STATISTICAL PROCESS MONITORING OF BIOREACTORS: A COMPARISON

Wu Long, Ognjen Marjanovic, Barry Lennox
Control Systems Centre, School of Electrical and Electronic Engineering,
University of Manchester, United Kingdom

Abstract: Batch processes, such as fermenters, generally require high levels of consistency in their operation to ensure minimal losses of raw materials and product. Recent application studies have indicated that multivariate statistical technology can provide some support when trying to maintain consistent operation in complex batch processes. This paper aims to compare four different approaches to batch process monitoring using statistical methods. The comparison is made in terms of their respective ability to tolerate normal process variation while detecting abnormal operation of a process. The comparison is performed using data sets obtained from one simulated bioreactor and two industrial fermentation processes.

Keywords: biotechnology, fault detection, performance monitoring, statistical process control.

1. INTRODUCTION

Batch processes, such as industrial fermenters, are used in the manufacturing of high-value products. As a result, it is particularly important to detect incipient degradation of batch performance in order to recover high-value product with minimal losses to raw materials, utilities and product. Furthermore, the pharmaceutical industry, together with the food and beverage industry are obliged to comply with increasingly stringent regulatory requirements enforced by agencies such as the Food and Drug Administration (FDA). For many compounds, these agencies demand proof that consistent operation is adhered to and without this proof the product cannot be sold.

One very popular approach for ensuring the consistency of batch process is to adopt methodology of the Multivariate Statistical Process Control (MSPC) (Nomikos and MacGregor, 1995; Wold, *et al.,* 1998). This methodology attempts to capture relationships that exist between different process variables and condenses this information into a small number of important metrics. These relatively few metrics can then be easily monitored in real-time in order to benchmark process performance and highlight potential problems, leading to continuous improvement of a manufacturing plant's reliability and profitability. It is worth noting that the emergence of powerful computational devices has allowed MSPC to emerge as an important new process monitoring technology.

However, along with the requirement for detection of incipient faults, there is, albeit somewhat implicit, requirement to maintain number of false alarms to the bare minimum. In fact, from the industrial experience of the authors, false alarms have often proved to be the crucial obstacle to the realization of the effective real-time batch process monitoring scheme. During the initial phases of the real-time application, frequent occurrences of false alarms were considered a serious nuisance for operational personnel. As a result, they undermined operators' confidence in the effectiveness of the overall batch process monitoring solutions.

Since the emergence of the statistical batch process monitoring techniques some 15 years ago (MacGregor and Nomikos, 1992) several variants and extensions to the original methodology have been proposed. In particular, Wold, *et al.* (1998) suggested alternative way of consolidating data from multiple batches, thereby resulting in fundamentally different statistical process model. Both of these approaches are included in this comparison as they represent the mainstream of statistical batch process monitoring. Additional multivariate method, included in this comparison, is a hybrid of the two standard multivariate methods, mentioned above. This method has already been suggested, most notably by Lee, *et al*. (2004), but has not yet been comprehensively compared with the main approaches in terms of false alarms sensitivity and fault detection capability using multitude of data sets. Alongside these multivariate statistical methods, there is also an extension of the standard univariate statistical process control to batch processes. This method is frequently used by people in industry and is readily available in modern commercially available software packages. Also, design and implementation of the univariate statistical process monitoring scheme is much simpler when compared to multivariate methods. However, quite surprisingly, this relatively simple statistical technique has not yet been comprehensively and explicitly compared with more involving multivariate methods.

There have been several published comparisons between different batch process monitoring techniques using industrial and simulated data sets (Westerhuis, *et al*., 1999; Gurden, *et al*., 2001; Van Sprang, *et al*., (2002); Chiang, *et al*., 2006).

However, in a great majority of the cases the focus of the comparison has been on the ability to detect a fault rather than on the sensitivity to false alarms. One notable exception is a comparison by Van Sprang, *et al*. (2002) although even they place much greater emphasis on the detection of erroneous batches rather than the sensitivity of a given scheme to normal process variation. Also, the explicit comparison between multivariate and univariate methodologies has not yet been presented to the best of the authors' knowledge.

