Fault Detection and Diagnosis in Industrial Fed-Batch Fermentation

Jon C. Gunther, Dale E. Seborg, and Jeffrey Baclaski

Abstract—This paper applies *multivariate statistical process control* (MSPC) techniques to pilot plant fermentation data for the purpose of fault detection and diagnosis. Data from ten batches, nine normal operating conditions (NOC) and one failed, were available. A *principal component analysis* (PCA) model was constructed from eight NOC batches, while the remaining NOC batch was used for model validation. Subsequently, the model was used to successfully detect (both offline and online) a process abnormality in the failed batch and diagnose the factors contributing to the fault. These monitoring results agree with the observed biological phenomena encountered during this batch.

I. INTRODUCTION

Many pharmaceuticals are produced by utilizing genetically modified microorganisms. These organisms require a specifically controlled environment for optimal growth and product synthesis. Many pharmaceutical manufacturers capitalize upon the beneficial features of fed-batch fermentation to achieve this requisite. This process allows for the maintenance of critical process variables (e.g., temperature, pH, and dissolved oxygen) at specific levels through feedback control. Yet, few first principle models accurately capture the biological activity of these organisms, and thus industrial fermentation installations seldom use these models for control purposes. At the same time, the pharmaceutical industry is under increasing pressure to reduce variability in their process by governmental entities, such as the U.S. Food and Drug Administration (e.g., the *Process Analytical Technology* (PAT) initiative [1]).

Data-driven modeling approaches, such as PCA, have been shown to be an effective alternative to first principles models for a variety of different processes in several industries [2]. A particularly valuable feature of PCA is its compatibility with many of the methods available in MSPC. This statistical methodology provides a means to detect the appearance, size, and instant of a process fault that causes a process to deviate from proper operation in multivariate systems [3]. Also, the source of the fault can often be diagnosed, if the fault is observable in the process data.

The objective of this research is to apply PCA and MSPC to industrial fermentation data (courtesy of Amgen, Inc.) in order to detect and diagnose process faults using both offline and online analysis.

J. C. Gunther is with the Department of Chemical Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106-5080, USA jgunther@engineering.ucsb.edu

D. E. Seborg is with the Department of Chemical Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106-5080, USA seborg@engineering.ucsb.edu

J. Baclaski is with Amgen, Inc., One Amgen Center Dr., Thousand Oaks, CA 93320-1799, USA jbacclask@amgen.com

II. BACKGROUND

For a batch process, J process variables are measured at K instances of time. In batch MSPC applications, it is assumed that I batches conducted under *normal operating conditions* (NOC) are available for the development of a PCA model. These data are typically represented in the 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 array. Several groups have evaluated different unfolding strategies [4], [5]. The two primary unfolding techniques preserve either the I (i.e., batch) or the J (i.e., variable) directions of the data. However, with variable-wise unfolding (i.e., unfolding the data into $\mathbf{X}$ ($IK \times J$)), the nonlinear, time-varying trajectories of these data are preserved after conventional preprocessing of the data [6]. Due to the fact that 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 situated side by side, starting with the first sampling instant [7]. 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 underlying nonlinear trajectories of the data from the PCA model.

In PCA $\mathbf{X}$ is expressed as the product of a scores matrix $\mathbf{T}$ ($I \times A$) and a transposed loadings matrix $\mathbf{P}'$ ($A \times JK$) plus a residual matrix $\mathbf{E}$.

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (1)$$

A represents the number of principal components, which is typically selected through a process of cross-validation [8]. A major advantage of PCA modeling is its ability to compare new batch data, $\mathbf{x}_{new}$ ($1 \times JK$), to old NOC data in a methodical fashion. This is achieved by projecting this new data set upon the PCA model generated from NOC data in order to determine the new batch scores, $\mathbf{t}_{new}$ ($1 \times A$).

