MULTI-SITE PERFORMANCE MONITORING IN BATCH PHARMACEUTICAL PRODUCTION

C. W. L. Wong, R. E. A. Escott*, A. J. Morris and E. B. Martin

Centre for Process Analytics and Control Technology,
School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK
**GlaxoSmithKline Chemical Development, Tonbridge, UK*

Abstract: A challenge facing the pharmaceutical and chemical industries is how to understand and identify differences in process behaviour where a product is manufactured at two different sites. Three approaches based on multi-group principal component analysis are investigated and benchmarked against single site models. The multi-group approach is shown to remove differences between sites such as operational scale thereby enabling the analysis to focus on identifying differences in variation between the two sites that are not a consequence of process configurations. From the analysis it is observed that the multi-group approach can assist in the understanding of manufacturing performance. *Copyright © 2003 IFAC*

Keywords: Biotechnology, Manufacturing Processes, Performance Monitoring, Statistical Analysis

1. INTRODUCTION

Manufacturing challenges facing the chemical and pharmaceutical industries include the need to reduce the time between product development and full-scale production, the achievement of right-first-time manufacture and the manufacture of consistently high quality product with minimal environmental impact. The second and third challenges are compounded by the need to transfer the manufacture of a product to different sites around the world in a robust manner. A contribution to these challenges is to utilise the data collected from the process and to convert it into information and ultimately knowledge, thereby enabling an enhanced understanding of the process to be achieved. This approach has resulted in process performance monitoring and its associated techniques becoming an integral part of process operation.

For many industrial processes, performance monitoring systems are developed for individual process units, as opposed to the complete process. The complexity of this problem is compounded when the product is manufactured at two or more sites, where independent monitoring systems can be developed. A major disadvantage of this situation is that the sources of the differences in process operation and product variation, between the sites, cannot readily be identified. Previously it has been conjectured that process operation and scale differences are responsible for variability, and cannot be removed through modelling. In this paper the multi-group methodology of Lane *et al.* (2001) helps address this situation in terms of multi-site process performance monitoring. It is shown that scale and processing differences can be removed thereby enabling the real differences between sites to be identified. The paper focuses on empirical, i.e. data based, approaches. However alternative techniques are possible including the use of hybrid modelling, i.e. the conjunction of a reduced complexity mechanistic model and an empirical model (McPherson *et al.*, 2001).

One of the characteristics of batch operations is the variation in duration as a consequence of the process itself, down-stream processing, etc. To apply the techniques described in the paper, it is necessary to perform batch length equalisation. Multivariate Dynamic Time Warping (DTW) and the cutting of the batch process data to a minimum length are considered.

A number of approaches are considered in the paper for the development of a multi-site monitoring scheme for a drug intermediate. The benchmark approach was based on the development of an individual model for each site. The data matrices comprising the common variables from the two sites were then combined and different scaling procedures applied. The first resulted in the removal of the global mean and standard deviation of each variable (calculated from the data for the two sites) whilst for the second approach, the local mean and standard deviation for each individual variable for each site was removed. Finally a multi-group model based on the pooled sample variance-covariance matrix was developed using all the variables monitored at both sites. Fig. 1 provides an overview of the different approaches.

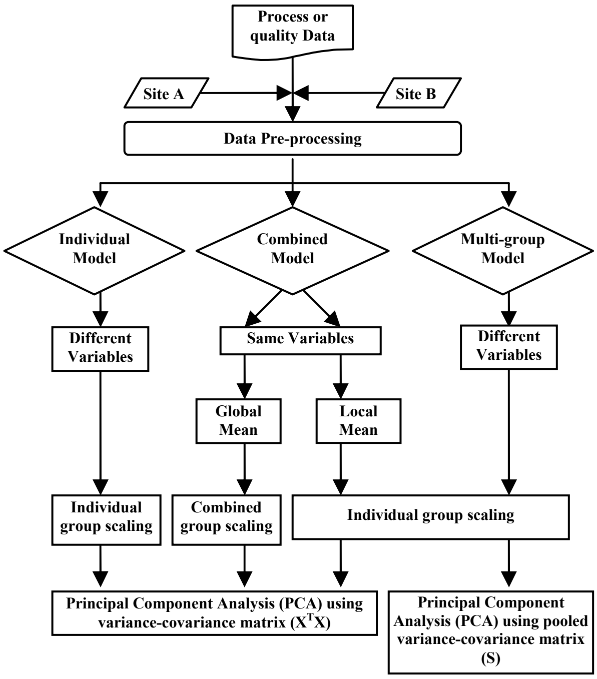

Fig. 1. Summary of different monitoring approaches.

