Process performance monitoring using multivariate statistical process control

E.B.Martin
A.J.Morris
J.Zhang

Indexing terms: Statistical process control, Multivariate projection techniques, Fault detection, Fault diagnosis, Nonlinear principal components analysis

Abstract: Statistical process control (SPC) is a tool for achieving and maintaining product quality. Classical univariate statistical techniques have focused on the monitoring of one quality variable at a time and are not appropriate for analysing process data where variables exhibit collinear behaviour. Minimal information is derived on the interactions between variables which are so important in complex manufacturing processes. These limitations are addressed through the application of multivariate statistical process control (MSPC). The bases of MSPC are the projection techniques of principal components analysis and projection to latent structures. The philosophy behind these approaches is to reduce the dimensionality of the problem by forming a new set of latent variables to obtain an enhanced understanding of the process behaviour. If the variables are highly correlated, then the process can be defined in terms of a reduced set of latent variables, which are a linear combination of the original variables. The authors present an overview of multivariate statistical process control and its nonlinear extension for process monitoring. The power of the methodology is demonstrated by application to two industrial processes.

1 Introduction

Statistical process control (SPC) forms the basis of process performance monitoring and the detection of process malfunctions. The objective of SPC is to monitor the performance of a process over time to verify that it remains in a state-of-statistical-control. Univariate SPC systems consider individual quality measurement sources, and as a consequence the interactions between variables which are so important in today's complex manufacturing plants are not considered. Implementing univariate SPC results in the majority of the variables collected on a process not being monitored. Furthermore, the monitored variables are not necessarily independent hence examining a limited group of variables, one at a time, makes the identification and interpretation of process malfunctions extremely difficult, and consequently the results of the analysis may provide misleading information.

Two types of information are monitored on a process; quality measurements (e.g. colour, texture, strength, weight, size, moisture content, material properties, etc.) and process information (e.g. temperature, pressure, flow rates, speed, etc.). Compared with the process measurements, where a large number of variables are monitored, only a limited number of quality variables are recorded and with a much lower and variable frequency. Univariate statistical process control charts, such as Shewhart charts ($\bar{X}$ and Range charts), CUSUM (cumulative sum) plots, and EWMA (exponentially weighted moving average) charts are typically used for the monitoring of a small number of quality variables. These charts compare current process performance against process behaviour when the product being produced is known to conform to pre-assigned specification limits or when the process continues to operate within statistically derived control limits. The only variation present is as a result of common cause variation which cannot be eliminated from the plant. That is, the process is said to be in a state of '*statistical control*'. Univariate charts are based upon the '*magnitude*' of the variable deviations. It is, therefore, not surprising that they can provide misleading information to operators as to when the process is in a state of statistical control since they ignore the interdependence between the variables. This limitation can have serious consequences since the operational failure of manufacturing plants and their associated instrumentation, controllers and control manipulators is of considerable importance in the management of the plant. Process malfunctions can lead to reduced product quality, reduced production, plant shutdowns, increased reworking, increased environmental impact and a low return on plant assets. Consequently, the early warning of deviations from nominal production can provide significant strategic advantages. In today's competitive markets, companies are required not only to react rapidly to changing market demands, but to improve upon current levels of consistency and reliability of production. The need for early warning of potential production problems becomes a key issue.

Multivariate statistical process control methods (MSPC) address some of the limitations of univariate monitoring techniques by considering all the data simultaneously and extracting information on the

IEE Proceedings online no. 19960321

Paper first received 15th August 1995 and in revised form 30th January 1996

The authors are with the Centre for Process Analysis, Chemometrics & Control, University of Newcastle, Newcastle-upon-Tyne NE1 7RU, UK

'*directionality*' of the process variations. That is, the behaviour of one variable relative to the others. These projection techniques allow the efficient handling of large amounts of monitored plant data which is subject to measurement errors, is ill conditioned and the variables exhibit collinear behaviour. Furthermore, they are not bound by the restrictive assumptions of variable independence and data normality. All the MSPC results presented in the paper have been produced using a multivariate data analysis package MULTIDAT (copyright: CPACC University of Newcastle) which has been written in MATLAB code.

2 Multivariate statistical projection methods

The cornerstones of MSPC are the projection methods of principal components analysis (PCA), [1], and projection to latent structures or partial least squares (PLS), [2]. The philosophy of these techniques is to reduce the dimensionality of the problem by forming a new set of variables. Principal components analysis (PCA) reduces the dimensionality of the problem by defining a series of new variables, principal components, which are a linear combination of the original measured variables and which explain the maximal amount of variability in the data. Adopting a similar approach to PCA, projection to latent structure (PLS) simultaneously reduces the dimensionality space of both the process variables and the product quality information, and a model is developed for the prediction of the quality variables in the reduced variable space.

2.1 Principal components analysis

The primary objectives of principal components analysis (PCA) are data summarisation, classification of variables, outlier detection, early warning of potential malfunctions and 'fingerprinting' for fault identification. PCA seeks to find a few linear combinations which can be used to summarise the data with a minimal loss of information. This reduction in dimensionality can be described as a 'parsimonious summarisation' of the data.

Let $X = \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_m$ be an m-dimensional data set describing either the process variables or the quality information. The first principal component is that linear combination of the columns of X, i.e. the variables, which describe the greatest amount of variability in X, $t_1 = Xp_1$ subject to $|p_1| = 1$. In the m-dimensional space, p_1 defines the direction of greatest variability, and t_1 represents the projection of each object onto p_1. The second principal component is the linear combination defined by $t_2 = E_1 p_2$ which has the next greatest variance subject to $|p_2| = 1$ and subject to the condition that it is orthogonal to the first principal component, t_1:

$$t_2 = E_1 p_2 \quad \text{where} \quad E_1 = (X - t_1 p_1^T)$$

This procedure is essentially repeated until m principal components are calculated. In effect, PCA decomposes the observation vector, X, as

$$X = TP^T = \sum_{i=1}^{m} t_i p_i^T$$

where p_i is an eigenvector of the covariance matrix of X. P is defined as the principal component loading matrix and T is defined to be the matrix of principal component scores. The loadings provide information as to which variables contribute the most to individual principal components, i.e. they are the coefficients in the principal components model; whilst information on the clustering of the samples and the identification of transitions between different operating regimes is obtained from the scores.

Principal component analysis depends critically upon the scales used to measure the variables. If we consider a set of multivariate data where the variables, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, ..., $\mathbf{x}_m$ are of completely different types, for example pressures, temperatures, flow rates, etc., then the structure of the principal components derived from this data set will depend essentially upon the arbitrary set of units of measurement. If there are large differences between the variances of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, ..., $\mathbf{x}_m$, those variables whose variances are large will tend to dominate the first few principal components. It is found that in practice these variables may not be of prime importance in detecting process malfunctions. This lack of scale invariance implies that care needs to be taken when scaling the data. Different scaling routines can produce different results. Three possible ways to scale the data are: select 'natural units' by ensuring all the variables measured are of the same type; variables can be mean-centred; or the variables can be scaled to zero mean and unit variance. The calculation of the principal components is then based upon the transformed matrix. There are no clear-cut rules as to which form of scaling should be adopted, it is totally problem dependent. It may be beneficial to examine the results using different scaling regimes.

One of the features of PCA is that the less important components often describe the noise in the data. If the process variables are colinear, k principal components ($k \leq m$) will explain the majority of the variability, i.e. a smaller number of principal components than original variables are required to explain the variability in the data. Consequently it is desirable to exclude these components.

$$X = TP^T + E = \sum_{i=1}^{k} t_i p_i^T + E$$

The number of principal components that provide an adequate description of the data can be assessed using a number of techniques. Typically, cross-validation is employed [3]. In practice two or three principal components are frequently sufficient for multivariate SPC with the latent variables generated from PCA forming the cornerstone of the multivariate statistical process control charts [4–6].

In process monitoring, once a plant malfunction has been detected it is important to identify those variables, or combination of variables, that characterise the problem. Typically, in the literature, it is usually emphasised that the first two principal components contain all the important information. For the early warning of a plant malfunction this may not necessarily be the case. In some situations, variables that indicate the onset of a plant problem have minimal impact on the first two components but are seen to dominate the lower-order principal components. Analysis and usage of these lower order components provides a valuable aid to plant fault detection.

