ADCHEM 2006

International Symposium on Advanced Control of Chemical Processes
Gramado, Brazil – April 2-5, 2006

DYNAMIC PCA FOR PHASE IDENTIFICATION OF RIFAMYCIN B FERMENTATION IN MULTI–SUBSTRATE COMPLEX MEDIA

Xuan–Tien Doan * and **R. Srinivasan** *,**

* *Institute of Chemical and Engineering Sciences, 1 Pesek Road, Jurong Island, Singapore 627833, e-mail: doan_xuan_tien@ices.a-star.edu.sg*

** *Department of Chemical and Biomolecular Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, e-mail: chergs@nus.edu.sg*

Prashant M Bapat *** and **Pramod P Wangikar** ***

*** *Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai Mumbai 400076 INDIA, e-mail: {prashan, pramodw}@iitb.ac.in*

Abstract: Information regarding when and how a fermentation process changes from one phase to the next is very useful to its modelling and hence control and optimization. In this study, we demonstrated that such information could be obtained by applying DPCA to online measurements of the fermentation process. The process under study is fermentation of Rifamycin B in a multi–substrate complex medium. We compare our observation to the results obtained from the simulation developed for the same system (Bapat *et al.*, in press). The analysis showed that for the first 100 hours or so, the progress of the fermentation experiment in the DPCA score space matched very well to the developed simulation, which had been validated with actual off–line data (Bapat *et al.*, in press). After that (ie. 100 hours onward), there is a significant difference between DPCA analysis result and the simulation result. The reason seemed to be that the simulation did not capture the effects of the secondary metabolism which becomes dominant at later stage of the fermentation.

Keywords: multivariate statistics, dynamic PCA, fermentation, cybernetic model, substitutable substrates, Rifamycin B

1. INTRODUCTION

There are a number of reasons which necessitate phase identification of fermentation. The first reason lies in the improved understanding of the process. The knowledge of when and how the process change from one phase to the next could give insights into which metabolic pathways the fermentation is undertaking. This is especially relevant to fermentation with multi–substrate complex media where there are many metabolic pathways (corresponding to multiple substrates) for the microorganism to proceed. In addition, phase recognition of fermentation process might also be useful in its optimization and control. A model with high accuracy and high robustness for a fermentation process is always desired but more often than not unavailable. The difficulty in modelling such a process is blamed on the complex dynamics of microorganisms, the variable/ill–defined fermentation media (Lopes and Menezes, 2004), and the multi–phase characteristic of the fermentation itself (Hanai and Honda, 2004). Accurate state identification could help to enable phase–wise process modelling for improved performance.

Multivariate statistical techniques and particularly Principal Component Analysis (PCA) have been used in many areas such as monitoring and supervision of continuous processes (MacGregor *et al.*, 1991) as well as batch processes (Nomikos and MacGregor, 1995); improving process understanding (Kosanovich *et al.*, 1996). In addition, PCA applications have been reported in (Gregersen and Jorensen, 1999; Albert and Kinley, 2001; Lopes and Menezes, 2004) for monitoring and supervision of fermentation process. In this paper, we will use dynamic PCA (DPCA) approach to analyze online data from the fermentation of Rifamycin B in a multi–substrate complex medium. DPCA, a variant of PCA technique, was proposed by (Ku *et al.*, 1995) to account for process dynamic behaviors more effectively. Results from DPCA analysis are compared to the corresponding ones from a simulation developed for the same system and described in (Bapat *et al.*, in press).

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique, optimal in terms of capturing the variability of the data. It determines a set of orthogonal vectors, called loading vectors, ordered by the amount of variance explained in the loading vector directions. The new variables, often referred to as *principal components* are uncorrelated (with each other) and are weighted, linear combinations of the original ones. The total variance of the variables remains unchanged from before to after the transformation. Rather, it is redistributed so that the most variance is explained in the first principal component (PC), the next largest amount goes to the second PC and so on. In such a redistribution of total variance, the least number of PCs is required to account for the most variability of the data sets. The development of PCA model, which can be found in numerous published literature including (Ralston *et al.*, 2001; Russell *et al.*, 2000) is summarized as follows. For a given data matrix $\mathbf{X}^o$ (raw data), which has n samples and m process variables as in (1), each row $\mathbf{x_i^T}$ is a sample of m variables associated with a given time.

$$\mathbf{X}^o = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \tag{1}$$

where: x_{ij} is the data value for the j^{th} variable at the i^{th} sample.

