2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

Prediction of Batch quality indices using functional space approximation and Partial Least Squares

Ramprasad Y., Shailesh Patel, Srikanth Ryali*, Ravi Gudi

Abstract—**Due to regulations and mandates in the pharmaceutical industry [9], it is important to be able to achieve real time release and achieve consistent batch runs. From a batch reporting perspective as well as for proactive scheduling of downstream steps in the manufacturing, accurate predictions of the end point quality indices of the batch is important. This paper presents an approach that is based on functional space approximation and multiway PLS for prediction of the batch quality indices. A novel yet simple method for completing a batch record during online batch monitoring and prediction is proposed. The proposed methodology has been validated on representative simulations involving a fed-batch fermentation for the prediction of final antibiotic yield as well as the batch durations.**

I INTRODUCTION

PROCESS monitoring is utmost importance in the operation of high value, low-volume batch processes towards meeting stringent quality constraints and minimize waste. The inherent nonlinear, time varying nature of the processes and varying initial as well as process conditions, result in batch-to-batch variation in general and unequal batch durations in particular. This nonlinearity and the variation in the batch durations inhibit the direct usage of multivariate statistical tools (PCA, PLS etc) and pose new challenges in monitoring and quality prediction. Both offline (model building steps) and online (deployment of model) tasks for batch process monitoring have their own unique challenges. During the model building steps some of the key challenges are related to (i) proper unfolding of the data matrix, (ii) synchronization of varying duration data and (iii) building the models for the prediction of yield and batch duration. In addition to these challenges, prediction models need to be developed for deployment during online batch runs so as to to predict the evolution in terms of batch durations and their yields.

To address time varying correlations of batch processes, Chen and Liu [1] proposed to partition the total batch duration into stages and develop PCA models for each of these stages]. This helps in better fault classification when compared with the application of a single PCA for the entire batch in offline monitoring. However the staged PCA approach requires the prediction of the end of each stage, which is not straightforward in an online monitoring context.

To account for the unequal data records during the offline model building step, various methods have been proposed in literature. Nomikos and MacGregor [2] proposed the use of an indicator variable against which the time evolution of a batch can be represented. However, such indicator variables, satisfying the conditions of monotonicity, may not be generically available for all batches. Kassidas [3] proposed a method based on dynamic time warping (DTW) algorithm to synchronize the variable trajectories. DTW appropriately translates, expands and contracts the patterns so that similar features within the patterns are matched. However, time warping of batch process data could lead to a loss of valuable information leading to relatively poorer discrimination and classification. Chen and Liu [4] have addressed the problem of unequal batch lengths through the use of polynomial based function space analysis to synchronize the variable trajectories. This method synchronizes all trajectories based on the concept of orthonormal function (Legendre) approximation.

On the other hand, the key difficulty during online monitoring is related to the completion of the measurement record for an on-going, evolving batch. This problem has been addressed in the literature [2] in a variety of ways. Nomikos and MacGregor proposed three methods to complete the batch data record, viz. (1) filling the future data in accordance with mean average trajectory calculated from the historical normal batch data, (2) assuming that the future deviations will remains same as at that at the current instant so as to this deviation to average normal trajectory and (iii) use the MPCA model to predict the future data from (or MPLS) model. The above methods are known to perform satisfactorily for equal batch duration data; however for unequal duration batches, one has to predict the age of new batch in order to complete the batch data. Kassidas et al. [3] proposed a DTW based method to predict the age of evolving batch. The DTW algorithm helps to find how old the batch is relative to the reference (average) trajectory.

From an online monitoring perspective, it is often times quite useful if a characterization of the quality indices at the end point of the batch is available early during the batch evolution [8]. Early prediction of the batch duration is useful in proactive scheduling of down stream processing

Manuscript received Sep 21, 2007

Shailesh Patel and Ramprasad Y. are with the Honeywell, R&T, Bangalore, India, 560076; (e-mail: Shaileshkumar.Patel@Honeywell.com).

