Newcastle University ePrints

Hong JJ, Zhang J, Morris J.

[Progressive multi-block modelling for enhanced fault isolation in batch processes.](#)

Journal of Process Control 2014, 24(1), 13-26.

Copyright:

NOTICE: This is the authors' version of a work that was accepted for publication in *Journal of Process Control*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Journal of Process Control*, 24(1), January 2014
DOI: 10.1016/j.jprocont.2013.10.013

Always use the definitive version when citing.

Further information on publisher website: <http://www.elsevier.com/>

Date deposited: 10th June 2014

Version of article: Author accepted manuscript

This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](http://creativecommons.org/licenses/by-nc/3.0/)

ePrints – Newcastle University ePrints
<http://eprint.ncl.ac.uk>

Progressive Multi-block Modelling for Enhanced Fault Isolation in Batch Processes

Jeong Jin Hong¹, Jie Zhang^{2*}, Julian Morris²

¹CAE Team, Display R&D Centre, Samsung Display Co., Ltd.
#95 Samsung 2-ro, Giheung-gu, Yongin-City, Gyeonggi-Do, 446-711, Korea
Email: jj.hong@samsung.com

²School of Chemical Engineering and Advance Materials
Newcastle University, Newcastle upon Tyne NE1 7RU, UK
Email: jie.zhang@newcastle.ac.uk

Abstract

A multi-block progressive modelling approach is proposed for enhanced fault isolation in batch processes. The unfolding of batch data typically leads to matrices with a large number of columns and this complicates contribution analysis. In order to rapidly focus fault isolation in batch processes, it would be desirable to employ multi-block modelling techniques. Multi-block model such as consensus principal component analysis (CPCA) can produce multiple monitoring charts for sub-blocks and block loadings and block scores can be obtained which can represent unique behaviour of each sub-block. CPCA model uses super score which is the same as score from normal principal component analysis (PCA) model and it does not produce enhanced monitoring performance. Multi-block PCA (MBPCA) model using block score for model calculation can represent sub-blocks' character but block scores are obtained from super loading so it may not be the best way to describe sub-blocks. A new MBPCA model is proposed for better expression of each sub-block. Through progressive modelling and contribution analysis, variables related to or affected by the fault, as well as the associated time information, are gradually identified. This enables a fault propagation path being established. The proposed method is applied to a benchmark simulated penicillin production process, PenSim.

Keywords: Fault Diagnosis, Process Monitoring, Batch Processes, On-line Progressive Modelling, Multi-Block Method

* Corresponding author

1. Introduction

Statistical process control (SPC) has been widely adopted to improve productivity and product quality without big investment of facilities. The most common SPC techniques such as Schewart chart and CUSUM statistics are generally referred to as 'Univariate Statistical Process Control' (USPC). They produce monitoring charts with control limits for a small number of variables to monitor those variables during the process (MacGregor and Kourti, 1995). However, this method is not very effective for processes with a large number of process variables because the USPC technique considers one variable each time and does not consider correlations among the process variables (Kourti, 2003; MacGregor and Kourti, 1995; Martin et al., 1996). Multivariate Statistical Process Control (MSPC) technique has been developed to overcome the limitation of the USPC. MSPC utilises multivariate statistical techniques, such as principal component analysis (PCA) and partial least square (PLS), to perform dimension reduction for the high dimensional process data with high level of correlations so that the nominal process behaviour can be represented by a smaller dimension latent variables or scores (Chiang et al., 2001; Kourti et al., 1995; Louwerse and Smilde, 2000; MacGregor and Kourti, 1995; Qin, 2003).

MSPC techniques were first developed for continuous processes and then extended to batch processes. Batch process is the widely used processing type in industries such as biochemical, pharmaceutical, semiconductor and display manufacturing industry (Gallagher et al., 1996; Nomikos and MacGregor, 1995a; Wise et al., 1999; Yoo et al., 2004; Zhang and Edgar, 2006). A big difference between continuous and batch processes is the process duration. A continuous process typically has a long processing duration once the process is started and the conversion from raw material to product is continuous. However, a batch process has shorter processing duration and the production is intermittent. In addition, batch processes have batch to batch variations meaning that each batch shows slightly different performances with the same recipe.

Batch process monitoring using the multivariate statistical process control (MSPC) method based on multi-way principal component analysis (MPCA) was proposed and developed by Nomikos and MacGregor (1994; 1995a; 1995b). Since then, there have

been many published researches on batch process monitoring using MSPC techniques (Gallagher et al., 1996; Kouti et al., 1995; Lennox et al., 2000; Nomikos and MacGregor, 1995a; Wise et al., 1999; Yoo et al., 2004; Zhang and Edgar, 2006). Monitoring charts in terms of the monitoring statistics, squared prediction errors (SPE) and the T^2 statistics, are produced and the monitoring statistics are checked with their control limits. Once a monitoring statistics exceeds its control limit, an abnormal situation is detected. Contribution plots (Miller et al., 1993) are typically used in fault diagnosis. Contributions from individual process variables to the monitoring statistics are produced to identify the process variables that are most related to the fault. However, this procedure is not able to determine if a variable is just affected by the fault or the variable is the cause of the fault. And it does not provide time information of when the abnormalities on the highly contributed variables occurred.