This paper presents results from three case studies in which both traditional univariate and advanced multivariate statistical analyses, namely Principal Component Analysis (PCA), have been applied to fermentation processes. Two of these studies have been conducted using industrial data while the third was performed using data from a simulated bioreactor. The main focus of this paper is to compare four different statistical monitoring approaches in terms of their respective ability to tolerate normal process operation while detecting anomalous process variation.

This paper is organized as follows. Firstly, a brief description of different statistical process monitoring techniques is provided in section 2. In section 3 multivariate statistical process monitoring charts are reviewed, followed by description of performance indices used in the comparison in section 4. Section 5 describes each of the three data sets used in this study. Section 6 contains results of the comparison. Finally, section 7 concludes the paper.

2. STATISTICAL PROCESS MONITORING TECHNIQUES

2.1 Univariate Statistical Process Control (USPC)

Univariate statistical process control (USPC) considers all the process variables to be independent of each other. The mean and standard deviation of the trajectory that each variable follows for a set of satisfactory batches are determined and these statistics are then used to establish quality control limits. These control limits in the USPC chart define an envelope of satisfactory operation for each recorded process variable. Consistent violation of these limits during a batch progression would then indicate that the conditions of the current batch are inconsistent with what is expected for satisfactory performance, suggesting that the batch may be of poor quality. A drawback with this approach is that it ignores any relationships that may exist between process variables and that many variables may need to be monitored.

2.2 Multivariate Statistical Process Monitoring

Multivariate statistical analysis captures relationships that exist between different process variables and condenses this information into a small number of important metrics. This analysis represents a more comprehensive attempt, when compared to USPC, to capture the nominal operation of the process in the form of a statistical model.

Multivariate statistics relies heavily upon the statistical routines referred to as Principal Component Analysis (PCA) and Partial Least Squares (PLS). In this paper, focus is on the Principal Component Analysis, which is generally used to develop a statistical model representing satisfactory process operation. More specifically, PCA model identifies the inter-variable relationships that exist during satisfactory process operation. PCA is then able to extract the main features of process operation, which can be stored in a small number of composite variables, commonly referred to as scores. These composite variables can then be easily monitored in real-time in order to benchmark process performance and highlight potential problems, leading to continuous improvement of the process operation.

Before reviewing multivariate methods, brief introduction to the problem of unfolding original batch data is discussed first. When the data is collected from I batches, each characterized by measurements of J process variables at K sampling instances, it is natural to arrange data matrix into 3-dimensional data matrix, denoted as $\underline{\mathbf{X}} \in \Re^{I \times K \times J}$. However, standard multivariate statistical methods require ubiquitous 2-dimensional data matrix format. Hence, there is a need to re-arrange the original 3-dimensional data matrix into more familiar 2-dimensional format in order to be able to apply multivariate methods already developed for continuous processes. This procedure is often referred to as "unfolding". However, there are several different ways of unfolding batch data. In fact, Westerhuis, *et al.* (1999) review in detail all six alternative methods and critically assess their implication on the nature of the resulting statistical models. In this paper, the focus is on methods based on the 2 most frequently used ways of unfolding batch data matrix.

Batch-Wise Principal Component Analysis (B-PCA). Nomikos and MacGregor (1995) unfold the original 3-dimensional data matrix into ubiquitous 2-dimensional format by preserving batch direction:

$$\underline{\mathbf{X}} \in \Re^{I \times K \times J} \Rightarrow \mathbf{X} \in \Re^{I \times JK} \qquad (1)$$

Note that the standard pre-processing procedure of scaling the resulting $\mathbf{X}$ matrix results in the removal of mean trajectories of each process variable from the data set. This is the crucial feature of this method that removes the main source of nonlinearity, thereby allowing accurate and yet linear statistical models to be developed, as argued by Westerhuis, *et al*. (1999) among others. After the unfolding and scaling procedures, standard Principal Component Analysis is performed.

