ELSEVIER

J. Proc. Cont. Vol. 6, No. 6, pp. 349–358, 1996

0959-1524/96 $15.00 + 0.00

PII: S0959-1524(96)00010-8

Non-parametric confidence bounds for process performance monitoring charts

E. B. Martin* and A. J. Morris†

*Centre for Process Analysis, Chemometrics and Control, *Department of Engineering Mathematics, †Department of Chemical and Process Engineering, University of Newcastle, Newcastle Upon Tyne, NE1 7RU, UK*

Statistical Process Control (SPC) provides a tool for achieving and maintaining product quality. In today's climate of major data monitoring campaigns there has been an increase in interest in the multivariate statistical projection techniques of principal components analysis and projection to latent structures for process performance monitoring. Within univariate SPC, techniques for identifying when a process is moving out of control are well established. Similar guidelines are required for multivariate statistical process control (MSPC). Two approaches will be discussed – Hotelling's T^2 statistic and a new approach, the M^2 statistic. Both approaches will be illustrated by application to a high pressure low density polyethylene tubular reactor and to a batch methyl methacrylate polymerisation reactor. Copyright © 1996 Elsevier Science Ltd

Keywords: multivariate statistical process control, fault detection and diagnosis, confidence bounds

Statistical process control charts, such as the Shewhart chart ($\bar{X}$ and range), CUSUM (cumulative sum) plot and EWMA (exponentially weighted moving average) chart are well established statistical tools for monitoring the behaviour of a process based on a small number of quality variables. These charts compare current process performance against process behaviour when the product being produced was known to conform to the customer's/factory specification. The detection of when the process is moving outside the 'in-control' limits is identified using well established techniques which assume the data is both normally distributed and independent, in conjunction with engineering judgement. This approach has been shown to be successful in a wide range of manufacturing industries for diagnosing non-conforming production.

Recently, the manufacturing industries have embarked upon major data collection programmes. In the process industries two types of information are collected, quality measurements (colour, texture, strength, material properties, etc.) and process information (temperature, pressure, flow rates, etc.). Compared with the process measurements, where perhaps hundreds of variables can be monitored, only a limited number of quality variables are recorded and at a much slower, and possibly variable frequency. Applying univariate monitoring techniques will result in the majority of the information contained within the data being discarded.

Additionally, univariate SPC systems only allow disturbances related to individual quality measurement sources to be detected. Thus, the information contained within the interactions between variables, which is so important in complex processes such as those found in the chemical, biochemical, food and materials processing industries, is ignored. These limitations can be addressed through the application of multivariate statistical process control (MSPC). The basis of MSPC is the multivariate projection techniques of principal components analysis and projection to latent structures. The philosophy of these techniques is to reduce the dimensionality of the problem by forming a new set of latent variables. If the variables are highly correlated, then the process can be defined in terms of a much reduced set of latent variables, which are a linear combination of the original variables. Principal component analysis (PCA) is an analysis tool which reduces the dimensionality of a single data matrix. The principal components generated from the analysis form the cornerstone of the multivariate statistical process control charts.

More recently, and as a result of economic pressures and changes in market perceptions of product quality and reliability, the necessity has arisen for the more slowly monitored 'quality' measures to be predicted from the more rapidly recorded process variables. This objective can be achieved through the application of regression type approaches, i.e. the identification of a linear relationship between the quality information and

A revised version of a paper originally presented at the IFAC Workshop on On-line Fault Detection and Supervision in the Chemical Process Industries held at Newcastle, UK in June 1995

the process variables. In this way, the final quality may be predicted in advance of it becoming available from the quality control laboratory. This can result in the prediction of the quality measure with increased frequency, and hence closer monitoring and tighter control can be achieved. Projection to latent structures, (PLS), utilises the information from both the process and quality variables and treats each set of information as dependent, thus reflecting the true nature of the process. Once again a new set of variable combinations can be derived and these form the basis of performance monitoring charts[1,2].

In univariate SPC, the control limits are constructed based upon the assumption of normality and independence. A similar approach can be adopted for multivariate charts, but the underlying assumptions appear to be restrictive and inappropriate for complex processes. A novel approach for constructing control limits is described based upon the density of the data. The methodology combines the techniques of the standard bootstrap and kernel density estimation.

This paper presents a brief overview of multivariate statistical process control for process monitoring using both the projection techniques of principal components analysis and projection to latent structures. A description of how confidence bounds are calculated using the statistical distribution, Hotelling's T^2, is presented and the limitations of this approach are identified. A new technique[3,4] is presented which we will call the M^2 statistic, which provides a non-parametric approach to the calculation of confidence bounds. The two methodologies are then compared by application to process fault detection in a tubular high pressure continuous polyethylene reactor and to a batch methyl methacrylate (MMA) polymerisation reactor.

Multivariate statistical process control

Effective and efficient utilisation of the large amounts of data collected by computers has become of increasing and strategic importance to a wide range of manufacturing industries. The main focus has been to analyse this information to improve the quality of the final product through the empirical modelling of the process and the extraction of process features from the data. The sheer volume and nature of the data, that is noisy measurements, missing observations/samples, variables not independent and small signal to noise ratios, necessitates the adoption of multivariate statistical projection techniques to gain a clearer understanding of the process behaviour and the detection of the occurrence of special or assignable events.