A progressive PCA modelling method was developed to overcome this problem (Hong et al., 2011). However, it deals with all the process variables in the process as one data block. This may not be efficient when the monitored process contains a large number of variables and/or data samples. Multi-block method such as consensus PCA (CPCA) model (Wold et al., 1987) can be applied as alternative method to MPCA as it has multiple sets of scores and prediction can be carried out for the data in each sub-block rather than the entire block of data. Therefore, each sub-block can be monitored individually and this can give the advantage of finding fault location more effectively in that the problem can be localised in one or a small number of data blocks (Wold et al., 1996, Qin et al., 2001). For a CPCA model, super score is used for computing block loadings for sub-blocks. As super score represents overall behaviour during a whole processing time, multiple sets of block loadings are all reflected by the overall behaviour. It means that the loadings for each block tend to exhibit a profile for a whole duration rather than a profile for each sub-block and it may not be ideal to analyze local process behaviours of sub-blocks.

A new multi-block PCA model is proposed in this study which uses block scores for computing a set of block loadings instead of the super score. Thus, a set of block scores computed by this method can address the behaviour existed in sub-blocks better than the one computed by CPCA method. The new multi-block PCA method is

used within the progressive modelling framework to establish fault propagation path efficiently for batch processes. The proposed method is applied to the benchmark simulation of a fed-batch penicillin production process, PenSim (Birol et al., 2002). Three fault batches were analysed using three different modelling methods, the conventional MPCA, CPCA and the new MBPCA. These modelling methods were applied with the progressive modelling procedure to compare the performances. The proposed multi-block PCA gives better results than other two methods for all cases when it combines with the progressive modelling scheme.

The paper is organised as follows. Section 2 presents the background of this study. Section 3 presents the new multi-block PCA method and the proposed progressive multi-block modelling approach to identify fault propagation path. Application results on a benchmark simulated fed-batch penicillin production process are presented in Section 4. The last section draws some concluding remarks.

2. Background

2.1 PCA for batch process monitoring

Batch process operation data are generally stored as three-dimension data arrays, where the three dimensions are batch numbers, variables and sample times. Such data are transformed to two dimension data structure which can be analysed using PCA. Multi-way PCA method is the most popular way to apply the batch process data into the PCA model. It has a step called data unfolding which transforms the three dimensional batch data into two a dimensional data structure. There are two popular ways for data unfolding: batch-wise unfolding proposed by Nomikos and Macgregor (1994; 1995) and variable-wise unfolding proposed by Wold et al. (1998). Under batch-wise unfolding, the data from an entire batch is transformed into a long row in the unfolded data matrix. With variable-wise unfolding, each column of the unfolded data matrix represents a variable. In this paper, the batch-wise unfolding method is used to remove non-linearity in the data as PCA algorithm itself is a linear method.

Consider an unfolded data matrix $\mathbf{X}$ with a size of $r \times c$. Each row of $\mathbf{X}$ represents a batch and a column represents a data point for each variable measured at each sampling time. A PCA model describes $\mathbf{X}$ as the sum of outer product of vectors, $\mathbf{t}$ and $\mathbf{p}$. Thus, $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T = \sum_{i=1}^c \mathbf{t}_i \mathbf{p}_i^T \quad (1)$$

In Eqn. 1, $\mathbf{t}_i$ is the i th score vector and $\mathbf{p}_i$ is the i th loading vector. The reconstruction of $\mathbf{X}$ using the first n ($n \leq c$) principal components can be calculated as

$$\hat{\mathbf{X}} = \mathbf{T}_n \mathbf{P}_n^T = \sum_{i=1}^n \mathbf{t}_i \mathbf{p}_i^T \quad (2)$$

As the loading matrix is obtained from the normal data, it describes correlations among the variables under the normal operating condition. When a fault presents, it will change the correlation structure among the process variables and/or change the magnitudes of some process variables. These will be detected by the SPE and/or the T^2 statistics respectively.

2.2 Multi-block Methods

Multi-block PCA modelling is an alternative way of the conventional PCA modelling for improved monitoring analysis and efficiency. Many batch chemical processes have several phases and different characteristics and data correlations may present in these phases (Lu and Gao, 2005; Yao and Gao, 2009). In the conventional MPCA with batch-wise unfolding, only one score value is calculated to describe one batch and it would be difficult to describe a whole system containing multiple characteristics, because the score is calculated from the one loading set containing different types of correlations. Multi-block methods can be considered to model multi-phase/stage behaviour of the process because multiple scores are calculated for a batch from the sub-groups (phases/stages) (Lu and Gao, 2005; Yao and Gao, 2009). It means that a score for a sub-block is calculated using the loading set representing the correlations only for the given sub-block. Thus, it can represent its sub-groups better than the score from a single data block. It basically divides the data into multiple sub-groups for the divided groups to have their own correlation structures. Therefore, each sub-block has its block loadings and scores.

There are several published multi-block methods. The first method called consensus PCA (CPCA) was proposed by Wold et al. (1987). This CPCA algorithm is based on the ordinary PCA algorithm. A matrix of super scores of CPCA is the same as the scores of ordinary PCA, so it does not give more information than ordinary PCA (Westerhuis et al., 1998; AlGhazzawi and Lennox, 2008). The only difference is whether data is arranged with multiple blocks or not. The algorithm of CPCA is shown below.