*Srikanth Ryali, is with the Honeywell R&T, Bangalore, India, 560076 (corresponding author phone: +91-80-26588360; fax: +91-80-26584750; e-mail: Srikanth.Ryali@Honeywell.com)

Ravi Gudi is with Chemical Department, Indian Institute of Technology, Bombay, India, 400076; (e-mail : ravigudi@iitb.ac.in)

equipment. Furthermore, early prediction of the end point quality is also useful from the context of estimating the loads that would be imposed on the subsequent purification steps, for example in an antibiotic fermentation.

In this paper, we propose to address some of the issues related to this task of early prediction of end-quality indices, using a framework that is based on functional space approximation and multi-way PLS. During the model building step, the functional space approximation of the batch trajectories generates a regressor matrix that is compact and parsimonious owing to the orthonormal property of the polynomials used. The multi-way PLS algorithm [10,11] is proposed in this framework to enable the development of a predictive model that relates the coefficients in the regressor matrix to the final quality indices of the batch. During on-line monitoring, it is necessary that an incomplete batch record be completed so that it can be subjected to functional space approximation. Here, we propose to use a simple Euclidean measure based strategy to select a batch from the repository that is closest to the evolving batch, and then perform the task of data filling. The methods proposed in this paper have been validated using data generated from simulations involving a representative fed-batch, nonlinear antibiotic fermentation process.

This paper is organized as follows: Issues related to data matrix unfolding and functional space approximation using Legendre polynomials are presented in the section 2. The multi-way PLS algorithm for regression of the end quality indices is briefly presented next in section 3. Issues related to completing a batch record of an evolving batch are discussed, and a simple Euclidean distance based method to assess similarities and fill the batch record, is presented in section 4. Finally, results for the prediction of the final quality indices using simulation data is presented in section 5, followed by a summary of the work.

II Statistical Model Building

As mentioned earlier, the key steps in statistical model building are related to the data unfolding, handling time varying correlations and varying batch durations. In the sequel, we discuss briefly the salient aspects of each step.

Batch processes exhibit three way variations in the space of variable, time and batch runs as shown in the Fig. 1. To apply multivariate statistical techniques for batch process monitoring, the three way array is generally unfolded into a two way data matrix. As is well known there are three possible ways to unfold the three way array data matrix slice by slice and two possible ways to rearrange the slices [5]. Each of these six possible arrangements of 3- way array data results in a large two dimensional matrix. The direction of unfolding influences the correlation structure that can be

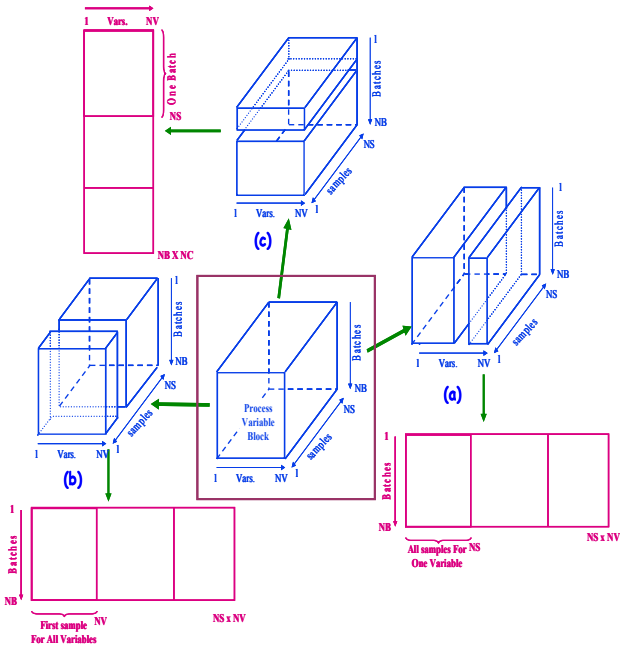

Figure 1 Unfolding of the 3 way batch data to 2 way batch data by slicing in (a) x – axis (b) y – axis and (c) z - axis

captured in subsequent steps of variance analysis. Unfolding the three way data matrix in the batch direction has a number of merits viz. the resulting mean centering along columns results in deviations around a mean trajectory and monitoring using these deviations is less likely affected by the nonlinearities present in the data [6].

