BATCH PROCESS MONITORING USING MULTIBLOCK MULTIWAY PRINCIPAL COMPONENT ANALYSIS

Sinem Perk * and Ali Çınar *

* *Illinois Institute of Technology, Department of Chemical and Environmental Engineering, 10 West 33rd Street, Chicago, IL, 60616, USA*

Abstract: Batch process monitoring to detect the existence and magnitude of changes that cause a deviation from the normal operation has gained considerable attention in the last decade. There are some batch processes that occur as a single step, whereas many others include multiple phases due to operational or phenomenological regimes or multiple stages where different processing units are employed. Having a single model for all different phases/stages with different covariance structures may not give a sufficient explanation of the system behavior and fault detection and diagnosis can be more challenging with increasing model size. Multiblock methods have been recently proposed to improve the capabilities of the existing statistical monitoring models. In this study, a multiblock algorithm based on consensus principal component analysis is applied to the benchmark fed-batch penicillin fermentation simulator data. The results of a static multiblock model and a sliding window multiblock model are compared. The need for data synchronization, and the effect of block size are discussed. Multiblock multiway principal component analysis methods are found to be effective in fault detection and localization.

Keywords: Batch monitoring, Fault detection and diagnosis, Multiblock models

1. INTRODUCTION

As the importance of batch processes in chemical and biotech process industries has increased in recent years, different modeling, monitoring, diagnosis and control techniques have been proposed by researchers. Since batch data has an additional dimension of batch number, the existing statistical control models, namely principal component analysis (PCA) and partial least squares (PLS) have been improved to account for the additional batch dimension. Nomikos and MacGregor (1994) and Kourti *et al.* (1995) presented process analysis, monitoring and diagnosis procedures based on multiway PCA and multiway PLS methods (Nomikos and MacGregor, 1995*b*), which

are widely adopted by the chemical industry. To account for the multistage/ multiphase batch processes multiblock methods based on PCA and PLS have been developed. Valuable information about the multiblock methods is available in the literature. A good summary and comparison of the popular multiblock methods is provided (Smilde *et al.*, 2003). In addition, these algorithms have been applied to monitoring, fault detection and diagnosis of continuous processes and new definitions for control limits for multiblock algorithms have been developed (Qin *et al.*, 2001). MacGregor *et al.* (1994) applied multiblock PLS methods on a multisection tubular reactor for the production of low density polyethylene and Kourti *et al.* (1995) divided the process variables into multiple blocks

and simultaneously related to the quality variables. A batch monitoring algorithm based on hierarchical PCA was proposed and evaluated on an industrial batch polymerization process (Rännar *et al.*, 1998). In their work, variables from each time slot are blocked in the corresponding time block, resulting in as many blocks as the batch time points. Another application of multiblock PCA algorithm is presented, where, an adaptive multiblock PCA for the monitoring of a sequencing batch reactor was proposed (Lee and Vanrolleghem, 2003). In this work, multiblock PCA based on consensus PCA is applied to a multiphase penicillin fermentation process. Monitoring, fault detection and diagnosis using a multiblock algorithm is studied.

2. MULTIWAY MONITORING

2.1 Principal Component Analysis (PCA) and Concensus PCA

PCA is a multivariable statistical technique that can extract the essential information from a data set reported as a single block of data such as process measurements. PCA decomposes the data matrix into a score matrix times a loading matrix plus a residual matrix. This decomposition is useful to explain the information contained in the data using fewer dimensions. In the PCA computations, mostly the SVD algorithms for simultaneous or the NIPALS algorithm for sequential computation of the dominant principal components is used (Wold *et al.*, 1987).

The consensus PCA algorithm for multiple blocks, based on a series of NIPALS steps is given in (Westerhuis *et al.*, 1998). The method was introduced to compare several blocks of variables measured on the same objects. The data are divided into B blocks. A column of one of the block is selected as a starting super score and this vector is regressed on all block data to find the block loadings, from which the block scores for all blocks are calculated. All block scores are augmented in a super block. The super score is then regressed on the super block to give the super weight. The super weight is normalized and used to calculate a new super score vector. If this new super scores converges to a predefined criteria, the iteration stops. Then, each block is deflated using the super scores and the procedure repeats for the next principal component dimension. Otherwise, the iteration continues until the super score vector converges. For monitoring purposes, the statistics can be calculated for both the super level and for lower block level.
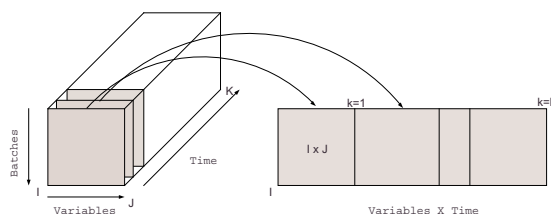

Fig. 1. Batch data representation and unfolding process.

