Process Biochemistry 51 (2016) 1919–1929

Contents lists available at ScienceDirect

Process Biochemistry

journal homepage: www.elsevier.com/locate/procbio

Bioprocess statistical control: Identification stage based on hierarchical clustering

Marco V. Cedeño, Leandro P.F. Rodríguez Aguilar, Mabel C. Sánchez*

Planta Piloto de Ingeniería Química (Universidad Nacional del Sur-CONICET), Camino La Carrindanga km 7, 8000, Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form 26 July 2016
Accepted 16 August 2016
Available online 17 August 2016

Keywords:
Multivariate process control
Hotelling statistic
Fault identification
Clusters
Fermentation

ABSTRACT

Bioprocesses are characterized by the fact that small variations in operating conditions may have a substantial impact on the final batch quality. Therefore, the early detection and isolation of faults allow implementing corrective actions before the effects of deviations from the normal operation have a detrimental effect on production. In this work a new strategy for the statistical monitoring of batch processes is presented, and it is applied to monitor the operation of a fermentation process. The methodology works in the original variable space, therefore it only uses the Hotelling statistic for detection purposes. To determine the set of measurements by which the fault is revealed, the nearest in control neighbor to the observation point is calculated, and the distance between these two points is used to evaluate the contribution of each observation to the inflated statistic. In contrast to the existing latent-variable and original-variable based approaches, a simple hierarchical clustering technique allows to identify the set of suspicious measurements, without assuming the probability density function of the variable contributions. Furthermore, the performance of the proposed identification procedure is compared to the one achieved using other monitoring techniques. A well-known fed-batch fermentation benchmark is employed with this purpose, and the comparison is based on the results of a comprehensive set of simulated fault scenarios.

1. Introduction

On-line monitoring of bioprocesses is essential for the safe and profitable production of high-quality products. Slow changes are common in biosystems, and may gradually grow until they turn into a serious operational problem. Therefore the early detection and isolation of faults allow implementing corrective actions before the effects of deviations from the normal operation have a detrimental effect on production.

Bioprocesses are usually batch oriented activities. Frequently, only basic operating signals (pH, acid or base addition, stirrer speed, temperature, pressure, liquid level, feed rate, etc.) can be measured on-line, and it is difficult (or even impossible) to build a mechanistic process model. Given the complexity of biological systems and the lack of quantitative analysis about the underlying mechanisms, multivariate statistical techniques have offered effective tools for bioprocess supervision [1,2].

Methods based on projecting the original variables (OVs) into latent structures are commonly used for monitoring batch units. Recently, Ge et al. [3] examined the state-of-the art of those data driven techniques. The authors categorized the monitoring strategies that have been proposed up to date into three groups: multiway, phase-based and two-dimensional dynamic monitoring methods.

At first, the traditional multiway techniques rearrange the aligned three dimensional data array collected from a batch process into a convenient number of two-way matrices. Then they extract the information contained in the correlated data performing linear transformations onto low dimensional spaces. These are defined by a small number of uncorrelated latent variables (LVs). The performance of different multiway methods has been addressed in several works. In this sense, Ramaker et al. [4] analyzed the fault detection properties of global, local and time evolving models for Principal Component Analysis (PCA). Furthermore Camacho et al. [5] discussed the same issue for techniques based on batch-wise and variable-wise unfoldings, local models and batch-dynamic ones with 1 lagged measurement- vector.

Phase based methods allow to reveal the multiphase/multistage characteristics of many batch processes. Different types of strategies have been proposed to take into account the unique characteristics of each phase (Yao and Gao [6]). Multiblock PCA methods separate the measurements associated to different oper-

* Corresponding author.
E-mail address: msanchez@plapiqui.edu.ar (M.C. Sánchez).

http://dx.doi.org/10.1016/j.procbio.2016.08.020
1359-5113/

ating stages/phases into meaningful blocks, and consider both the variable correlations within each block and among blocks. The Adaptive Hierarchical PCA is a particular multiblock technique that forms as many blocks as time intervals, and controls the amount of weight given to the new information with respect to the recent history. Other approach uses a clustering technique to divide the process into modeling phases taking into account the changes in variables correlation structure, and models each phase separately. Extensions of the sub-PCA modeling method to monitor processes with limited referenced data, uneven-length batches and transitions are also provided. The Multiphase PCA strategy identifies the phases by an iterative procedure that recognizes portions of the batch operation which can be represented by a linear PCA model.

To take into account the dynamic characteristics that may also exist from batch to batch, Lu et al. [7] developed a two-dimensional dynamic PCA method. Later on, Yao and Gao [8] extended the technique to multiphase processes.

Furthermore batch processes are inherently nonlinear. To deal with this kind of behavior, Lee et al. [9] proposed a multiway technique based on Kernel PCA (KPCA), which has the ability to capture nonlinearities. The KPCA maps the OVs to a higher dimensional feature space where PCA is performed. The linear monitoring approaches mentioned above can be extended to their kernel counterparts.

Monitoring strategies based on PCA declare the faulty state when The Hotelling Statistic (D), calculated using the retained PCs, or The Square Prediction Error (SPE), which monitors the residual space, or both of them are greater than their critical values. A combination of D and SPE is also used for fault detection purposes (Alcala and Qin [10]).

When the faulty state is declared, the identification of the measurements that reveal the fault is performed. These are the observations that contribute more to the inflated statistic value. For linear PCA, Westerhuis et al. [11] extended the use of contribution plots to LV models with correlated scores, and introduced control limits for the variable contributions (VCs). These thresholds allow automatizing the identification of the faulty variables. Later on, Alcala and Qin [10] determined the set of suspicious variables using the concept of the Reconstruction-Based Contribution (RBC), and also defined the control limit for the fault-free RBC. Alcala and Qin [12] extended their previous work to address the monitoring of non-linear processes using KPCA. For this problem, they could not state the control limit for the RBC, and proposed to perform the identification task by the visual inspection of the contribution plots. In contrast, Godoy et al. [13] defined the VC in terms of the partial-derivative of the normalized statistic with respect to the observation, and arbitrarily set the control limit of the VC to the value of one.

Other strategies for the statistical monitoring of batch processes work in the original measurement space. They make only use of The Hotelling's Statistic, defined in terms of all the OVs (T^2), to determine if the batch is out of control. The performance of the T^2 chart for observations taken from a common multivariate normal distribution, or a series of multivariate normal distributions with different mean vectors but identical covariance matrices, was evaluated by Mason and Young [14]. Later on, Alvarez et al. [15] developed the Original Space Strategy for Batch Processes, OSSBP, for monitoring batches such that the mean observation vector and the covariance matrix change with time during the run. Local models were used to take into account those variations in a simple way (Ramaker et al. [4]). The VCs to the statistic were calculated using the unique decomposition of T^2 proposed by Alvarez et al. [16]. Furthermore the VC control limit was evaluated using the reference population data. The strategy has been successfully applied for the statistical monitoring of batch processes when the number of variables involved is not extremely high and there exist strong non-linear relationships among them that prevent the measurements from being linear combinations. It should be noticed that these features are shared by many bioprocesses. In these cases the covariance matrix is not singular and the process can be monitored using only one statistic.

All the aforementioned LV- and OV-based strategies calculate the contribution of one observation to the inflated Hotelling statistic considering that the values of the remaining variables are equal to their measurements. Therefore, there is a sole parametric curve defining all the possible values of the statistic as function of that particular variable. Cedeño et al. [17] did not impose that restriction on the identification. The authors proposed to calculate initially the coordinates of the nearest point to the observation that is located on the boundary of the normal operating region. That point is called the Nearest In-Control Neighbour (NICN). Then the relative influence of each measurement on the inflated T^2 value is evaluated in terms of the distance between the observation and the NICN points. Those variables whose distance measures exceed a certain threshold are considered suspicious. A data driven technique is presented to determine those threshold values. The application of this procedure for batch process monitoring is not straightforward because a set of faults should be simulated for each time interval. This increases the computational burden and relies on the availability of a process simulation code.

In this work, the statistical monitoring of batch processes for fault diagnosis is addressed by means of a simple strategy that works in the original measurement space. Therefore the detection task is performed evaluating only the T^2. To identify the measurements that signal the fault, it is proposed a methodology based on the NICN concept and the application of hierarchical-clustering. In contrast to existing techniques, the new identification method avoids making assumptions about the probability density function of the VCs. The performance of the identification technique is compared with respect to the behavior of other LVs- and OVs-based methods using a well-known fed-batch penicillin fermentation benchmark (Birol et al. [18]). To obtain robust conclusions, the results of a large amount of simulated faults are used to calculate a set of identification performance indexes.

