Bio-reactor monitoring with multiway-PCA and Model Based-PCA

Yang Zhang and Thomas F. Edgar
Department of Chemical Engineering
University of Texas at Austin, TX 78712

Contributed to AIChE 2006 Annual Meeting, San Francisco, CA.

1. Abstract

With the rapid growth of biotechnology and the PAT (Process Analytical Technology) initiative in the pharmaceutical industry, more attention is being focused on monitoring bioreactor production to create a safe production environment and obtain a high-quality product. However, a bioreactor is difficult to monitor mainly due to the following reasons: 1) The process is always batch or semi-batch rather than continuous. 2) The dynamic behavior is highly nonlinear and rarely is a high fidelity model available to describe the dynamic behavior of the process. 3) The micro-organisms can be affected when operating conditions change unpredictably.

Typically, process monitoring methods can be divided into data-driven and knowledge-driven techniques. multiway-PCA developed by Nomikos and MacGregor [1-3] is the first and most widely used data-driven method in batch process monitoring. The basic idea of MPCA is unfolding the three dimensional batch data to two dimensions so as to perform PCA on the data matrix. Based on this pioneering work, more efforts have been made to make the technique more powerful and applicable including: 1) New data unfolding methods (batch-wise[1]; variable-wise[4]; hybrid-wise[5]); and 2) Batch data synchronization (variable indicator[2]; dynamic time warping[6]; correlation optimized warping[7]). Besides data-driven methods, Model Based-PCA (MB-PCA)[8] is based on

the fundamental knowledge of process behavior and can be successfully used in batch and continuous processes. If the process model is accurate, then the data unfolding and synchronization steps can be avoided by applying the MB-PCA method.

In this work, we focus on finding an efficient and effective way to perform PCA on a penicillin fermenter simulation model. The detailed fermenter model was developed by G. Birol et al.[9]. MB-PCA method is also applied and compared with MPCA with DTW. The effect of the coupling of manipulated and controlled variables on PCA-based fault detection is estimated.

2. Short review of multiway-PCA and Model-Based PCA (MBPCA)

a) multiway-PCA

For batch and semi-batch processes, there is no steady state and usually the historical trajectories contain considerable nonlinearity, and no effective online PCA monitoring technique existed prior to 1995. Nomikos and MacGregor [1-3] originally introduced the basic ideas of multiway PCA and PLS methods to monitor batch processes in real time. Since multiway-PCA requires the normal operating condition data to build PCA model, it is called a data-driven method.

The multiway-PCA procedure to monitor and analyze batch processes can be summarized as follows [10]:

Step A: PCA model building

1. Batch trajectory synchronization (Optional)

2. Unfold the normal operating condition's historical data $\overline{X}$ ($I \times J \times K$) into a two-dimensional array. I is the number of batches; J is the variable number and K is the sampling time.
3. Normalize the data. (mean centering and rescaling each column to unit variance)
4. Calculate the principal components and extract score and loading matrices.
5. Obtain upper control limits of *SPE* and Hotelling's T^2.

Step B: Fault Detection

The data preprocessing (first three) steps are the same as those in step A

Offline test: calculate the score of a new batch at the end of a run and compare it with the upper control limits. This procedure is usually used to check product quality.

Online test: calculate the score of a new batch at regular time intervals during the batch and compare the results with upper control limits. An online test is used to perform real-time monitoring and if there is an alarm triggered, the contribution plot, which shows the contribution of each variable to the scores of *SPE* and Hotelling's T^2 at a specific time, is used to perform fault diagnosis.

Batch trajectory synchronization can be crucial for batch process monitoring. However, since simulation data are used in this test, different trajectories are already synchronized. As a result, DTW method is not a key consideration in this paper.

The unfolding approach leads directly to the variation information that PCA must extract. Batch-wise unfolding focuses on analyzing the differences among batches; variable-wise unfolding attempts to discover the variability between variables, and time-wise unfolding is used to extract the correlation among samples (at different times). These two methods are widely used in batch monitoring.

When batch-wise unfolding is used, the dynamic behavior of the batch process is removed by this method, which is an advantage, but the row vector data will not be complete until the end of a batch. In PCA online monitoring, the score and loading calculations need a complete dataset. Thus, one has to predict the future values for the whole batch, which is time-consuming and can add uncertainty, especially during the initial period of a batch. The variable-wise method proposed by Wold et al.[4] does not have this problem because only the current time data matrix ($I \times K$) is needed for each time. The shortcoming of the variable-wise approach is that the system dynamics are still included in the dataset after preprocessing.

Lee et al. [5] combined batch-wise and variable-wise methods together, which leads to hybrid-wise unfolding method. At first, the dataset is unfolded batch-wise and the mean centering and scaling steps are performed. After that, the data are rearranged to variable-wise. The advantage of hybrid-wise unfolding is that the time dependency is cancelled and future data prediction is also avoided.

