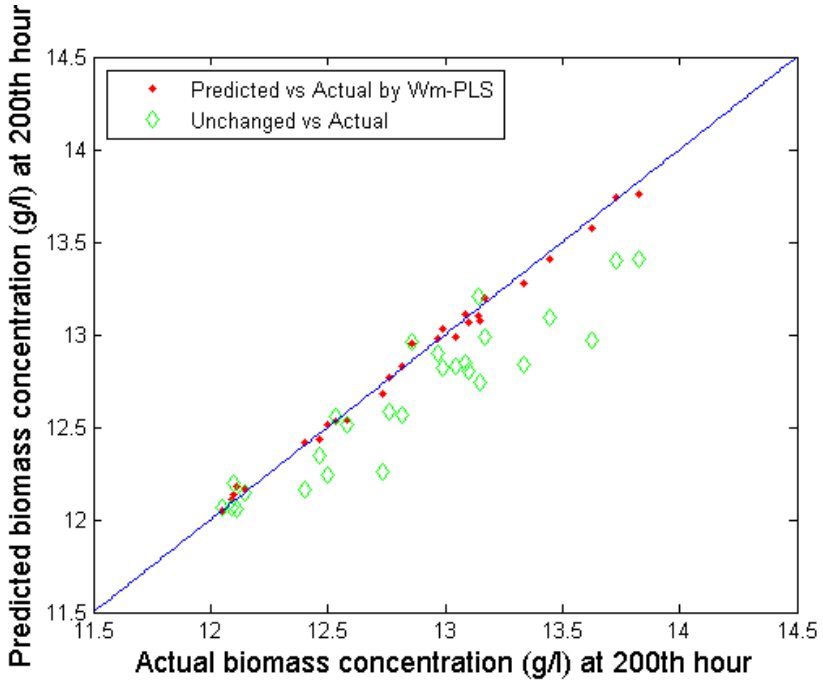

Figure 8: The prediction of future batch end-product quality for shorter batches

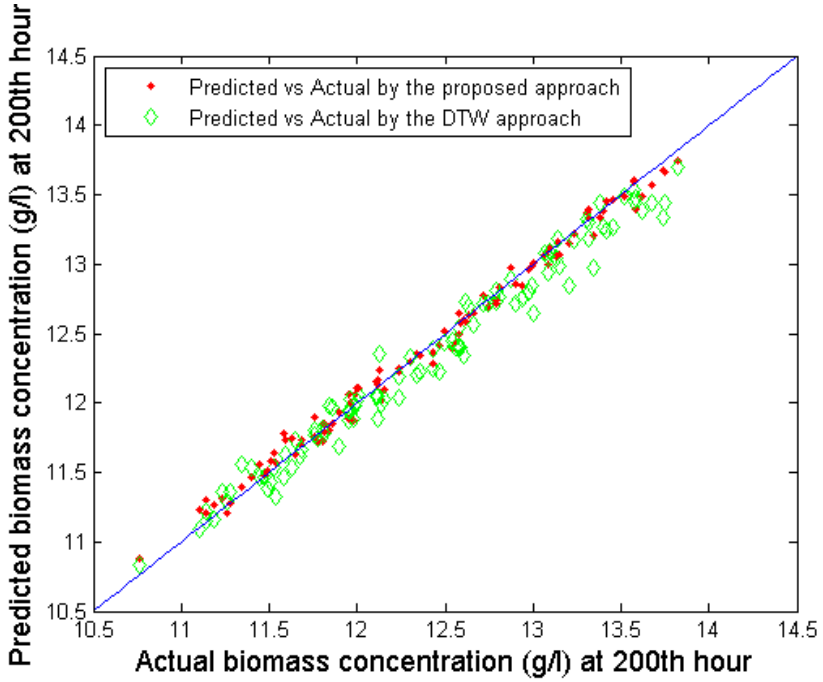

Figure 9: The test of the identified FBL-PLS model from the aligned batch data

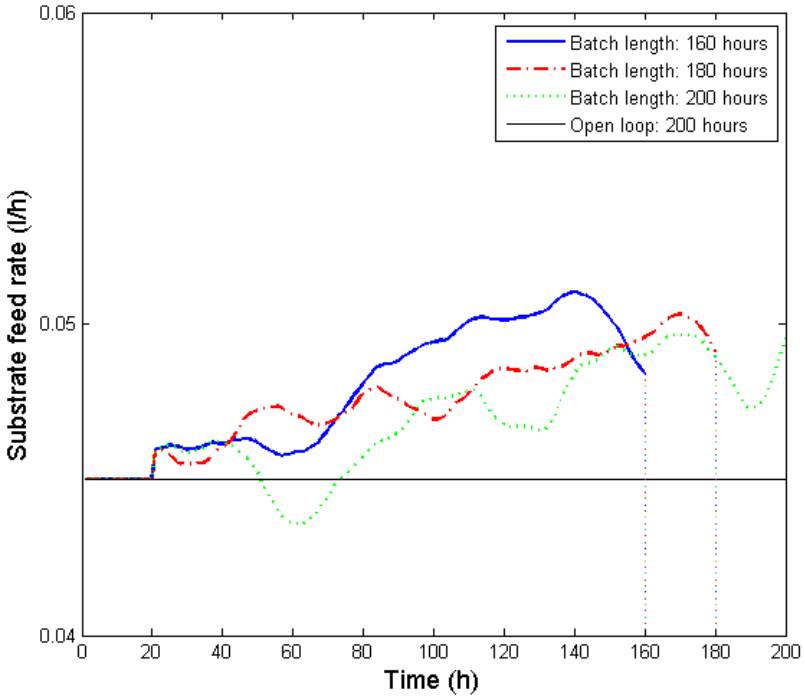

Figure 10: Three options of the future substrate feed rate trajectories

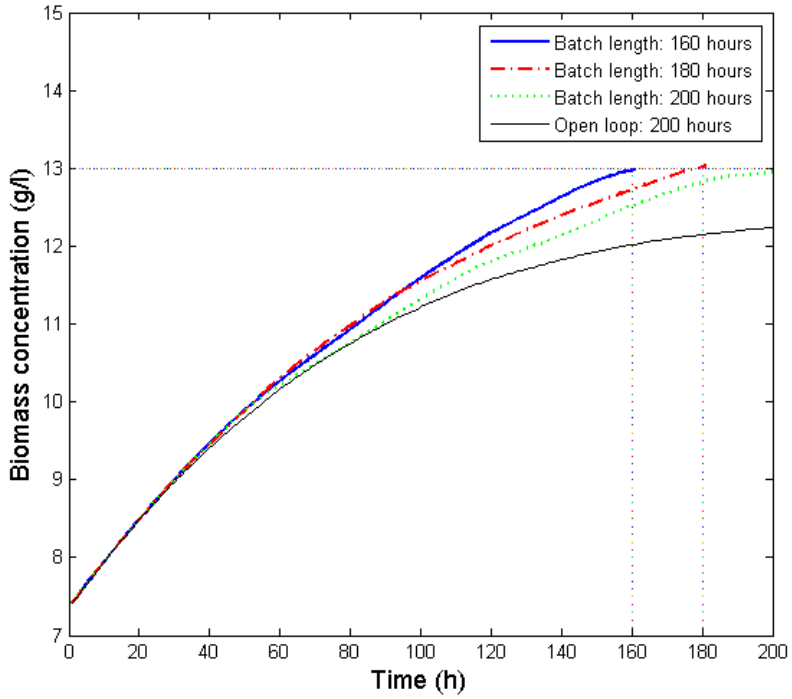

Figure 11: Three options of the future biomass concentration trajectories