The rest of this paper is organized as follows. In Section 2, the new strategy for batch process monitoring and a short description of the fed-batch fermentation benchmark are presented. Section 3 comprises an analysis of the results obtained when different approaches are applied to identify the observations that reveal a comprehensive set of simulated faults. Conclusions are summarized at the end of the article.

2. Material and methods

2.1. Batch process monitoring: a strategy based on original variables

The statistical monitoring of batch processes comprises two stages. The first one is performed offline, and makes use of the information contained in the Normal Operating Condition Data Set (NOC). This stage provides an empirical model of the process and threshold values, which will be used then to test the progress of new batches. The second stage is accomplished during the batch run. At each time interval, measurement values are analyzed to determine whether the process remains in control. If the faulty state is declared, then the identification of the observations that reveal the fault follows.

Time evolution of a batch manufacturing process can be registered measuring the same J variables during K time intervals. Each of them is represented by an index k, such that $k = 1$ when the batch begins and $k = K$ when the operation finishes. Hence the information

of the I successful batch runs contained in the NOC can be grouped into a three way data matrix $\mathbf{Z}$ (batch × variables × time). At first $\mathbf{Z}$ is centered and scaled to obtain the matrix $\mathbf{X}$ ($I \times J \times K$). Standardized data is used to study the variation of the variable trajectory around the mean trajectory, and to handle differences in the measurement units. Because slab and tube scaling have similar performances for batch process monitoring using latent projection-based techniques [19], in this work tube scaling is used for the sake of simplicity.

If the correlation matrix of the data contained in the NOC is not singular, the operation of the process can be monitored in the OV space, i.e., no transformation of the data into a LV space is required (Mason and Young [14]; Alvarez et al. [16]). Let us assume that the covariance matrix $\mathbf{R}_k$ states the relationship among the variables at the k-th time interval. It is calculated using the observations contained in $\mathbf{X}_k$ ($I \times J$), the k-th time slice of $\mathbf{X}$, as follows

$$\mathbf{R}_k = \frac{\mathbf{X}_k^T \mathbf{X}_k}{(I-1)} \tag{1}$$

This matrix is used to estimate the T^2 critical value for the k-th time interval, $T^2_{C,k}$. In this sense, at first the T^2 of the k-th observation is evaluated for each batch contained in the NOC as follows

$$T^2_{i,k} = \mathbf{x}^T_{i,k} \mathbf{R}_k^{-1} \mathbf{x}_{i,k} \qquad (i = 1 \ldots I) \tag{2}$$

where $\mathbf{x}_{i,k}$ ($J \times 1$) is the standardized measurement vector. Then, all the $T^2_{i,k}$ values are employed to estimate the $T^2_{C,k}$, which is the (1-α) quantile of the T^2_k probability density function. The Kernel Density Estimation (KDE) approach [20,21] is applied with this purpose to avoid the errors incurred by assuming a specific probability distribution for T^2_k.

The matrix $\mathbf{R}_k$ and the value of $T^2_{C,k}$ constitute the local process model for the k-th time instant. They are stored to be used during the on-line monitoring [4]. A graphical scheme of the tasks performed off-line is presented in Fig. 1.

From Eq. (2), it is evident that all $\mathbf{R}_k$ matrices should be non-singular to be inverted. If two or more measured variables are near correlated, the calculation of T^2 provides inaccurate results. For identifying collinearities, it is recommended to calculate the condition indexes, which are defined as the square root of the ratio of the maximum eigenvalue to each of the others eigenvalues. If a condition index is greater than 30, the presence of a severe collinearity is pointed out. Since it is not advisable to use the statistic value in this case, several alternatives are suggested to overcome the problem. The simplest one is to eliminate one of the variables involved in the collinearity. To determine the set of collinear variables, the linear combination of variables provided by the eigenvector associated to the smallest eigenvalue is examined. After ignoring the variables with small coefficients, the linear relationship between those that are producing the collinearity problem is obtained. Another method for removing collinearities is to re-construct the inverse of the correlation matrix by excluding the eigenvectors related to the near zero eigenvalues. This technique should be applied with caution because the ability to identify shifts in some directions in terms of the full set of the original variables may be lost (Mason and Young [14]). Furthermore, there exist different types of matrix regularization techniques in the literature that can be applied to correct for the instability of the R_k^{-1} before calculating T^2.

When the process is operating, the elements of the measurement vector obtained at the k-th time interval, $\mathbf{zn}_k$, are centered and scaled as follows

$$xn_{j,k} = \frac{(zn_{j,k} - \bar{z}_{j,k})}{\delta_{j,k}} \tag{3}$$

where $zn_{j,k}$ is the measured value of the j-th variable at the k-th observation and $xn_{j,k}$ stands for its standardized measurement value, $\bar{z}_{j,k}$ and $\delta_{j,k}$ are the mean and standard deviation of the j-th column contained in the k-th vertical slice of $\mathbf{Z}$. After that, T^2 for the new measurement is calculated as:

$$T^2_{n,k} = \mathbf{xn}_k^{\mathrm{T}} \mathbf{R}_k^{-1} \mathbf{xn}_k \tag{4}$$

and compared with the corresponding critical value, $T^2_{C,k}$. If $T^2_{n,k} > T^2_{C,k}$, the faulty state is declared; then VCs to the inflated T^2 are calculated and analyzed to point out the measurements that signal the out of control state.

To estimate the VCs, the NICN [17] (nearest neighbor of the observation point which is in statistical control) is determined. This information allows estimating how far the faulty observation is from an in control location, and which directions explain more about the occurrence of the anomalous situation.

The NICN coordinates are calculated minimizing the Mahalanobis distance between the observation point and another one which is in statistical control, that is, the T^2 value for the NICN point is equal to $T^2_{C,k}$. The NICN is the solution of the following optimization problem

$$\begin{aligned} &Min \quad {}^{\mathbf{T}}\mathbf{R}_k^{-1}(\mathbf{x}nic\mathbf{n}_k - \mathbf{xn}_k) \\ &s.t. \\ &\qquad \mathbf{x}nic\mathbf{n}_k^{\mathrm{T}} \mathbf{R}_k^{-1} \mathbf{x}nic\mathbf{n}_k \;=\; T^2_{C,k} \end{aligned} \tag{5}$$

where $\mathbf{x}nic\mathbf{n}_k$ is the NICN vector at the k-th time interval. It should be noticed that the NICN is not an observation, but the closest point to $\mathbf{xn}_k$ located on the J dimensional ellipsoid of probability (1-α). There exist two local minima that satisfy the necessary conditions of optimality for Problem (5). They are represented by the following expression:

$$\mathbf{x}nic\mathbf{n}_k = \pm \left(\frac{T^2_{C,k}}{T^2_{n,k}} \right)^{1/2} \mathbf{xn}_k \tag{6}$$

where $T^2_{n,k}$ is the T^2 for the observation vector $\mathbf{xn}_k$. The objective function values of both solutions are compared to decide which one is the NICN. This is much easier than evaluating the second order optimality conditions.