Initially, some scaling is required. The most common approach is to scale the data using its mean and standard deviation

$$\mathbf{X} = (\mathbf{X}^o - \mathbf{1}_n \mu^T)\mathbf{D}^{-1} \tag{2}$$

where: $\mathbf{X}^o$ is a $n \times m$ data set of m process variables and n samples.

μ is the $m \times 1$ mean vector of the dataset.

$\mathbf{1}_n = [1, 1, \dots, 1]^T \in \mathbf{R}^n$.

$\mathbf{D} = diag(sd_1, sd_2, \dots, sd_m)$ whose i^{th} element is standard deviation of the i^{th} variable.

After appropriate scaling, the loading vectors can be determined by singular value decomposition (SVD) of the data matrix

$$\frac{1}{\sqrt{n-1}}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3}$$

where: $\mathbf{U} \in R^{n \times n}$ and $\mathbf{V} \in R^{m \times m}$ are unitary matrices.

$\mathbf{\Sigma} \in R^{n \times m}$ is diagonal matrix.

Solving Equation 3 is equivalent to solving an eigenvalue decomposition of the sample covariance matrix $\mathbf{S}$

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T \tag{4}$$

The matrix $\mathbf{\Sigma}$ contains the nonnegative real singular values of decreasing magnitude along its main diagonal ($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{min(m,n)}$), and zero off–diagonal elements. Column vectors

in the matrix $\mathbf{V}$ are the loading vectors. Upon retaining the first a singular values, the loading matrix $\mathbf{P} \in R^{m \times a}$ is obtained by selecting the corresponding loading vectors.

The projections of the observations in $\mathbf{X}$ into the lower dimensional space are contained in the score matrix

$$\mathbf{T} = \mathbf{XP} \tag{5}$$

and the projection $\hat{\mathbf{X}}$ of $\mathbf{T}$ back into the m–dimensional observation space

$$\hat{\mathbf{X}} = \mathbf{TP}^T \tag{6}$$

The residual matrix $\mathbf{E}$ is the difference between $\mathbf{X}$ and $\hat{\mathbf{X}}$

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \tag{7}$$

The residual matrix $\mathbf{E}$ contains that part of the data not explained by the PCA model with a principal components and usually associated with "*noise*", the uncontrolled process and/or instrument variation arising from random influences. The removal of this data from $\mathbf{X}$ can produce a more accurate representation of the process, $\hat{\mathbf{X}}$ (Russell *et al.*, 2000).

Dynamic Principal Component Analysis (DPCA)

Ordinary PCA presented above is essentially a linear technique, and hence its best applications are limited to steady state data with linear relationships between variables (Misra *et al.*, 2002). To analyze a dynamic system, *Dynamic Principal Component Analysis* (DPCA) is required. The concept of DPCA was based on applying PCA to time lagged input data (Ku *et al.*, 1995).

Mathematically, DPCA starts with forming a time–lagged version of the input data $\mathbf{X}$

$$\mathbf{X}_d^o = \begin{pmatrix} \mathbf{x}(d+1)^T & \mathbf{x}(d)^T & \dots & \mathbf{x}(1)^T \\ \mathbf{x}(d+2)^T & \mathbf{x}(d+1)^T & \dots & \mathbf{x}(2)^T \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \mathbf{x}(n)^T & \mathbf{x}(n-1)^T & \dots & \mathbf{x}(n-d)^T \end{pmatrix} \tag{8}$$

where: $\mathbf{x}(k) = [x_{k,1} x_{k,2} \dots x_{k,m}]^T$ is the m–dimensional observation vector at time k. n is the number of data samples. d is the time lag.