Algorithm of CPCA (in the case of 2 sub-blocks):

1. Set initial super scores, $\mathbf{t}_3$
2. Calculate block loadings: $\mathbf{p}_1 = \mathbf{X}_1^T \mathbf{t}_3 / (\mathbf{t}_3^T \mathbf{t}_3)$, $\mathbf{p}_2 = \mathbf{X}_2^T \mathbf{t}_3 / (\mathbf{t}_3^T \mathbf{t}_3)$
3. Normalised block loadings, $\mathbf{p}_1$ and $\mathbf{p}_2$, to unit length.
4. Calculate block scores: $\mathbf{t}_1 = \mathbf{X}_1 \mathbf{p}_1$, $\mathbf{t}_2 = \mathbf{X}_2 \mathbf{p}_2$
5. Combine block scores into one block $\mathbf{T}$: $\mathbf{T} = [\mathbf{t}_1 \mathbf{t}_2]$
6. Calculate super weight $\mathbf{w}_3$: $\mathbf{w}_3 = \mathbf{T}^T \mathbf{t}_3 / \mathbf{t}_3^T \mathbf{t}_3$
7. Normalise $\mathbf{w}_3$ to unit length.
8. Calculate super score $\mathbf{t}_3$: $\mathbf{t}_3 = \mathbf{T} \mathbf{w}_3$
9. Go back to step 2 until $\mathbf{t}_3$ is converged.
10. Obtain residuals: $\mathbf{E}_1 = \mathbf{X}_1 - \mathbf{t}_3 \mathbf{p}_1^T$, $\mathbf{E}_2 = \mathbf{X}_2 - \mathbf{t}_3 \mathbf{p}_2^T$
11. Replace $\mathbf{X}_1$ and $\mathbf{X}_2$ by $\mathbf{E}_1$ and $\mathbf{E}_2$ to compute the next principal component.
12. Check if the required model accuracy or the number of required PC is reached, if yes, stop the calculation, if no, go back to step 1.

By using the CPCA method, it is possible to analyse each block which may be able to find out where the fault comes from (AlGhazzawi and Lennox, 2008). Another algorithm proposed by Chen and McAvoy (1997) is called multi-block PCA (MBPCA). In this method, neither block loadings nor block scores are normalised. It is firstly performed by the ordinary PCA model, and then the obtained loadings are divided in the same order as data $\mathbf{X}$ is separated. After that, new block scores and loadings are calculated and, by using those, residuals are obtained. This algorithm is very similar to multi-block PLS (MBPLS) algorithm with deflation by block scores as it also deflates by using block scores (Qin et al., 2001). The computation procedure of MBPCA can be outlined below.

Algorithm of MBPCA (in the case of 2 sub-blocks):

1. Set initial super scores, $\mathbf{t}_3$
2. Compute super loadings: $\mathbf{p}_3 = \mathbf{X}_1^T \mathbf{t}_3 / (\mathbf{t}_3^T \mathbf{t}_3)$
3. Normalised super loadings, $\mathbf{p}_3$ to unit length.
4. Separate $\mathbf{p}_3$ into two sub-blocks ($\mathbf{p}_1$ and $\mathbf{p}_2$) in the same order as $\mathbf{X}$ is blocked.
5. Calculate block scores: $\mathbf{t}_1 = \mathbf{X}_1 \mathbf{p}_1$, $\mathbf{t}_2 = \mathbf{X}_2 \mathbf{p}_2$
6. Obtain block loadings: $\mathbf{p}_1 = \mathbf{X}_1^T \mathbf{t}_1 / (\mathbf{t}_1^T \mathbf{t}_1)$, $\mathbf{p}_2 = \mathbf{X}_2^T \mathbf{t}_2 / (\mathbf{t}_2^T \mathbf{t}_2)$
7. Go back to step 2 until block scores are converged.
8. Obtain residuals: $\mathbf{E}_1 = \mathbf{X}_1 - \mathbf{t}_1 \mathbf{p}_1^T$, $\mathbf{E}_2 = \mathbf{X}_2 - \mathbf{t}_2 \mathbf{p}_2^T$
9. Replace $\mathbf{X}_1$ and $\mathbf{X}_2$ by $\mathbf{E}_1$ and $\mathbf{E}_2$ to compute the next principal component.
10. Check if the required model accuracy or the number of required PC is reached, if yes, stop the calculation, if no, go back to step 1.