2.2 Multiway Principal Component Analysis

Batch processes usually generate three-way data X ($I \times J \times K$), where I is the number of batches, J is the number of variables and K is the number of sampling times in a batch. MPCA is equivalent to performing ordinary PCA on an unfolded and properly scaled two-way array. Unfolding of the batch data can be done in six possible ways. For instance, each of its vertical slices ($I \times J$) is put side by side to the right starting with the slice corresponding to the first time interval forming an ($I \times JK$) dimensional matrix (Figure 1). This particular unfolding enables the analysis of variability amongst batches (Nomikos and MacGregor, 1994). Mean centering of the unfolded matrix, in other words subtraction of the mean trajectories from each variable, removes most of the nonlinearity contained in the data. And after proper scaling, PCA is applied. The ith elements of the score vectors correspond to the ith batch with respect to the other batches in the database over the entire history of the batch. The loadings matrices summarize the time variation of the measured variables about their average trajectories.

The statistics used for monitoring multivariate batch processes are the statistical distance T^2, also known as the D-statistic in literature when it is based on batch process data scores, and the squared prediction error, SPE (Nomikos and MacGregor, 1995a). If a new batch is good and consistent with the normal batches, its scores should fall within the normal range and the SPE of Q-statistic should be small.

2.3 Multiblock Consensus PCA

In the multiblock MPCA the data matrix is divided into blocks in the time dimension according to the start and end times of the different phases. In other words, each block data is $I \times J \times K_b, b = 1, ..., B$ (B : *total number of blocks*) (Figure 2). These three-way block data are unfolded and scaled for consensus PCA calculation.
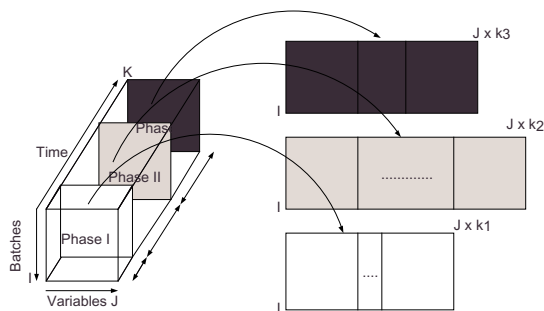
Fig. 2. Multiphase batch data unfolding.

Fig. 3. Different phases of penicillin production.

3. PENICILLIN FERMENTATION SIMULATOR: PENSIM

The penicillin fermentation simulator was developed by Birol *et al.* (2002) and has been used as a testbed for process monitoring. More information is given in Çınar *et al.* (2002). The nonlinear dynamics and multiphase characteristics of the process also make it a good candidate for multiblock model applications.

The effects of environmental variables such as pH and temperature, and input variables such as aeration rate, agitation power, feed flow rate of substrate on biomass formation have been considered for completeness. Biomass growth, CO_2, penicillin production, substrate consumption and heat generation are included in the model equations.

In a typical penicillin fermentation process the formation of the target product, the antibiotic, is usually not associated with cell growth. It is common practice to grow the cells in a batch culture followed by a fed-batch operation with continuous glucose addition to promote the synthesis of the antibiotic.

The data from the simulator have three phases. The first phase is the batch culture. Simulator can switch to the fed-batch mode either manually at a preset time or after a threshold value for the glucose is reached. The fed-batch mode consists of two phases the exponential growth phase and the stationary phase where the penicillin concentration reaches and stays at the maximum value. These three phases are shown in Figure 3.

4. MULTIBLOCK PROCESS MONITORING AND DATA ANALYSIS

Two different approaches based on multiblock consensus PCA algorithm are applied to the monitoring of the fermentation simulation data. The multiblock consensus PCA, as described above, is called the static model. Here, the reference batches from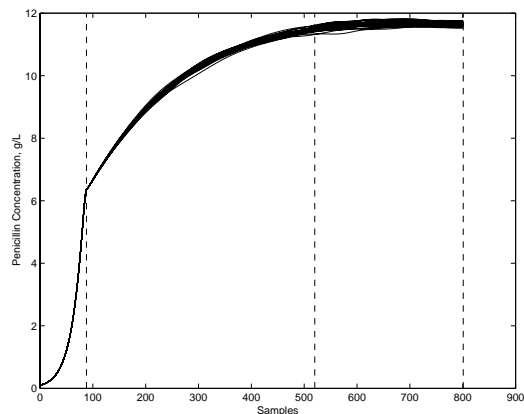 normal operation are used in the model building and a new batch is projected on to this model. It is called static model because only one multiblock model is formed and used in fault detection and diagnosis. In the sliding window model, a window of batches is slid on the reference data. A model is formed using only the batches in the window. The new batch is projected onto this model and if the system does not give an out-of-control signal, the new batch is included in the model and the first batch in the window is excluded from the model. In the following cases, both methods are employed. Each model is built using the first four dominant principal components and each block is given equal weight in the calculations.

