ELSEVIER

Chemical Engineering Journal 75 (1999) 69–76

Chemical Engineering Journal

Supervision of fed-batch fermentations

Lars Gregersen*, Sten Bay Jørgensen

CAPE Centre, Department of Chemical Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received 4 July 1997; received in revised form 29 May 1998; accepted 4 January 1999

Abstract

Process faults may be detected on-line using existing measurements based upon modelling that is entirely data driven. A multivariate statistical model is developed and used for fault diagnosis of an industrial fed-batch fermentation process. Data from several (25) batches are used to develop a model for cultivation behaviour. This model is validated against 13 data sets and demonstrated to explain a significant amount of variation in the data. The multivariate model may directly be used for process monitoring. With this method faults are detected in real time and the responsible measurements are directly identified. The fault detection and identification is enabled through inspection of a few simple plots. Thus, the presented methodology allows the process operator to actively monitor data from several cultivations simultaneously.

Keywords: Batch processes; Statistical process monitoring (SPM); Process chemometrics; Projection to latent structures (PLS); Principal component analysis (PCA)

1. Introduction

Batch processes are usually very difficult to model due to the circumstances under which they are used. Short runs and large batch to batch differences in process conditions make it difficult and time consuming to develop first principles models for the versatile reactor that the batch reactor actually is. Statistical process monitoring (SPM) is commonly used for monitoring continuous processes, where statistical methods are used to monitor that process variables are kept at a stationary level [1]. Fed-batch (or semi-batch) processes are, however, non-stationary and the process variables are, therefore, not constant. Thus, it is more difficult to develop a model for normal behaviour and to detect deviations from standard operation.

This paper shows an application of a multivariate statistical method for fault diagnosis. The method uses data which are obtained from existing standard measurements from an industrial process. Hence, no measurements have to be added in order to establish the described modelling method than normally would exist in equipment utilised by the fermentation process industries. When advanced analytical equipment is available (e.g., NIR spectra of the broth) it can be included, though. Data are used to develop a model of the normal behaviour of the process. Process knowledge enters the model development process when the measurements and the types of batches are specified and selected. The developed model can be used for on-line fault diagnosis and it is also demonstrated that the model can be used for prediction of the product concentration at the end of the batch.

A short introduction to the process described in this paper is given in Section 2. The crucial part of the process chemometrical way of modelling is the availability of process data and as long as this requirement is fulfilled the methods can be used for any process. The data handling is described in Section 3. The results of applying the methods to a fed-batch fermentation are described in Section 4 and the paper ends with a discussion and conclusions. A list of symbols is given in Appendix A.

2. Process description

The process investigated in this paper is an industrial fed-batch fermentation process where a *Bacillus* species produces enzymes. In this article the focus is on the main fermentor where the product is produced. The previous steps, spore propagation and seed tank, which have as purpose to produce biomass will not be dealt with, but the methods can be used for those as well to ascertain consistency and error propagation between steps.

*Corresponding author.

1385-8947/99/$ – see front matter
PII: S1385-8947(99)00018-2

The operating procedure for the modelled fed-batch fermentation is to start with a small amount of biomass and substrate in the main fermentor. When most of the initially added substrate has been consumed by the microorganisms the substrate feed is started. This operating procedure is used in order to keep the substrate concentration low during the fermentation. A low substrate concentration in the fermentor is necessary for achieving a high product formation rate due to the catabolite repressor effect.

The fed-batch operating procedure leads to a highly nonlinear process behaviour. Small changes in concentrations or flows can have a large effect on the kinetics of the internal reactions in the cells leading to large batch to batch differences in cell growth and product formation. Almost every key variable (volume, biomass, product concentration, etc.) is changing as the process progresses. This behaviour distinguishes batch and fed-batch processes from continuous processes where process control can be performed by maintaining key variables constant. In fermentation processes it is customary to keep the pH and temperature level constant in order to give the cells the best possible conditions for making the desired product. This is also the case for the process described in this paper and the pH and temperature data, therefore, show very little variation.

The nonlinear behaviour together with limited duration of the process makes it difficult to develop dynamic models of the system. Such modelling requires detailed knowledge about the microorganisms their metabolism and reaction rates and quantitative data from carefully planned experiments specific to the particular fermentation. Due to the large number of different microorganisms and products used by industry the effort that is needed to develop dynamic models for simulation, fault diagnosis and control seems too large to be overcome in the near future.