CPCA and MBPCA model can produce independent monitoring chart for each sub-block using the block scores which is useful to find the location of the detected fault. As mentioned before, block scores in multi-block methods can be advantageous in delivering enhanced fault diagnosis and analysis performance than the conventional PCA method containing one score matrix representing the behaviour of the monitored batch through its whole batch duration as block scores can express their own behaviour better than the MPCA score. CPCA uses the super scores for deflations for sub-blocks. As super score matrix for the testing batch is designed to describe all existing behaviours for the batch through whole processing time, deflations of sub-blocks by using the super score matrix can be affected by variations from the other sub-blocks. This could make the CPCA method not to provide appropriate monitoring results for sub-blocks. Consequently, CPCA monitoring result would be same as that of the MPCA. It just has a MPCA monitoring chart divided into multiple monitoring charts for sub-blocks. However, in case of MBPCA method, block scores are used for deflation of sub-blocks. It can lead to better monitoring performance as the block scores only consider the behaviours for its own sub-block and is not affected by the behaviours from other sub-blocks. It is more like two different MPCA models running at same time. For example, if a process containing two phases has a problem in the 2nd phase, CPCA and MBPCA modelling methods can be built with two sub-blocks of

data separated by different phases. As CPCA uses the super score matrix for deflation of all sub-blocks as mentioned before, the fault is detected from monitoring charts of all sub-blocks which is not appropriate analysis as the fault is actually existed only in the 2nd sub-block. In contrast, MBPCA method using the blocks scores for the deflation step, fault can be detected from the monitoring chart for the 2nd sub-block only. However, MBPCA method produces block loadings by dividing the normalised super loading according to same order as the data is divided, and it can make the MBPCA method to be not good enough for fault analysis. The loadings represent correlation structure among the variables through the process time, therefore, for enhanced fault diagnosis/analysis performance, block loadings are very important as they should contain correlations for certain time period of sub-blocks. As block loadings for MBPCA method are just sub-sets of the super loadings, correlations in block loadings are already influenced each other and they are not actually specified for their own sub-blocks.

3. Progressive Modelling based on a New Multi-block PCA Method

3.1 A New Multi-block PCA Method

A new multi-block PCA model is proposed to achieve enhanced monitoring performance using the progressive modelling method. The procedure is very similar to the MBPCA modelling method. However, the proposed method does not compute any information matrix on super level. It computes block loadings using data from sub-blocks not using the super loadings whereas the MBPCA method uses information matrix on super level to compute both block loadings and super loadings. The super score matrix can only provide overall monitoring information for the whole process time which is not focused on a specific behaviour. Thus, the proposed method does not involve super score or super loading calculation step. Block loadings calculation from sub-block data sets can produce correlation structures specific to the behaviour in sub-blocks. Therefore, the proposed method can compute block loadings contain more accurate information for each block than the conventional multi-block methods. Figure 1 briefly compares the CPCA, MBPCA and the proposed MBPCA methods.

(a)

(b)

(c)

Figure 1. Brief description of three different multi-block methods: a). CPCA; b). MBPCA; c). the proposed MBPCA

The CPCA method uses super scores to obtain block loadings whereas the MBPCA method separates super loadings into the same order as the original data is divided to form block loadings. These two methods use information on super level to obtain information on sub-block level. This might affect the analysis results as information

on super level is not specific for certain behaviour. Better results might be achieved if only information specified to certain process behaviours is used for model building. Therefore the proposed method does not compute super level information but only uses sub-block data to compute sub-block scores and loadings. Thus, it is very similar to MPCA modelling, but it can produce multiple MPCA monitoring charts rather than one monitoring chart from the conventional MPCA model for easy fault localisation. Data can be separated by various criteria such as different operation units, variables, or processing time. To determine how data to be divided, process knowledge and the purpose of modelling should be considered. In this study, modelling is carried out for fault analysis to find out dynamics of how the fault affects the variables over processing time so that data is divided by operation time. A brief description of the calculation steps is given below.

Algorithm of the proposed new multi-block PCA (in the case of 2 sub-blocks):

1. Perform PCA for each sub-block, $\mathbf{X}_1$ and $\mathbf{X}_2$
 - 1.1 Set initial block scores, $\mathbf{t}_1$ and $\mathbf{t}_2$ by selecting the first column of $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively.
 - 1.2 Calculate block loadings: $\mathbf{p}_1 = \mathbf{X}_1^T \mathbf{t}_1 / (\mathbf{t}_1^T \mathbf{t}_1)$, $\mathbf{p}_2 = \mathbf{X}_2^T \mathbf{t}_2 / (\mathbf{t}_2^T \mathbf{t}_2)$
 - 1.3 Normalised block loadings, $\mathbf{p}_1$ and $\mathbf{p}_2$, to unit length.
 - 1.4 Calculate block scores: $\mathbf{t}_1 = \mathbf{X}_1 \mathbf{p}_1$, $\mathbf{t}_2 = \mathbf{X}_2 \mathbf{p}_2$
 - 1.5 Go back to step 1.2 until $\mathbf{t}_1$ and $\mathbf{t}_2$ have both converged.
2. Obtain residuals: $\mathbf{E}_1 = \mathbf{X}_1 - \mathbf{t}_1 \mathbf{p}_1^T$, $\mathbf{E}_2 = \mathbf{X}_2 - \mathbf{t}_2 \mathbf{p}_2^T$
3. $\mathbf{X}_1 = \mathbf{E}_1$, $\mathbf{X}_2 = \mathbf{E}_2$
4. Check if the required model accuracy or the number of required PC is reached, if yes, stop the calculation, if no, go back to step 1

3.2 Progressive Modelling with Multi-block PCA

The progressive modelling described in (Hong et al., 2011) is applied with multi-block PCA to obtain all the variables influenced by the fault as procedure is described in Figure 2. This procedure allows all variables influenced by the fault to be identified. Mean trajectories from the normal data is used to replace the future data point for on-line monitoring. So that this proposed procedure can be carried on before the batch is completed.

