8th International IFAC Symposium on
Dynamics and Control of Process Systems
Preprints Vol.1, June 6-8, 2007, Cancún, Mexico

MODELLING INDUSTRIAL FERMENTATION DATA WITH MULTIWAY MULTIVARIATE TECHNIQUES

Ana Patrícia Ferreira, João Almeida Lopes*, José Cardoso de Menezes

Centre for Biological and Chemical Engineering, IST, Technical University of Lisbon
Av. Rovisco Pais, P-1049-001 Lisbon, Portugal, Tel. (+351) 21 841 9838;
Fax. (+351) 21 841 9197; bsel@ist.utl.pt

Abstract: Several multivariate statistical techniques have been extensively proposed for monitoring industrial processes. In this paper, multiway extensions of two such techniques: multiway principal component analysis (MPCA) and multiway partial least squares regression (MPLS) were applied to a large data set from an industrial pilot-scale fermentation process to improve process knowledge. The MPCA model is able to diagnose faults occurring in the process whether they affect or not process productivity while the MPLS model enables the prediction of final product concentration and the detection of faults that will influence the fermentation productivity. Copyright © 2007 IFAC

Keywords: fermentation processes, industrial, modelling, batch, multivariate analysis

1. INTRODUCTION

Batch and semi-batch processes play an important role in the chemical industry, mainly because of their flexibility to produce low-volume, high-value products. Examples of batch processes include the production of polymers, pharmaceuticals and biochemicals and the separation and transformation of materials by batch distillation and crystallization. Successful batch operation means being able to maintain process variable trajectories with a high degree of reproducibility from batch to batch (Kosanovich et al, 1996; Nomikos and MacGregor, 1995b). Batch processes generally exhibit some batch-to-batch variation arising from variation in raw materials quality, seeding, variability in charging of the reactor and unnoticed deviation in instrumentation performance. Since these variations lead to low reproducibility, adequate monitoring and control techniques are essential to ensure safe operation and to assure the production of consistently high quality products.

The main characteristics of batch processes are related to both their success and their incompatibility with the conventional mathematical or empirical modelling for monitoring and controlling continuous processes. An alternative approach for monitoring and control of batch processes based on the use of multivariate statistical techniques and in the philosophy of statistical process control (SPC) (Doty, 1996) was proposed by Nomikos and MacGregor (1994). Following this approach, the behaviour of the process is modelled using data from an historical data set of past successful batches (which are assumed to be in a state of control) and, subsequently, future unusual events are detected by referencing the progress of a new batch against the "in-control" model and its statistical properties.

To construct the process model a multivariate statistical projection method, multiway principal component analysis (MPCA), is used to compress the data matrix (X) and to extract the information by projecting the original set of highly correlated variables

* Current affiliation: Requimte, Faculty of Pharmacy, University of Porto, Rua Aníbal Cunha, 164, 4050-047 Porto, Portugal

into a low-dimensional space that summarizes both the variables and their time histories during successful batches. MPCA only makes use of the process variable trajectory measurements (X) but this multivariate SPC (MSPC) approach was latter extended by Nomikos and MacGregor (1995b) to include measurements on product quality variables (Y) taken at the end of each batch by using multiway partial least squares regression (MPLS). Rather than focusing only on the variance of X, MPLS focuses on the variance of X that is more predictive for the product quality, Y. Thus, MPCA is applied to understand and monitor the variability in process variables while MPLS enables the study and monitoring of variations in the process variables that are most influential on the quality and productivity variables. Several applications of the MSPC approach to different fields of chemical industry have been described in the literature. Examples of applications include monitoring polymerization reactions (Nomikos and MacGregor, 1995a; Kosanovich et al, 1996), an industrial ceramic melter (Wise and Gallagher, 1996), industrial fed-batch fermentation processes (Gregersen and Jørgesen, 1999; Albert and Kinley, 2001; Lennox et al, 2001), industrial batch drying processes (García-Muñoz et al, 2003) and pharmaceutical industry processes (Westerhuis and Coenegracht, 1997).