2.2 Prediction of product quality

Economic pressures on market perceptions of product quality and reliability has necessitated that the more slowly monitored quality measures be predicted from

the more rapidly monitored process variables; inferential estimation or software sensor. This can be achieved through the application of regression methods, i.e. the identification of a relationship between the quality information and the process variables, hence knowledge of the final quality of the product may be predicted prior to that obtained from the quality control laboratory, for some processes this can result in a saving of many hours. The most widely used technique is multiple linear regression (MLR). However, this approach is inappropriate in multidimensional systems where variables are typically highly correlated, since the model coefficients can be numerically unstable, small perturbations in the data can result in major changes to the model. Alternative regression approaches such as ridge regression and regularisation methods resolve the problem of singularity but do not automatically reduce the dimensionality of the problem; this is of core importance in process performance monitoring in order not to overload process operators with a large number of charts.

Principal component regression (PCR), based upon the regression of the scores, calculated from PCA, on the quality variable of interest, addresses both the singularity and dimensionality problem but it treats the quality variables as though they are independent. The technique of projection to latent structures (PLS) utilises the information on both the process and quality variables and treats both sets of information as being dependent. Nonlinear versions of these approaches, nonlinear PCR and nonlinear PLS, have recently been developed [7–9] which incorporate a neural network within the inner model of the algorithm.

2.3 Projection to latent structures (PLS)

At the present time, statistical quality control methods based upon the product quality data have been the standard approach to process monitoring. Consequently the majority of the data collected on process variables is wasted. In multivariate SPC, the process data can be used in conjunction with the quality information. The process variables are typically collated in a data matrix X, with the quality information contained in a data matrix Y, these two sets of information can be related using the multivariate statistical technique of projection to latent structures (PLS).

Projection to latent structures is a tool for solving regression problems with highly collinear process variables. It simultaneously reduces the dimensionality of the X and Y spaces to find the latent vectors for the X and Y spaces which are most highly correlated, i.e. those that not only explain the variation in X, but that variation in X which is most predictive of Y. The technique is well referenced in the literature [2, 10]. Given a set of information on m process variables, X, and k quality variables, a factor from the X and Y data are evaluated, t_1 and u_1,

$$t_1 = Xw_1 \quad \text{and} \quad u_1 = Yc_1$$

These equations are referred to as the outer relation for the X block and Y block, respectively. The vectors w_1 and c_1 are called factor weights. PLS finds the factor weights in such a way that t_1 and u_1, are most correlated to one another. A linear regression is then performed between the first pair of factors t_1 and u_1,

$$u_1 = b_1 t_1 + \varepsilon$$

This relationship is termed the inner relation of the X and Y blocks. The final stage of the algorithm is the regression of X on its factor t_1 and Y on its factor u_1:

$$X = t_1 p_1^T + E \quad \text{and} \quad Y = u_1 q_1^T + F$$

where E and F are the residuals. The above three steps are repeated with the residuals E and F replacing X and Y, respectively. Up to m pairs of factors can be derived, but in a similar way to PCA, only the first few contain the important information relevant to the operation of the process. Cross validation or other related approaches can be used to select the number of factors [3].

PLS can also be viewed as a biased regression method and the final model can be expressed in terms of the X data as the regression model:

$$Y = X\beta + F$$
$$\beta = W(P^T W)^{-1} Q^T$$

Projection to latent structures enables information on the final quality of the product to be made available to plant operators on a more regular and timely basis. Just as in PCA, the scores of the new variable combinations resulting from the application of PLS can form the basis of plant performance monitoring charts. An interesting question is then 'Which approach is the more appropriate for performance monitoring?'. The PLS-based approach is used when Y is of high quality and measured as frequently as X, or when a good representation exists between X and Y; otherwise PCA should be used.

3 Performance monitoring charts

The primary requirement for the development of multivariate SPC charts is the acquisition of data that is representative of nominal process operation, that is when the plant is producing 'within specification product'. The data population can be obtained from historical data bases or designed experiments. A model, based on historical data collected when only common cause variation was present, can be constructed using either PCA or PLS. Future behaviour is then referenced against this 'nominal' or 'in-control' model. The basis of the success of this approach is the recognition that many of the measurements are highly correlated and thus different combinations of the variables may define the same underlying disturbances or events occurring in the process. Consequently, it can be assumed that when the process is producing a product within predefined specification limits, the dimensionality of the process can be substantially reduced to a few latent variables. Typically, this information is presented graphically in terms of time series plots, and two and three dimensional representations of the new latent variables. The three most common forms of monitoring charts are those of the scores against time, two and three dimensional plots of the scores and the squared prediction error.

3.1 Squared prediction error plot

We initially focus upon PCA, although the concepts are directly transferable to PLS as will be seen later. Once a model has been developed from the nominal data set using k principal components, $X = TP^T$, the fitted values, $\hat{x}_{ij}$, can be calculated for each new multivariate observation. These values are then used to evaluate the squared prediction error, SPE, for each new observation, x_{ij} i.e. the squared difference between the observed values and the predicted values from the

nominal or reference model:

$$\sum_{j=1}^{k}(x_{ij} - \hat{x}_{ij})^2$$

The squared prediction error plot provides the user with a facility to identify a previously unidentified event. Using the latent variable relationships developed from the historical data set, the scores for each new observation are located within the score plane. Typically, although not necessarily, principal components one and two and the calculated value of the SPE are plotted in a three-dimensional diagram. Figs. 1–3 show a number of performance monitoring chart configurations involving plots of SPE and the scores, [4]. In Fig. 1 the principal components T_1 and T_2 form the x and y-axes of the monitoring chart, respectively, with the squared prediction error (SPE) defining the z-axis. Each observation is located on this plot via its scores and SPE. Fig. 2 gives three possible two-dimensional representations of the scores and the SPE. Fig. 3 can be used to monitor the SPE at each sampling time point.

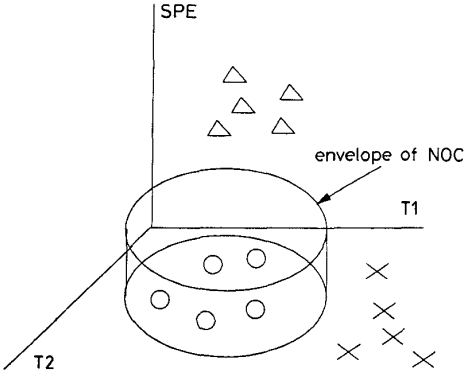

Fig. 1 *Three-dimensional representation of squared prediction error and scores*

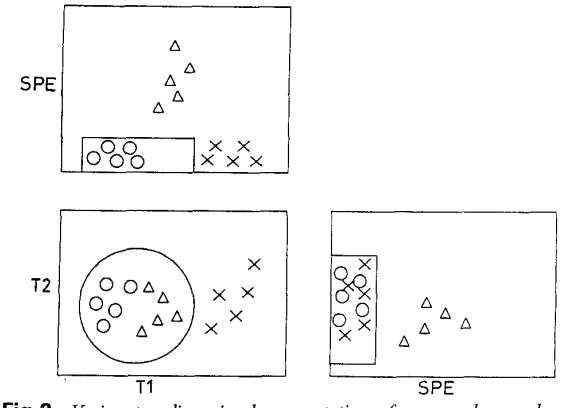

Fig. 2 *Various two-dimensional representations of scores and squared prediction error*

A process malfunction can lead to one of the following two situations. On the one hand, the fault can change the correlation structure amongst the measured process variables. In this case, the nominal PCA model is no longer valid and significant prediction errors will result. This can be detected from the SPE plot and is represented by 'Δ' in Figs. 1–3. On the other hand, the process malfunction may not alter the correlation structure among the process variables. For this scenario, the SPE will remain within the control limits but the scores will move outside the envelope of nominal operation. This situation is represented by '×' in Figs. 1–3. Acceptable process performance would fall within the envelope of normal operation ('○' in Figs. 1–3).

