Industrial experiences with multivariate statistical analysis of batch process data. Chemom Intell Lab Syst

ARTICLE *in* CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS · APRIL 2006

Impact Factor: 2.32 · DOI: 10.1016/j.chemolab.2005.10.006

Riccardo Leardi
Università degli Studi di Genova
121 PUBLICATIONS **4,018** CITATIONS

Randy Pell
Dow Chemical Company
27 PUBLICATIONS **1,048** CITATIONS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 81 (2006) 109 – 119

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Industrial experiences with multivariate statistical analysis of batch process data

Leo H. Chiang [a,*], Riccardo Leardi [b], Randy J. Pell [c], Mary Beth Seasholtz [c]

[a] *The Dow Chemical Company, Core R&D, 2301 Brazosport Blvd., Freeport, TX 77541, U.S.A.*
[b] *University of Genoa, Department of Pharmaceutical and Food Chemistry and Technology, Via Brigata Salerno (ponte), I-16147 Genova, Italy*
[c] *The Dow Chemical Company, Core R&D, 1897 Building, Midland, MI 48642, U.S.A.*

Received 23 June 2005; received in revised form 19 October 2005; accepted 20 October 2005
Available online 20 December 2005

Abstract

The data collected from a batch process over time from multiple sensors can be arranged in a matrix of J-variables × K-time points. Data collected on multiple batches can be arranged in a cube of I-batches × J-variables × K-time points. The analysis of a cube of data can be performed by unfolding in two different ways, batch unfolding giving an $I \times JK$ data matrix or observation unfolding resulting in an $IK \times J$ data matrix, followed by PCA. The data can also be analyzed directly using three-way methods such as PARAFAC or Tucker3. In the literature there is no clear agreement as to the most effective approach for the analysis of batch data.

This paper makes detailed comparisons between the two unfolding approaches and the Tucker3 method. Batch data from a fermentation process at The Dow Chemical Company San Diego facility is used for this study. The three methods were found to be complementary to each other and a well-trained chemometrician/practitioner will find all three methods to be useful for batch data analysis. The batch unfolding MPCA is more sensitive to the overall batch variation while the observation unfolding MPLS is more sensitive to the localized batch variation. The Tucker3 method is in good balance in terms of detecting both variations.

Keywords: Multi-way analysis; Tucker3 analysis; Multi-way principal component analysis; Multi-way partial least squares; Batch process monitoring; Chemometrics; Fault detection

1. Introduction

To achieve consistent product quality from a batch process, minimizing batch-to-batch variability is important. In off-line applications, multivariate analysis techniques such as principal component analysis (PCA) or partial least squares (PLS) can identify and pinpoint the root causes of batch-to-batch variability. In on-line applications, these techniques are used to monitor batch conditions. The objective is to identify and correct abnormal conditions early enough to avoid out-of-specification product.

There are three commonly used approaches for the analysis of batch process data. The first approach is the pioneering work of Nomikos and MacGregor [1]. For a data set with I batches, J process variables, and K time points, a three-way $I \times J \times K$ cube of data is unfolded into a two-way $I \times JK$ data as illustrated in Fig. 1. PCA is applied to the unfolded data. This approach is referred to as batch unfolding multi-way PCA (MPCA) in this paper.

A modification of the batch unfolding MPCA approach was proposed by Wold et al. [2]. A three-way $I \times J \times K$ cube of data is unfolded into a two-way $IK \times J$ data as illustrated in Fig. 2. Then, a PLS model is developed to relate the unfolded data to a monotonically increasing/decreasing variable (i.e., maturity variable) that measures the percent completion of a batch. This analysis is referred to as observation unfolding multi-way PLS (MPLS) in this paper. This approach has been shown to be effective in improving batch quality of an industrial fermentation process at the San Diego biotech facility of The Dow Chemical Company [3].

The third approach belongs to the family of three-way methods, including PARAllel FACtor analysis (PARAFAC)

[*] Corresponding author. Tel.: +1 979 238 5377; fax: +1 979 238 0100.
E-mail address: hchiang@dow.com (L.H. Chiang).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.10.006

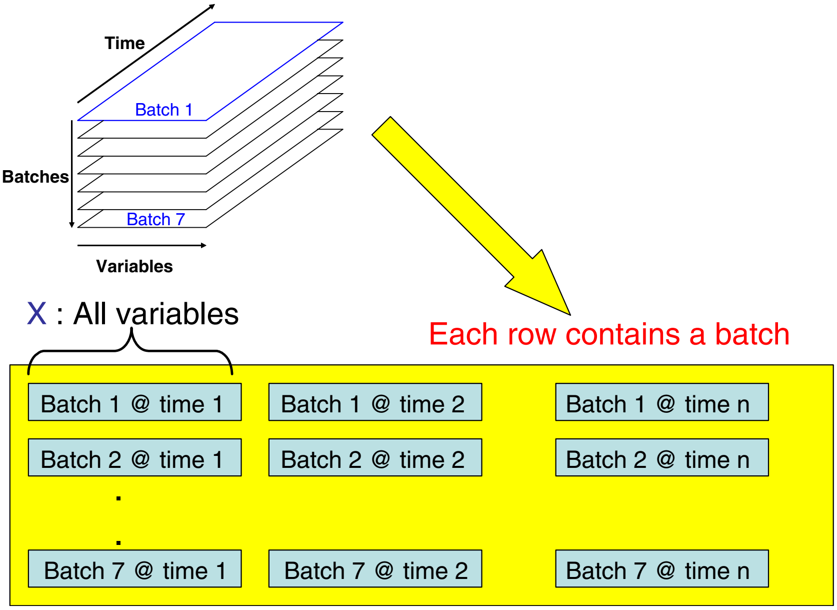

Fig. 1. Illustration of batch unfolding multi-way PCA analysis.

and the Tucker3 method [4–7]. In these approaches, unfolding is not required. Instead, a three-way cube of data is decomposed into three loading matrices (modes). Illustration of the Tucker3 method is shown in Fig. 3. Analysis of these three matrices often reveals useful batch information.

In this paper these three approaches are compared in detail using batch data sets from an industrial fermentation process. Multivariate statistical analysis has been applied successfully in various industrial bioprocesses for fault detection and diagnosis, and product quality prediction. Interested readers are referred to Albert and Kinley [8] for a tylosin biosynthesis process, Gregersen and Jorgensen [9] for a fed-batch fermentation process, Lopes et al. [10] for a pharmaceutical fermentation process, and Undey et al. [11] for a fed-batch penicillin cultivation process. Background material on process monitoring for fermentation can be found in Cinar et al. [12].

