FORECASTING FOR FERMENTATION OPERATIONAL DECISION MAKING

G.A. Montague and E.B. Martin

School of Chemical Engineering and Advanced Materials,
*Newcastle University, Newcastle upon Tyne, NE1 7RU, Engla*nd.

Abstract: An awareness of the likely future behaviour of a batch or fed-batch fermentation process is valuable information that can be exploited to improve product consistency and maximise profitability. For example, by making operational policy changes in a feedforward control sense, improved consistency can be facilitated, whilst prior knowledge of batch productivity, or the end time, can help determine the downstream processing configuration and upstream process scheduling. In this paper, forecasting methods based on multivariate batch statistical data analysis procedures are contrasted with case based reasoning (CBR). Two case studies are considered, fed-batch pharmaceutical fermentation and batch beer fermentation. It is demonstrated that following appropriate statistical pre-screening of the data, CBR is comparable to linear projection to latent structures (PLS) for the more straightforward forecasting problem whilst for the more complex problem, CBR is preferable. Copyright © 2007 IFAC

Keywords: Industrial case studies; Projection to latent structures; Case based reasoning

1. INTRODUCTION

Batch to batch variation is an issue that all fermentation using industries face. Some degree of variation is inevitable for processes employing biological steps but for large scale fermentation processes, that make use of undefined media components, the variation can be more significant. Variation is accepted, to an extent, through its accommodation within validated bioprocess operating windows, for those processes subject to formal validation. In these, and other, bioprocesses, variability is a problem that is tackled to a degree with on-line feedback control. For example, the regulation of the environment of the fermenter through temperature, pH and dissolved oxygen control is one means of reducing variability but such schemes in some sense tackle the symptoms as opposed to the causes of variation.

Tackling the problem from a control engineer's perspective, the first step is to identify the sources of variation, thereafter a means by which their impact can be mitigated are investigated and finally, with reduced variability achieved, operational modifications to enhance financial performance can be made. In fermentations, disturbances to process operation arise from changes in seed state (Ignova *et al*, 1999), media composition and environmental conditions / operational policy. The latter problem is minimised by feedback regulation of the environment and strict adherence to standard operating policy by process operators. To minimise seed state variation there again needs to be strict adherence to operating policy and the adoption of seed transfer criteria that are linked to the physiological state rather than simply elapsed time (Neves *et al*, 2001). Media composition variations arise naturally in undefined media and tend to be a source of change in bioprocess performance. To minimise the impact, often the number of suppliers are few and raw material assessed with spectroscopic measurements to identify suitability based on a raw material fingerprint (Scarff *et al*, 2006). Such measures are now more widely applied to determine media quality and also identify fermentation endpoint. The reason that endpoint detection is necessary is that despite stringent operating policies and practices, variability still occurs. This can compromise decision making with regard to overall process plant behaviour and impact on financial performance with uncertainty leading to cautiousness.

The ideal control engineering solution to problems caused by variation is to employ a feedforward compensation policy, where the impact of measured disturbances is predicted using a model and mitigated using appropriate process manipulations. In the fermentation case, the presence of disturbances is detected not through their direct measurement but through their impact on performance thus enabling the exploitation of a feedforward strategy. If the future behaviour of batches can be anticipated then this information can be used proactively to make operational decisions that maximise batch performance. Methods to achieve potential improvements in fermentation batch profitability are discussed in Xue and Yuan (2005). Whilst their approach considers the strategies to maximise batch profitability, it is critical to consider the process as a whole. Moving towards process optimisation, rather than batch optimisation, requires a predictive capability so that, for example, seed vessels can be scheduled and downstream processing operations operated to capacity. This requires prior knowledge of batch performance that can only be gleaned from batch behaviour forecasts. Pollanen *et al* (2001) reinforce this observation and discuss how a fermentation forecasting tool could be used to optimise the process schedule.

Two case studies demonstrate how forecasting algorithms can be used to assess future bioprocess conditions. This information can then be utilised by operators to make operational policy changes or in a decision support role. The first study concerns the prediction of performance in the latter stages of the batch to modify feed profiles; the second considers late batch performance assessment to allow improved batch scheduling on a multi-batch vessel site.