4.1 Case 1: Multiblock monitoring with equal batch lengths

Fifty normal operating batches with small random variations are generated. The phase change times are the same for all batches. Total batch duration is 400hr and sampling time is 0.5hr, making 801 samples, including the initial conditions. The data are broken into three phases based on the preset time points and the blocks are formed. A -2% step change in aeration rate (variable 1) is introduced at 70hr and the disturbance stays in the system until the 90th hour. Since the batch/fed-batch switch takes place at 44hr, this change should affect mostly the second block, and thus second block model should signal. For the static model, all of the normal operation batches are employed in the model development. In the sliding window model, a window size of forty batches is slid through the data.

In order to detect the occurrence of a fault in a process, super model control charts are the first ones to look at. If an alarm is given at the super level, block diagnostics are the second information source to check and find out where the fault is originating from. Block contributions to the statistics is also used to detect the blocks that
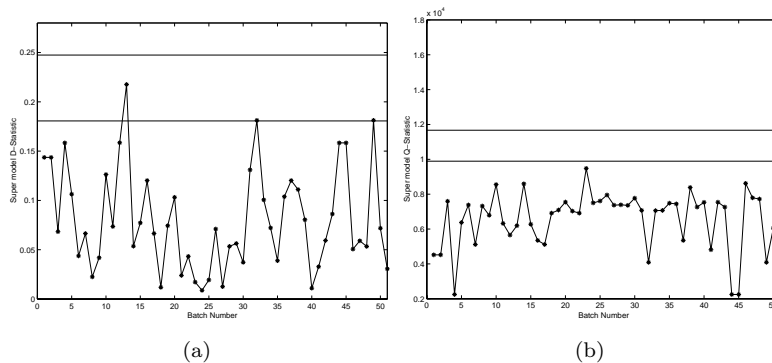
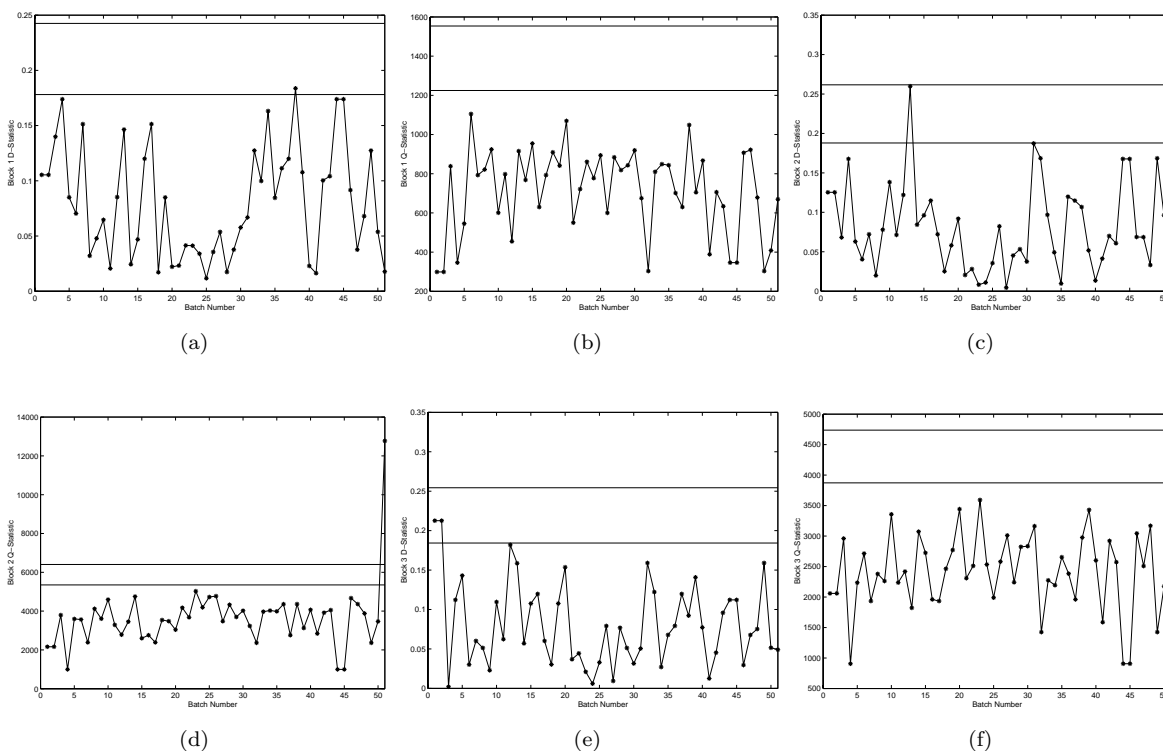Fig. 4. Case 1:Statistics for the equal length batch super model (a) D-Statistic (b) Q-Statistic.