Figure 2. Flow diagram of progressive MBPCA modelling

When a faulty batch is detected from the SPE and/or T^2 statistics at a certain time point on a certain sub-block, process variables on that sub-block showing high contribution to the detected fault are identified from the contribution plots at that time. These variables are then removed from that sub-block and another MBPCA model is built with the remaining variables. If the model built with the reduced variables can still detect the fault, SPE and/or T^2 contribution plot is again used to identify more variables related to the fault in that sub-block. This procedure is iterated until no fault is detected on both monitoring charts for all sub-blocks. Then time series SPE plots for the identified variables are produced to get time information when abnormalities are observed which is used to obtain the fault propagation path through the variables. As the model has sub-groups, this procedure is applied for each sub-group.

4. Application to a Simulated Fed-batch Penicillin Production Process

4.1 Process and data

The proposed method is applied to a benchmark simulated fed-batch penicillin production process. The fed-batch fermentation process simulation software, PenSim v2.0, was developed by Birol et al. (2002) at Illinois Institute of Technology and is used to produce both normal operating data and faulty data. For simplicity, every batch runs for 400hrs to have the same batch length. There are 11 variables measured every half hour. Table 1 lists the variables.

Table 1. Process variables

Number	Variables
1	Aeration rate (L/h)
2	Agitator power (W)
3	Substrate feed rate (L/h)
4	Substrate feed temperature (K)
5	Dissolved oxygen (mmol/L)
6	Culture volume (L)
7	CO ₂ concentration (mmol/L)
8	pH
9	Temperature (K)
10	Generated heat (kcal/h)
11	Cold water flow rate (L/h)

This process runs in batch mode from the beginning until the amount of biomass reaches certain level and it then changes to fed-batch mode by having the main reaction with continuous feeding of substrate into the reactor. Batch mode normally operates during the first 45 hours when the total batch duration is 400 hours. Sub-blocks are formed by these two different operating modes. Sub-block 1 contains data collected during batch mode running whereas sub-block 2 contains data from fed-batch mode. Multi-block PCA models were built using 56 normal batches and an additional 15 normal batches were used to determine the appropriate number of principal components. Three faulty batches shown in Table 2 are studied using the proposed method. Data is unfolded by using the batch-wise unfolding method and auto-scaled before model building. Zero-deviation is applied to fill the missing future samples for on-line monitoring.

Table 2. Three faulty batches

Fault No.	Variable that disturbance is introduced
1	Variable #3 (substrate feed rate): -0.03%, ramp, 160h ~ 250h
2	Variable #3 (substrate feed rate): -10%, step, 205h ~ end
3	Variable #8 (pH): pH controller failure, 0.5h ~ end

4.2 Results and discussions

Figures 3 and 4 show the monitoring charts by the CPCA model applied with the normal validation data on both sub-blocks. On both monitoring charts, points indicate normal training data whereas crosses represent normal validation data. As shown in the figures all training data and validation data points are locating inside of the confidence bounds when 1 PC is used for both sub-blocks in the CPCA model. This is the maximum number of PCs that keeps all normal batch data inside of the confidence bounds, thus in this study, a CPCA model built with 1 PC on all sub-blocks is used to monitor the process. MBPCA uses 6 PCs for the 1st sub-block and 13 PCs for the 2nd sub-block whereas the proposed new MBPCA uses 1 PC for both the 1st and the 2nd sub-blocks.

Figure 3. Monitoring Charts for normal validation batches on sub-block 1

Figure 4. Monitoring charts for normal validation batches on sub-block 2

For on-line monitoring, the future data to the end of the batch need to be estimated. There are several ways to fill in the future data such as zero-variance method, current variance method and prediction. For this case study, zero-variance method is used to generate unknown future data. This method simply uses the mean trajectory of the reference batches to fill unknown data positions.

4.2.1 Fault Case 1: Abnormal change on substrate feed rate (-0.03%, Ramp)

The first faulty batch has a problem on substrate feed rate. The substrate feed rate is gradually decreased by -0.03% per hour. The lower substrate feed rate leads to less cell mass concentration than required and it results in higher DO level and lower carbon dioxide concentration and eventually lower penicillin concentration (Birol et al., 2002). On-line CPCA can detect the fault on the Q statistics monitoring chart of sub-block 2 from operation time of 200.5h as shown in Figure 5. However to avoid false alarm, the time point of 201.5h is used to obtain a SPE contribution plot to see which variable is highly responsible for this result. Both monitoring charts in block 1 are displayed from the batch beginning whereas the monitoring charts in the 2nd block are displayed from 43.5h. This is because the monitoring charts for the 1st block and the 2nd block are computed using data from the batch beginning to 43h and from 43.5h to the end of the process, respectively.

Figure 5. On-line monitoring charts for 1st faulty batch (CPCA model)

The MBPCA model was also used in the progressive modelling scheme for comparison purpose. Monitoring charts of the MBPCA method are displayed in Figure 6. As shown in the figure, no point locating outside of the confidence limits was observed so that MBPCA did not detect the 1st fault. However, T^2 for the 2nd block is gradually increasing. Thus, it is expected that the fault can be detected if the process runs a bit longer.