There have been comparisons made on the unfolding methods and the three-way methods. Westerhuis et al. [13] compared PARAFAC, Tucker3, observation unfolding MPLS, and batch unfolding MPCA on two data sets and concluded that batch unfolding MPCA is the preferred method for batch analysis. Wise et al. [14] concluded that PARAFAC performed slightly better than batch unfolding MPCA for on-line fault detection of a semiconductor etch process. Louwerse and Smilde [15] and Smilde [5] outlined

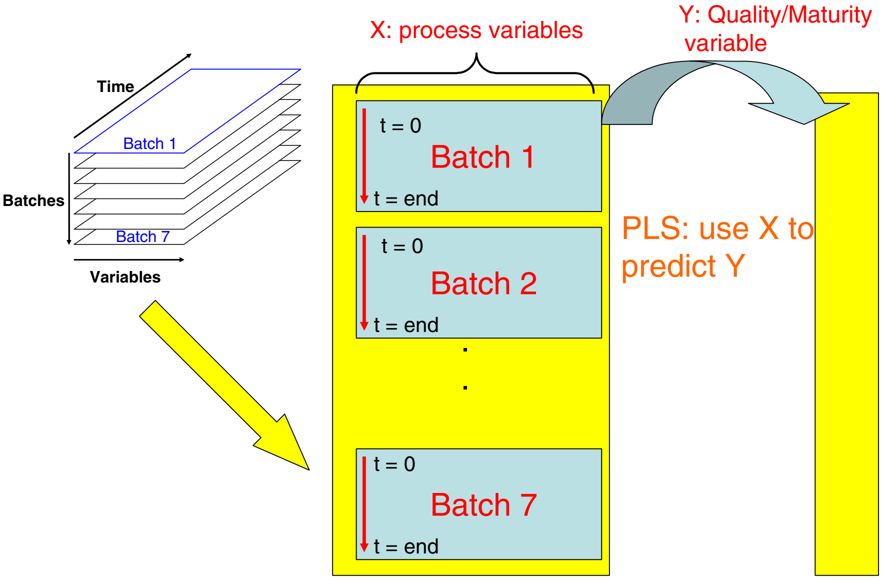

Fig. 2. Illustration of observation unfolding multi-way PLS analysis.

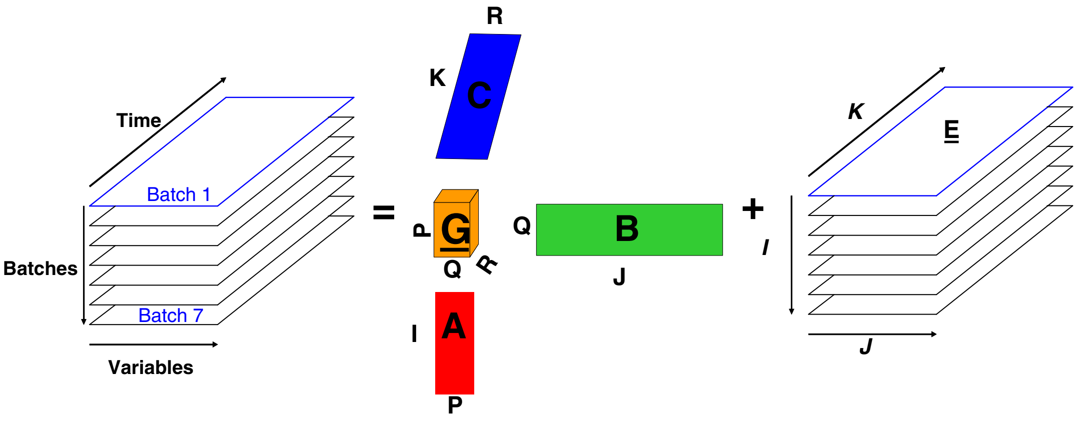

Fig. 3. Illustration of the Tucker3 analysis.

theoretical aspects and comparative results of Tucker3, PARAFAC, batch unfolding PCA and determined that no clear conclusions can be drawn on the most effective method for batch analysis. Van Sprang et al. [16] observed comparable results for on-line fault detection using batch unfolding MPCA and observation unfolding MPLS for five data sets. Kourti [17] pointed out a potential problem of observation unfolding MPLS when process variables are not correlated with maturity variable and suggested that batch unfolding MPCA properly identified batch-to-batch variation. Undey et al. [18] commented that observation unfolding MPLS is advantageous for on-line process monitoring, while batch unfolding MPLS is useful for end-of-batch product quality prediction.

In addition to these three approaches, other multivariate batch analysis methods have been proposed recently, including moving window PCA [19], batch dynamic PCA [20], time-varying state space modeling [21], multi-way kernel PCA [22], independent component analysis [23], and stage-based PCA [24]. It is clear that researchers have not come to an agreement as to which approach is most effective for batch analysis. The purpose of this paper is to illustrate the usefulness of the three main approaches for solving an industrial batch process problem.

2. Method

2.1. Batch unfolding MPCA

In batch unfolding MPCA (see Fig. 1), the rows of the unfolded **X** matrix represent the batches. The model is expressed as:

$$\mathbf{X}^{I\times JK} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E}$$

where **T** is a score matrix of $I\times R$, R is the rank, **P** is a loading matrix of $(JK\times R)$, and **E** is a residual matrix of $I\times JK$ [1]. Additional batch information such as initial batch condition or final product quality can be incorporated with a PLS model:

$$\boldsymbol{y}^{I\times 1} = \mathbf{T}\boldsymbol{c} + \boldsymbol{f}$$

where $\boldsymbol{c}$ is an inner coefficient vector of $R\times 1$ and $\boldsymbol{f}$ is a residual vector of $I\times 1$.

Note that the score calculation can only be performed when the process variables for the entire batch are measured. This is attractive for off-line analysis because the score represents the overall batch-to-batch variation. However, this poses a problem for on-line implementation. Garcia-Munoz et al. [25] demon-

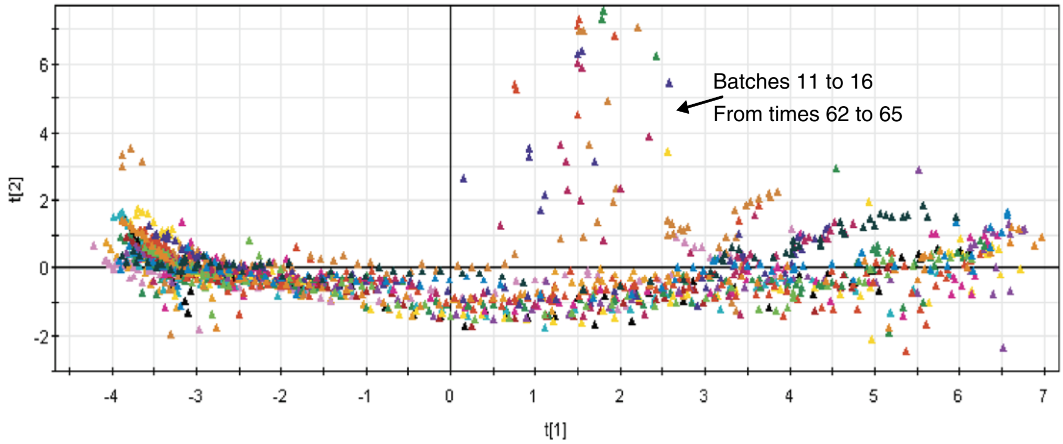

Fig. 4. Observation unfolding MPLS score plot for all variables and all batches. Each colored trajectory summarizes the dynamics of a given batch. If all batches are consistent, then all trajectories will overlap.

