Process Monitoring and Diagnosis by Multiblock PLS Methods

John F. MacGregor and Christiane Jaeckle
Dept. of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada

Costas Kiparissides and M. Koutoudi
Dept. of Chemical Engineering, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece

Schemes for monitoring the operating performance of large continuous processes using multivariate statistical projection methods such as principal component analysis (PCA) and projection to latent structures (PLS) are extended to situations where the processes can be naturally blocked into subsections. The multiblock projection methods allow one to establish monitoring charts for the individual process subsections as well as for the entire process. When a special event or fault occurs in a subsection of the process, these multiblock methods can generally detect the event earlier and reveal the subsection within which the event has occurred. More detailed diagnostic methods based on interrogating the underlying PCA/PLS models are also developed. These methods show those process variables which are the main contributors to any deviations that have occurred, thereby allowing one to diagnose the cause of the event more easily. These ideas are demonstrated using detailed simulation studies on a multisection tubular reactor for the production of low-density polyethylene.

Introduction

The on-line monitoring and diagnosis of process operating performance are an extremely important part of programs aimed at improving processes and product quality over time. Such programs have led to the extensive use of statistical process control (SPC) charts (for example, Shewhart, EWMA, and CUSUM charts) to monitor key product quality variables. Although the basic concepts behind these SPC methods, as laid out by Shewhart (1931), are still as valid as ever, the charting methods utilized to implement them do not recognize the tremendous changes in measurement and data collection technology that have occurred in the last few decades. With process computers hooked up to nearly every industrial process, massive amounts of data are being collected routinely on hundreds of variables. Temperatures, flow rates, pressures, and some concentrations are usually measured throughout the process on a second or minute basis. However, many important final product quality and productivity variables, such as polymer molecular weight and process yields, are still only measured off-line and are available only infrequently on an hourly or daily basis. To effectively monitor process operating performance one must utilize all the information contained in the large number of routine measurements on the process variables (X) as well as that in the infrequent final property measurements (Y). The difficulties in accomplishing this arise not only from the need to handle measurements on such a large number of variables, but also from the fact that such process operating data are extremely ill-conditioned (variables are highly correlated and often collinear), the signal-to-noise ratio is often low, and there are usually missing data.

Multivariate statistical process control charts based on multivariate statistical projection methods such as principal component analysis (PCA) and projections to latent structures or partial least squares (PLS) have recently been proposed to treat these situations. Kresta et al. (1991), and MacGregor et al. (1991a) laid out the basic methodology for monitoring continuous processes, and MacGregor and Nomikos (1992) extended it to the case of batch processes. In related work, Wise et al. (1991a) utilized such PCA/PLS-based charts to monitor an industrial Ceramic melter, and Piovoso et al. (1992) used them to monitor an industrial polymerization process. A thorough treatment of PCA methods in quality control is given in Jackson (1991).

The major advantages of these multivariate projection methods are their ability to handle large numbers of highly correlated variables, measurement errors, and missing data. Of particular importance is their ability to reduce the dimensionality of the monitoring space by projecting the information in the data down into low-dimensional spaces defined by a few latent variables. The processes are then monitored in these reduced spaces by using one-, two-, or three-dimensional control charts that retain all the simplicity of presentation and interpretation of conventional single variable SPC charts. However, by utilizing the information contained in all the measured variables simultaneously, they are much more powerful for detecting deviations from normal operating conditions.

The purpose behind SPC charts for monitoring processes is to provide a timely detection of any deviation from normal or "in-control" behavior so that operators and engineers can then examine the data around this point in time and diagnose an assignable cause for the deviation. By finding and eliminating such assignable causes through implementing process modification, changing operating procedures, and so on, a steady improvement in productivity and product quality can be achieved. However, neither the classical SPC charts nor these multivariate generalizations of them provide any diagnostic capabilities that might help one isolate an assignable cause once it occurs. This task of diagnosis has been simply left up to the troubleshooting skills of the process operators and engineers. The purpose of this article is to present diagnostic methods for multivariate SPC charts based on projection methods (PCA/PLS) and to show that their diagnostic capability is another great advantage that multivariate procedures have over classical univariate charts.

The diagnostic procedures will utilize the fact that underlying the multivariate monitoring scheme is a PCA or PLS projection model that characterizes the relationships among all the process variables under normal operating conditions. Once a deviation is detected by the monitoring chart, the variables which make major contributions to this deviation are easily isolated using the underlying projection model. Once the major contributing variables are known, the diagnosis problem is much easier.

In very large processes involving many processing units with many variables in each unit, the number of potential causes can be very large, making the diagnosis more difficult. For these situations, we develop new hierarchical multivariate monitoring methods based on multiblock PLS algorithms that allow one to monitor not only the overall process, but each unit or identifiable subsection of the process. These methods are often capable of clearly isolating in which subsection of the process a deviation has occurred, and then further revealing which variables within this subsection are the major contributors to the event. Although they are still not capable of unambiguously diagnosing the cause of an event (this can only be done by methods incorporating exhaustive fundamental process knowledge), they can clearly focus on a particular section of the plant and on a small number of specific variables in that section, thereby making it much easier for operators to use their knowledge to diagnose possible assignable causes.

To illustrate these concepts we consider a high-pressure tubular reactor process for the manufacture of low-density polyethylene. Previous studies applying multivariate statistical projection methods to analyze data from these reactors, to develop predictive models, and to establish monitoring schemes have been presented in MacGregor et al. (1991b) and Skagerberg et al. (1992).

Low-Density Polyethylene (LDPE) Process

Low-density polyethylene is produced at high pressures in tubular and autoclave reactors. A thorough overview of the literature, the reaction kinetics and the fundamental modeling of these LDPE processes is presented in Kiparissides et al. (1993). In this article we consider the monitoring and diagnosis of the operating performance of the tubular reactor process.

Most tubular LDPE reactors consist of three to four sections totaling 500 to 1,500 m in length. Ethylene and a solvent are preheated and fed, along with initiator, to the first section. Additional injections of cool ethylene, solvent and initiators are made between the sections to cool down the reaction mixture and to provide fresh initiators for further reaction. Each section is jacketed and cooled by a cooling medium flowing countercurrently. In this article we consider only the first two sections of such an industrial reactor to illustrate the methods more clearly. However, the value of the multivariate and multiblock PLS methods presented here will increase substantially as the number of variables and blocks increases. Figure 1 shows the first two sections together with a typical temperature profile.

Operating conditions in these reactors influence many of the molecular properties of the polymer produced, and these in turn affect the behavior of the polymer in its final application. The major productivity variable of interest is the conversion per pass (CONV). The molecular properties of interest include

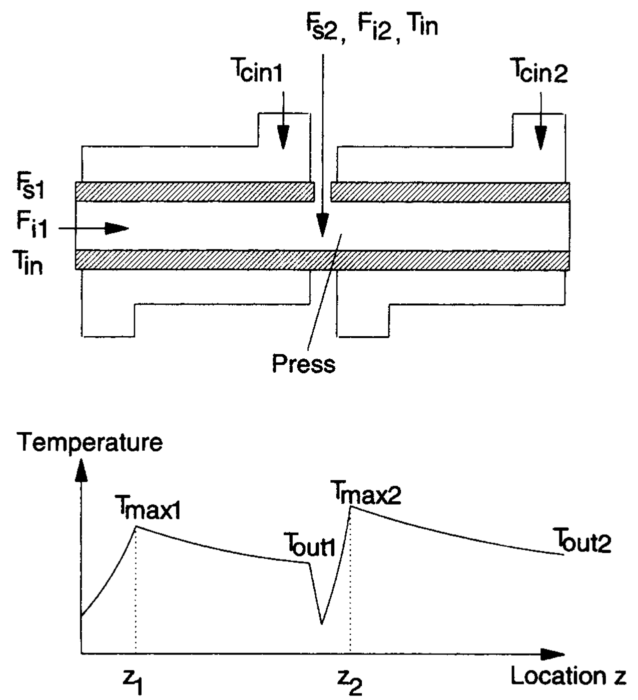

Figure 1. Two-zone LDPE reactor with a typical temperature profile.