Due to the lack of models for model-based control the standard operating procedure of fermentation processes is to run with a predetermined feed profile that has been determined as the result of multiple optimization experiments, where high productivity and high reproducibility are the major objectives in these experiments. The optimization experiments can be very versatile when trying to optimise operating procedures, substrates and the microorganisms itself, but also takes long time. This is an ongoing task for most processes.

As a result of disturbances and differences in initial conditions the measurement trajectories can deviate from the expected optimal course. If the deviations are not compensated for the batch may have a reduced performance in terms of lower yield and production of unwanted by-products. On the other hand, many upsets of the process that can be seen through changes in the measurements will not have an effect on the quality of the process. In order to detect a fault, it is of course necessary that the fault affects, directly or indirectly, the measurements.

Abrupt, gross faults in single variables can easily and reliably be detected by a conventional process control system. Drift of variables and faults involving multiple variables are not easily detected. These more complicated faults, even if they are small, can have a large effect on the quality of the process; an effect that is difficult to predict without advanced tools.

It is conventionally up to the process operator to determine when deviations are unacceptably large and will have an effect on the product quality and the productivity. The reliability of this highly manual procedure depends on the training that the process operator has received, his experience and the character and number of processes that he has to supervise simultaneously.

The aim of the methods presented in this paper is to provide the process operators with a tool for detection and isolation of faults by limiting the amount of data that the process operator has to monitor in order to evaluate the present and future operation of the process. This tool is especially beneficial when process operators are monitoring several processes at the same time.

3. Data analysis

A set of on-line measurements are obtained from the process at regular sample intervals. These measurements have been stored for many past fermentations forming a database of historical information about the process. The measurements available for the considered process are shown in Table 1. Note that all the measurements are unsophisticated standard measurements. Thus, no advanced or expensive measurement devices has to be installed in order to make the methods work; the particular type of measurement is not important as long as the measurements represent the state of the process.

The model is entirely data based as opposed to conventional first principles modelling. It is up to the modeller to choose data in a way such that the data describes the behaviour of the system. The selection of the correct data for the modelling work is the most important step of the modelling phase. When a process data base already exists the task is reduced to selection among the data. Better results can usually be obtained if the data are obtained from designed experiments, but such an approach is costly and time consuming and is not considered in this paper. Designed experiment are more valuable and sometimes

Table 1
On-line measurements obtained from the fermentation process

1	Total amount of substrate	2	Agitator power input
3	Total amount of antifoam	4	Weight
5	pH	6	Temperature
7	Dissolved oxygen	8	Air flow
9	CO_2% in off-gas	10	O_2% in off-gas

essential when optimization experiments are performed on a process in order to perturb key variables.

The data used in this paper are obtained from a historical data base. It is desired that the model will work for all future batches. Therefore, a model is built using all available data sets (except for the data sets used for validation). A few of these data sets had to be left out because the course of the batches are so incompatible with the other batches that a single model cannot be formed that includes the variation of all the batches. The nonconforming data are left out in order to develop a model of the desired behaviour of the process. When a new batch is monitored one can then investigate if the batch is operating within the window of the desired behaviour given by the model. If the batch deviates too much one can say that a fault has occurred and that some action must be taken. As the model has been built on historical data where process faults have not been treated as suggested in this paper some faults are unfortunately allowed to persist. The goal is that once fault diagnosis has been implemented and the process variation has decreased, a new and better model will be developed leading to an ever improving process as this iteration progresses.

Other types of models can be developed that describe more specialised types of behaviour. Separate models can be built using data from batches that have a given high or low yield. Models can be built using data from batches that have experienced a certain type of fault. Building such specialised models gives higher specificity towards the type of error, but at the same time the specificity towards previously unseen errors is diminished.

Data from the batch and fed-batch processes can conveniently be put into a three-way matrix $\underline{\boldsymbol{X}}$ ($I \times J \times K$). I is the number of batches, J is the number of variables (10), and K is the number of samples from each batch (114). The numbers in parentheses refer the numbers actually used in this paper. The size can vary by orders of magnitude depending on the process duration and available measurements when other processes are modelled.