The following steps are the same as continuous process PCA (steps 3-5), for more details, refer to [1, 10].

b) MB-PCA

Compared with mutliway-PCA, MB-PCA does not need these steps, which can save computational resources. In MB-PCA, a first principles model is used to describe the nonlinearity and dynamics of normal operating conditions. The sample data are compared with the calculated data using a first principles model and PCA is performed on the residual between the model predictions and the data. If the first principles model is perfect, there will be no dynamics left in the residual dataset and the linear system assumption is satisfied.

As a result, the residual vectors of every batch can be adjusted to have the same length without performing complex calculations. Then, the residual data matrix is unfolded ($E(I \times KJ)$ etc.) and scaled according to:

$\tilde{e}_{ij} = (e_{ij} - \bar{e}_i) / \sigma(e_j)$, $\bar{e}_i$ is the mean value of each column (1)

Finally, PCA is performed on the unfolded data as continuous process. However, if the first principles model is imperfect, the residuals will still include dynamic effects and can lead to incorrect monitoring results (discussed later in section 4).

In general, the algorithm of MB-PCA can be summarized as:

Step A: Model building

1. The normal operation condition values of J process variables are stored regularly in matrix Y (J×K). In other words, Y is a 'slice' of the three dimensional matrix 'X'.

2. The model calculated values are stored in matrix Y_M.

3. The residual matrix (E) is calculated and normalized by:

$$e_{ij} = Y_{ij} - Y_{M,ij}$$

$$\tilde{e}_{ij} = (e_{ij} - \bar{e}_j) / \sigma(e_j)$$

4. $\tilde{e}_{ij}$ has zero mean and unit variance which is used in PCA model building.

Further steps are the same as those for continuous processes.

Step B: Fault detection

1. The tested values are stored in matrix Y_t.

2. The model predicted values under same control strategy are saved in $Y_{M,t}$

3. Data normalization:

$$z_{ij} = Y_{t,ij} - Y_{M,t,ij}$$

$$\tilde{z}_{ij} = (z_{ij} - \bar{e}_j) / \sigma(e_j)$$

4. $\tilde{z}_{ij}$ is the tested data and subsequent steps are the same as those in continuous process.

3. Results and discussion.

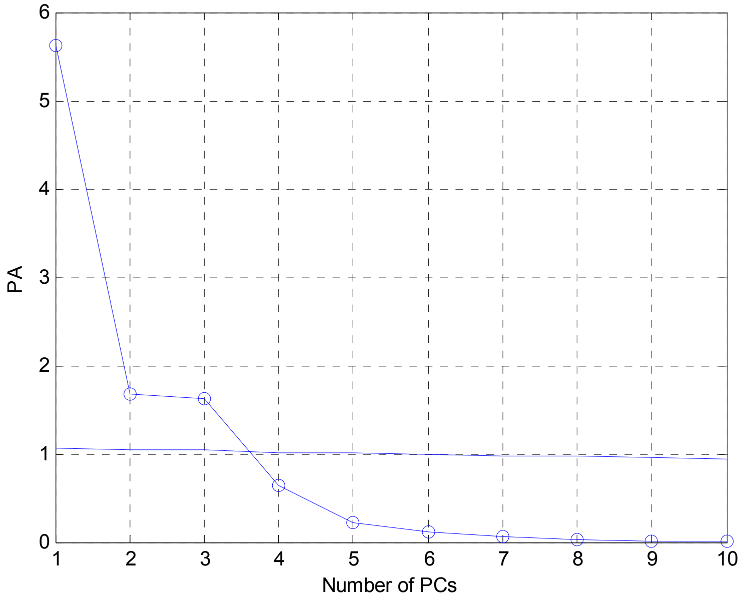
The detailed bioreactor model was developed by G. Birol et al.[9]. The simulated bioreactor volume is 1000L and the time scale is one thousand times faster than real time and variations are added to the process variables to mimic real industrial plants. In this way, all the normal batches will fluctuate around a mean trajectory. Process variable values are available very 2 seconds which is close to 1 hour real time. Ten process variables are monitored regularly, which are batch time, base reagent flow rate, head pressure, vent flow rate, biomass concentration, product concentration, broth pH, dissolved oxygen concentration, current product yield and broth temperature. Two types of process faults are generated: 1). Bioreactor vessel pressure sensor failure at time zero; 2). Bioreactor pH sensor has small deviation at time zero. Another normal batch is generated for testing.