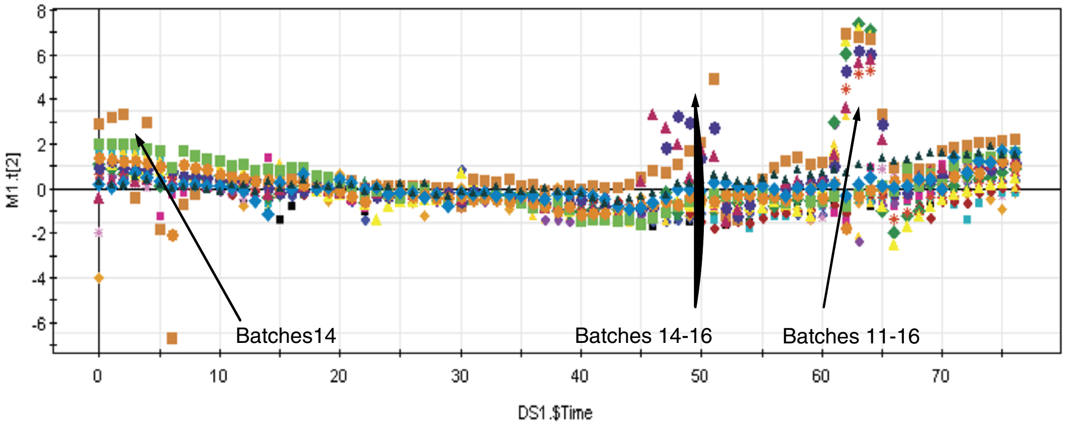

Fig. 5. Observation unfolding MPLS second principal component versus elapsed time for all variables and all batches.

strated that missing value estimation is effective in resolving this issue.

2.2. Observation unfolding MPLS

As illustrated in Fig. 2, the observation unfolding MPLS model is written as:

$$\mathbf{X}^{KI\times J} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}$$

where **T** is a score matrix of $KI\times R$, **P** is a loading matrix of $J\times R$, and **E** is a residual matrix of $KI\times J$ [2]. The unfolded **X** matrix is then related to the maturity/quality variable (a monotonically increasing/decreasing variable that measures the percent completion of a batch) using PLS:

$$\boldsymbol{y}^{KI\times 1} = \mathbf{T}\boldsymbol{c} + \boldsymbol{f}$$

where $\boldsymbol{c}$ is an inner coefficient vector of $R\times 1$ and $\boldsymbol{f}$ is a residual vector of $KI\times 1$.

Note that the score calculation can be performed when the process variables are measured at any given time. This is attractive for on-line implementation because an updated model prediction is obtained based on current process measurements. The T^2 and Q statistics contribution charts can also be used to identify process faults on-line [26]. It is preferable to use a quality variable as a maturity variable. If this is not possible, batch elapsed time can be used instead. Initial batch conditions and any final batch quality variables can be incorporated using batch unfolding MPLS with **X** block either as a $I\times JK$ matrix of raw measurements or a $I\times RK$ matrix of scores from observation unfolding MPLS.

2.3. Tucker3 analysis

Unlike the multi-way analysis, matrix unfolding is not required in the Tucker3 analysis. Illustration of the Tucker3 analysis is shown in Fig. 3, in which a batch data set (I batches × J variables × K time points) is decomposed in three loading matrices (matrix **A** for batch mode: $I\times P$; matrix **B** for variable mode: $J\times Q$, and matrix **C** for time mode: $K\times R$), one core matrix **G** (factors: $P\times Q\times R$), and one three-way residual matrix **E** ($I\times J\times K$) [4,27]. Mathematically, the Tucker3 analysis can be expressed as:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk}.$$

Analysis of the three loading matrices **A**, **B**, and **C** often reveals useful batch information. Because all three modes are reduced, this analysis is commonly referred to as Tucker3 (P, Q, R) model. When only batch mode is reduced, Tucker3 (P, Q, R) model becomes Tucker1 (P) model, which is mathematically the same as batch unfolding MPCA. In other words, Tucker3 (P, Q, R) model is a restricted Tucker1 (P) model. As such, a better fit is always observed in batch unfolding MPCA than the Tucker3 model.

3. Experimental

3.1. Application

The multivariate analysis was performed on 20 fermentation batches from The Dow Chemical Company San Diego facility. The key objective was to eliminate problematic fermentation issues that prevent 100% fermentation success rate. For the fermentation process, 17 variables were measured every minute. A fifteen-minute average was applied to the data set, resulting in a matrix 20 batches × 17 variables × 77 time points. A quality variable (optical density, OD) was measured during fermentation every few hours. This quality variable, measure of the dynamics and the maturity of cell growth, will be used as the Y-block in observation-unfolding MPLS. Linear interpolation was applied to obtain OD measurement every 15 min. Duration of each batch varies and all batches were cut to the minimum length for batch alignment.

Table 1
Summary statistics of the observation unfolding PLS model

PLS rank	$R^2_{X,\text{ cum}}$ (%)	$R^2_{Y,\text{ cum}}$ (%)	$Q^2{}_{\text{cum}}$ (%)
1	62.4	94.4	94.3
2	70.9	96.1	96.0
3	75.7	97	96.8
4	79.7	97.4	97.2

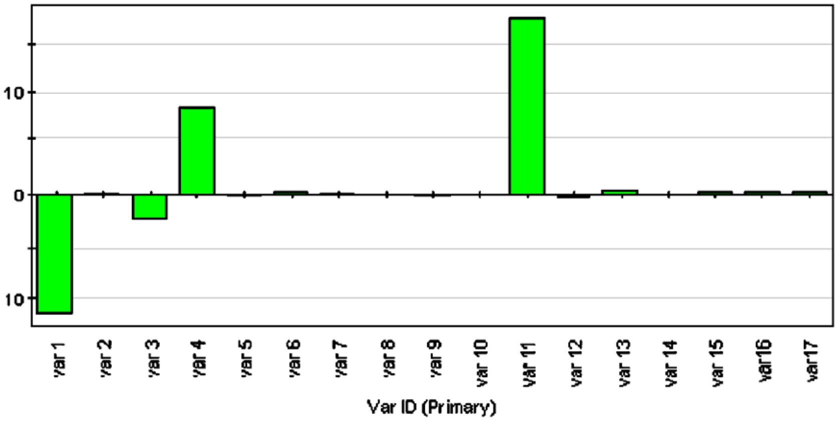

Fig. 6. $\mathbf{T}^2$ contribution plot for batch 14 at time 63.

3.2. Pre-processing

There are two commonly used autoscaling methods for batch analysis, namely *j*-scaling and *jk*-scaling.