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

For offline PCA analysis, $\mathbf{P}'\mathbf{P}$ is the identity matrix. It should be noted that for online monitoring, a reduced portion of the loading matrix $\mathbf{P}$ is used, and therefore $\mathbf{P}'\mathbf{P}$ is not equal to the identity matrix. Equation 2 can be used to generate the predicted value, $\hat{\mathbf{x}}_{new}$ and the new batch residual, $\mathbf{e}_{new}$ ($1 \times JK$).

$$\mathbf{e}_{new} = \mathbf{x}_{new} - \mathbf{t}_{new}\mathbf{P}' \quad (3)$$

Two statistical metrics are widely used to monitor disparities between the new data set and the NOC batches. *Hotelling's*

T^2 statistic captures differences in the systematic part of the PCA model (i.e., $\mathbf{TP}'$). It is defined as follows:

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

$$\mathbf{S} = \frac{\mathbf{T}\mathbf{T}'}{I-1} \quad (5)$$

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 [4]. If the $\mathbf{X}$ data are from a multivariate normal distribution, T^2 follows an F distribution and $(1 - \alpha) \times 100\%$ confidence limits can be calculated accordingly [9]:

$$T_{\alpha}^2 = \frac{A(I^2 - 1)}{I(I - A)} F_{A, I-A, \alpha} \quad (6)$$

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

$$Q_{new} = \mathbf{e}_{new}\mathbf{e}_{new}' \quad (7)$$

where Q_{new} is χ^2 distributed. A method for approximating $(1 - \alpha) \times 100\%$ confidence limits where the data is assumed to possess a multivariate normal distribution is used in this paper [10]:

$$Q_{\alpha} = \theta_1 \left[1 - \frac{\theta_2 h_0 (1 - h_0)}{\theta_1^2} + \frac{z_{\alpha} (2\theta_2 h_0^2)^{1/2}}{\theta_1} \right]^{1/h_0} \quad (8)$$

$$\mathbf{V} = \frac{\mathbf{E}\mathbf{E}'}{I-1}$$

$$\theta_i = \text{trace}(\mathbf{V}^i) \text{ for } i = 1, 2, \text{ and } 3$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$$

From an operational perspective, it is desirable 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, 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., $\mathbf{x}_{new}$ has Jk elements where $k \leq K$). For the unfinished new batch, the missing data must be estimated for each process variable in order to proceed. Several solutions to this problems have been proposed and tested [4]. The *PCA Projection* method is used in this paper. This method only uses the portion of the loading matrix corresponding to the elapsed time until the current moment to calculate the new score vector, $\mathbf{t}_{new}(k)$ ($1 \times Jk$).

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

At sample k , $\mathbf{e}_{new}(k)$ and $T_{new}^2(k)$ can be evaluated in a fashion similar to that in equations 3 and 4 noting that $T_{new}^2(k)$ is calculated from a time-varying scores covariance

matrix, $\mathbf{S}(k)$. For monitoring the residuals, the Squared Prediction Error, $SPE_{new}(k)$ was utilized

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

Another significant benefit of PCA is its ability to determine process variable contributions from $T_{new}^2(k)$ and $SPE_{new}(k)$. These contributions can then be used for fault diagnosis. For online monitoring, these contributions can be calculated in the following manner [9]:

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

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

where $\mathbf{S}_{k,aa}^{-1}$ is the a th diagonal element of the time-varying covariance matrix. Confidence limits for $C_{SPE_{jk}}$ are determined in the same way as for $SPE_{new}(k)$, while confidence limits for $C_{T_{jk}^2}$ are calculated in a *jackknife* procedure where each NOC batch is omitted and contributions for each batch are calculated yielding an estimated mean and standard deviation that are used as 3σ limits [9].

III. PROCESS DESCRIPTION

The fed-batch fermentation experiments were performed in a 300 L reactor at Amgen Process Development. The host cell species was *Escherichia coli*, which had been modified by recombinant DNA techniques to produce a desired protein. To induce protein production, a chemical species was added to the reactor after approximately one third of the batch duration. A controlled environment within the reactor was maintained with cascaded PID control loops. The key process variables that were available for PCA analysis are summarized in Table I.

TABLE I
PROCESS VARIABLE MEASUREMENTS USED IN THE PCA MODEL.

Variable	Abbreviation
Agitation	AG
Inlet air flow	Air
Dissolved oxygen	DO
Dissolved oxygen controller output	DOc
Jacket temperature	TJ
Jacket temperature controller output	TJc
Inlet O ₂ flow	O2
Vessel temperature	T
Vessel temperature controller output	Tc
pH	pH
pH controller output	pHc

Data for ten batches were available and each batch contains 1084 samples. NOC batches 1-8 were used in PCA

model development, while NOC batch 9 was used for model validation and batch 10 was used for failure detection. Amgen personnel categorized batch 10 as a failure due to its elevated acetate levels, resulting in a low cell concentration (see Fig. 1).