The matrix $\underline{\boldsymbol{X}}$ can be unfolded to a two-way matrix, see Fig. 1. This two-way matrix is called $\boldsymbol{X}$ ($I \times KJ$). For each fermentation (a row in $\boldsymbol{X}$) a quality measure is recorded and stored in a matrix $\boldsymbol{Y}$. The measure used here will be the final product concentration, but one could also use, e.g., the productivity. Each column of $\boldsymbol{X}$ corresponds to a certain variable at a certain point in time. If the process is carried out following a predetermined feed profile it is expected that the trajectories of the measurements are similar and that the mean value of a variable at a certain point in time can be used as a reference value for future processes. The goal of the monitoring is to observe and minimise deviations from this reference value in future batches. Thus, to facilitate the analysis the columns are centred and scaled to unit variance.

The matrix $\boldsymbol{X}$ is rather large, but the columns of $\boldsymbol{X}$ are not independent. They describe similar events in the process and the dimension of the space spanned by $\boldsymbol{X}$ is usually very low.

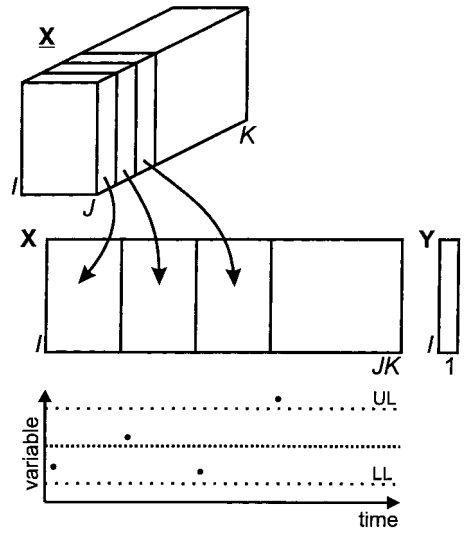

Fig. 1. Unfolding of a three-way matrix to form a two-way matrix. $\boldsymbol{X}$ contains the on-line variables and $\boldsymbol{Y}$ some measure of the quality of the process (here: the final product concentration). The principle behind process chemometrics is shown in the lower part of the figure for a single variable. Every time a new measurement is obtained it will be compared to the expected level. If the deviation is too large (above upper (UL) or below lower (LL) limits) the process is behaving abnormally and the process operator should take action.

Thus, by using a multivariate statistical technique to reduce the dimensionality of the variable space the problem of describing the process becomes much simpler to handle. Principal component analysis (PCA) is frequently used for this purpose and is recommended if no quality variables are available, which is frequently the case for many biotechnological processes. When quality variables are available one can use principal component regression (PCR) or preferably projection to latent structure (PLS) which is a linear regression method that optimally utilises the information in $\boldsymbol{X}$ and $\boldsymbol{Y}$ at the same time [2].

When the variable space is compressed using either PCA or PLS the process can be monitored in a low-dimensional space using simple plots [3,4]. The modelling methods requires data from several batches in order to produce a good model. Sometimes 20 batches are sufficient, but 50 or even 100 can be necessary if the dimensionality of the model space is large. The number can be reduced if some, possibly designed, experiments have been carried out that fills the model space in an efficient way, but this is a rare possibility for production scale fermentations.

In the present application a quality variable is available for the described process and, therefore, a PLS model will be developed. PLS is defined by a bilinear model that is used to model the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{W}^{\mathsf{T}} + \boldsymbol{E}, \quad \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\mathsf{T}} + \boldsymbol{F}, \quad \boldsymbol{u}_a = b_a \boldsymbol{t}_a. \tag{1}$$

PLS maximises the covariance between $\boldsymbol{u}_a$ and $\boldsymbol{t}_a$ and the number of components A (number of columns in $\boldsymbol{T}$) is chosen such that $\boldsymbol{E}$ and $\boldsymbol{F}$ are small in some sense. The

data are in other words reduced to a number of scores (either $\boldsymbol{T}$ or $\boldsymbol{U}$) that lie in a low dimensional space of the data, but describes a large fraction of the variation of the data. The expression in (1) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{F}^{*}. \tag{2}$$

This expression is in many cases easier to work with. The regression parameter matrix is given by

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}^{\mathsf{T}}\boldsymbol{W})^{-1}\boldsymbol{Q}^{\mathsf{T}}, \tag{3}$$

where the loading matrices $\boldsymbol{W}$, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are determined by the PLS algorithm [5,6]. $\boldsymbol{Y}$ can be predicted using:

$$\hat{\boldsymbol{Y}} = \boldsymbol{XB} = \boldsymbol{XW}(\boldsymbol{P}^{\mathsf{T}}\boldsymbol{W})^{-1}\boldsymbol{Q}^{\mathsf{T}} \tag{4}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{TQ}^{\mathsf{T}}. \tag{5}$$