Figure 6. On-line monitoring charts for the 1st faulty batch (MBPCA model)

Figure 7. On-line monitoring charts for the 1st faulty batch (new MBPCA model)

Figure 7 shows the monitoring charts from the proposed new MBPCA method. The Q statistics from the 2nd block exceed its control limit at time 203 h. The SPE contribution plot at 204 h is generated to identify the related variables.

Figure 8 shows the SPE contribution plot at 201.5 h of sub-block 2 under the CPCA model. The dashed line in Figure 8 indicates the 99% control limits for the contributions (Hong et al., 2011). As monitoring charts for sub-block 1 do not detect any fault, it can be considered that the fault did not occur when the process was running under batch mode but it occurred in the fed-batch mode. Therefore there is no need to consider data from batch mode for fault analysis and fault diagnosis can be rapidly focused to sub-block 2. In the SPE contribution plot for sub-block 2, two variables, substrate feed rate (No. 3) and dissolved oxygen (No. 5), are identified as its SPE value is higher than its confidence limit.

Figure 8. SPE contribution plot from the on-line CPCA model

Figure 9 shows the SPE contribution plot obtained from the new MBPCA model at 204 h. Two variables (No. 3 and No. 5) are identified in sub-block 2 having high contributions (exceeding their 99% control limit). To find out time information when abnormal behaviour is observed for each identified variable, the time series SPE plots are produced. If five successive points locating outside of the confidence limits are observed, it is then considered that this profile represents abnormal behaviour otherwise it is considered as false alarm.

Figure 9. SPE contribution plot from the on-line new MBPCA model

As both CPCA model and the new MBPCA model identified the same two variables, these variables were removed and another model was build. Both CPCA and the new MBPCA models use the same variables. Another CPCA model built with the remaining variables for on-line monitoring detects the fault at 227.5h. The monitoring charts and the SPE contribution plot at this time point are displayed in Figures 10 and 11 respectively. Three variables can be identified at this time by looking at the SPE contribution plot. The time series SPE plots for those variables can provide time information when abnormal changes are observed. The new MBPCA model built with the remaining variables also detects the fault at 229.5h point and identifies three variables on the SPE contribution plot. Monitoring charts for the new MBPCA model and SPE contribution plot are displayed in Figures 12 and 13 respectively. Progressive modelling using CPCA and the new MBPCA method show very similar results but the time information obtained from the time series SPE plots is different as shown in Table 4.

Figure 10. On-line monitoring charts for the 1st faulty batch (CPCA, 2nd Step)

Figure 11. SPE contribution plot for the 1st faulty batch (CPCA, 2nd Step)

Figure 12. On-line monitoring charts for the 1st faulty batch (new MBPCA, 2nd step)

Figure 13. SPE contribution plot for the 1st faulty batch (new MBPCA, 2nd Step)

These three variables are then removed and a new CPCA model and a new MBPCA model are built with the remaining variables to see if there are any other variables having abnormality. Both newly built models with the remaining variables did not

detect the fault as shown in Figures 14 and 15. Thus progressive CPCA modelling scheme and progressive new MBPCA modelling scheme are finished and all the variables associated with the fault have been identified. Time series SPE plots for the identified variables are then produced to gain time information on when the abnormal behaviours are observed as shown in Figure 16.

Figure 14. On-line monitoring charts for the 1st faulty batch (CPCA, 3rd step)

Figure 15. On-line monitoring charts for the 1st faulty batch (new MBPCA, 3rd step)

(a) Substrate Feed Rate

(b) Dissolved Oxygen

(c) Culture Volume

(d) Generated Heat

(e) Cold Water Flow Rate

Figure 16. Time series SPE plots for the identified variables by the progressive CPCA model (1st Fault)

From Figure 16, the time points when the SPE values start locating outside of the control limits for the identified variables can be worked out and the same approach is carried out for the new MBPCA based modelling scheme. The obtained results are

summarised in Table 3 where the times when the corresponding variables are detected to be abnormal are provided.

Table 3. On-line monitoring results for the 1st faulty batch

Variable \ Model	CPCA	MBPCA	New MBPCA
Variable #3 Substrate feed rate (L/h)	164h	Not Detected (N/D)	164h
Variable #5 Dissolved oxygen (mmol/L)	179.5h		179.5h
Variable #6 Culture volume (L)	190.5h		193h
Variable #10 Generated heat (kcal/h)	182h		183.5h
Variable #11 Cold water flow rate (L/h)	182h		183.5h

As shown in Table 3, CPCA and the new MBPCA can identify the same variables. The MBPCA method does not detect any abnormality on monitoring charts. The progressive modelling approach with different multi-block methods can produce time information. CPCA and the new MBPCA based progressive modelling produced similar results as time information obtained from those two approaches for some identified variables are almost the same. In addition, sequence propagations produced based on the time information from the two approaches are the same as well. Figure 17 shows sequence produced by using time information from the new MBPCA model.

Figure 17. Fault propagation sequence for the 1st faulty batch obtained from the new MBPCA

This on-line progressive modelling procedure with three different modelling methods is also applied to other faulty batches.