However, this approach suffers from two main drawbacks. Firstly, it requires batches to be of similar lengths (i.e. similar durations), which is not always the case in process industries. This is because varying batch duration causes unfolded data matrix to have varying number of columns. Secondly, during the real-time application, it is required to estimate future portion of process evolution in order to evaluate the model at each sampling instant. This issue arises because each row of the unfolded data matrix contains measurements of all of the process variables across all of the sampling instances for a given batch. This problem is generally solved using crude estimation models suggested by Nomikos and MacGregor (1995). In this case study, all of the batches are of equal length and the so-called "current deviations approach" is used to estimate future portion of the batch progression (Nomikos and MacGregor, 1995; Van Sprang, *et al.,* 2002). Briefly, the current deviations approach assumes that the future measurements will continue to deviate from their mean trajectories at the same level as present at a current time instant k.

Variable-Wise Principal Component Analysis (V-PCA). Another popular method was originally proposed by Wold, et al. (1998), which unfolds the original 3-dimensional data matrix by preserving variable direction:

$$\underline{\mathbf{X}} \in \Re^{I \times K \times J} \Rightarrow \mathbf{X} \in \Re^{IK \times J} \qquad (2)$$

In this case, scaling of the resulting matrix does not remove mean trajectory and therefore, the main source of non-linearities remains in the data set. After the unfolding and scaling procedures, standard Principal Component Analysis is performed.

However, this approach does not suffer from the drawbacks of the B-PCA method. In particular, varying batch lengths do not render resulting data matrix unusable. This is because number of columns in the unfolded data matrix is not dependent on the number of sampling instances during a batch. Hence, this method can be applied to processes that exhibit variation in terms of batch duration. Also, during the real-time application, there is no requirement to estimate future process evolution. This is because each row of the unfolded data matrix contains measurements of process variables at a single sampling instant, as opposed to containing measurements across the entire batch duration.

Batch-Variable-Wise Principal Component Analysis (B-V-PCA). Batch-Variable-Wise Principal Component Analysis (B-V-PCA) was suggested in several publications, most notably by Lee, et al. (2004). In this case original 3-dimensional data matrix is first of all unfolded by preserving batch direction, resulting in . Each column of this matrix is scaled, thereupon removing the mean trajectory of each process variable. However, unlike B-PCA method, this approach then folds auto-scaled data matrix back into 3-dimensional format and then re-unfolds it in the same way as V-PCA method, i.e. by preserving variable direction: . Hence, data matrix used for identification/evaluation of B-V-PCA model is of the same dimensions as the data matrix used in the case of V-PCA method. However, in the case of B-V-PCA mean trajectory is removed from each process variable, which is not the case with V-PCA approach.

It is worth noting that, unlike V-PCA approach, this method does require batches to be of equal lengths because the scaling procedure is performed on the data matrix obtained using "batch-wise unfolding approach. However, real-time application of this method does not involve estimation of the future batch progression because the data used for model identification is unfolded such that each row contains measurements at a single sampling instant. Hence, B-V-PCA can be viewed as a hybrid of B-PCA and V-PCA, attempting to combine their comparative advantages and mitigate their disadvantages.

3. MULTIVARIATE STATISTICAL MONITORING CHARTS

In general, on-line monitoring of batch processes, using multivariate statistical methods, is performed using two types of control charts: the T^2 chart to monitor deviation of a process from a center-point as defined by statistical model and SPE chart to monitor process deviation for a statistical model. These metrics are calculated at each sampling instant of an evolving batch.

One of the most frequently used metrics in multivariate statistical process monitoring is the composite prediction error of a model, termed Squared Prediction Error (SPE). SPE is computed by comparing predictions made by statistical model with the actual values of process variables. This metric provides a single measure of process deviation from the statistical model. Hence, the SPE metric is expected to have high values during abnormal operation and low values when the process operates in a satisfactory manner.

Another frequently used metric is the T^2 statistic. This metric is composed of all of the retained scores, which are obtained by projection of a data point onto the plane defined by means of PCA or PLS analysis. This metric represents the main features of the process. Hence, the T^2 metric provides a measure of how far away the current operating conditions of the plant are from the conditions present in the data that was used for statistical model development, assuming mismatch between process and a related statistical model is minimal.

