ADCHEM 2006

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

FAULT DETECTION AND DIAGNOSIS IN INDUSTRIAL FED-BATCH CELL CULTURE

Jon C. Gunther * **Dale E. Seborg** *
Jeremy S. Conner **

* *Department of Chemical Engineering, University of California, Santa Barbara*
** *Amgen, Inc., Thousand Oaks, California*

Abstract: *Multivariate statistical process monitoring* techniques are applied to pilot-plant, cell culture data for the purpose of fault detection and diagnosis. Data from 23 batches, 20 *normal operating conditions* (NOC) and three abnormal, were available. A PCA model was constructed from 19 NOC batches, while the remaining NOC batch was used for model validation. Subsequently, the model was used to successfully detect (both offline and online) abnormal process conditions and to diagnose the root causes.

Keywords: Monitoring, cell culture processes, batch control, process control, biocontrol, biotechnology, multivariable systems

1. INTRODUCTION

Protein production cell culture has progressed significantly in recent years and is now a major source of industrially produced therapeutic agents. Because this process is sensitive to environmental conditions, successful cell culture requires precise maintenance of critical process variables (e.g., temperature, pH, and dissolved oxygen). In addition, the pharmaceutical industry is under increasing governmental pressure, such as the *Process Analytical Technology* (PAT) initiative (U.S. Food and Drug Administration, 2004), to reduce process variability.

Data-driven monitoring approaches, such as *Principal Component Analysis* (PCA), have proven to be an effective method for detecting abnormal process conditions and reducing process variability (Kourti, 2005). A particularly valuable feature of PCA is its compatibility with many of the methods available in *multivariate statistical process control* (MSPC). This statistical methodology provides a means to detect the appearance, magnitude, and duration of a process fault that causes a process to depart from proper operation (Cinar *et al.*, 2003). Also, the source of the fault can be diagnosed, assuming that the fault is observable from process data.

The objective of this research is to apply PCA and MSPC to industrial fed-batch cell culture data (courtesy of Amgen, Inc.) in an attempt to detect and diagnose abnormal process conditions using both offline and online analysis. These abnormal conditions were indicated during discussion with Amgen engineers.

2. BACKGROUND

Consider a batch process, where J process variables are measured at K instances of time. In batch MSPC applications, it is assumed that I batches conducted at *normal operating conditions* (NOC) are available for the development of a PCA model. These data are typically represented in a three-dimensional data array $\underline{\mathbf{X}}$ $(I \times J \times K)$.

For standard PCA analysis, three-dimensional array data are *unfolded* into a two-dimensional matrix. Several groups have evaluated different unfolding strategies (Nomikos and MacGregor, 1995; Wold *et al.*, 1998). The two primary unfolding techniques preserve either the I direction (i.e., batches) or the J direction (i.e., variables) of the data. For variable-wise unfolding (i.e., unfolding the data into $\mathbf{X}$ $(IK \times J)$), the nonlinear, time-varying trajectories of these data are preserved (Westerhuis *et al.*, 1999). Because batch-wise unfolding avoids this complication, it was chosen for this research. Hence, $\underline{\mathbf{X}}$ was unfolded into a matrix $\mathbf{X}$ $(I \times JK)$, such that each $I \times J$ slice is located side by side, starting with the first sampling instant. Subsequently, these data were *autoscaled* (i.e., the columns of $\mathbf{X}$ were mean-centered and scaled to unit variance) in an attempt to remove the dominance of large magnitude measurements and the nonlinear trajectories of the data from the PCA model.

For PCA $\mathbf{X}$ is expressed as the summation of the product of a score matrix $\mathbf{T}$ $(I \times A)$ and a transposed loadings matrix $\mathbf{P}^{'}$ $(A \times JK)$ plus a residual matrix $\mathbf{E}$ $(I \times JK)$, where A denotes the number of principal components, which is typically selected through a process of cross-validation (Wold, 1978):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{'} + \mathbf{E} \tag{1}$$

A major advantage of PCA modeling is its ability to compare new batch data, $\mathbf{x}_{new}$ $(1 \times JK)$, to the NOC data in a systematic fashion. PCA achieves this comparison by projecting this new data set on the PCA model generated from NOC data in order to determine the new batch scores, $\mathbf{t}_{new}$ $(1 \times A)$:

$$\mathbf{t}_{new} = \mathbf{x}_{new}\mathbf{P}(\mathbf{P}^{'}\mathbf{P})^{-1} \tag{2}$$

2.1 Offline Monitoring

For offline PCA analysis, Eq. 2 can be used to calculate $\mathbf{t}_{new}$. Note that $\mathbf{P}^{'}\mathbf{P}$ is by definition the identity matrix due to the orthonormality of $\mathbf{P}$ (Nomikos and MacGregor, 1995). After determining $\mathbf{t}_{new}$, Eq. 3 can be used to calculate the new batch residual, $\mathbf{e}_{new}$ $(1 \times JK)$.