2. PROCESS DESCRIPTION

The process interrogated is a single stage within a multi-stage synthetic route for the production of an active pharmaceutical ingredient (API). The process is carried out at two manufacturing sites by a regulated batch procedure. The process data have been acquired at both sites from reactor probes that are linked to data historians and that have been subsequently extracted for analysis. The chemistry step involves an exothermic addition that is controlled by reactant addition rate and the reactor temperature and has a duration period of approximately 4 hours. Although different plant configurations have been employed at the two sites, similar process variables are monitored, alongside coincident quality control measures. The process data variables include reactant addition rate (maturity), reactor temperature, reactor pressure, agitation rate and vapour temperature. The quality variables include input and output material activity, process yield and various impurity levels. Data from 57 batches from Site A and 152 batches from Site B were included in the analysis.

3. DATA PRE-PROCESSING

The raw data collected were initially pre-screened for missing observations, outliers, small signal to noise ratios, etc. Once data anomalies were identified, an appropriate in-filling algorithm was applied such as data deletion or linear interpolation. The next stage was to examine the resulting time series plots of the individual variables to attain good process operation understanding. It is essential that this stage is undertaken in collaboration with process personnel.

Batch process data collected on a number of batches is typically arranged in a three-way matrix, batch (I) x variables (J) x time (K). After equalisation of batch lengths, multi-way principal component analysis (MPCA) (Nomikos and MacGregor, 1994) was applied. The data matrix is first unfolded to give a two-dimensional array as shown in Fig. 2 and PCA is applied to the unfolded data matrix.

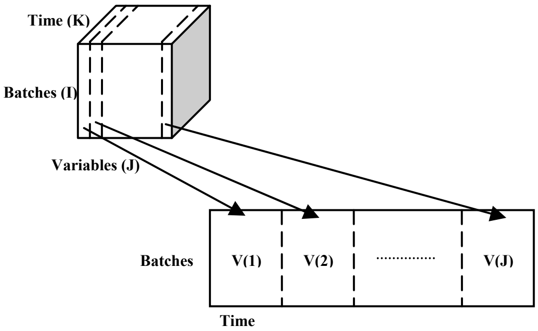

Fig. 2. Schematic representation of the unfolding of a three-way matrix.

To apply the bi-linear technique of multi-way principal component analysis illustrated in Fig. 2 batch lengths are required to be of equal duration. Two methods proposed to standardise batch length are cutting to a minimum length and multivariate Dynamic Time Warping (DTW) (Gollmer and Posten, 1996; Kassidas *et al.,* 1998). DTW is a method that matches features in a data pattern, or profile, to a reference profile. An optimal batch profile is first identified and the other batches are aligned against this reference batch. Fig. 3 illustrates the resulting synchronisation for the variables, reactor temperature and pressure for all batches at site A. Of particular note is the extraction of the underlying structure in the pressure variable that was masked prior to the application of DTW.

The second step was to remove data during periods of operation that were not deemed to be important in the subsequent analysis. For this specific application, the most important period with respect to product quality, is during the reactant addition period and hence this period defined the time period over which the data was analysed.

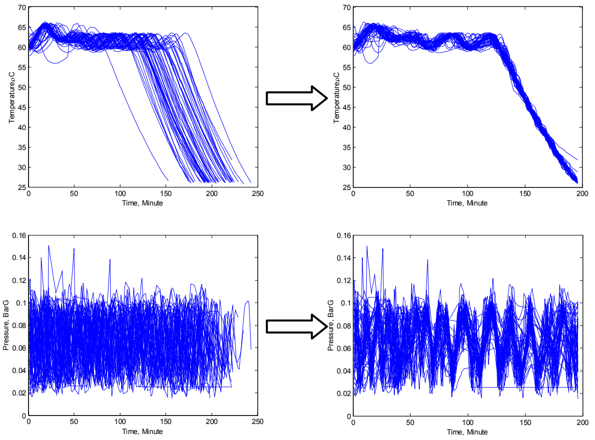

Fig. 3. Synchronisation of the time trajectories by DTW for reactor temperature and pressure for all batches at site A.

4. STATISTICAL DATA ANALYSIS

Both process and quality data were investigated but the results reported are for the process data only. A total of five process variables are monitored at site A and four at site B with three variables common to the two sites.