The corresponding covariance matrix $\mathbf{S}$ for the time–lagged data is

$$\mathbf{S} = \frac{(\mathbf{X}_d)^T (\mathbf{X}_d)}{n - d - 1} \tag{9}$$

Solving the eigen–decomposition of the covariance matrix $\mathbf{S}$ (Equation 4) and retaining a principal components gives the DPCA model for $\mathbf{X}$.

3. RIFAMYCIN B FERMENTATION MODEL

The fermentation model that we used in this study was developed by P. Wangikar and his colleagues at Indian Institute of Technology (Chemical Engineering Department) and reported in (Bapat *et al.*, in press). It is a dynamic model for the fermentation of Rifamycin B, an antibiotic which is produced on industrial scale, in a multi–substrate complex medium. The model considers the organism to be an optimal strategist (maximizing growth and product formation) with a built–in mechanism that regulates the sequential and simultaneous uptake of multiple substrate combinations. The uptake of individual substrate is assumed to be dependent on the level of a key enzyme or a set of enzymes. In addition, the fraction of flux through a given metabolic branch is estimated by solving the constraint multivariate optimization problem.

4. EXPERIMENT DATA

A detailed description of the Rifamycin B fermentation experiment can be found in (Bapat *et al.*, in press). In the experiment, a combination of different substrates were employed. In this study, we analyze *GLU_AMS_SFCSL_FEDBATCH* experiment which had *GLU*cose, *AM*monium *S*ulphate, *S*oya *F*lour and *C*orn *S*teep *L*iquor. Initial conditions for the experiment are outlined in Table 1.

Table 1. Initial conditions of Rifamycin B fermentation experiment

Variables (g/L)	GLU_AMS_SFCSL_FEDBATCH
Biomass	0.65
Amino acid	4
Glucose	70.43
$(NH_4)_2SO_4$	3.4
Insoluble	20

From the online data collected from the experiments, we selected the measurements for a number of variables which correspond to the experimental conditions (cf. Table 2), to form a data matrix input to DPCA analysis.

5. DPCA ANALYSIS

5.1 Methodology

Procedure to carry out DPCA analysis is summarized below

Table 2. Variables in DPCA analysis

No.	Variables
1	Age (hour)
2	exhaust CO_2 concentration (%)
3	exhaust O_2 concentration (%)
4	pH
5	dissolved O_2 concentration (%)
6	stirring rate (rpm)

(1) The data set is initially augmented (ie. transform into the lagged form $\mathbf{X}_d$) as shown in Equation 8. Several time lags were studied and based on the findings in (Bapat *et al.*, in press), the time lag is set at $t = 8$ hour.
(2) Auto–scaling is applied to $\mathbf{X}_d$ (Equation 2).
(3) The covariance matrix $\mathbf{S}$ of the augmented data is evaluated (Equation 9)
(4) Eigen–decomposition of $\mathbf{S}$ is performed and $a = 2$ principal component vectors are retained.

5.2 Results and Discussion

Fig. 1 shows the fermentation progress in DPCA score space. When there is a change in the progress's direction, the point is marked as a red dot and corresponding time is shown. The simulation developed by P. Wangikar and his colleagues was run for the same initial condition as that for the experiment. Its results are presented in Fig. 3. For better visualization, the result for amino acid predicted by the simulation is shown in a separate plot (ie. Fig. 2).

Observing Figs. 1, 2 and 3, we can conclude that the results from DPCA analysis agree very well with the simulation results for the first 100 hours. As the simulation results shown in Fig. 3 indicates, among the three substrates, amino acid has the largest consumption rate at the beginning of the fermentation. When its consumption rate slows down, corresponding rates of other substrates start to increase. This is reflected in Fig. 1 as a turning point at $t = 20$ hr. The next significant change in the fermentation progress occurs at $t = 27$ hr when the amino acid actually starts being reproduced. Around $t = 60$ hr, Fig. 3 shows that the fermentation media runs out of ammonium sulfate and this results in the turning point at $t = 60$ hr in the score plot Fig. 1. During 60 to 92 hr, both amino acid and glucose are consumed but from 92 hr, the prior substrate is reproduced while the latter continues being consumed. Again, DPCA detects the change and reflects in a turning in the fermentation progress (cf. Fig. 1). At $t = 135$ hour it seems that DPCA results could be implying the depletion of amino acid in the media, which is also predicted by the developed simulation.