The model can be used for calculating a vector $\boldsymbol{t}$ (a t-score) for a new data set $\boldsymbol{X}_{\text{new}}$

$$\boldsymbol{t}_{\text{new}} = (\boldsymbol{X}_{\text{new}}\boldsymbol{W}(\boldsymbol{P}^{\mathsf{T}}\boldsymbol{W})^{-1})^{\mathsf{T}}. \tag{6}$$

This expression can be used only when all data from a fermentation is available. For on-line purposes a full $\boldsymbol{X}$ matrix has to be constructed. In this paper $\boldsymbol{X}$ will be constructed by using all of the available information collected up to the current time and the remaining part of $\boldsymbol{X}$ will be filled with a copy of the most recently obtained measurement. This method results in good fault detection properties and reasonably well-behaved estimation of $\boldsymbol{Y}$ as well. This way of filling $\boldsymbol{X}$ corresponds to predicting what would happen if a fault is allowed to remain unchanged for the remaining duration of the batch and is a way to evaluate the seriousness of faults. This procedure is justified because the process dynamics become increasingly slower as the tank is filled and less change of the concentration variables is observed especially as the product concentration stabilises during the last part of the batch.

3.1. Fault diagnosis

Fault diagnosis consists of three steps: fault detection, isolation and identification (FDII). The methods presented in this section will readily detect faults and isolate the measurements that are behaving abnormally and the methods may facilitate the identification of the fault, i.e., to determine the physical origin of the fault in the process.

For detection, two statistics, the T_f^2 and the standard prediction error, can be calculated. The T_f^2 statistic (based on the Hotelling T^2 statistic [7]) is calculated using the scores

$$T_f^2 = \boldsymbol{t}_{\text{new}}^{\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{t}_{\text{new}} \sim \frac{A(I^2-1)}{I(I-A)}F_{A,I-A,} \tag{7}$$

where $\boldsymbol{S}$ is the covariance matrix of the t-scores contained in the matrix $\boldsymbol{T}$ calculated during the model development [8], I is the number of batches used for modelling and A is the number of components. F denotes the F distribution.

The squared prediction error (SPE) is calculated by

$$\text{SPE}_k = \sum_{r=(k-1)J+1}^{Jk} \boldsymbol{e}_r^{\mathsf{T}}\boldsymbol{e}_r, \tag{8}$$

where $\boldsymbol{e}_r$ is the rth column of the matrix $\boldsymbol{E} = \boldsymbol{X}_{new} - \boldsymbol{t}_{new}\boldsymbol{P}^{\mathsf{T}}$. In the simple case where there is only one new batch to be considered $\boldsymbol{E}$ is a row vector and $\boldsymbol{e}_r$ is a scalar and $\boldsymbol{e}_r^{\mathsf{T}}\boldsymbol{e}_r$ condenses into $\boldsymbol{e}_r^2$. The distribution of the SPE can be approximated by a weighted χ^2 distribution $\text{SPE}_k \sim (v_k/2m_k)\chi^2_{2m_k^2/v_k}$, where m_k and v_k are the mean and variance of the SPE obtained for the data set used for the model development at time instant k [4].

A fault is detected whenever the T_f^2 statistic or the SPE exceeds, e.g., a 95% confidence limit. The 95% limit is usually taken to be a warning level only and action is taken when the statistic exceeds a 99% limit. The T_f^2 statistic reveals faults that can be described by the model. The SPE will show if a totally new event is occurring in the process. This measure includes unusual variation of the controlled variables stabilised by simple control.

The process can also be monitored using the scores in a so-called score plot. Usually the number of components is low (2–3) and, therefore, a single plot is usually sufficient to display the state of the process. If the model contains more than two components one can either construct three-dimensional plots or make several two-dimensional plots to show the variation. Confidence limits can be established using the ellipsis defined by Eq. (7). These limits are suitable when developing the model since the data used in the development are all from similar batches (in some sense). When the model has been based on only a few data sets it is often seen that the ellipses are not filled evenly. This can be due to violation of the normality assumptions of the scores or simply a result of basing the model on historic data that do not cover the entire score space because the data were obtained under normal operating conditions. It has been proposed that kernel density estimates of the confidence limits can be beneficial when monitoring the process [9]. Using kernel density estimates for the limits it is assured that confidence is given to those scores that are in an area that actually have been encountered when building the model.

