Copyright © 2002 IFAC
15th Triennial World Congress, Barcelona, Spain

MONITORING OF BATCH PHARMACEUTICAL FERMENTATIONS: DATA SYNCHRONIZATION, LANDMARK ALIGNMENT, AND REAL-TIME MONITORING

Cenk Ündey, Bruce A. Williams, and Ali Çınar [1]

Illinois Institute of Technology
Department of Chemical and Environmental Engineering
10 W 33rd Street, Chicago, IL, 60616 USA

Abstract: Most batch pharmaceutical fermentations have successive phases of operation. Detection of process phase landmarks is important for improving the performance of process monitoring and control. This study considers techniques for adjustment of batch data lengths to match landmarks of phases during the progress of the batch. Time synchronization and landmark alignment techniques are presented and their integration with on-line monitoring is discussed through illustrative examples. *Copyright © 2002 IFAC.*

Keywords: On-line process monitoring, time alignment, curve registration, dynamic time warping

1. INTRODUCTION

Batch fermentations of pharmaceuticals usually have complex reaction mechanisms and non-linear, time-variant process dynamics that make their modeling, monitoring and control challenging. In a batch pharmaceutical fermentation process that lasts several days, some organisms may have generation times that are shorter than one hour. Slight changes in operating conditions during critical periods may have a significant influence on growth and differentiation of organisms, and impact final product quality and yield. Changes in raw material quality and impurity levels in the feed also affect the final product. Furthermore, most batch fermentations proceed through a number of production phases. Fluctuations in operation may cause variations in both temporal occurrence and magnitude of process events. The landmarks for these phases may shift in time for various runs and impact the computation of precise reference mean trajectories. Consequently, batch fermentation data sets contain unequal and unsynchronized data for various batch runs that need to be pre-processed prior to multivariate modeling to prevent inconsistent statistical results and possibility of false alarms in multivariate statistical control charts.

In this work, data length equalization and trajectory synchronization techniques are integrated with on-line multivariate statistical monitoring framework for monitoring batch fermentations in the pharmaceutical industry. While the methods have been used with industrial data, their performance will be illustrated by using simulated fed-batch penicillin fermentation data. The data length adjustment and phase landmark alignment techniques include: (1) indicator variable, (2) dynamic time warping (DTW), and (3) curve registration. Curve registration is based on functional data analysis where batch trajectory data are interpreted as sampled from continuous functions. Functional data analysis and curve registration are discussed and their performance are compared with indicator variable and DTW methods.

[1] Corresponding author (e-mail : *cinar@iit.edu*).

Multiway principal components analysis (MPCA) technique is used to develop empirical models out of time-aligned fermentation data for analyzing completed batches and adaptive hierarchical PCA (AHPCA) is used for on-line monitoring. The number of false alarms has been reduced after time alignment in penicillin fermentation case studies. Another important ramification of the integration of time alignment algorithms with MSPM framework is the availability of information about the locations of the important process events including microbial phases in each batch. This information can be used to determine the necessary control actions as well as the required operating policy changes that apply for different microbial phases to enhance overall productivity.

Different approaches are found in the literature for the physiological phase detection in fermentation processes based on knowledge-based pattern recognition and fuzzy logic (Konstantinov and Yoshida, 1992). These solutions rely on the use of temporal shape libraries and extensive rule bases to explain physiological alterations.

The presentation will focus on the description of various landmark detection and alignment techniques, the integration of trajectory alignment and multivariate statistical process monitoring (SPM), and the illustration of the proposed unified framework by monitoring penicillin fermentation by using dynamic models and data generated by simulations. Case studies illustrate the advantages of the proposed techniques. Fed-batch penicillin fermentation is used as a case study. Process data are simulated using the modified mechanistic model of Bajpai and Reuss (1980). Details of the extended process model and the simulator are reported by Undey *et al.* (2000).