Principal components analysis[5] (PCA) is a tool which reduces the dimensionality of the problem by defining a series of new variables, principal components, which explain the maximal amount of variability in the process data. The new variables are a linear combination of the original variables which are constrained to be orthogonal. The first principal component is that linear combination of the original variables which explains the greatest amount of variability ($\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$) in the $\mathbf{X}$ matrix. The loadings, $\mathbf{p}_i$, define the direction of greatest variability, and the score vector, $\mathbf{t}_i$, represents the projection of each object on to $\mathbf{p}_i$. The second principal component is defined to be orthogonal to the first and explains the next greatest amount of variability, i.e. $\mathbf{t}_2 = \mathbf{E}_1\mathbf{p}_2$ where $\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T$. One proceeds in this manner until m principal components are obtained.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T = \sum_{i=1}^{m} \mathbf{t}_i \mathbf{p}_i^T \tag{1}$$

One of the features of PCA is that the less important components often describe the noise in the data. If the process variables are collinear, 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.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = \sum_{i=1}^{K} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} \tag{2}$$

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[6]. In practice, two or three principal components are frequently sufficient for multivariate SPC with the principal components generated from PCA forming the cornerstone of the multivariate statistical process control charts.

Increasingly, there is major interest in inferring the final product quality from the process data. Multiple linear regression, although often used, is inappropriate since the solution can be numerically unstable with small perturbations in the data causing potentially large changes in the regression coefficients. Alternative regression approaches such as ridge regression and regularisation methods solve the problem of singularity, but do not reduce the dimensionality of the problem. This latter point is of core importance in process performance monitoring, otherwise process operators and plant management would be overloaded with lots of meaningless or repeated information. Principal components regression (PCR) redresses both the singularity and dimensionality problem, but still treats the quality variables as though they were independent. Such an approach can be questionable in processes with highly correlated quality measures.

The technique of projection to latent structures[7,8] (PLS) goes one step further and utilises the information on both the process and quality variables, and treats both sets of information as though they were dependent. Projection to latent structures is a regression method based upon projecting the information in high dimensional space ($\mathbf{X}$,$\mathbf{Y}$) down on to a lower dimensional space defined by a small number of latent vari-

ables $\mathbf{t}_1$, $\mathbf{t}_2$,...,$\mathbf{t}_a$. These new latent variables summarise the important information contained in the original data sets and can form the basis of monitoring charts. Given a set of information on m process variables, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_m]$, and k quality variables, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2,\ldots\mathbf{y}_k]$, a factor from the $\mathbf{X}$ data (which is a combination of the m process variables) and the $\mathbf{Y}$ data (which is a combination of the k quality variables) are evaluated, $\mathbf{t}_1$ and $\mathbf{u}_1$, respectively:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \text{ and } \mathbf{u}_1 = \mathbf{Y}\mathbf{q}_1 \tag{3}$$

These equations are referred to as the outer relations for the $\mathbf{X}$ block and $\mathbf{Y}$ block, respectively. The vectors $\mathbf{w}_i$ and $\mathbf{q}_i$ are called factor weights. PLS finds the factor weights in such a way that $\mathbf{t}_i$ and $\mathbf{u}_i$ are most closely linearly correlated to one another. A linear regression is then performed between the first pairs of factors $\mathbf{t}_i$ and $\mathbf{u}_i$.

$$\mathbf{u}_1 = \mathbf{b}_1\mathbf{t}_1 + \mathbf{e}_1 \tag{4}$$

This relationship is termed the inner relation of the $\mathbf{X}$ and $\mathbf{Y}$ blocks, and it links the two blocks together through the latent variables $\mathbf{t}_i$ and $\mathbf{u}_i$. The third stage of the algorithm is the regression of $\mathbf{X}$ on its factor $\mathbf{t}_i$ and $\mathbf{Y}$ on its factor $\mathbf{u}_i$.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{E} \text{ and } \mathbf{Y} = \mathbf{u}_1\mathbf{q}_1^T + \mathbf{F} \tag{5}$$

where $\mathbf{E}$ and $\mathbf{F}$ are the residuals. The coefficients $\mathbf{p}_i$ and $\mathbf{q}_i$ are found using least squares regression, once again.

$$\mathbf{p}_1^T = (\mathbf{t}_1\mathbf{t}_1)^{-1}\mathbf{t}_1\mathbf{X} \qquad \mathbf{q}_1^T = (\mathbf{u}_1\mathbf{u})^{-1}\mathbf{u}_1^T\mathbf{Y} \tag{6}$$

The above three steps are repeated with the residuals $\mathbf{E}$ and $\mathbf{F}$ replacing $\mathbf{X}$ and $\mathbf{Y}$, respectively. Up to m pairs of factors can be derived, but in a similar way to PCA, only the first 'a' contain the information relevant to the operation of the process. Cross-validation or other related approaches can be used to select the number of factors[5]. The major advantage of projection to latent structures is that although it is a linear methodology, because of the manner in which the analysis is performed it is capable of successfully modelling slightly non-linear situations. This makes the technique extremely versatile.

Performance monitoring charts

The implementations of multivariate SPC and the associated charts are similar to those for univariate SPC. The ideas described within the paper are equally applicable to PLS based performance monitoring, but here attention will initially focus on PCA. A reference region is first identified using PCA. It is based upon historical data which was collected when within specification, product was being manufactured and only common cause variation was present. Future process behaviour is then referenced against this 'nominal' or 'in-control' representation[1,2]. The basis of the success of this approach is the recognition that many of the measurements taken on processes are highly correlated and thus different combinations of the variables define the same underlying events. Consequently, it can be assumed that when process production is within predefined specification limits, the dimensionality of the process can be reduced to a few latent variables. For a plant with an underlying dimension of two, the information can be presented in a bivariate plot. The axes of the monitoring chart being the first two principal components with each point located via its score ordinates. Adopting a similar approach to that for univariate charting methodology, nominal operating regions can be defined based on standard statistical distributional theory. The same rules apply for identifying when a process is out, or moving out of control:

(i) two points lie outside the warning limits,
(ii) one point lies outside the action limits or
(iii) seven points consistently increase or decrease.