A general estimator for the kernel density can be defined as [10]

$$f(\boldsymbol{t}) = \frac{1}{n|\boldsymbol{H}|}\sum_{i=1}^{I} K_d\big(\boldsymbol{H}^{-1}(\boldsymbol{t}-\boldsymbol{t}_i)\big), \tag{9}$$

where $|\boldsymbol{H}|$ is the absolute value of the determinant of the matrix $\boldsymbol{H}$, which is a bandwidth matrix. K_d is the multivariate kernel function. One way of creating K_d from a univariate kernel K is by using a product kernel

$$K_d(\boldsymbol{u}) = \prod_{j=1}^{d} K(u_j). \tag{10}$$

The Gaussian kernel $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$ will be used

here. Other kernels may be used, but the resulting confidence bounds do not vary much when other kernels are chosen. The bandwidth matrix is chosen to be a diagonal matrix. This complies with the scores being independent. The bandwidth has to be chosen based on the shape of the data and the coarseness that is desired in the plot. A default value for the bandwidth is $h_i = 1.059 s_i I^{-1/5}$, where s_i is the standard deviation of the ith score [10]. For confidence limits in the score plots it has been found that the bandwidth usually has to be slightly larger than the default value leading to a coarse confidence region.

4. Experimental results

A model has been developed by carefully selecting data sets from the historical data base that reflect the normal desired operation of the fermentations. This has been done by first discarding any batch that has very large undesired or unusual behaviour compared to the desired batch behaviour (e.g., because of experiments or infections). When a batch is very short or very long it is discarded, too. The data suitable for the model development is truncated such that 114 time samples are included in the model. As mentioned, the dynamics of the process become slower towards the end of the batch and there are usually few corrective measures to be performed if a fault occurs near the end of a batch, anyway. An initial model is estimated and batches are removed from the modelling data set if they are lying outside a 99% confidence bound in a score plot using ellipses as confidence bounds. The procedure is carried out iteratively until all batches remain within the 99% confidence bound. It is important here to identify why a batch does not lie within the confidence bound in order to make sure that only those batches that are really not conforming, are eliminated from the normal data set.

After the reduction in the number of data sets 25 data sets were used for model development and 13 for validation of the model. Using the prediction error sum of squares (PRESS) as validation criterion two components are found to be sufficient for describing the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The obtained model uses only 28% of the information in $\boldsymbol{X}$, but explains 80% of the variation of $\boldsymbol{Y}$. The low percentage of used variation in $\boldsymbol{X}$ is due to the inclusion of controlled variables that have low variation (e.g., pH and temperature). These variables are known to have large influence on the product formation and that is the reason they are controlled. If the influence of the controlled variables on the product formation is to be modelled by the PLS model these variables must be perturbed in designed experiments. These experiments have not been performed since that would be expected to lead to a decrease in product formation and gross variation in these variables is not encountered due to the control. If the PLS model was to be used entirely for prediction purposes and it is assumed that the control is perfect the model performance could be improved by not including the controlled variables in the model. In the present case we are interested in fault diagnosis of the controlled variables, too, and therefore, leave the controlled variables in the model.

4.1. On-line estimation of final product concentration

Using Eqs. (4)–(6) the final product concentration can be predicted in real time. Fig. 2 shows the performance of this method for a low producing fermentation. The data from this batch, which was used for model validation, but not for the modelling itself, will be used in the following section. The final product concentration is 0.40, whereas the average value for the batches used for model development is 0.46. Fig. 2 shows that the model is able to predict the final product concentration within 10% during most of the fermentation except during the interval from 70 to 80, where there is a large deviation due to a process fault. The accuracy of the estimation is slightly larger than the accuracy provided by the lab when chemical analyses are performed. This deviation can be interpreted further by the fault diagnosis as described in the following section.

The T_f^2 statistic in Fig. 3 plotted as a function of time for the same batch as in Fig. 2, shows that this particular batch has a large deviation from $t = 70$ to $t = 80$. The slow drift of the T_f^2 that can be noted is difficult to detect by looking at the raw measurements. It has already been indicated (in Fig. 2) that this process drift results in a much lower than average product concentration at the end of the fermentation. Fig. 2 illustrates the importance of this type of monitoring to enable the prediction of the consequences of deviations.

The SPE in Fig. 4 indicates that this process is deviating from the average process almost throughout the entire fermentation.