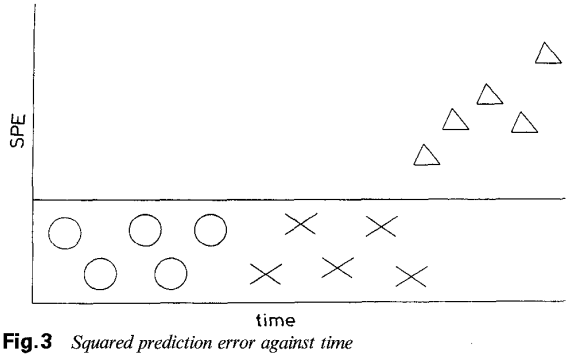

Fig. 3 *Squared prediction error against time*

3.2 Confidence bounds

Once a model has been developed which reflects nominal plant production, it is necessary to detect any departure of the k-dimensional process from the reference model. Adopting an approach similar to that for univariate statistical process control, nominal operating regions can be defined for each principal components scores plot based upon standard statistical distributional theory. One approach is to assume that the underlying k-dimensional process is normally distributed. It is possible to determine whether the process is in control by calculating Hotelling's [11] squared distance for each pair of principal components of interest:

$$T^2 = n(\mu_0 - \bar{x})^T S^{-1}(\mu_0 - \bar{x})$$

T^2 is distributed as the statistic $(n-1)kF/(n-k)$, where F has a central F-distribution with k, $n-k$ degrees of freedom. This relationship is used to establish control limits where there is an $100\alpha\%$ chance of a false alarm, where α is typically of the value 0.05 or 0.01. An out of control signal is identified if

$$T_0^2 > \frac{(n-1)kF_{k,n-k,\alpha}}{n(n-k)}$$

Confidence bounds for a scores monitoring chart are comparatively straightforward to evaluate. Rules similar to those adopted for univariate SPC charting methodology for identifying when a process is out of control, or moving out of control, can be applied: (i) two points lie outside the warning limits (i.e. $\alpha = 0.05$), (ii) one point lies outside the action limits (i.e. $\alpha = 0.01$), (iii) seven points consistently increase or (iv) seven points consistently decrease.

3.3 Interpreting the 'out of control' signal and 'causation variables' selection

When a process is recognised to be out-of-control or moving out-of-control, operational personnel need a tool to identify the variables, or combination of variables, that are responsible for, or indicative of, a drift in the process operating conditions. Adjustments can then be made to the process to avoid the continued manufacture of non-conforming product. Typically, once the process has been identified as not being in a state of statistical control, it is usually the responsibility of the process operators or process engineers to diagnose the assignable cause(s) and implement the appropriate cor-

rective action. This stage is more easily carried out through the interrogation of the multivariate control charts constructed using either PCA or PLS and by reverting back to the original process variables and examining their contribution to the calculated scores and the squared prediction error. At this juncture it may be possible to identify the most likely set of original variables whose contribution has increased over that defined in the nominal model and which are reflective of the non-conforming behaviour. One possible graphical approach is through the implementation of a variable contribution plot, [12], which describes the change in the new observation variables relative to the average value calculated from the PCA/PLS model.

However, it is not necessarily those variables which exhibit the greatest changes which are solely reflective of an out-of-control signal. It may be a conjunction of small, or jointly small and large, changes in the process variables which indicate the probable cause of the problem. In practice, such diagnostic plots could become prohibitively complex in processes where there are a large number of monitored variables and potentially there will be many variable combinations to interrogate. In this case a random search optimisation procedure could provide an automatic method of handling a potentially combinatory explosive problem and should help identify the critical combination of variables. This is currently being explored.

4 Example of a continuous manufacturing plant

We consider for illustration of these techniques a manufacturing process where there are 14 online monitored variables (Table 1) and 5 property (quality) variables (Table 2) which are measured offline in the quality assurance laboratory, Table 1.

Table 1: On-line measured process variables of the continuous manufacturing plant

Process variables	Definition
T_{in1}	inlet temperature
T_{max1}	maximum temperature in zone 1
T_{out1}	outlet temperature from zone 1
T_{max2}	maximum temperature in zone 2
T_{out2}	outlet temperature from zone 2
T_{cin1}	inlet temperature of zone 1 coolant
T_{cin2}	inlet temperature of zone 1 coolant
X_{tmax1}	position where Tmax1 occurs
X_{tmax2}	Position where Tmax2 occurs
F_{i01}	inlet flow of feed 1
F_{s01}	inlet flow of intermediate feed 1
F_{i02}	inlet flow of feed 2
F_{s02}	inlet flow of intermediate feed 1
Pres	vessel pressure

Table 2: Off-line measured process variables of the continuous manufacturing plant

Quality variables	Definition
Conv	cumulative conversion of monomer
MW_n	number of average molecular weight
MW_a	weight average molecular weight
LCB	long chain branching
SCB	short chain branching

The economic operation of this process usually requires that un-reacted species be recovered and recycled back into the process. This can introduce impurities into the manufacturing process. A further identified problem is the occurrence of vessel fouling which limits the heat transfer capability thus making the temperature control system less effective. Data from the stable operating regions is available from a historical database and provides the nominal (reference) data set.

Table 3: Cumulative percentage of variability explained in X-block using principal components analysis

Principal component	1	2	3	4	5	6	7
Cumulative percentage of variability explained in X	33.3	51.1	67.1	76.6	84.4	90.0	95.0

For this problem, a PCA analysis was performed on the nominal data and the principal components evaluated, Table 3. The first four principal components explain approximately 80% of the variation in the process variables whilst seven components explain 95% of the variability in the data.

Using cross-validation, seven components were identified to be the 'optimal' number of components to be used in the monitoring of the process. Multivariate monitoring charts were initially constructed using the scores from the first two principal components together with the squared prediction error (SPE). Fig. 4 shows the resultant plot of the scores for principal component 1 versus principal component 2, calculated from the nominal data; '*' define the nominal data set. As new data is monitored the effect of a process malfunction (in this case a fouling problem) is observed on both charts '+'. In this process, the effect of fouling on the temperature profile results in the new scores appearing outwith the nominal data region. This would have alerted the plant operators of a problem on the process. It is interesting to observe that the plot of principal component two versus principal component three, Fig. 5 provides even more conclusive evidence of the occurrence of a problem within the process. Although a problem within the reactor was identified in principal component two, principal component three provides stronger evidence of a problem. The value of the scores for principal component three are larger in magnitude for the non-conforming data compared with those from principal component two. In terms of the early warning of plant malfunctions it is not always the first two principal components which aid the identification of the process moving out of control. Fig. 6 demonstrates that the effect of a problem within the process can also be identified using the squared prediction error plot.

The confidence bounds plotted on the scores plot are calculated using Hotelling's T^2 statistic. A problem with the T^2 confidence bounds is that one percent of the samples, known to conform to the customers requirements, can potentially lie outwith the bounds because of the statistical interpretation of confidence bounds. A further problem with the T^2 bounds is that regions of no information are identified within the bounds, the assumption made is that if a sample falls within the T^2 bounds then there is no problem with the process. This feature of confidence bounds has been identified by operators and engineers as a limitation of the current methodoloy. A modified form of confidence bounds which totally encapsulates the nominal data set

has recently been developed to overcome these problems [7–9].

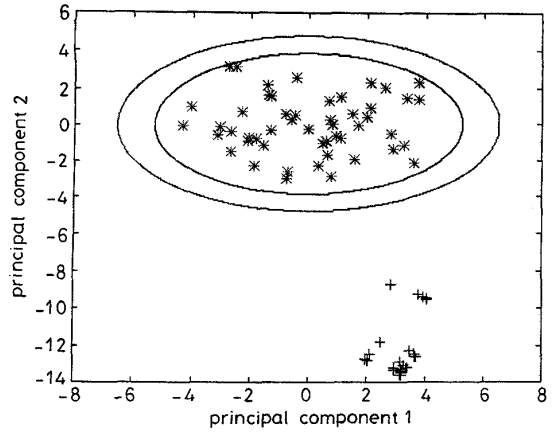

Fig. 4 *Scores plot for principal components 1 and 2*
* = nominal data, + = data from process malfunction

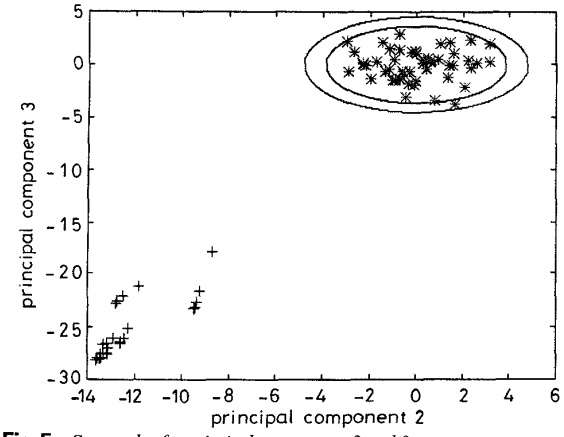

Fig. 5 *Scores plot for principal components 2 and 3*
* = nominal data, + = data from process malfunction