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \mathbf{t}_{new}\mathbf{P}^{'} \tag{3}$$

Two statistical metrics are widely used to monitor disparities between the new batch and the NOC batches. *Hotelling's* T^2 statistic captures differences in the systematic part of the PCA model (i.e., $\mathbf{T}\mathbf{P}^{'}$). It is defined as follows:

$$T^2_{new} = \mathbf{t}_{new}(\mathbf{S})^{-1}\mathbf{t}^{'}_{new} \tag{4}$$

$$\mathbf{S} = \frac{\mathbf{T}^{'}\mathbf{T}}{I-1} \tag{5}$$

where $\mathbf{S}$ is the covariance matrix of the model score matrix, $\mathbf{T}$, (cf. Eq. 1) and I is the number of NOC batches. If the $\mathbf{X}$ data are from a multivariate normal distribution, T^2 follows an F distribution and α confidence limits can be calculated accordingly (Westerhuis *et al.*, 2000):

$$T^2_{\alpha} = \frac{A(I^2-1)}{I(I-A)}F_{A,I-A,\alpha} \tag{6}$$

A second metric, the *Sum of Squared Residuals* Q, captures the information in the residuals,

$$Q_{new} = \mathbf{e}_{new}\mathbf{e}^{'}_{new} \tag{7}$$

where Q_{new} is assumed to be χ^2 distributed. A method for approximating α confidence limits based upon this assumption (Jackson and Mudholkar, 1979) is used in this paper:

$$\begin{aligned} Q_{\alpha} &= \theta_1[1 - \frac{\theta_2 h_0(1-h_0)}{\theta_1^2} + \\ &\quad \frac{z_{\alpha}(2\theta_2 h_0^2)^{1/2}}{\theta_1}]^{1/h_0} \\ \mathbf{V} &= \frac{\mathbf{E}\mathbf{E}^{'}}{I-1} \\ \theta_i &= trace(\mathbf{V}^i) \text{ for } i = 1, 2, \text{ and } 3 \\ h_0 &= 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \end{aligned} \tag{8}$$

2.2 Online Monitoring

From an operational perspective, it is preferable to monitor the batch online, as it progresses, so that corrective or terminative action can be taken in a timely manner. However, to evaluate process data from a new batch using Eqs. 2-8, the new batch is required to have the same number of columns as the NOC data (i.e., JK columns). This is not possible when the batch is incomplete and thus future measurements are missing from the new batch (i.e., $\mathbf{x}_{new}$ only has Jk columns where $k \leq K$). For the unfinished new batch, the missing future data must be estimated in order to proceed. Several solutions to this problem have been proposed and evaluated (Nomikos and MacGregor, 1995). The *PCA Projection* method is used in this paper. It only uses the portion of the loading matrix corresponding to the elapsed time period until the current sampling instant k to calculate the new score vector, $\mathbf{t}_{new}(k)$ $(1 \times Jk)$:

$$\mathbf{t}_{new}(k) = \mathbf{x}_{new,1:Jk}\mathbf{P}_{1:Jk}(\mathbf{P}^{'}_{1:Jk}\mathbf{P}_{1:Jk})^{-1} \tag{9}$$

For online monitoring, the term $\mathbf{P}^{'}_{1:Jk}\mathbf{P}_{1:Jk}$ in Eq. 9 is not necessarily identity until $k = K$. At

sample k, $\mathbf{e}_{new}(k)$ and $T^2_{new}(k)$ can be evaluated in a manner similar to Eqs. 3 and 4 noting that $T^2_{new}(k)$ is calculated from a time-varying scores covariance matrix, $\mathbf{S}(k)$. To monitor the residuals, the *Squared Prediction Error*, $SPE_{new}(k)$, was utilized:

$$SPE_{new}(k) = \sum_{j=1}^{J} \mathbf{e}_{new,jk}(k)^2 \tag{10}$$

Another significant benefit of PCA is its ability to determine process variable contributions from $T^2_{new}(k)$ and $SPE_{new}(k)$. These contributions can then be used for fault diagnosis. For online monitoring, the contributions can be calculated in the following manner (Westerhuis *et al.*, 2000):

$$C_{T^2_{jk}} = \sum_{a=1}^{A} \mathbf{S}^{-1}_{k,aa} \mathbf{t}_{new,a}(k) \mathbf{x}_{new,jk} \mathbf{P}_{jk,a} \tag{11}$$

$$C_{SPE_{jk}} = \mathbf{e}_{new,jk}(k)^2 \tag{12}$$

where $\mathbf{S}_{k,aa}$ is the ath diagonal element of the time-varying covariance matrix and the subscripts j and k represent a single process variable and a single sampling instant, respectively. Confidence limits for $C_{SPE_{jk}}$ are determined in the same way as for $SPE_{new}(k)$. However, confidence limits for $C_{T^2_{jk}}$ are calculated in a *jackknife* procedure. In this approach, each NOC batch is omitted in a sequential manner and contributions for each batch are calculated. The estimated mean and standard deviation are then used as 3σ limits (Westerhuis *et al.*, 2000).

3. PROCESS DESCRIPTION

The fed-batch cell culture experiments were performed at Amgen Process Development. A controlled environment within the reactor was maintained with cascaded PID feedback loops for DO and pH. The key process variables used in this PCA research are summarized in Table 1.

Data for 23 batches were available. These batches were all conducted at nearly identical process conditions and possessed approximately equal time duration. NOC batches 1-19 were used in PCA model development, while NOC batch 20 was used for model validation and batches 21-23 were used for detection of abnormal situations. Amgen personnel categorized batches 21-23 as abnormal due to irregular thermal heating (21), DO controller problems (22), and agitator problems that led to a future device failure (23).

Table 1. Process variable measurements used in the PCA model.

Variable	Abbreviation
Agitation	AG
Agitation controller output	AGc
Inlet air flow	AF
Inlet air flow controller output	AFc
Inlet CO_2 flow	CO2
Inlet CO_2 flow controller output	CO2c
Dissolved oxygen	DO
Dissolved oxygen controller output	DOc
Inlet O_2 flow	O2
Inlet O_2 flow controller output	O2c
Vessel temperature	T
Vessel temperature controller output	Tc
pH	pH
pH controller output	pHc

4. RESULTS

A PCA model was constructed using NOC batches 1-19 data for the 14 process variables in Table 1. Cross-validation was performed in order to select an appropriate number of principal components, three.

To evaluate the ability of the PCA model to detect process abnormalities and reject false positives, overall batch T^2 and Q values were determined (see Figs. 1 and 2). It is clear that abnormal batches 21-23 exceed the 99% confidence limits for Q, while NOC validation batch 20 does not.

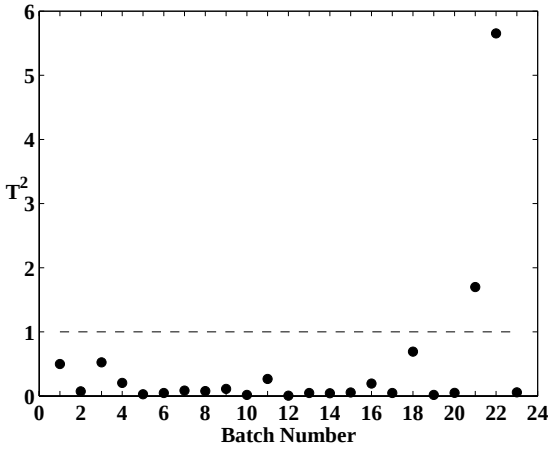

Fig. 1. Overall batch T^2. Batches 1-19 were calibration, batch 20 was validation, and batches 21-23 were abnormal. The dashed line denotes the 99% confidence limits.

