Real time phase detection based online monitoring of batch fermentation processes

Soumen K. Maiti \textsuperscript{a}, Rajesh K. Srivastava \textsuperscript{b}, Mani Bhushan \textsuperscript{a}, Pramod P. Wangikar \textsuperscript{a,b}

\textsuperscript{a}Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
\textsuperscript{b}School of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

**ABSTRACT**

Industrial fermentations conducted in a batch or semi-batch mode demonstrate significant batch-to-batch variability. Current batch process monitoring strategies involve manual interpretation of highly informative but low frequency offline measurements such as concentrations of products, biomass and substrates. Fermentors are also fitted with computer interfaced instrumentation, enabling high frequency online measurements of several variables and automated techniques which can utilize this data would be desirable. Evolution of a batch fermentation, which typically uses complex medium, can be conceptualized as a sequence of several distinct metabolic phases. Monitoring of batch processes can then be achieved by detecting the phase change events, also termed as singular points (SP). In this work, we propose a novel moving window based real-time monitoring strategy for SP detection based only on online measurements. The key hypothesis of the strategy is that the statistical properties of the online data undergo a significant change around an SP. The strategy is easily implementable and does not require past data or prior knowledge of the number or time of occurrence of SPs. The efficacy of the proposed approach has been demonstrated to be superior compared to that of reported techniques for industrially relevant model organisms. The proposed approach can be used to decide offline sampling timings in real time.

**Keywords:**
Principal component analysis
Mean
Covariance
Moving window
Singular point

1. Introduction

Fermentation processes are widely used in food, pharmaceutical, agrochemical and chemical industries. The production units range from small scale for biopharmaceuticals to large scale for bulk chemicals. A majority of the processes are operated in a batch or semi-batch mode. Intense competition and regulatory requirements pose severe demands on consistency of these batches in terms of the end of batch productivity and product quality [1]. However, fermentation processes are subject to intrinsic batch-to-batch variability due to variability in raw material quality, state of the seed culture and operator skills. It is therefore desirable to automate monitoring, fault detection and diagnosis and control of fermentation processes. This can lead to improved process reliability, product quality and productivity as well as reduced development time, manpower inputs and cost of production [2].

Typically, during operation, the product quality and batch performance are monitored via off-line measurements of concentrations of the product, byproducts, biomass and substrates. These measurements are expensive, labor intensive and time consuming, are obtained at low frequencies (e.g., every few hours) at pre-defined intervals and hence, may not always lead to timely information about the status of the batch. Further, in some processes, the product formation begins only towards the later parts of the batch and this leads to additional difficulty in adequately monitoring the process using these offline measurements [3]. Fermentors are typically equipped with several on-line sensors such as pH, temperature, concentrations of dissolved oxygen (DO) and carbon dioxide and partial pressure of oxygen and carbon dioxide in the exhaust gas. These measurements are inexpensive, usually available at high frequencies (e.g., every few seconds) and are obtained in an automated fashion. Hence, there is enormous potential to use these measurements to effectively monitor batch fermentation processes.

In the general process systems engineering literature, several different techniques have been reported for process monitoring and fault diagnosis [4]. These can be broadly classified as process model based, knowledge based and historical data based. The success of any model based strategy depends critically on the
adequacy of the underlying model. Industrial fermentation processes typically employ complex media with multiple substitutable carbon and nitrogen substrates, which leads to difficulties in developing adequate process models. Further, several aspects of fermentation processes such as the dynamic evolution of pH and concentration of dissolved oxygen, are not well understood in general and this may lead to additional difficulties in developing reliable process models. Hence, model based strategies may not be suitable for monitoring of majority of industrial fermentation processes. Knowledge based monitoring techniques such as those based on fuzzy logic require expert knowledge of the system and therefore are system specific [4,5]. Such expert knowledge may not be available for the system of interest. Further, even for cases where such knowledge exists in terms of the manpower knowledgeable about the system, it is not straightforward to translate such knowledge to a form that can be readily utilized by automated monitoring systems. Historical data based methods rely on large amount of past data to capture the underlying relationships between the process variables [4,6,7]. However, due to batch-to-batch variability intrinsic to fermentation processes, it is difficult for these techniques to delineate between normal and abnormal variations.

Another set of methods, based on ideas from statistical control literature, have been proposed that rely only on data available from the current batch [8–10]. Fermentation processes typically utilize complex organic substrates such as yeast extract in addition to defined components such as glucose and ammonia. This provides a substitutable multisubstrate milieu, which may result in sequential and/or simultaneous utilization of the substrates. The cellular metabolism may be different in each such substrate uptake phase [11]. Evolution of a batch fermentation process can then be conceptualized as a sequence of such phases, each with its own duration and dynamics. It is expected that batch-to-batch variability would therefore, among other things, translate to variations in switching times between the phases [12]. Hence, effective monitoring can be achieved by detecting the time of occurrence of these various phases. The reported technique based on this philosophy detects the phase change time by identifying qualitative changes in trajectories of the test statistic $T^2$ and principal component score plots [9]. Being qualitative in nature, this technique is difficult to automate. While other statistical process monitoring techniques such as Shewhart Charts, Cumulative sum (CUSUM) and Exponentially weighted moving average (EWMA) [8,10,13] have been applied for monitoring batch processes in general, they have not been specifically applied for monitoring fermentation processes characterized by multiple phases since it is typically assumed in these techniques that the entire batch data is characterized by single set of statistical properties (such as mean and covariance).

In this article, we present a real time phase detection based process monitoring scheme that does not require process model or historical data. The scheme is inspired from statistical control literature, is multivariate in nature, relies only on online measurements and can be easily automated to work with industrial processes. The basic premise in our approach is that statistical properties of online measured data are different in different phases. Hence, the problem of phase change detection is treated to be equivalent to that of detection of changes in statistical properties of the data. To be consistent with earlier work [9,12], we refer to a point where phase change is detected as a singular point (SP).