Fig. 5. Case 1:Static block models for equal length batch data (a)D-Statistics for Block 1,(b) Q-Statistics for Block 1, (c)D-Statistics for Block 2,(d) Q-Statistics for Block 2, (e)D-Statistics for Block 3,(f) Q-Statistics for Block 3.

are mainly contributing to the fault. The variable that is responsible for the fault and the variables that are affected most are found from the variable contribution charts. Super model diagnostics are given in Figure 4, the corresponding block diagnostics are shown in Figure 5, the time of occurrence of the fault for the faulty batch is given in Figure 6 and the contribution plots are provided in Figure 7. From these graphics, it is easy to conclude that the fault is in the second phase and variable 1 is the main cause of the fault.

Sliding window model generally gives better results than a static model in the sense of false alarms. On the other hand, a poor choice of window length can result in an insensitive model,

since a slow drift of batches from normal operation can be missed in a small window. In this study, sliding window model also detects the fault and correctly identifies the main contributors to the fault. Because of space concerns the sliding window model results are not shown.

In these calculations, the blocks were given equal weight. If the block containing the fault is given a smaller weight compared to the other blocks, it leads to missed alarms. Looking at the super level diagnostics one can think the system is in-control, however, in the lower level one of the blocks may be experiencing a disturbance that need to be corrected as soon as possible. Block
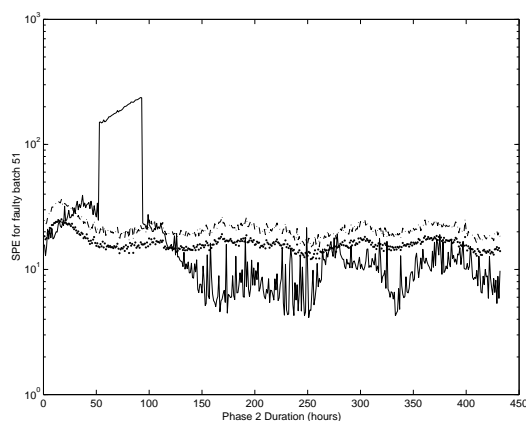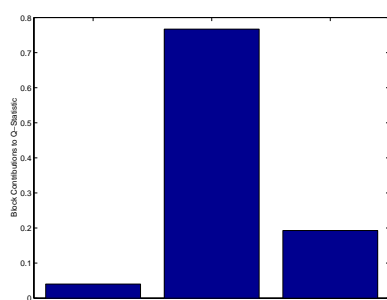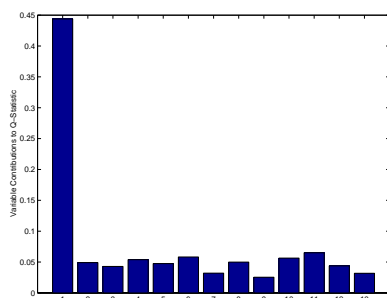
Fig. 6. Case 1: SPE for the faulty batch during the evolution of the batch.


Fig. 7. Contribution plots for equal length batch data: (a)Block contributions to Q-Statistic and (b)Variable contributions to Q-Statistic.

weighting should be given extra attention to build a reliable model.