The primary goal of this study was to demonstrate how MPCA and MPLS can be used to model industrial fermentation processes. These statistical techniques were applied to a data set of industrial pilot-scale fermentation batches for the production of an active pharmaceutical ingredient: clavulanic acid. In this work, only on-line measured variables related to biomass and product quality were considered in the development of process models. It is intended to illustrate how MPCA can be used to discriminate between similar and dissimilar batches and to understand some of the major sources of batch-to-batch variations. This work is also aimed at studying the ability to predict fermentation yield from the same variables through the use of MPLS regression. The availability of an accurate process model will enable 1) the development of an inferential sensor for product concentration and 2) to study the contribution of each variable to process productivity. Both actions can enable significant improvements in the process.

2. MATERIALS AND METHODS

2.1 Data

2.1.1 Microorganism and Culture Conditions

The bioprocess studied is an industrial process for the production of clavulanic acid using a high-producing strain of Streptomyces clavuligerus supplied by CIPAN, S.A. (Vala do Carregado, Portugal). Cultivation was carried out using a medium containing complex carbon and nitrogen sources and appropriate precursors. The operating conditions used were typical of those employed routinely in the fermentation industry for aerobic submerged cultivations for the production of secondary metabolites (Neves et al, 2001).

2.1.2 Data acquisition

A set of 16 fermentation batches was monitored by performing on-line measurements on five variables: capacitance and conductance of the culture broth, concentration of carbon dioxide and oxygen in the exhausted gases and dissolved oxygen in the culture broth. Capacitance and conductance readings were performed by a Biomass Monitor 214M (Aber Instruments, Aberystwyth, UK), equipped with an annular probe (Ferreira et al, 2005). The capacitance of the fermentation broth is directly proportional to the viable biomass concentration while the conductance is a measure of the concentration of ions present in the broth (Spierings, 1998). The composition of the outlet gas stream was analysed by an infrared carbon dioxide analyser SIFOR 200 (Maihak, Hamburg, Germany) and a paramagnetic oxygen analyser PMA-25 (M&C, Ratingen, Germany). The derived variables carbon dioxide production rate (CER) and oxygen uptake rate (OUR) were calculated as indicated by Heinzle and Dunn (1991). These derived variables were used for model development in replacement of the raw gas composition measurements. The dissolved oxygen concentration was measured with a standard O_2 electrode inserted directly in the fermentor. The acquisition of data is controlled by an application developed in the graphical programming environment LabView® (National Instruments, Austin, TX, USA).

2.1.3 Reference method for clavulanic acid determination

Clavulanic acid on the fermentation broth was assayed by a colorimetric method based in the absorption at 312 nm of the product of the reaction between clavulanic acid and imidazole (Bird et al, 1982).

2.2 Data processing

All variables were pre-processed by detection and removal of outliers followed by smoothing with Savitzky-Golay filter for noise reduction. Times series were synchronized by interpolation. The batches had slightly different durations and so, only those observations taken from 3 h of growth until termination time of the shortest batch, 91 h, were used for model development. The three-dimensional array $\underline{\mathbf{X}}$ containing the process data (16 × 5 × 181) was unfolded into a two-dimensional array $\mathbf{X}$ (16 × 905) by preserving the batch direction (Westerhuis et al, 1999). MPCA and MPLS models were developed based on algorithms available from the PLS toolbox v3.0 for Matlab. The number of principal components to include in the MPCA model was selected based on the average criterion (Jobson, 1992), according to which those PC which capture an amount of variance greater than the average should be retained. The eigenvalues (λj) are used as a measure of the amount of variance captured by each

PC, so the average criterion consists in retaining the j PCs for which $\lambda_j > \overline{\lambda}$. The number of latent variables to include in a MBPLS model was determined by leave-one-out cross-validation.

3. RESULTS

3.1 Multiway principal component analysis

To determine the number of components to include in the MPCA model, the eigenvalues of the first 15 PC were computed. The eigenvalues of the first five PC are greater than the average for all PC. According to the average criterion, 5 PC should be considered for model development.

