ELSEVIER

Chemometrics and Intelligent Laboratory Systems 30 (1995) 97–108

Chemometrics and intelligent laboratory systems

Multi-way partial least squares in monitoring batch processes

Paul Nomikos *, John F. MacGregor

Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7

Received 21 December 1994; accepted 10 May 1995

Abstract

Multivariate statistical procedures for monitoring the progress of batch processes are developed. Multi-way partial least squares (MPLS) is used to extract the information from the process measurement variable trajectories that is more relevant to the final quality variables of the product. The only information needed is a historical database of past successful batches. New batches can be monitored through simple monitoring charts which are consistent with the philosophy of statistical process control. These charts monitor the batch operation and provide on-line predictions of the final product qualities. Approximate confidence intervals for the predictions from PLS models are developed. The approach is illustrated using a simulation study of a styrene–butadiene batch reactor.

Keywords: Multivariate analysis; Partial least squares; Batch processes; Statistical process control

1. Introduction

Batch and semi-batch processes play an important role in the chemical industry due to their low volume–high value products. Examples include reactors, crystallization, distillation, and injection molding processes; the manufacturing of polymers, herbicides, insecticides, pharmaceuticals, and biochemicals. Batch processes are characterized by a prescribed processing of materials for a finite duration. Successful batch operation means tracking this prescribed recipe with high degree of reproducibility. Feedrates, temperature and pressure profiles are implemented with servo-controllers, and precise sequencing operations are produced with tools such as programmable logic controllers. Monitoring these batch processes is very important to ensure their safe operation and to assure that they produce consistently high quality products.

Batch processes suffer a lack of reproducibility from batch to batch variations due to disturbances and the absence of on-line quality measurements. These variations may be difficult for an operator to discern, but could have an adverse effect on the final product quality. Often, a disturbance or an operational problem can be undetected and the poor product quality may remain undetected until significant expense has been incurred. A system that monitors the time evolution of a batch and can detect variations from normal operation in the familiar manner of a statistical process control (SPC) chart might allow for corrective action early in the batch, the quick disposition of batches not salvaged, and the diagnosis of assignable causes that can be eliminated from future batches.

* Corresponding author. Present address: DuPont Canada Inc., Research and Development, 461 Front Rd., Kingston, Ontario, K7L 5A5, Canada. email: nomikos@king5.dnet.dupont.com

SSDI 0169-7439(95)00043-7

The main characteristics of batch processes, flexibility, finite duration, nonlinear behavior, and unsteady state, are related both to their success and their incompatibility with the conventional mathematical or empirical modeling for monitoring and controlling continuous processes. For on-line monitoring of batch processes there are two general methodologies. One is based on fundamental mathematical models (Kalman filters; e.g. [1]), and the second on knowledge based models (expert systems, artificial intelligence methods; e.g. [2]). These methods are reviewed by Nomikos and MacGregor [3] and contrasted with statistical approaches based on multi-way principal component analysis (MPCA).

SPC methods in batch processes usually are limited only to end product quality measurements [4,5], or to a single variable measured throughout the batch [6]. The use of statistical methods such as principal component analysis (PCA) and partial least squares (PLS) to the analysis of multivariate continuous chemical process data has been well documented [7–9]. Nomikos and MacGregor [3,10] extended these ideas to the nonlinear and finite duration batch processes. They proposed SPC schemes for batch processes based on MPCA, which utilize directly the information of the on-line measurements and systematically and scientifically recognize significant deviations from the normal operating behavior of the process through simple SPC charts. The behavior of the process is characterized using an empirical model based on the MPCA analysis of data obtained when the process is operating well and is in a state of statistical control. Subsequently, future unusual events are detected by referencing the measured process behavior against this 'in-control' model and its statistical properties.