Since offline detection is only used to check the final product quality, in this research only online monitoring is applied. Furthermore, in multiway-PCA, with batch-wise unfolding online, one needs to predict future process variable values. As a result, the other two unfolding methods (variable-wise and hybrid-wise unfolding) are tested and compared. The advantage of these two methods are future value prediction is avoided when performing online monitoring. Ten normal batch data are used to build a multiway-PCA model. For MB-PCA, the same process model is used to generate model values and the only difference between model value and real process value is there is no randomness in the model simulation. SPE control charts are used to monitor the process.

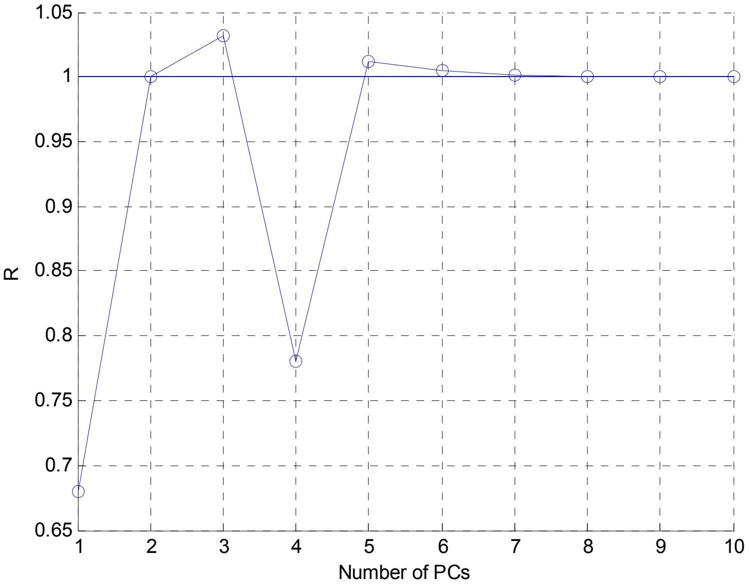
In the following sections, multiway-PCA is applied and both unfolding methods are compared. After that, the pro's and con's of MB-PCA are discussed.

3.1 multiway-PCA

The PC number selection methods are imbedded with the NIPALS algorithm to determine the number of PCs that should be retained in a PCA model. Two types of methods (Parallel Analysis (PA) and R ratio) are used to decide the number of PCs retained in PCA model and Figures 1 and 2 show the results for both methods. In Figure 4, the PA method indicates three PCs are needed while R ratio indicates one is enough. The process variance retained is 56.20% and 89.23% respectively. In this research, the number of PC's is set to three.

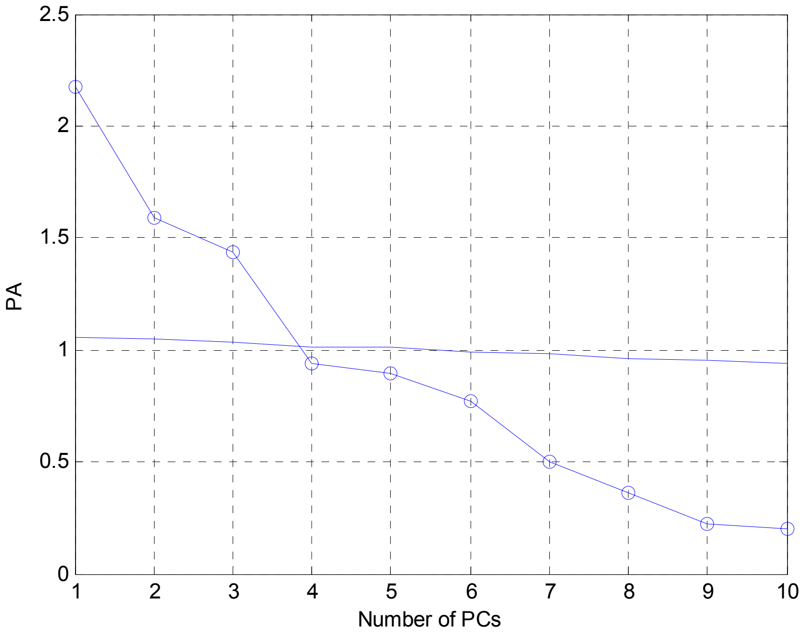


(a) PA method result
Figure 1. PC number selection with variable-wise unfolding method