A MPCA model with 5 PC captures 74.6% of the variance contained in the data. The scores plot of the two first PCs, which account for 45% of the variance in the data, is displayed in Fig. 1. All the batches fall inside the 99% confidence ellipse and all but one batch (batch 3) fall inside the 95% confidence ellipse. From this figure is clear that certain batches exhibit similar variable trajectories while others form separate clusters (e.g., batches 5, 6 and 7), thus the MPCA algorithm proves to have the power to discriminate among batches based on the trajectories of variables measured on-line. Batches 3 and 5 to 7 showed to be significantly dissimilar from the remaining. Comparing the variables trajectories along all the batches (data not shown), it was possible to detect the events causing the dissimilarities.

Batch 3 presents significantly higher conductance profile when compared with the remaining batches, although it was not possible to devise a reason for this fact. In addition, for a long time period (around 40 h) the temperature in the bioreactor was kept at a lower level. In respect to batches 5, 6 and 7, different conditions were used for substrate addition and this produced the alterations in the quality variables, particularly for biomass, dissolved oxygen profile (since it caused alterations in the microorganism growth) and conductance.

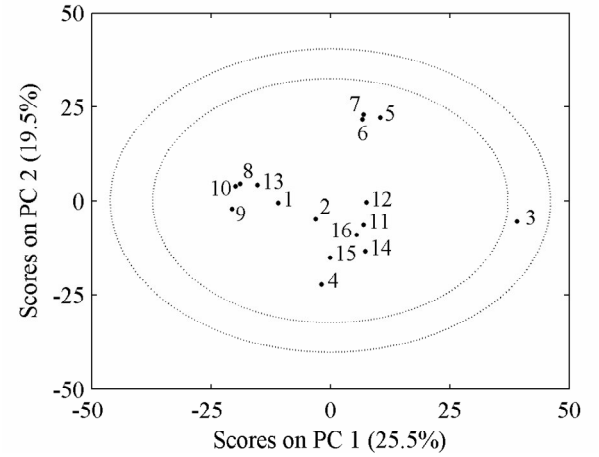
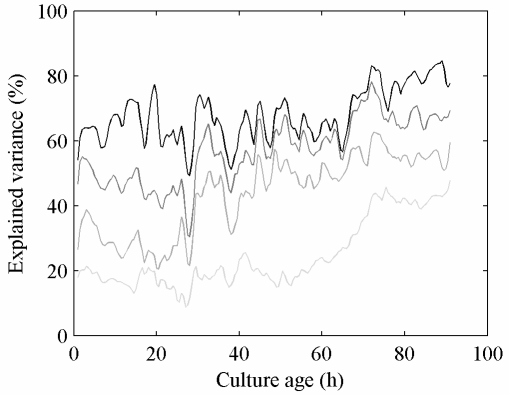
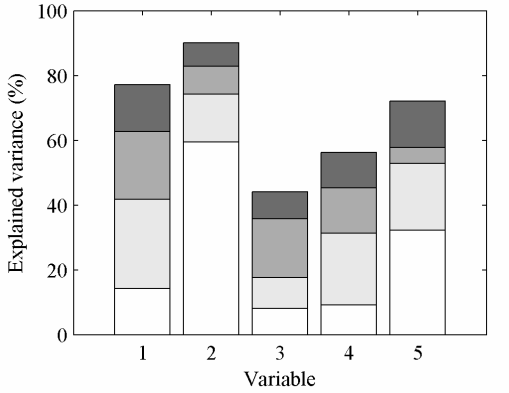


Fig. 1 – Scores plot of the two first PC of the MPCA model (the dotted lines represent the confidence ellipsoids at a significance level of 95 and 99%).



A



B

Fig. 2 - Percent of explained variance by each PC plotted on a cumulative basis: A) over time; B) by variable (CAP: capacitance, COND: conductance, CER: CO_2 evolution rate, OUR – O_2 uptake rate, DO – percent dissolved O_2). From lightest to darkest: PC1 to PC5.