4.2.2 Fault Case 2: Abnormal change on substrate feed rate (-10%, Step)

Table 4 shows the results of on-line modelling for the second faulty batch that has problem on variable #3 (substrate feed rate). It was suddenly reduced by 10% and kept at that level from 205h until the end of the process. Therefore, the sequence of behaviours occurred by this fault is expected to be similar with the one of the 1st fault. CPCA and the new MBPCA can give information that allows the appropriate analysis to be carried out. Sequence obtained from new MBPCA model is similar to the sequence for the 1st faulty batch. However, MBPCA model does not provide good results as the fault was not identified. Even though the time information obtained from CPCA and the new MBPCA method are slightly different, it can be observed that results from both methods are very similar as the fault propagation paths identified by both methods are the same which can well describe how the given fault affects the variables over the operating time. One big difference is time when abnormal behaviour of dissolved oxygen level is observed for these two approaches. It seems that CPCA based modelling scheme takes a bit long time to be able to observe abnormality of the variable of dissolved oxygen level. Based on time information obtained by the new MBPCA based progressive modelling scheme, fault propagation diagram can be obtained as shown in Figure 18.

Table 4. On-line monitoring results for the 2nd faulty batch

Variable \ Model	CPCA	MBPCA	New MBPCA
Variable #3 Substrate feed rate (L/h)	205.5h	Not Detected (N/D)	205.5h
Variable #5 Dissolved oxygen (mmol/L)	307.5h		275h
Variable #6 Culture volume (L)	224.5h		227.5h
Variable #10 Generated heat (kcal/h)	223h		214.5h
Variable #11 Cold water flow rate (L/h)	223h		215.5h

Figure 18. Fault propagation sequence for the 2nd faulty batch obtained from the new MBPCA

This fault propagation diagram provides information on how the fault can affect the variables during the process. In this case, substrate feed rate has been detected the earliest so that it can be considered as the cause variable. The faulty behaviour on substrate feed rate, which is the reduction of substrate flow rate into the reactor, reduces growth rate. As a result, culture volume and carbon dioxide concentration are decreased and the generated heat and cold water flow rate are also reduced due to lower activation energy and temperature control. The fault propagation diagram shown in Figure 18 can support explanation how the fault affects the process mentioned before (Birol et al., 2002).

4.2.3 Fault Case 3: pH controller failure

The 3rd faulty batch has a problem on variable #8 (pH) as a result of switching off the pH controller from the beginning of the batch. As the reaction tends to have lower pH level, it results a reduction of the mass growth rate leading to lower production rate (Birol et al., 2002). Table 5 shows the results by on-line progressive method with three different modelling methods. CPCA and new MBPCA identify eight variables as fault related variables whereas MBPCA model has only two variables identified. Both CPCA and new proposed MBPCA methods can deliver appropriate time information and it can support the description by Birol et al. (2002) whereas MBPCA method does not.

Table 5. On-line monitoring results for the 3rd faulty batch

Variable \ Model	CPCA	MBPCA	New MBPCA
Variable #3 Substrate feed rate (L/h)	47h	N/D	47.5h
Variable #4 Substrate feed temperature (K)	N/D	8.5h	N/D
Variable #5 Dissolved oxygen (mmol/L)	26.5h	N/D	29h
Variable #6 Culture volume (L)	3.5h	N/D	3.5h
Variable #7 CO ₂ concentration (mmol/L)	38.5h	N/D	39.5h
Variable #8 pH	2.5h	2.5h	2.5h
Variable #9 Temperature (K)	31h	N/D	30.5h
Variable #10 Generated heat (kcal/h)	21.5h	N/D	21.5h
Variable #11 Cold water flow rate (L/h)	39h	N/D	39.5h

Figure 19 displays the fault propagation path gained based on time information from the proposed MBPCA based progressive modelling scheme. It can be found that reduction of pH level causes reduction of the culture volume due to reduced mass growth rate as mentioned before. This change affects generated heat and dissolved oxygen level due to lower reaction activity. And it sequentially affects other identified variables as shown in Figure 19.

Figure 19. Fault propagation sequence for the 3rd faulty batch obtained from the new MBPCA

5. Conclusions

A multi-block progressive modelling scheme is proposed for the enhanced fault isolation in batch and fed-batch processes. As multi-block method can provide individual monitoring chart for each sub-block, fault isolation can be rapidly focused to blocks with detected abnormality. CPCA uses super score for deflation and prediction of sub-blocks and MBPCA uses block scores for deflation of sub-blocks. These two methods generate blocks loadings from the super loading and it might not provide clear correlation specific for each sub-block. The deflation step in CPCA use super score and this might not be good enough as super score represents general behaviour of the batch through the whole process time, not specific for sub-block behaviours. For better analysis performance, model should well describe correlation structure for each sub-block and block score should be used for deflation and prediction of each sub-block. The new multi-block method proposed here has no calculation regarding super level, as it is not required to express behaviours of each sub-block. But it has block loading and block score calculations using data from each sub-block to capture behaviour of each sub-block without interrupting by behaviour of other sub-blocks. Progressive modelling with different multi-block modelling methods is tested on the benchmark simulated fed-batch penicillin production process, PenSim. The new MBPCA method and CPCA method can deliver better results for all given cases than the MBPCA method.