The projection techniques of PCA and PLS depend critically upon the scales used to measure the variables. If we consider a set of multivariate data where the variables, $\mathbf{X}_1$, $\mathbf{X}_2$,..., $\mathbf{X}_n$ are of completely different types, for example pressures, temperatures, flow rates, etc. then the structure of the latent variables derived from this data set will depend essentially upon the arbitrary set of units of measurement. 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; mean-centre the data; or alternatively, scale the variables to zero mean and unit variance, i.e. normalised variables. There are no clear-cut rules as to which form of scaling should be adopted – it is entirely problem dependent. Currently, it may be beneficial to examine the results using different scaling regimes.

Confidence bounds

Once a model has been developed which is reflective of the nominal operating region, it is necessary to detect any departure of the k dimensional process from its standard behaviour. Typically in univariate SPC, these decisions are based on the confidence limits. The construction of confidence limits requires knowledge of the underlying distribution of the data. However, in complex systems the only information available consists of samples from distributions rather than exact knowledge of the distributions involved. Consequently, it is necessary to assume an underlying distribution or alternatively, use a density based approach.

An idealised form of the problem under consideration can be formulated as follows. Given a sample $\mathbf{X}_1, \mathbf{X}_2,\ldots, \mathbf{X}_n$ from an unknown density, f, we are required to construct a sequence of independent observations $\mathbf{Z}_1, \mathbf{Z}_2,\ldots$ from f. This is impossible to achieve exactly in practice because full information about f is not available. The observations in the sample $\{\mathbf{X}_i\}$ and the required realisations $\{\mathbf{Z}_j\}$ will be assumed to be k-dimensional vectors.

There are two potential approaches to the problem. A parametric form for f could be assumed, such as the normal distribution with unknown parameters; the sample $\{\mathbf{X}_i\}$ is then used to estimate the unknown parameters, and a standard simulation method is then used to generate the required simulated observations. Alternatively, the realisations $\{\mathbf{Z}_j\}$ are generated directly by successive random sampling, with replacement from the sample $\{\mathbf{X}_1, \mathbf{X}_2,\ldots,\mathbf{X}_n\}$, bootstrapping. This latter approach has the advantage of freeing the procedure from the parametric assumptions, but the disadvantage of making it impossible for any value to occur in the simulated data that has not occurred exactly in the original sample $\{\mathbf{X}_i\}$.

It is therefore natural to consider an intermediate approach based on density estimation. The observations $\mathbf{X}_1, \mathbf{X}_2,\ldots, \mathbf{X}_n$ can be used to construct a non-parametric estimate $\hat{f}$ of the density f, and then as many independent realisations as required can be drawn from $\hat{f}$. Depending on the problem it is occasionally desirable to simulate not from $\hat{f}$ itself, but from a version transformed to have the same mean vector and covariance matrix as the observed data.

Hotelling's squared distance

A widely used approach, as mentioned above, is to assume that the underlying k-dimensional process is normally distributed. Let $\mathbf{X}_i = \{\mathbf{X}_{i1}, \mathbf{X}_{i2},\ldots, \mathbf{X}_{ik}\}$ represent a k-dimensional vector of principal components calculated on a process at time point i. The value $\mathbf{X}_{ij}$ represents the score for the j^{th} principal component. When the process is in control, the $\mathbf{X}_i$ will be independent and we shall assume that they follow a multivariate normal distribution with mean vector μ and covariance matrix Σ. The values μ and Σ are estimated from a reference sample having n observations using $\bar{\mathbf{X}}$ and $\mathbf{S}$, the mean of vector μ is zero. Based on the k-dimensional vector of principal components, it is possible to determine whether the process is in control by calculating Hotelling's (1947) square distance[9], T_0^2, for the principal components of primary interest.

$$T_0^2 = n\,(\mathbf{X}_i - \bar{\mathbf{X}})^{\mathrm{T}}\mathbf{S}^{-1}\,(\mathbf{X}_i - \bar{\mathbf{X}}) \tag{7}$$

T_0^2, is distributed as the statistic $(n-1)kF/n(n-k)$, where F has a central F-distribution with k and $n-k$ degrees of freedom. Using this relationship an out-of-control limit can be established with a $100\alpha\%$ chance of a false alarm, α is typically 0.01, i.e. one out of every 100 alarms will be spurious. An out of control signal is given by:

$$T_0^2 > \frac{(n-1)kF_{k,n-k;\alpha}}{n(n-k)} \tag{8}$$

Confidence bounds for a scores monitoring chart are comparatively straightforward to evaluate. However, the major drawback of such an approach is the underlying assumption of multivariate normality. For many industrial processes the authors have found through tests for multivariate normality on the scores, rarely have they been identified as following a multivariate normal distribution. Information fed back to the operators based on confidence limits evaluated using Hotelling's T^2 approach, therefore may well be inaccurate and misleading.

Likelihood-based confidence regions

An alternative approach to defining the nominal operating region is to construct a likelihood based confidence region for a vector parameter θ of length k, using the bootstrap and non-parametric density estimation[3,4,10]. In MSPC, the vector parameter θ will be the zero vector. In practice, the aim is to continue to produce product around the mean of previous good production.

Before developing a likelihood based confidence region, some basic notation is briefly introduced. Let $\hat{\theta}$ be an estimate of an unknown vector parameter of length k, based on a sample X of size n. Let V be the asymptotic variance matrix of $n^{1/2}(\hat{\theta} - \theta)$, and $\hat{V}$ is an estimate of V. V is seldom known and so procedures for constructing confidence regions are usually based in some way on the distribution of Y:

$$Y = n^{1/2}\hat{V}^{-1/2}(\hat{\theta} - \theta) \tag{9}$$

A likelihood based confidence region can be calculated by first drawing B independent samples from $\mathbf{X}$, the score vectors, using the non-parametric bootstrap, i.e. sampling directly from the original vectors, with replacement. The variable $\mathbf{Y}_i$ is then calculated for each of the bootstrap samples, Equation 9.