Table. 1 Process Variables (X) for the LDPE Reactor

Variables	Definition
T_{in}	Inlet temperature of the reaction mixture (K)
T_{max1}	Maximum temperature of the reaction mixture in the first zone (K)
T_{out1}	Outlet temperature of the reaction mixture in the first zone (K)
T_{max2}	Maximum temperature of the reaction mixture in the second zone (K)
T_{out2}	Outlet temperature of the reaction mixture in the second zone (K)
T_{cin1}	Inlet temperature of the coolant in the first zone (K)*
T_{cin2}	Inlet temperature of the coolant in the second zone (K)*
z_1	Position of the reactor where T_{max1} appears (% of reactor length)
z_2	Position of the reactor where T_{max2} appears (% of reactor length)
F_{i1}	Total inlet flow rate of the initiators to the reactor (g/s)
F_{i2}	Total flow rate of the initiators in the intermediate feedstream (g/s)
F_{s1}	Inlet flow of the solvent in the reactor (% of ethylene)
F_{s2}	Flow of the solvent in the intermediate feed (% of ethylene)
Press	Pressure of the reactor (atm)

* The outlet temperature of the coolant in the two zones is fixed.

the weight and number average molecular weights (MW_w, MN_n), and the long- and short-chain branching frequencies (LCB, SCB). However, none of these properties are available on-line, and many of them are either not measured at all or are measured only infrequently. More often variables such as melt index (MI) and density (ρ) which are related to these fundamental properties are measured periodically in the laboratory. However, many on-line measurements such as the temperature profile down the reactor, the coolant temperatures, and the solvent and initiator flow rates are available on a frequent basis.

The process measurements assumed to be available in this study are listed in Table 1. Although the entire temperature profile is usually available for each reactor section, we use a common industrial practice of summarizing the profile in each section by its inlet, maximum and outlet temperatures together with the position of the maximum (z). This is again done to simplify the presentation of the results in this article. In practice it might be better to retain all the temperature measurements along the profile, since the PLS methods presented are capable of handling such highly collinear data, and in this way no detailed information arising from subtle changes in the profile shape will be lost. Such data on the entire profile were used in earlier studies (MacGregor et al., 1991b; Skagerberg et al., 1992).

There are a number of reasons for wanting to monitor LDPE reactors. Changes in impurity levels (ppm) in the feeds can greatly influence the rate of reaction and the temperature profile, and hence the conversion and polymer quality. Sudden drops in impurity levels or increases in initiator efficiencies or initiator feed rates could lead to a rapid temperature rise and possible decomposition reactions. Reactor fouling reduces the wall heat-transfer coefficient and requires periodic defouling cycles. Chain transfer agents entering with the monomer or solvent can influence the molecular weight development. Other events due to malfunctioning equipment and sensors or due to poor operating procedures should also be alarmed and corrected. The methods should not only be able to detect such events, but also provide information on their possible causes.

The data used to illustrate the methods developed here are from a fundamental steady-state simulation model presented in Kiparissides et al. (1993) and adjusted to match typical data from an industrial process. These steady-state data might reasonably represent measurements collected from an industrial process at time intervals longer than the process time constant or averages of measurements taken over some time period.

Projection to Latent Structures (PLS)

The basic concepts and algorithms of PLS have been thoroughly presented in the chemometrics literature (Wold et al., 1984; Geladi and Kowalski, 1986; Martens and Naes, 1991; Höskuldsson, 1988). Their use in multivariate monitoring of process operating performance has also been treated in recent articles (Kresta et al., 1991; MacGregor et al., 1991a,b; Wise et al., 1991b; MacGregor and Nomikos, 1992). Therefore, in this section we will provide only a basic overview of the methods and their application to monitoring.

Consider the situation where one has measurements on k process variables (Table 1) taken at n different times. This data can be arranged into a $(n\times k)$ process data matrix X. Measurements on m key quality and productivity variables are usually made much less frequently and can be arranged into an $(\ell\times m)$ output matrix Y where ℓ is the number of output observations.

If one were interested in understanding and monitoring the variability in only the process variables (X), then one could perform a PCA on X. If one were more interested in studying and monitoring the variations in the process variables that are most influential on the quality and productivity variables (Y), then one should perform a PLS analysis using both Y and the $(\ell\times k)$ submatrix of X corresponding to the times at which the Y samples were taken. Both these procedures are based on projecting the information in the high-dimensional data spaces (X, Y) down onto low-dimensional spaces defined by a small number of latent variables ($t_1, t_2, \ldots, t_A$). These new latent variables summarize all the important information contained in the original data sets. In PLS, the scaled and mean-centered X and Y matrices are decomposed as:

$$X=\sum_{a=1}^{A} t_a p_a^T + E \quad (1)$$

$$Y=\sum_{a=1}^{A} t_a q_a^T + F \quad (2)$$

The latent vectors t_a are computed sequentially from the data for each PLS dimension ($a=1, 2, \ldots, A$) such that the linear combination of the xs defined by the latent variable $t_a = w_a^T x$ and the linear combination of the ys defined by the latent variable $u_a = q_a^T y$ maximize the covariance between X and Y that is explained at each dimension. The common nonlinear iterative partial least-squares (NIPALS) algorithm is described in Appendix A. The vectors w_a and q_a are loading vectors whose elements w_{aj} and q_{aj} express the contribution of each variable x_j and y_j, respectively, toward defining the new latent

Table 2. Nominal Product Quality Variables (Y) for the Reference Set

Variables	Definition	Nominal Values
CONV	Cumulative conversion of monomer	0.1324
MW_n	Number of average molecular weight	27,361
MW_w	Weight average molecular weight	163,680
LCB	Long Chain Branching/1,000 atom C	0.792
SCB	Short Chain Branching/1,000 atom C	26.08

variables t_a and u_a. The total number of PLS components (A) needed to extract the information from X and Y is usually low (typically 2 to 5) and can be determined by cross-validation (Wold, 1978). The first few PLS dimensions are usually the most important and dominate the model. Hence, for monitoring it is often sufficient to consider only the first few dimensions ($a=1$, 2) in Eqs. 1 and 2, while for prediction purposes we should use all the dimensions determined through cross-validation.

PLS can also be viewed as a biased regression method and the final model in Eq. 2 can be expressed in terms of the X data as the regression model:

$$Y = X\beta + F \tag{3}$$

where the matrix of regression coefficients is given by:

$$\beta = W(P^T W)^{-1} Q^T \tag{4}$$

where W, P, and Q are $(k \times A)$, $(k \times A)$ and $(m \times A)$ matrices whose columns are the vectors w_a, p_a and q_a; $a = 1, 2, \ldots, A$, as defined in the PLS algorithm in Appendix A.

PLS model for the LDPE process

Although a number of LDPE grades are made in any industrial reactor, we consider here the development of a monitoring and diagnosis scheme for one such grade. Table 2 shows nominal product quality variables (Y) at the exit of the second reactor section for this grade, and Table 3 the range of settings for some of the main process variables during normal production of this grade. A reference set of data that is meant to represent normal process variations was generated using Monte Carlo simulation to select process variables within these ranges. Note that variations were introduced not only into some of the measured process variables but also into unmeasured factors such as the wall fouling factor coefficients and the impurity feed rates to the reactor. These data were selected to represent

Table 3. Process Conditions for the Reference Set

Variables	Range of Variation
F_{s1}	5.95-6.05%
F_{s2}	5.95-6.05%
Press	2,965-3,035 atm
T_{in}	477-483 K
Fouling Factor Coefficient	15-25 $(cal/cm^2/s/K)^{-1}$
Impurities	15-35% of initiator flow rates
Initiator Flow Rate	0.408-0.510 g/s

Table 4. Cumulative % Sum of Squares Explained by the PLS and the Multiblock PLS Models

PLS Dimension	PLS		Multiblock PLS $X1$ = reactor section 1 $X2$ = reactor section 2	
1	X: 28%	Y: 63%	$X1$: 39% $X2$: 41%	Y: 61%
2	X: 44%	Y: 84%	$X1$: 57% $X2$: 56%	Y: 83%
3	X: 56%	Y: 90%	$X1$: 69% $X2$: 70%	Y: 86%
4	X: 66%	Y: 93%	$X1$: 87% $X2$: 89%	Y: 90%
5	X: 78%	Y: 94%	$X1$: 98% $X2$: 98%	Y: 92%

normal process operation when only common cause variations were present and when acceptable product quality was achieved.

Using 50 simulated steady-state conditions, a PLS model was developed to represent the process behavior for this "in-control" situation. In effect, the PLS model attempts to explain the "common-cause" variations in the process data and to exclude the random variations and measurement errors that are uncorrelated with other X and Y variables. The results are summarized in Table 4. The first two dimensions were most significant, explaining 84% of the variation in the Y variables using 44% of the variation in X.

Monitoring the LDPE reactor using the PLS model

Conventional SPC approaches would probably plot five individual charts on the Y variables that are measured only every few hours and ignore the 14 frequently available process variables (X). Even in this simplified problem, if another 14 charts on the xs were added, it would be almost impossible to interpret so many graphs simultaneously. However, we have already seen from the PLS analysis that under normal operations most of the information in the process variables (X) that explains the variations in the Y space (84%) is captured in the first two latent vectors (t_1, t_2). If the process continues to operate in the same manner in which it did when the normal operating data were collected, then new observations should continue to be explained by the same PLS model. That is, they should fall very close to the plane defined by the dominant latent vectors and project into the same region of the plane as did the past data.