Acknowledgement

The work is partially supported by Syngenta Ltd.

References

- AlGhazzawi, A., B. Lennox (2008). Monitoring a complex refining process using multivariate statistics, *Control Engineering Practice*, 16(3), 294-307.
- Birol, G., C. Undey and A. Cinar, (2002). "A modular simulation package for fed-batch fermentation: penicillin production", *Computers and Chemical Engineering*, 26, 1553-1565.
- Chen, G. and McAvoy, T.J., (1997). "Multi-block predictive monitoring of continuous processes", *IFAC ADCHEM International Symposium on Advanced Control of Chemical Processes*, Banff, Canada, 73-77.
- Chiang, L.H., Russell, E.L., Braatz, R.D., (2001). *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag London Limited.
- Gallagher, N.B., Wise, B.M., Stewart, C.W., (1996). "Application of multi-way principal components analysis to nuclear waste storage tank monitoring", *Computers and Chemical Engineering*, 20, S739-S744
- Hong, J.J., Zhang, J., Morris, A.J., (2011). "Fault Localisation in Batch Processes through Progressive PCA Modelling", *Ind. Eng. Chem. Res.*, Vol.50(13), 8153-8162
- Kourti, T., Nomicos, P., MacGregor, J.F., (1995). "Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS", *Journal of Process Control*, 5(4), 277-284.
- Kourti, T., Lee, J., MacGregor, J.F., (1996). "Experiences with industrial applications of projection methods for multivariate statistical process control", *Computers and Chemical Engineering*, 20, S745-S750.
- Kourti, T. (2003). Multivariate dynamic data modelling for analysis and statistical process control of batch processes, start-ups and grad transitions, *Journal of Chemometrics*, 17, 93-109.
- Lennox, B., Hiden, H.G., Montague, G.A., Kornfeld, G., Goulding, P.R., (2000). "Application of multivariate statistical process control to batch operations", *Computers and Chemical Engineering*, 24, 291-296.

- Louwerse, D.J. and Smilde, A.K., (2000). "Multivariate statistical process control of batch processes based on three-way models", *Chemical Engineering Science*, 55, 1225-1235.
- Lu, N., F. Gao (2005). "Stage-based process analysis and quality prediction for batch processes", *Ind. Eng. Chem. Res.*, 44, 3547-3555.
- MacGregor, J.F. and Kourti, T., (1995). "Statistical process control of multivariate processes", *Control Engineering and Practice*, 3(3), 403-414.
- Martin, E.B., Morris, A.J., Zhang, J., (1996). "Process performance monitoring using multivariate statistical process control", *IEE Proceedings - Control Theory and Applications*, 143(2), 132-144.
- Miller, P.; Swanson, R. E.; Heckler, C. F. "Contribution plots: the missing link in multivariate quality control", *Fall Conf. of ASQC and ASA*, Milwaukee, WI, USA, 1993.
- Nomikos, P. and MacGregor, J.F., (1994). "Monitoring of batch processes using multi-way principal component analysis", *AIChE Journal*, 40, 1361-1375.
- Nomikos, P., and MacGregor, J.F., (1995a). "Multivariate SPC charts for monitoring batch processes", *Tecnometrics*, 37, 41-59.
- Nomikos, P. and MacGregor, J.F., (1995b). "Multi-way partial least squares in monitoring batch process", *Chemometrics and Intelligent Laboratory Systems*, 30, 331-340.
- Qin, S.J., Valle, S., Piovoso, M.J., (2001). "On unifying multiblock analysis with application to decentralized process monitoring", *Journal of Chemometrics*, 15, 715-742.
- Qin, S.J., (2003). "Statistical process monitoring: basics and beyond", *Journal of Chemometrics*, 17, 480-502.
- Westerhuis, J. A., T. Kourti, J. F. MacGregor (1998). Analysis of multiblock and hierarchical PCA and PLS models, *Journal of Chemometrics*, 12, 301-321.
- Wise, B.M., Gallagher, N.B., Butler, S.W., White, JR, D.D., Barna, G.G., (1999). "Comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process", *Journal of Chemometrics*, 13, 379-396.

- Wold, H., L. Antti, Lindgren, O. Jerker (1998). Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, 44, 175-185.
- Wold, S., Hellberg, S., Lundstedt, T., Sjoström, M., Wold, H., (1987). “PLS modeling with latent variables in two or more dimensions”, *Proceedings of Symposium on PLS Model Building: Theory and Application*, Frankfurt am Main, Germany, 23-25.
- Wold, S., Kettaneh, N., Tjessem, K., (1996). “Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection”, *Journal of Chemometrics*, 10, 463-482.
- Yao, Y., Gao, F. R. (2009). “A survey on multistage/multiphase statistical modeling methods for batch processes”, *Annual Reviews in Control*, 33, 172–183
- Yoo, C.K., Lee, J.M., Vanrolleghem, P.A, Lee, I.B., (2004). “On-line monitoring of batch processes using multiway independent component analysis”, *Chemometrics and Intelligent Laboratory Systems*, 71, 151-163.
- Zhang, Y., Edgar, T.F., (2006). “Bio-reactor monitoring with multiway-PCA and model based-PCA”, *AIChE Annual Meeting, San Francisco, CA*.