3.2.1. j-scaling

In *j*-scaling, the three-way batch data are first unfolded as illustrated in Fig. 2. Then, autoscaling is applied to the unfolded data. With *j*-scaling, the differences between variables have been removed (i.e., each variable has zero mean and unit variance), while the differences between batches and the differences between times have been preserved (i.e., the average of each variable computed on all the batches at the same time will be different from 0). In other words, the dynamic behavior of the variables is retained. This scaling is commonly used in observation unfolding MPCA.

3.2.2. jk-scaling

In *jk*-scaling, three-way batch data are first unfolded as illustrated in Fig. 1. Then, autoscaling is applied to the unfolded data. With *jk*-scaling, the differences among the variables at any time are removed. This means that the “ideal” batch will be made by a vector of zeros. The focus of this scaling is on the differences among batches. This scaling is commonly used in batch unfolding MPCA.

3.3. Software

The observation unfolding multi-way analysis was completed using Umetrics SIMCA-P version 10 software [28]. The batch unfolding multi-way analysis was completed using BatchSPC version 2.0 software [29]. The Tucker3 analysis was completed using Matlab™ code written by the authors.

4. Results and discussion

4.1. Observation unfolding MPLS

To identify consistency among batches using observation unfolding MPLS, the score plot is used (see Fig. 4). The first two principal components capture 71% of the total variation,

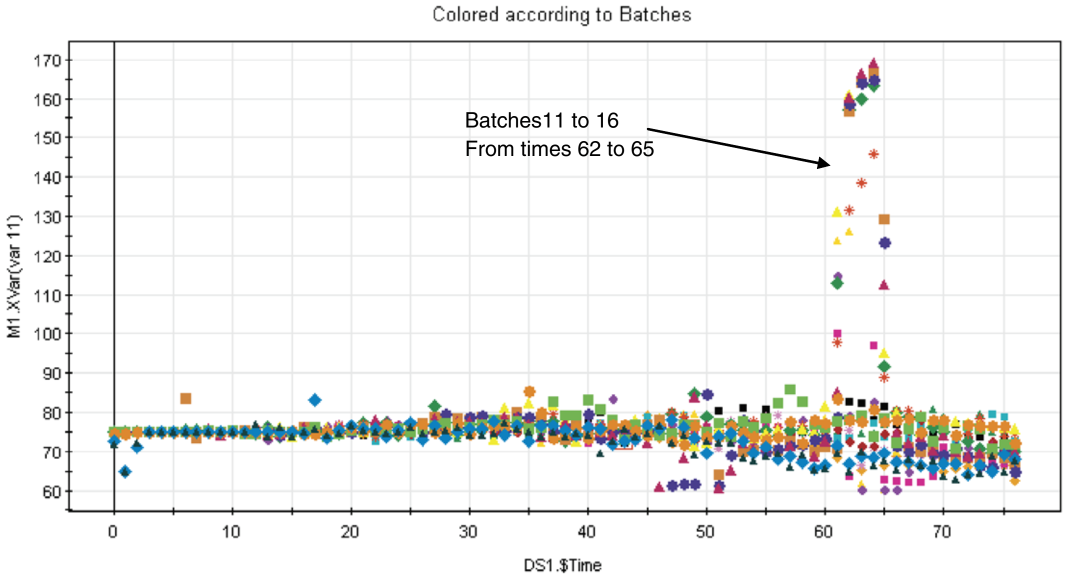

Fig. 7. Time series plot of variable 11.

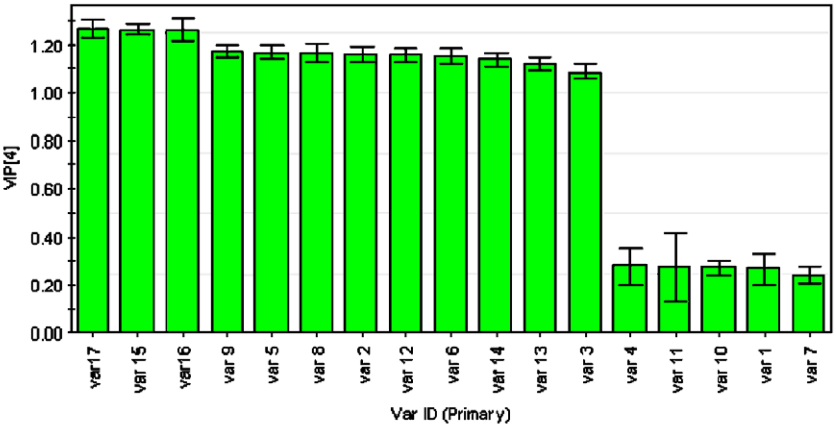

Fig. 8. Variable importance in the projection (VIP) plot for all batches.

indicating that the score plot is a reasonable summary of the batch trajectory.

Each colored trajectory summarizes the dynamics of a given batch. If all batches are consistent, then all trajectories will overlap. Spikes in the trajectories are observed in the second principal component. This indicates abnormal operating conditions for batches 11 to 16 from times 62 to 65. After discussion with the process engineers, it is confirmed that batches 11 to 16 experienced feed control issues during times 62 to 65 [3]. Joint analysis of the loading and score plots indicates that the first principal component summarizes time variation. To emphasize the batch variation, it is meaningful to examine the second principal component versus the elapsed time in Fig. 5. Unusual operating conditions are also observed for batches 14–16 from times 45 to 53 and for batch 14 in the beginning of the reaction. From these two figures, localized problems can be easily identified.

The rank of the model is determined to be four using leave 1/7 out cross validation and the overall summary statistics are shown in Table 1. See Martens and Dardenne for a discussion of the benefits of cross validation [30]. The T_2 contribution chart for batch 14, time 63 is shown in Fig. 6. It indicates that variables 1, 4, and 11 are associated with the abnormal behavior at time 63 for batch 14. A time series plot of variable 11 shown in Fig. 7 confirms this finding. The same conclusion is obtained for the other 5 abnormal batches around times 62 to 65. The variable importance in the projection (VIP) plot in Fig. 8 ranks the importance of each variable in terms of its correlation with the maturity variable. It is clear that variables 1, 4, 7, 10, and 11 correlate weakly with the maturity variable. It will be more interesting to focus only on variables that are related to the quality variable for further analyses.

For the second model, batches 11 to 16 and variables 1, 4, 7, 10, and 11 are removed. After leave 1/7 out cross validation, the rank is determined to be one. Comparison between this model ($R^2_{X,\text{cum}}$=90.3%; $R^2_{Y,\text{cum}}$=94.8%; Q^2_{cum}=94.7%) and the original model indicates that this reduced model fits better. However, model diagnostics in Fig. 9 indicate that batches 3 and 4 are abnormal throughout the fermentation process.

For the final model, batches 3 and 4 are also removed and the remaining 12 batches show consistent batch trajectories in the score plot (not shown here). This is confirmed by the improvement in the model fitness statistics ($R^2_{X,\text{cum}}$=91.9%; $R^2_{Y,\text{cum}}$=95.5%; Q^2_{cum}=95.4% at rank 1, determined by cross validation). Batches 3, 4, and 11 to 16 are then projected back to the model and the Q statistics are plotted in Fig. 10. The Q statistics for batches 3 and 4 are high for the entire phase, while the Q statistics for batches 11 to 16 are high for some duration of the fermentation process. This is a consistent conclusion from the previous two rounds of batch analyses.