2. Experimental methods

In this study, experimental data has been collected for two different strains, Amycolatopsis balhimycina DSM5908 and Bacillus pumilus ATCC 21951 while the data for Amycolatopsis mediterranei S699 was taken from Doan et al. [9]. For A. balhimycina and B. pumilus, the fermentation experiments were performed in a 2.5 l fermentor equipped with various sensors and data acquisition system (Model: BINOS1002 M, Rosemount Analytical, Germany). The fermentor was aerated at a constant flow rate of 1.0vvm (volume of air per unit volume of medium per minute) using a mass flow controller. Dissolved oxygen (DO) concentration in the fermentor was maintained at 40% of saturation value by controlling the stirrer speed in cascade mode with DO. The concentrations of oxygen and carbon dioxide in the exhaust gas were measured by infrared spectroscopy and paramagnetic analysis, respectively (Analyzer BINOS1002 M, Rosemount Analytical, Germany). The online measurements were stored at 5 min intervals.

The Amycolatopsis balhimycina strain was a gift from Prof Anna Eliasson Lantz of Denmark’s Technical University, Denmark, and was stored on Bennett agar plates at 4 °C. Seed culture was grown in 100 ml medium in a 500 ml capacity Erlenmeyer flask with single baffle and incubated at 30 °C and 150 rpm. The seed medium contained per liter of distilled water: glucose: 15 g, glycerol: 5 g, soya peptone: 15 g, NaCl: 5 g and yeast extract: 3 g. Upon reaching an optical density of ~12 at 600 nm, 25 ml of the seed culture was transferred to a fermentor containing 1 l of production medium. The production medium contained, per liter of distilled water, glucose: 54–100 g, glycerol: 0–16 g, ammonium sulfate: 3–6.6 g, trisodium citrate: 0.25 g, MgSO4·7H2O: 1.5 g, MnSO4·H2O: 0.001 g, NaCl: 1 g, MES: 0.145 g and KH2PO4: 0.2 g. In addition, the following vitamins were added: biotin: 0.0005 g, calcium-pantothenate: 0.001 g, nicotinic acid: 0.001 g, myo-inositol: 0.025 g, thiamin HCl: 0.001 g, pyridoxine HCl: 0.001 g and para-aminobenzoic acid: 0.0002 g. Temperature was maintained at 30 °C and pH was maintained at 7.0 by adding 1.5N NaOH solution by using a pH controller. The online measurements included NaOH flow rate, pH, agitator speed and DO concentration. Oxygen, agitator speed and CO2 and O2 concentration in exhaust gas.

For both the strains, samples were drawn from the fermentation medium at regular intervals to obtain the time profiles of concentrations of dry cell weight (DCW), product(s) and substrate(s). Glucose, glycerol, D-ribose, acetate, acetoin and 2,3-butanediol were analyzed via RI detector on HPLC (Hitachi, Merck KGaA, Darmstadt, Germany) using HP-Aminex-87-H column (Biorad, Hercules, CA, USA) with column temperature maintained at 60 °C. A mobile phase of 5 mM sulfuric acid with flow rate of 0.6 ml/min was used. The concentration of free amino acids was estimated via the ninhydrin method. The details are described in earlier works [11,14,15]. Ammonia was measured using Nessler’s reagent [16]. For A. balhimycina, DCW was measured by filtering 1.5 ml of the fermentor broth through Whatman filter paper (Whatman, Brentford, Middlesex, UK) as reported elsewhere [11]. Micrococcus luteus was used as a test organism to measure antimicrobial activity of balhimycin [17]. For this purpose, agar test plates in the agar medium and filter with fermentor samples were prepared. Then the plates were incubated for two days at 30 °C. The growth inhibition diameter around the holes was measured and concentration of balhimycin was determined using pre-computed calibration curve.

The data for Amycolatopsis mediterranei S699 was taken from literature [9] and consisted of the following online measurements: pH, dissolved oxygen, agitator speed and CO2 and O2 concentration in exhaust gas.

3. Phase detection technique

3.1. Algorithm

In this work, the problem of monitoring of fermentation process has been posed as that of detection of singular points (SPs). We assume that the underlying characteristic dynamics and in turn the statistical properties of the online data vary from one phase to another. Thus, we propose that an SP can be detected by appropriately detecting the change in the statistical properties of the available online data as described below (Fig. 1). For the current phase $i$, a new data point $x_k$ ($x_k = [x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kp}]$ where $p$ is the number of variables being measured), the following hypothesis is checked:

$$H_0 : x_k \in \Phi_i$$
$$H_1 : x_k \notin \Phi_i$$

(1)
Let the data belonging to phase $f_i$ correspond to a probability distribution $P_i$. Then these hypotheses can be tested by constructing an appropriate test statistic depending on the nature of the distribution $P_i$. In this work, we assume that $P_i$ is a normal distribution, i.e., $P_i = N(\mu_i, \Sigma_i)$, where $\mu_i$ and $\Sigma_i$ are the mean and covariance matrix of the data corresponding to phase $f_i$ which can be approximated by sample average $\bar{x}_i$ and sample covariance $S_i$ calculated from the available data belonging to phase $f_i$ as

$$
\bar{x}_i = \frac{1}{n_i} \sum_{x_j \in f_i} x_j, \quad S_i = \frac{1}{n_i - 1} \sum_{x_j \in f_i} (x_j - \bar{x}_i)^T (x_j - \bar{x}_i)
$$