A density estimator, for example kernel density estimator, can then be fitted to the distribution of the $\mathbf{Y}_i$'s. A univariate kernel estimator with kernel K is defined by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{10}$$

where h is the window width, also called the smoothing parameter or bandwidth. The kernel estimator is a sum of the 'bumps' placed at the observations. The kernel K defines the shape of the bumps while the smoothing parameter determines their width. An illustration of a kernel density estimator is given in *Figure 1* where the individual bumps are shown as well as the estimate $\hat{f}$.

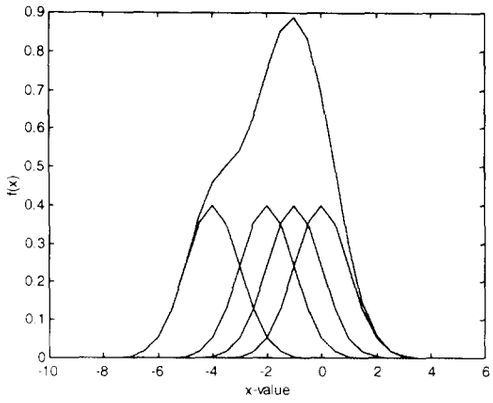

Figure 1 Kernel estimates showing individual kernels

In the k-dimensional case, let K be a simple, known k-variate density function, such as the standard normal density. For a given smoothing parameter $h > 0$, a multivariate kernel density estimate is defined to be:

$$\hat{f}(\mathbf{X}/h) \equiv \frac{1}{nh^k}\sum_{i=1}^{n} K\left\{\frac{(\mathbf{X}-\mathbf{Z}_i)}{h}\right\}, \mathbf{X} \in \Re^k \tag{11}$$

The distribution with density $\hat{f}$ is an approximation to the common distribution of the $\mathbf{Z}_i$'s conditional on $\mathbf{X}$, the original data set.

A number of alternative measures exist to estimate h, the window width or smoothing parameter. The problem of choosing how much to smooth is of crucial importance in density estimation. The effect of varying the window width is as follows, as the value of h tends to zero, the density estimate becomes a sum of Dirac delta function spikes at the observations, while as h becomes large, all detail, spurious or otherwise, is obscured. The appropriate choice of smoothing parameter is influenced by the purpose for which the density estimate is to be used. A natural choice for choosing the smoothing parameter is to plot out several nominal operating regions and select the estimate which is most in accordance with one's prior ideas about the density.

An alternative approach is to assume some underlying distribution, e.g. the standard k-variate normal density and estimate a smoothing parameter based upon this assumption. A third approach described in the literature[11] is to use likelihood cross-validation. However, one difficulty with such an approach is the effect of outliers and their influence on the window width. In this work the smoothing parameter was selected by least squares cross-validation[12]. This approach is completely automatic and computationally not excessive. Given any estimator $\hat{f}$ of a density f, the integrated squared error can be written as:

$$\int(\hat{f}-f)^2 = \int \hat{f}^2 - 2\int \hat{f}f + \int f^2 \tag{12}$$

The last term does not depend upon $\hat{f}$, and so the ideal choice of window width (in the sense of minimising integrated square error) will correspond to the choice which minimises the quantity R defined by:

$$R(\hat{f}) = \int \hat{f}^2 - 2\int \hat{f}f \tag{13}$$

The basic principle of least squares cross-validation is to construct an estimate of $R(\hat{f})$ from the data themselves and then to minimise this estimate over h to give the choice of window width. The term $\int \hat{f}^2$ can be found from the estimate $\hat{f}$. The density estimate, $\hat{f}_{-i}$, constructed from all the data points except X_i, is defined as:

$$\hat{f}_{-i}(x) = (n-1)^{-1}h^{-k}\sum_i\sum_{j} K^*\left\{h^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\} \tag{14}$$

We can now define a function M_0,

$$M_0(h) = \int f^2 - 2n^{-1}\sum_i \hat{f}_{-1}(\mathbf{X}_i) \tag{15}$$

The advantage of the score M_0 is that it depends only on the data. The idea of least squares cross-validation is to minimise the score M_0 over h.

To express the score M_0 in a form which is more suitable for computation, first define $K^{(2)}$ to be the convolution of the kernel with itself. If, for example, K is the standard Gaussian kernel, then $K^{(2)}$ will be the Gaussian density with variance 2. Expressions for $\int \hat{f}^2$ and $n^{-1}\Sigma\hat{f}_{-1}(\mathbf{X}_i)$ can be developed and substituted into expression (15). A very closely related score function $M_1(h)$, easier to calculate than M_0 is given by

$$M_1(h) = n^{-2}h^{-d}\sum\sum K^*\left\{h^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\} + 2n^{-1}h^{-1}K(0) \tag{16}$$

where the function $K^{(*)}$ is defined by

$$K^*(t) = K^{(2)}(t) - 2K(t) \tag{17}$$

Our ultimate aim is not so much to fit the density $\hat{f}$ to the entire data set, but to use $\hat{f}$ as a basis for defining the limits of the nominal operating zone such that precisely 1% of the vectors $\mathbf{Y}_i$ lie outside the nominal operating region. Such a region generally falls in an area where the data vectors are relatively sparse, and a kernel estimate of $\hat{f}$ which is 'optimal' from the point of view of minimising a global measure of loss can be unduly bumpy. A density estimate which is slightly over-smooth relative to that produced by techniques such as squared-error cross-validation, can result in better contours in the tails.