$$[\mathbf{Z}]_{I\times JK} = \begin{pmatrix} \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{K1}}}^{Variable\,J=1} & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{K1}}}^{J=2} & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{K1}}}^{J=3} & \cdots & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{K1}}}^{J=J} & Batch-1 \\ \overbrace{v_{t_1} v_{t_2} v_{t_3} \ldots v_{t_{K2}}}^{Variable\,J=1} & \overbrace{v_{t_1} v_{t_2} v_{t_3} \ldots v_{t_{K2}}}^{J=2} & \overbrace{v_{t_1} v_{t_2} v_{t_3} \ldots v_{t_{K2}}}^{J=3} & \cdots & \overbrace{v_{t_1} v_{t_2} v_{t_3} \ldots v_{t_{K2}}}^{J=J} & Batch-2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{KI}}}^{Variable\,J=1} & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{KI}}}^{J=2} & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{KI}}}^{J=2} & \cdots & \overbrace{v_{t_1} v_{t_2} \ldots v_{t_{KI}}}^{J=J} & Batch-I \end{pmatrix} \quad (1)$$

Assuming the three way data matrix to be represented by X (I x J x K) with I,J,K being the indices for the batch, variable and time, respectively, the unfolding step along the batch direction yields a two way data matrix Z consisting of batches as samples as shown in Equation 1.

It must be noted that due to unequal batch durations, the rows in the above matrix Z has different lengths; also for a typical batch set, the column space is of higher dimension than the row space. Towards a more compact representation of the time series variable profiles, Chen & Liu [4] proposed the functional space approximations using orthonormal polynomials, which we briefly discuss next.

For each variable V in a batch, the time series behavior can be approximated as below.

$$V_{\{1,2,3,\ldots K_1\}} = \sum_{p=1}^{P} c_p \phi_p \quad (2)$$

where K_1 is the time duration of the variable in a batch
c_p is the coefficient associated with the basis function
Φ_p is the orthonormal basis function.

Each of the variables that are measured in the batch can likewise be expressed in a similar expansion of the orthonormal polynomials. Accordingly the time series of each variable in the matrix Z can be rewritten in a more compact form using the coefficients *c* associated with the functional expansion as in Equation (3).

$$[Z] = \begin{pmatrix} \overbrace{c_{1,1}c_{2,1}\ldots c_{nv_1,1}}^{Variable\ 1} & \overbrace{c_{1,1}c_{2,1}\ldots c_{nv_2,1}}^{Variable\ 2} & \cdots & \overbrace{c_{1,1}c_{2,1}\ldots c_{nv_J,1}}^{Variable\ J} \\ \overbrace{c_{1,2}c_{2,2}\ldots c_{nv_1,2}}^{Variable\ 1} & \overbrace{c_{1,2}c_{2,2}\ldots c_{nv_2,2}}^{Variable\ 2} & \cdots & \overbrace{c_{1,2}c_{2,2}\ldots c_{nv_J,2}}^{Variable\ J} \\ \vdots & \vdots & \ddots & \vdots \\ \overbrace{c_{1,I}c_{2,I}\ldots c_{nv_1,I}}^{Variable\ 1} & \overbrace{c_{1,I}c_{2,I}\ldots c_{nv_2,I}}^{Variable\ 2} & \cdots & \overbrace{c_{1,I}c_{2,I}\ldots c_{nv_J,I}}^{Variable\ J} \end{pmatrix} \quad (3)$$

To arrive at an optimal order for the approximation of different variables, we propose to use the relative reconstruction error (RRE) as a criterion. The RRE is an akaike-information like criteria that weighs the reconstruction error of a variable and the number of coefficients used for representation of its time domain variable profiles appropriately as,

$$\mathrm{RRE} = w \times \mathrm{MSAE} + (1-w) \times \mathrm{order} \quad (4)$$

where MSAE is the mean Square of approximation error and *w* is the weighting factor. The model order that minimizes the RRE is chosen for the functional approximation of the batch profiles. Having generated a compact representation for the batch measurements, the regression model relating these measurements to the quality indices needs to be developed. This development is briefly discussed next.