After calculating the coordinates of the NICN, the influence of each variable to the inflated statistic value is estimated. Since all the variables have been previously standardized to be dimensionless, the distance in which each measured variable should be modified to reach the NICN is considered as an estimation of the contribution of that variable to the T^2. Thus the j-th VC for the k-th observation is defined as follows

$$cn_{j,k}^{T^2_{n,k}} = |xn_{j,k} - xnicn_{j,k}| \tag{7}$$

The information contained in the contribution vector, $\mathbf{cn}^{T^2_{n,k}}$, is necessary but not enough to automatically determine the set of measurements that reveal the fault. In this regard, control limits for the VCs are required. These cannot be estimated using the NOC, as it is usually done by other methodologies, because the NOC is associated with successful batches and the VC definition (Eq. (7)) is related to a fault state. Also the procedure used to estimate the control limits for continuous processes, developed by Cedeño et al. [17], is not recommended for the batch ones because it requires many fault simulations. To overcome these difficulties, it is proposed to concentrate the VCs to the $T^2_{n,k}$ in two regions according with their similarities using a clustering technique, and to assume that the contributions of the measurements that signal the abnormal situation are in one of the clusters. This assumption is based on the fact that the $cn_j^{T^2_{n,k}}$ ($j = 1 \ldots J$) indicate how apart the process behavior is with respect to its normal operation, therefore the larger

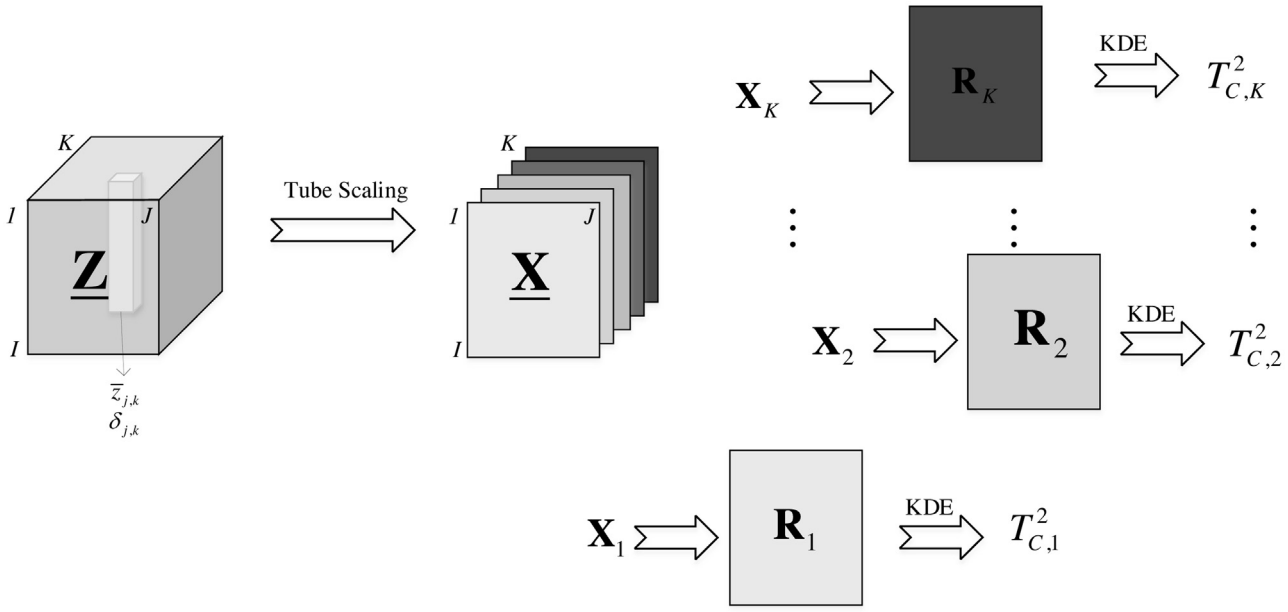

Fig. 1. Graphical scheme for the offline stage.

contributions are associated to the measured variables that signal the abnormal behavior.

Clustering procedures have been applied on many research areas as biology, botany, medicine, psychology, geography, image processing, etc. [22]. Eventhough clustering have being applied for fault detection [23], their use for the automatic analysis of VCs has not been reported until now.

In this work a simple hierarchical clustering procedure is applied. In contrast to existing methods, it does not rely on assumptions about the probability density function of the VCs. At first, the technique calculates the Euclidean distance between pairs of contributions (objects), and stores this information into a dissimilarity vector of length $J(J-1)/2$. This information is used to determine the proximity among objects. As objects are paired into binary clusters, the newly formed clusters are grouped into larger ones until a hierarchical tree takes shape. If two lines of the associated dendrogram are intersected by a horizontal line, the objects below the left-hand line belong to one cluster while the objects below the right-hand line correspond to the other cluster. For example, let us consider that for a certain $\mathbf{x}$ of dimension (13×1), $\mathbf{cn}^{T^2_{n,k}} = [0.638\ 0.003\ 0.003\ 0.053\ 0.028\ 0.125\ 0.133\ 0.010\ 0.160\ 0.045\ 0.106\ 0.279\ 0.004]$. The dendrogram related to those values is shown in Fig. 2, where it can be seen that one cluster includes the contributions associated with the 2nd–13th observations while the other group only contains the contribution corresponding to the first measurement. Therefore this is consider the suspicious variable. The dendrogram is only presented for illustrative purposes. The results of the identification task can be obtained without visualizing this diagram. Furthermore, no assumptions are imposed to separate the faulty variables from the other ones.

In this work the performance of the proposed technique, called Nearest in Control Neighbour for Batch Processes (NICNBP), to deal with the identification stage of bioprocess monitoring is analyzed. That performance is also compared with the corresponding ones to other existing procedures. Because different types of kernel methods have become popular for the monitoring of batch processes in recent years, two identification strategies employed in kernel techniques are included in the comparison analysis. These were proposed by Alcala et al. [12] and Godoy et al. [13], and they are denominated KPCA-RBC and KPCA-CN, respectively, in this work. A brief description about the application of KPCA for fault detection and identification is presented in the Appendix. Also the identification method used by OSSBP [15] is included in the comparative performance study.

To establish a fair comparison basis, it is a common practice that all the techniques involve in the comparison provide similar values for the average number of Type I errors (AVTI) when no faults are present. In this work the approach proposed by Ramaker and Van Sprang [4] is used to calculate the AVTI that is defined as

$$AVTI = \frac{\sum_{i=1}^{I} \rho_i}{I} \times 100 \tag{8}$$

where ρ_i is equal to one if a batch contained in the NOC gives an out-of-control signal and zero otherwise.

The comparison of different identification techniques is carried out using a large set of simulated faults. To test the behavior of the m-th technique ($m = 1 \ldots 4$) when it detects the p-th simulated fault, the following procedure is performed:

1) The monitoring technique is run until the value of its statistical test exceeds the critical one for three consecutive time intervals. This time instant is represented by Td (detection interval). Then the VCs to the statistic are calculated and used to determine the Set of Suspicious Variables (SSV_{mp}). For NICNBP, this task is carried out using the clustering technique previously explained. Regarding OSSBP, the measurements whose contributions exceed their control limits are considered as suspicious variables. Those threshold values are calculated using the normal operation data and the three sigma rule [15]. The methodology KPCA-CN automatically incorporates an observation to the SSV_{mp} if its normalized contribution is greater than one [13]. For KPCA-RBC, VCs control limits are not defined in the literature, therefore to implement an automatic identification procedure those are evaluated experimentally using the RBCs for the normal process operation and the three sigma rule [11]. This has been a common industrial practice for process monitoring. Also, in this work it is especially supported by the fact that the NOC is built using many batch runs.
2) Variable trajectories are used to determine the Set of Faulty Variables (SFV_{mp}) at Td. A variable is included in SFV_{mp} if its measured

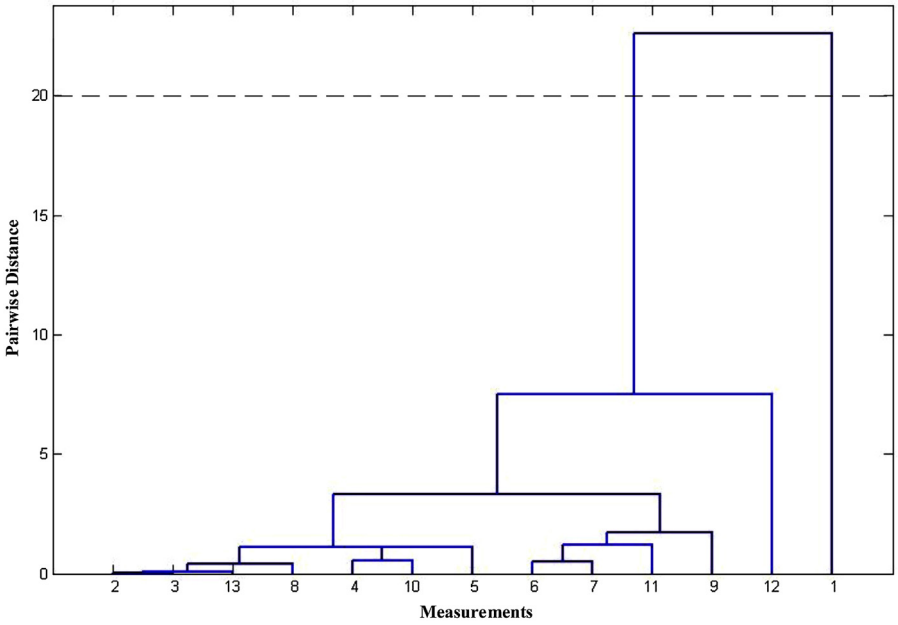

Fig. 2. Dendrogram for the identification example.

value is out of its nominal operating range at the detection time. This is the 99% confidence range of the empirical distribution of the measurement in the NOC at *Td*.