2. TIME NORMALIZATION TECHNIQUES

Batch fermentation processes are often accompanied by physiological phase changes. The landmarks of the process reflect physiological transition points that are to be used both to monitor the biological evolution of the process and to decide upon the appropriate control strategy. The occurrence times of these transitions represent variability due to expected or unexpected alterations in cell behavior. Some of these landmarks may be delayed and others are advanced in different batch runs. For instance, the first landmark of a variable x from a hypothetical batch run c_1 (dashed line) occurs at $t = \tau_1(c_1)$ and the second at $t = \tau_2(c_1)$ whereas in another run operated under the same conditions the occurrence of the first landmark of x is advanced and the second delayed such that $\tau_1(c_2) > \tau_1(c_1)$ and $\tau_2(c_2) < \tau_2(c_1)$.

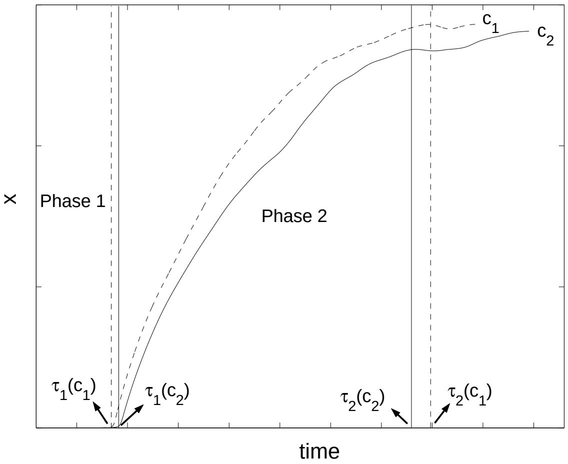


Fig. 1. Physiological phase differences.

To develop a more sound SPM framework, those unsynchronized curves should be aligned. This will provide a consistent comparison of process observations since the time axis difference are minimized. Another important benefit of this preprocessing will be the equalization of the batch lengths to a common duration that is necessary prior to matrix and vector calculations used in empirical model development for SPM.

Indicator Variable Technique (IVT) is based on selecting a process variable to indicate the progress of the batch instead of time. This variable should be chosen such that it also shows the *maturity* or *percent completion* of each batch. Some candidates are percent conversion or percent of a component fed to the fermenter. A measure of the *maturity* of a batch is provided by the percentage of its final value attained by the indicator variable at the current point in time. Observations about the progress of all other variables are taken relative to the progress of the indicator variable (IV). The indicator variable should be smooth, continuous, monotonic and spanning the range of all other process variables within the batch data set (i.e. $iv(t_k) > iv(t_{k-1})$ for monotonic increase, where iv denotes the indicator variable). Linear interpolation techniques are used to transform batch-time dimension into indicator variable dimension. For monitoring new batches, data are collected from all process variables and adjusted with respect to the indicator variable such that $\mathbf{x} \rightarrow \mathbf{x}(t_k)$, where $\mathbf{x}$ denotes process variables and t_k the corresponding time stamp of the indicator variable sampling instance. This technique has been used for batch/semi-batch polymerization and batch fermentation processes, where reaction extent or percent of a component fed are used as indicator variable (Kourti *et al.*, 1996; Neogi and Schlags, 1998; Rothwell *et al.*, 1998). IVT is appealing because of its ease of implementation, but it does not account for the locations and the alignment of physiological landmarks. Another likely problem with IVT is the lack of an appropriate