Therefore, following Kresta et al. (1991), one can plot the behavior of the process in the t_1-t_2 plane (or as individual control charts in these orthogonal latent variables if more than 2 are used) and define a control region in this plane within which new process data (X) should continue to project as long as the plant continues to operate normally. Such a two-dimensional latent variable plot is shown in Figure 2 for the LDPE process, where the contours define the control boundaries corresponding to 5% and 1% significance levels. Under the assumption that the latent vectors are normally distributed with zero means, these regions are defined by the ellipses (Anderson, 1958):

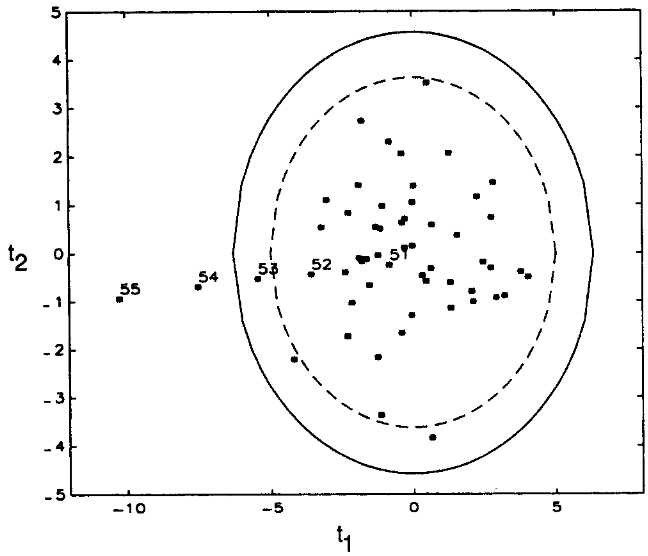

Figure 2. PLS: t_1-t_2 plane; impurities entering both reactor sections.

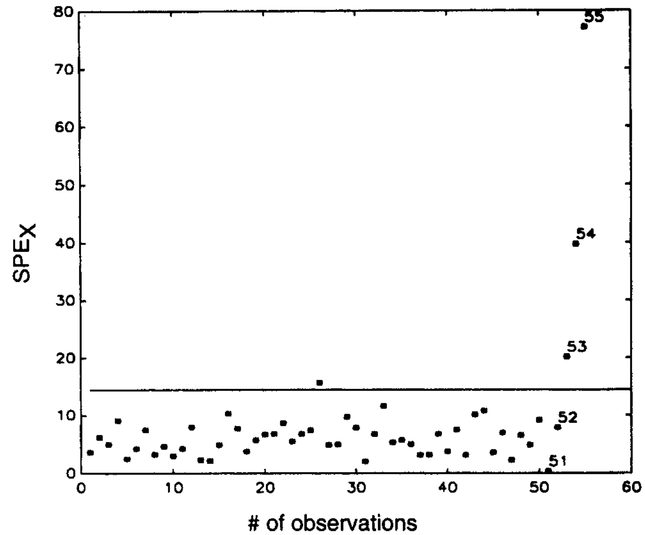

Figure 3. PLS: SPE$_x$; impurities entering both reactor sections.

$$\tau^T S^{-1} \tau = \left(\frac{2 \times 49}{48}\right) F_\alpha(2, 48) \quad (5)$$

where S is the estimated covariance matrix of $\tau = (t_1, t_2)^T$, $F(2, 48)$ is F-distribution with 2 and 48 degrees of freedom, and α is the significance level. When the t_is are orthogonal as in the standard PLS algorithm (Appendix A), the covariance matrix $S = \text{diag}\{s_1^2, s_2^2\}$. For highly nonlinear systems, a reference distribution approach as discussed in Kresta et al. (1991) can be used to define the control region boundaries.

The unnumbered points in Figure 2 represent the projection of the normal operating data into this plane and illustrate that under normal operation most points are contained within these control contours. However, when a fault occurs, such as a gradual increasing level of impurities in the feed ethylene to both sections the projected process data points 51, 52, 53, 54, and 55 progressively move outside the acceptable region, thereby alarming a special event. A similar projection plot for the much less frequently available product quality data (Y) could be presented using the first two latent variables (u_1, u_2) of the Y-space. New y-data, when obtained, should fall within a similar region in this space.

If the process continues to operate in a normal manner, then new observations should not only continue to project into these control regions of the latent variable planes, but also should lie in or very close to these planes. Kresta et al. (1991) suggested that the squared perpendicular distance of new observations (x_i) from these planes (squared prediction errors or SPE) could be calculated as:

$$\text{SPE}_{x,i} = \sum_{j=1}^{k} (x_{ij} - \hat{x}_{ij})^2 \quad (6)$$

$$\text{SPE}_{y,i} = \sum_{j=1}^{m} (y_{ij} - \hat{y}_{ij})^2$$

where $\hat{x}_{ij}$ and $\hat{y}_{ij}$ are the values predicted by the PLS model. These can be plotted vs. time much as a conventional range or s^2-chart to detect the occurrence of any new source of variation not present in the reference set. Such new sources of variation would give rise to new latent variables and therefore would result in the new observations moving away from the plane defined by the original latent variables and therefore the SPE would increase. Such a plot of SPE$_x$ is shown in Figure 3 for the data of Figure 2. The data coming from normal process operations fall well below the control limit, but as the impurities start to increase the SPE$_x$ rapidly violates the control limit indicating that a new event has occurred. The control limits for the SPE correspond to a hypothesis test on the model adequacy at approximately a 1% level of significance. They were determined by approximating the distribution of the SPE by a Chi-Squared distribution with parameters determined by matching moments of this distribution with those of the "in-control" reference data (Nomikos and MacGregor, 1993).

Diagnosing faults by interrogating the PLS model

Although the above monitoring charts are very powerful ways for detecting deviations from normal operations, they do not indicate reasons for such deviations. The remainder of this section is concerned with developing diagnostic tools for this purpose.

A simple approach to diagnosis might be to develop an expert system based on the behavior of the data projections in the latent variable space. Once one has experience with using two- and three-dimensional plots in the latent variable space, such as Figure 2, it becomes apparent that certain types of faults such as impurity contamination, reactor fouling, and so on are characterized by the data projections moving into specific regions of the latent variable space. This would therefore imply that an expert system might be developed from the behavior of the latent variables during past faults. For example, the next time a movement outside the control region in the negative t_1 direction occurred (as in Figure 2) impurity contamination might again be suspected.

However, more detailed information about possible causes can be obtained by a closer interrogation of the underlying PLS model. Consider the problem of diagnosing a possible cause for the event detected in Figures 2 and 3. The event is clearly detected by point 53 on the SPE_x plot and by point 54 on the latent variable (t_1-t_2) plot. Consider that the point 54 has just been obtained. To diagnose the event one can examine the contributions of the individual variables to this larger than normal value of the SPE_x at point 54, that is:

$$SPE_{x,54} = \sum_{j=1}^{k} (x_{54,j} - \hat{x}_{54,j})^2.$$

Here the predictions $\hat{x}_{54,j}$ are made from the PLS model developed for the "in-control" operating data as:

$$\hat{x}_{54,j} = \sum_{a=1}^{A} t_{a,54} p_{aj}$$

where the new latent variable projections for the 54th observation are given by:

$$t_{a,54} = \sum_{j=1}^{k} w_{a,j} x_{54,j}$$

The prediction errors in each process variable $(x_{54,j} - \hat{x}_{54,j})$; $j = 1, 2, \ldots, k$ which form the individual contributions to this $SPE_{x,54}$ are plotted in Figure 4. The major contributions can be seen to come from the positions of the hot spot in each section (z_1 and z_2) and from the feed rates of the initiators to each section (F_{i1}, F_{i2}). What this reveals is that the hot spot positions have moved further down the reactor and the initiator feed rates are higher than they should be according to the PLS model. These observations are consistent with a situation in which the extent of reaction in both sections is lower than expected. The given initiator flows should have produced more reaction in both sections, and with lower extents of reaction the location of the hot spots tend to move down the reactor. An obvious explanation is that either radical scavenging impurities have entered both sections or the initiator efficiencies have dropped.

A second way of diagnosing the event is to note that in the latent variable space (t_1-t_2) shown in Figure 2 the main reason for the alarm being given is that there has been a large decrease in the t_1 direction between points 51 and 54. The importance of each process variable to this latent variable (t_1) is given by the PLS loading vector w_1. Therefore, the contribution of each variable (x_j) to this large movement in t_1 can be computed as $\{(w_{1,j} \cdot \Delta x_j); j = 1, 2, \ldots, k\}$ where $\Delta x_j = (x_{54,j} - x_{51,j})$. These contributions are shown in Figure 5. It can be seen that the major contributors to the movement in t_1 are again movement of the positions of the two hot spots (z_1 and z_2) further down the reactor (multiplied by negative ws), and decreases in the two hot-spot temperatures (T_{max1} and T_{max2}) (multiplied by positive ws). These again confirm a change in the extent of the reaction in both sections leading to the same diagnosis as above.