On-going studies are addressing the question of the use of the smoothed bootstrap as opposed to the stan-

dard bootstrap. The samples constructed from F_n in the bootstrap simulations will have some strange properties. All the values are drawn from the original values and will contain some values repeated a number of times. An approach that does not lead to samples with these properties is the smoothed bootstrap[3]. Here the simulations are constructed not from F_n, but by using an algorithm to simulate from a smoothed version of F_n. If $\hat{F}$ is the distribution function of the density estimate $\hat{f}$, then the effect of the smoothed bootstrap will be used to estimate $\rho(F)$ by $\rho(\hat{F})$. Whether $\rho(F)$ is better estimated by $\rho(F_n)$ or $\rho(\hat{F})$ is currently under investigation.

Engineering implications of the M^2 statistic

In selecting a reference or nominal data set the inherent assumption is that all production defined by this data set is acceptable to the customer. With Hotellings T^2 statistic, by definition, a proportion of the data will theoretically lie outside the confidence bounds, e.g. for a sample of size 20, one point would lie outside the 95% confidence bound if the data arose from a normal distribution. In the proposed M^2 approach, there is built in flexibility to the technique which enables the user to select the proportion of data, i.e. production which might be borderline and which should lie outside the bounds. In practice if all the data is from valid production then it could be argued that the bounds should enclose all the nominal data. Selection of the bounds then becomes a production engineering issue as opposed to a statistical phenomena.

Application to a low density polyethylene (LDPE) reactor

We consider as an example process, the monitoring of a continuous, multi-section, tubular reactor used in the manufacture of low density polyethylene[13]. The operating conditions in large continuous polymer reactors influence the many molecular properties of the polymer being produced. These properties tend to be difficult to measure, even off-line in the laboratory. However, other variables such as melt flow index and density can be measured, and are related to the fundamental polymer properties albeit in a complex way. In addition, measurements of reactor temperature along its length, reactant concentrations, and reactant and coolant flows are usually available on-line.

Fourteen process variables (reactor temperature profiles on each reactor section, initiator and solvent flowrates, coolant flows and temperatures and reactor pressure) are frequently monitored. Five polymer quality variables (weight and number average molecular weights, M_w and M_n, long and short chain branching properties, LCB and SCB, and cumulative conversion) are available from infrequent off-line measurements. The process produces different grades of polymer. The manufacture of one particular grade is considered here. Data from this operating region are available and provide a nominal (reference) data set of good production defined by a set of fifty samples. In practice, such nominal data sets would be available from process data bases of monitored plant variables or from designed experiments. Data are also available during which the process was subject to various malfunctions – the effects of reactor wall fouling which affects the reactor wall heat transfer coefficient, initiator impurity changes at ppm levels which modify the rate of reaction and the chain transfer agents in the monomer, and solvent feed rate which affects the molecular development. These are considered as process faults which need to be detected as early as possible through the monitoring of process operation. We focus in this paper on the first problem, i.e. reactor wall fouling.

The on-line observed variables and the polymer property variables were amalgamated into an 'input' matrix [**X**] and an 'output' matrix [**Y**], respectively. PCA and PLS representations were developed from the nominal data set. Using three latent variables, approximately 80% of the variation in the quality variables can be explained using PLS, whilst five principal components were required to explain 80% of the variability in the **X**-matrix alone using PCA. Multivariate monitoring charts were formed initially using the first four principal components. *Figures 2* and *3* show the scores plots for principal components one and two, and principal components

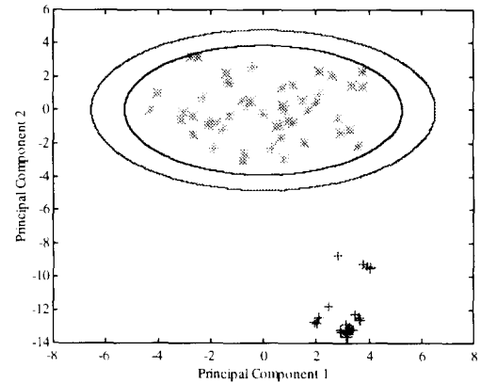

Figure 2 Scores plot for principal component 1 and 2 with nominal region defined using Hotelling's T^2 statistic

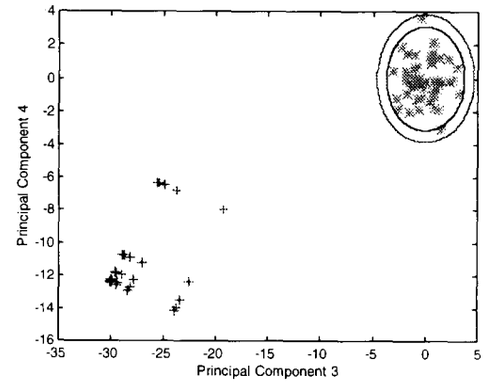

Figure 3 Scores plot for principal component 3 and 4 with nominal region defined using Hotelling's T^2 statistic

three and four, respectively (* defines the nominal data set). As new data are monitored the effect of a process malfunction (in this case a fouling problem) is clearly observed, '+'. The effect of fouling on the reactor temperature profile is observed in both figures. Although the occurrence of fouling can be identified by principal component two, the evidence of a problem is more clearly highlighted by principal component three.

The confidence bounds plotted in *Figures 2* and *3* are those calculated using Hotelling's T^2 statistic. Inspection of the region within the confidence bounds indicates quite large areas where no production information on the process was available. But by definition, any production in these regions would have been perceived to be satisfactory by the operators. The question therefore arises as to whether the Hotelling's bounds are appropriate in a production engineering context where data are not independent and do not arise from a multivariate normal distribution. By making such assumptions, the bounds will tend to be wider than required as indicated from the scores plots reported here.

In comparison, *Figure 4* shows the principal component scores plot for PC1 and PC2 with the confidence bounds calculated using the M^2 multivariate statistic. Clearly the bounds are defined with respect to the underlying density of the data. Here, regions of production data sparsity are significantly reduced, especially on the periphery of the data cluster.