III Multiway Partial Least Square (MPLS)

Multi-way Partial Least Squares (MPLS) [10, 11] is equivalent to the application of the regular PLS algorithm on the unfolded matrices. In this procedure, the end quality indices (yield and batch duration in our case) for the normal batches are aggregated into a matrix Y and are regressed onto the matrix of features Z shown in Equation 3.

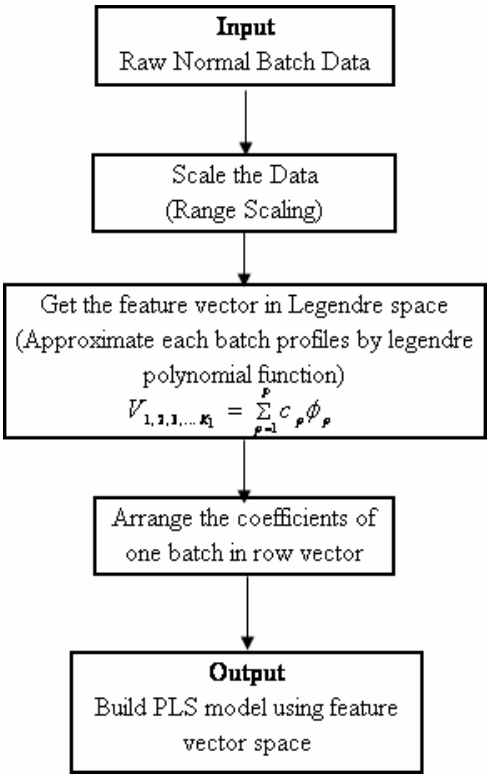

Figure 2: Overall approach for the prediction of batch yield and duration

The overall architecture for the batch process monitoring and quality prediction as depicted in Fig. 2 is described below:

Step A: Data Preprocessing

Scaling

The new batch data is scaled by the average mean and average range estimated from the normal historian batch data. Let B_{NEW} be the resultant scaled new batch data.

Function Space Analysis – Legendre Polynomial Approximation

The new batch B_{NEW} is projected on legendre orthonormal functional space. The coefficients of projection on functional space are arranged in a row by placing the coefficients for all variables. Let B_{LNEW} be the feature vector for new batch.

Step B: Prediction of batch yield and batch duration

The new scaled batch data is projected on PLS directions and the predictions of the batch yield and batch durations are obtained.

IV. Online Monitoring – Completion of Batch Record

The online monitoring and prediction step for an evolving batch requires the completion of a batch record at each instant of time, via a future prediction of how the batch is expected to evolve. While a number of methods have been proposed earlier here we propose to exploit information related to the similarity of the evolving batch to a set of batches in the archived repository of past batches [5]. Towards this end, we use a Euclidean distance based measure to assess this similarity and also reconstruct the batches as a weighted sum of the profiles of similar batches. For brevity, we illustrate this reconstruction method for only the univariate case as follows:

Consider a variable y in an online batch K that has evolved up to the current instant k. The Euclidean distance of this variable profile in the evolving batch from the variable profiles in the batch repository can be calculated as,

$$D_j = \sum_{t=1}^{k} (y_{t,K} - y_{t,j})^2 \text{ for j} = 1,2...\text{NB} \quad (5)$$

where NB is the number of batches in the repository. We next sort the batches in the repository based on these distances and consider the first L batches using the minimum distance criteria. The reconstructed trajectory of the variable y in the evolving batch K is then defined as a weighted average of the L batches in the repository as,

$$\hat{y}_{t,K} = \sum_{i=1}^{L} \frac{y_{t,j}}{w_i} \quad (6)$$

where the weights w_i are calculated in terms of the distances D_j as,

$$w_j = \frac{D_j}{\sum_{i=1}^{L} D_i} \quad (7)$$

The duration of the reconstructed batch trajectory can be likewise written as a weighted average of the L individual batch profiles. The variable L would obviously influence the accuracy of reconstruction and needs to be carefully chosen. Also, the profiles of all the variables in the batch need to be considered while calculating the weights for the reconstruction. In our work, we propose to choose L by a visual examination of the reconstructed profiles for all the variables in the known batches, and the weights w are calculated by considering a multivariate extension of the distances discussed above.