4.1 Individual PCA Model

Having pre-screened and equalised the duration of the batch data, the next step was to build individual multi-way PCA models for each site. By extracting the principal component score vectors, batch behaviour could be investigated. The leverage plot for the individual batches for the first two principal components, Fig. 4, clearly illustrates the impact of batch 15.

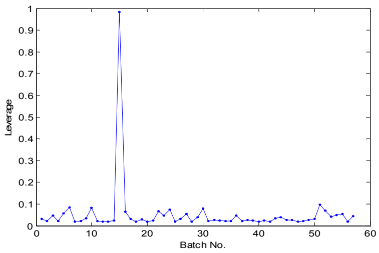

Fig. 4. Leverage for scores of principal component 1 and principal component 2.

By interrogating the data, it was observed that there had been an agitator failure during this batch run. Consequently to develop an appropriate monitoring model, it was necessary to remove batch 15 from the data matrix. From Fig. 4 it is not apparent that any other batches will have a major impact on the analysis, thus multi-way PCA was applied to the remaining 56 batches, Fig. 5. Ten principal components were retained in the subsequent analysis explaining 68% of the underlying variability. From Fig. 5 it can be observed that the scatter of the batches is random with a number lying out with the action limits. These batches were interrogated and issues relating to the data acquisition system were identified.

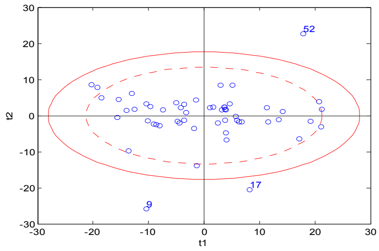

Fig. 5. Bivariate scores plot of principal component 1 and 2 after removal of batch 15.

From the loadings plot, process behaviour over time for different variables can be examined. Fig. 6 and 7 show the univariate loadings plot for principal component 1 and principal component 3, respectively. The dotted line is used to differentiate between the five variables through a batch run (reactor temperature, pressure, level, agitator speed and reactant addition rate). Variable three is observed to have a high loading throughout the duration of the batch for principal component one. It is interesting to observe from the loadings how the influence of variable changes over batch duration. This is particularly evident from principal component 3, Fig. 7.

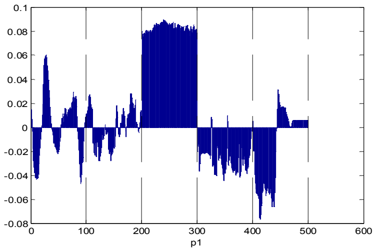

Fig. 6. Univariate loadings plot of principal component 1 for 5 process variables.

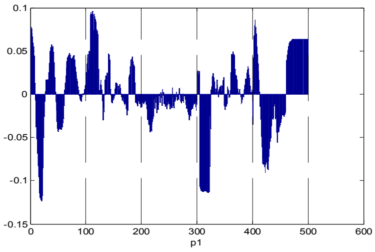

Fig. 7. Univariate loadings plot of principal component 3 for 5 process variables.

The same analysis was undertaken for site B. From examination of the leverage plot (not shown) two batches, 34 and 47, were identified as having high leverage. After the removal of these batches, a randomly scattered scores plot was obtained (Fig. 8). Retention of 10 principal components in this case resulted in 86% of the underlying variation in the data being explained. Examining the contribution plot of batch 127 (Fig. 9) for principal component one and the time series plots of the process variables it was noted that there was an abnormal reactant addition rate for this batch. This information can be used by process personnel who can either take corrective action or else ensure that subsequent batches are not affected by a similar problem.

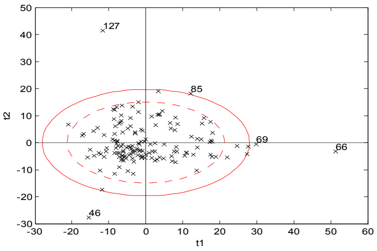

Fig. 8. Bivariate scores plot of principal component 1 and 2 after removal of batch 34 and 47.

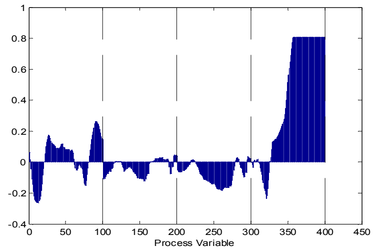

Fig. 9. Contribution plot for batch 127.