Confidence intervals and, therefore, control limits of the SPE and T^2 chart are computed using normal operating data and assuming that SPE metric and T^2 metric are characterized by chi-squared distribution and F-distribution, respectively, under normal

operating conditions. In this paper, control limits are calculated with confidence interval of 99%.

4. PERFORMANCE INDICES

Two performance indices are used to evaluate how well each approach performs in terms of detecting abnormal variation while tolerating normal process operation. The two indices are the overall Type I error and the overall Type II error.

4.1 Type I Error

Formally defined, Type I error, also known as "error of the first kind" or "false positive" is the error of rejecting a null hypothesis when it represents a normal state of nature. In other words, Type I error occurs when something that should have been accepted as normal was rejected. In terms of batch process monitoring, Type I error arises whenever normal operating condition is classified as faulty.

In this paper, Type I error is calculated using the following formula:

$$\text{Type I Error} = \frac{\sum \text{false alarms}}{IK} \quad (3)$$

where I is the number of normal operation batches and K is the number of sampling instances during each batch.

The following procedure is used to calculate Type I error estimate. One batch is removed from the set of normal operating batches, i.e. training data set, and a statistical model is built on the remaining normal operation batches. As this batch is taken from the set of normal operation batches, it is assumed to be classified as normal by monitoring control charts. Therefore, crossing of a control limit in this case is assumed to be an instance of a false alarm.

4.2 Type II Error

In formal terms, Type II error, also known as "error of the second kind" or "false negative", is the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, Type II error occurs when something that should have been rejected was accepted as normal.

In terms of batch process monitoring, Type II error occurs whenever faulty process operation is classified as normal.

In this paper, Type II error is calculated using the formula, given in equation (4).

$$\text{Type II Error} = \frac{\sum \text{undetected faulty samples}}{IK} \quad (4)$$

where "undetected faulty samples" corresponds to samples of process variables or metrics for which it is known that the fault has taken place and yet this process variable or metric remains within its control limits.

The ideal case is the one in which both Type I and Type II error are equal to zero. However, this is hardly ever achieved and the attempt is made instead to strike a compromise between conflicting objectives. In some situations it may be critical to reduce Type II error from 100% down to, say, 50% regardless of the Type I error. This situation may arise in some critical systems where benefit of the fault detection offsets any costs incurred as a result of false alarms. In these cases the designer would choose the most sensitive approaches to achieve this aim.

5. DESCRIPTION OF THE DATA

In order to evaluate different approaches for on-line process monitoring, the approaches are applied to three data sets related to bioscience industry.

5.1 Industrial Fermenter No. 1 – Data1

The first case study is conducted on the data set from a bioreactor used in pharmaceutical industry. This data set consists of 10 normal batches and 1 faulty batch. For each batch, 3 process variables are continuously measured across 181 sampling instances. The faulty batch is characterized by a highly subtle incipient fault on one of the critical sensors. This data set will from now be referred to as Data1.

5.2 Industrial Fermenter No. 2 – Data2

The second case study is conducted on the data set from a fermenter used in food and beverage industry. This data set consists of 20 satisfactory batches and 2 abnormal batches for which 7 process variables are continuously measured at 83 sampling instances. Abnormal batches are caused by the underlying, i.e. unmeasured, disturbance that is reflected in terms of a multitude of process variables. This data set will from now be referred to as Data2.

5.3 Simulated Bioreactor – Data3

Third case study is performed using simulated penicillin production process, which is documented in detail by Birol, *et al*. (2002). Measured process variables include: substrate feed-rate, dissolved oxygen and CO_2 concentrations, reactor temperature and volume, flow-rate of cooling water, pH inside the reactor and the flow-rate of base/alkaline solution into the reactor. Data set corresponding to normal process operation comprises 30 batches, each

containing measurements of 8 process variables at the 157 sampling instances during which substrate flows into the bioreactor. Batch-to-batch and within-batch variation are created by applying filtered pseudo-random binary sequence to the substrate feed-rate. Three faulty batches are created by reducing substrate feed concentration from the nominal 600 g/l to540 g/l. This type of fault was simulated on 3 batches, each having different realization of substrate feed-rate sequence. This data set will from now be referred to as Data3.