2. CASE STUDY EXAMPLES

2.1 Lager fermentation

In the lager brewing process, wort is mixed with yeast to initiate the production of alcohol. The reaction is exothermic and the temperature rises until it reaches a control point where it ideally remains throughout the batch until the completion criteria are satisfied and chilling occurs. The end-point of the fermentation is defined by two key variables. The first is the present gravity (PG) which measures the alcohol content of the brew. More specifically as the fermentation progresses, the sugars in the wort are converted to alcohol, and the density of the brew decreases. Once all of the sugars have been used up, the reaction slows down and the PG reaches a plateau. The second key variable is diacetyl concentration (2,3-pentanedione). Diacetyl is a small flavour active molecule often described as a butterscotch or honey tone, and is a by-product of yeast amino-acid metabolism. High concentrations of diacetyl are undesirable, especially when brewing lagers, as lagers tend not to have a strong flavour. However, once the yeast has finished fermenting the sugars, it will re-absorb the diacetyl from the beer. This process is known as the 'diacetyl rest'.

Although the main processing steps involved in the industrial scale production of beer are well established, the length of the brewing process is subject to natural variation. This is mainly due to changes in fermentation length as most other stages have well defined processing times.

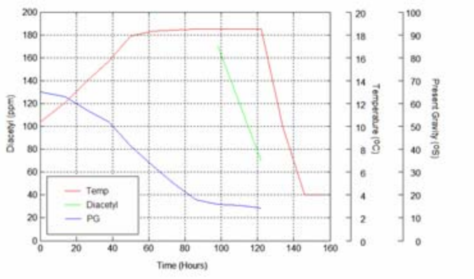

Fig. 1. Process batch trends (lager fermentation)

Early prediction of the fermentation endpoint is thus valuable information allowing tighter scheduling of downstream operations such as bottling and canning, reduced turnaround time for fermentation vessels and ultimately, an increase in annual plant throughput. In this study, the need to predict the end-point arising from the PG specification was more critical from a timing perspective than diacetyl considerations and the prediction results focus on PG forecasting.

Typical batch trends from the lager fermentation are shown in Fig. 1. The rise in temperature to the controlled level as the fermentation progresses along with the fall in PG to satisfy the endpoint specification at around 120 hours can be observed. PG measurements are obtained through off-line analysis at a frequency of approximately eight hours. The diacetyl measurements are less frequent and, in this case, only two were made, with the latter confirming that levels had fallen below the diacetyl constraint. This measurement is obtained from off-line analysis and the satisfaction of the diacetyl endpoint is also confirmed by taste testing.

2.2 Antibiotic fermentation

The second case study addresses improvements in the operational policy for a fed-batch fermentation process producing an antibiotic. A substrate is fed to the batch following the initial biomass accumulation period. The substrate concentration in the batch is controlled by the operators taking samples for off-line analysis and adjusting the substrate feed-rate to maintain the concentration at a level that maximises product formation. Towards the end of the batch, product formation slows down and the batch is terminated and the broth is sent to downstream processing for product recovery.

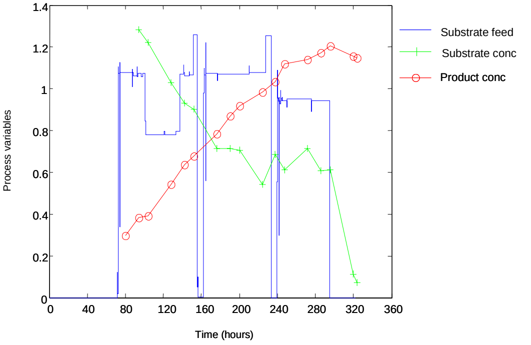

Fig. 2. Typical batch trends (antibiotic fermentation)

To minimise the substrate concentration in downstream processing, the substrate feed is terminated before the end of the batch and concentration falls as it is consumed for product formation. Typical batch trends are shown in Fig. 2.

A target end of batch substrate concentration has been specified. If the concentration falls below the target value, product accumulation rates fall significantly. This effect can be observed in Fig. 2 where the product concentration falls, between 300 and 320 hours, as a result of the substrate feed being terminated too early. Alternatively, if the substrate feed had been cut too late the substrate concentration would have been above the target value at the end of the batch. In this case, in addition to substrate wastage, downstream recovery of product becomes less efficient. The substrate feed termination time is currently set to be a fixed number of hours before the planned end of the batch, implicitly assuming that the batch behaves in a consistent manner. However, batch to batch variability results in changes in substrate concentration and usage rates. As a consequence, it is necessary to adjust the substrate feed termination time to compensate for batch variations. To do so requires a prediction of usage rates up to the end of the batch.