Once the future data points are filled, the batch record is complete and the steps mentioned in the section III are performed. The above method is repeated as soon as a new measurement in the evolving batch is available.

V Results

In this section, we present validation results of the proposed framework using data from simulations involving an antibiotic fermentation process. The mathematical model of the fed-batch fermentation process presented in [7] was modified to incorporate varying batch durations and yields/antibiotic titer. The data consisted of time profiles of 10 variables (such as temperature, dissolved oxygen, pH etc.) varying over batch durations of approximately 150 to 350 hours and batch yields of 0.45 g/l to 1.1 g/l, across the batches. A total of 210 normal batches were generated. The time domain information of all the variables in all the batches were projected on to the functional space and the representative feature vectors for these 210 batches were obtained. Among the 210 normal batches, 170 batches were used for training the PLS algorithm and rest 40 batches were used for cross validation.

Batch verification tests:

From the perspective of batch reporting and consistency-checks necessary for validating the batch (in terms of issues related to real time release outlined in PAT guidelines for the pharma manufacturing[9]), it is important that the performance of the batch be checked after completion. Towards this end we present the validation result for the completed test batches.

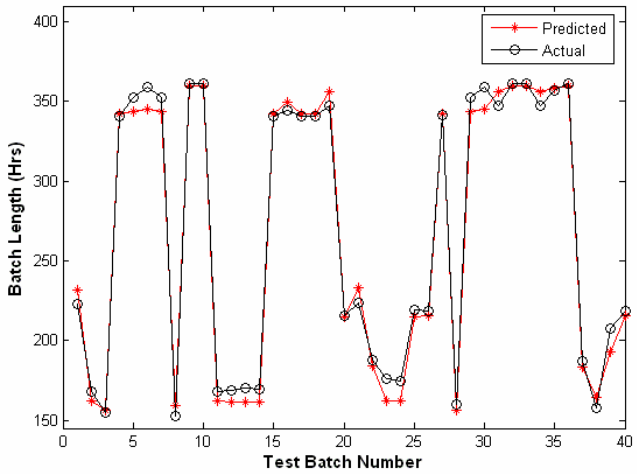

Figure 3. . Performance of PLS algorithm in predicting the batch duration for completed batches (test cases)

Figures 3 & 4 depict the performance of the PLS framework towards predicting the final batch quality indices, viz. the batch duration and yields respectively. It can be seen that the batch durations are predicted very accurately compared to the batch yields. The predictions of the batch yield are particularly poor in the low yield region, which could be attributed to lack of sufficient training data in this region.

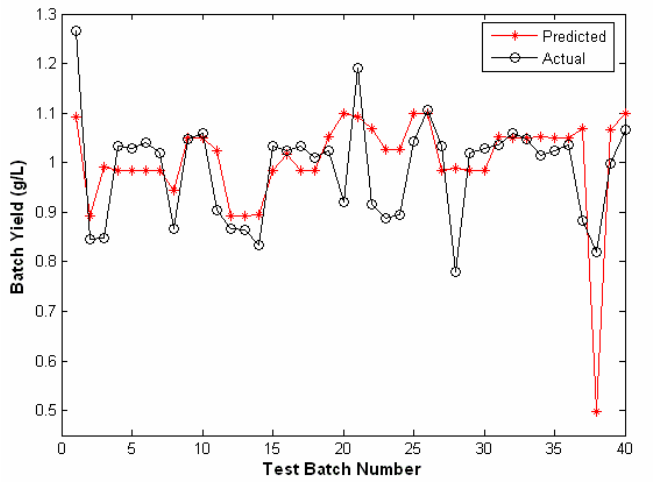

Figure 4. Performance of PLS algorithm in predicting the batch yield for completed batches (test cases)