4.2 Case 2: Multiblock monitoring with different batch lengths

Forty normal operating batches with different batch/fed-batch switch times and different durations are simulated with small random variations. The system automatically switches to the fed-batch mode after the glucose concentration drops to a threshold value, 0.3g/l. Because of the random variations introduced, each batch has a different batch/fed-batch switch time. Also, the
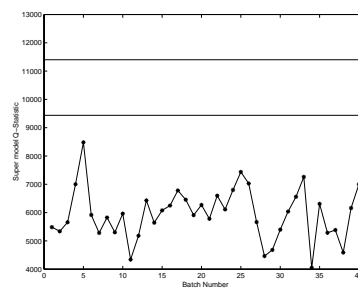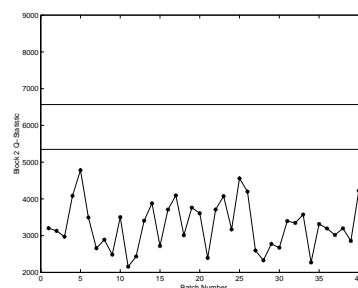
Fig. 8. Case 2:Statistics for the static block models for unaligned batch data (a)Super model Q-Statistics, (b) Block 2 Q-Statistic.

time required to reach the maximum penicillin concentration varies from batch to batch. For the penicillin fermentation it is easy to detect the phase-change points in time, since they correspond to the changes in measured variables of the system. The second phase starts with substrate addition and the third phase starts when penicillin concentration reaches its maximum value.

MPCA model development requires batches having the same length. However, in most processes, variations in process operation yield batch data of unequal length. In order to form the three-way array and apply batch monitoring tools, batch lengths should be equalized. Batch length equalization methods should also account for the alignment of process landmarks. Dynamic time warping (DTW) locally translates, compresses and expands the patterns so that similar features are aligned (Kassidas *et al.*, 1998; Ündey and Çınar, 2002).

In our case study, the start and end times for each phase during each batch are recorded at the end of a simulation. Then, each batch is divided into three different phases. Each batch in each phase is synchronized using DTW and equalized batches are put in their respective blocks.

A faulty batch is generated and projected on the models, and statistical control charts are employed to see if the fault is detected. A – 2% step change is introduced to the aeration rate (variable 1) between samples 140 and 180 (correspond to

samples 50 and 90 in the second block) and a window length of thirty batches is used in the sliding window model. The fault is detected and correctly diagnosed by both methods.

Super model Q-Statistics detects the fault and block statistics in Figure 8 show that the fault is again originated from the second phase and variable 1 is diagnosed as the main cause of the fault (not shown).

5. CONCLUSION

The multiblock PCA algorithm based on consensus PCA is used on a multiphase batch/fed-batch penicillin fermentation simulator data for monitoring, fault detection and diagnosis purposes. A static model and a constantly updating sliding window model are built on two different batch data. Both methods were effective in detecting and localizing the fault in the system, However, none of them proved to be superior to the other.

Original MPCA is applied to both data sets considering the whole batch as a single block data, and the results proved that MPCA is an effective tool for multiphase/ multistage data as well. The benefit of multiblock algorithms in localizing and isolating the fault can be better experienced in very large processes involving many processing units with many process variables. With multiblock methods, the overall process and also each different unit or subsections of a unit can be monitored. This enables the isolation of the processing unit in which the deviation occurred and detection of the major contributors to the event.

REFERENCES

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

Çınar, A., S.J. Parulekar, C. Ündey and Í. Birol (2002). *Batch Fermentation: Modeling, Monitoring and Control*. Marcel Dekker.

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

Kourti, T., P. Nomikos and J.F. MacGregor (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock, multiway PLS. *Journal of Process Control* **5**, 277–284.

Lee, D.S. and P.A. Vanrolleghem (2003). Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnology and Bioengineering* **82**, 489–497.

MacGregor, J.F., C. Jaeckle, C. Kiparissides and M. Koutoudi (1994). Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal* **40**, 826–838.

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

Nomikos, P. and J.F. MacGregor (1995a). Multivariate SPC charts for batch processes. *Technometrics* **37**, 41–59.

Nomikos, P. and J.F. MacGregor (1995b). Multiway partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* **30**, 97–108.

Qin, S.J., S. Valle and M.J. Piovoso (2001). On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics* **15**, 715–742.

Rännar, S., J.F. MacGregor and S. Wold (1998). Adaptive batch monitoring using hierarchical PCA. *Chemometrics and Intelligent Laboratory Systems* **41**, 73–81.

Smilde, A.K., J.A. Westerhuis and S. de Jong (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics* **17**, 323–337.

Ündey, C. and A. Çınar (2002). Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Systems Magazine* **22**(5), 40–52.

Westerhuis, J.A., T. Kourti and J.F. MacGregor (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* **12**, 301–321.

Wold, S., P. Geladi, K. Esbensen and J. Öhman (1987). Multi-way principal components and PLS analysis. *Journal of Chemometrics* **1**, 41–56.