(2)

where $n_i$ are the number of data points belonging to phase $f_i$. In order to obtain reliable estimates of $\mu_i$ and $\Sigma_i$, the hypothesis testing is performed only after collecting data for a minimum window length ($W_{\text{min}}$), i.e., $n_i \geq W_{\text{min}}$. The relevant test statistic is then [18]:

$$
T_k^2 = (x_k - \bar{x}_i)S_i^{-1}(x_k - \bar{x}_i)^T
$$

(3)

which represents the Mahalanobis distance of the current point $x_k$ from the mean of the data corresponding to phase $f_i$. The null hypothesis is rejected when $T_k^2$ violates the upper or lower control limits, $T_{UCL}^2$ and $T_{LCL}^2$, respectively as:

$$
T_k^2 \leq T_{UCL}^2 \quad \text{or} \quad T_k^2 \geq T_{LCL}^2
$$

(4)

where

$$
T_{UCL}^2 = \frac{p(n_i - 1)(n_i + 1)}{n_i(n_i - p)} F(\alpha/2; p, n_i - p)
$$

(5)

$$
T_{LCL}^2 = \frac{p(n_i - 1)(n_i + 1)}{n_i(n_i - p)} F(1 - \alpha/2; p, n_i - p)
$$

(6)

where $\alpha$ is the significance level. For this study, $\alpha = 0.01$ has been used. When the null hypothesis is rejected, a phase change event is declared and the index $i$, which keeps track of the number of phases detected so far, is incremented by 1: $i = i + 1$. Data corresponding to the new phase is then collected afresh and the procedure continued. On the other hand, when the null hypothesis is accepted, the current point is appended to the sample available for phase $f_i$ and statistical properties of this phase are recomputed by Eq. (2) before testing the hypothesis (i.e., Eq. (1)) for next available measurement.

**Fig. 1.** Schematic representation of the proposed “Moving window-dynamic principal component analysis (MW-DPCA)” approach for singular point (SP) detection.
The proposed approach, if implemented directly, can suffer from following drawbacks: (a) high sensitivity to measurement and process noise: this can lead to high false alarm rate (detection of SP even if there has been no phase change in the process), (b) inability to capture dynamic relationships in the measured data, and (c) unnecessary computational overload if online data is very frequent since time constants of fermentation process may be much larger. To deal with these drawbacks, we incorporate the following modifications to the basic approach.

3.1.1. Incorporating robustness to noise

(i) A phase change event is declared only if the null hypothesis is rejected for at least \( n \) out of \( \xi \) consecutive data points. The parameters \( \eta \) and \( \xi \) can be tuned to achieve an acceptable tradeoff between false alarm rate and speed of phase change detection.

(ii) Principal component analysis (PCA): PCA involves projecting the measured data onto few orthogonal directions (referred to as loadings) and monitoring the projection of the data (scores) only on those directions. These orthogonal directions are the eigenvectors of the covariance matrix corresponding to its largest eigenvalues. The number of directions used depends on the fraction of variability of the data captured in those directions. Based on this number \( b \), \( T^2 \) and control limits used for monitoring are changed as [18]:

\[
T^2_k = (\bar{x}_k - \bar{x})P_b[\text{diag}(\lambda_b)]^{-1}P^T_b(\bar{x}_k - \bar{x})\]

(7)

where \( \text{diag}(\lambda_b) \) is the diagonal matrix of \( b \) largest eigenvalues of covariance matrix \( \Sigma \) and the columns of the matrix \( P_b \) are the corresponding \( b \) number of eigenvectors of \( \Sigma \). The corresponding expressions for upper and lower control limit are as [19]

\[
T^2_{ULCL} = \frac{b(n_b - 1)(n_b + 1)}{n_b(n_b - b)}F(\alpha/2; b, n_b - b)
\]

(8)

\[
T^2_{ULCL} = \frac{b(n_b - 1)(n_b + 1)}{n_b(n_b - b)}F(1 - \alpha/2; b, n_b - b)
\]

(9)

While the covariance matrix was utilized in the above discussion on PCA, use of the correlation matrix for PCA has also been reported [9]. We provide results based on both the techniques.

3.1.2. Incorporation of dynamic relationships among variables

The current online measurements may depend on the past online measurements. To capture such dynamic relationships, appropriately lagged data can be added to the current measurement [12]. Let the data at current time be related to data up to \( d \) samples in the past, where \( d \) is known as the lag, then the current data \( x_k \) is modified as: \( x_{k-d}^d = [x_k, x_{k-1}, \ldots, x_{k-d}] \). Accordingly, the current phase data matrix \( X_t \) is changed to \( X_t^d \) and the mean and covariance matrix of current phase, and the upper and lower control limits are also changed to reflect this modification. The parameter \( d \) needs to be tuned to obtain a balance between the predictive ability of the model, computational cost and speed of SP detection. To be consistent with the nomenclature used in literature [18], PCA when applied to the lagged data will be referred to as Dynamic PCA (DPCA).

3.1.3. Reducing the computational requirement

Typically the online data is measured at time scales (order of seconds) which are much faster than the time constants (typically order of hours) of fermentation processes. We consider data sampled at a lower frequency (sampling rate \( \tau \)) for detecting phase shifts. This sampling rate is then a tuning parameter which should be chosen to be consistent with the time constants of the fermentation process.