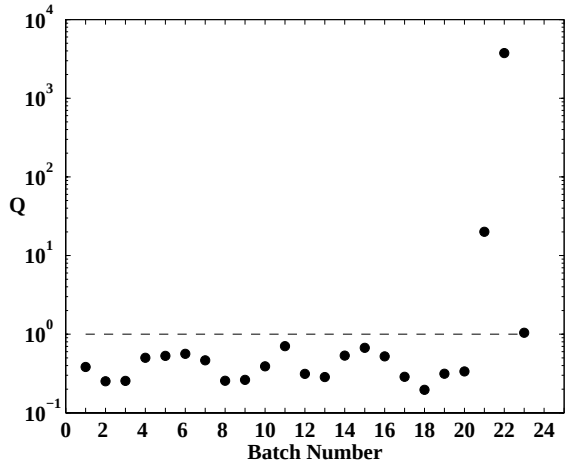

Fig. 2. Overall batch Q. Note that a semilog scale is used. Batches 1-19 were calibration, batch 20 was validation, and batches 21-23 were abnormal. The dashed line denotes the 99% confidence limits.

From an operational perspective, it is desirable to detect the onset of abnormal operation before the batch is finished. To fulfill this objective, online $T^2(k)$ and $SPE(k)$ were calculated. In Figure 3 the results for validation batch 20 are displayed. The $T^2(k)$ confidence limit is not violated, while the seven $SPE(k)$ confidence limit violations that occur are not exceptional for 1966 samples.

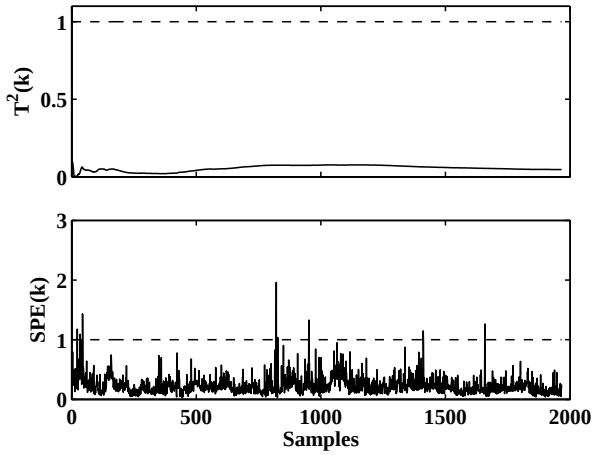

Fig. 3. Online normalized $T^2(k)$ and $SPE(k)$ for batch 20. The dashed lines denote the 99% confidence limits.

In Figure 4 it is evident that the $SPE(k)$ confidence limits are violated for batch 21 for the entire duration of the batch, while a sustained $T^2(k)$ violation occurs for all samples $k < 1100$. To diagnose the cause of this abnormal situation, a contribution plot (Fig. 5) was generated and identifies the temperature controller output as being the major source of abnormal process conditions. From inspection of the vessel temperature controller output (Tc) time-series data in Figure 6, it is clear that batch 21 is abnormal in comparison to the average NOC batch trajectory for batch 21. Amgen engineers indicated that for batch 21 the reactor possessed a unique thermal heating jacket that resulted in elevated Tc values.

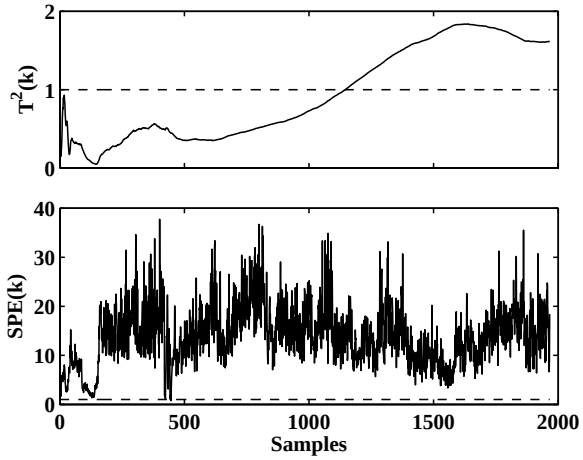

Fig. 4. Online normalized $T^2(k)$ and $SPE(k)$ for batch 21. The dashed lines denote the 99% confidence limits.

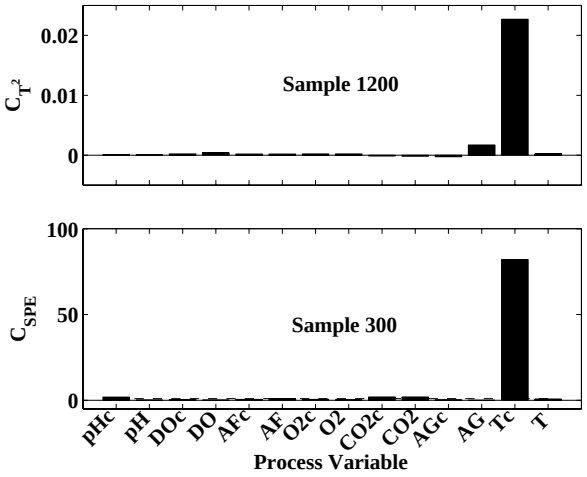

Fig. 5. Process variable contributions to online $T^2(k)$ and $SPE(k)$ at designated samples for batch 21.