3. FORECASTING PROCEDURES

Both case studies fall within the framework of a forecasting problem. In Case Study 1, the problem involves forecasting the time a brew will take to reach a specified PG using information from the early stages of the batch. Case Study 2 concerns the forecasting of the rate of substrate usage, again using information available from early stages in the batch.

Two fundamental strategies to address these challenges are available: model based or pattern based forecasting. In model based forecasting, a mathematical description of the process is determined from past process batches and used with data from the existing batch to identify likely future behaviour. The model can be either empirical with the parameter values determined using process data or it can have mechanistic structure where the parameters are related to physical/chemical attributes. Trelea *et al* (2001) compared the alternative approaches for the predictive modelling of a brewing fermentation and concluded that if mechanistic understanding is available it should be employed. Karim et al (2003) demonstrated through four case studies, the utility of both principal component analysis and neural networks in fermentation batch performance assessment. The black-box neural network approach was found to be acceptable even when the number of batches was limited.

Lopes and Menezes (2003) investigated the application of tri-linear PLS for the prediction of the end-batch concentration of product in an antibiotic fermentation. They observed that final productivity could be forecast with some accuracy using information from a critical mid-stage of the batch. Undey *et al* (2003) applied a similar unfolding approach utilising multi-way PLS (MPLS) to predict the end-batch concentration in a penicillin fermentation in the presence of disturbances or faults. The forecasting of end of batch performance to reduce batch time and minimise product degradation was considered by Sankpal *et al* (2001). Data from a *Aspergillus niger* fermentation was used to demonstrate that productivity can be forecast fifteen hours ahead. The forecasting model was developed using a symbolic regression method with the resulting relationship being a simple autoregressive time series.

For bioprocesses in general, the lack of intimate knowledge of structure usually precludes the usage of mechanistic approaches. For this reason an empirical structure is utilised in this paper. The methodology adopted is that of multi-way projection to latent structures (PLS). Multi-way PLS is the extension of PLS and allows the modelling of batch processes.

The alternative tactic is to adopt a pattern recognition strategy. For example, if the process trends follow a particular functional form, but with varying parameters, then patterns can be captured from the early batch behaviour, functional parameters determined and forecasts made using the functional patterns identified. Pattern recognition methods have been applied to fermentation processes to extract key features. For example, Stephanopoulos *et al* (1997) considered a number of algorithms for pattern extraction and demonstrated that benefits can be gained through their application to multiple fermentation examples.

Rather than using current batch behaviour and matching it or fitting some functional form, the alternative is to omit this step and match the pattern of the current batch to the most similar previous batch or batches and use past batches as a predictor of current batch performance, i.e. Case Based Reasoning (CBR).

Watson and Marir (1994) presented a review of CBR and discussed potential areas of application. Among the characteristics of problems suited for the application of CBR are that a model of the system is not available but numerous historical examples of behaviour exist. The fundamental principle behind CBR is that the historical records are examined and the most similar situations to the current condition are used to make decisions. Wastewater treatment is a sector that has attracted a significant number of applications of CBR. For example, Ruiz et al (2006) considered how traditional distance metrics could be complemented with multiway principal component analysis performance metrics. For the comparison of batches They observed that not only does MPCA provide useful comparison metrics, it is also appropriate for the detection of batches that demonstrate abnormal behaviour.

Despite numerous publications concerning the modelling of fermentation processes, very few have considered the application of CBR. This is surprising given the complexity of the fermentation process and the difficulty in specifying a model based on the mechanisms. These sentiments were echoed by Roger et al (2002) who studied the application of pattern recognition methods to the fermentation of wine. It is clear that sophisticated algorithms do allow informative patterns to be extracted but if an approach such as CBR is available it does have benefits, particularly in terms of ease of development. In this paper, the utility of the CBR approach is contrasted with MPLS.

3.1 Multi-Way Partial Least Squares

Multi-way partial least squares (Wold *et al*, 1987) relates early batch behaviour to a performance or end-point objective. Essentially, a time series of early batch behaviour is used to predict the final batch condition. Leave-one-out cross validation can be applied to make maximum use of limited batch data.