Explained variance plots were used to study variability contained in the data as a function of time and original variable. The amount of variance explained by the model is calculated by comparing the true process data with the estimates computed from the MPCA model (Kosanovich et al, 1996). It can be computed from Eq. 1, where $\hat{\sigma}^2$ and σ^2 are the estimated and true sum of squares, respectively.

$$\text{Explained variance (\%)} = \frac{\hat{\sigma}^2}{\sigma^2} \times 100 \qquad \text{Eq. 1}$$

A large percentage of explained variance indicates that the variability in the data and the correlations among variables are captured by the model. Variance plots over time can be used as an indicator of the phenomenological / operational changes that occur during process evolution, identified by changes in the variance captured by each PC which signal alterations on the correlation structure among the variables. Fig. 2A displays the variance explained over time by the five PCs. From this figure, the major phenomenological alterations on the process take place around 35 and 50 h of growth. This is consistent with previous process knowledge according to which the period of transition from exponential growth to stationary phase occurs

between 30 and 40 h and ends after 50 to 55 h of growth. It is more difficult to attribute a defined meaning to the remaining alterations identified in the covariance structure of the data along time (from Fig. 2A). They are most probably due to intrinsic variations in the culture which are not easily inferred from the variables used to develop the model (e.g., morphology changes). It is possible that the changes identified at 20 and 72 h of growth are due to morphological alterations in the microorganism (cf. Ferreira et al (2006)) but there is no clear evidence to support this hypothesis.

The variance explained for each variable included in the model over the five PCs is displayed in Fig. 2B. The MPCA model explains over 80% of the variance in capacitance, conductance and dissolved oxygen concentration and close to 60% of the variance in the variables derived from exhaust gas analysis (CER and OUR). PC1 accounts for most of the variance in the conductance (this is why the amount of variance in batch 3 is captured mainly by this PC) and explains also a significant amount of the variance contained in the capacitance and dissolved oxygen. PC2 explains the largest amount of variation in the biomass concentration and also a relevant amount of the variance contained in OUR and dissolved oxygen concentration. The variables for which PC3 accounts for a greater amount of explained variance are the exhaust gas derived variables (CER and OUR) and the dissolved oxygen concentration. PC4 and PC5 contribute, in general, with lower amounts of explained variance.

3.2 Multiway partial least squares regression

Following the analysis of the MPCA model, a MPLS model was developed to investigate the performance of the method in the prediction of the final clavulanic acid concentration for each batch and also to assess which quality variables are the most influential on this productivity variable.

Table 1 - Variance captured by the MPLS model in the X- and Y- blocks

	% Variance captured			
	X block		Y block	
LV	This LV	Total	This LV	Total
1	19.4	19.4	91.9	91.9
2	15.8	35.1	3.7	95.6
3	17.6	52.8	1.6	97.2
4	6.3	59.1	2.1	99.3

The optimal number of latent variables (LV) to include in the model was determined through leave-one-block-out cross-validation, after dividing batches in the data set in 4 contiguous blocks. The minimum RMSECV is attained when the model is built considering 4 LV. Table 1 presents the amount of variance captured by each LV and the total amount of variance captured for both the X and Y blocks of data. Most of the variance in Y-block is captured by the first LV, which models over 90% of the variance in the concentration data.