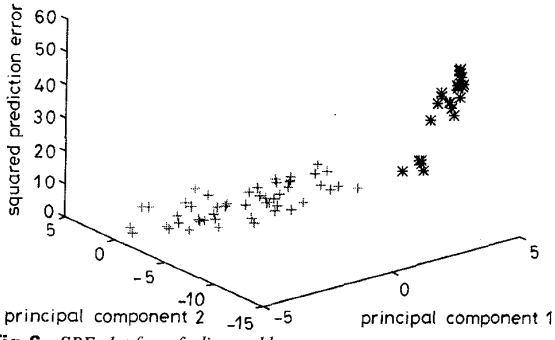

Fig. 6 *SPE plot for a fouling problem*
* = nominal data, + = data from process malfunction

The variables which primarily determine the direction of the individual principal components, t_1, t_2, etc., are those which have the 'largest' absolute loadings. If a process problem is identified, then the principal component in which non-conforming behaviour has been most clearly identified is believed to contain information which is reflective of why a process has moved away from the nominal operating region. For this problem, the identification of those variables which may be reflective of the problem is contained within principal component three and to a lesser extent principal component 2. Principal components one, two and three are assessed more closely:

$$t_1 = 0.184T_{in1} + 0.394T_{max1} + 0.114T_{out1} + 0.382T_{max2} + 0.273T_{out2} - 0.105T_{cin1} - 0.012T_{cin2} - 0.371X_{tmax1} - 0.358X_{tmax2} + 0.341F_{i01} + 0.325F_{s01} + 0.033F_{i02} - 0.114F_{s02} + 0.256Pres$$

$$t_2 = -0.212T_{in1} - 0.259T_{max1} - 0.428T_{out1} + 0.319T_{max2} + 0.101T_{out2} - 0.319T_{cin1} - 0.162T_{cin2} - 0.263X_{tmax1} - 0.283X_{tmax2} - 0.234F_{i01} + 0.327F_{s01} - 0.029F_{i02} + 0.384F_{s02} + 0.074Pres$$

$$t_3 = 0.266T_{in1} + 0.119T_{max1} - 0.349T_{out1} - 0.137T_{max2} - 0.418T_{out2} - 0.442T_{cin1} - 0.420T_{cin2} - 0.242X_{tmax1} + 0.148X_{tmax2} + 0.089F_{i01} - 0.142F_{s01} + 0.331F_{i02} - 0.025F_{s02} + 0.093Pres$$

The dominant variables for principal component one are variables 2, 4, 8, 9, 10 and 11, T_{max1}, T_{max2}, X_{tmax1}, X_{tmax2}, F_{i01} and F_{s01}, respectively, whilst principal component two primarily relates to variables 3, 4, 6, 11 and 13, T_{out1}, T_{max2}, T_{cin1}, F_{s01} and F_{s02}. Principal component three explains the majority of the variability arising from variables, 3, 5, 6, 7 and 12, T_{out1}, T_{out2}, T_{cin1}, T_{cin2} and F_{i02}. This breakdown clarifies the reasoning as to why it is not until the third component that the occurrence of fouling is conclusively recognised, the dominant variables are in fact all temperature measurements. In discussions with engineers, it was acknowledged that the combination of variables identified as determining the direction of principal component three are those which would change with the onset of fouling. It is therefore very important not to neglect lower-order plots of the scores.

Figs. 7 and 8 are known as contribution plots and represent the scaled variables for each of the past four data samples from the process i.e. the contribution from the current point is displayed along with the three previous calculated contributions to the score plot. Fig. 7 is a contribution plot typical of when the process is operating within the bounds of its nominal operating region, whilst Fig. 8 is an example of a plot where the process is experiencing operational problems. The main difference between the two plots is the scale of the contributions. For a process 'in control' the range of values will be as expected for good plant performance (+/– 2σ). In contrast, 'out-of-control' production will be identified by certain variables making larger contributions than those anticipated, with the prediction error being comparatively larger. By identifying those variable combinations which have experienced the greatest change, in conjunction with the operator's expertise, it is possible to relate a particular sequence of changes to a particular process malfunction. Variables 3, 5, 6 and 7, T_{out1}, T_{out2}, T_{cin1} and T_{cin2} respectively, exhibit the greatest deviations during the time span monitored for the fouling problem. This information allows the operator to interpret the process behaviour in terms of the original plant measured variables.

Instead of just looking at the process variables, it is often interesting to monitor the effect the quality of the resultant product. This is accomplished by using the latent variables of PLS to monitor the final quality.

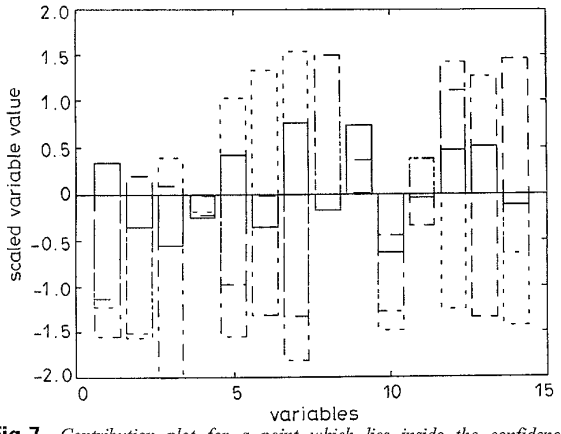

Fig. 7 *Contribution plot for a point which lies inside the confidence bounds*

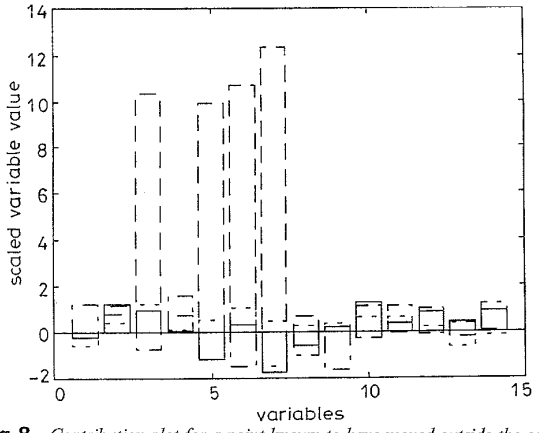

Fig. 8 *Contribution plot for a point known to have moved outside the control limits*

For this approach, four latent variables were identified using cross-validation as being sufficient to describe the process satisfactorily, Table 4.

Table 4: Cumulative percentage of variability explained in Y-block using projection to latent structures

Latent variables	1	2	3	4	5	6	7
Cumulative percentage of variability explained in Y	66.1	87.0	91.6	94.6	95.7	96.7	97.2

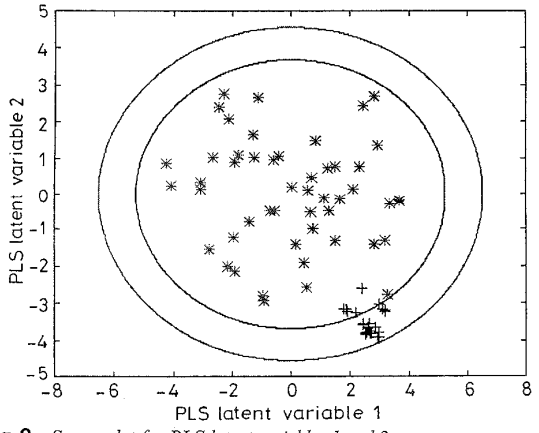

Fig. 9 *Scores plot for PLS latent variables 1 and 2*
* = nominal data, + = data from process malfunction

The plots for latent variables one and two, and three and four, are shown in Figs. 9 and 10, respectively. From the plot of latent variable one versus latent variable two, there is no clear indication of a process problem. However, when the lower order components are monitored, there is strong evidence that the quality of the final product is being degraded by the fouling problem. This time the problem is identified most clearly through latent variable 4.

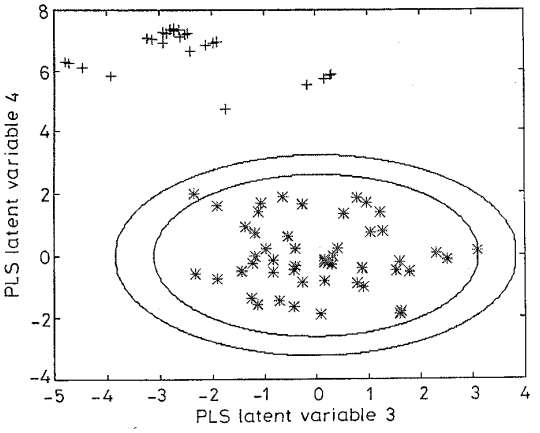

Fig. 10 *Scores plot for PLS latent variables 3 and 4*
* = nominal data, + = data from process malfunction

5 Multivariate SPC in batch processes

The monitoring of batch processes is strategically important to ensure their safe operation and the production of consistently high-quality products. Currently some of the difficulties contributing to the failure to provide adequate monitoring include, the lack of reliable online sensors for measuring product quality variables, the presence of nonlinearities and the absence of steady-state operation. Most of the existing industrial approaches for achieving consistent and reproducible results from batch processes are based upon the precise sequencing and automation of all the stages in a batch operation. Monitoring is usually confined to checking that these sequences are followed and that certain reactor variables, such as temperatures and reactant feedrates, are following acceptable trajectories.