3.2 Case Based Reasoning Procedure

In CBR, a library of previous process behaviour is established. Current process behaviour is then compared with previous experience and based on past observations and decisions taken in those instances, the 'best' action to take for the current instance is determined. A critical metric in CBR is how to determine similarity between cases. In this paper, the time progression of the batch needs to be taken into account. Two metrics are considered. The first is a process variable distance metric:

$$\text{Distance} = \sqrt{\left(\sum_{i=N1}^{N2} (ynew(i) - yhist(i) - bias)^2 + \lambda(unew(i) - uhist(i))^2\right)/(N2 - N1)} \tag{1}$$

The bias term acts to adjust for the offset between output profiles. λ acts as a weighting between the output and input deviation and *N1* and *N2* specify the window of samples over which the comparison is made. *ynew* and *unew* are the output and input measured variables of the new batch and *yhist* and *uhist* refer to the CBR library batches. The bias is calculated as follows:

$$bias = \left(\sum_{i=N1}^{N2} ynew(i) - yhist(i)\right)/(N2 - N1) \tag{2}$$

Predictions of future behaviour of the current batch are then made:

$$ynew = yhist_{closest} + bias_{closest} \tag{3}$$

where the subscript refers to the closest batch selected from the library. An additional threshold requirement was imposed to increase accuracy. That is the library of examples was constrained to comprise those for which the reconstruction predictor errors determined from a PLS regression forecast was below a threshold. This limitation excluded those batches which were significantly different but which was not apparent from the weighted distance metric.

The alternative distance metric considered was based on multiway PCA scores. In statistical process control, PCA scores plots are frequently used to judge similarity between samples. Singhal and Seborg (2001) extended this concept and the approach taken is a modified metric inspired by their studies:

$$Dis\tan ce = \sqrt{\sum_{i=1}^{Nscores} VarExplained(i) \times (score_{new}(i) - score_{hist}(i))^2} \tag{4}$$

Nscores refers to the number of principal components retained. The metric assesses, in the scores space, batch variability similarity with the distances weighted by the variance explained.

4. RESULTS

4.1 Case Study 1 Results

Data from 100 beer batches was utilised with PG and temperature logged manually. Samples were taken at different frequencies, with the sampling interval varying between 8 and 24 hours. From a process operational perspective, it is important to gain insight into the batch completion from around 48 hours. Typically 4 to 5 samples have been taken by this time and it is this information that is available to forecast batch completion time. To equalise the sampling frequency for subsequent computational purposes, a cubic spline was used to interpolate between samples and the data was reconstructed at a 5 hourly sampling frequency. Investigations considered the predictive ability using information available up to 45 hours.

A data matrix containing the PG and temperature profiles for the first 45 hours was established for all 100 batches and the time at which each batch achieved a PG of 18 determined. Initial analysis of the data was carried out using multi-way PCA and the results indicated a number of outlying batches. Subsequent investigation revealed that temperature control problems arose in some batches. Forecasting behaviour in the presence of unpredictable temperature change is not a practical proposition. Consequently to investigate the performance of the forecasting algorithms, batches with temperature control issues were removed leaving 64 batches.

The RMS errors associated with PLS and CBR forecasts are shown in Table 1. It is clear that the CBR approach using the weighted distance metric produces results of similar accuracy to PLS. The CBR algorithm using the scores based distance metric performs less well. In Table 1 CBR1 refers to distance metric case selection without employing bias removal, i.e. the bias term is set to zero (eqn 1), CBR2 is with bias removal (eqs 1, 2 & 3) and CBR3 are the results for the PCA based metric (eq. 4).

Table 1 - RMS errors for PG18 forecast

	CBR1	CBR2	CBR3	PLS
Full (100 batches)	6.58	6.58	8.94	6.05
Subset (64 batches)	6.33	6.12	8.66	6.34

To assist in interpreting predictive ability, Fig. 3 shows the PLS predictions against the actual time to achieve PG18 on a subset of data. The increase in prediction error with time is expected as forecasts further into the future are being made.

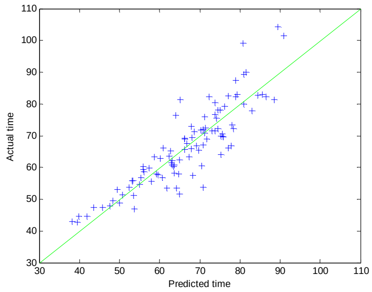

Fig. 3. PLS predictions of time to PG 18

4.2 Case Study 2 Results

In contrast to the first case study, the fed-batch nature of the antibiotic fermentation means that a number of manipulated variables can be modified throughout the batch. Furthermore, samples are routinely taken for analysis to track progress with multiple broth components analysed. This more comprehensive data set provides greater opportunity to predict behaviour but also, the higher data dimensionality means that forecast development is more complex.