Most batch and semi-batch processes operate in open loop with respect to product quality variables, simply because few, if any, on-line sensors exist for tracking these variables. Upon completion of the batch a range of quality measurements are usually made on a sample of the product in the quality control laboratory. The MPCA in the proposed SPC schemes [3,10] only makes use of the process variable trajectory measurements ($\underline{\mathbf{X}}$) taken throughout the duration of the batch. Measurements on product quality variables ($\mathbf{Y}$) taken at the end of each batch were used only to help classify a batch as 'good' or 'bad'. However, such product quality data can be used in a much more direct fashion. Multi-way partial least squares (MPLS) can be performed using both the process data ($\underline{\mathbf{X}}$) and the product quality data ($\mathbf{Y}$). Rather than focusing only on the variance of $\underline{\mathbf{X}}$, MPLS focuses more on the variance of $\underline{\mathbf{X}}$ that is more predictive for the product quality $\mathbf{Y}$.

The same multivariate SPC monitoring ideas that were developed using MPCA can be extended directly using MPLS when product quality data ($\mathbf{Y}$) are available. The additional information that one can get from MPLS is an on-line inference of the final quality of the product. This paper describes the monitoring scheme for batch processes based on MPLS, which is analogous to that based on MPCA, and focuses on some statistical properties of PLS for predictions and on the development of control charts for on-line predictions of the final product qualities. A simulation of a styrene–butadiene batch reactor is used to illustrate the ideas of the proposed SPC method.

2. MPLS analysis of batch data

The number of measurements (feedrates, temperatures, pressures, etc.) being made every few seconds over several hours of the batch duration creates a data overload. Some of the variables measured may be redundant and most of them are highly correlated with one another. Not only is the relationship among all the variables at any one time important, but so is the entire past history of the trajectories of all these variables. The aim is to build an empirical model based on the measurements of a reference batch database, which will describe the normal operation of the process ($\underline{\mathbf{X}}$) when it produces good quality product ($\mathbf{Y}$). This empirical model will be used to monitor the evolution of future batch runs.

MPLS [11] is an extension of PLS [12] to handle data in three-dimensional arrays. A historical dataset of batch trajectory data consists of $i = 1,2,..,I$ batch runs where each of them has the on-line measurements of $j = 1,2,..,J$ variables over $k = 1,2,..,K$ time intervals throughout the batch. This vast amount of

data is organized into a three-way array $\underline{\mathbf{X}}$ ($I \times J \times K$) as it is shown in Fig. 1. Different batch runs are organized along the vertical side, the measurement variables along the horizontal side, and their time evolution occupies the third dimension. The final quality variables ($m = 1,2,..,M$) for each batch are summarized in a ($I \times M$) matrix $\mathbf{Y}$. The relation between MPLS and PLS is that MPLS is equivalent to performing ordinary PLS on a large two dimensional matrix $\mathbf{X}$ formed by unfolding the three way array $\underline{\mathbf{X}}$ in one of the six possible ways (two are degenerate cases). For analyzing and monitoring batch processes [3,10], the most meaningful way of unfolding the array $\underline{\mathbf{X}}$ is to put each of its vertical slices ($I \times J$) side by side to the right to create the matrix $\mathbf{X}$ ($I \times JK$), starting with the one corresponding to the first time interval (Fig. 1). Each of the vertical slices of $\underline{\mathbf{X}}$ is a ($I \times J$) matrix representing the values of all the process variables for all the batches at a common time interval k. After the unfolding of $\underline{\mathbf{X}}$, each column of $\mathbf{X}$ and $\mathbf{Y}$ are mean centered and scaled to unit variance prior to perform the ordinary PLS analysis. MPLS in this framework explains the variation of a process variable about its average trajectory at each point of time, as this has been defined from the reference normal database, which is most closely related to the end quality of the product. This subtraction of the average trajectories removes the major nonlinear behavior of the process, leading to a more linear and stationary problem which is suitable for statistical analysis and inference [3,10]. Although other multiway methods [13–16] have been proposed for decomposing such three-way arrays, we focus on MPLS exclusively because of its simplicity and ease of interpretation.