Similarly, the monitoring plots in Figures 2 and 3 are capable of detecting other faults and combinations of faults, and the diagnostic bar charts will reveal different combinations of

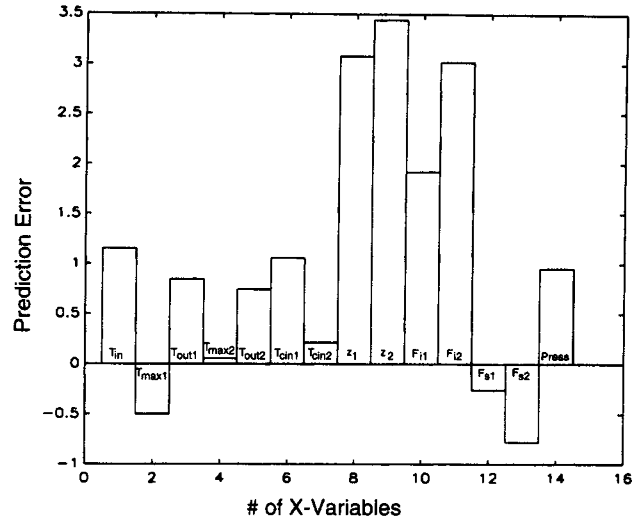

Figure 4. PLS: prediction errors in the individual process variables contributing to SPE_x at time point 54.

process variables as being the major contributors to the SPE_x and to the t_1, t_2 movements. For example, when fouling of the reactor walls in any section becomes excessive it is detected rapidly by the monitoring charts. Figure 6 shows the results for increased fouling in both sections. As fouling increases, the projection of the new observations (points 51, 52, 53, 54, 55) into the latent variable plane (t_1-t_2) does not reveal a problem. However, the values of SPE_x for these points clearly and quickly reveal that a deviation from normal behavior has occurred. The diagnostic bar chart (Figure 6c) for the variable contributions to the SPE_x at point 53 reveals that the major contributing process variables are the coolant inlet and reactor outlet temperatures, exactly what one might expect in a fouled reactor.

Although the diagnostic bar plots presented here do not

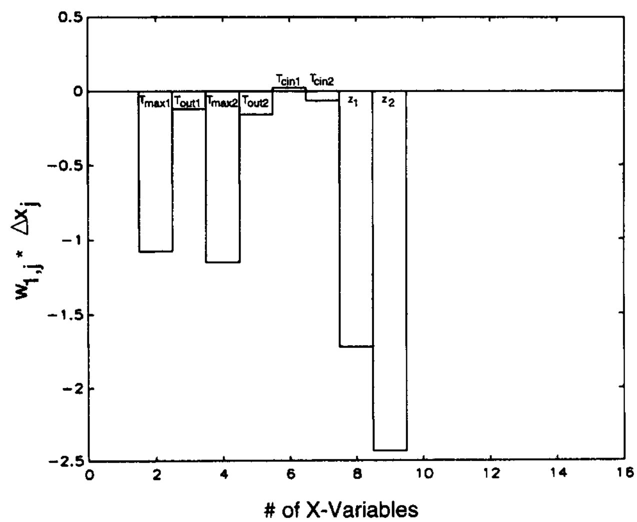

Figure 5. PLS: variable contributions to the change in t_1 from point 51 to 54.

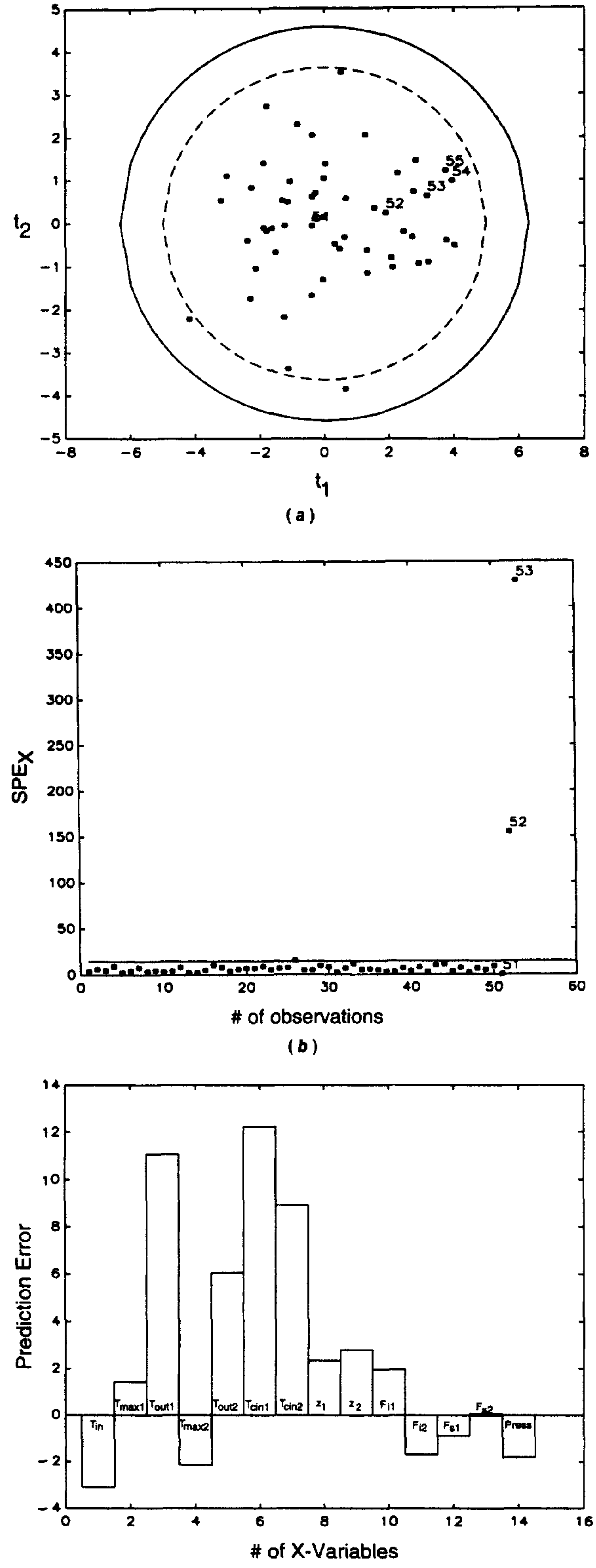

Figure 6. PLS: fouling in both reactor sections.

(a) t_1-t_2 plane; (b) SPE_x; (c) prediction errors contributing to SPE_x at time point 53.

unambiguously reveal the cause of the event, they clearly point to the group of process variables that are no longer consistent with normal operating conditions, and thereby focus the attention of the operators or engineers, and allow them to use their process knowledge to deduce possible causes. The value of such plots as diagnostic tools becomes much greater as the process becomes more complex and the number of process variables grows. Miller et al. (1993) illustrate their usefulness in some applications to complex manufacturing processes. More precise statistical tests on the magnitude of the individual residual contributions could also be performed (for example, Wise et al., 1991b), but since we have already confirmed that a statistically significant deviation has occurred, we use them here mainly as a qualitative diagnostic tool.

Multiblock PLS Methods

When the number of variables is very large, then even the above monitoring and diagnosis schemes can become unwieldy and difficult to interpret. However, when large processes consist of many distinct sections with only a few streams connecting each section, it is logical to break up the process variables into several X-blocks, one corresponding to each section. Furthermore, if there are several product streams exiting from different sections of the plant, it is also logical to split up the quality and productivity variables into several Y-blocks. Analysis of such blocked data can then be accomplished by multiblock PCA and PLS methods. The advantage of such blocking is mainly to allow for easier interpretation of the data by looking at smaller meaningful blocks and at the relationship between blocks.

Slama (1991) analyzed several months of data from a fluid-bed catalytic cracking process and the subsequent fractionation section of an oil refinery. Since several hundred variables were involved it was found to be convenient to break them up into blocks and use multiblock PLS. X-blocks corresponded to process variables in the feed system, the reactor, the two regenerators, and several sections of the fractionation train, and Y-blocks corresponded to yields of light gases, heavy components, and so on. In this article, we propose the use of such multiblock projection methods to develop on-line multivariate SPC schemes for monitoring and diagnosing process operating performance with blocked processes.