3) The test is categorized as a perfect, ambiguous, incorrect or void identification, taking into account the following definitions:

Precise identification (PI): SSV_{mp} is equal to SFV_{mp}.

Ambiguous identification (AI): at least one of the observations contained in SSV_{mp} belongs to SFV_{mp}.

Incorrect identification (II): no variable contained in SSV_{mp} belongs to SFV_{mp}.

Void identification (VI): SSV_{mp} is empty.

For the *m*-th technique, the total number of detected faults, that is, the number of identification tests, is quantified and used to calculate the percentages of PIs, AIs, IIs and VIs (%PI, %AI, %II and%VI) for that strategy. The procedure outlined before avoids the identification performance measures being influenced by the particular features of the detection stage.

2.2. Case study

In this work the performance of the proposed identification technique is evaluated using simulated faults of the penicillin production process described by Birol et al. [18]. The reaction takes place in the cultivation reactor shown in Fig. 3. Because the formation of the antibiotic is usually not associated with the cell growth, in practice cells grow in a batch culture followed by a fed-batch operation that promotes the synthesis of the product. In this sense, a small amount of biomass and substrate are added at first, and the reactor operates in batch mode during 45 h. The substrate feed is started when most of the initially added substrate has been consumed by the microorganisms. The fed-batch mode lasts 355 h.

Nine measurements are taken into consideration for the first stage whereas 13 process observations are used to describe the second one. Measurement information is provided in Table 1. The temperature of the culture mass is automatically controlled by means of a proportional-integral-derivative controller; instead the pH is adjusted using an on-off control strategy. The addition of substrate during the second stage is performed in an open-loop.

Table 1
Measurements Description.

Tag	Variable description	Comments
1	Aeration rate	Stages 1 y 2/Input variable
2	Agitator power input	Stages 1 y 2/Input variable
3	Substrate feed rate	Stage 2/Input variable
4	Substrate feed temperature	Stage 2/Input variable
5	Substrate concentration	Stages 1 y 2
6	Dissolved oxygen concentration	Stages 1 y 2
7	Biomass concentration	Stages 1 y 2
8	Penicillin concentration	Stage 2
9	Volume	Stages 1 y 2
10	Carbon dioxide	Stages 1 y 2
11	pH	Stages 1 y 2
12	Temperature	Stages 1 y 2
13	Cooling water flow rate	Stage 2/Input variable

To carry out the performance analysis, the NOC is built using the results of 181 batch simulations. They are performed by running the simulation code PenSim v.2.0 (http://simulator.iit.edu/web/pensim/simul.html) repeatedly under normal operating conditions with small random variations. Also the noise of measurements is simulated. The sample interval and the batch run time are set to the values of 1 h and 400 h, respectively, therefore $\mathbf{Z} \in \mathbb{R}^{181\times13\times400}$. In Fig. 4, the nominal trajectories of the variables are shown.

Then 444 extra failures [15] are simulated taking into account different fault types (step S; ramp R; initial condition IC), magnitudes, directions, occurrence times and durations. Variables' trajectories for two faults, selected as examples, are displayed in Figs. 5 and 6. One of them is related to the occurrence of a step change in the agitator power input (magnitude: −10% of the variable nominal value at 100 h; duration: 100 h). As it can be seen in Fig. 5, that change affects the dissolved oxygen concentration. The second fault corresponds to a ramp change in the substrate feed rate (maximum magnitude: −5% of the variable nominal value at 245 h, duration: 100 h). Fig. 6 shows that all the output variable values deviate from the nominal ones.

The proposed failures are divided into eight groups, which are summarized in Table 2. Regarding the first two fault sets, the simulation procedure is run by setting the IC of a variable to a value different from the one recommended in the batch recipe. The

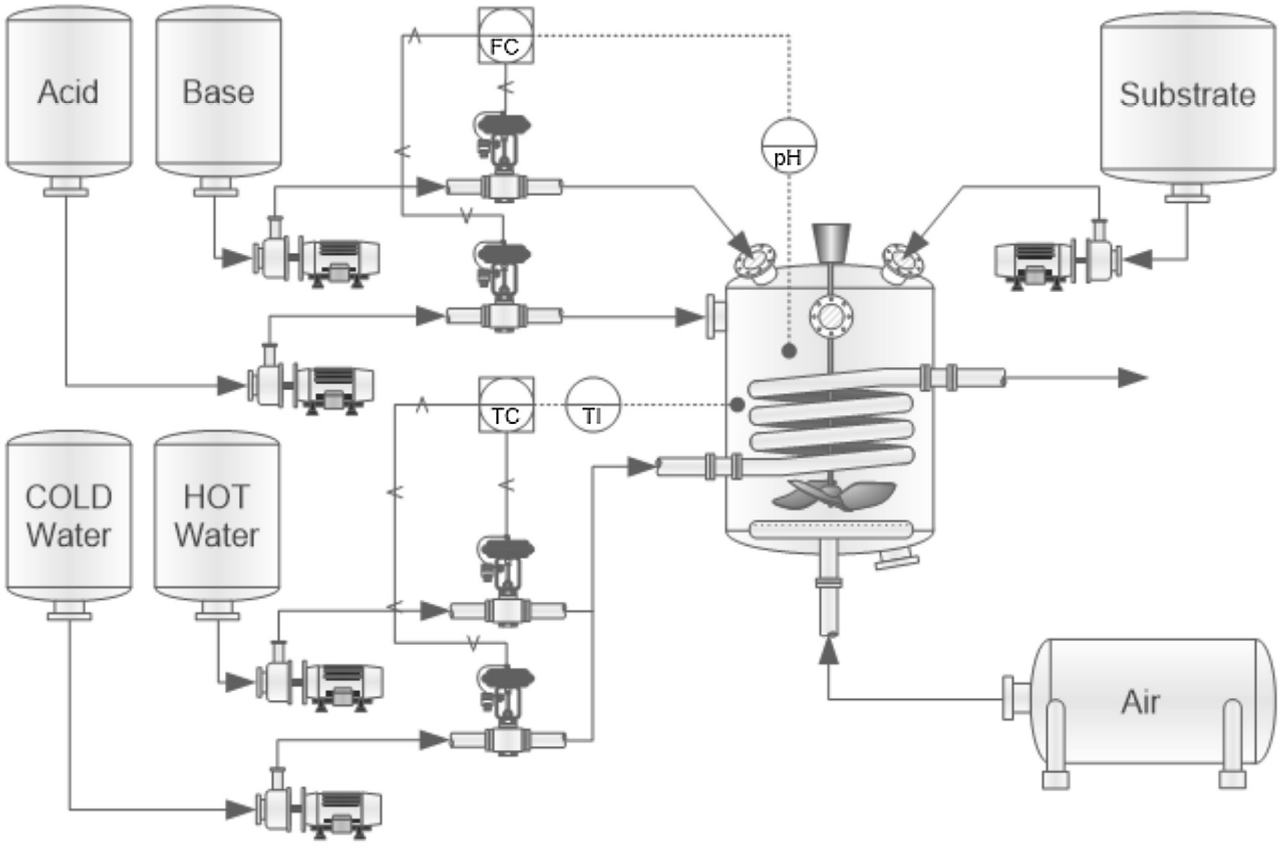

Fig. 3. Fed-batch penicillin cultivation process flowsheet.

Table 2
Faults description.