For batch data, PLS decomposes the $\mathbf{X}$ ($I \times JK$) and $\mathbf{Y}$ ($I \times M$) matrices into a summation of R score vectors ($\mathbf{t}$ ($I \times 1$)) and loading vectors ($\mathbf{p}$ ($JK \times 1$), $\mathbf{q}$ ($M \times 1$), plus some residual matrices ($\mathbf{E}$ ($I \times JK$), $\mathbf{F}$ ($I \times M$)):

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r' + \mathbf{E}, \quad \mathbf{Y} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{q}_r' + \mathbf{F} \tag{1}$$

or if we combine the $\mathbf{t}$, $\mathbf{p}$, and $\mathbf{q}$ vectors into $\mathbf{T}$ ($I \times R$), $\mathbf{P}$ ($JK \times R$), and $\mathbf{Q}$ ($M \times R$) matrices

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}, \quad \mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F} \tag{2}$$

where $\mathbf{T}$ is given by:

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \tag{3}$$

The $\mathbf{w}$ vectors are orthonormal, the $\mathbf{t}$ vectors are orthogonal, and the matrix ($\mathbf{P}'\mathbf{W}$) is upper triangular with ones as diagonal elements [17]. This decomposition summarizes and compresses the data with respect to both x and y variables and time into low dimensional spaces that describe the operation of the process which is most relevant to final product quality. Each row of the $\mathbf{T}$ matrix corresponds to a single batch and depicts the overall variability of this batch with respect to the other batches in the database. The $\mathbf{P}$ and $\mathbf{W}$ matrices summarize the time variation of the measurement variables about their average trajectories, and their elements give the weights applied to each variable at each time interval within a batch to give the t-scores for that batch. The $\mathbf{Q}$ matrix relates the variability of the process measurements to the final product qualities.

As in MPCA, batches with unusual operation will appear in MPLS either as batches with large t-scores, or with large residuals in the x-space ($Q_x = \sum_{c=1}^{KJ} \mathbf{E}(i, c)^2$), or with both. Additionally in MPLS, if the residuals for a batch in the y-space ($Q_Y = \sum_{c=i}^{M} \mathbf{F}(i, c)^2$) are large, it means that its final product qualities are not well predicted by its process measurements. MPCA has proven very useful in the post analysis of batch runs and has shown its abilities, both in simulated examples and in real industrial data [3,10,18], to discriminate between normal and abnormal batches and detect abnormalities which are difficult to detect by visual inspection only of the measurement trajectories. MPLS shares these benefits. The power of both MPCA and MPLS results from using the covariance matrix of the variable trajectories. By doing this, both methods utilize not just the magnitude of the deviations of each process variable from its mean trajectory but also both their simultaneous and temporal correlations.

The assumptions behind these methods, as in all inferential methods, are that one has 'comparable' runs and 'observable' events of interest. The first assumption states that future batches will operate in a similar way with those in the reference database. If something changes in the operation of the batch (e.g. amount or type of catalyst), it will make all subsequent batches operationally different from the previ-