By selecting the smoothing parameter well in excess of the optimal value selected by least-squares cross-validation, a comparable result to that produced by Hotelling's T^2 can be obtained. This is shown in *Figure 5*. However, because of the flexibility of placing the contour bounds to incorporate the statistically optimal number of points, the final result is still of potentially greater interest and once again does not include large regions of no information.

Figures 6 and *7* show the corresponding plots constructed using the latent variables derived from PLS. Here, the fouling problem does not manifest itself until latent variables three and four. Consideration of the plot of latent variables 1 versus 2 would indicate that the quality of the final product does not appear to be too questionable since the projected points lie within the

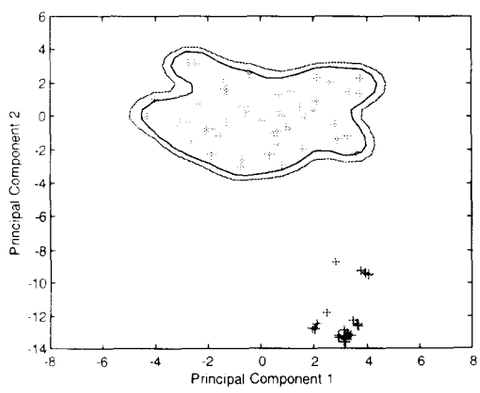

Figure 4 Scores plot for principal component 1 and 2 with nominal region defined using the M^2 approach

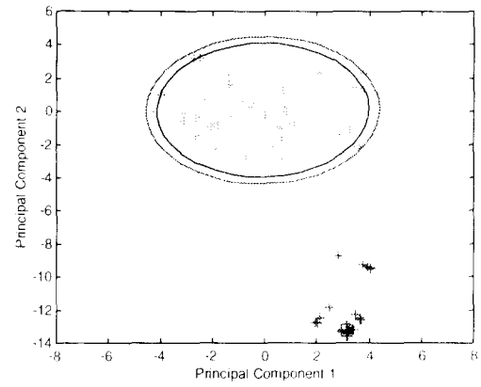

Figure 5 Scores plot for principal component 1 and 2 with nominal region defined using the M^2 approach

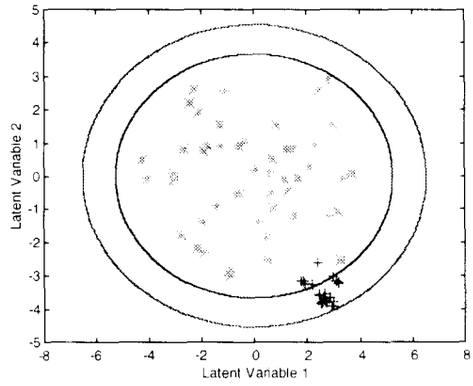

Figure 6 Scores plot for latent variable 1 and 2 with nominal region defined using Hotelling's T^2 statistic

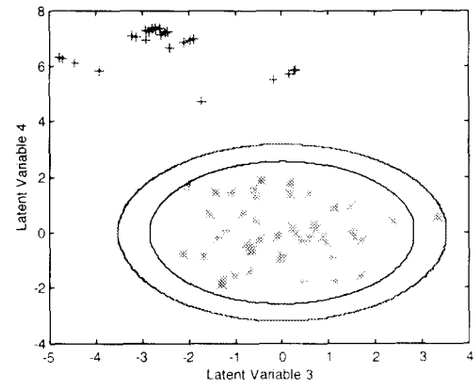

Figure 7 Scores plot for latent variable 3 and 4 with nominal region defined using Hotelling's T^2 statistic

99% limits. However, the plot of latent variable 3 versus 4 provides an extremely visual picture that a problem has occurred, reflecting in a degradation of the quality of the polymer being produced. By constructing the bounds using the M^2 statistic (*Figures 8* and *9*), this potential confusion does not occur. In contrast to *Figure 6*, *Figure 8* shows that the fouling problem is no longer captured within the bounds, hence the visual identification of a problem in the final product quality would have been instantly recognisable from any permutation of latent variables.

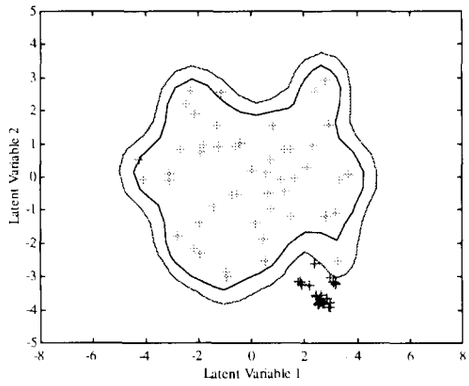

Figure 8 Scores plot for latent variable 1 and 2 with nominal region defined using the M^2 approach

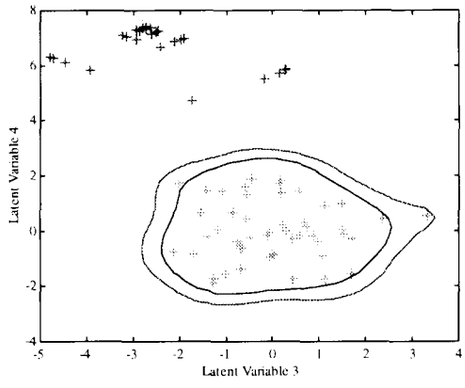

Figure 9 Scores plot for latent variable 3 and 4 with nominal region defined using the M^2 approach

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

When a process is defined to be moving out of control, operator personnel need a facility to identify which variable, or combination of variables, are reflective of the process going out of control in order to make adjustment to the process or to correct a process malfunction. It is usually left to the process operators or process engineers to assess the contributions from the measured variables in order to diagnose the assignable cause(s). This stage is much more easily carried out by interrogation of the multivariate charts constructed from PCA and/or PLS through examination of the contributions of the process variables to the score plots. One possible graphical procedure is the contribution plot. This is the change in the newly observed variables relative to the average value calculated from the PCA/PLS model[15].