6. RESULTS AND DISCUSSION

Prior to analysis, the number of principal components for each method is determined using well known cross-validation technique, see (Wold, 1978). This exercise was performed for each data set with the results given in Table 1.

Table 1 Number of retained principal components

	Data1	Data2	Data3
V-PCA	1	3	3
B-V-PCA	1	2	2
B-PCA	4	7	8

Type I error of the four compared methods for three data sets are presented in Tables 2 and 3. These tables express false alarm rates as percentages of the overall number of measured samples during normal operation batches. Table 2 displays Type I error for USPC charts and T^2 charts, while Table 3 displays Type I error for USPC and SPE charts.

Table 2 Type I Error for the USPC charts and T^2 charts

	Data1	Data2	Data3
USPC	3.79%	2.73%	0.83%
V-PCA	3.2%	2.29%	2.78%
B-V-PCA	3.26%	2.35%	1.06%
B-PCA	0.11%	0.72%	0.53%

Table 3 Type I Error for the USPC and SPE charts

	Data1	Data2	Data3
USPC	3.79%	2.73%	0.83%
V-PCA	6.08%	5.12%	1.36%
B-V-PCA	12.87%	9.28%	1.78%
B-PCA	25.30%	25.48%	4.20%

The first observation is that relative sensitivities of the four compared methods to normal process variation are consistent across all of the three data sets. In particular, T^2 of the B-PCA method is the least sensitive, of all of the monitoring charts, to normal process variation. This chart has very small values of Type I error for each of the three data sets as observed in Table 2. On the other hand, SPE chart of the B-PCA method is clearly and consistently by far the most susceptible to false alarms, as observed in Table 3. Also, T^2 charts of both V-PCA and B-V-PCA perform similarly to each other and to the USPC charts, while their SPE charts are consistently more sensitive to normal process variation when compared to USPC charts. Overall, T^2 charts of all of the three multivariate approaches as well as the USPC charts are clearly and consistently less sensitive to normal process operation, when compared to SPE charts. In other words, SPE chart of any of the three compared methods is likely to generate more false alarms than either the corresponding T^2 chart or USPC chart.

One anomaly in Table 3 is the low value of Type I error for the SPE chart corresponding to B-PCA in the case of the simulated bioreactor (Data3). The main reason for this behaviour lies in the fact that the simulated bioreactor has a high degree of batch repeatability. In fact, dynamics of this process are time-invariant, i.e. simulation is not modified on a batch-to-batch basis. On the other hand, industrial processes generally exhibit continuous changes in their dynamic behaviour and their repeatability is relatively smaller when compared to simulated time-invariant systems, resulting in a larger number of false alarms.

Type II error of the four compared methods for three data sets are presented in Tables 4 and 5. These tables express number of correct control limit violations by process variables or multivariate metrics as a percentage of the overall number of samples collected throughout the fault duration.

Table 4 Type II Error for the USPC and T^2 charts.

	Data1	Data2	Data3
USPC	100%	77.53%	81.37%
V-PCA	100%	57.23%	98.73%
B-V-PCA	100%	59.04%	92.57%
B-PCA	100%	72.89%	95.75%

Table 5 Type II Error for the USPC and SPE charts

	Data1	Data2	Data3
USPC	100%	77.53%	81.37%
V-PCA	92.09%	42.17%	66.03%
B-V-PCA	61.15%	10.24%	1.27%
B-PCA	39.57%	7.83%	1.06%

First of all, the faulty batch in Data1 is not detected at all by either USPC or T^2 charts, as shown in Table 4 (Type II error=100%). This observation confirms incipient nature of the fault. The most successful in

detecting this highly subtle fault is SPE chart of the B-PCA method. This is somewhat unsurprising since this chart is shown in Table 3 to be highly sensitive to normal process operation, i.e. having large Type I error. On the other hand, faulty batches corresponding to Data2 are detected by all of the four methods and most clearly by the SPE chart of the B-PCA method. Faulty batches of Data3 (simulated bioreactor) were detect to some extent by USPC and V-PCA. The clearest detection of the faults is provided by the SPE charts of B-V-PCA and B-PCA methods.