Multiblock data analysis methods have their origins in path analysis and path modeling in the fields of sociology and econometrics. Multivariate projection methods for analyzing such blocked data are largely due to Herman Wold (1982) and Svante Wold (1987). In this article we will use a variation of the multiblock or hierarchical PLS algorithms of S. Wold et al. (1987) and Wangen and Kowalski (1988). The algorithm is presented and discussed in Appendix B. The basic concepts will be illustrated here using the case shown in Figure B1 where there is a single Y-block and two X-blocks. This simple structure will be used also in the LDPE process example where the $X1$ and $X2$ blocks will contain the variables in sections 1 and 2 of the reactor, respectively.

Although there are many possible variations of multiblock PLS algorithms, the one we employ here leads to a set of orthogonal loading vectors ($w\ell_a$, $a = 1, 2, \ldots$) and orthogonal latent vectors ($t\ell_a$, $a = 1, 2, \ldots$) for each block $X\ell$. The $X\ell$

blocks are then represented in terms of their leading A PLS components as:

$$X1 = \sum_{a=1}^{A} t1_a p1_a^T + E1$$

$$X2 = \sum_{a=1}^{A} t2_a p2_a^T + E2$$

This allows one to establish monitoring and diagnostic plots for each block separately as previously described for single block PLS. An overall monitoring space for the process can be obtained by using projections in the latent vector space (tc_a, $a = 1, 2, \ldots$) of the consensus T matrix formed by collecting the latent vectors from the individual blocks (see Appendix B and Figure B1). Although the score vectors $tc_a (a = 1, 2, \ldots, A)$ of this consensus T matrix are no longer orthogonal to one another, experience has shown that if the plant is blocked in a meaningful fashion where most variable interactions are within the blocks, these vectors continue to define the same plane as the latent vectors obtained by single block PLS, and they provide essentially the same predictions of Y:

$$\hat{Y} = \sum_{a=1}^{A} tc_a q_a^T$$

The problem of selecting the block structure of the process to use in the multiblock methods is not something for which specific rules can easily be given. The choice of blocks will depend upon engineering judgment and the objectives of the study. However, some general guidelines are as follows. Blocks should correspond as closely as possible to distinct units of the process (Slama, 1991) where all the variables within a block or process unit may be highly coupled, but where there is minimal coupling among variables in different blocks. Variables associated with streams that leave one block and enter another (feed or recycle streams) should generally be included in both blocks. A check on whether or not the blocking has been well done is to compare the predictions of Y obtained from the single and multiblock algorithms for the same number of dimensions (A). These should be comparable. If the multiblock algorithm provides significantly poorer predictions or needs more latent vectors (A) to provide the same quality of prediction, then the blocking has probably been poorly done.

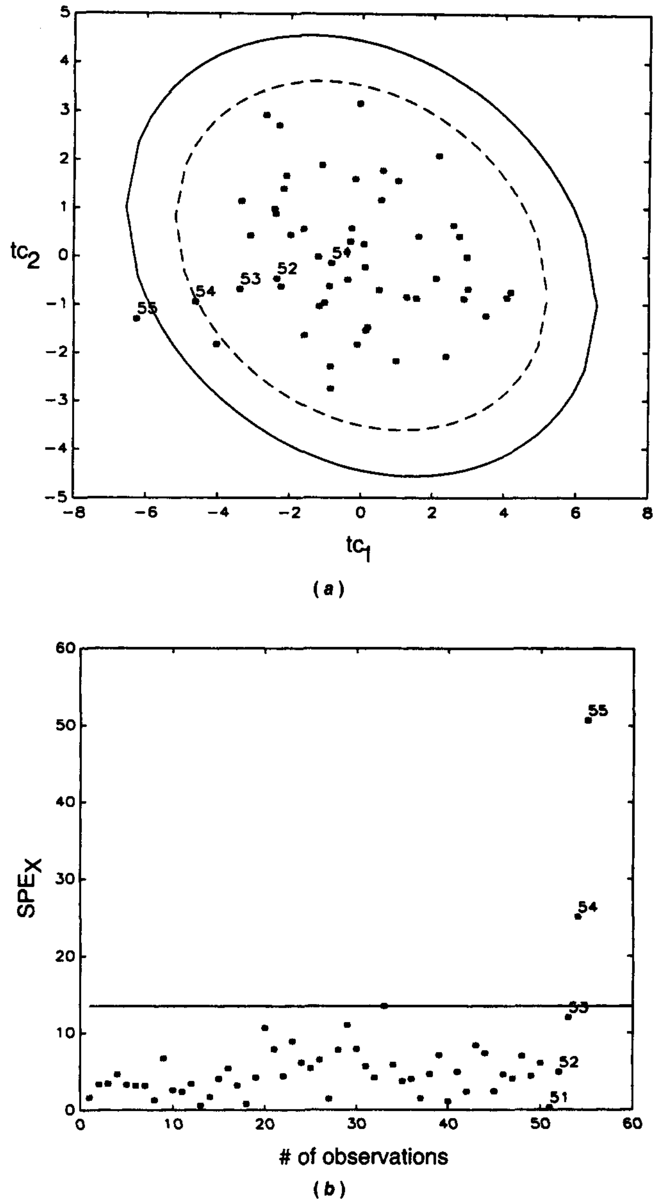

Figure 7. Multiblock PLS: impurities entering the second reactor section.

(a) Consensus tc plane for overall process; (b) SPE_x for overall process.

Multiblock monitoring and diagnosis of the LDPE process

Tubular LDPE reactors consist of three to four distinct reactor sections each with separate cooling sections and additional side feedstreams. All the process variables within a section are highly coupled (for example, temperature measurements along a profile) but variables between sections are much less coupled. Furthermore, each section can encounter its own set of special events such as fouling of the walls, impurities entering in a sidestream, and so on. Therefore, it is logical to break up the process into blocks, one identified with each reactor section, and to develop monitoring and diagnosis charts for each section as well as an overall monitoring chart. In this way, we will see that the ability to detect and then diagnose faults is enhanced.

The process variable blocks $X1$ and $X2$ were defined as containing all the variables in sections one and two, respectively, as listed in Table 1. However, since in the simulation used for this study, the reactor inlet temperature (T_{in}) and the pressure were common to both sections they were included in each block giving eight variables per block. In general, the pressure and inlet temperature ("light-off temperature") for each reactor section would be slightly different and therefore distinct to that block. The results of the multiblock PLS analysis are shown in Table 4. Contrasting these with the results of ordinary PLS (also given in Table 4), one can see that the sum of squares in Y explained by the multiblock model is

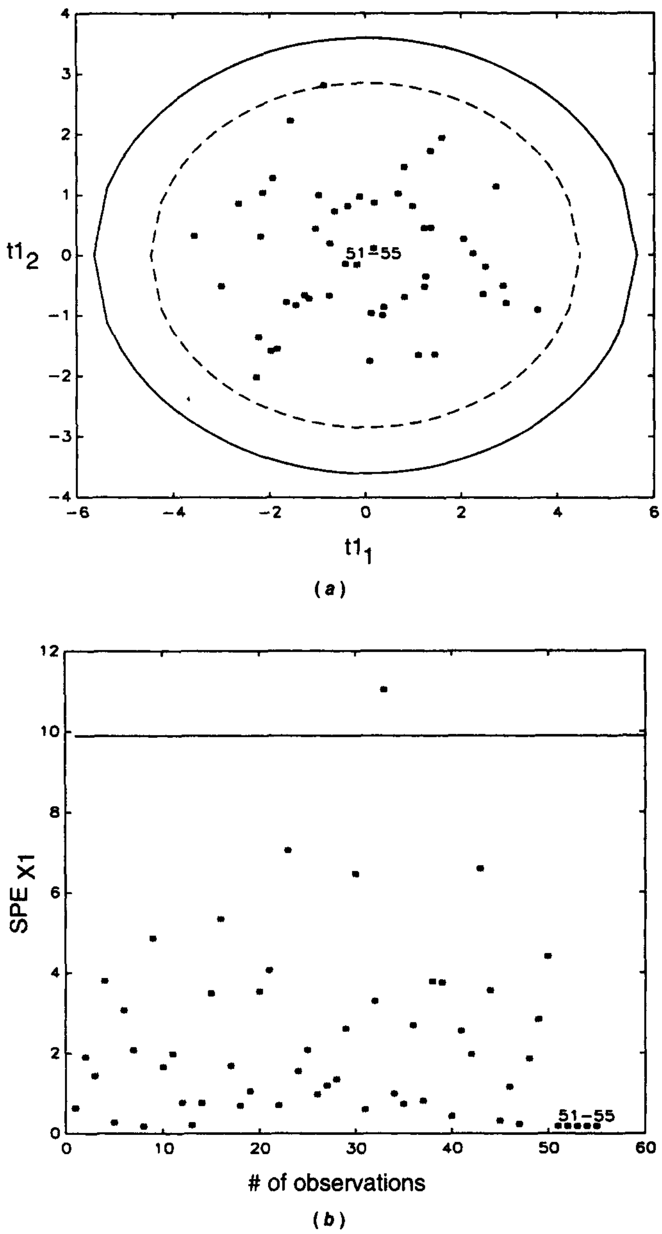

Figure 8. Multiblock PLS: impurities entering the second reactor section.