Fig. 1. Time series plot for acetate concentration for all ten batches.

IV. RESULTS

In another industrial PCA fermentation study [11], the data were divided into two regimes (i.e., pre-induction and post-induction) and modeled separately. This approach was also used in this investigation. But given the greater importance of early batch phenomena, only the pre-induction period data is considered here.

Also, the Eigenvector, Inc. *PLS Toolbox 3.5* [14] was used to generate these results.

A. Fault Detection and Diagnosis

A PCA model was constructed using data from the seven process variables in Table I. Table II provides a summary of the variance captured by each principal component.

In order to select an appropriate number of principal components for the PCA model, cross-validation was conducted and two principal components were selected. Because only nine NOC batches were available for this research, the choice of which eight batches were used in the calibration set, with the remaining batch being used for validation, could have a significant effect upon the research results. Therefore, each batch was individually excluded from the calibration set. To compare how the different combinations of batches affected results, the number of false positives for the NOC validation batch and incorrect detection of the failed batch,

TABLE II
PERCENT VARIANCE CAPTURED BY PCA MODEL.

Principal Component Number	Eigenval. of Covar. Matrix	Variance Capt. by this PC [%]	Variance Capt. Total [%]
1	3140	26.4	26.4
2	2460	20.6	47.0
3	1620	13.6	60.6
4	1350	11.3	71.9
5	1260	10.6	82.4
6	1090	9.1	91.6
7	1010	8.4	100

using the overall T^2 and Q metrics, were tabulated. Both T^2 and Q successfully detected the failed batch and no false positives for all of the batch combinations occurred for T^2 . However, in four of the nine combinations, false positives appeared in Q . As in other studies [12], it appears that the validation residuals are predisposed to be larger than those in the calibration set, and thus, false positives for Q should be expected especially for such a small number of batches. Figs. 2-8 were generated for a subset of batches that did not yield a Q false positive for the validation batch.

In Figs. 2 and 3, it is clear that failed batch 10 is very different from the batches in the calibration set, since this batch exceeds 99% confidence limits of the scores, T^2 , and Q plots. However, NOC batch 9, which was omitted from the calibration set, is within these 99% confidence limits, thus validating the PCA model.

Fig. 2. Scores plot.

In addition to performing this offline investigation, online $T^2(k)$ and $SPE(k)$ were considered for batches 9 and 10

Fig. 3. Offline batch T^2 and Q . Note that logarithmic vertical axes were used for each metric so that failed batch 10 could be included.

(see Figs. 4 and 5). It should be noted that these the data of these figures were normalized to their respective 99% confidence limits.

Fig. 4. Online normalized $T^2(k)$ and $SPE(k)$ for NOC batch 9. The horizontal lines are the 99% confidence limits.

In Fig. 4 it is evident that batch 9 does not violate the $T^2(k)$ 99% confidence limit, while there is a large number of violations of the $SPE(k)$ 99% confidence limit. This is likely due to the small number of NOC batches. Although there are some violations for batch 9, they are small compared to the number and size of violations for batch 10. In Fig. 5 long term sustained violations starting at at sample 600 in $SPE(k)$ and sample 700 in $T^2(k)$ are evident.

Fig. 5. Online normalized $T^2(k)$ and $SPE(k)$ for failed batch 10. Note that logarithmic axes were used for each metric to show greater detail.

In order to diagnose batch 10, contribution plots for $T^2(k)$ and $SPE(k)$ plots are shown in Fig. 6. It is apparent that the process variables associated with the DO control loop (i.e., DO, DOc, and Air) are most responsible for the large $T^2(k)$ and $SPE(k)$ values. Upon closer examination, DO and DOc deviate significantly from the calibration set average trajectories (see Fig. 7).

Fig. 6. Process variable contributions to online $T^2(k)$ and $SPE(k)$ at designated samples for failed batch 10.

Discussions with Amgen engineers indicated that a large acetate concentration contributed to the batch 10 failure due to the inhibitive effects of acetate on cell growth. Acetate is excessively produced by *E. coli* cells during two

Fig. 7. Time series plots for DO and DO controller output for batches 10 and the average batch trajectory.

environmental conditions [13]:

- 1) Under anaerobic conditions.
- 2) Under aerobic conditions with excessive glucose present.

In batch 10, there is a ~ 50 sample period in which DO levels are nearly zero, thus creating a momentary anaerobic environment. The abnormal process variables detected by the PCA contribution plots (see Fig. 6) are consistent with the biological conditions that engineers reported led to the process failure.