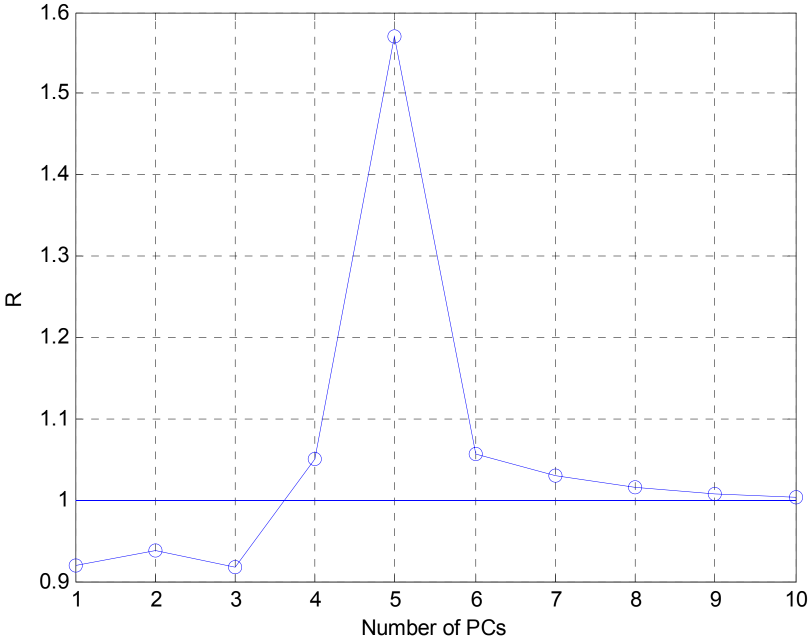


(b) R ratio method result
Figure 1. PC number selection with variable-wise unfolding method

In the same way, Figure 2 shows the results of hybrid-wise unfolding. In this case both methods suggest three PCs with 57.16% of the process variation retained. When comparing Figures 1 (a) and 2 (a), it is easy to see that variable-wise unfolding has larger eigenvalues. The larger the eigenvalues, the more process information indicate. However, after a closer look one can discover that since variable-wise unfolding does not remove process dynamics, the covariance may be dominated by a few variables instead of all variables. In other words, the covariance matrix does not fully represent the process variability. Furthermore, the covariance matrix cannot be used to explain the nonlinear relationship, hence nearly all PC selection methods are based on a linear system assumption. In this case, hybrid-wise is preferred over variable-wise unfolding.



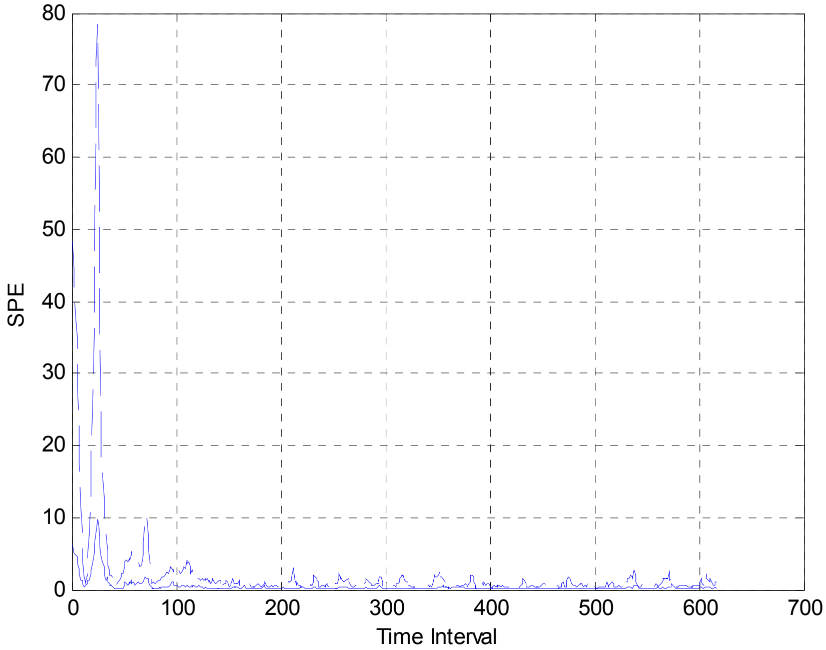
(a) PA method result
Figure 2. PC number selection with hybrid-wise unfolding method