Previous approaches to the monitoring of batch data has focused on the use of either fundamental mathematical models (based on state estimation methods) or detailed knowledge based models (using expert systems or artificial intelligence methods) to process the data. An alternative approach based on empirical models has been developed using multiway principal component analysis (MPCA) and multiway projection to latent structures (MPLS). MPCA and MPLS are extensions to the projection techniques of PCA and PLS and are based on the philosophy of compressing the process information into a few latent variables which are a linear combination of the original variables. The only information needed to develop an SPC monitoring procedure is a historical database of past successful batches. Batch data differs from continuous data in that the problem is now a three-way problem, the added level being that of time [13, 14].

The major issue which arises is how to handle the large number of measurements taken on the process which are themselves not independent. The measured variables are also autocorrelated in time. It is not sim-

ply the relationship between all the variables which is important, it is the entire past histories of the trajectories. The technique of principal components analysis can be used to reduce the dimensionality of the problem by projecting the information down onto a lower dimensional space which summarises the behaviour of the variables relative to one another and their time history during previous successful batches. A simple way to view MPCA is to consider opening out the three-way matrix into a two-way array, by placing each two-dimensional time slab (batches × variables) side-by-side and performing a standard PCA, Fig. 11.

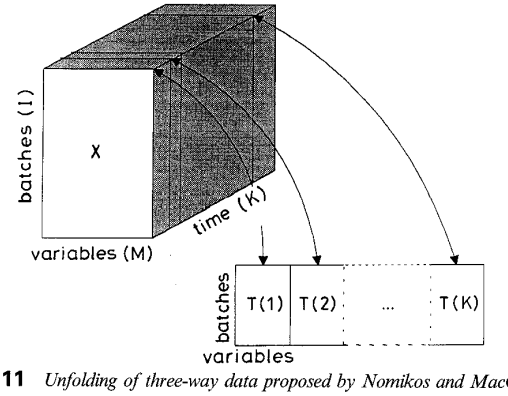

Fig. 11 *Unfolding of three-way data proposed by Nomikos and MacGregor*

Again it is possible to perform a version of PLS on the batch data, multiway partial least squares (MPLS). This approach is a combination of PLS and MPCA. The objective is to extract information from the process measurement variable trajectories that is more relevant to the final quality variables of the product.

An example of using MPCA to analyse batch performance is shown in terms of the monitoring charts constructed from a multiway principal components analysis on a nominal data set of 40 batches. A similar series of plots to those used for continuous processes monitoring can be used for batch monitoring. Fig. 12 illustrates a plot of the scores for the first two principal components. A number of unsatisfactory batches were also present in the data set as indicated by the movement of the scores outside the confidence bounds for specific batches (Fig. 13). Hotelling's T^2 was once again used to evaluate the bounds. Although monitoring charts are in their infancy for the monitoring of batch processes, preliminary results are indicative of a very powerful tool for industry to monitor their batch processes.

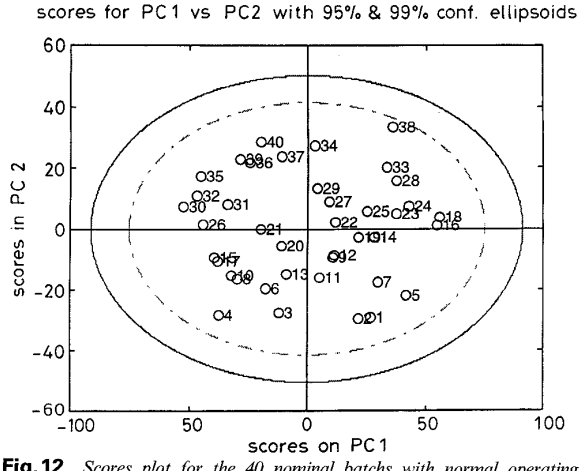

Fig. 12 *Scores plot for the 40 nominal batchs with normal operating region defined by 95% and 99% confidence limits*

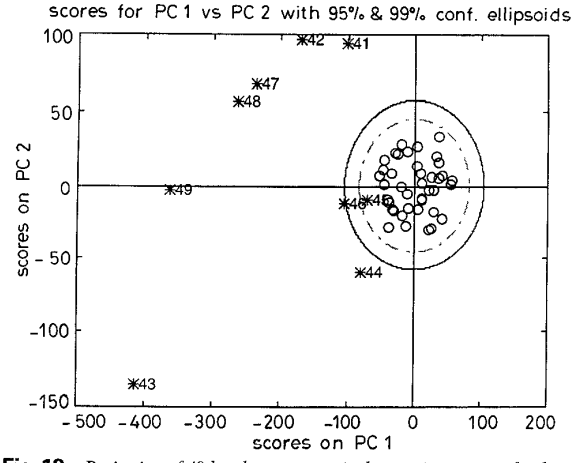

Fig. 13 *Projection of 49 batches onto nominal operating region of reduced variable space*

6 Nonlinear principal component analysis

Principal component analysis is now widely used for reducing the dimensionality of the problem to obtain an enhanced understanding of process behaviour. However, the linearity assumption can lead to misleading conclusions in the analysis of data from highly nonlinear processes. Conventional PCA is not effective when the variables are nonlinearly correlated and in such situations nonlinear principal component analysis (NLPCA) is more appropriate.

Nonlinear principal components analysis can be used in a similar way to PCA, that is data summarisation, data visualisation and data exploration for nonlinearly correlated data. The concept of extracting features from highly nonlinear data has been discussed by a number of researchers, with most techniques being based upon artificial neural networks. Of particular interest is the ability of the autoassociative neural network topology (Fig. 14) to provide a transformation into a nonlinear feature space [16]. The architecture of this network comprises five layers: an input layer, mapping layer, bottleneck layer, demapping layer and an output layer. The use of nonlinear features has been shown to successfully describe the underlying structure of nonlinear data, [17]. Extension of the autoassociative neural network architecture to allow the generation of nonlinear principal components requires the use of the statistical procedure of principal curves [18].

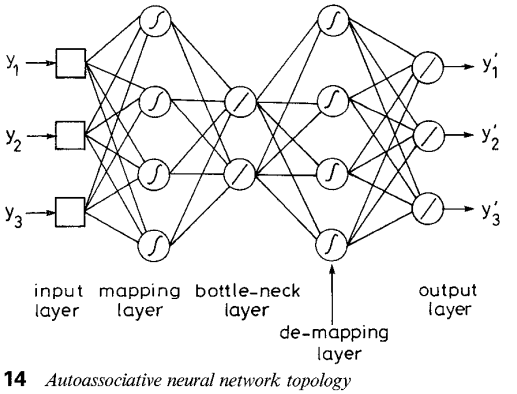

Fig. 14 *Autoassociative neural network topology*

The main difference between PCA and nonlinear PCA is the introduction of nonlinear mappings between the original and reduced dimensional-space. A commonly used nonlinear function is the sigmoid function

$$\sigma(x) = \frac{1}{1 - \exp^{-x}}$$

A linear principal component minimises the sum of the orthogonal deviations between a straight line and the data whilst the nonlinear approach summarises the data by a smooth curve (a principal curve). Principal curves are generalisations of principal components. The calculation of principal curves essentially contains two steps, a projection step and a smoothing step. The calculation is generally started with the principal component as the initial curve. In the projection step, data points are projected down onto the curve. Then, in the smoothing step, the curve is smoothed using techniques such as the locally weighted regression smoother [19] or kernel smoothers [20]. The procedure is iterated between the two steps until convergence results.

In principal component analysis, the principal loading vectors are used as a model to generate principal scores for new data. However, the principal curve procedure does not calculate any nonlinear loadings. In industrial process applications, it is desirable to have a nonlinear principal model which can be used to generate nonlinear principal components for new data. Dong and McAvoy [21] proposed using neural networks to learn nonlinear principal component models. Two networks are required. The first maps the input data (m dimensional) onto one-dimensional principal scores, evaluated from the principal curve. The second model defines the relationship between the principal scores and the m-dimensional corrected data set. If a nonlinear function can be used to express the curve, this function is equivalent to the principal loadings for linear PCA. When the data is projected onto the curve, indexes can be found which express the projected points. These are equivalent to the principal scores for linear PCA. In nonlinear PCA, just as with linear PCA, there is no response variable hence it is more suited to feature extraction than prediction.