Based on this analysis, 12 of the 20 batches were identified as golden batches in a sense that no major operating problems

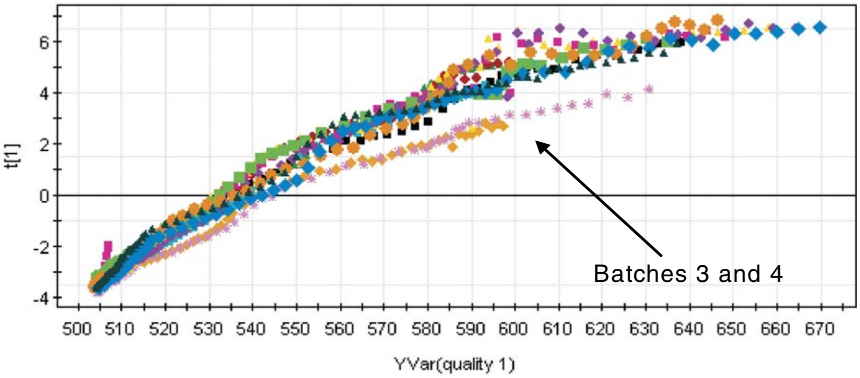

Fig. 9. Observation unfolding MPLS diagnostics (first principal component versus quality variable) after 6 abnormal batches and 5 unimportant variables are removed.

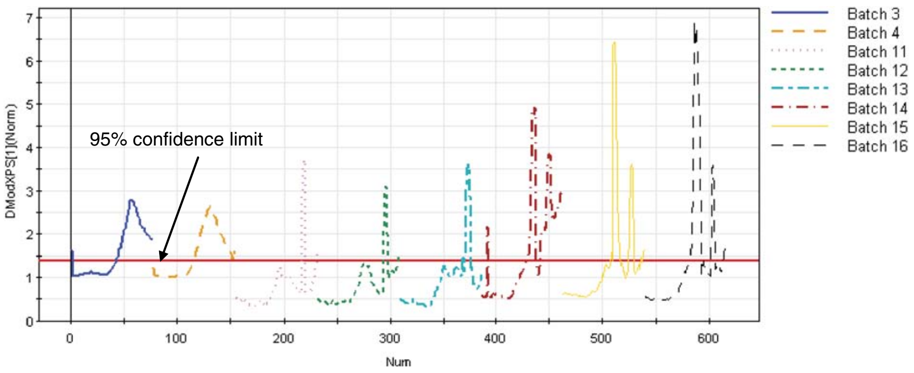

Fig. 10. Q statistics plot for the eight abnormal batches.

were encountered and that good correlation between the process variables and the maturity variables was observed in both phases. Two batches were known to be bad and the other six batches were listed as questionable in a sense that some operating issues were encountered and that poor correlation between the process variables and the maturity variables was observed for some batches.

Westerhuis et al. [13] and Kourti [17] comment that observation-unfolding MPLS assumes there is sufficient information in the trajectories to obtain a good prediction of maturity variable. In other words, this approach only works when there is some linear combination of the variables in each region of time that is either decreasing or increasing. This assumption is often satisfied when a maturity variable can be determined in a batch process. For the data set used in this paper, optical density is measured on-line to determine the maturity of the fermentation process. Based on the physical nature of fermentation, many process variables are correlated to optical density (as evident in the VIP plot shown in Fig. 8). Observation-unfolding MPLS correctly focuses these variables in the model. It is especially important to monitor these variables because any abnormality will often lead to process or product quality problems.

It is possible that a monotonically increasing or decreasing maturity variable does not correlate with any of the process variables in a batch process. This is a strong indication that key process variables are missing in the system. It is important to validate the assumption that some process variables are correlated with the maturity variable by examining the model fitness and model diagnostics. If the goal is to obtain on-line quality prediction, then it is premature to apply the model. If the goal is to detect and diagnose process faults on-line, then it is useful to implement observation unfolding MPLS by examining the scores and residuals of the model.

4.2. Batch unfolding MPCA

To identify consistency among batches in the batch unfolding MPCA, the score plot is used (see Fig. 11). Each dot represents a batch. If some batches are consistent with the rest, then clustering of batches is expected. If bad batches exist, then it is expected that the bad batches will be outside the critical limit. As shown in the score plot, all batches are inside the 95% confidence limit and no clear clustering of batches is observed.

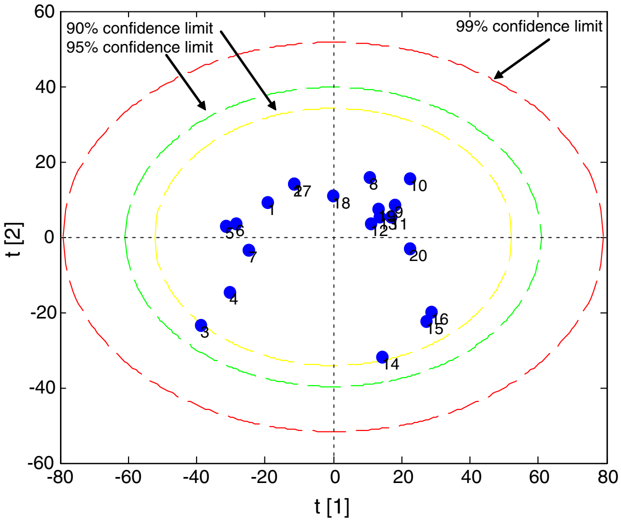

Fig. 11. Batch unfolding MPCA score plot for all variables and all batches.

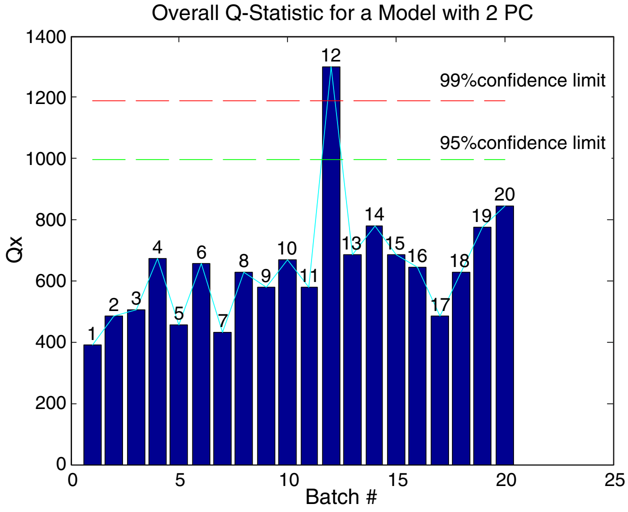

Fig. 12. Batch unfolding MPCA Q statistic for all batches.