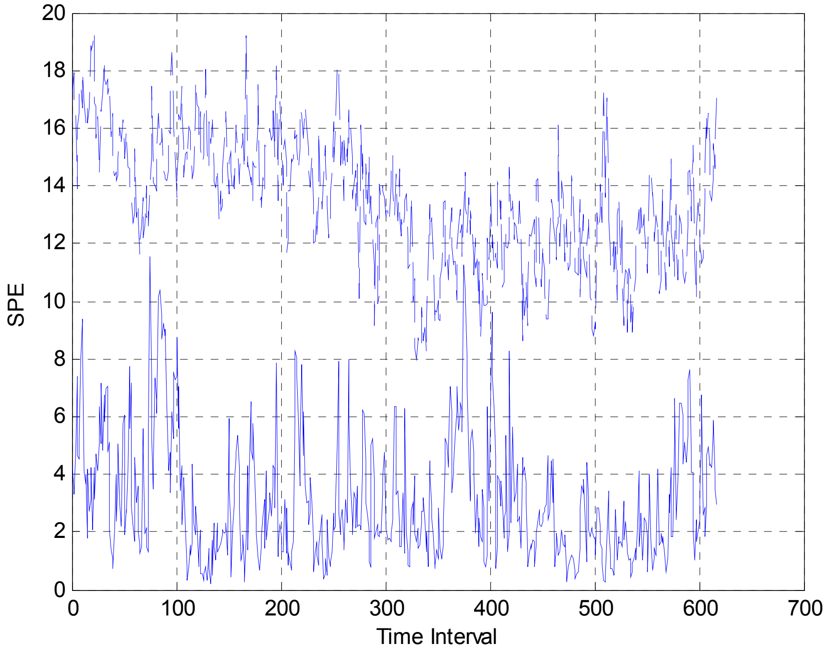
(b) R ratio method result

Figure 2. PC number selection with hybrid-wise unfolding method

After deciding how many of PCs should be retained, Figures 3 to 5 compare the results of variable- and hybrid-wise unfolding on three different cases. In Figure 3, one normal batch is monitored by both unfolding methods. It can be seen that two methods indicate there is no fault during the batch.

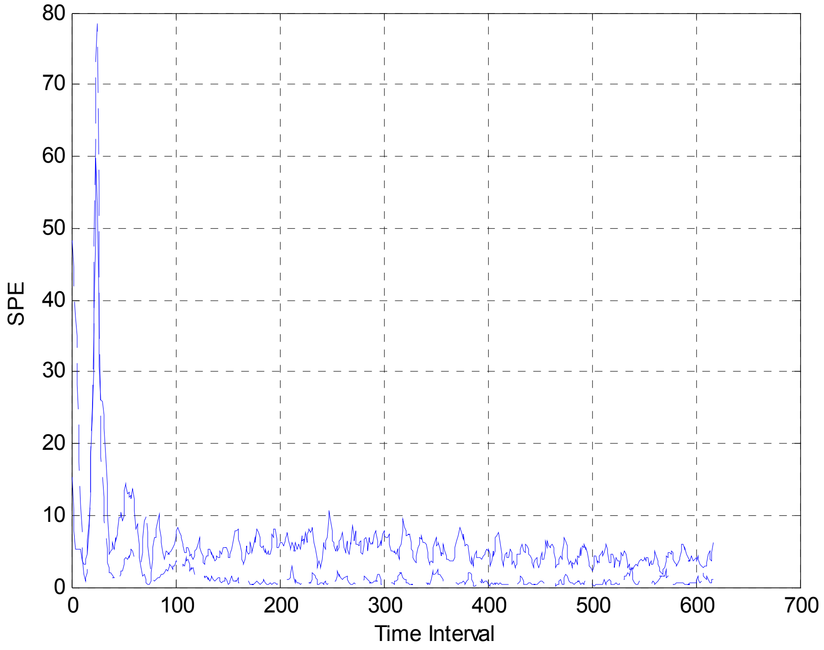


(a) Variable-wise unfolding
Figure 3. SPE control chart of a normal batch. '--' indicates upper control limit and '—' represents monitored batch.



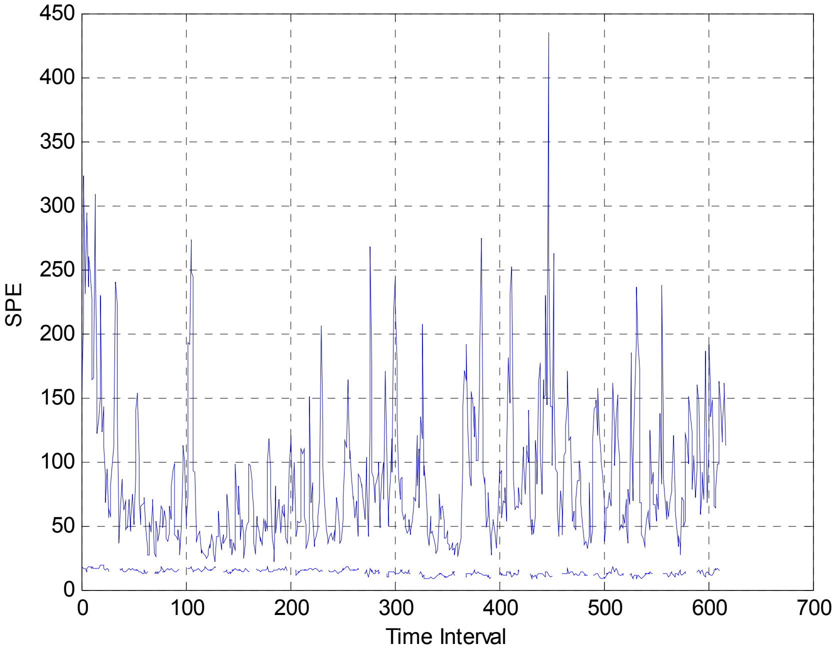
(b) Hybrid-wise unfolding
Figure 3. SPE control chart of a normal batch. '--' indicates upper control limit and '—' represents monitored batch.