(a) $t1_1$-$t1_2$ plane for reactor section 1; (b) SPE_{X1} for reactor section 1.

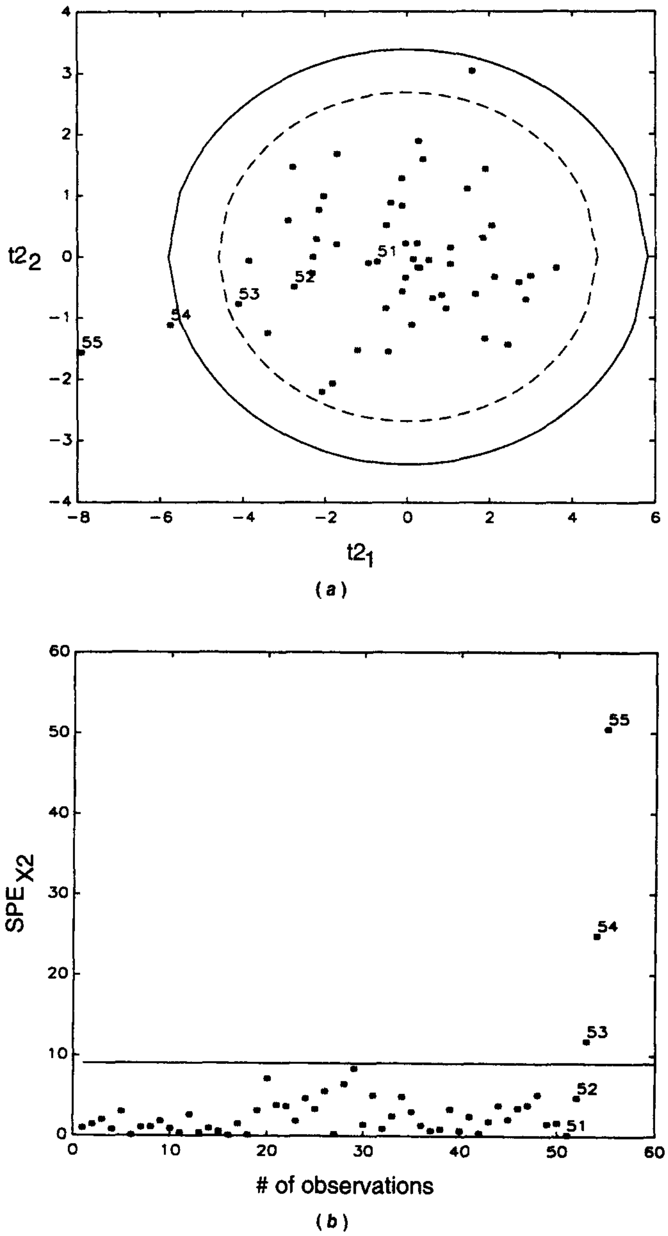

Figure 9. Multiblock PLS: impurities entering the second reactor section.

(a) $t2_1$-$t2_2$ plane for reactor section 2; (b) SPE_{X2} for reactor section 2.

almost equivalent to that of the single block model. In both cases, most of the variation in the *Y*-space is explained by two latent vectors (83% and 84%).

Consider again the problem of monitoring the LDPE process. From the multiblock formulation we get three monitoring spaces: an overall monitoring space based on the consensus latent variables tc_a, and one monitoring space for each section based on the section latent variables ($t1_a$, $t2_a$; $a=1, 2, \ldots$). SPC monitoring plots for each of these three spaces are shown in Figures 7, 8, and 9. The points without numbers in the (t_1-t_2) and SPE_x plots represent normal or "in-control" observations. The points 51, 52, 53, 54, and 55 represent points of increasing impurity levels entering the side feedstream to the second section of the reactor. In the monitoring plots for the overall process (Figure 7) this problem is detected, but only at point 55 in the (tc_1-tc_2) plane, and at point 54 in the SPE_x-plot. (The control region in the (tc_1-tc_2) plane, defined again by Eq. 5, is now an oblique ellipsoid because tc_1 and tc_2 are no longer orthogonal.) However, in the monitoring space for the second reactor section by itself (Figure 9) the detection of the problem is more rapid (at points 54 and 53, respectively). Furthermore, in the monitoring space for the first reactor section (Figure 8) no indication of any problem is given since there is no abnormal behavior in this section. Hence, the multiblock monitoring charts are clearly able to pinpoint that the problem is occurring in the second section of the reactor and

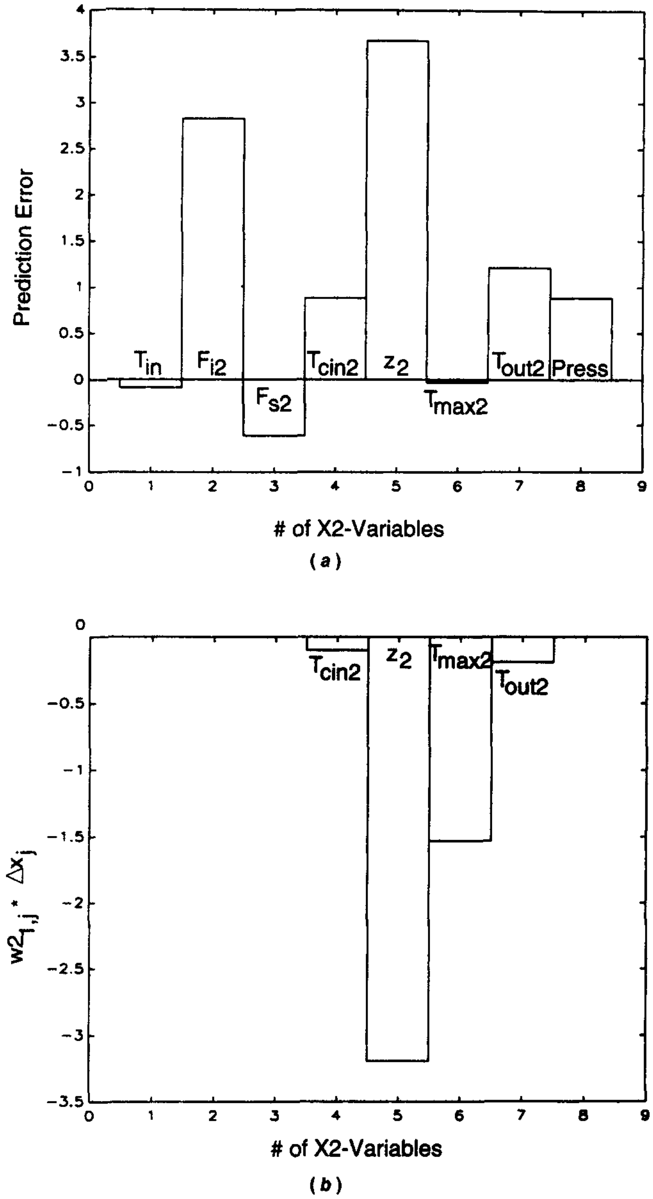

Figure 10. Multiblock PLS: impurities entering the second reactor section.

(a) Prediction errors contributing to the SPE_{X2} for reactor section 2 at time point 54; (b) variable contributions to the change in $t2_1$ from point 51 to 54.

not in the first section, and furthermore, they appear able to detect the problem earlier than the overall PLS charts.

The increased sensitivity of the multiblock procedure arises because the monitoring plots for the individual blocks are assessing the magnitude of the deviations in the (t_1-t_2) plane and in the $SPE_{X\ell}$ relative to normal, or "in control" variations on that part of the process only, and not with respect to variations in all the variables of the process as must be done in the charts for monitoring the overall process. As illustrated in Table 4, they also accomplished this by utilizing more of the information in the individual sub-blocks ($X\ell$) than is utilized by the single block PLS on the whole system. Therefore, if indeed the fault is localized to one section then the monitoring plots for that section will be more sensitive in detecting it. In general, if faults occur in one section of a plant they may propagate their effects through to subsequent sections thereby setting off alarms in the monitoring charts for these sections as well. In this situation, it would be logical to suspect the location of the fault to be in that section whose monitoring charts first alarmed the event.