Fault Set	Variable Tag	Fault Type	Fault Magnitude (% design value)	Starting Time (h)	Duration (h)	Number of faults
1	5, 7, 11	IC	±5, ±10, ±15	0	–	18
2	12	IC	±0.5°, ±2°, ±5°	0	–	6
3	1	S	±5, ±10, ±20	0, 100, 200, 300	10, 50, 100	72
4	2	S	±5, ±10, ±20	0, 100, 200, 300	10, 50, 100	72
5	3	S	±5, ±10, ±20	45, 145, 245, 345	10, 50, 100	66
6	1	R	±5, ±10, ±20	0, 100, 200, 300	10, 50, 100	72
7	2	R	±5, ±10, ±20	0, 100, 200, 300	10, 50, 100	72
8	3	R	±5, ±10, ±20	45, 145, 245, 345	10, 50, 100	66

variables selected for this study are the substrate and biomass concentrations, and the pH and temperature of the reactive medium, which are changed one at a time. All these variables affect the biomass growth. Experimental findings suggest that the microorganisms' growth strongly depends on the carbon source and oxygen as substrates, and it is inhibited by high amounts of biomass itself in penicillin fermentation. Moreover, environmental variables, such as pH and temperature, also play an important role on the quality and quantity of the final product. If the pH of the culture medium decreases, a reduction in the total cell mass concentration results, which is associated to a decrease in the penicillin concentration. Regarding the temperature, biomass growth shows an increasing tendency with the increment of this variable up to a certain value, and a rapid decrease is observed beyond that value [18].

With respect to the remaining fault sets, step and ramp faults are simulated in the aeration rate, substrate feed rate and agitator power input, one at a time. The first two variables affect the supplement of oxygen and glucose, which are essential for biomass growth. The agitator power input has an effect on the overall mass transfer coefficient of the reactive medium. Many works in the area of fault diagnosis for fed-batch fermentation have considered deviations in the nominal values of those three variables [18,24,25].

3. Results and discussion

In this section the application results of the aforementioned strategies for the fault identification of the fermentation process are presented and analyzed.

The same type of model is used to describe the relationships among the process variables. In this regard, both kernel methods and OV-methodologies use K local models [4,15]. For KPCA, a second order polynomial kernel is used because it captures the nonlinearity of this system better than other functions [9]. The number of retained principal components is selected in such a way that the 80% of the sample total variance is reconstructed.

Twenty four samples are used to analyze changes in the ICs with respect to those defined in the batch recipe. In general, the smallest changes in substrate and biomass concentrations, temperature and pH cannot be detected or their detection is delayed. In particular, Fault Set 1 represents 18 fault events in substrate and biomass

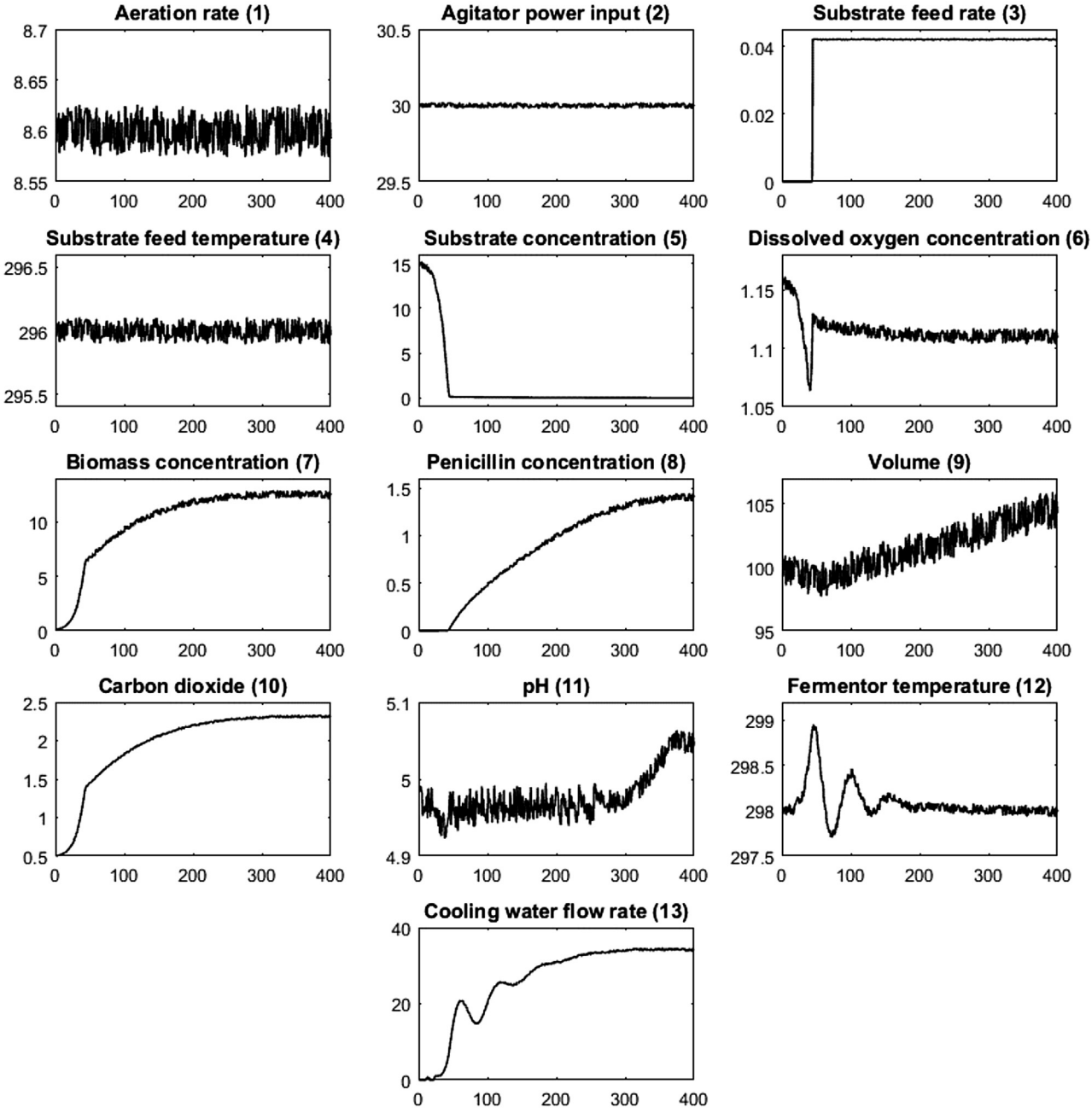

Fig. 4. Nominal trajectories of the process variables.

concentrations and pH in the range [±5%–±15%]. The OV-based strategies only detect 11 faults whereas LV-based methodologies recognize 15 failures. In general, the SFV_{mp} only contains the variable whose value has been modified. The numbers of PIs, AIs and IIs are [6, 4, 1], [8, 2, 1], [6, 9, 0] and [6, 9, 0] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. Results show that kernel methods are more suitable to detect and identify changes in biomass concentration. In contrast, they have problems to correctly identify changes in the substrate concentration for medium- and large-magnitude faults. The same tendency is observed for pH changes. In these tests, SSV_{mp} contains almost all the variables when kernel methods are applied. The identification performance of NICNBP is the best one, because this strategy gives 4 PIs out of a total of 6 pH faults, while OSSBP provides 2 perfect tests.

Fault Set 2 comprises changes in the initial fermentor temperature in the range [±0.5 °C–±5 °C]. No technique detects the small magnitude faults. In general, the remaining failures are noticed immediately after they occur, and SFV_{mp} only contains that variable. The numbers of PIs, AIs and IIs are [3, 1, 0], [4, 0, 0], [3, 1, 0] and [2, 2, 0] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. It can be seen that NICNBP behaves better than the other strategies for this type of fault. The analysis of the results indicates that AIs are obtained for the highest temperature deviations.

Fault Set 3 and Fault Set 4 are related to S changes in the aeration rate and the agitator power input, respectively. There is a little correlation between the aeration rate and the rest of the variables. Therefore a disturbance in that variable does not strongly affect the measured values of the remaining ones, at least for the variation range considered in this work. In consequence, a fault in the aeration rate mostly affects its own measured value. The same conclusion arises for the agitator power input.