4.2 Combined PCA Model - Removal of Global Mean

The first combined model was constructed by applying multi-way PCA to the standardised data matrix based on the batch process data from the two sites. Only identical variables were selected to be included for analysis (reactor temperature, pressure and reactant addition rate). Examining Hotelling's T^2, three non-conforming batches were identified, batch 15 at site A and batches 17 and 125 at site B (not shown). It is interesting to observe that the batches from site B differed to those identified in the individual site analysis, demonstrating the potential limitation of this approach in terms of it providing conflicting information to the previous analysis. Following the removal of these batches, multi-way principal component analysis was applied to the remaining data, Fig. 10.

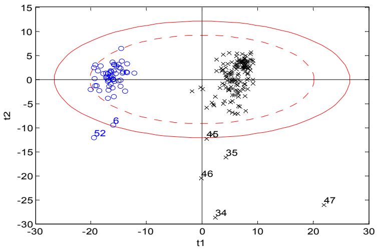

Fig. 10. Bivariate scores plot of principal component 1 and 2 after removal of batch 15 at site A and batch 17 and 125 at site B. Site A, 'o', Site B, 'x'.

From the figure, two clusters can be observed. More specifically, principal component 1 identifies the variation about the global mean for the two sites (Fig. 11) and thus both "within" and "between" group variation is captured.

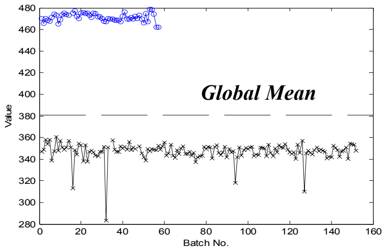

Fig. 11. Variation for one variable.

The lower order components do not exhibit this behaviour and display a more random scatter. Fig. 12 shows the bivariate scores plot of principal component 3 and principal component 4. A total of 77% of the variation was explained by the ten retained principal components.

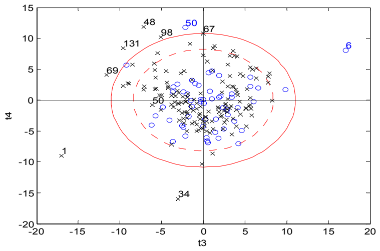

Fig. 12. Bivariate scores plot of principal component 3 and principal component 4.

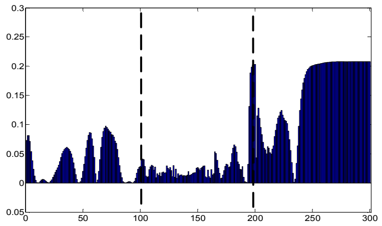

Fig. 13. Differential contribution plot between the two clusters.

Fig. 13 shows the differential contribution plot for the first principal component. The differential contribution plot calculates the difference between the contribution for a group of points from site A and a group of points from site B. From the resulting representation, it was observed that the differences were mainly related to the reactant addition rate. The rates and the total amounts of addition differ between the two sites due to operational differences, i.e. different reactor sizes and configurations. Thus it was conjectured that by removing the scale effect, a single model could realistically be developed for the two sites.

4.3 Combined PCA Model -Removal of the Local Mean

A second combined multi-way PCA model was built from the same data sets as in Section 4.2. However, the data was standardised specifically for each site. By standardising the data matrix in this way, the variation of each variable from its mean value relative to the individual site, Fig. 14, is considered.

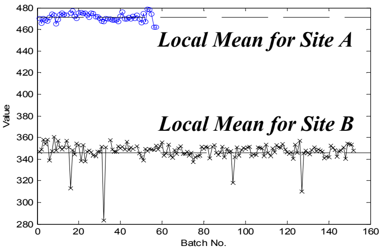

Fig. 14. Local variation for one variable.

From the bivariate scores plot of principal component 1 and principal component 2, batch 15 at site A and batch 34 and 47 at site B were again observed to have a strong influence on the process representation. Removing these batches, the subsequent analysis resulted in 75% of the underlying variation being explained following the inclusion of ten principal components.

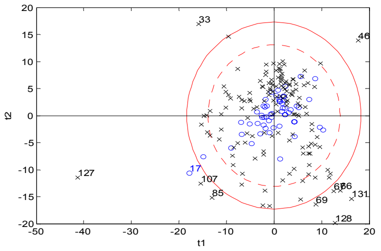

Fig. 15. Bivariate scores plot of principal component 1 and 2 after removal of batch 15 at site A and batch 34 and 47 at site B.

Examining the loadings plot, it can be observed that the key variable in terms of defining the main source of variation associated with principal component one is that of the reactant addition rate.