However, from $t = 97$ hour, DPCA results start to deviate from what is predicted by the simulation. For example, DPCA score plot clearly indicates that phase changes occur at $t = 105$ hr, $t = 125$ hr, $t = 146$ hr but no such changes could be observed from the simulation results shown in Fig 3.

The reason for this discrepancy needs further investigation and especially verification directly with actual experimental data (which will be available in the near future), instead of simulation results. Nevertheless, it should be noted that as the nitrogen source starts to deplete (i.e., both ammonia and amino acids) around $t = 90$ hr, the fermentation goes into a mode of endogenous metabolism, where some cell lysis occurs and cells grow on the nitrogen available from protein released by lysis. Glucose uptake continues for growth and for maintenance. The secondary metabolic product formation, which is more significant in this phase, was not accounted for by the developed model. This explains the observation that the simulation model fit experimental data very well until the depletion of nitrogen source from the medium (Bapat *et al.*, in press). Toward the end of fermentation run, as the secondary metabolism becomes dominant, the simulation results appear significantly deviate from the actual data (Bapat *et al.*, in press). Consequently, comparison between DPCA results and the simulation results for close–to–end fermentation experiment might not give any valid conclusion.

6. CONCLUSION

We applied DPCA to online measurements of Rifamycin B fermentation data to study the fermentation progress. We compared our observation to the results obtained from the simulation developed for the same system (Bapat *et al.*, in press). The analysis showed that for the first 100 hours or so, the progress of the fermentation experiment in the DPCA score space matched very well to the developed simulation, which had been validated with actual off–line data (Bapat *et al.*, in press). After that (ie. 100 hours onward), there is a significant difference between DPCA analysis result and the simulation result. The reason seemed to be that the simulation did not capture the effects of the secondary metabolism which becomes dominant at later stage of the fermentation.

The study demonstrated the capability of DPCA in identifying phase changes, which could be useful in fermentation process optimization and control. For further work, we are going to validate the DPCA results with the actual off–line data, which as believed will further support the capability of DPCA. In addition, data from fermentation of Rifamycin B in other complex media are