Nonlinear PCA now enables the multivariate process performance monitoring of nonlinear processes. However, if linear principal component analysis reduces the dimensionality of the problem satisfactorily, then a linear approach to monitoring must be adopted since the interpretation of a linear technique is more straightforward for engineers and operators. As described previously, movement of the scores from regions of nominal operation, together with increasing squared prediction error values, identify changes in process operation that are other than 'common cause'. In some cases, especially with linear systems, the movement of the scores can be visibly different from those of the nominal operating region with different faults causing the scores to move off in different directions. The directions of process variable movement can be monitored by studying the projected score movement in the reduced score space. However, in some situations, especially in nonlinear systems, the movement of the score plots are not as distinctive. Consequently it can be difficult to locate the score movements since they are buried within the nominal scores region. Here we propose using an accumulated scores plot to distinguish between different fault situations, [22]. The accumulated scores are defined as follows:

$$A = \int_0^t (x - \bar{x})\, dt$$

where x is the nonlinear score, $\bar{x}$ is the mean of the nominal nonlinear score, and A is the accumulated nonlinear score. The accumulated scores for the nominal operating region cluster around zero with the accumulated scores during process malfunctions moving away from the score plot defining the nominal operating region. This approach is somewhat analogous to the cumulative sum (CUSUM) approach.

6.1 Application study — number 1

An industrial processing unit provides material for further processing into machinable components. The process is subject to raw material and energy supply changes as well as internal chemical changes. These changes affect the process operation and consequently the resulting material. The detection of process movement into different regions of operation, due to changes in process physics and chemistry, is vital to ensure the consistent production of material essential for subsequent processing phases. The process operators are known to have observed a number of different operating regions during production campaigns. It is difficult for the plant operators to control the movement between these different operating regions given the limited knowledge of the process operation and the multitude of reasons for the changes. If the drift from one operating region to another could be identified and the cause-effect relationship established, then corrective action could be taken, leading to more consistent production and enhancing subsequent manufacturing operations. Principal components analysis was applied and the distribution of the first three principal components examined.

Fig. 15 shows the plot of the scores for principal components 2 and 3 which were calculated using linear PCA. Interestingly once again it appears that the higher-order principal components (plot of the scores for principal component 1 and 2 is not shown) extract more information on the process behaviour indicating the possibility of at least two major regions of operation, if not three. Here, the points denoted by 'O' are for the first data set and the points denoted by '+' are for the second set of data. In comparison, the results of the nonlinear PCA analysis are shown in Figs. 16 and 17; plots of the nonlinear scores for principal component 1 versus principal component 2, and the nonlinear scores for principal component 1 versus principal component 3, respectively. Clearly six operating regions are now identifiable with the possibility of more during periods of operating region transition.

Fig. 18 shows the plot of the 'accumulated scores for nonlinear principal component 2' for the two data sets. Again there is clear indication of a change in the process operating conditions over the first period of monitored operation where the plant was observed to be moving into a difficult region of operation. Following changes to the plant operating conditions the process slowly moves back into a more acceptable region of operation near to the original nominal region. It is very encouraging that the nonlinear approach is able to provide conclusive evidence of the plant behaviour that had been perceived by the operating personnel.

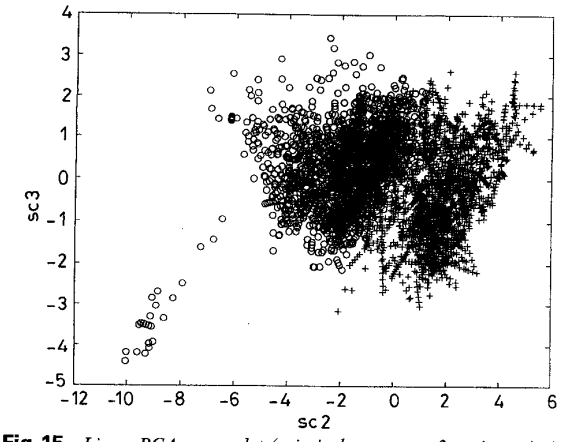

Fig. 15 *Linear PCA scores plot (principal component 2 against principal component 3)*
○ = first set, + = second set

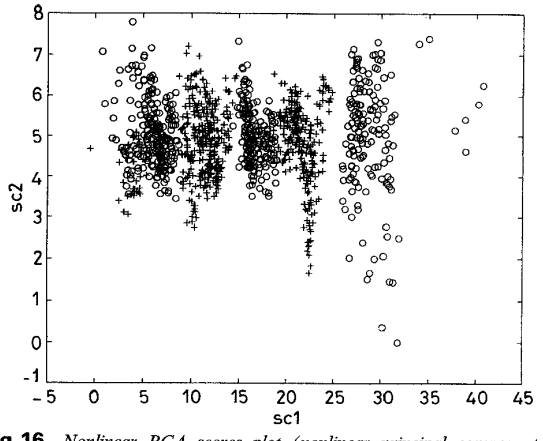

Fig. 16 *Nonlinear PCA scores plot (nonlinear principal component 1 against nonlinear principal component 2)*
○ = first set, + = second set

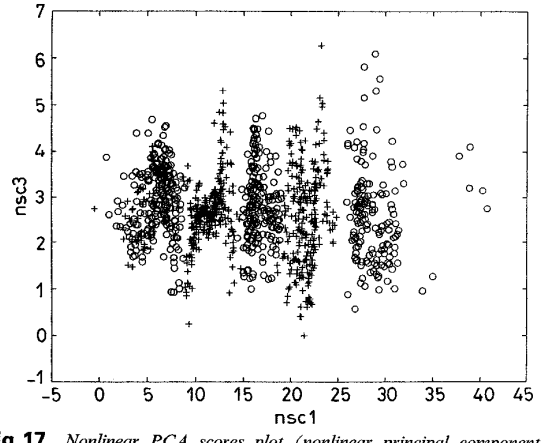

Fig. 17 *Nonlinear PCA scores plot (nonlinear principal component 1 against nonlinear principal component 3)*
○ = first set, + = second set

6.2 Application study — number 2

Returning to the problem described in section four, a linear principal component analysis was initially performed on the process data. Two principal components were found to only explain 51.1% of the variability in the X-block, with three principal components explaining 67%, and seven components explaining 95%. In this case, there is a substantial amount of variance which cannot be explained by the first two, or indeed three, principal components.

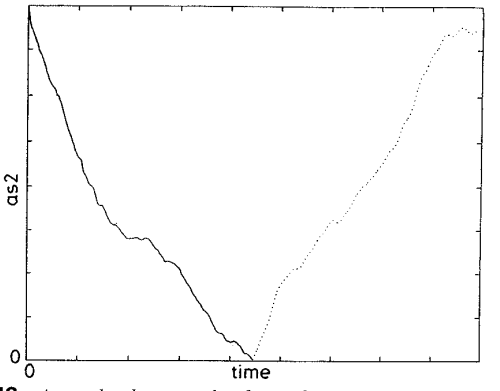

Fig. 18 *Accumulated scores plot for nonlinear principal component 2 showing different operating regions*
—— first condition, ····· second condition

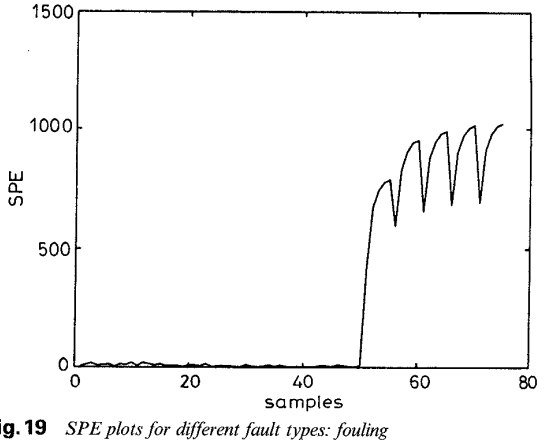

Fig. 19 *SPE plots for different fault types: fouling*

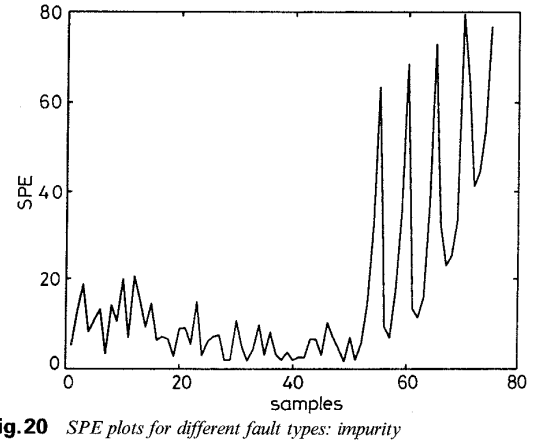

Fig. 20 *SPE plots for different fault types: impurity*

Nonlinear principal component analysis was then used to analyse the reactor data. Using two nonlinear principal components 74.8% of the variability in the data is explained, whilst three nonlinear principal components explain 89.8% of the data variance. This clearly indicates that nonlinear principal component analysis is able to extract more information from the data. A nonlinear principal component model with three principal components was subsequently devel-

oped. Using linear principal component analysis, the squared prediction error was calculated for the four different fault situations and the subsequent plots are presented in Figs. 19–22. The four fault situations are reactor fouling, reactive impurity, solvent problems, and combined reactor fouling and reactive impurities. In these plots, the first 50 data points represent the nominal operating conditions and the remaining 25 represent the phase when a process malfunction had occured within the reactor. It can be seen that all the various faults can be detected by monitoring the SPE.