Recognising that the prediction of future behaviour must take into account batch trajectories, multi-way PCA approach was applied to data from the first 280 hours of the 21 available fermentations to investigate whether batches can be identified where a drop in productivity occurs at the end of the batch. In such cases, it would have been prudent to keep feeding substrate for a longer period. MPCA was applied to the concentration and federate data to identify batches exhibiting a change from nominal behaviour.

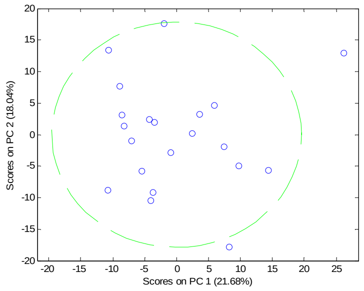

Fig. 4. Bivariate scores plot of PC1 versus PC2

Fig. 4 shows the principal component bivariate scores plot. The multi-way PCA approach provides an indication of whether patterns exist within the data thereby indicating dissimilarity between batches. In this case, PC 1 v PC2 plot identified that some batches deviated from nominal behaviour. In Fig. 4, two batches lie outside the 95% confidence interval and as such exhibit behaviour different to nominal operation. One of the batches is the most productive in the latter stages of all the batches analysed and the other is the lowest productivity batch. These results suggest that there may be patterns present in the data that explain deviations from nominal behaviour. Also of the 21 batches, at least one could lie outside the 95% confidence interval by chance and still be characteristic of nominal operation.

To investigate whether it is possible to forecast final product concentration from batch trajectories to 280 hours, multi-way PLS was applied to the same variable set and using the same sampling frequency as for MPCA. Leave-one-out cross validation was implemented and Fig. 5 shows the validation results. It can be observed from Fig. 5 that the predictions of final product concentration fail to reasonably forecast the final value, falling well off the 45^o parity line.

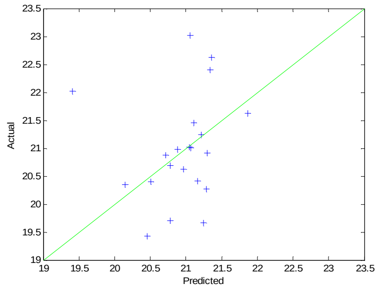

Fig. 5. MPLS forecast of end product concentration

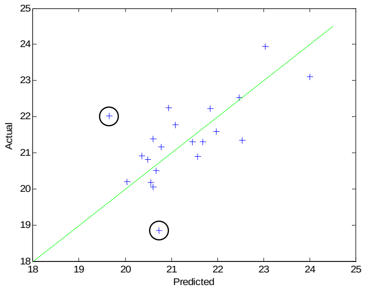

Fig. 6. CBR using product and feed profile

Table 2 – Comparison of CBR1 and MPLS performance using average RMS prediction errors

	CBR (product)	CBR (product & feed)	MPLS
Full data set	0.90	0.93	1.15
Outliers excluded	0.64	0.68	0.98

Table 2 shows the RMS error of the predictions resulting from the application of CBR and MPLS corresponding to Figs 5 and 6. It is evident that the outlying data points dominate the prediction error in the CBR analysis and whilst they are also significant in the MPLS results, the poorer performance of MPLS remains evident once they are removed.

Two important observations arise from these results. When attempting to predict future batch behaviour, it is essential to check whether the behaviour of the current batch is consistent in form to past experience. If behaviour is different to that experienced previously, then predictions are likely to be inaccurate. PCA provides an effective tool to make such judgements. Secondly, the CBR approach provides a more accurate prediction of future batch performance in this example. In the case of MPLS, process behaviour is captured in a model and the model then used to make predictions. The loss of accuracy arises in the step from data to model as the error in model fit due to model structural inaccuracy is avoided in CBR. The process is sufficiently complex for this error to be significant.

In implementing a CBR scheme, the first step is to create a database / library of 'cases' capturing progression of the batch through the course of the fermentation. This involves identifying those variables that are indicative of changes in behaviour and then building a library from which it is possible to compare the behaviour of the current batch. As more experience accumulates, the library of batches can be updated enhancing the capability of the CBR tool. The CBR method used here has been developed with a relatively small number of batches and hence the results can be improved on, when additional batches become available.