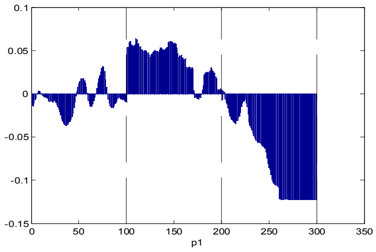

Fig. 15. Univariate Loadings plot of principal component 1.

4.4 Multi-group PCA Model

An extension to traditional multi-way PCA, multi-group multi-way PCA, was then investigated for the simultaneous monitoring of different manufacturing sites. Multi-group modelling is based on the assumption that a common eigenvector subspace exists for the sample variance-covariance matrix of individual sites. Through the pooled sample variance-covariance matrix, the principal component loadings are calculated. The pooled sample variance-covariance matrix (**S**), which forms the basis of the multi-group model is defined as a weighted sum of the *g* individual variance-covariance matrices $s_1, s_2, \ldots, s_g$:

$$S = \frac{(n_1 - 1)s_1 + (n_2 - 1)s_2 + \ldots + (n_g - 1)s_g}{(N - g)} \qquad (1)$$

for *i* = 1, …, *g*. *N* is the total number of observations (batches), *g* is the number of groups and n_i is the number of observations within group *i*. Consider the data set for site A, containing variables 1 to 5 and data set for site B comprising variables 1, 2, 3 and 6 in which variables 1, 2 and 3 are identical. The individual variance-covariance matrices for site A and B are given in Table 1 and the pooled variance-covariance matrix is defined in Table 2.

Table 1 Variance-covariance matrix for site A and B.

Variance-covariance matrix for Site A					
Variable	1	2	3	4	5
1	**A_{11}**	**A_{12}**	**A_{13}**	**A_{14}**	**A_{15}**
2	A_{12}	**A_{22}**	**A_{23}**	**A_{24}**	**A_{25}**
3	A_{13}	A_{23}	**A_{33}**	**A_{34}**	**A_{35}**
4	A_{14}	A_{24}	A_{34}	**A_{44}**	**A_{45}**
5	A_{15}	A_{25}	A_{34}	A_{45}	**A_{55}**

Variance-covariance matrix for Site B				
Variable	1	2	3	6
1	**B_{11}**	**B_{12}**	**B_{13}**	**B_{16}**
2	B_{12}	**B_{22}**	**B_{23}**	**B_{26}**
3	B_{13}	B_{23}	**B_{33}**	**B_{36}**
6	B_{16}	B_{26}	B_{36}	**B_{66}**

Table 2 Pooled variance-covariance matrix.

Pooled variance-covariance matrix						
Variable	1	2	3	4	5	6
1	**C_{11}**	**C_{12}**	**C_{13}**	**C_{14}**	**C_{15}**	**C_{16}**
2	C_{12}	**C_{22}**	**C_{23}**	**C_{24}**	**C_{25}**	**C_{26}**
3	C_{13}	C_{23}	**C_{33}**	**C_{34}**	**C_{35}**	**C_{36}**
4	C_{14}	C_{24}	C_{34}	**C_{44}**	**C_{45}**	**C_{46}**
5	C_{15}	C_{25}	C_{35}	C_{45}	**C_{55}**	**C_{56}**
6	C_{16}	C_{26}	C_{36}	C_{46}	C_{56}	**C_{66}**

where

$$C_{11} = \frac{(n_1 - 1)A_{11} + (n_2 - 1)B_{11}}{(n_1 + n_2 - 2)}$$

$$C_{12} = \frac{(n_1 - 1)A_{12} + (n_2 - 1)B_{12}}{(n_1 + n_2 - 2)}$$

$$C_{13} = \frac{(n_1 - 1)A_{13} + (n_2 - 1)B_{13}}{(n_1 + n_2 - 2)}$$

$C_{14} = A_{14}$
$C_{15} = A_{15}$
$C_{16} = B_{16}$

$$C_{22} = \frac{(n_1 - 1)A_{22} + (n_2 - 1)B_{22}}{(n_1 + n_2 - 2)}$$

$$C_{23} = \frac{(n_1 - 1)A_{23} + (n_2 - 1)B_{23}}{(n_1 + n_2 - 2)}$$

$C_{24} = A_{24}$
$C_{25} = A_{25}$
$C_{26} = B_{26}$

$$C_{33} = \frac{(n_1 - 1)A_{33} + (n_2 - 1)B_{33}}{(n_1 + n_2 - 2)}$$

$C_{34} = A_{34}$

$C_{35} = A_{35}$

$C_{36} = B_{36}$
$C_{44} = A_{44}$
$C_{45} = A_{45}$
$C_{46} = 0$

$C_{55} = A_{55}$

$C_{56} = 0$
$C_{66} = B_{66}$

Multi-group PCA was then applied to the pooled variance-covariance matrix. Batch 15 from site A and batch 34 and 47 from site B were removed from the analysis as they have a major influence on the model. Reapplying multi-group multi-way PCA resulted in 63% of the variation being explained by ten principal components, Fig. 16.