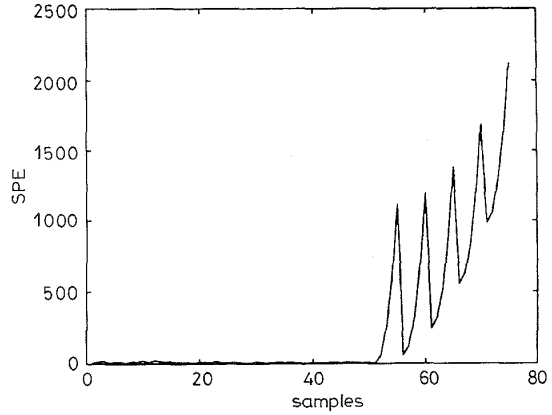

Fig. 21 *SPE plots for different fault types: solvent*

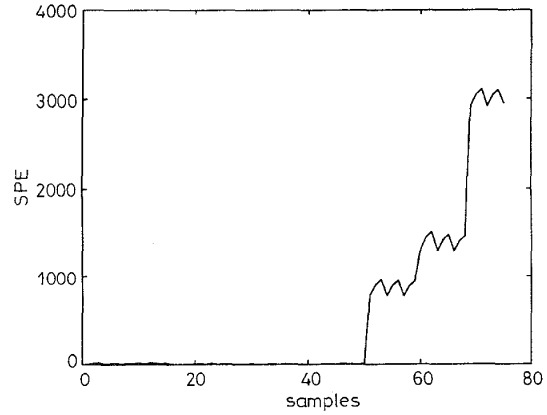

Fig. 22 *SPE plots for different fault types: combined*

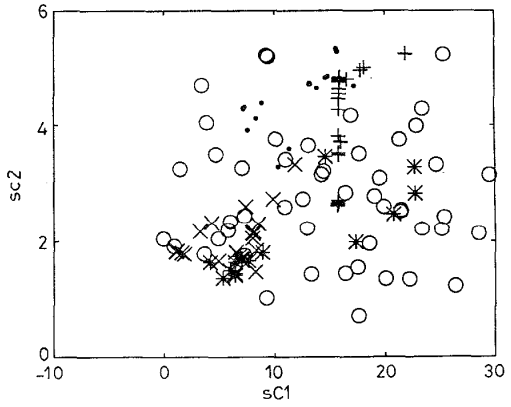

Fig. 23 *Nonlinear score plots for different fault types: fouling*
○ nominal * solvent
+ fouling ● combined
× impurity

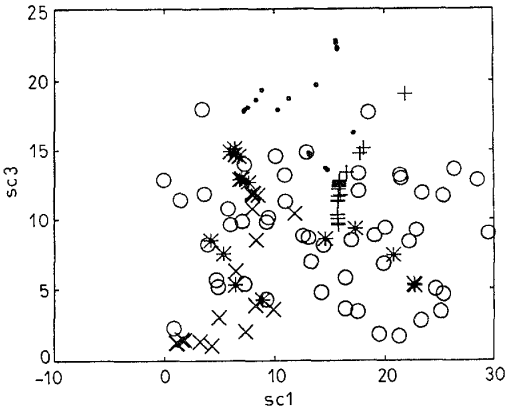

Fig. 24 *Nonlinear score plots for different fault types: impurity*
○ nominal * solvent
+ fouling ● combined
× impurity

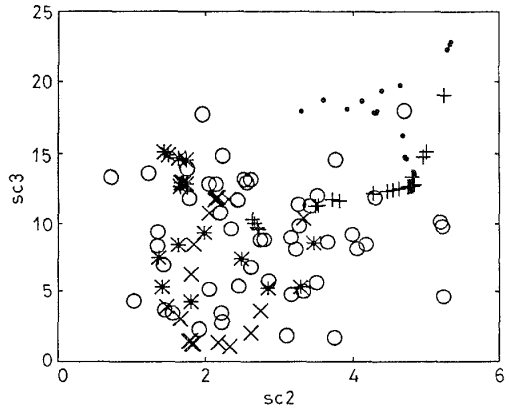

Fig. 25 *Nonlinear score plots for different fault types: solvent*
○ nominal * solvent
+ fouling ● combined
× impurity

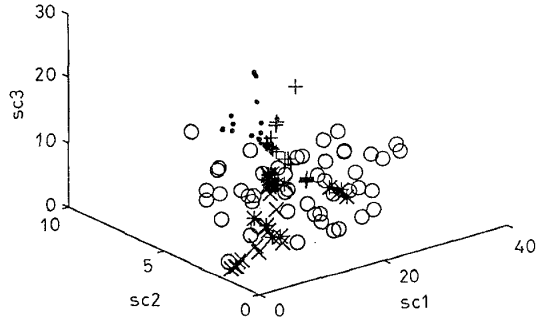

Fig. 26 *Nonlinear score plots for different fault types: combined*
○ nominal * solvent
+ fouling ● combined
× impurity

Score plots of the linear principal components have previously been used to used to identify the onset of the different faults. The motivation of this work is to identify the type of fault which has occurred. It is believed that different faults can result in different process measurement values which could be projected into different areas of the score space. In this study the use of nonlinear principal component score plots to localise the four different fault situations resulted in the distribution observed in Figs. 23–26. It can be seen here that the faults cannot be distinguished from the

scores produced for the nominal data set. In comparison, Figs. 27–30 show the plot of the accumulated scores. The different fault situations can now be clearly distinguished. Fig. 27 shows the trajectories of the accumulated scores for the nonlinear principal components 1 and 2. From Figs. 27 and 28 it can be seen that fouling is characterised by the movement of the scores in the north-east direction while the combined impurity and fouling faults are characterised by the score movement in the north-west direction. Both impurity and solvent problems result in the score moving in the south-west direction. However, they can be more clearly distinguished from Fig. 29 which plots the integration of the scores for the second and the third nonlinear principal components. Here the impurity problem corresponds to the score movement in a south-west direction while the flow problem corresponds to the score movement in a north-west direction. The results indicate that the accumulated scores can be effectively used to extract information resulting from the change in process operation and as a result can contribute to the localisation of different process faults. Fig. 30 shows a 3D plot of principal components 1, 2 and 3.