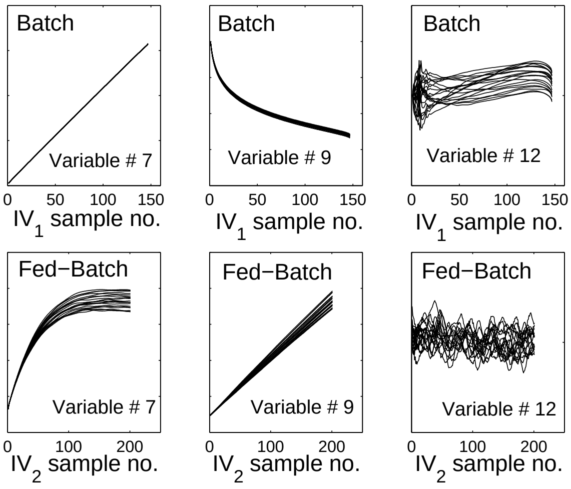


Fig. 2. Equalized batch lengths using mixed-IVT. Indicator variable 1: Substrate concentration decrease during batch operation (upper figures), 2: percent substrate added during fed-batch operation (lower figures)

candidate for a single indicator variable for all phases of a batch. In such cases, further processing is required to apply the technique such as dividing process phases so that an indicator variable can be identified in each phase. This mixed approach was applied in the penicillin case study for IVT based SPM. The difficulty with the batch/fed-batch operation is the lack of an appropriate variable that spans the whole range of process variables. Process data from 55 successful batches were divided into two parts representing batch operation and followed by the fed-batch operation. Two different IVs are chosen for each operation phase in penicillin fermentation and results are shown for variables 7, 9 and 12, biomass concentration, culture volume and fermenter temperature, respectively. For the batch operation, substrate concentration decrease is found to be the best candidate as an indicator variable (upper figures in Figure 2). Percent addition of substrate into the fermenter is selected as indicator variable for the fed-batch operation (lower figures in Figure 2).

Dynamic Time Warping (DTW) has its origins in speech recognition and is a flexible, deterministic, pattern matching scheme which works with pairs of patterns. It is able to locally translate, compress, and expand the patterns so that similar features in the patterns are matched. DTW nonlinearly warps two trajectories in such a way that similar events are aligned and a minimum distance between them is obtained. Basic description of DTW and different algorithms are given by Sakoe and Chiba (1978) and Rabiner *et al.* (1978). Gollmer and Posten (1996) have implemented this technique to detect process phase changes and faults in fed-batch *E. coli* fermentations. A recent application of DTW for monitoring and diagnosis in a batch polymerization process has been reported by Kassidas *et al.* (1998).

The objective of DTW is to find the nonlinear mapping function $C(k) = [i(k), j(k)]$, $k = 1, \ldots, K$ between two multivariate observation sets, the reference set $\mathbf{R}$ $(m \times p)$ and the test set $\mathbf{T}$ $(n \times p)$, where p denotes the number of variables, and m and n the number of observations on each set. This is done subject to a set of path and end point constraints to minimize the following accumulated distance

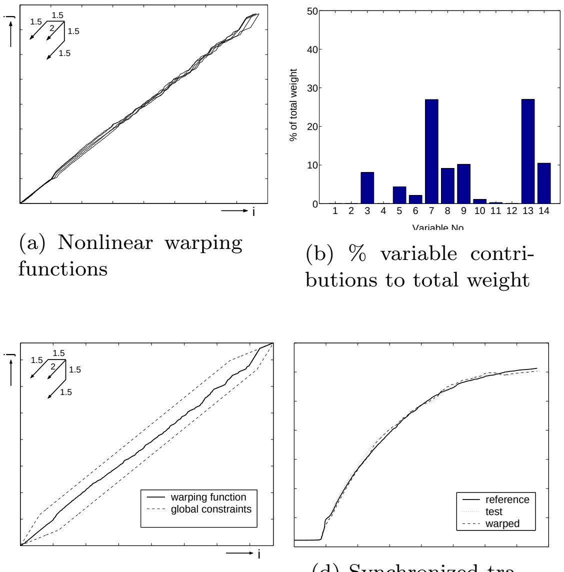
$$D^*(C) = \frac{1}{N(w)} \min_C \sum_{k=1}^{K} d[i(k), j(k)] w(k) \qquad (1)$$