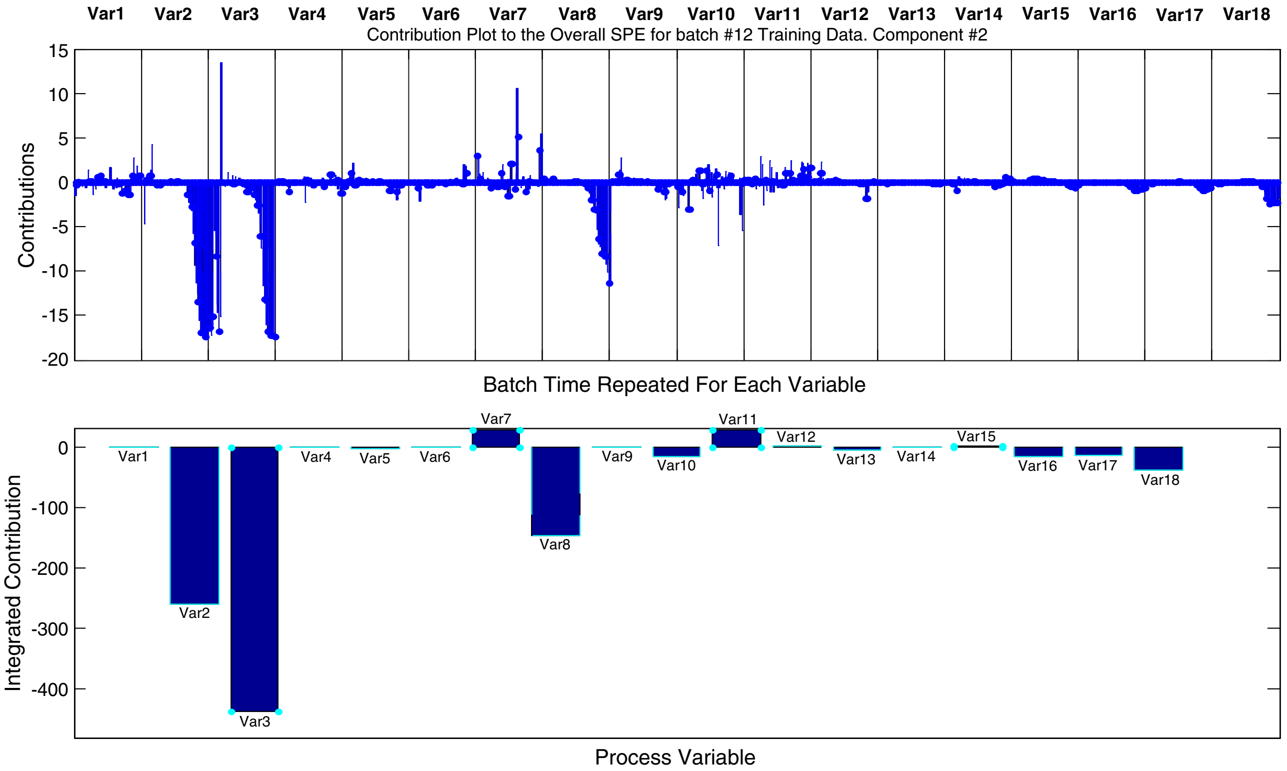

Fig. 13. Batch unfolding MPCA Q statistic contribution chart for batch 12.

The Q statistic plot in Fig. 12 shows that batch 12 is above the 99% confidence limit, while the rest of the batches are below the 95% confidence limit. Based on the score plot and the Q statistic, it can be concluded from the batch unfolding MPCA that only batch 12 is abnormal and that the rest of the batches are consistent.

The Q statistic contribution chart for batch 12 is plotted in Fig. 13. Overall, variables 2, 3, and 8 are shown to be abnormal. Recall from Fig. 6 that variables 1, 4, and 11 at times 62 to 65 are abnormal for batches 11 to 16. The batch unfolding MPCA did not come to the same conclusion. In other words, the localized problem in batches 11–16 are masked by the overall problem in variables 2, 3, and 8 in batch 12. From the statistical point of view, it is clear that the presence of abnormal batches in the model makes it difficult to correctly define the confidence limits.

To investigate whether abnormal batches can be identified on-line using batch unfolding MPCA, the 12 golden batches (as identified using the observation unfolding MPLS and confirmed by the process engineers) were used to build a model and the results show that all golden batches are inside the 95% confidence limit (see Fig. 14).

Batches 3 and 4 are projected onto the MPCA model and the Q statistics are plotted in Fig. 15. Note that almost all of the data points of the normal batches satisfy the Q statistic confidence limit (not shown here) and it is clear that batches 3 and 4 are different than the normal batches throughout the entire runs. Comparison between Figs. 10 and 15 indicates that both approaches identify these two batches as abnormal. At any time during the batch run, the contribution chart can be used to identify process variables that are related to the abnormal batch behavior. Batches 11 to 16 are projected onto the MPCA model and the Q statistics are plotted in Fig. 16. Similarly to Fig. 10, localized problems for these batches are now clearly detected.

4.3. Tucker3 analysis

Tucker3 applied to the j-scaled data leads to a (2,2,2) model explaining 68.8% of the total variance. In order to have a

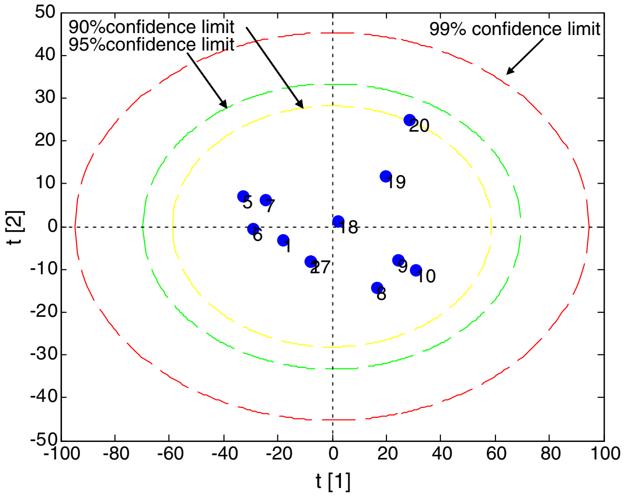

Fig. 14. Batch unfolding MPCA score plot after 8 abnormal batches and 5 unimportant variables are removed.