Typical contribution plots are shown in *Figures 10* and *11*. The contribution from the current point is displayed along with the three previously calculated contributions to the score plot. *Figure 10* is an example of a contribution plot when the process is operating within its bounds, whilst *Figure 11* shows the results from a process which has ceased to operate within the desired limits. The major difference between the two plots is in the scale of the contributions. For a process in control, the range of values will be within the region of ±2 (two

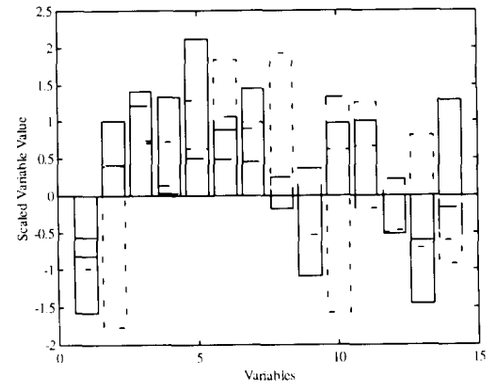

Figure 10 Contribution plot for a point which lies inside the confidence bounds: [--- point n; –.–. –. point $(n - 1)$; — point $(n - 2)$ ---- point $(n - 3)$]

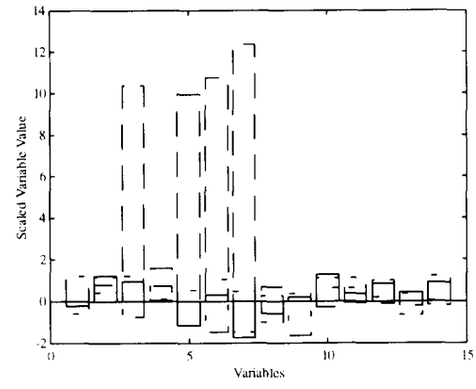

Figure 11 Contribution plot for a point which is known to have moved outside the control limits: [--- point n; –.–. –., point $(n - 1)$; — point $(n - 2)$ ---- point $(n - 3)$]

standard deviations on either side of the mean). In contrast, 'out of control' production is identified by certain variables making larger contributions than those anticipated. By identifying those variables which have experienced the greatest change in conjunction with the production engineer's and operator's expertise, it is possible to relate a particular sequence of changes to a particular process malfunction. Variables 3, 5, 6 and 7, outlet temperature from zone 1, outlet temperature from zone 2, inlet temperature of zone 1 coolant and inlet temperature of zone 2 coolant, respectively, exhibit the greatest deviations during the time span monitored for the fouling problem. This information, i.e. the combination of variables, can provide sufficient information to allow operational personnel to focus on the potential cause(s) of the process problem.

Multivariate SPC in batch processes

There is increasing strategic interest in the manufacturing of high value added chemicals. Examples include, for example, speciality chemicals, resins and polymers, pharmaceuticals and biochemicals. There are also many other batch type operations, such as drying, crystallisation and injection moulding, which are very important to the chemical and manufacturing industries. Most of

the existing industrial approaches for achieving consistent and reproducible results from batch processes are based upon careful implementation of an *a priori* derived sequencing of operations.

Recently, techniques for monitoring batch processes more closely have been developed, again based on extensions to the projection techniques of PCA and PLS, known as multi-way PCA and multi-way PLS[16]. Multi-way PCA (MPCA) is an extension to the projection technique of PCA and is based upon the philosophy of compressing information into a few latent variables which are a linear combination of the original variables. Again, 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 three-way, the added dimension being that of time.

The major issue which arises is how to handle the large number of measurements taken on the process which, in addition to not being independent, are also autocorrelated in time. It is not simply the relationship between all the variables which is important, it is the entire past histories of the trajectories. The data reduction technique of principal components analysis can be used to project the information on to a lower dimensional space which summarises the variables and their time history during previous batches. A simple way to view MPCA is to consider opening out the three-way matrix into a two-dimensional array, by placing each two-dimensional time slab (samples × variables) consecutively and performing a standard PCA.

A modification of PLS can also be applied to the batch data, multi-way projection to latent structures (MPLS). The approach adopted is a combination of PLS and MPCA. The final objective is to extract information from the process measured variable trajectories that is more reflective of the final quality parameters of the product.

Application to a batch polymerisation reactor

The batch polymerisation reactor studies are based on a pilot scale methyl methacrylate (MMA) reactor installed at the University of Thessaloniki in Greece. Heating and cooling of the reaction mixture is achieved by circulating water at an appropriate temperature through the reactor jacket. The reactor temperature is controlled by a cascaded regulator system consisting of a primary PI and two secondary PI control loops. The manipulated variables for the two secondary regulators are hot and cold water flow rates. These streams are mixed prior to entry to the reactor jacket and provide a flexible heating/cooling system. A detailed mathematical model which includes reaction kinetics, heat and mass balances has been developed to provide a rigorous simulation which has been validated against the pilot plant. Using this simulation, representative studies of reactor operation and the effects of different process malfunctions and faults can be realistically studied. In this study, a number of 'good' production batches is obtained by monte-carlo simulation to provide a nominal (or reference) data set. Two types of malfunction are studied – initiator impurity problems and reactor fouling problems. Multi-way PCA was separately carried out on both the reference data set and the data set collected during process malfunctions.

Figures 12 and *13* present the scores plots for multi-way PCA. *Figure 12* shows the results with the nominal operating region defined using Hotelling's T^2 statistic, whilst *Figure 13* presents the results based upon the M^2 statistic. The three batches (+) lying close to the nominal operating region and T^2 bounds are associated with a fouling problem. It is interesting to again observe that the standard approach draws in one of the batches identified to be associated with fouling, but that the new M^2 technique does not incorporate the non-conforming batches within the region of nominal production; clearly an important consideration in batch supervised performance monitoring.