where $w(k)$ denotes a weighting function that is used to impose local continuity constraints. A symmetric and smoothed version of this function is used in this work because it gave better performance. While the choice of $w(k)$ is arbitrary, it depends on the degree of allowable warping for a particular application. Neither too steep, nor too gentle local moves on the warping path should be allowed. $N(w)$ is a normalization factor. $d[i(k), j(k)]$ is the distance between the two points in test and reference sets. Mahalanobis distance can be used as a measure of local similarity between the point in signal $\mathbf{T}(i,p)$ and the reference point $\mathbf{R}(j,p)$ as follows

$$d[C(k)] = [\mathbf{T}(i(k), :) - \mathbf{R}(j(k), :)] \mathbf{W} [\mathbf{T}(i(k), :) - \mathbf{R}(j(k), :)]^T$$

where a $(p \times p)$ positive definite matrix $\mathbf{W}$ is reflecting the relative importance of the variables preferably based on their resemblance to time axis. Multivariate DTW is applied to batch trajectories iteratively so that $\mathbf{W}$ is updated at each iteration to align patterns more precisely. After each iteration $\mathbf{W}$ is adjusted so that variables that show smaller deviations from the mean profiles are given higher weights. Figure 3(b) shows % variable contributions to $\mathbf{W}$ after 10th iteration. Variables 7 (biomass concentration) and 13 (generated heat by biomass production) received the highest contributions since their profiles are smoother, monotonically increasing and continuous hence resembling the time axis to some degree than the rest of the variables.

Curve Registration casts the landmark alignment problem in the functional data analysis (FDA) framework (Ramsay and Silverman, 1997). FDA involves the estimation of mth order linear differential operators $L = w_0 I + w_1 D + \ldots + w_{m-1} D^{m-1} + D^m$ where $Lx = 0$, D^m denotes the mth derivative, and the weights w are functions of the independent variable t. Let N functions x_i be defined on closed real interval $[0, T_0]$ and $h_i(t)$ be a transform of t for the case i with domain $[0, T_0]$. The time of events must remain in the same order regardless of the time scale, $h_i(t_1) > h_i(t_2)$ for $t_1 > t_2$. Define $y(t)$ to be a fixed (reference)



(a) Nonlinear warping functions

(b) % variable contributions to total weight

(c) A sample warping function

(d) Synchronized trajectories

Fig. 3. Multivariate DTW example

function defined over $[0, T_0]$ to act as a template (for example a reference batch trajectory) for individual curves x_i such that after registration, the features of x_i will be aligned with the features of y. In discrete values y_i, $k = 1, \ldots, K$,

$$y_i = x_i\left[h_i(t_k)\right] + \epsilon_{ik} \qquad (2)$$

where ϵ_{ik} is a small residual relative to x_i and roughly centered about 0 (Ramsay and Silverman, 1997). The curve registration task is to determine the time warping functions h_i so that trajectories $x_i[h_i(t_k)]$ can be interpreted more accurately. The h_i can be determined by using a smooth monotone transformation family consisting of functions that are strictly increasing (monotone) and have an integrable second derivative (Ramsay, 1998):

$$D^2 h = qDh \qquad (3)$$

A strictly increasing function has a nonzero derivative and consequently the weight function $q = D^2h/Dh$ or the curvature of h. h can be estimated by minimizing a measure of the fit Υ_η of $x_i[h_i(t_k)]$ to y. A penalty term in Υ_η based on q permits the adjustment of the smoothness of h_i (Ramsay and Silverman, 1997). To estimate the warping function h_i, one minimizes

$$\Upsilon_\eta(y, x|h) = \sum_{\ell=0}^{L} \int \alpha_\ell(t)\, \left\| D^\ell y(t) - D^\ell x\left[h(t)\right] \right\|_\ell^2 \, dt + \eta \int q^2(t)dt \qquad (4)$$

where $\alpha_\ell(t)$'s are weight functions, and L denotes the highest order of the derivative

$$\left\| D^\ell y(t) - D^\ell x\left[h(t)\right] \right\|_\ell^2 = (D^\ell y(t) - D^\ell x\left[h(t)\right])^T \mathbf{W}_\ell (D^\ell y(t) - D^\ell x\left[h(t)\right]). \qquad (5)$$