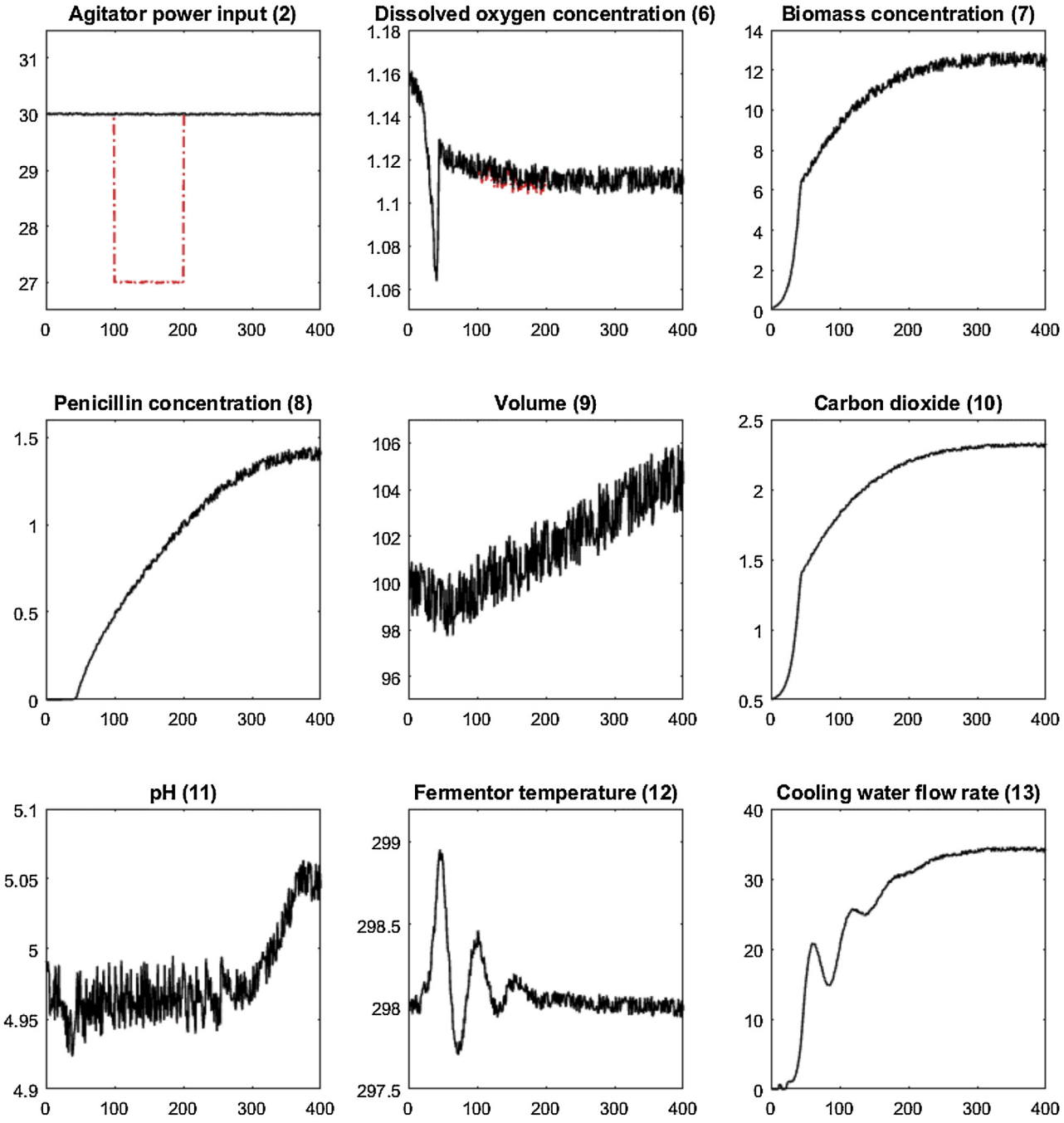

Fig. 5. Trajectories of the process variables for a step change in the agitator power input (Normal operation: full black line, faulty state: dash-dotted red line (-.)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Regarding Fault Set 3, the Ss in the aeration rate are detected immediately after they occur. The OV-based strategies detect 71 failures out of a total of 72 simulated ones. The numbers of PIs, AIs and IIs are [59, 12, 0], [67, 4, 0], [32, 38, 2] and [17, 53, 2] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. For this group, the performance of the NICNBP is highlighted. In addition to the aeration rate, kernel methods repeatedly signal other variables as suspicious. Some of them are output variables (dissolved oxygen concentration, biomass concentration, pH) which have remained within their normal operating range at the detection time, and others are input variables whose values have not been modified.

With respect to Fault Set 4, small S changes in the agitator power input are hardly detected by the OV-strategies. They recognize 48 failures out of a total of 72 simulated ones, while LV-strategies detect 66 faults of the same set. The numbers of PIs, AIs and IIs are [42, 6, 0], [46, 2, 0], [34, 32, 0] and [20, 46, 0] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. Once the faults are detected, OV-based strategies identify many of them correctly and provide AIs for a small number of large-magnitude failures. Many AIs for medium- and large-magnitude faults are obtained using kernel methods. Besides the agitator power input, they point out variables which are within their normal operation range.

In general, the detected small-magnitude changes are perfectly identified by OSSBP and KPCA-based strategies for the previous two fault sets. But, if the fault magnitude increases, it is noticed a false smearing of the fault. Those methodologies calculate the contribu-

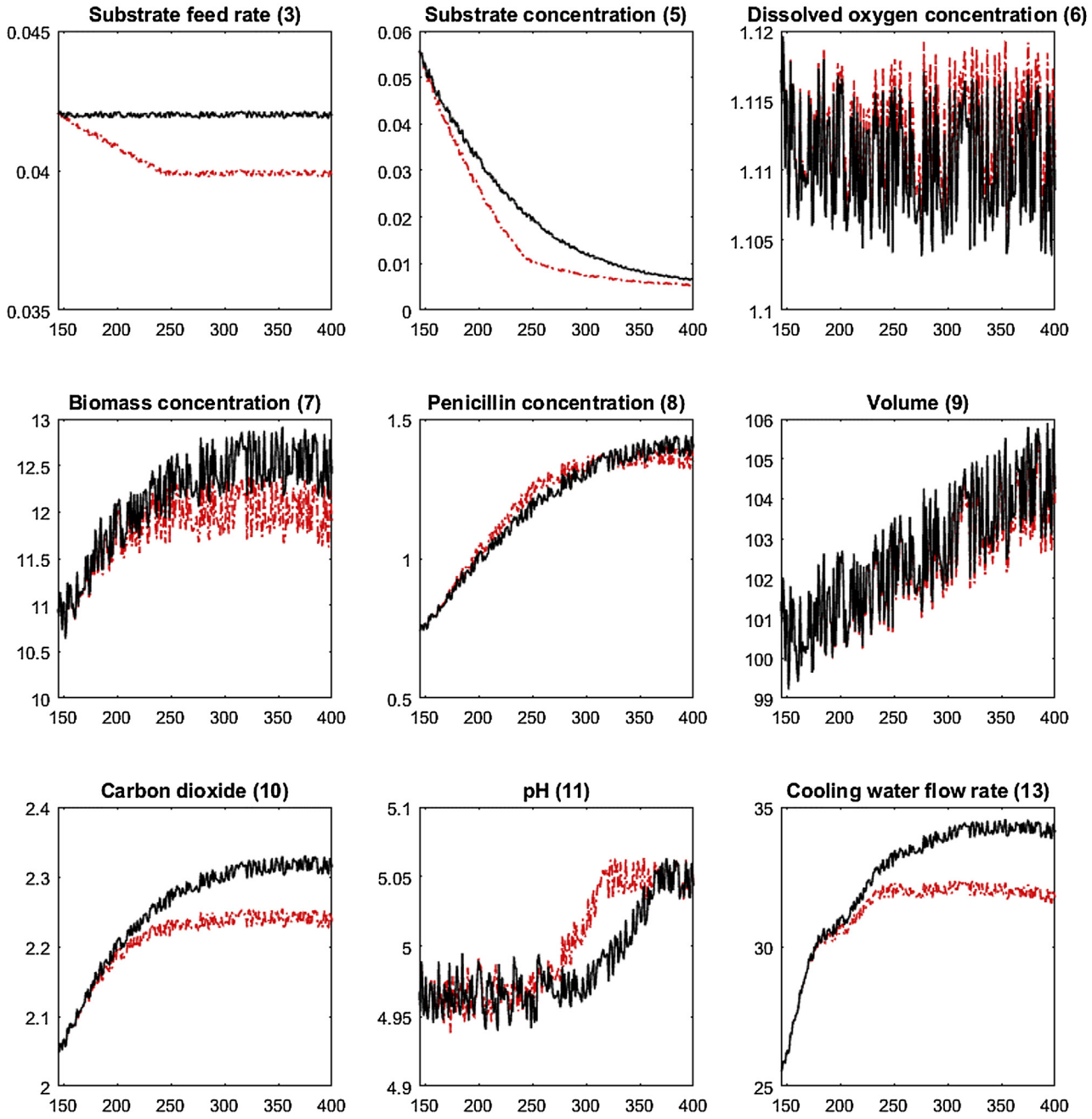

Fig. 6. Trajectories of the process variables for a ramp fault in the substrate feed rate (Normal operation: full black line, faulty state: dash-dotted red line (-.)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tion of each variable using the measured values of the remaining ones. Therefore high VCs are obtained for measurements which are in their normal operating range because they are calculated as function of the simulated fault magnitude. In these cases, the fault is pointed out by the variable in which the failure has been simulated, and also by other ones. The KPCA-CN indicates a large set of suspicious variables in comparison with that provided by KPCA-RBC because it uses an arbitrary control limit for the VCs.