In Figure 4, bioreactor vessel pressure sensor failure results are shown and both methods detect the fault. Furthermore, hybrid-wise unfolding detects the fault from time zero while variable-wise has a 50 unit time interval delay. Hybrid-wise unfolding is also favored since the monitored value violates the upper control limit much more compared with variable-wise.



(a) Variable-wise unfolding

Figure 4. SPE control chart of vessel pressure sensor failure case. '--' indicates upper control limit and '—' represents monitored batch.

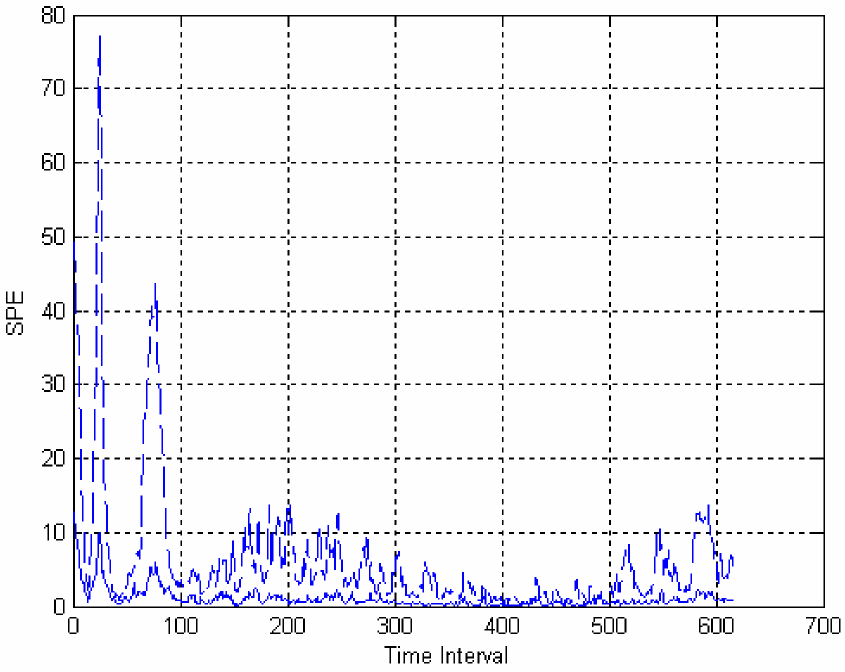


(b) Hybrid-wise unfolding

Figure 4. SPE control chart of vessel pressure sensor failure case. '--' indicates upper control limit and '—' represents monitored batch.

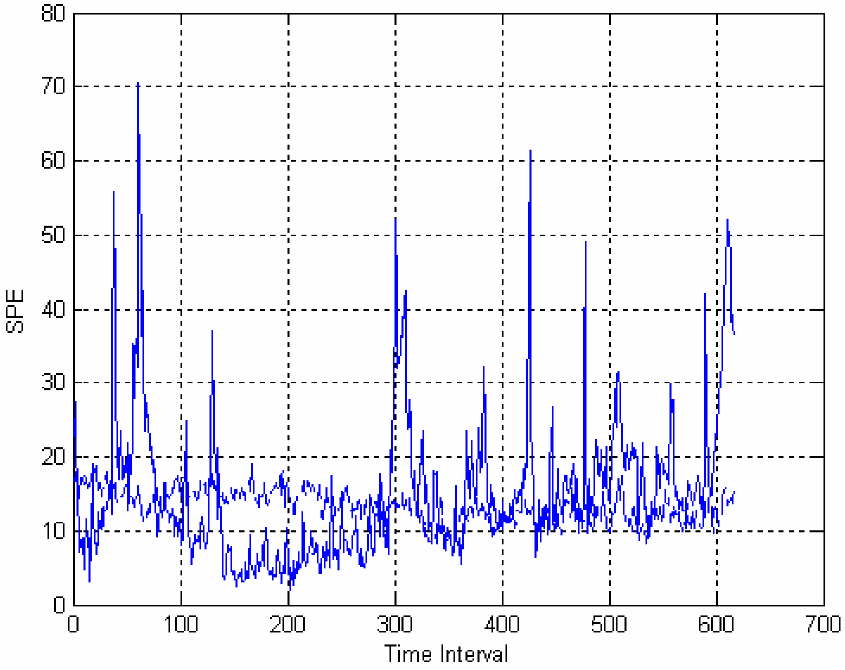
Different from the pressure sensor failure case, the pH sensor has only a small deviation compared with normal operating condition. The SPE control charts of both methods are shown in Figure 5. It can be seen that hybrid-wise unfolding detects the small deviations along the batch while variable-wise unfolding treats the batch as normal for the whole batch.