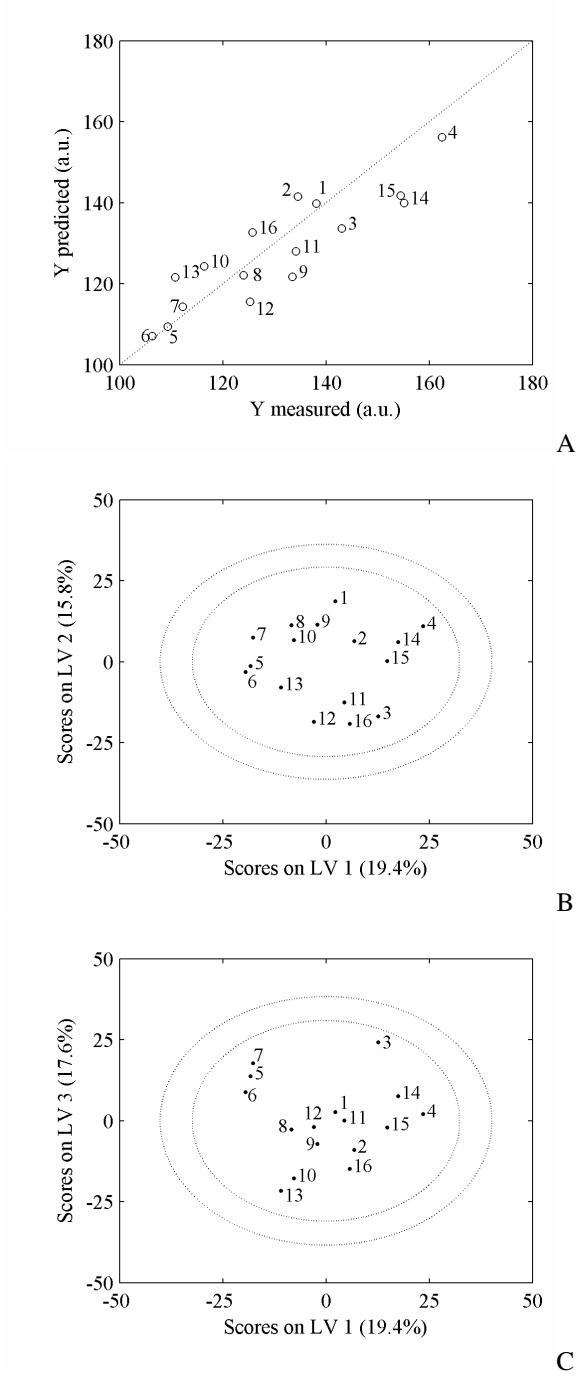


Fig. 3 - MPLS model results. A) Correlation between the measured and predicted (cross-validation) final product concentration; B and C) Scores plots of the model (the dotted lines represent the confidence ellipsoids at a significance level of 95 and 99%): B) LV2 vs. LV1, C) LV3 vs. LV1

This 4 LV model has RMSECV of 8.2 a.u. (normalized concentration values are presented for confidentiality reasons) and predicts 76.2% of the variance in Y, on cross-validation. The correlation between measured concentration values and cross-validation predictions is depicted in Fig. 3A. The relative mean prediction error is 5.1%, on cross-validation. The scores plot of the two first LV of the MPLS model is displayed in Fig. 3B. Clearly, the relationship among batch scores is very different than the one observed for the MPCA model (Fig. 1).

Batch 3 does not appear to be an outlier in the MPLS scores plot, the samples are more uniformly

distributed in the latent variable space and even though batches 5, 6 and 7 still lie close, they do not appear to be clearly separated from the remaining batches. The explanation for these differences lies on the theoretical bases of each method. MPCA focuses solely on the covariance among variables while MPLS focuses on the covariance in the X-block that is most correlated with the Y-block. It is possible to conclude that the causes for the differences between batches observed in the MPCA scores plot have little influence in product yield prediction, since they are only revealed in the third LV of the MPLS model (see Figure 3C), which accounts for less than 2% of the variance in Y (cf. Table 1).

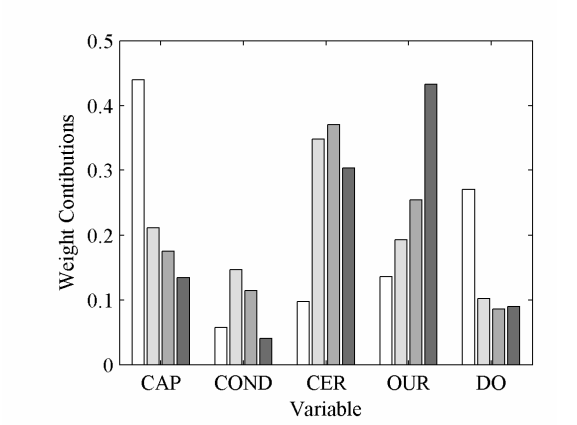
From the analysis of variable contributions to the MPLS model it is possible to assess which variables are the most influent in product concentration prediction. For the sake of clarity, variable contributions must be analysed for each latent variable or over process time (Louwerse et al, 1999). The weight contributions of each original variable, j, to each latent variable, r, are computed from Eq. 2, where K is the number of time points and J the number of variables. Matrix C_{var} has dimensions J × R and describes the weight contributions of each variable for each latent variable, summed over all time points. The contribution of each time point, k, for each latent variable, r, is computed in a similar way following Eq. 3, where the matrix C_{time} has dimensions K × R and describes the weight contributions of each time point for each latent variable, summed over all variables.