For batch 22 the $T^2(k)$ and $SPE(k)$ confidence limits in Figure 7 are violated from the onset of the batch. From the contribution plot in Figure 8, it is obvious that DO was abnormal in both the score and residual spaces. Figure 6 reveals that for the early period of operation ($k < 700$) the DO values were indeed large in comparison to the average NOC batch trajectory.

For batch 23 the abnormal process conditions are more difficult to detect. An abnormally large number of confidence limit violations occur for $SPE(k)$ in Figure 9, but none occur for $T^2(k)$. However, in Figure 10 the contribution plot clearly indicates abnormal agitation. In Figure 6 batch

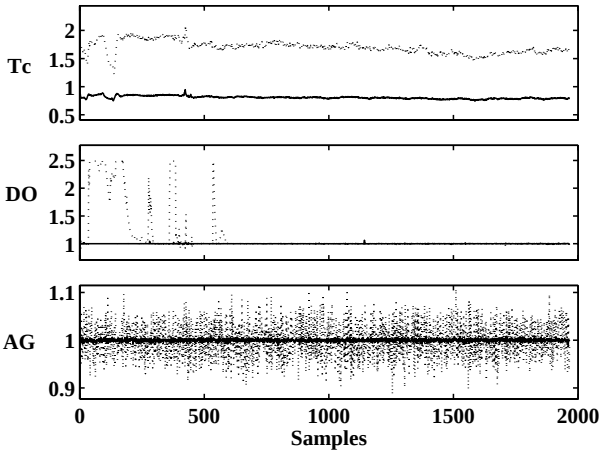

Fig. 6. Time-series plots for process variables most affected by abnormal process conditions for batches 21 (top), 22 (middle), and 23 (bottom). Solid line represents average NOC batch trajectory, while the dotted line represents the particular batch.

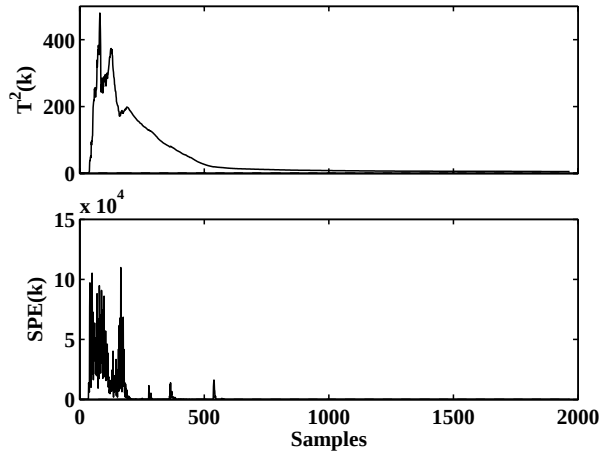

Fig. 7. Online normalized $T^2(k)$ and $SPE(k)$ for batch 22. The dashed lines denote the 99% confidence limits.

23 appears to possess considerably more agitation variation than the average batch trajectory. Amgen engineers reported that the agitator for this reactor failed during the next period of operation.

5. CONCLUSIONS

In this paper, MSPC and PCA techniques are applied to industrial fed-batch cell culture data. It was shown that a PCA model can successfully detect abnormal process conditions resulting from differences in the equipment (batch 21), operational issues (batch 22), and imminent device failure (batch 23). Analysis of contribution plots indicated that abnormal Tc levels, elevated DO values, and large agitation variation were the major sources of abnormal process conditions

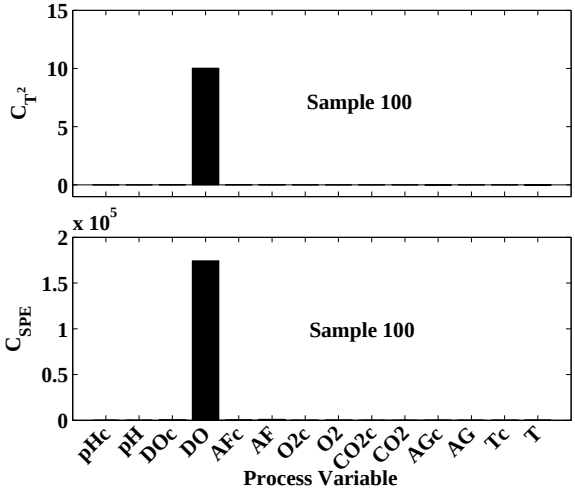

Fig. 8. Process variable contributions to online $T^2(k)$ and $SPE(k)$ at designated samples for batch 22.