The overall approach with the above modifications is summarized in Fig. 1 and this technique will be referred to as “moving window dynamic PCA” (MW-DPCA). For the sake of comparison, in the results section, we have also considered SP detection without reducing the dimensionality of the data. For this purpose, Eqs. (3), (5) and (6) are used. This approach will be referred to as “moving window all dimensions” (MW-AD). The results are also compared with conventional PCA (with and without lag) approach where the entire data is assumed to correspond to a single mean vector and covariance matrix [9]. For such a scenario, the \( T^2 \) for every data point follows beta distribution since each data point is used to estimate the mean and covariance [19]. Then, the upper control limit (\( T^2_{ULCL} \)) and lower control limit (\( T^2_{ULCL} \)) are determined as:

\[
T^2_{ULCL} = \frac{(m - 1)^2}{m} B(\alpha/2; b/2, (m - b - 1)/2)
\]

(10)

\[
T^2_{ULCL} = \frac{(m - 1)^2}{m} B(1 - \alpha/2; b/2, (m - b - 1)/2)
\]

(11)

Similar to the modifications adopted for the strategies proposed in this article, the SP detection algorithm based on these conventional PCA techniques also utilizes the heuristic that for an SP to be declared, \( n \) out of \( \xi \) consecutive points should be out of control limits where data points are assumed to be available at sampling frequency \( \tau \).

3.2. Selection of model parameters

Implementation of the proposed algorithm in an effective manner requires the specification of the tuning parameters \( d, \tau, \xi, \) and \( \eta \). In general, the optimal values of the parameters will vary from one organism to another due to significant differences in their fermentation physiology. While searching through the parameter space, the following values for these parameters have been considered regardless of the organism: \( d: [0, 4, 8, 12, 16], \tau: [5, 10, 15, 20] \text{min} \), \( \xi: [4, 5, 6, 7, 8], \) and \( \eta: [2, 3, 4, 5, 6, 7, 8] \). We first construct the receiver operating characteristic (ROC) curve for all possible models (i.e. all combinations of parameters). ROC captures the trade off between the sensitivity and specificity for a binary classifier system [20,21]. The problem of SP detection can also be considered to be a binary classification problem where each data point needs to be classified as either a normal point (not an SP) or an SP as shown in Eq. (12).

Null hypothesis \( H_0 \) : Point is not SP

Alternative hypothesis \( H_1 \) : Point is SP

(12)

Four types of outcomes are possible while testing these competing hypotheses: (i) true negative: \( H_0 \) is actually true and it is not rejected by the model, (ii) true positive: \( H_1 \) is actually true and \( H_0 \) is rejected by the model, (iii) false negative: \( H_0 \) is actually false but it is not rejected by the model, and (iv) false positive: \( H_1 \) is actually true but is rejected by the model. For a given model and batch data, let the number of instances of each of the above outcomes be denoted by TN, TP, FN and FP respectively. For the given batch, knowing the true status of each time point (whether it is normal point or SP) from offline measurement data and/or expert knowledge, these numbers are then computed for each model (combination of parameters) under consideration.

All these models are then represented on the ROC curve which is a plot of sensitivity versus 1-specificity, where sensitivity and specificity are defined as:

Sensitivity = \( \frac{TP}{TP + FN} \)

Specificity = \( \frac{TN}{TN + FP} \)

In the ROC curve, models lying on the top left hand corner indicate the optimal trade off between high sensitivity and specificity.
Based on the specific requirements (related to sensitivity and specificity) of the user, any of the models which capture the best trade off, can be used. However, in absence of such requirements this choice is not straightforward and, single metrics which are a combination of specificity and sensitivity can be used to rank these models [22]. One such popular metric is Matthews Correlation Coefficient (MCC) which is defined as [23]

\[
MCC = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

The MCC takes values between $-1$ and $+1$. Value of $-1$ indicates complete mismatch between model predictions and true nature of every point (SP or not an SP) while value of $+1$ indicates complete agreement. For finding a common (across different batches), high-performing model for a given organism, models which have MCC values greater than $\%$ of maximum MCC (for that batch) and which do not miss more than some specified number $n_m$ of true SPs (for that batch) for all batches, are considered. Here $\%$ and $n_m$ are user specified parameters. If more than one model satisfies these criteria, any of these models can be selected.

Fig. 2. Detection of singular points (SPs) in batch fermentation of D-ribose producing, transketolase deficient strain of Bacillus pumilus ATCC 21951. A) Profiles of the online measurements; B) $T^2$ plot for the MW-DPCA approach with covariance matrix using model parameters as $d = 4$, $\tau = 5$ min, $\xi = 6$ and $\eta = 2$; and D) Profiles for the offline measurement of concentrations of the substrates, products and dry cell weight. The measured SPs ($MSP$s) are $q$: end of lag phase and start of exponential growth phase, $r$: ribose production start, $s$: end of exponential growth phase and start of stationary phase, $u$: end of stationary phase and $v$: glucose exhaustion. Initial media composition for the batch: glucose 200 g l$^{-1}$, corn steep liquor (CSL) 12 g l$^{-1}$, ammonium sulfate 5 g l$^{-1}$, CaCO$_3$ 16 g l$^{-1}$, MnSO$_4$ 0.5 g l$^{-1}$, leucine 0.5 g l$^{-1}$ and tryptophan 0.05 g l$^{-1}$. Gray filled circle symbols indicate the $MSP$, dark filled circle symbols (i.e., $q'$, $r'$, $u'$ and $v'$) denote the model predicted SPs ($PSP$s) which match with $MSP$ and open circle symbols (i.e., $w$, $z$, $y$ and $t$) denote the PSPs which do not match with $MSP$. In B, the arrows are indicative of the presence of points which go out of the range of the plot. In this particular batch, acetate and acetoin were not detected.
4. Results

We present the results for the proposed moving window based techniques: (i) MW-DPCA using covariance; (ii) MW-DPCA using correlation; and (iii) MW-AD. The results are also compared with conventional approaches which use the data for the entire batch duration, namely, (iv) DPCA using covariance; (v) DPCA using correlation; and (vi) using all dimensions (AD). For the PCA based approaches, three PCs were used. For all the approaches, $W_{\text{min}}$ is chosen to be 30. A brief discussion about the rationale behind this choice is presented in Appendix A. Efficacy of these six approaches is compared for SP detection for case studies involving three different microorganisms. The predicted SPs (PSP) were compared with SPs identified manually (MSP) based on the offline measurements. Note that due to the low frequency of the offline measurements, the occurrence of true SPs may differ from the identified MSP by some tolerance $\delta$. Hence, the following strategy was used:

- If $P_{\text{SP}} \in [M_{\text{SP}} \pm \delta]$ then $P_{\text{SP}}$ is considered as a true SP.
- If $P_{\text{SP}} \notin [M_{\text{SP}} \pm \delta]$ then $P_{\text{SP}}$ is considered as a false SP.