5. CONCLUSIONS

This paper has considered the application of forecasting methods for fermentation end point prediction. For the more straightforward forecasting problem of lager PG, PLS and CBR were comparable. However, forecasts of acceptable accuracy could only be achieved when the batch temperature was under control. This is not surprising given that unpredictable changes in temperature have a considerable impact on fermentation performance. From a practical perspective the implication is that the temperature control system needs improvement if consistent prediction accuracy is to be achieved. Indeed, temperature control to achieve desired taste objectives is demonstrated in Kobayashi *et al* (2006) and a forecasting element to a control scheme offers many benefits.

In the antibiotic manufacturing process, temperature and other variables are regulated more precisely but batch to batch variations still occur. The prediction of behaviour at the end of the batch is more complex and as a result the linear PLS model structure compromises accuracy. The non-model based CBR approach does not suffer from such limitations and therefore produces more accurate forecasts. It is nevertheless important to recognise that outlying batches can seriously degrade predictive performance so multivariate batch outlier removal is an important first step.

6. REFERENCES

Ignova, M., G.A. Montague, A.C. Ward and J. Glassey (1999). Fermentation seed quality analysis with self-organising neural networks. *Biotech. Bioeng.* **64(1)**, 82-91.

Karim M.N., D. Hodge and S. Laurent (2003). Data-based modeling and analysis of bioprocesses: Some real experiences, *Biotechnol. Prog*. **19**, 1591-1605.

Kobayashi, M; Nagahisa, K; Shimizu, H; Shioya, S. (2006). Simultaneous control of apparent extract and volatile compounds concentrations in low-malt beer fermentation. *Applied Microbiology and Biotechnology*, **73**, 3, 549-558(10)

Lopes J.A. and J.C. Menezes (2003). Industrial fermentation end-product modelling with multilinear PLS. *Chemometrics and Intelligent Laboratory Systems,* **68,** 75– 81.

Neves A.A., L. M. Vieira and J.C. Menezes (2001). Effects of preculture variability on clavulanic acid fermentation, *Biotech. Bioeng.* **72(6)**, 628-633

Roger J.M. Sablayrolles J.M., Steyer J.P. and Bellon-Maurel V. (2002). Pattern analysis techniques to process fermentation curves. Biotech Bioeng, Vol 79, 7, pp804-815

Ruiz, M. C. Colomer and J. Meléndez. (2006). Multiway principal component analysis and case based-reasoning approach to situation assessment in a wastewater treatment plant, *2es Jornades UPC de Recerca en Automàtica, Visió i Robòtica, Barcelona. Juliol*

Sankpal N. J. Cheema, S. Tambe. and B.D. Kulkarni (2001). An artificial intelligence tool for bioprocess monitoring: application to continuous production of gluconic acid by immobilized Aspergillus niger *Biotech. Letters* **23** 911–916

Scarff M., S. Arnold, L. Harvey and B. McNeil (2006). Near infrared spectroscopy for bioprocess monitoring and control: Current status and future trends. *Critical Reviews in Biotech.*, **26**:17–39.

Singhal A. and D.E. Seborg (2001). Pattern matching in historical batch data using PCA. IEEE Control Systems Magazine, Vol 22, 10, pp 53-63

Trelea I., M. Titica, S. Landaud, E. Latrille, G. Corrieu, A. Cheruy, (2001). Predictive modelling of brewing fermentation: from knowledge-based to black-box models, *Mathematics and Computers in Simulation* **56**, 405-424.

Undey C, S. Ertunc and A. Cinar. (2003). Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. *Ind. Eng. Chem. Res*, **42**, 4645-4658

Watson I. and Marir, F. (1994). Case-based reasoning: A review, *The Knowledge Engineering Review* **9(4)** 327-354.

Wiese J., A. Stahl and J. Hansen (2005). Applying and optimizing case-based reasoning for wastewater treatment systems. *AI Communications, Special Issue: Binding Environmental Sciences and AI.* **18(4)**, IOS Press.

Xue Y F. and J. Q. Yuan (2005). On-line application oriented scheduling for fed-batch antibiotic fermentation, *American Control Conference*, June 8-10, 2005. Portland, OR, USA

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