On-line quality prediction

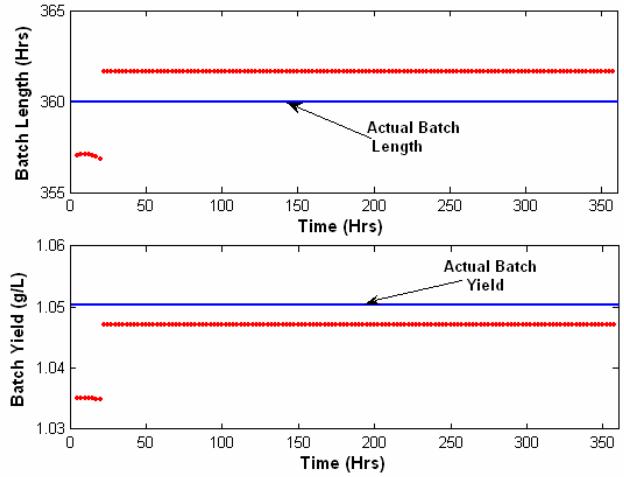

Figure 5. Performance of PLS in the prediction of batch length and yield in online monitoring

Figure 5 presents the results of on-line prediction of the end quality indices of the batch. For this prediction, the value of L (see Equation 6) was chosen to be 5. A higher value of L caused the evolving batch to be compared quite favorably with batches in the repository that were quite dissimilar and resulted in a bias in the prediction of the batch durations during the step related to the completion of batch records. On the other hand, a smaller value of L was found to adversely affect the reconstruction error and the quality of the yield predictions. The initial transients seen in Figure 5 can be explained in terms of overlapping signatures of the batch trajectories during the initial phase of the batch. However, as the batch progressed in time, more measurements for the current batch were available and the signatures were more clearly resolved. As can be seen from Figure 5 , after approximately 25 hours of batch operation, both end-point batch yields as well as batch durations were accurately predicted.

VI CONCLUSION

A framework based on functional space approximation and multi-way PLS has been proposed in this paper, for early on-line, prediction of the end quality indices of the batch. The functional space approximation was used to address the problems of varying batch duration and generate a compact representation of the batch trajectories. The multiway PLS algorithm was used to generate a regression model relating the functional space feature vectors to the end quality indices of the batch. A simple Euclidean distance based method of data filling for an incomplete batch (during online operation) was proposed and found to satisfactorily predict the future batch evolution. The proposed methodology was validated using representative simulations involving a fed-batch fermentation, and demonstrated the practicality of the proposed methodology.

REFERENCES

[1] Chen, J., and Liu, J., "*Derivation of Function Space Analysis Based PCA Control Charts for Batch Process Monitoring*", Chemical Engineering Science, 56, 2001, 3289-3304.

[2] Nomikos, P., and MacGregor, J.F., "*Monitoring Batch Processes using Multiway Principal Component Analysis*" AIChE Journal, 40(8), 1994, 1361-1375.

[3] Kassidas, A., Macgregor, J.F., and Taylor, P. A., "*Synchronization of Batch Trajectories using Dynamic Time Warping*", AIChE Journal, 44(4), 1998, 864-875.

[4] Chen, Q., Kruger, U., Meronk, M., and Leung, A.Y.T., "Synthesis of T2 and Q statistics for process monitoring", Control Engineering Practice, 12, 2004, 745-755.

[5] Nomikos, P., and MacGregor, J.F., "*Multivariate SPC Charts for Monitoring Batch Processes*", Technometrics, 37, 1995a ,41-59.

[6] Nomikos, P., and MacGregor, J.F., "*Multiway Partial Least Squares in Monitoring Batch Processes*", Chemometrics Intell. Lab. Syst., 30, 1995(b), 97-108.

[7] Birol G., Undey C., Ali Cinar, "*A Modular Simulation Package for Fed-batch Fermentation: Penicillin productions*", Computers and Chemical Engineering, 26, 2002, 1553-1565.

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

[9] Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development,

Manufacturing, and Quality Assurance, US-FDA report (2004).

[10] Kourti, T., and MacGregor, J.F., "*Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods*", Chemometrics Intell. Lab. Syst., 28, 1995, 3-21.

[11] Wold, S., Geladi, P., Esbensen, K., and Ohman, J., "*Multiway Principal components and PLS Analysis*", Journal of Chemomertrices, 1, 41-56 (1987).