In this work, the tolerance $\delta$ was chosen to be 2.5 h.

4.1. Case study 1: Bacillus pumilus

The tkt deficient strains of bacillus are reported to be commercially important for the production of D-ribose. Two batches I and II with different initial conditions were conducted, where the following online variables were recorded: pH, DO, agitator speed and concentrations of CO$_2$ and O$_2$ in the exhaust gas. Batch II was conducted in triplicate (labeled as IIa, IIb and IIc). As explained in the section on model parameter selection, for a given approach, MCC values and the number of missed SPs for all combinations of model parameters were calculated for these batches. The SP detection results for batch I with $t = 50$ and $n_m = 2$ are presented in Fig. 2. Other satisfactory models had only minor variations in model parameters such as $\eta$ and $\xi$ and are therefore not presented. The MSPs shown in this figure were identified based on the physiological characteristics as captured by the offline measurements shown in Fig. 2D. Fig. 2A shows the raw online data.
while Fig. 2B shows the T2 values for the MW-DPCA approach along with the corresponding time varying values for T2LCL and T2UCL. By monitoring the control limit violation of the T2 value, the proposed MW-DPCA with covariance approach has detected four of the five MSPs. Two additional SPs are also detected. These may be false alarms or true SPs that are not captured as MSPs due to the low frequency of offline measurements. This type of prediction accuracy would be difficult to achieve simply by visual inspection of the online data which shows sharp changes at several time points. For example, the plot of agitator speed has several visible troughs and peaks. However, not all these sharp changes correspond to SPs. Results for MW-DPCA with correlation and MW-AD approaches are not presented since these were inferior to those of MW-DPCA with covariance for this microorganism. For the sake of comparison, results for conventional single model techniques viz. DPCA with covariance, DPCA with correlation and AD were also generated. Fig. 2C shows the T2 values along with the corresponding T2UCL and T2LCL for DPCA with covariance, which was found to be the best among these three conventional approaches. The static nature of DPCA is reflected in the time invariant nature of the T2 control limits. It is seen from Fig. 2B and C that MW-DPCA performs better than DPCA.

Offline data and results for Batch II are shown in Fig. 3. From the offline data it can be seen that there is significant variability in these batches. In particular, D-ribose, acetate and acetoin production in batch IIc appears to be delayed compared to the other two batches. This variability is reflected in different times of occurrence of MSPs corresponding to similar events such as the start of death phase. The MW-DPCA model applied to batch I is able to capture the batch-to-batch variability in terms of the PSpS in batch II as well. Moreover, the results are superior to those of conventional DPCA (data not shown).

4.2. Case study 2: Amycolatopsis balhimycina

Balhimycin (a glycopeptide antibiotic) producer strain of A. balhimycina was cultivated in media containing multiple carbon and nitrogen substrates including complex sources such as yeast extract and defatted soybean flour. The substitutable substrates may be taken up sequentially or simultaneously, thereby complicating the task of SP detection from the raw online data. Seven batches (labeled I-VII) were conducted. Results for batches I-IV, with $t = 50$ and $n_m = 2$, and common model parameter values across batches for a given approach, are shown in Figs. 4 and 5. Due to space limitations, results for batches V-VII are presented as supplementary material. Batch I is characterized by three MSPs based on the offline data (Fig. 4B). The three moving window based approaches have successfully identified these SPs whereas the

![Fig. 4. Comparison of different monitoring techniques for batch fermentation of balhimycin producing strain of Amycolatopsis balhimycina, batch I. (A) The MSPs (gray filled circle symbols) and PSpS (dark filled circle symbols when the time of PSpS matches with that of MSP and open circle symbols otherwise) by different approaches: (i) MW-DPCA using covariance matrix, $d = 0$, $t = 5$ min, $\xi = 8$ and $\eta = 5$, (ii) MW-DPCA using correlation matrix, $d = 4$, $t = 10$ min, $\xi = 7$ and $\eta = 2$, (iii) MW-AD, $d = 0$, $t = 5$ min, $\xi = 5$ and $\eta = 3$, (iv) DPCA using covariance matrix, $d = 8$, $t = 5$ min, $\xi = 4$ and $\eta = 2$, (v) DPCA using correlation matrix, $d = 8$, $t = 5$ min, $\xi = 4$ and $\eta = 2$, (vi) using all dimensions (AD), $d = 8$, $t = 5$ min, $\xi = 8$ and $\eta = 3$ and (vii) MSPs. (B) Profiles for the offline measurement of concentrations of the substrates, products and dry cell weight. (C) Profiles of the online measurements. The MSPs are: q: start of antibiotic production, r: glycerol exhaustion, s: glucose consumption stops and stationary phase starts. Media composition for batch I: Glucose 100 g l$^{-1}$, glycerol 10 g l$^{-1}$, yeast extract 1 g l$^{-1}$, ammonium sulfate 4.95 g l$^{-1}$, defatted soybean flour 1 g l$^{-1}$, and micronutrients ZnSO4 0.02 g l$^{-1}$, FeSO4 0.02 g l$^{-1}$, trisodium citrate 0.025 g l$^{-1}$, MgSO4 1.5 g l$^{-1}$, MnSO4 0.01 g l$^{-1}$, NaCl 1 g l$^{-1}$, KH2PO4 0.16 g l$^{-1}$, biotin 0.00005 g l$^{-1}$, calcium-pantothenate 0.001 g l$^{-1}$, nicotinic acid 0.001 g l$^{-1}$, myo-inositol 0.025 g l$^{-1}$, thiamin HCL 0.001 g l$^{-1}$, pyridoxine HCL 0.001 g l$^{-1}$, para-aminobenzoic acid 0.0002 g l$^{-1}$ and MES buffer 1.045 g l$^{-1}$.](image-url)
three single model conventional approaches have missed one or more of these SPs (Fig. 4A). For example, strategy (v) has missed all MSPs. It is interesting to note that between 36 and 48 h, most of the techniques have predicted two additional SPs, which may correspond to the diauxic nature of the growth as the organism begins to utilize glucose in this interval. Identification of such events, which fall in the interval between two offline sampling times, is possible only based on online data. The online data for this batch, in particular DO and agitator speed, is more noisy compared to that for B. pumilus (Fig. 2A). This precludes manual identification of SPs by visual inspection of online data for this batch. However, the automated MW techniques have successfully identified the SPs.