$$C_{\mathrm{var},jr} = \sum_{k=1}^{K} w_{(k-1)J+j,r}^2 \qquad \text{Eq. 2}$$

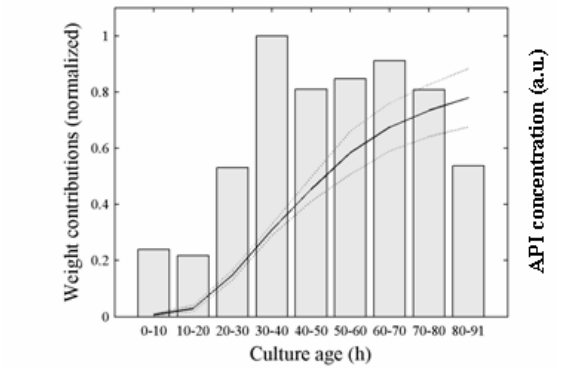
$$C_{time,kr} = \sum_{j=1}^{J} w_{(k-1)J+j,r}^2 \qquad \text{Eq. 3}$$

The weight contribution plot of the variables for each LV is displayed in Figure VIII.6A. Biomass concentration (measured as capacitance) is the dominant variable for predicting final product concentration, since it is the variable with highest contribution to LV1, which captures most of the variance in the Y block. Conductance, on the other hand and on the opposite of what was observed for the MPCA model, is the variable with the lowest influence in the model predictions. Both CER and OUR, the exhaust gas analysis derived variables, exhibit a moderate influence in the model while dissolved oxygen concentration shows to have a significant contribution to LV1 and thus to the prediction of the final product concentration.

For the sake of clarity, the contribution over time is presented as the contribution of several defined time intervals for the prediction of process yield with the purpose of identifying the fermentation stages with greater influence in the PLS model. The weight contribution of defined time windows is presented in Figure VIII.6B, for the first LV. The mean product concentration profile is also depicted on this figure.



A



B

Fig. 4–Contributions to the MPLS model: A) Weight contributions of each variable to each latent variable (CAP: capacitance, COND: conductance, CER: CO_2 evolution rate, OUR – O_2 uptake rate, DO – percent dissolved O_2), from lightest to darkest: LV1 to LV4; B) Weight contributions of each time point to LV1. Lines in B represent the product concentration profile: solid line – mean profile; dotted lines – mean profile ± one standard deviation.

The first 20 h of growth have a small influence on the fermentation productivity. The most influential period of the fermentation ranges from 30 to 80 h of growth, while the influence of the final part of the process is more moderate. The low influence of the early fermentation period is easily understood considering that during this period the dominant event is biomass growth and there is little correlation with production of clavulanic acid. For this particular process and strain, the onset of production occurs after around 20 h of growth (see Figure VIII.6B), which explains the increase in weight contributions after that time since the information contained in the process quality variables becomes more correlated with the final product concentration.

4. CONCLUSIONS

Multiway principal component analysis and multiway partial least squares regression were applied to a data set of industrial pilot-scale fermentation batches with the purpose of improving the knowledge on the process and to assess the discriminant capacity of the methods. A MPCA

model was developed considering five principal components which captured nearly 75% of the variance contained in the data. Analysis with MPCA enables the detection of dissimilarities among batches as well as the identification of abnormal variation in the quality variables. Additionally, the analysis of the amount of variance explained as a function of time allows the detection of state transitions along the course of fermentation thus increasing process knowledge.