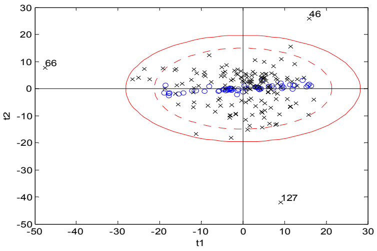

Fig. 16. Bivariate scores plot of principal component 1 and 2 of multi-group model after removal of batch 15 at site A and batch 34 and 47 at site B.

From the loadings plot, Fig. 17, variable three, reactant addition rate, was identified as the most important variable in terms of defining the main source of variation for principal component one. This variable was one of the three common to the two sites along with variable one and two, reactor temperature and pressure. Variable four and five related to those monitored only at site A, level and agitator speed, and variable six related to vapour temperature that was only monitored at site B.

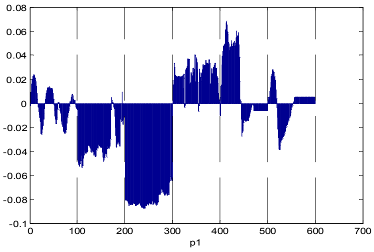

Fig. 17. Univariate loadings plot of principal component 1.

The advantage of being able to develop a single model for two, or more, sites is that it enables an enhanced understanding of the subtle differences in performance between the two manufacturing processes. In addition it can help facilitate the transfer of a process to a new site by providing a baseline monitoring model with the model being updated as new batches are manufactured. The scores plot clearly detects those batches which move outside the statistical control region for the two sites on one chart and the corresponding scores contribution plots identifies the combination of variables responsible for the out of control signal. Thus, the application has demonstrated that the multi-group model has acceptable detection and diagnostic properties though the overall sensitivity may be affected compared with those of the corresponding individual plant models.

5. CONCLUSION

The capabilities of multi-group models, to model different process configurations on two sites, based on the pooled sample variance-covariance matrix has been demonstrated by its application to data from a drug intermediate batch process. Pre-screening of the data was initially performed to remove any abnormal variability. Batch length equalisation was achieved through the application of multivariate DTW to the process data. The DTW batch data was further reduced to ensure that the analysis focused on the main area of interest. Multi-way principal component analysis was then applied to the pre-processed data. The first approach used analysed the data from each plant individually. Two combined models where the data was scaled differently were also studied. The multi-group models developed not only eliminates between cluster variations but also allows the process monitoring of two different plants by a single model. This development provides a powerful monitoring tool for understanding and hence minimising the differences in product quality and process operation across different manufacturing plants. In addition based on the proposed approach, it is possible to utilise the approach to assist in the transfer of a process to a new site.

6. ACKNOWLEDGEMENTS

Chris Wong would like to acknowledge the EPSRC, GlaxoSmithKline, the UK ORS Scheme and CPACT for financial support of his PhD.

7. REFERENCES

Gollmer, K. and C. Posten (1996). Supervision of bioprocesses using a dynamic time warping algorithm. *Control Engineering Practice* **4(9)**, pp. 1287-1295.

Jolliffe, I.T. (1986). *Principal component analysis*. Springer-Verlag, New York.

Kassidas, A., J. MacGregor and P.A. Taylor (1998). Synchronisation of batch trajectories using dynamic time wraping. *AIChE* Journal, **44(4)**, pp. 864-875.

Lane, S., E.B. Martin, R. Kooijmans and A.J. Morris (2001). Performance monitoring of a multi-product semi-batch process. *Journal of Process Control*, **11**, pp. 1-11.

Martin, E.B., A.J. Morris and C. Kiparissides (1999). Manufacturing performance enhancement through multivariate statistical process control. *Annual Reviews in Control* **23(1)**, pp. 35-44.

McPherson, L.A., E.B. Martin and A.J. Morris (2002). Super model-based techniques for batch performance monitoring, *ESCAPE-12*, pp. 523-528.

Nomikos, P. and J. MacGregor (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal* **40(8)**, pp. 1361-1373.