The weight matrices $\mathbf{W}_\ell$'s allow for general weighting of the elements and the weight functions $\alpha_\ell(t)$'s permit unequal weighting of the fit to a certain target over time (Ramsay and Silverman, 1997). η adjusts the penalty on the degree of smoothness. B-splines $q(u) = \sum_{p=0}^{P} c_p B_p(u)$ are used in this study as the polynomial basis for performing the curve registration because calculating the coefficients of the polynomial is well defined. In addition, when estimating the solution to transforming particular waveforms into the B-spline domain, the required number of calculations increases linearly with the number of data points (Ramsay, 1998). The derivative of Υ_η with respect to the B-spline coefficient vector $\mathbf{c}$ is

$$\frac{\partial \Upsilon_\eta(y, x|h)}{\partial \mathbf{c}} = -2 \sum_{\ell=0}^{L} \alpha_\ell(t) \frac{\partial h(t)}{\partial \mathbf{c}} \left[\frac{\partial D^\ell x(h)}{\partial h} \right]^T \mathbf{W}_\ell \times (D^\ell y(t) - D^\ell x[h(t)])dt + \eta \int \left(\frac{\partial q(t)}{\partial \mathbf{c}} \right)^2 dt \qquad (6)$$

The derivative $[\partial D^\ell x(h)/\partial h]$ must be estimated with a smoothing technique to ensure monotonic increasing (Ramsay and Silverman, 1997).

To determine the number of landmarks and their locations from a set of trajectories, process knowledge and/or numerical techniques can be used. The challenge of implementing multivariate landmarking is that landmarks may be different (in number and location) for different process variables. Critical issues are the selection of landmarks among the process variables that define the phase phenomena of the process, and the number of landmarks to define clearly the progress of the batch. One solution to these issues is to use an iterative approach which will reconcile the identification of process landmarks with respect to particular trajectory landmarks. This procedure could be implemented as follows:

(1) Find the landmarks (ℓm) of the most important variable trajectory ℓm_1. Align all other variable trajectories with respect to the landmarks ℓm_1.
(2) Calculate the principal components of the aligned set of process variables. Determine the landmarks of the first principal component ℓm_{PCA}.
(3) Realign the process trajectories with respect to ℓm_{PCA}.
(4) Recalculate the principal components of the realigned set of process variables. Determine the landmarks of the first principal component ℓm_{PCAnew}.
(5) Determine if ℓm_{PCAnew} are reasonably close to ℓm_{PCA}. If so, the process landmarks are defined by ℓm_{PCAnew}. If not, then return to Step 3.

The implementation of this procedure will also depend upon the interpretation of the results. Once ℓm_{PCAnew} has converged, one may proceed with statistical analysis using the data warped with respect to ℓm_{PCAnew}. As an alternative, only the data identified as "most significant" (either by user or principal components) may be warped with respect to ℓm_{PCAnew}, and other process data may be warped with respect to its own optimal landmarks.

When landmarking a test trajectory with respect to a reference trajectory, two distinct cases occur. The first case is called uniform landmark case because all the landmarks are delayed (or advanced) by a constant time. The second is the mixed case that represents a more general framework where some landmarks are delayed and others are advanced yielding a more challenging landmark detection problem (Figure 1). Furthermore, the time shifts of the landmarks will vary, preventing the use of an assumption of a constant time shift τ between calculated and mean-value landmarks.