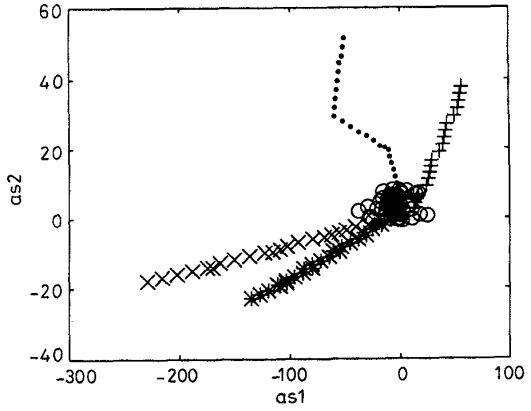

Fig. 27 *Accumulated nonlinear scores plot*
○ nominal * solvent
+ fouling ● combined
× impurity

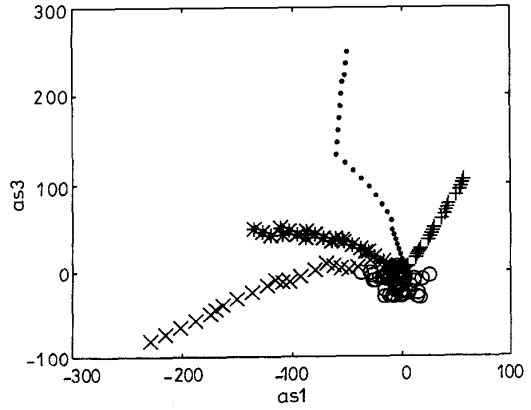

Fig. 28 *Accumulated nonlinear scores plot*
○ nominal * solvent
+ fouling ● combined
× impurity

7 Conclusions

The paper has provided an overview of linear and nonlinear multivariate statistical process control based on the statistical projection techniques of principal compo-

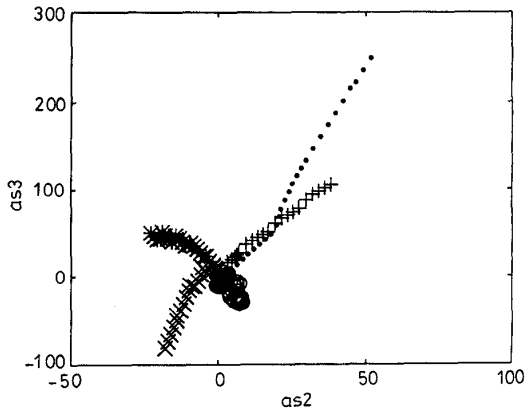

Fig. 29 *Accumulated nonlinear scores plot*
○ nominal * solvent
+ fouling ● combined
× impurity

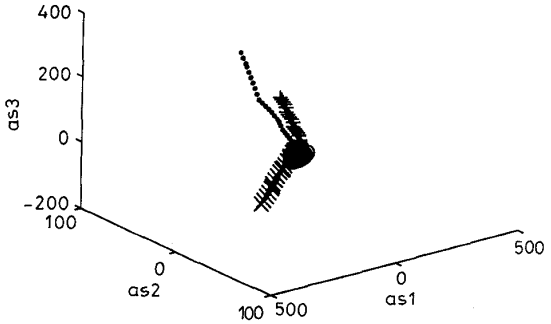

Fig. 30 *Accumulated nonlinear scores plot*
○ nominal * solvent
+ fouling ● combined
× impurity

nents analysis and to a lesser extent projection to latent structures. The methods take full advantage of the multivariate nature of the data and as a result are more likely to provide early warning of process faults than if univariate SPC techniques, which consider one variable at a time, were applied. The two areas of particular focus were the identification of the process moving away from the nominal operating region using either bivariate plots of the scores or alternatively squared prediction error plots, together with contribution plots to aid the identification of the variables reflective of the cause of the process moving out of control. It was also proposed that greater use should be made of the loadings in conjunction with the scores.

The use of nonlinear principal component analysis techniques was shown to be able to effectively extract significantly more information from nonlinearly correlated variables than conventional linear methods. It can provide a more effective reduction in the dimensionality of the problem than linear principal component analysis. The studies in this paper have demonstrated that nonlinear feature extraction from complex data can result in the explanation of more of the data variance than a similar number of linear principal components. The use of accumulated nonlinear scores has been proposed as an aid to the localisation of process faults. The accumulated scores of the nominal data set will cluster around zero with the accumulated scores of the 'fault-data' representing the directions of score movement due to various faults. It is believed that the further development of the techniques proposed here

will make an important contribution to the performance monitoring of complex nonlinear processes.

In both application studies it would have been very difficult, if not impossible, to detect the process problems using univariate methods. Multivariate SPC and plant performance monitoring offers significant strategic improvements in terms of product quality, product consistency, reduced plant down-time, reduced rework, the early detection of process malfunctions and increased use of existing plant assets. Clearly its implementation will involve a change in operational and management cultures commensurate with those associated with univariate SPC; it will incur financial and manpower investment. The question will then arise 'what will it cost to implement'? Perhaps the question that should be asked is 'can we afford not to do it'?

8 Acknowledgments

The authors would like to acknowledge the support of the Department of Engineering Mathematics and the Department of Chemical & Process Engineering, EC BRITE/EURAM Projects, Project no. BE - 7009/93 Intelligent Manufacture of Polymers, and project no. BE 8104/93, Intelligent Equipment & Process Monitoring for Consistent Finished Product Quality. Special thanks are also due to Prof. Costas Kiparissides and M. Papazoglou for providing the continuous reactor and batch reactor data.

9 References

1 WOLD, S., ESBENSEN, K., and GELADI, P.: 'Principal components analysis', *Chemometrics Intelligent Lab. Syst.*, 1987, **2**, pp. 37–52
2 HÖSKULDSSON, A.: 'PLS regression methods', *J. Chemometrics*, 1988, **2**, pp. 211–228.
3 WOLD, S.: 'Cross validatory estimation of the number of components in factor and principal components models', *Technometrics*, 1978, **20**, (4), pp. 397–404
4 KRESTA, J.V., MACGREGOR, J.F., and MARLIN, T.E.: 'Multivariate statistical monitoring of process operating performance', *Canadian J. Chem. Eng.*, 1991, **69**, pp. 35–47
5 MACGREGOR, J.: 'Statistical process control of multivariate processes', *Control Eng. Practice*, 1994, **3**, (3), pp. 403–414
6 MACGREGOR, J.F., NOMIKOS, P., and KOURTI, T.: 'Multivariate statistical process control of batch processes using PCA and PLS'. IFAC ADCHEM'94 conference on *Advanced control of chemical processes*, May 1994, Kyoto, Japan, pp. 525–530
7 MARTIN, E.B., and MORRIS, A.J.: 'Non-parametric confidence bounds'. Internal Report, Department of Chemical & Process Engineering, University of Newcastle, UK, 1995
8 MARTIN, E.B., MORRIS, A.J., XU, D.L., and ZHANG, J.: 'Non-linear PCA and non-linear PLS'. Internal Report, Department of Chemical & Process Engineering, University of Newcastle, United Kingdom, 1995
9 MARTIN, E.B., MORRIS, A.J., and PAPAZOGLOU, M.: 'Confidence bounds for multivariate performance monitoring charts'. IFAC workshop on *On-line fault detection and supervision in the chemical process industries*, 1995, Newcastle, United Kingdom, pp. 33–42
10 GELADI, P., and KOWALSKI, B.R.: 'Partial least squares regression: a tutorial', *Analytica Chimica Acta*, **185**, pp. 1–17
11 HOTELLING, H.: 'Multivariate quality control' *in* Eisenhart, Hastay and Wallis (Eds.) 'Techniques of statistical analysis' (1947)
12 MILLER, P., SWANSON, R.E., and HECKLER, C.F.: 'Contribution plots: the missing link in multivariate quality control'. 37th annual fall conference ASQC, 1993, Rochester, NY
13 NOMIKOS, P., and MACGREGOR, J.F.: 'Monitoring of batch processes using multi-way principal components analysis', *AIChE J.*, 1994, **40**, (8), pp. 1361–1375
14 NOMIKOS, P., and MACGREGOR, J.F.: 'Multivariate SPC charts for monitoring batch processes', *Technometrics*, 1995, **37**, pp. 41–59
15 KRAMER, M.A.: 'Non-linear principal component analysis using autoassociative neural networks', *AIChE J.*, 1991, **37**, pp. 233–243
16 IGNOVA, M., GLASSEY, J., MONTAGUE, G., MORRIS, A.J., and KIPARISSIDES, C.: 'Neural networks and non-linear SPC'. 3rd IEEE conference on *Control applications*, Glasgow, August 1994, pp. 1271–1276
17 HASTIE, T.J., and STUETZLE, W.: 'Principal curves', *J. Am. Stat. Assoc.*, 1989, **84**, pp. 502–516
18 CLEVELAND, W.S.: 'Robust locally weighted regression and smoothing scatter plots', *J. Am. Stat. Assoc.*, 1979, **74**, pp. 829–836
19 GASSER, T., and MULLER, H.G.: 'Kernel estimation of regression function' *in* 'Smoothing techniques for curve estimation' (Springer Verlag, 1979)
20 DONG, D., and MCAVOY, T.J.: 'Non-linear principal component analysis based on principal curve and neural networks'. Proc. ACC, 1994, Baltimore, USA, pp. 1284–1288
21 ZHANG, J., MARTIN, E.B., and MORRIS, A.J.: 'Non-linear principal components analysis and accumulated score plots'. Internal Report, Department of Chemical & Process Engineering, University of Newcastle, United Kingdom, 1995