To further diagnose the cause of the event, diagnostic bar charts can be constructed as in the previous section of the article, showing the contribution of the individual variables to the SPE_x and to the changes in the latent variables. However, now the search can be substantially reduced since only those variables in the section where the fault was detected need to be examined. For the LDPE example considered above the contribution of each of the variables (x_j, $j = 1, 2, \ldots, 8$) in section 2 of the reactor to the SPE_{x2} and to the first latent variable movement ($w2_{1j}\,\Delta x_j$) for that section are shown in Figure 10. As before, the major contributors to the SPE_{x2} are the position of the hot spot (z_2) and the initiator flow rate (F_{i2}) in section 2 and the major contributors to the negative movement of the first latent vector in the monitoring space for section 2 are the position of the hot spot (z_2) and the maximum temperature (T_{max2}). Both these results again point to a change in the extent of reaction in section 2 (with no corresponding change in section 1), and one would immediately suspect an impurity contamination or an initiator efficiency drop localized to the reactor's second section.

Similar results were obtained when impurities entered only the first section of the reactor, when fouling occurred in either section of the reactor, and when either of these faults occurred in both sections of the reactor simultaneously. In all cases the multiblock PLS monitoring charts were slightly more sensitive than the ordinary PLS charts in detecting a change, they clearly revealed the section(s) within which the fault(s) had occurred, and diagnosis was again straightforward by examining the bar charts showing the contribution of the variables in each section to the SPE_x and to the changes in the latent variables for that section. For faults of different types occurring simultaneously in one section, the monitoring charts again rapidly detect the presence of a fault, but diagnosis is more difficult since many variables now contribute to the changes.

Conclusions

In this article, the multivariate SPC procedures for monitoring process operating performance that are based on multivariate statistical projection methods (PCA/PLS) have been extended to allow for a blocking of large processes into smaller sections. A multiblock PLS method is used to develop latent variable subspaces for each block as well as for the overall process. This enables one to establish monitoring charts in these latent variable subspaces, and in the squared prediction errors (SPE_xs) for each of the subspaces. In general, if faults occur in certain sections of the plant, then it is shown that these multiblock monitoring methods are usually able to clearly isolate the section in which the fault occurred, and in general, they are able to detect these faults earlier than their single block PLS counterparts.

Once an event has been detected it is important to diagnose an assignable cause for it. In this article, we present some tools based on interrogating the underlying PLS or multiblock PLS model, that can help in the diagnosis of these events. Although they do not directly identify the assignable cause, they clearly reveal those process variables whose behavior is exceptional with respect to past "in control" behavior. By focusing the engineers' or the operators' attention on a few variables in a certain section of the plant, they can more easily use their underlying process knowledge to diagnose a possible cause.

These methods were evaluated on a simulation of a low density polyethylene tubular reactor system with two reaction sections. The number of variables and blocks was kept small enough in this example to clearly illustrate the approach. However, the power of the methods lies in their ability to treat large processes with many highly correlated variables.

Although we have used multiblock PLS methods, if only process data (X) were available, or if too little productivity and quality data (Y) were available, then multiblock PCA methods could be used instead (Wold et al., 1987). All the monitoring and diagnostic procedures developed here would be equally applicable. The only difference would be that all the variation in the process variables would be used, rather than only that variation that is most predictive of the Y-variables.

Since alternative process monitoring and fault diagnosis methods are available, some comments on the weaknesses and strengths of this approach are warranted. The approach advocated here is a direct extension of Shewhart's SPC philosophy to large multivariable systems. The only knowledge required to develop the PLS models and the monitoring and diagnostic charts is a good historical database of normal or "in control" process operation where good productivity and quality have been achieved. The PLS models are then used to capture the "good" process behavior by defining regions within the low dimensional spaces defined by the dominant latent vectors that contain the projections of these data. The monitoring and diagnostic charts are based on the assumption that future "good" process behavior should continue to resemble past "good" behavior and therefore projections of future data should lie in the same region of these PLS hyperplanes. Any statistically significant deviation from this hypothesis is then classified as a special event or a fault. In this way, the approach is "nondirectional" (Box and Kramer, 1992), that is, it is not aimed at detecting faults of a specific nature. It will detect any abnormal behavior. For this reason, the diagnostic, multiblock methods proposed here, that attempt to isolate possible causes for the event, are a necessary complement to the monitoring charts.

Other monitoring methods which use fundamental models or which have been trained on databases containing specific faults are "directional" methods. They may be more powerful at detecting specific faults than the multivariate statistical methods of this article because they employ additional information. However, comprehensive theoretical models for large processes are difficult to develop and rarely do they model the entire process (such as mechanical equipment) nor all the possible faults. Furthermore, plants with comprehensive databases of faults, required in training knowledge-based models, are rare. Therefore, such "nondirectional" statistically-based approaches are very appealing. The most important factor in making them powerful is that they are truly multivariable. The multivariable statistical methods used to develop the monitoring and diagnosis procedures of this article are particularly appealing, because they compress the information into low dimensional spaces where the interpretation and presentation of the results is as straightforward as conventional univariate SPC charts.

Acknowledgments

The authors wish to thank the Ernest-Solvay-Stiftung for its support in the form of a scholarship for C. Jaeckle, and the Natural Science and Engineering Research Council of Canada for supporting this research.

Notation

a = index for PLS dimensions
A = number of PLS dimensions that are used
E = residual matrix for X in PLS
$E\ell$ = residual matrix for $X\ell$ in multiblock PLS
F = residual matrix for Y
F_α = F-distribution with significance level α
i = index for observations (rows of X, Y)
j = index for variables (columns of X, Y)
p, p_a = loading vector for X in one PLS dimension, in PLS dimension a
$p\ell_a$ = loading vector for block $X\ell$ in multiblock PLS dimension a
P = matrix of loading vectors for X
q, q_a = loading vector for Y in one PLS dimension, in PLS dimension a
Q = matrix of loading vectors for Y
s_i^2 = variance of t_a
S = estimated covariance matrix of $(t_1, t_2)^T$
SPE_X, SPE_Y = squared prediction error calculated on X, Y
$SPE_{X,i}$, $SPE_{Y,i}$ = squared prediction error calculated on X, Y for observation i
$SPE_{X\ell}$ = squared prediction calculated on block $X\ell$
t_a = latent variable
t, t_a = score vector for X in one PLS dimension, in PLS dimension a
T = matrix of score vectors
$t\ell_a$ = score vector for block $X\ell$ in multiblock PLS dimension a
tc_a = consensus score vector in multiblock PLS dimension a
u, u_a = score vector for Y in one PLS dimension, in PLS dimension a
w, w_a = vector of weights for X variables in one PLS dimension, in PLS dimension a
$w\ell_a$ = vector of weights for block $X\ell$ variables in multiblock PLS dimension a
$w\ell_{aj}$ = weight for $X\ell$ block variable j in multiblock PLS dimension a
W = matrix of weight vectors
x = vector of process variables representing one observation
X = matrix of process data (no. of rows = no. of observations; no. of columns = no. of process variables)
$X\ell$ = block ℓ of process data in multiblock PLS
y = vector of quality variables representing one observation
Y = matrix of quality data (no. of rows = no. of observations; no. of columns = no. of quality variables)

Greek letters

β = matrix of regression coefficients between the mean centered and autoscaled X and Y
τ = vector of latent variables $(t_1, t_2, \ldots, t_a)^T$

Literature Cited

Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, Wiley, New York (1958).