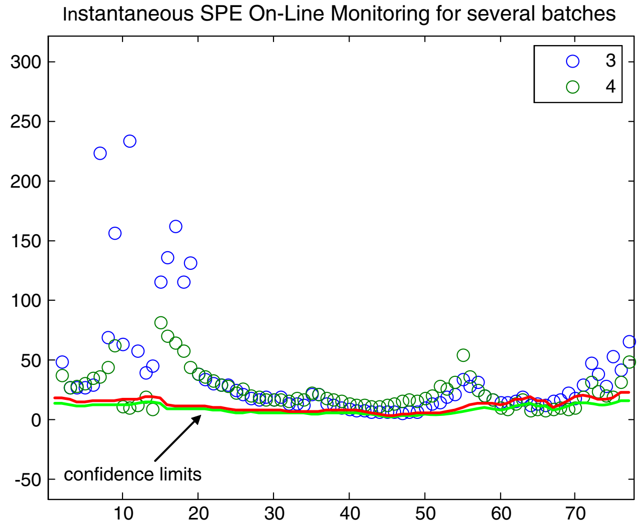

Fig. 15. On-line detection for batches 3 and 4 using the Q statistic. The solid lines represent the 95% and 99% confidence limits.

simpler and more interpretable model, the core matrix has been rotated to maximize its superdiagonality. The following core matrix has therefore been obtained:

124.07	−2.59	1.88	20.29
5.37	9.66	−8.93	38.19

From it the relevance of each possible triplet of loadings can be deduced. Here the core matrix has been rearranged according to the following structure:

$g_{1,1,1}$	$g_{2,1,1}$	$g_{1,1,2}$	$g_{2,1,2}$
$g_{1,2,1}$	$g_{2,2,1}$	$g_{1,2,2}$	$g_{2,2,2}$

The square of each element corresponds to the variance explained by taking into account the corresponding triplet of loadings. As an example, by looking at the same time at the loadings on the first factor for each of the modes we have an explained variance equal to 124.072, i.e. 15,395. Because the total variance of the data set is 26,163 (i.e., 17 × (77 × 20 − 1)), this combination explains 58.8% of the total variance. The second most important term in the core matrix is $g_{2,2,2}$ (38.19), by itself accounting for 5.6% of the total variance. The most relevant non-superdiagonal element is $g_{2,1,2}$ (20.29), explaining just 1.6% of the variance.

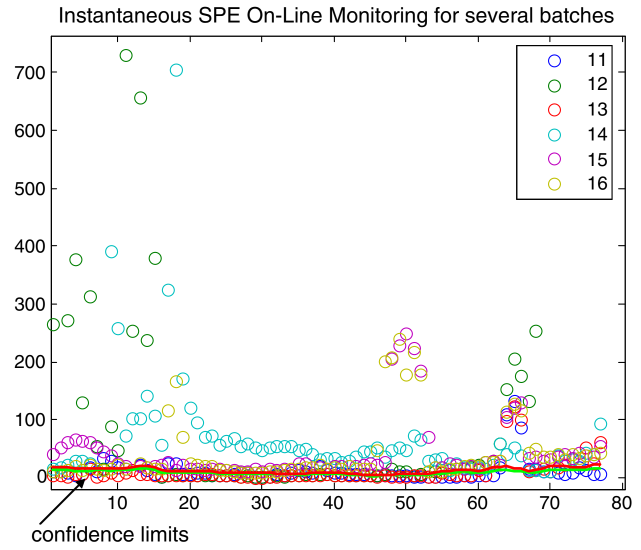

Fig. 16. On-line detection for abnormal batches 11–16 using the Q statistic. The solid lines represent the 95% and 99% confidence limits and the x axis represents time.

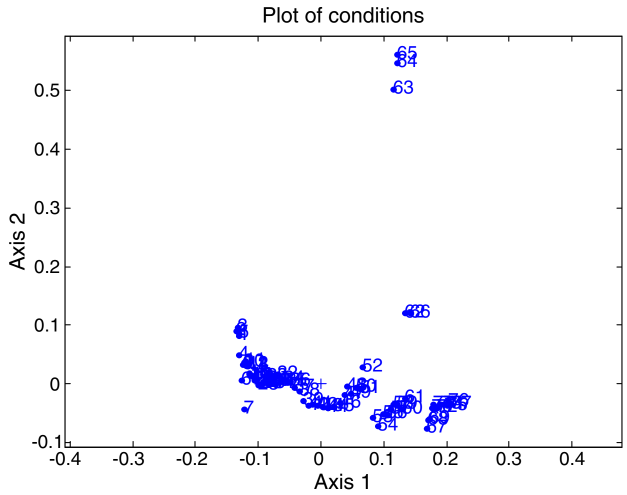

Fig. 17. Condition loading (time) plot.

The fact that the elements $g_{1,1,1}$ and $g_{2,2,2}$ are by far the most important ones means that the projections of the three sets of loadings on the two factors can be interpreted jointly, exactly in the same way as in a "standard" PCA.

From the plot of the conditions (i.e., reaction times) in Fig. 17 it can be seen that the loadings are dispersed mainly along the first axis, with a regular trend from low values to high values. Times 1–8 have very similar loadings on the first axis, this meaning that until time 8 the reaction did not actually start. Times 62–66 (mainly 63–65) have instead much higher loadings on the second axis, indicating that something unusual has happened at those times.

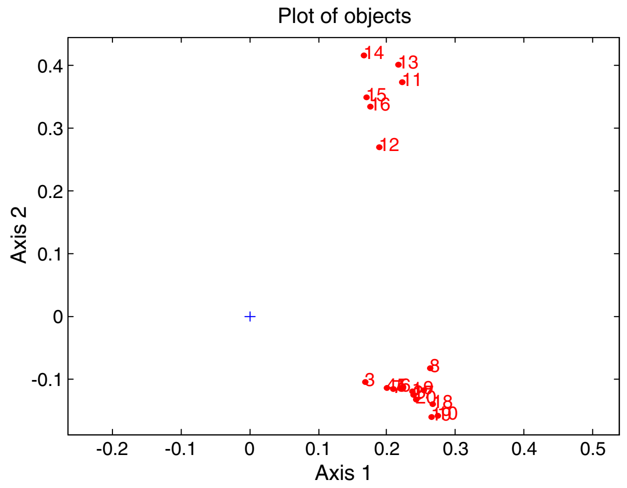

Fig. 18. Object loading (batch) plot.

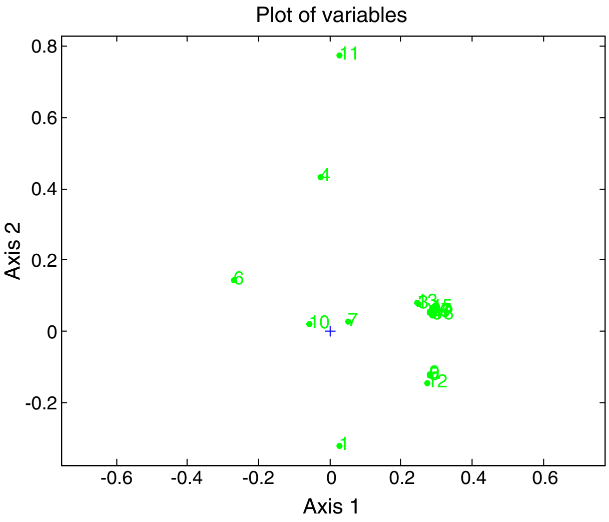

Fig. 19. Variable loading (variable) plot.

The plot of the objects (i.e., batches) in Fig. 18 shows clearly two main groups. All the batches having high loadings on the second axis (i.e., batches 11–16) are those batches that have been described by the operators as being characterized by “feed control issues”.