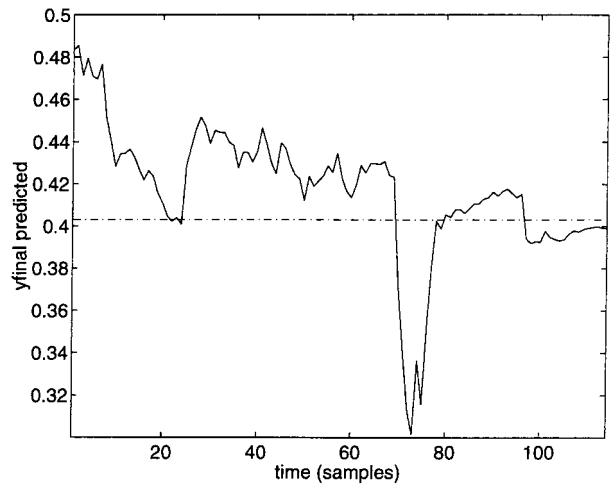

Fig. 2. Prediction of final product concentration. The dotted line indicates the actual product concentration as it was measured at the end of the batch. If the large fault at $t = 70$ was allowed to persist throughout the fermentation the product concentration was estimated to be much lower than the one actually obtained.

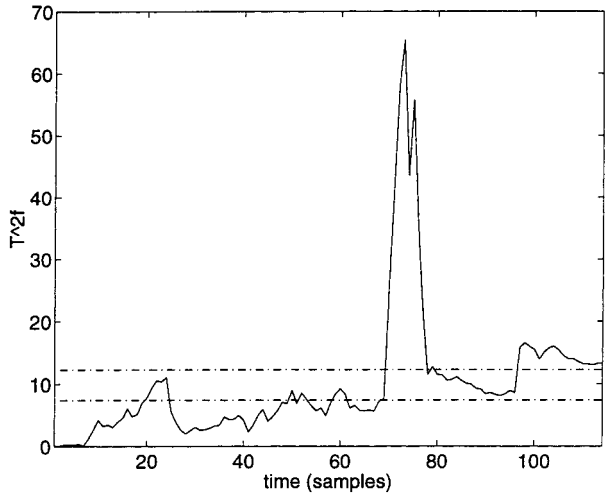

Fig. 3. T_f^2 statistic. Dash-dotted lines indicate 95% and 99% confidence limits.

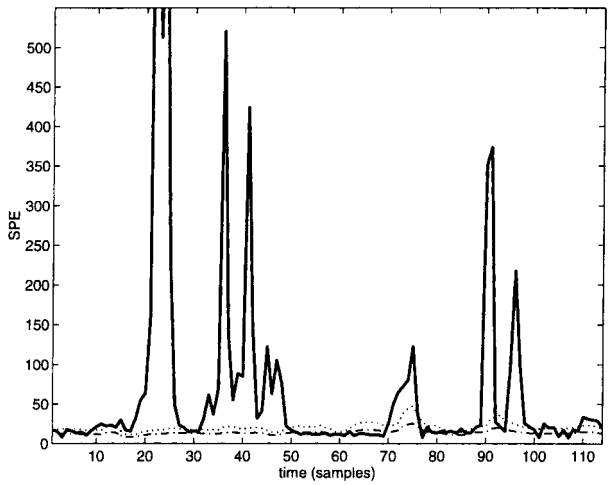

Fig. 4. Squared prediction error (SPE). Max value at peak when $t = 20$ is about 1400. The dash-dotted line is the 95% and the dotted line is the 99% confidence limits.

Contribution plots, which indicate the variables that are contributing most to the T_f^2 statistic or the SPE, can easily be constructed and can, thus, be used in the fault identification [11,4]. Contribution plots can either be used to find the change in the contribution from one point in time to another or the contribution plot can be used to find the deviation of the current batch when compared to the normal batch behaviour described by the model. We here choose to look at the fault which has been detected around $t = 70$. Fig. 5 shows the change in contribution of the variables from a point in time just before the fault could be detected in the SPE and T_f^2 plots ($t = 67$) to the point where the fault is at its highest ($t = 73$). The figure shows that there is a large change in the contribution of the CO_2 and O_2 measurements. Thus, the task of isolating faulty measurement has been reduced to looking only at a few plots that show that a fault has occurred and contribution plots to identify the variables that contribute to the fault.