Fault Set 5 represents S changes in the substrate feed rate. The OV-based strategies detect 64 faults out of a total of 66 whereas LV-based methodologies recognize 65 failures. In general, failures are detected immediately after they occur. The SFV_{mp} for small-magnitude faults mainly contains the substrate feed rate. But the SFV_{mp} for the other failures frequently includes at least two variables because the substrate feed rate is strongly correlated with the substrate concentration at the detection time. The numbers of PIs, AIs and IIs are [21, 40, 3], [29, 32, 3], [7, 57, 1] and [10, 54, 1] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. A reduction in the number of PIs is observed with respect to the one obtained for Fault Sets 3 and 4. The strategies give AIs for the three fault magnitudes, but especially those outcomes arise for the large ones. The *SSVs* obtained by kernel methods often contain variables, such as the dissolved oxygen concentration, the carbon dioxide content, the cooling water flowrate, the aeration rate, etc., which are within their normal operating ranges. That is, the effect of the false smearing is observed. In contrast, the AIs for NICNBP arise when both the substrate feed rate and the substrate composition are out of their normal operating range at the fault detection time, but the strategy signals only one of these variables.

Table 3
Performance indexes for fault identification.

	Fault Type	OSSBP	NICNBP	KPCA-RBC	KPCA-CN
%PI	Global	67,8	79,3	41,8	36,9
	S	66,7	77,6	36,0	23,2
	R	69,6	81,0	47,1	50,0
	IC	60,0	80,0	45,0	40,0
%AI	Global	28,0	16,5	55,0	59,9
	S	31,7	20,8	62,6	75,4
	R	23,9	12,5	48,1	45,2
	IC	33,3	13,3	50,0	55,0
%II	Global	4,2	4,2	3,2	3,2
	S	1,6	1,6	1,5	1,5
	R	6,5	6,5	4,8	4,8
	IC	6,7	6,7	5,0	5,0
%VI	Global	0,0	0,0	0,0	0,0
	S	0,0	0,0	0,0	0,0
	R	0,0	0,0	0,0	0,0
	IC	0,0	0,0	0,0	0,0

Next, R changes are simulated in the aeration rate (Fault Set 6), agitator power input (Fault Set 7) and substrate feed rate (Fault Set 8). Regarding Fault Set 6, the OV-based strategies detect 70 faults out of a total of 72 whereas LV-based methodologies recognize 71 failures. The numbers of PIs, AIs and IIs are [57, 13, 0], [63, 7, 0], [46, 23, 2] and [38, 31, 2] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. With respect to Fault Set 7, the detection performance of kernel techniques is highlighted because they detect 71 faults out of a total of 72, while OV-methods only recognize 49 failures of the same set. The numbers of PIs, AIs and IIs are [43, 6, 0], [46, 3, 0], [34, 34, 3] and [41, 27, 3] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. It is observed that the best number of PIs is obtained using NICNBP, and the identification performance of KPCA-CN increases for Fault Set 7.

From the analysis of the results of Set Faults 3, 4, 6 and 7, it can be noticed that OV-techniques detect changes in the aeration rate better than in the agitator power input. Furthermore kernel methods outperform OV-strategies for detecting faults in the last mentioned variable. The largest number of PIs is obtained using NICNBP for both variables and fault types. Also, if the final fault magnitude of the R is equal to the S magnitude, kernel techniques identify correctly more R faults than S failures in the aforementioned variables. This happens because R changes are slower than the S ones. Therefore, at the R detection time its fault magnitude is lower than the corresponding one to the S fault, and the false smearing effect is reduced.

Finally, the Step Fault 8 represents ramp changes in the substrate feed rate. The OV-based strategies detect 65 faults out of a total of 66 whereas kernel methods recognize all the simulated faults. The numbers of PIs, AIs and IIs are [28, 25, 12], [40, 13, 12], [18, 43, 5] and [25, 36, 5] for OSSBP, NICNBP, KPCA-RBC and KPCA-CN, respectively. For all the strategies, the number of PIs increases with respect to the one obtained for the fifth fault set. The *SFV*s mainly contain only one variable because the fault magnitudes are lower than the corresponding ones to Ss simulations at the detection time. This also originates a reduction in the false smearing of kernel methods and OSSBP, and a better identification performance for NICNBP.

Table 3 shows a summary of the identification performance of the techniques, i.e., the results obtained when each technique detects a fault and the identification procedure is run. The performance indexes are calculated considering both each fault type and all the simulated faults as a whole. For this last case the performance measures are reported as global indexes. These highlight the NICNBP behavior. The highest %PI value as well as the lowest%AI one are obtained using the proposed strategy.

Regarding the identification of the different fault types, it can be seen that the best %PI and %AI values are also obtained using NICNBP. Furthermore it is noticed that all the strategies achieve the best %PI for simulated Rs. These values are similar to the ones obtained for simulated Ss when NICNBP and OSSBP are used. In contrast, the %PI achieved by kernel methods for simulated Rs is significantly better than the same index for Ss because the fault magnitudes are lower at the detection time. This reduces the false smearing effect.

For the selected case study and set of simulated faults, an II arises when SFV_{mp} is null. That is, all the measurement values are within their normal operating ranges at the detection time, even though some values are near the bounds of the interval. Incorrect identifications are obtained for 14 and 16 identification tests when kernel methods and OV-based strategies are applied, respectively. Incorrect identifications are related with faults which do not have a clear effect on the process variables. Because the detection is performed using the same procedure for OSSBP and NICNBP, the %IIs of both methods are equal. The same happens for kernel strategies. Furthermore no methodology gives VIs for all the analyzed fault cases.

4. Conclusion

A new strategy for the statistical monitoring of batch processes is presented and applied to follow the operation of a fed-batch penicillin process. The methodology works in the original variable space, therefore it only uses the T^2 for detection purposes. To determine the set of observations that signal the fault, the NICN to the observation point is calculated, and the distance between these two points is used to evaluate the contribution of each variable to the inflated statistic. Then contributions are divided into two groups using a simple hierarchical clustering technique that allows isolating the suspicious measurements into one of the clusters. In comparison with other methods, the proposed identification procedure is very simple. It can be used for any process since no parameters should be set. This avoids the errors associated to the assumption of the VCs' probability density functions.

Given that NICNBP works in the original variable space, the inversion of the correlation matrix is required to calculate T^2 statistic. If the numbers of variables involved in the process is not extremely high and the non-linear relationships among them prevent the measurements from being linear combinations, no application problems related to the inversion of the correlation matrix appear. These conditions arise in the case study presented in this work and in several other batch processes presented in the literature.

In contrast to other works, a broad comparative performance study is presented in this contribution. The identification performance measures used to evaluate the behavior of different methodologies are calculated using a huge amount of simulated cases, which comprise very diverse fault scenarios. Furthermore the empirical control limits of the VCs for the strategies OSSBP and KPCA-RBC are estimated employing a NOC, which includes many samples. For KPCA-RBC, those limits have not been reported previously; therefore they provide a sensible approximation to run the identification procedure completely on line.

In comparison with KPCA-based strategies and OSSBP, the results show that NICNBP provides the best% PI and%AI indexes for the penicillin fermentation process. This methodology allows to estimate how far the faulty observation is from an in control location, and which directions explain more about the occurrence of the anomalous situation. The VC of a measurement is not affected by the influence of other measurements.

In future works an extension of the proposed identification scheme to other multivariate monitoring strategies will be analysed.

Acknowledgments

The authors wish to thank the financial support of CONICET (National Research Council of Argentina), and UNS (Universidad Nacional del Sur, Bahía Blanca, Argentina).