The plot of the variables in Fig. 19 shows that variables 2, 3, 5, 8, 9, 12–17 have high positive loadings on the first axis and are opposite to variable 6, having a high negative loading on the first axis; variables 11 and 4 have high positive loading on the second axis, while variable 1 has a high negative loading on the second axis; variables 7 and 10 have low loadings on both axes and therefore are not relevant to the model.

Since, as previously shown, the core matrix is almost completely superdiagonal, a joint interpretation of the three loading plots can be obtained. First of all, it can be noticed that the loadings of the times are dispersed along the first axis, while the loadings of the batches are dispersed along the second axis. From the discussion about the core elements, it can now be concluded that the variability among times is more than eight times greater than the variability among batches. This is quite logical, since the same batch at the beginning and at the end of the process is much more different than two different batches at the same stage of the process. By taking into account the loadings of the times and the loadings of the variables, it can be said that variables 2, 3, 5, 8, 9, 12–17 increase with time, while variable 6 decreases with time and variables 1, 4, 7, 10 and 11 stay relatively constant. This is the same conclusion reached with the observation unfolding MPLS.

By looking at the three loading plots, it is clear that the main difference between the batches having feed control issues and the rest of the samples is at times 62–66, when variables 4 and 11 were much higher and variable 1 was much lower than the average.

Batches 3 and 4 are known to be bad batches. It can be seen that, among the batches not affected by the “feed control issues” problem, they are the two batches with the lowest loadings. This means that they had consistently lower values of variables 2, 3, 5, 8, 9, 12–15 and higher values of variable 6 than the rest of the batches. This small difference, occurring throughout the whole reaction time, led to significantly poorer quality products.

5. Conclusions

The Tucker3 analysis, the observation unfolding MPLS, and the batch unfolding MPCA are applied to the batch data from the San Diego Fermentation process. It is concluded that:

- All three methods are complementary to each other and a well-trained chemometrician/practitioner will find all three methods to be useful for batch data analysis.
- The batch unfolding MPCA is more sensitive to the overall batch variation while the observation unfolding MPLS is more sensitive to the localized batch variation. The Tucker3 method is in good balance in terms of detecting both variations.
- For this fermentation data set, it is important to focus on both the overall and localized variations. Out of the 20 batches, batches 11–16 experienced temporary control issues during the runs, but went back to normal at the end. It is important to identify and remove these batches for further analyses. Both the Tucker3 method and the observation unfolding MPLS clearly identified these six batches, but the batch unfolding MPCA did not.
- Batches 3 and 4 were bad throughout the entire runs. This conclusion can be easily obtained using the observation unfolding MPLS. This conclusion is less clear in the Tucker3 analysis and the batch unfolding MPCA. However, once these two batches are labeled as “bad”, it will be easier to look for signs in both methods that these two batches are indeed bad.
- All methods work well in detecting the overall problems for the two bad batches and the localized problems for the other six batches once models are rebuilt using only the 12 good batches.
- The observation and interpretation of the three loading plots obtained by Tucker3 often directly lead to the same findings that can be reached by the unfolding approaches.

Acknowledgement

The authors want to thank Prof. John MacGregor at McMaster University for providing BathchSPC software.

References

[1] P. Nomikos, J.F. MacGregor, AIChE J. 40 (1994) 1361–1375.

[2] S. Wold, N. Kettaneh, H. Friden, A. Holmberg, Chemometr. Intell. Lab. Syst. 44 (1998) 331–340.

[3] L. Chiang, A. Kordon, L. Chew, D. Coffey, R. Waldron, K. Haney, T. Bruck, A. Jenings, H. Talbot, in: S. Shah, J.F. MacGregor (Eds.), Proceedings of the Dynamics Control of Process System-7, 2004.

[4] R. Bro, Chemometr. Intell. Lab. Syst. 38 (1997) 149–171.

[5] A.K. Smilde, J. Chemometr. 15 (2001) 19–27.

[6] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis: Applications in the Chemical Sciences, John Wiley & Sons, 2004.

[7] R. Leardi, C. Armanino, S. Lanteri, L. Alberotanza, J. Chemometr. 14 (2000) 187–195.

[8] S. Albert, R.D. Kinley, Trends Biotech. 19 (2001) 53–62.

[9] L. Gregersen, S.B. Jorgensen, Chem. Eng. J. 71 (1999) 69–76.

[10] J.A. Lopes, J.C. Menezes, J.A. Westerhuis, A.K. Smilde, Biotechnol. Bioeng. 80 (2002) 419–427.

[11] C. Undey, E. Tatara, A. Cinar, J. Biotech. 108 (2004) 61–77.

[12] A. Cinar, S.J. Parulekar, C. Undey, B. Gulnur (Eds.), Batch Fermentation: Modeling, Monitoring, and Control, Marcel Dekker, 2003.

[13] J.A. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemometr. 13 (1999) 397–413.

[14] B.M. Wise, N.B. Gallagher, S.W. Butler, D.D. White, G.G. Barna, J. Chemometr. 13 (1999) 379–396.

[15] D.J. Louwerse, A.K. Smilde, Chem. Eng. Sci. 55 (2000) 1225–1235.

[16] E.N. van Sprang, H. Ramaker, J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Chem. Eng. Sci. 57 (2002) 3979–3991.

[17] T. Kourti, Annu. Rev. Control (2003) 131–139.

[18] C. Undey, S. Ertunc, A. Cinar, Ind. Eng. Chem. Res. 42 (2003) 4645–4658.

[19] B. Lennox, G.A. Montague, H. Hiden, G. Kornfeld, P.R. Goulding, Biotechnol. Bioeng. 74 (2001) 125–135.

[20] J. Chen, K.C. Liu, Chem. Eng. Sci.. 57 (2002) 63–75.

[21] A. Simoglou, E.B. Martin, A.J. Morris, Comp. Chem. Eng. 26 (2002) 909–920.

[22] J. Lee, C. Yoo, I. Lee, Comp. Chem. Eng. 28 (2004) 1837–1847.

[23] C.K. Yoo, J. Lee, P.A. Vanrolleghema, I. Lee, Chemometr. Intell. Lab. Syst 71 (2004) 151–163.

[24] N. Lu, Y. Yang, F. Gao, F. Wang, in: F. Allgower, F. Gao (Eds.), Proceedings of the 7th International Symposium on Advanced Control of Chemical Processes, 2004, pp. 471–476.

[25] S. Garcia-Munoz, T. Kourti, J.F. MacGregor, Ind. Eng. Chem. Res. 43 (2004) 5929–5941.

[26] P. Miller, R.E. Swanson, Appl. Math. Comp. Sci. 8 (1998) 775–792.

[27] A.K. Smilde, Chemometr. Intell. Lab. Syst. 15 (1992) 143–157.

[28] Umetrics, Inc., SIMCA-P+, version 10, www.umetrics.com, 2003.

[29] MACC McMaster Advanced Control Consortium, BatchSPC, version 2.0, www.chemeng.mcmaster.ca/MACC, 2004.

[30] H.A. Martens, P. Dardenne, Chemometr. Intell. Lab. Syst. 44 (1998) 99–121.