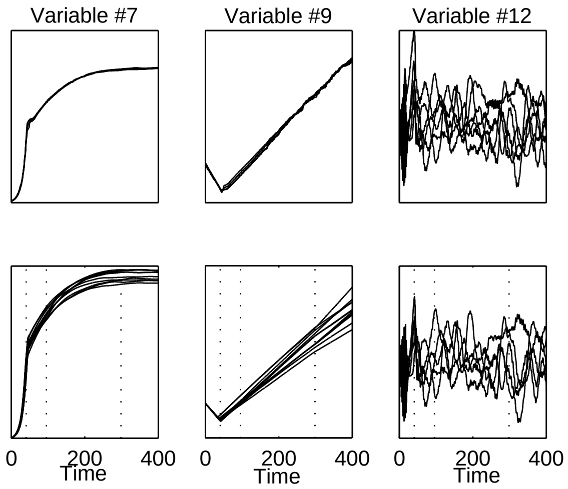


Fig. 4. Comparison of alignments of several profiles in the reference set using DTW (upper figures) and curve registration (lower figures).

3. INTEGRATION AND DEPLOYMENT OF TIME ALIGNMENT TECHNIQUES IN ON-LINE SPM FRAMEWORK

Once the batch trajectories are equalized and synchronized using one of the techniques explained in Section 2, an on-line real time SPM framework can be developed upon integration of these techniques into monitoring methods. Since batch processes generate three-dimensional arrays (Figure 5), equalized and synchronized data from the batches are arranged into a three-dimensional array $\mathbf{X}$ $(I \times J \times K)$ where I is the number of batches, J is the number of variables and the K is the number of sampling times in a given batch.

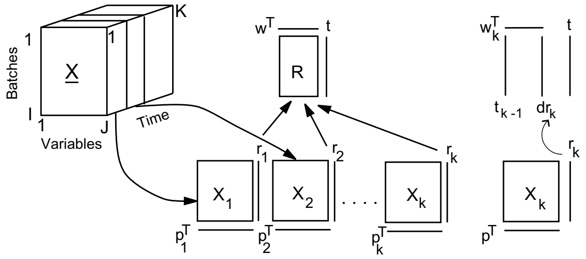


Fig. 5. Batch data representation, unfolding process and adaptive hierarchical PCA technique

The on-line monitoring method used in this work is based on a variation of MPCA technique called adaptive hierarchical PCA (AHPCA) recently suggested by Rannar *et al.* (1998). This technique also works with the unfolded and properly scaled three-way array of normal operation (NOC) data except one important difference that it does not require the estimates of the future values.

The integration of time alignment techniques with AHPCA is performed in two stages. In model development stage, the reference set (NOC data) is equalized/synchronized and an AHPCA model is developed using these preprocessed data (Figure 5). The monitoring stage differs depending on the synchronization technique used. If an indicator variable is used for time alignment, each variable is sampled whenever sampling is made on the indicator variable. DTW is implemented by using an expanding window. As new data become available, DTW aligns new points with the reference trajectory. It will produce more accurate results as the batch progresses. Both IVT and DTW techniques are implemented regardless of the landmarks. Curve registration technique however, aligns the new observations based on the landmarks. Details on MPCA and AHPCA can be found elsewhere (Wold *et al.*, 1987; Rannar *et al.*, 1998).

Both DTW and curve registration techniques align the reference batches successfully to a uniform batch length (Figure 4). The DTW aligned curves for variables 7 (biomass) and 9 (volume) are not smooth, whereas the curve registration aligned curves show smooth behavior during all process phases. This is because DTW batch length reconciliation is completed by skipping or repeating points. The more abrupt curve alignment of DTW also affects the precise monitoring of new batches.

In an instance where there is a ramp decrease in the substrate feed rate at 150 hours (Figure 6), the T^2 chart using DTW generates many large false alarms with large T^2 values before the occurrence of the fault, whereas the T^2 values based on curve registration yield fewer smaller false alarms in the initial 100 hours, before the fault occurs. Both methods detect correctly the fault around 200 hours, when the drift of substrate feed rate reaches values outside of normal operation. In new batch runs where the batch length is significantly different than the mean batch length, regularization method reduces significantly the number of false alarms. When phase detection is implemented, smooth phase transitions produce aligned multivariable batch runs so that SPM will not give false alarms.