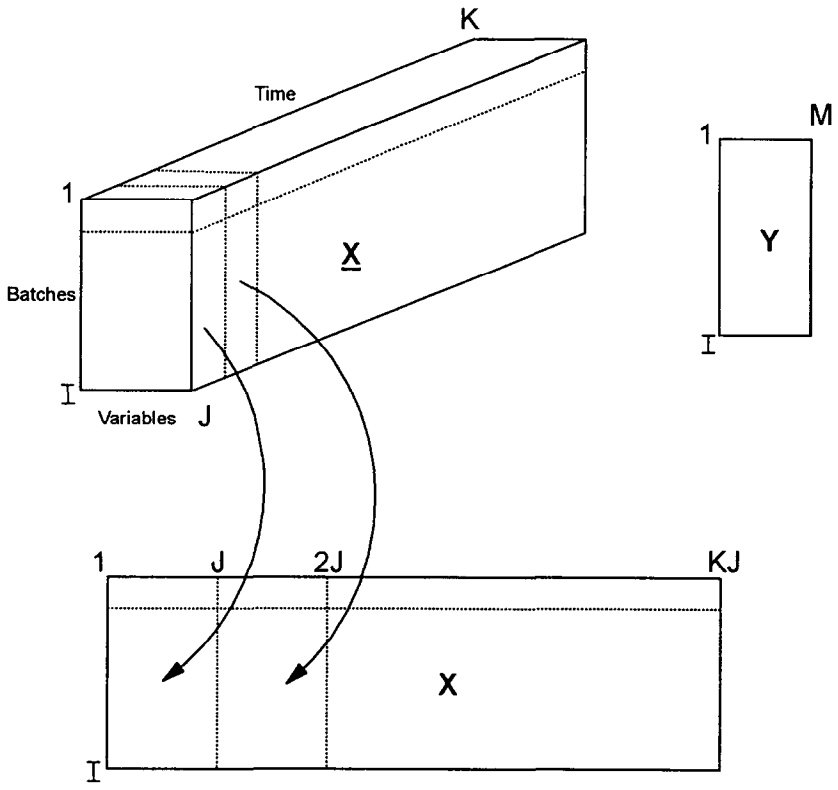

Fig. 1. Arrangement of batch data in MPLS. I is the number of batches, J is the number of process measurement variables, K is the number of time intervals, and M is the number of product quality variables. The three-way array $\underline{\mathbf{X}}(I \times J \times K)$ unfolds into a matrix $\mathbf{X}(I \times JK)$, and a normal PLS can be performed between the $\mathbf{X}$ and $\mathbf{Y}$ matrices.

ous ones. In this case one has to build a new database which incorporates the change and re-apply the method. The second assumption requires the measurements to contain some information about an abnormality, in order that the method to be able to detect it. If there is no information in the data about a fault, then no method can detect it.

3. SBR example

The application of MPLS in batch data is illustrated here with a detailed mechanistic model for semi-batch emulsion polymerization of styrene–butadiene rubber (SBR) [19]. Using typical variations in the initial charge of materials and impurities, and in the process operations, a number of batches were simulated. Details of the simulations can be found in [3,20], and the data used in this article are available from the authors upon request. Fifty batches which gave final latex with quality properties within an acceptable region were selected to provide a reference data array. On-line measurements were assumed to be available on nine variables: the feedrates of styrene and butadiene monomers, the temperature of the feed stream, the reactor contents, the cooling water, and the reactor jacket, the latex density, the total conversion, and the instantaneous heat release from an energy balance. Using 200 time increments over the duration of the batch, the reference dataset $\underline{\mathbf{X}}$ was a $(50 \times 9 \times 200)$ array. The resulting latex and polymer properties of the product were summarized in five quality variables ($\mathbf{Y}$ (50×5)): composition (% styrene), particle size (Å), branching (branches/reacted monomer units), crosslinking (crosslinks/reacted monomer units), and polydispersity.

The ability of MPLS to discriminate between

batches with acceptable product and 'bad' batches was tested through a post analysis of the 50 'good' batches plus some 'bad' ones with product quality barely outside the acceptable region. MPLS was able to detect clearly these 'bad' batches by placing them in the reduced space (t-plots) away from the main central cluster formed by the 50 'good' batches. Having established the observability of faults, an MPLS model was built from the 50 'good' batches, which summarizes the information contained in them about the normal operation of the process. This model will be used as the statistical reference to classify new batches as 'good' or 'bad' and give on-line predictions of their final qualities.

Two PLS components were needed, as determined by cross-validation [21], to capture the variation of the process variables about their average trajectories which is most predictive for the final product qualities. The cumulative percentage sum of squares explained (%SS) by the two principal components of the **X** and **Y** matrices and of each quality variable separately, is given in Table 1. One should always use cross-validation to determine the number of components in the PLS model and to assert its predictability. To rely only on the percentage of explained **Y** will be misleading because of the large number of predictor x-variables ($9 \times 200 = 1800$ in the SBR example). Any regression model could have accounted for a large portion of the variability in the **Y**.

The last row in Table 1 is the regression statistic mean sum of squares due to regression (MSR) over the mean squared error (MSE) (see Section 5) with its 95% critical value, which shows how well the x-data account for the variation in each y-variable. These F-tests provide another way from a regression point of view to check how well each of the y-variables is explained by the MPLS model. As it can be seen from Table 1, quality variables 3 and 4 are explained very well from the MPLS model and only quality variable 2 (particle size) is poorly explained by the process measurements. This arises because the particle size is determined largely by the variation in the number of seeded particles charged initially in the reactor, and it is not influenced much by resulting process conditions.