The MW-DPCA results for batches II, III and IV along with the corresponding offline data are presented in Fig. 5A–C respectively. For batch II, all the MSPs are captured by the MW-DPCA approach. Some of the additional PSpS detected in this batch may correspond to phase change events not identified by the offline measurements. For batch III, three of the four MSPs are correctly identified while the MSP u does not match any PSp. However, it should be noted that the time of occurrence of u may be inaccurate due to lack of offline measurements between 108 to 120 h. It is interesting to note that there is a PSp at ~112 h, which may indicate the actual time of exhaustion of glucose (the event corresponding to MSP u). For batch IV, four of the five MSPs are captured by MW-DPCA. The missed MSP r corresponds to the end of first exponential phase. However, due to lack of offline measurements, this event could have actually occurred anytime in the 24–36 h interval. Once again, it is interesting to note that there is a PSp in this interval, which may correspond to the actual time of occurrence of this event.

4.3. Case study 3: Amycolatopsis mediterranei S699

Data for two batches I and II for this case study has been taken from Doan et al. [9]. The results for batch I for MW-DPCA with t = 75, n_m = 2 and DPCA with t = 50, n_m = 1 along with online and offline data, and the T2 plots are shown in Fig. 6. For DPCA there was no model parameter combination which met the t = 75 and n_m = 2 criteria for both batches and hence lower t value had to be used to obtain a common model. MW-DPCA approach has successfully predicted the three MSPs in batch I while DPCA
approach has predicted only one \( M_{SP} \). From the online data (Fig. 6A), it is seen that there are no sharp changes corresponding to \( M_{SP} q \) and \( r \) and this may be the reason for the failure of DPCA approach in capturing these events. However, the use of multiple models in MW-DPCA enables successful prediction of these events.

Batch II contains two alternate nitrogen substrates namely ammonia and nitrate leading to sequential uptake. MW-DPCA successfully predicted three out of four \( M_{SPs} \) whereas the conventional DPCA is able to predict only one (Fig. 7A). Note that in the neighbourhood of \( M_{SPs} q \) and \( r \), there are no noticeable sharp changes in the online profiles. Despite this, MW-DPCA has captured \( M_{SP} r \) and has predicted an SP approximately 8 h prior to \( M_{SP} q \).

5. Discussion

In this work we have presented a moving window based approach for the detection of SPs in fermentation processes. The approach has the following salient features: (i) The method does not need to assume that a single statistical model is applicable for the entire batch duration, (ii) the switching times from one statistical model to the next are not decided a priori and are instead decided in real time based on the dynamic evolution of the batch under consideration, (iii) similarly, the \( T^2 \) control limits are not fixed a priori but are decided in real time based on the amount of data available in the corresponding phase, and (iv) the approach can be used even in the absence of historical data. To demonstrate
the efficacy of our approach, we have presented a comparison with the conventional single model based approach. The results were found to be superior to the conventional single model approach even though the latter utilize the data for the entire batch duration. In contrast, our approach utilizes only the currently available data in an evolving batch in real time. This feature makes our approach amenable for real time implementation which is not possible with the conventional single model approach.

Note that the proposed SP detection is based on violation of either the upper or the lower control limit by the $T^2$ statistic (Eqs. (5) and (6)). In contrast, monitoring techniques have conventionally relied only on violation of $T^2_{UCL}$ alone [24,25]. Note that while changes in $\mu$ mainly lead to violation of $T^2_{UCL}$, changes in $\Sigma$ can manifest as violations of either $T^2_{UCL}$ or $T^2_{LCL}$ [19]. Since a phase change can correspond to either a change in the mean (operating level of the variables) or the covariance (relationships between variables), we chose to use both $T^2_{UCL}$ and $T^2_{LCL}$ to detect SPs. Indeed, we have observed several cases where violation of $T^2_{LCL}$ detects the SP (data not shown). Additionally, to provide insight into the nature of the SP, we perform statistical tests to check if the means and the covariances of adjacent phases are identical. Details about these statistical tests are presented in Appendix B. Based on these tests; it was observed that several of the phase shift points which were detected due to $T^2_{LCL}$ violation corresponded to changes only in the covariance and not in the mean (data not shown). This indicates the utility of using $T^2_{LCL}$ apart from $T^2_{UCL}$ as a bound on $T^2$ to improve SP detection.