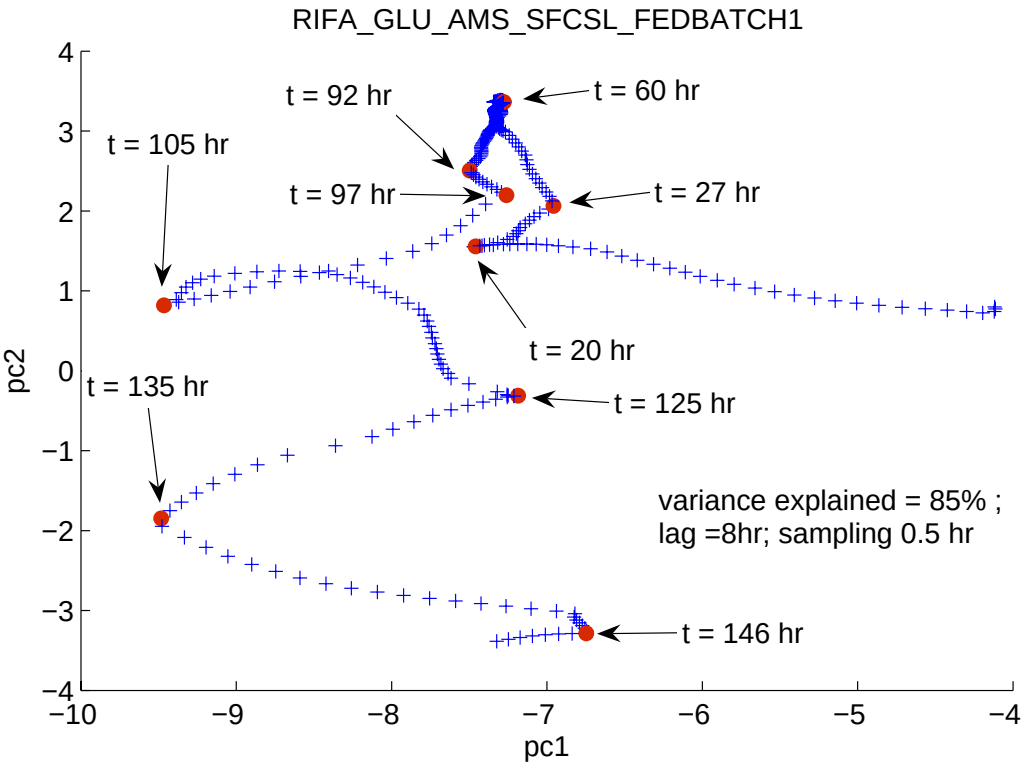

Fig. 1. Score plot from DPCA analysis: all red dots correspond to the time where phase changes in fermentation are likely to occur

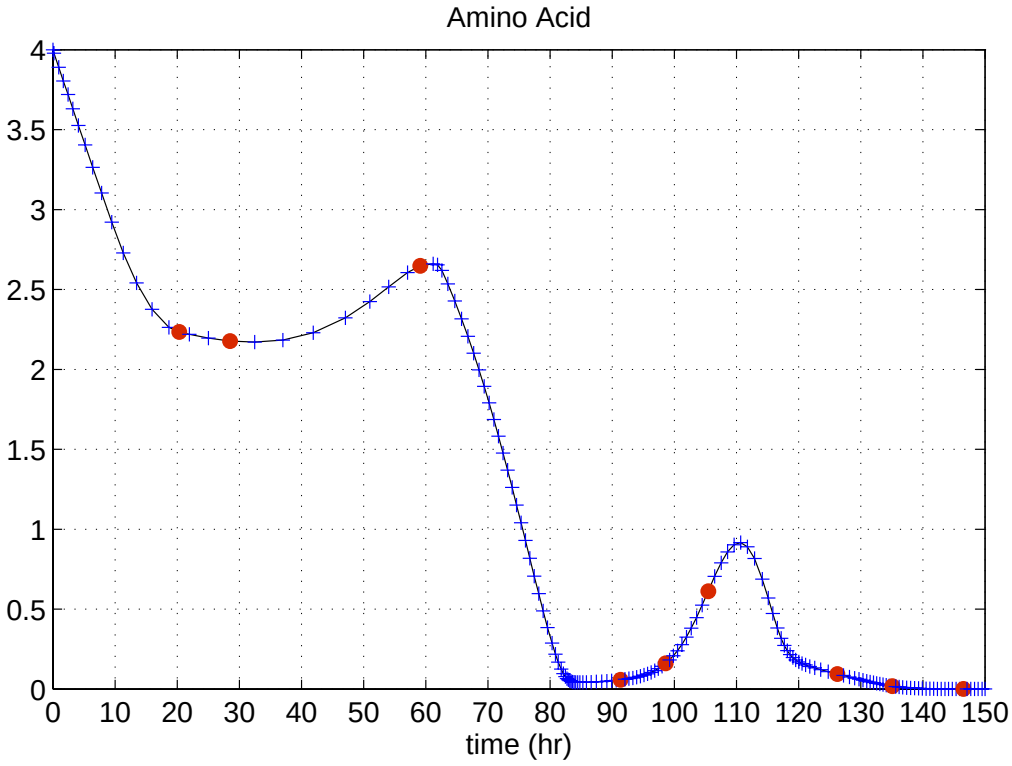

Fig. 2. Simulation results for amino acids in the same experiment with corresponding red dots as in Figure 1

also available and will be analyzed in the same way. These works would establish the ground for further studies such as building inferential PLS model and integrate it with the developed simulation for better optimization and control.