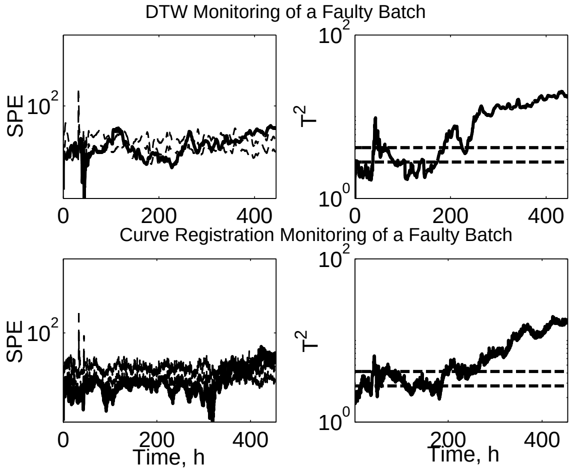


Fig. 6. MSPM charts for a faulty batch run aligned using DTW (upper frames) and curve registration (lower frames).

4. CONCLUSIONS

The emergence of batch process operations in many high value-added manufacturing operations increased the importance of rapid and accurate monitoring of batch process operations. A number of time alignment techniques have been integrated into on-line real time statistical process monitoring framework. Given the on-line monitoring techniques, the process should be monitored while equalizing the different batch lengths on-line by implementing a unified monitoring framework. We propose a regularization procedure based on mixed landmarking to be implemented along with an AHPCA structure. Such monitoring systems can handle phase landmarks and fault detection as well as fault diagnosis activities. Integration of landmark detection and time alignment techniques provides more effective process supervision. This information can be used to determine the necessary control actions and operating policy changes for different microbial phases to enhance overall productivity in pharmaceutical processes.

Acknowledgement. Financial support provided by NSF (BES-0084749) is gratefully acknowledged.

5. REFERENCES

Bajpai, R.K. and M. Reuss (1980). A mechanistic model for penicillin production. *J. Chem. Technol. Biotechnol.* **30**, 332–344.

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

Kassidas, A., J.F. MacGregor and P.A. Taylor (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE Journal* **44**(4), 864–875.

Konstantinov, K. B. and T. Yoshida (1992). Real-time qualitative analysis of the temporal shapes of (bio)process variables. *AIChE Journal* **38**(11), 1703–1715.

Kourti, T., J. Lee and J.F. MacGregor (1996). Experiences with industrial applications of projection methods for multivariate statistical process control. *Comp. and Chem. Engng.* **20**(Suppl. A), 745.

Neogi, D. and C. Schlags (1998). Multivariate statistical analysis of an emulsion batch process. *Ind. Eng. Chem. Res.* **37**(10), 3971–3979.

Rabiner, L.R., A.E. Rosenberg and S.E. Levinson (1978). Consideration in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. on Acoustics, Speech and Signal Process.* **6**(26), 575.

Ramsay, J.O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society - Series* B **60**, 365–375.

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis.* Springer-Verlag.

Rannar, S., J.F. MacGregor and S. Wold (1998). Adaptive batch monitoring using hierarchical PCA. *Chemometrics Intell. Lab. Syst.* **41**, 73–81.

Rothwell, S.G., E.B. Martin and A.J. Morris (1998). Comparison of methods for handling unequal length batches. In: *IFAC DYCOPS5.* Corfu, Greece. pp. 66–71.

Sakoe, H. and S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech and Signal Process.* **2**(26), 43–49.

Undey, C., G. Birol, I. Birol and A. Cinar (2000). An educational simulation package for penicillin fermentation. In: *AIChE Annual Meeting.* Los Angeles, CA.

Wold, S., P. Geladi, K. Esbensen and J. Ohman (1987). Multi-way principal component and PLS analysis. *Journal of Chemometrics* **1**, 41–56.