Appendix A. Kernel Principal Component Analysis

Let $\mathbf{x}$ be a sample vector of J measurements, which is mapped into a high dimensional space, called the feature space, via a mapping function $\phi = \Phi(\mathbf{x})$. In that space the inner product of two vectors ϕ_i and ϕ_n is defined using a kernel function k as follows:

$$\phi_i^T \phi_n = k(\mathbf{x}_i, \mathbf{x}_n)$$

Also the representation of the training data in the feature space is denoted as:

$$X = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_I]^T$$

The goal of the KPCA is to eigen-decomposed the covariance matrix of a set of I mapped samples ϕ_i in order to obtain the scores and loadings of the PCA model. That covariance matrix, $\mathbf{S}$, is defined in the next equation

$$(I-1)\mathbf{S} = X^T X = \sum_{i=1}^{I} \phi_i \phi_i^T$$

Because neither ϕ_i nor the inner product $\phi_i \phi_i^T$ are explicitly stated, the kernel trick is applied to solve the problem. The Gram Kernel matrix, $\mathbf{K}$, is calculated as function of the measurement vectors

$$\mathbf{K} = XX^T = \begin{bmatrix} \phi_1^T\phi_1 & \cdots & \phi_1^T\phi_I \\ \vdots & \ddots & \vdots \\ \phi_I^T\phi_1 & \cdots & \phi_I^T\phi_I \end{bmatrix} = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k(\mathbf{x}_1,\mathbf{x}_I) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_I,\mathbf{x}_1) & \cdots & k(\mathbf{x}_I,\mathbf{x}_I) \end{bmatrix}$$

and it is demonstrated that its allows to evaluate the scores of the model as follows:

$$\mathbf{t} = \mathbf{\Lambda}^{-1/2}\mathbf{P}^T\mathbf{k}(\mathbf{x})$$

where $\mathbf{P}$ contains the retained eigenvectors of $\mathbf{K}$, $\mathbf{k}(\mathbf{x})$ stands for the inner product of $\mathbf{x}$ with the entire set of I measurement vectors, and it is assumed that ϕ_i vectors have zero mean. A detailed explanation of the procedure can be found in Alcala and Qin [12].

To identify the measurements that signal the fault Alcala and Qin [10] formulated the RBC. The reconstructed observation vector along the direction $\boldsymbol{\xi}_j$, $\hat{\boldsymbol{x}}_j$, is formulated as

$$\hat{\boldsymbol{x}}_j = \mathbf{x}_m - \boldsymbol{\xi}_j f_j$$

where f_j is the fault magnitude. The reconstruction task consists in calculating f_j such that it minimizes the value of the hypothesis statistical test for $\hat{\boldsymbol{x}}_j$. Regarding KPCA [12], this implies that

$$f_j = arg\min \quad Statistic(\mathbf{k}(\mathbf{x}_m - \boldsymbol{\xi}_j f_j))$$

The RBC is defined as the square of the fault magnitude, that is, $RBC_j = f_j^2$.

On the other hand Godoy et al. [13] proposed to approximate a normalized statistic ($Statistic\_N$) by the sum of J VCs, $nc_j^{Statistic\_N}$ ($j = 1 \ldots J$). The j-th term is evaluated as function of the partial derivative of $Statistic\_N$ with respect to the j-th observation considering that the values of the other variables are constant and equal to the measured ones, as it is shown next

$$Statistic\_N \approx \sum_{j=1}^{J} \left(\frac{x_j}{2} \frac{\partial Statistic\_N}{\partial x_j} \right) = \sum_{j=1}^{J} nc_j^{Statistic\_N}$$

References

[1] A.J. Gupta, J.A. Hageman, P.A. Wierenga, J.W. Boots, H. Gruppen, Chemometric analysis of soy protein hydrolysates used in animal cell culture for IgG production—an untargeted metabolomics approach, Process Biochem. 49 (2014) 309–317.

[2] P. Van Den Kerkhof, G. Gins, R. Van Den Broeck, Van Impe, JFM./Multivariate assessment of activated sludge stability in lab-scale experiments, Process Biochem. 48 (2013) 1789–1793.

[3] Z. Ge, Z. Song, F. Gao, Review of recent research on data-based process monitoring, Ind. Eng. Chem. Res. 52 (2013) 3543–3562.

[4] H.J. Ramaker, E.N.M. Van Sprang, J.A. Westerhuis, A.K. Smilde, Fault detection properties of global, local and time evolving models for batch process monitoring, J. Process Contr. 15 (7) (2005) 799–805.

[5] J. Camacho, J. Pico, A. Ferrer, The best approaches in the online monitoring of batch processes based on PCA: does the modeling structure matter? Anal. Chim. Acta 642 (2009) 59–69.

[6] Y. Yao, F. Gao, A survey on multistage/multiphase statistical modeling methods for batch processes, Annu. Rev. Control 33 (2009) 172–183.

[7] N. Lu, Y. Yao, F. Gao, Two-dimensional dynamic PCA for batch process monitoring, AIChE J. 51 (2005) 3300–3304.

[8] Y. Yao, F. Gao, Multivariate statistical monitoring of multiphase two dimensional dynamic batch processes, J. Process Contr. 1 (2009) 1716–1724.

[9] J.M. Lee, Yoo Ch, I.B. Lee, Fault detection of batch processes using multiway kernel principal component analysis, Comput. Chem. Eng. 28 (2004) 1837–1847.

[10] C. Alcala, S.J. Qin, Reconstruction-based contribution for process monitoring, Automatica 45 (2009) 1593–1600.

[11] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Generalized contribution plots in multivariate statistical process monitoring, Chemom. Intell. Lab. Syst. 51 (1) (2000) 95–114.

[12] C. Alcala, S. Qin, Reconstruction-based contribution for process monitoring with kernel principal component analysis, Ind. Eng. Chem. Res. 49 (2010) 7849–7857.

[13] J. Godoy, D. Zumoffen, J. Vega, J. Marchetti, New contributions to non-linear process monitoring trough kernel partial least squares, Chemom. Intell. Lab. Syst. 135 (2014) 76–89.

[14] R.L. Mason, J.C. Young, Multivariate Statistical Process Control with Industrial Applications, SIAM, Philadelphia, USA, 2002.

[15] C.R. Alvarez, A. Brandolin, M. Sánchez, Batch process monitoring in the original measurement's space, J. Process Contr. 20 (2010) 716–725.

[16] C.R. Alvarez, A. Brandolin, M.C. Sánchez, On the variable contributions to the D-statistic, Chemometr. Intell. Lab. 88 (2) (2007) 189–196.

[17] M.V. Cedeño, L.P. Rodriguez, C.R. Alvarez, M.C. Sánchez, A new approach to estimate variable contributions to hotelling's statistic, Chemometr. Intell. Lab. 118 (2012) 120–126.

[18] G. Birol, C. Ündey, A. Cinar, A modular simulation package for fed-batch fermentation: penicillin production, Comput. Chem. Eng. 26 (2002) 1553–1565.

[19] S. Gurden, J. Westerhuis, R. Bro, A. Smilde, A comparison of multiway regression and scaling methods, Chemometr. Intell. Lab. (2001) 121–136.

[20] G. Robertson, M. Thomas, J. Romagnoli, Topological preservation techniques for nonlinear process monitoring, Comput. Chem. Eng. 76 (2015) 1–16.

[21] E. García-Portugués, R. Crujeiras, W. González-Manteiga, Kernel density estimation for directional-linear data, J. Multivariate Anal. 121 (2013) 152–175.

[22] B.S. Everitt, S. Landau, M. Leese, D. Stahl, Cluster Analysis, 5th edn., John Wiley &Sons Inc., USA, 2011.

[23] M. Maestri, A. Farall, P. Groisman, M. Cassanello, G. Horowitz, A robust clustering method for detection of abnormal situations in a process with multiple steady-state operation modes, Comput. Chem. Eng. 34 (2010) 223–231.

[24] X. Zhang, W. Yan, X. Zhao, H. Shao, Nonlinear biological batch process monitoring and fault identification based on kernel fisher discriminant analysis, Process Biochem. 42 (2007) 1200–1210.

[25] J. Van Impe, G. Gins, An extensive reference dataset for fault detection and identification in batch processes, Chemom. Intell. Lab. Syst. (2015) 20–31.