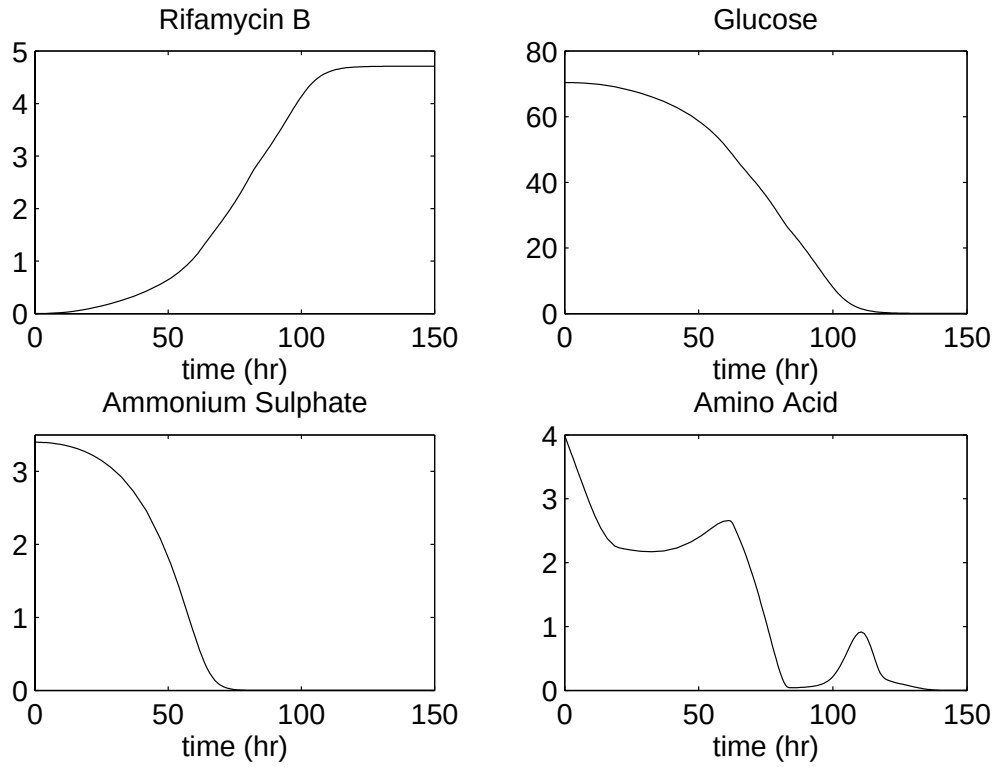

Fig. 3. Simulation results for the same system

REFERENCES

Albert, S. and R. D. Kinley (2001). Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision. TRENDS *in Biotechnology* **19**, 53–62.

Bapat, Prashant M., Sharad Bhartiya, K. V. Venkatesh and Pramod P. Wangikar (in press). A structured kinetic model to represent the utilization of multiple substrates in complex media during rifamycin b fermentation. *Biotechnology & Bioengineering*.

Gregersen, Lars and Sten Bay Jorensen (1999). Super-
vision of fed–batch fermentations. *Chemical Engineering Journal* **75**, 69–76.

Hanai, Taizo and Hiroyuki Honda (2004). Application of knowledge information processing methods to biochemical engineering biomedical and bioinformatics fields. *Adv Biochem Engin/Biotechnol* **91**, 51–73.

Kosanovich, Karlene A., Kenneth S. Dahl and Michael J. Piovoso (1996). Improved process understanding using multiway principal component analysis. *Ind. Eng. Chem. Res.* **35**, 138–146.

Ku, W., R. H. Storer and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **30**, 179–196.

Lopes, J. A. and J. C. Menezes (2004). Multivariate monitoring of fermentation processes with non–linear modelling methods. *Analytica Chimica Acta* **515**, 101–108.

MacGregor, J. F., T. E. Marlin, J. Kresta and B. Skagerberg (1991). Multivariate statistical methods in process analysis and control. In: *Cheical process control* – CPCIV. pp. 79–100.

Misra, M., H. H. Yue, S. J. Qin and C. Ling (2002). Multivariate process monitoring and fault diagnosis by multi-scale PCA. *Computers and Chemical Engineering* (26), 1281–1293.

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

Ralston, P., G. DePuy and J. H. Graham (2001). Computer–based monitoring and fault diagnosis: a chemical process case study. *ISA Transactions* (40), 85–98.

Russell, E. L., L. H. Chiang and R. D. Braatz (2000). *Data–driven Techniques for Fault Detection and Diagnosis in Chemical Process*. Springer–Verlag London.