From the point of view of sensor selection, it would be of interest to determine the utility of the various online measurements in SP detection. To this end, the contribution of various online measurements in SP detection was quantified for the case studies presented in this work (see Appendix C). Fig. 8 shows the contribution plots for SP prediction by MW-DPCA for batches in case study I. Note that the contributions of agitator speed and DO were more significant than those of other online variables in majority of the batches. This is consistent with the fact that the rate of aerobic growth dictates the oxygen requirement which in turn determines the DO and agitator speed as seen from Eq. (14).

$$\frac{dC_0}{dt} = k_l a(C^+ - C_0) \quad \frac{\mu X_{BM}}{Y_{B/O}}$$  (14)
At pseudosteady state condition $dC_0/dt$ becomes zero and Eq. (15) holds:

$$k_L a \propto f(\mu X_{BM})$$  \hspace{2cm} (15)

But since

$$k_L a \propto h(N)$$  \hspace{2cm} (16)

we get,

$$N \propto g(\mu X_{BM})$$  \hspace{2cm} (17)

The proposed SP detection technique can in principle be used in conjunction with online data collected via a variety of sensors [26–29]. These sensors may range from simple probes such as those for pH, dissolved oxygen concentration, and optical density to complex probes which can acquire near infrared (NIR) or fluorescence spectroscopy based measurements [30,31]. Some of the complex probes are more informative but may suffer from drawbacks such as limited measurement range and high cost. We believe that the proposed approach can be used to evaluate the utility of a given sensor in SP detection. Additionally, the proposed approach can be used as a guide in real time in making decisions about the timings of offline sampling. In particular, an offline sample can be collected whenever an SP based on online data is detected instead of collecting offline samples based on arbitrarily specified timings.

**Acknowledgements**

The authors acknowledge the generous gift of the Amycolatopsis balhimycina strain from Anna Eliasson Lantz of Denmark’s Technical University, Denmark. The work was partially supported by a grant from the Department of Biotechnology, Government of India.

**Appendix A. Selection of $W_{\text{min}}$**

The case studies considered in this article involved either four ($A. \text{ Balhimycina DSM5908}$) or five ($B. \text{ pumilus ATCC 21951 and A. mediterranei S699}$) online measurements. For these case studies, $W_{\text{min}}$ was taken to be 30 since it was found (Fig. A1) that there was not much variation in the $T_{LCL}$ and $T_{UCL}$ values with respect to $W_{\text{min}}$ beyond 30 for $p = 4$ and 5.

**Appendix B. Covariance and mean comparison of adjacent phases**

The key idea in our proposed algorithm is that occurrence of an SP corresponds to change in statistical properties of online data. Under the assumption of normally distributed data, this change in statistical properties will be reflected as differences in the mean vectors and/or covariance matrices of adjacent data sets (before and after detection of an SP). The data in the two adjacent...
populations is considered to be normal with means $\mu_1, \mu_2$ and covariances $\Sigma_1, \Sigma_2$ respectively. To identify changes in these parameters, the following statistical tests were used:

The modified likelihood test [32] was used for testing whether the covariance matrices of adjacent populations are identical. In particular, the following hypotheses were tested:

null hypothesis : $H_0 : \Sigma_1 = \Sigma_2$

alternative hypothesis : $H_1 : \Sigma_1 \neq \Sigma_2$

An approximate test of $H_0$ at significance level $\alpha$ based on the modified likelihood ratio statistic is to reject $H_0$ if $-2 \log L > c_\alpha(1 - \alpha)$ where $c_\alpha(1 - \alpha)$ denotes the percentage point from the $\chi^2$ distribution such that area to the left is $1 - \alpha$, $\rho = 1 - (2p^2 + 3p - 1)/(6(p + 1)m)/\sum_{k=1}^{p^2} 1/k_1 - 1$, $f = p(p + 1)/2$ is degrees of freedom and $\Lambda^* = \prod_{i=1}^{n} (\det S_1)^{(n_i - 1)/2}/(\det S_2)^{(n - 2)/2}$ is the modified likelihood ratio.

In these expressions, $S_i$ is the $i$th population sample covariance matrix, $S = \sum_{i=1}^{p^2} S_i$, $n_i$ is the size of the $i$th population sample, $n = \sum_{i=1}^{p^2} n_i$, $k_i = (n_i - 1)(n - 2)$. For checking whether the mean vectors of adjacent populations are identical, the following hypothesis was tested [33].

null hypothesis : $H_0 : \mu_1 = \mu_2$

alternative hypothesis : $H_1 : \mu_1 \neq \mu_2$

The null hypothesis was rejected if the test statistic $T_2^p > T_2^c$ [assuming $n_1 \leq n_2$, where $T_2^p = n_1(x_1 - x_2)'C^{-1}(x_1 - x_2)$]

$\bar{x}_1 = \sum_{i=1}^{n_1} x_{1i}/n_1$, $x_{1i}$ is the $i$th data point of 1st population.

$\bar{x}_2 = \sum_{i=1}^{n_2} x_{2i}/n_2$, $x_{2i}$ is the $i$th data point of 2nd population

$C = \sum_{i=1}^{n_1} (u_{1i} - \bar{u})(u_{1i} - \bar{u})'$

$u_{1i} = x_{1i} - (n_1/n_2)^{1/2}x_{2i}$, $\gamma = 1, 2 \ldots n_1$

$\bar{u} = \sum_{i=1}^{n_1} u_{1i}/n_1$, $T_2^c = n_1p/n_1 - p + 1 F(1 - \alpha; p,n_1 - p + 1)$

For both covariance and mean checking, the value of $\alpha$ was taken as 0.01.