All these results suggest that hybrid-wise is preferred over variable-wise unfolding and mutliway-PCA, and a small number of batch data (ten in this case) is still effective in monitoring.



(a) Variable-wise unfolding

Figure 5. SPE control chart of pH sensor failure case. '--' indicates upper control limit and '—' represents monitored batch.

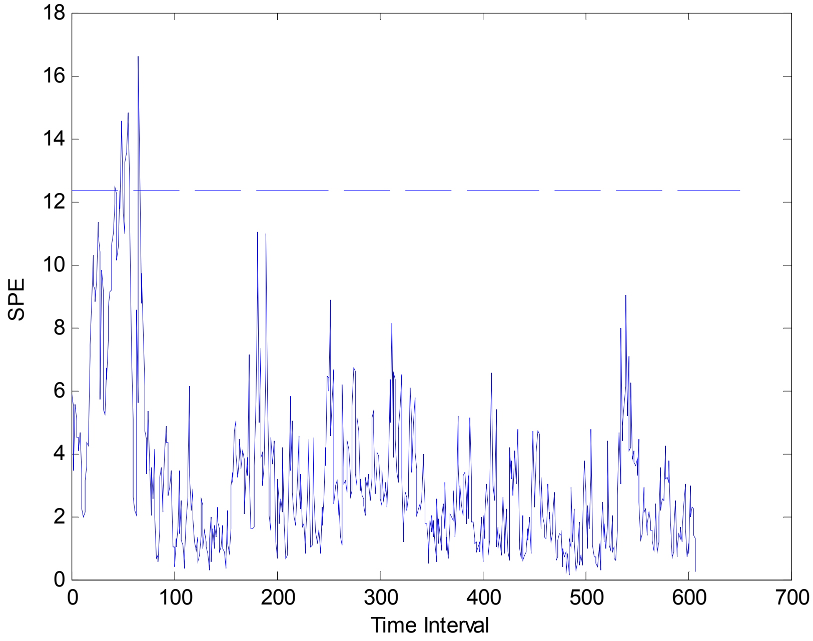


(b) Hybrid-wise unfolding

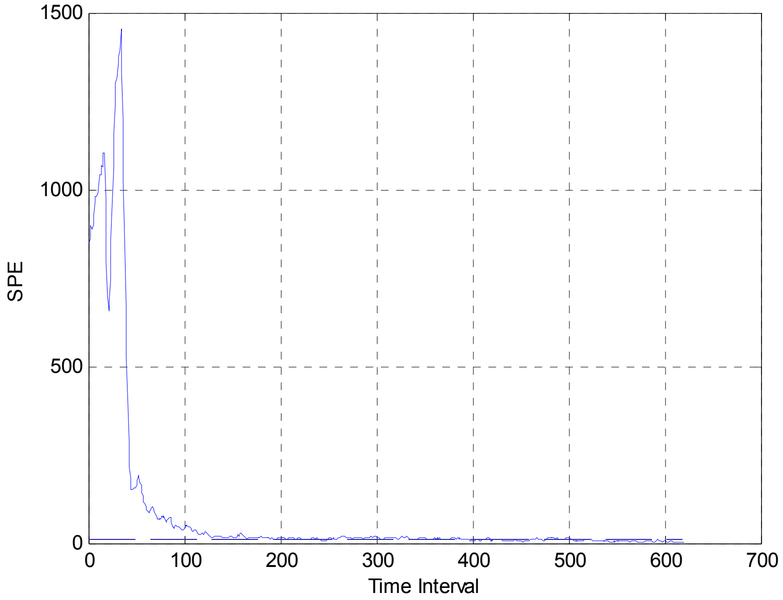
Figure 5. SPE control chart of pH sensor failure case. '--' indicates upper control limit and '—' represents monitored batch.