Plots of the latent vectors $\mathbf{t}_r$ versus $\mathbf{u}_r$ ($\mathbf{u}_r = \mathbf{Y}_{r-1}\mathbf{q}_r/(\mathbf{q}_r'\mathbf{q}_r)$, where $\mathbf{Y}_{r-1}$ is the residual matrix of **Y** after extracting $r-1$ components) are shown in Fig. 2. The linear nature of these plots suggests that nonlinear PLS [22] would probably not be needed. Indeed, performing such a nonlinear PLS gave essentially identical results to the linear analysis. The particular unfolding of X that is been used and the subtraction of the average trajectories from the process measurements have apparently eliminated most nonlinear effects in the data.

4. On-line monitoring

The central idea in on-line monitoring is as follows. An MPLS model is built using a database of 'good' batches which yielded acceptable product and did not exhibit any operational problems. Subsequent batches are then referenced against this 'in-control' model. The **W**, **P**, and **Q** matrices from such an analysis contain all the structural information about how the process measurements deviate from their mean values at each time interval and how these are related to the final quality variables. The predicted t-scores ($\hat{\mathbf{t}}$ $(1 \times R)$), the predicted quality variables ($\hat{\mathbf{y}}$ $(1 \times M)$), and the residuals ($\mathbf{e}$ $(1 \times KJ)$, $\mathbf{f}$ $(1 \times M)$) for a new batch $\mathbf{X}_{\text{new}}$ $(K \times J)$ are given by:

unfold and scale $\mathbf{X}_{\text{new}}(K \times J)$ to $\mathbf{x}_{\text{new}}(1 \times KJ)$

$$\hat{\mathbf{t}} = \mathbf{x}_{\text{new}}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1},\ \hat{\mathbf{y}} = \hat{\mathbf{t}}\mathbf{Q}',\ \mathbf{e} = \mathbf{x}_{\text{new}} - \hat{\mathbf{t}}\mathbf{P}',$$
$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{y}} \qquad (4)$$

Table 1
Cumulative percentage sum of squares explained by the two principal components of the **X** and **Y** matrices

%SS	**X**	**Y**	$Y1$	$Y2$	$Y3$	$Y4$	$Y5$
PC1	14.82	57.10	52.87	7.93	91.21	91.23	42.24
PC2	23.05	65.08	54.30	20.79	91.28	91.29	67.74
MSR/MSE ($F_{2,47,0.05} = 3.20$)			27.92	6.17	245.97	246.21	49.93