Appendix C. Contribution of variables towards SP detection

When an SP is detected, the variables primarily responsible for occurrence of SP can be identified based on the contribution plot of the variables. The procedure is as follows [18]:

When an SP is detected at the $i$th time point corresponding to observation $x_k$, then $T_2^k > T_{2,LCL}$ or $T_2^k < T_{2,UCL}$. The normalized score $t_i/\lambda_i$ are then computed for the $i$th principal component ($i = 1, 2, \ldots, b$) where $t_i$ is the score of the projection of $x_k$ to the $i$th loading vector. The principal components for which $(t_i^2/\lambda_i) > (1/b)T_{2,UCL}$ (in case $T_2^k > T_{2,UCL}$) or $(t_i^2/\lambda_i) < (1/b)T_{2,LCL}$ (in case $T_2^k < T_{2,LCL}$) are determined to be responsible for the out of control status. Let the number of such principal components be $r$. Then the contribution of each variable $j$ to the out of control score $c_i$ can be defined as $cont_{ij} = (t_i/\lambda_i)(p_{ij}/(K_j - \mu_j))$, where $p_{ij}$ is the $(i,j)$th element of the loading matrix $P$. If $cont_{ij}$ is negative, it is set equal to zero. The total contribution of the $j$th process variables is then:

$CONT_j = \sum_{i=1}^{r} cont_{ij}$ The variables with large values of $CONT$ are identified as primary causes for phase change detection. This information can potentially aid the process operator in determining the nature of phase change as well as in taking any control action if required.

Appendix D. Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>number of eigen values to be considered for PCA</td>
</tr>
<tr>
<td>$B(\alpha/2; a, b)$</td>
<td>beta-distribution with $\alpha$% significant level and $a$ and $b$ degree of freedom</td>
</tr>
<tr>
<td>$C_s$</td>
<td>saturation concentration of oxygen</td>
</tr>
<tr>
<td>$C_{st}$</td>
<td>dissolved oxygen concentration in medium</td>
</tr>
<tr>
<td>$d$</td>
<td>dynamic lag</td>
</tr>
<tr>
<td>$\text{diag}(\lambda_b)$</td>
<td>the diagonal matrix of $b$ largest eigenvalues of covariance matrix $S$</td>
</tr>
<tr>
<td>$\text{DPCA}$</td>
<td>dynamic principal component analysis</td>
</tr>
<tr>
<td>$F(\alpha/2; a, b)$</td>
<td>percentage point from F-distribution with $a$ and $b$ degrees of freedom such that the area to the left is $\alpha/2$</td>
</tr>
<tr>
<td>$k_s$</td>
<td>volumetric oxygen mass transfer coefficient</td>
</tr>
<tr>
<td>$m$</td>
<td>total number of data points for the entire batch</td>
</tr>
<tr>
<td>$M_{SP}$</td>
<td>SP based on offline measurements</td>
</tr>
<tr>
<td>$\text{MW-AD}$</td>
<td>moving window all dimensions</td>
</tr>
<tr>
<td>$\text{MW-DPCA}$</td>
<td>moving window dynamic PCA</td>
</tr>
<tr>
<td>$n_i$</td>
<td>the number of data points belonging to phase $\phi_i$</td>
</tr>
<tr>
<td>$n_m$</td>
<td>number of missed SPs during prediction</td>
</tr>
<tr>
<td>$N$</td>
<td>agitator speed</td>
</tr>
<tr>
<td>$p$</td>
<td>number of variables being measured</td>
</tr>
<tr>
<td>$P_i$</td>
<td>probability distribution corresponding to $i$th phase</td>
</tr>
<tr>
<td>$P_b$</td>
<td>matrix of $b$ eigenvectors of $S$, corresponding to the largest $b$ eigenvalues</td>
</tr>
<tr>
<td>$P_{SP}$</td>
<td>predicted SPs</td>
</tr>
<tr>
<td>$S_1$</td>
<td>sample covariance matrix of $i$th phase</td>
</tr>
<tr>
<td>$\text{SP}$</td>
<td>singular point</td>
</tr>
<tr>
<td>$\tau$</td>
<td>percent of MCC considered to find the common model</td>
</tr>
<tr>
<td>$T_{2,UCL}$</td>
<td>Mahalanobis distance of the current point $x_k$ from the mean of the data corresponding to phase $\phi_i$</td>
</tr>
<tr>
<td>$T_{2,LCL}$</td>
<td>lower control limit of $T^2$</td>
</tr>
<tr>
<td>$T_2$</td>
<td>upper control limit of $T^2$</td>
</tr>
<tr>
<td>$W_{\text{min}}$</td>
<td>minimum window length for calculating covariance matrix</td>
</tr>
<tr>
<td>$X_k$</td>
<td>data (row) vector at $k$th time</td>
</tr>
<tr>
<td>$X_{pk}$</td>
<td>value of $p$th variable at $k$th time</td>
</tr>
<tr>
<td>$X_{d1}^k$</td>
<td>data (row) vector with lag $d$ at $k$th time</td>
</tr>
<tr>
<td>$X_{d1}$</td>
<td>data matrix of $i$th phase with lag $d$</td>
</tr>
<tr>
<td>$X_{BM}$</td>
<td>biomass concentration</td>
</tr>
<tr>
<td>$V_{B/O}$</td>
<td>yield of biomass per unit of oxygen consumed</td>
</tr>
</tbody>
</table>

Greek letters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>the significance level</td>
</tr>
<tr>
<td>$\phi_i$</td>
<td>$i$th phase</td>
</tr>
<tr>
<td>$\eta$</td>
<td>number of points required to violate the control limits for SP detection</td>
</tr>
<tr>
<td>$\mu$</td>
<td>specific growth rate</td>
</tr>
<tr>
<td>$\tau$</td>
<td>sampling rate</td>
</tr>
<tr>
<td>$\xi$</td>
<td>number of consecutive points checked for violation of control limits for SP detection</td>
</tr>
</tbody>
</table>
δ tolerance for comparing $M_{SP}$ and $P_{SP}$

$\mu_i$ population mean of ith phase

$\Sigma_i$ population covariance matrix of ith phase

Appendix E. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.procbio.2009.03.008.

References