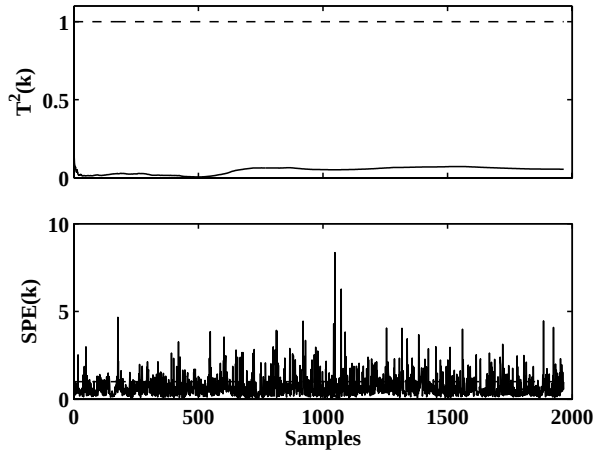

Fig. 9. Online normalized $T^2(k)$ and $SPE(k)$ for batch 23. The dashed lines denote the 99% confidence limits.

found in batches 21, 22, and 23 respectively. The PCA explanation of these process abnormalities is consistent with the process behavior reported by Amgen engineers.

ACKNOWLEDGEMENTS

Financial and technical support provided by Amgen, Inc. is gratefully acknowledged.

REFERENCES

Cho, H.-W. and K.-J. Kim (2003). A method for predicting future observations in the monitoring of a batch process. *J. Qual. Tech.* **35**(1), 59–69.

Cinar, A., S. J. Parulekar, C. Ündey and G. Birol (2003). *Batch Fermentation: Modeling, Monitoring, and Control*. Marcel Dekker. New York.

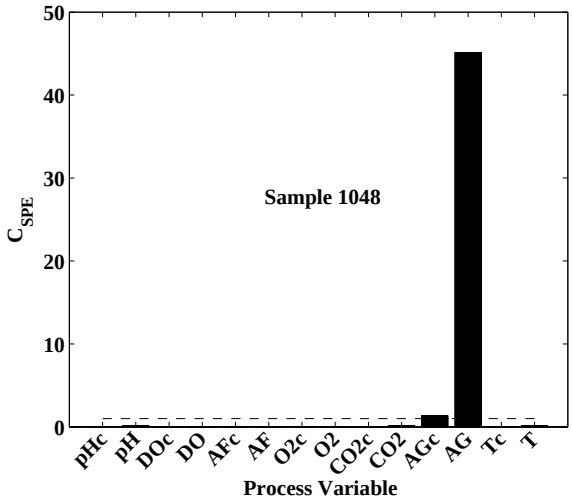

Fig. 10. Process variable contributions to online $T^2(k)$ and $SPE(k)$ at designated samples for batch 23.

Jackson, J. E. and G. S. Mudholkar (1979). Control procedures for residual associated with principal component analysis. *Technometrics* **21**(3), 341–349.

Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *Internat. J. Adapt. Control Signal Process.* **19**, 213–246.

Nomikos, P. and J. F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics* **37**(1), 41–59.

U.S. Food and Drug Administration (2004). Guidance for industry PAT: A framework for innovative pharmaceutical development, manufacturing, and quality assurance.

Vinci, V. A. and S. R. Parekh (2003). *Handbook of Industrial Cell Culture*. Humana Press. Totowa, New Jersey.

Westerhuis, J. A., S. P. Gurden and A. K. Smilde (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* **51**, 95–114.

Westerhuis, J. A., T. Kourti and J. F. MacGregor (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemometrics* **13**, 397–413.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**(4), 397–405.

Wold, S., N. Kettaneh, H. Fridén and A. Holmberg (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **44**, 331–340.