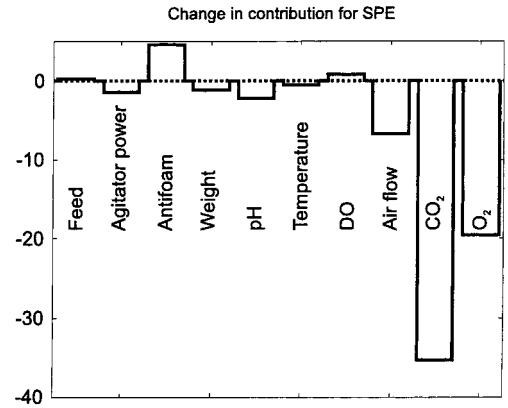

Fig. 5. Contribution plot showing the change of the process from $t = 67$ to $t = 73$. It is seen that the variables CO_2 and O_2 give highest contribution to the fault and that they are lower than normal. From the plot of the predicted final product concentration (Fig. 2), it can be seen that the fault has a negative effect on the quality.

4.2. Score plots

A score plot can be used to monitor the process. Since the model only contains two components the plot in Fig. 6 is the only one needed to monitor the major variations of a normally operating batch. The figure shows the variation of the process in the reduced space of the two components.

The score plot describes the present state of the process and allows the operator to interpret the development of the process. Emphasis must be put on the word *interpret*

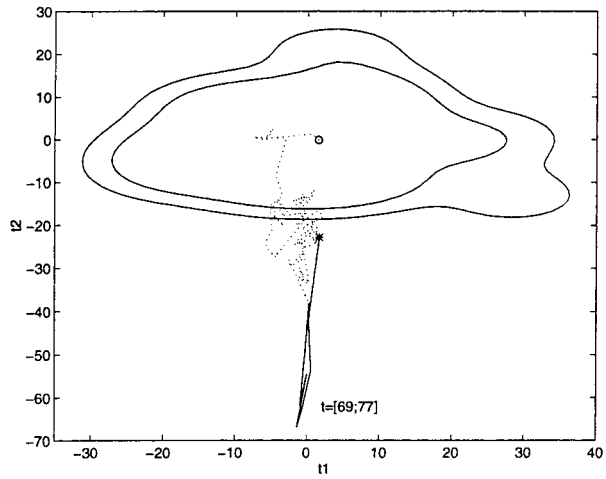

Fig. 6. Score plot for a faulty batch illustrating the development of the process in a reduced space. The beginning of the fermentation is marked with a 'o'. The time interval [69;77] (starting with a '*') is shown as solid lines. The score t_1 varies mainly when there are large oscillations in the temperature. t_2 varies mainly when the CO_2 and O_2 measurements change. Confidence limits are kernel density estimates.

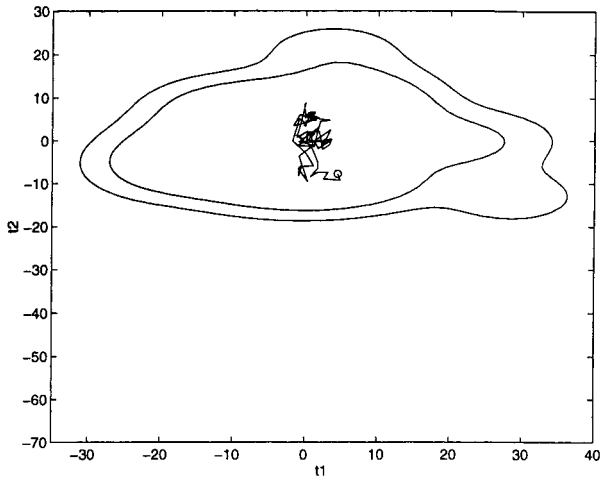

Fig. 7. Score plot for a well-behaved batch. The scores remain near the point (0,0) suggesting that the product concentration will end on the average value of the batches that where used to form the model. Confidence limits are kernel density estimates.

because the scores usually lack any direct physical meaning. One way of interpreting the score plot is to ascribe different phenomena to the movement of scores. For example, this model shows large variation of the t_2 score when deviations in the O_2 uptake and CO_2 production occur. Another way of finding a physical relationship is to investigate the loading matrices, which directly show the relationship between the measurements and the scores. Both interpretations can be useful when the behaviour of a batch is to be described and current or future faults are to be eliminated.

Eq. (5) shows the relationship between the scores and the dependent variable y as $\hat{\boldsymbol{Y}} = \boldsymbol{TQ}^{\mathrm{T}}$. Since in this case $\boldsymbol{Q} = [0.05\ 0.08]$ it can be inferred that batches will have a higher than average product concentration at the end of the batch when the score values all are positive, i.e., the scores are moving around in the upper-right corner of the score plot.