Box, G. E. P., and T. Kramer, "Statistical Process Monitoring and Feedback Adjustment," *Technometrics*, **34**, 259 (1992).
Geladi, P., and B. R. Kowalski, "Partial Least-Squares Regression: A Tutorial," *Anal. Chim. Acta*, **185**, 1 (1986).
Höskuldsson, A., "PLS Regression Methods," *J. Chemometrics*, **2**, 211 (1988).
Jackson, J. E., *A User's Guide to Principal Components*, Wiley Interscience, New York (1991).
Kiparissides, C., G. Verros, and J. F. MacGregor, "Mathematical Modeling, Optimization and Control of High-Pressure Ethylene Polymerization Reactors," *J. Macromol. Sci.-Rev. Macromol. Chem. Phys.*, **C33**(4), 437 (1993).
Kresta, J., J. F. MacGregor, and T. E. Marlin, "Multivariate Statistical Monitoring of Process Operating Performance," *Can. J. Chem. Eng.*, **69**, 35 (1991).
MacGregor, J. F., T. E. Marlin, J. Kresta, and B. Skagerberg, "Multivariate Statistical Methods in Process Analysis and Control," *AIChE Symp., Proc. Int. Conf. on Chemical Process Control*, Y. Arkun and W. H. Ray, eds., Padre Island, TX, p. 79 (Feb., 1991a).
MacGregor, J. F., B. Skagerberg, and C. Kiparissides, "Multivariate Statistical Process Control and Property Inference Applied to Low Density Polyethylene Reactors," *IFAC Symp.*, ADCHEM'91, Toulouse, France, p. 131 (Oct., 1991b).
MacGregor, J. F., and P. Nomikos, "Monitoring Batch Processes," NATO Adv. Study Inst. (ASI) for Batch Processing Systems Eng., Antalya, Turkey (May-June, 1992).
Martens, H., and T. Næs, *Multivariate Calibration*, Wiley-Interscience, New York (1989).
Miller, P., R. E. Swanson, and C. F. Heckler, "Contribution Plots: The Missing Link in Multivariate Quality Control, submitted to *J. Qual. Tech.* (1993).
Nomikos, P., and J. F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Processes," submitted to *Technometrics* (1993).
Piovoso, M. J., K. A. Kosanovich, and R. K. Pearson, "Monitoring Process Performance in Real Time," *Proc. Amer. Control Conf.* p. 2359 (1992).
Shewart, W. A., *Economic Control of Quality*, Van Nostrand, New York (1931).
Skagerberg, B., J. F. MacGregor, and C. Kiparissides, "Multivariate Data Analysis Applied to Low-Density Polyethylene Reactors," *Chemometrics and Int. Lab. Syst.*, **14**, 341 (1992).
Slama, C. F., "Multivariate Statistical Analysis of Data from an Industrial Fluidized Catalytic Process Using PCA and PLS," M. Eng. Thesis, Dept. of Chemical Engineering, McMaster Univ., Canada (1991).
Wangen, L. E., and B. R. Kowalski, "A Multiblock Partial Least Squares Algorithm for Investigating Complex Chemical Systems," *J. Chemometrics*, **3**, 3 (1988).
Wise, B. M., D. J. Veltkamp, N. L. Ricker, B. R. Kowalski, S. M. Barnes, and V. Arakali, "Application of Multivariate Statistical Process Control to the West Valley Slurry-Red Ceramic Melter Process," *Waste Management Proc.*, Tuscon (1991a).
Wise, B. M., and N. L. Ricker, "Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity," *Int. Fed. Aut. Cont. Symp.*, ADCHEM '91 Proc., Toulouse, France, p. 125 (1991b).
Wold, H., "Soft Modeling. The Basic Design and Some Extensions," *Systems Under Indirect Observations*, Chap. 1, Vol. 2, K. G. Goreskog and H. Wold, eds., North Holland, Amsterdam (1982).
Wold, S., "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, **20**(4), 397 (1978).
Wold, S., A. Ruhe, H. Wold, and W. J. Dunn, "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM J. Sci. Stat. Comput.*, **5**(3), 735 (1984).
Wold, S., S. Hellberg, T. Lundstedt, and M. Sjöström, "PLS Modeling with Latent Variables in Two or More Dimensions," PLS Meeting, Frankfurt (Sept. 1987).

Appendix A

PLS algorithm

Projection to Latent Structures or PLS is really a class of algorithms that attempts to summarize the variation in a matrix X (of observations on many variables) that is in some way predictive of a corresponding matrix Y (of observations on other variables). Variables in both matrices are usually mean centered and scaled (to unit variance).

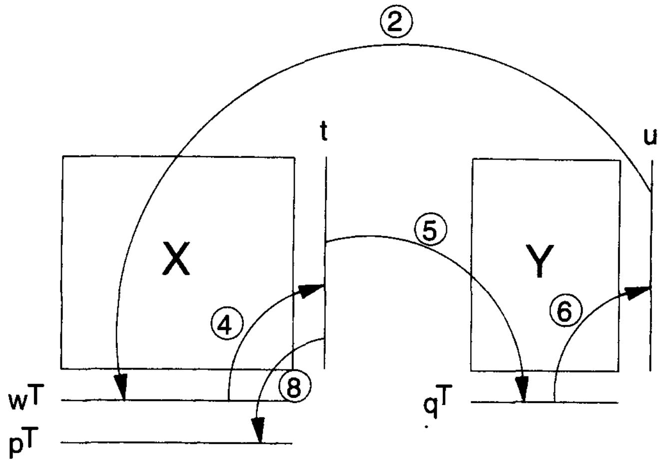

Figure A1. PLS iteration.

The most common form of PLS and the corresponding NIPALS algorithm is summarized below and illustrated in Figure A1:

(1) Start: Set u equal to any column of Y.
(2) Regress columns of X on u to get loadings: $w^T = u^T X / u^T u$.
(3) Normalize w to unit length.
(4) Calculate to scores $t = Xw/w^T w$.
(5) Regress the columns of Y on t: $q^T = t^T Y / t^T t$.
(6) Calculate the new score vector for Y: $u = Yq/q^T q$.
(7) Check convergence of u: if yes go to 8; if no go to 2.
(8) Calculate X matrix loadings by regressing columns of X on t: $p^T = t^T X / t^T t$.
(9) Calculate residual matrices: $E = X - tp^T$; $F = Y - tq^T$.
(10) To calculate the next set of latent vectors replace X and Y by E and F and repeat.

Various interpretations of the PLS algorithm and its properties are discussed by Höskuldsson (1988). An important property is that the t_as are orthogonal to one another.

Appendix B

Multiblock PLS algorithm

The multiblock or hierarchical PLS algorithm presented here is a variation on those of Wold et al. (1987) and Wangen and Kowalski (1988). In this article we only consider the case of multiple X blocks and a single Y block. The iteration sequence is illustrated in Figure B1 for the case of two X blocks and presented below:

(1) Start: Set u equal to a column of Y.
(2) Perform part of a PLS round on each of the blocks ($X1$ and $X2$) to get ($w1$, $t1$) and ($w2$, $t2$) as in steps 2 to 4 in Appendix A.
(3) Collect all the score vectors $t1$, $t2$ in the consensus matrix T.

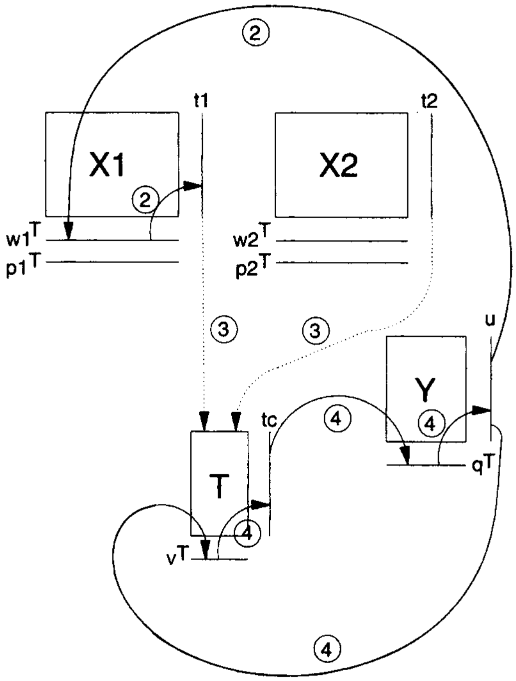

Figure B1. Multiblock PLS iteration.

(4) Make one round of PLS with T as X (steps 2 to 6 in Appendix A) to get a loading vector $\boldsymbol{v}$ and a score vector $\boldsymbol{tc}$ for the T matrix, as well as a loading vector $\boldsymbol{q}$ and a new score vector $\boldsymbol{u}$ for the Y matrix.

(5) Return to step 2 and iterate until convergence of $\boldsymbol{u}$.

(6) Compute the loadings $\boldsymbol{p1} = \boldsymbol{X1}^T\boldsymbol{t1}/\boldsymbol{t1}^T\boldsymbol{t1}$ and $\boldsymbol{p2} = \boldsymbol{X2}^T\boldsymbol{t2}/\boldsymbol{t2}^T\boldsymbol{t2}$ for the $X1$ and $X2$ matrices.

(7) Compute the residual matrices: $\boldsymbol{E1} = \boldsymbol{X1} - \boldsymbol{t1p1}^T$; $\boldsymbol{E2} = \boldsymbol{X2} - \boldsymbol{t2p2}^T$; $\boldsymbol{F} = \boldsymbol{Y} - \boldsymbol{tcq}^T$.

(8) Calculate the next set of latent vectors by replacing $\boldsymbol{X1}$, $\boldsymbol{X2}$ and $\boldsymbol{Y}$ by their residual matrices $\boldsymbol{E1}$, $\boldsymbol{E2}$ and $\boldsymbol{F}$ and repeating from step 1.

Among the properties of this algorithm are that the score vectors $\boldsymbol{t\ell}_a$ ($a = 1, 2, \ldots, A$) for each block ℓ are orthogonal to one another, but the overall score vectors $\boldsymbol{tc}_a$ are not orthogonal to one another. This follows from the computation of the residual matrices $E1$ and $E2$ in step 7 (Wangen and Kowalski, 1988). Alternatively, one could formulate the algorithm to yield orthogonal consensus block scores ($\boldsymbol{tc}_a$), but nonorthogonal block scores ($\boldsymbol{t\ell}_a$) (see Wold et al., 1987).

Manuscript received Apr. 14, 1993, and revision received Jul. 15, 1993.