Overall, SPE chart of the B-PCA method is by far the most sensitive to both normal and abnormal process operation. Use of this chart is likely to result in many false alarms although it is also likely that this chart will be most consistent in terms of fault detection. SPE charts of all of the three multivariate methods are found to be more sensitive to process variation when compared to USPC and T^2 charts. In fact, no clear difference in terms of Type I and Type II error is observed between T^2 and USPC charts. Finally, performance of the B-V-PCA, particularly in terms of its SPE chart performance, is in between performances of V-PCA and B-PCA methods. Hence, B-V-PCA can be viewed as conciliation between V-PCA and B-PCA methods.

7. CONCLUSIONS

This paper presents results from three case studies in which both traditional univariate and advanced multivariate statistical analyses, based on the Principal Component Analysis, have been applied to bioreactors. Two standard multivariate methods are included in this comparison. These are the approach that preserves batch direction, B-PCA, and the one that preserves variable direction, V-PCA, during the unfolding procedure. Third multivariate method, B-V-PCA, is a hybrid of these two approaches. Two of the studies are conducted using industrial data while the third is performed using data from a simulated bioreactor. The main focus of this paper is to compare four different statistical monitoring approaches in terms of their respective ability to tolerate normal process operation while detecting anomalous process variation.

Results show that the Squared Prediction Error chart of the standard multivariate approach to batch process monitoring that unfolds data by preserving batch direction is by far the most sensitive of all the charts to both normal and abnormal process operation. Therefore, this chart needs to be used with great caution if frequent false alarms are to be avoided. On the other hand, no clear advantage of the so-called T^2 charts when compared to USPC charts is observed. Their performances are very similar while design and interpretation issues are much simpler in the case of the USPC chart. Finally, hybrid approach, B-V-PCA, that combines features of the two main multivariate methods, namely B-PCA and V-PCA, provides compromise in terms of Type I error and Type II error between these two standard approaches. Therefore, this new hybrid approach may prove to be the right compromise between conflicting objectives related to maximization of correct fault detection and minimization of false alarms, combining advantages and mitigating disadvantages of B-PCA and V-PCA.

REFERENCES

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

Chiang, L. H., R. Leardi, R. J. Pell and M. B. Seasholtz (2006). Industrial experiences with multivariate statistical analysis of batch process data. *Chemometrics and Intelligent Laboratory Systems*, Vol. **81**, pp. 109- 119.

Gurden, S. P., J. A. Westerhuis, R. Bro and A. K. Smilde (2001). A comparison of multiway regression and scaling methods. *Chemometrics and Intelligent Laboratory Systems*, Vol. **59**, pp. 121- 136.

Lee, J. M, C. K. Yoo and I. B. Lee (2004). Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis. *Journal of Biotechnology*, Vol. **110**, pp. 119- 136.

MacGregor, J. F. and P. Nomikos (1992). Monitoring batch processes. In: *Batch processing systems engineering: Current status and future directions, NATO ASI Series F* (Reklaitis, Rippin, Hortasco, Sunol. (Ed)), Vol. **143**, pp. 242- 258.

Nomikos, P. and J. F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, Vol. **37**, pp. 41- 59.

Van Sprang, E. N. M., H. J. Ramaker, J. A. Westerhuis, S. P. Gurden and A. K. Smilde (2002). Critical evaluation of approaches for on-line batch process monitoring. *Chemical Engineering Science*, Vol. **57**, pp. 3979- 3991.

Westerhuis, J. A., T. Kourti and J. F. MacGregor (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics*, Vol. **13**, pp. 397-413.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics*, Vol. **4**, pp. 397- 405.

Wold, S., N. Kettaneh, H. Friden and A. Holmberg (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory Systems*, Vol. **44**, pp. 331- 340.