The MPLS model was computed with four latent variables and explains 59 and 99% of the variance in the X- and Y-blocks, respectively. The model is reasonably accurate, given the large batch-to-batch variations commonly encountered in biological processes. Analysis of the contributions to the MPLS method gives a clear indication of which variables and time windows are more relevant in process productivity, i.e. what and when to monitor the process. For this bioprocess, biomass concentration and dissolved oxygen concentration were determined to be the most influent variables for productivity prediction and the process stage with higher contribution for the MPLS model is the period between 30 and 80 h of growth, when product formation rate is higher. On the contrary to what happened when using MPCA, the occurrence of faults without direct influence in the process productivity are not detected by MPLS method.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to CIPAN S.A. for supporting this work and for permission to publish the above results. APF and JAL gratefully acknowledge financial support from the Portuguese Foundation for Science and Technology (research grants POCI BD/8807/2002 and POCTI BPD/7194/ 2001, respectively).

REFERENCES

Albert S, Kinley RD (2001), Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision, Trends in Biotechnology 19: 53-62.

Bird AE, Bellis JM, Gasson BC (1982), Spectrophotometric assay of clavulanic acid by reaction with imidazole, Analyst 107:1241-1245.

Doty LA (1996), Statistical Process Control, 2nd Ed, New York (USA): Industrial Press Inc.

Ferreira AP, Vieira LM, Cardoso JP, Menezes JC (2005), Evaluation of a new annular capacitance probe for biomass monitoring in industrial pilot-scale fermentations, Journal of Biotechnology 116:403-409.

Ferreira AP and Menezes JC (2006), Monitoring a complex media fermentation with sample-sample two-dimensional FT-NIR correlation spectroscopy, Biotechnology Progress 22: 866 – 872.

Garcia-Muñoz S, Kourti T, MacGregor JF, Mateos AG, Murphy G (2003), Troubleshooting of an industrial batch process using multivariate methods, Industrial & Engineering Chemistry Research 42: 3592 – 3601.

Gregersen L, Jørgensen SB (1999), Supervision of fed-batch fermentations, Chemical Engineering Journal 75: 69 – 76.

Kosanovich KA, Dahl KS, Piovoso MJ (1996), Improved process understanding using multiway principal component analysis, Industrial & Engineering Chemistry Research 35: 138 – 146.

Jobson JD (1992), Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods, New York (USA): Springer-Verlag.

Lennox B, Montague GA, Hiden HG, Kornfeld G, Goulding PR (2001), Process monitoring of an industrial fed-batch fermentation, Biotechnology and Bioengineering 74: 125 – 135.

Louwerse DJ, Tates AA, Smilde AK, Koot GLM, Berndt H (1999), PLS discriminant analysis with contribution plots to determine differences between parallel batch reactors in the process industry, Chemometrics and Intelligent Laboratory Systems 46: 197-206.

Martens H, Næs T (1991), Multivariate calibration, Chichester: John Wiley & Sons, pp 254 – 258.

Neves AA, Vieira LM, Menezes JC (2001), Effects of pre-culture variability on clavulanic acid fermentation, Biotechnology and Bioengineering 72:628-633.

Nomikos P, MacGregor JF (1994), Monitoring batch processes using multiway principal component analysis, AIChE Journal 40: 1361 – 1375.

Nomikos P, MacGregor JF (1995a), Multivariate SPC charts for monitoring batch processes, Technometrics 37: 41 – 59

Nomikos P, MacGregor JF (1995b), Multi-way partial least squares in monitoring batch processes, Chemometrics and Intelligent Laboratory Systems 30: 97 – 108.

Rothwell SG, Martin EB, Morris AJ (1998), Comparison of methods for dealing with uneven length batches, Preprints of the 7th International Conference on Computer Applications in Biotechnology (Japan), pp 411-416.

Spierings AJC (1998), On-line Measurement of Viable Biomass, in: Van Impe JF, Vanrolleghem PA and Iserentant DM (Eds.), Advanced Instrumentation, Data Interpretation, and Control of Biotechnological Processes, Dordrecht (Netherlands): Kluwer Academic Publishers, pp 41.

Westerhuis JA, Coenegracht PMJ (1997) Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, Journal of Chemometrics 11: 379 – 392.

Westerhuis JA, Kourti T, MacGregor JF (1999), Comparing alternative approaches for multivariate statistical analysis of batch process data, Journal of Chemometrics 13: 397 - 413

Wise BM, Gallagher NB (1996), The process chemometrics approach to process monitoring and fault detection, Journal of Process Control 6: 329 – 348.