3.2 MB-PCA

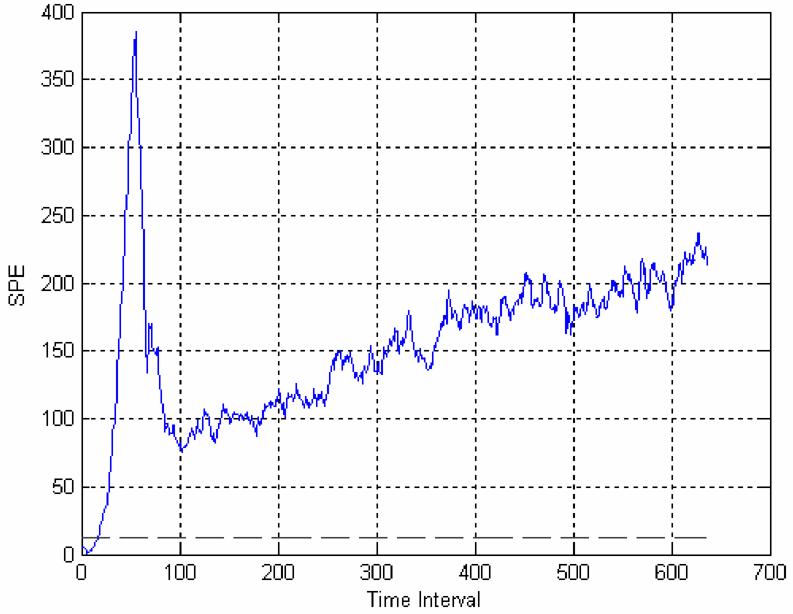
Three cases tested above are also calculated by MB-PCA and the results are summarized in Figure 6. It can be seen that MB-PCA gives satisfactory results on all cases. In addition, MB-PCA does not require data synchronization and data unfolding, these advantages can save the computation resource a lot. However, in order to obtain correct results, MB-PCA needs a process physical model that needs more engineering effort to build in order to describe the process accurately. Figures 3 to 6 also indicate both methods can be very accurate in batch process monitoring. Table 1 compares both methods from different aspects and, from the point of view of efficiency and accuracy, multiway-PCA is favored for a batch process. Meanwhile, it is worthwhile to point out that for a bioprocess where the recipe always changes, MB-PCA can be useful since multiway-PCA needs operating condition data, which can be time-consuming to generate (sometimes a few months). For MB-PCA, one only needs to update a few kinetic and physical constants.



(a) Normal batch
Figure 6. SPE control chart by MB-PCA. '--' indicates upper control limit and '— ' represents monitored batch.



(b) Vessel pressure sensor failure
Figure 6. SPE control chart by MB-PCA. '--' indicates upper control limit and '— ' represents monitored batch.



(c) pH sensor failure
Figure 6. SPE control chart by MB-PCA. '--' indicates upper control limit and '— ' represents monitored batch.

Table 1. multiway-PCA and MB-PCA comparison

	multiway-PCA	MB-PCA
Physical model	Not needed	Accurate physical model information
Synchronization (such as DTW)	Suggested	Not needed
Different unfolding method	Further treatment (multiple options)	Does not matter
Applicable system	Does not matter	Simple system
Normal operating data	Many batches are needed	One batch

4. Conclusions

Two different unfolding methods of multiway PCA are applied to a simulated industrial fermerter data. Hybrid-wise unfolding is preferred over variable-wise since it can remove process dynamics and our calculation results verified this. MB-PCA is also applied to the same system and comparisons are made between MB-PCA and multiway-PCA; multiway is preferred in terms of efficiency and accuracy most of the time.

Future work will focus on batch process diagnosis based on multiway-PCA and other MSPC methods.

Acknowledgement:

The authors thank Emerson Process Management for their funding support.

References:

1. Nomikos, P. and J.F. MacGregor, *Multivariate SPC Charts for Monitoring Batch Processes.* Technometrics, 1995. **37**: p. 41-59.
2. Nomikos, P. and J.F. MacGregor, *Monitoring of batch processes using multi-way principal component analysis.* AIChE Journal, 1994. **40**: p. 1361-1375.
3. Nomikos, P. and J.F. MacGregor, *Multi-way partial least squares in monitoring batch process.* Chemometrics and Intelligent Laboratory Systems, 1995. **30**: p. 97-108.
4. Wold, S., et al., *Modelling and diagnostics of batch processes and analogous kinetic experiments.* Chemometrics and Intelligent Laboratory Systems, 1998. **44**: p. 331-340.
5. Lee, J.-m., C.K. Yoo, and I.-B. Lee, *Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis.* Journal of Biotechnology, 2004. **110**: p. 119-136.

6. Kassidas, A., J.F. MacGregor, and P.A. Taylor, *Synchronization of Batch Trajectories Using Dynamic Time Warping.* AIChE Journal, 1998. **44**(4): p. 864-875.
7. Pravdova, V., B. Walczak, and D.L. Massart, *A comparison of two algorithms for warping of analytical signals.* Analytica Chimica Acta, 2002. **456**: p. 77-92.
8. Wachs, A. and D.R. Lewin. *Process Monitoring Using Model-based PCA*. in *Proc. IFAC Symp. on Dynamics and Control of Process Systems*. 1998. Corfu.
9. Birol G., C. Undey, and A. Cinar, *A modular simulation package for fed-batch fermentation: penicillin production.* Computers and Chemical Engineering, 2002, **26**, p. 1553-1565.
10. Zhang, Y. and Edgar, T. F. Multivariate Statistical Process Control. In: *New Directions in Bioprocess Modeling and Control* (Eds M. Boudreau and G. McMillan). ISA, 2006.