Score plots can, furthermore, be used as a fingerprint of the batch. Instead of investigating plots of the different measured variables one can use a score plot for an entire batch to investigate if something unusual has happened during the fermentation. Fig. 7 shows such a score plot of a well-behaved batch. The scores stay in this figure close to the point (0,0) which shows us that this batch did not have any faults that affected the product concentration. It would have been much harder to interpret the original measurements due to their time varying nature.

5. Discussion and conclusion

The methods shown above are powerful tools for compressing and displaying process information in a meaningful way. The methods can be used both for fault diagnosis and for prediction purposes. It should be noticed that the predictions are obtained in real time as opposed to wet chemical analyses and with almost the same accuracy.

Using the demonstrated methods the operator is provided with a clear view of the process performance. Instead of watching 10 (correlated) variables at the same time it is sufficient to inspect only two simple plots in order to evaluate the present and future behaviour of the process.

The relationship between the measurements and the quality variable (product concentration) is utilised by the model such that measurement deviation that do not signify a quality change are not marked as fault in the T_f^2 and score plots. If quality variables are unavailable or it is believed that any measurement deviation should be marked as a fault a process model using PCA instead of PLS may be developed. Such a model will lead to the same type of fault diagnosis plots.

The displayed figures are intended as process operator tools to facilitate monitoring process performance. With these tools operator attention can be directed mainly at faulty processes instead of constantly watching all concurrently running processes.

As displayed here these data-driven methods are only used for supervision and not directly for control. It will be a relevant future step to develop an expert system that can be used to interpret the score plots and automatically take appropriate action when certain kinds of (known) faults occur. Without any doubt, the described chemometrical tools will lead to a higher utilisation of process data in modelling, optimisation and control of complicated processes because process chemometrics provide the process industries with data handling methods that can utilise the process knowledge contained in process data, which currently are obtained and stored for many processes.

Appendix A

Scalars and functions

A	Number of components in the model.
I	Number of batches.
J	Number of variables.
K	Number of samples within a batch.
m_k	Mean of the SPE.
t_{new}	Score vector for a new batch.
v_k	Variance of the SPE.
T_f^2	Hotelling's T^2 statistic for a future batch.
$K_d(u)$	Multivariate kernel density function.
$K(u)$	Univariate (Gaussian) kernel density function.

Matrices and vectors

B	Regression matrix.
E	Residual in modelling X.
F	Residual in modelling Y.
F^*	Residual in modelling the regression $Y = XB$.

H	Bandwidth matrix (diagonal) used for kernel density estimation.
P	X loading.
Q	Y loading.
S	Covariance matrix of the scores T.
T	Score matrix for X.
U	Score matrix for U.
W	Normalised X Loading.
$\underline{\boldsymbol{X}}$	Three-way data matrix ($I \times J \times K$).
X	Unfolded data matrix ($I \times JK$).
X_{new}	Unfolded data set for a new batch ($1 \times JK$).
Y	Quality variable ($I \times 1$).
$\hat{\boldsymbol{Y}}$	Prediction of Y

References

[1] R.J. Pond, Fundamentals of Statistical Quality Control, Prentice-Hall, Inc., New York, 1994.

[2] J. Edward Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.

[3] P. Nomikos, J.F. MacGregor, Multivariate spc charts for monitoring batch processes, Technometrics 37(1) (1995) 41–59.

[4] P. Nomikos, Statistical Process Control of Batch Processes, Ph.D. Thesis, McMaster University, 1995.

[5] A. Höskuldsson, Regression Methods in Science and Technology, vol. 1, Thor Publishing, Arnegaards Allé 7, Denmark, 1996.

[6] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1989.

[7] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, 2nd ed., Academic Press, New York, 1980.

[8] N.D. Tracy, J.C. Young, Multivariate control charts for individual observations, J. Qual. Technol. 24(2) (1992) 88–95.

[9] E.B. Martin, A.J. Morris, M.C. Papazoglou, C. Kiparissides, Batch process monitoring for consistent production, Computers Chem. Eng. 20 (1996) S599-S604, Suppl.

[10] J.S. Simonoff, Smoothing Methods in Statistics, Springer, Berlin, 1996.

[11] P. Miller, R.E. Swanson, C.F. Heckler, Contribution plots: a missing link in multivariate quality control, 37th Annual Conference, ASCQ, NY, 1993.