B. Online vs. Offline Analysis

The results in Section IV-A were based on online monitoring calculations, although offline monitoring calculations could have also been performed more easily because all batch data were available for investigation (unlike in an industrial setting). However, Westerhuis et al. [9] warn of problems with offline monitoring when inspecting $SPE(k)$. A *smearing* phenomenon, where residual information is spread across samples of the entire batch, can occur when calculating an offline $SPE(k)$. This phenomenon is apparent in Fig. 8.

In the bottom portion of Fig. 8, the $SPE(k)$ exceeds the confidence limits for the entire duration of the batch, while in the top figure, a confidence limit violation does not occur until $\sim$ sample 600. In the offline approach, the effect of the process fault has been distributed across all samples. Thus, it is recommended that when attempting to determine the instant of fault appearance, $SPE(k)$ should be calculated in an online fashion.

Fig. 8. Online (top) and offline (bottom) normalized $SPE(k)$ for batch 9. Note the appearance of smearing in the offline plot.

V. CONCLUSIONS

In this paper, MSPC and PCA techniques have been applied to industrial fed-batch fermentation data. It has been shown that a PCA model constructed from a small number of data sets (i.e., eight) can successfully detect and diagnose the circumstances leading to a failed batch. Analysis of contribution plots indicated that abnormal DO and DOc levels were the predominant causes for the PCA chart violations. These low DO (or anaerobic) conditions likely contributed to the observed high acetate concentration and low cell density for the failed batch. Online calculation of $SPE(k)$ is preferred over the offline alternative because *information smearing* can occur in the offline calculation of $SPE(k)$.

VI. ACKNOWLEDGMENTS

This research was generously supported by Amgen, Inc. Their financial and technical support is gratefully acknowledged.

Also, Eigenvector, Inc. is recognized for generously providing consultation.

REFERENCES

- [1] U.S. Food and Drug Administration, "Guidance for industry PAT: A framework for innovative pharmaceutical development, manufacturing, and quality assurance," 2004. [Online]. Available: <http://www.fda.gov/cder/guidance/6419fnl.pdf>
- [2] T. Kourti, "Application of latent variable methods to process control and multivariate statistical process control in industry," *Internat. J. Adapt. Control Signal Process.*, vol. 19, pp. 213–246, 2005.
- [3] A. Çinar, S. J. Parulekar, C. Ündey, and G. Birol, *Batch Fermentation: Modeling, Monitoring, and Control*. New York: Marcel Dekker, 2003.

- [4] P. Nomikos and J. F. MacGregor, "Multivariate SPC charts for monitoring batch processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995.
- [5] S. Wold, N. Kettaneh, H. Fridén, and A. Holmberg, "Modelling and diagnostics of batch processes and analogous kinetic experiments," *Chemom. Intell. Lab. Syst.*, vol. 44, pp. 331–340, 1998.
- [6] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Comparing alternative approaches for multivariate statistical analysis of batch process data," *J. Chemom.*, vol. 13, pp. 397–413, 1999.
- [7] H.-W. Cho and K.-J. Kim, "A method for predicting future observations in the monitoring of a batch process," *J. Qual. Tech.*, vol. 35, no. 1, pp. 59–69, 2003.
- [8] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models," *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978.
- [9] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, "Generalized contribution plots in multivariate statistical process monitoring," *Chemom. Intell. Lab. Syst.*, vol. 51, pp. 95–114, 2000.
- [10] J. E. Jackson and G. S. Mudholkar, "Control procedures for residual associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.
- [11] L. Chiang, A. Kordon, L. Chew, D. Coffey, R. Waldron, T. Bruck, K. Haney, A. Jenings, and H. Talbot, "Multivariate analysis for quality improvement of an industrial fermentation process," in *7th International Symposium on Dynamics and Control of Process Systems (DYCOPS)*. Boston: IFAC, 2004.
- [12] E. N. M. van Sprang, H.-J. Ramaker, J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, "Critical evaluation of approaches for on-line batch process monitoring," vol. 57, pp. 3979–3991, 2002.
- [13] F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umabarger, Eds., *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd ed. Washington, D.C.: American Society of Microbiology Press, 1996, vol. 1-2.
- [14] B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, and R. S. Koch, *PLS Toolbox 3.5*, Eigenvector Research, Inc., 2005.