Conclusions

The paper has presented a new approach to the generation of confidence bounds for use in the analysis of process data and in multivariate statistical process control. The approach provides an alternative statistic (the

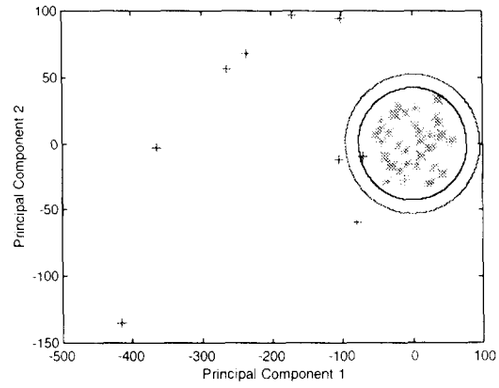

Figure 12 Scores plot for principal component 1 and 2 with nominal region defined using Hotelling's T^2 statistic

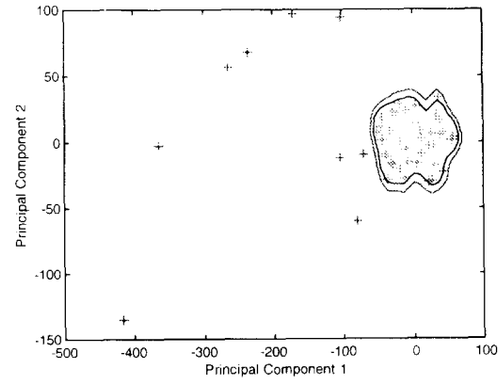

Figure 13 Scores plot for principal component 1 and 2 with nominal region defined using M^2 approach

M^2 statistic) to Hotelling's T^2, that describes the state of the data more closely. It also acknowledges the natural distributional nature of the data rather than forcing a normal distribution on to it. It is applicable to both continuous (PCA and PLS) and batch (multiway PCA and PLS) analysis.

One of the major advantages of the likelihood based confidence region is that it follows the data more closely and is less likely to incorporate regions of unknown operation which is necessarily the case for regions based upon Hotelling's T^2 statistic. A further advantage is that by adopting the M^2 approach, confidence regions are essentially preselected by identifying the 'contour' which excludes the required number of production data, if indeed any.

A question which arises now is how often should these bounds be updated? This is a question which is currently under investigation in conjunction with industrial collaborators.

One drawback, which is synonymous with all techniques which are based upon the use of kernel density estimation, is the selection of the smoothing parameter. Increasing the smoothness of the resultant plot defines limits which are smooth and therefore not so data-tied. By increasing the smoothing parameter it is possible to reduce the likelihood based approach to approximate Hotelling's T^2. However, once again the option to preselect the confidence bound to incorporate the appropriate number of data values results in much tighter and more realistic bounds. Research is ongoing, but the indications are that the M^2 statistic may be more appropriate for defining nominal operating regions especially in on-line monitoring situations.

Acknowledgements

The authors would like to acknowledge the support of the Departments of Engineering Mathematics and Chemical and Process Engineering, and funding from the EU BRITE/EURAM programme Intelligent Manufacture of Polymers No. 7009. Special thanks are also due to Professor Costas Kiparisides and Mike Papazoglou for supplying the LDPE and batch methyl methacrylate reactor data.

References

1 MacGregor, J. F., Marlin, T. E., Kresta, J. and Skagerberg, B. in 'Proceedings CPC-IV 1991 Conference' South Padre Island, Texas, p. 18, 199
2 MacGregor, J. 'Pre-prints of the IFAC ADCHEM 1994 Conference on Advanced Control of Chemical Processes' May, Kyoto, Japan, 1994
3 Martin, E. B. and Morris, A. J. 'Confidence bounds for multivariate process performance monitoring charts' Internal Report, CPACC, University of Newcastle, 1995
4 Martin, E. B., Morris, A. J. and Papazoglou, M. 'Pre-prints of the IFAC Workshop on On-line Fault Detection and Supervision in the Chemical Process Industries' Newcastle, *UK*, p. 33, 1995
5 Wold, S. 'Chemometrics and Intelligent Laboratory Systems', Vol. 2, p. 37
6 Wold, S. *Technometrics* 1978, **20** (4) 397
7 Wold, S. 'PLS modelling with latent variables in two or more dimensions', Frankfurt PLS Meeting, 1987
8 Geladi, P. and Kowalski, B. R. *Anal. Chim. Acta* 1986, **185,** 1
9 Hotelling, H. (Eds. Eisenhart, Hastay and Wallis) in *Techniques of Statistical Analysis* New York, 1947
10 Hall, P. *Biometrika* 1987, **74**, 481
11 Stone, M. *J. Royal Statistical Society B* 1974, **36**, 111
12 Bowman, A. W. *Biometrika* 1984, **71,** 353
13 Skagerberg, B., MacGregor, J. J. and Kiparissides, C. *Chemonetrics and Int, Lab. Sys.* 1991, **14**, 341
14 Martin, E. B., Morris, A. J. and Zhang, J. 'IEE Proceedings Control Theory and Applications', (Accepted March 1996)
15 Miller, P., Swanson, P. E. and Heckler, C. F. 'Contribution plot: the missing link in multivariate quality control' Presented at the 37th Annual Fall Conference ASQC, Rochester, NY, 1993
16 MacGregor, J. F., Nomikos, P. and Kourti, T. 'Pre-prints of the IFAC ADCHEM'94 Conference on Advanced Control of Chemical Processes', May, Kyoto, Japan, 1994