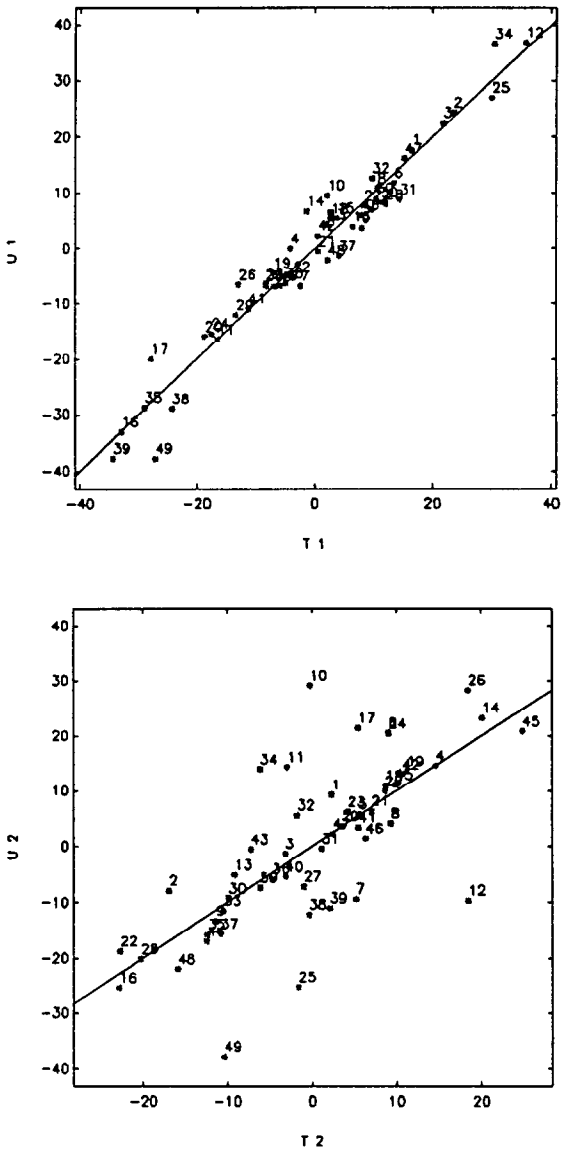

Fig. 2. **t** versus **u** plots. Each point represents one of the fifty batches in the reference database. The t and u observations, for both PLS components, fall close to the diagonal line of the graph. This indicates that the PLS latent variables in the x- and y-space (t, u) are well correlated.

The problem which arises in the on-line application of the above equations is that the $\mathbf{X}_{new}$ matrix is not complete until the end of the batch operation. At time interval k, the matrix $\mathbf{X}_{new}$ has only its first k rows complete and it is missing all the future observations ($K-k$ rows). Several approaches have been studied to overcome this problem [10]. The approach shown here uses the ability of PLS to handle missing data [23]. PLS does this by projecting the already known observations up to time interval k ($\mathbf{x}_{new,k}(1 \times kJ)$) into the reduced space defined by the **W** and **P** matrices in a sequential manner as following:

at each time interval k

for $r = 1$ to R

$$\hat{\mathbf{t}}(1,r) = \mathbf{x}_{new,k}\mathbf{W}(1{:}kJ,r) / (\mathbf{W}(1{:}kJ,r)'\mathbf{W}(1{:}kJ,r)) \quad (5)$$

$$\mathbf{x}_{new,k} = \mathbf{x}_{new,k} - \hat{\mathbf{t}}(1,r)\mathbf{P}(1{:}kJ,r)'$$

end

where the symbol $(1{:}kJ,r)$ indicates the elements of the rth column from the first row up and to the kJth row. PLS, essentially, predicts these missing values by restricting them to be consistent with the already known values, and with the correlation structure of the process variables as defined by the **W** and **P** matrices. This approach gives t-scores very close to their final values as the $\mathbf{X}_{new}$ becomes complete, but during the first few time intervals may give poor estimates of the t-scores since there is so little information to work with. However, in our experience with MPCA and MPLS on this and other examples [3,10,24], this method works well by the time one has about 10% of the batch history. The reasons for this is that one is not building a PLS model based on large amounts of missing data, but using an already well established PLS model to predict the future behavior of a new batch. Furthermore, the early data are complete for all the measurement variables up to the current time interval, and are very good for predicting the future trajectory deviations which arise from variations in the initial batch charge conditions (i.e. impurities, particle concentrations, etc.).

Now, one can calculate at each time interval the predicted t-scores, the predicted final quality variables, and the residuals. Note that there are no residuals (**f**) in the y-space during the on-line monitoring since the actual values of the quality variables will be known only at the end of the batch. If a new batch is still operating in the same way as the batches in the normal database, but has a larger than normal variation in its measurements, this will show up in the t-scores which will place the new batch away from the

origin of the reduced x-space (t-plots). In the case where a new type of variation occurs, the new batch data will move away from the reduced x-space defined by the MPLS model, and its residuals will be large. In this case, the squared prediction error (SPE) associated with the latest on-line measurements at time interval k ($\mathrm{SPE}_k = \sum_{c=(k-1)J+1}^{kJ} \mathbf{e}(1, c)^2$) which represents the perpendicular distance of the instantaneous batch process measurements from the reduced x-space, will indicate the particular instant that something behaves abnormally [10]. Thus, one has to monitor the t-scores and the SPE for a new batch by using SPC charts. These charts are easily implemented and easily interpreted. Multivariate hotelling

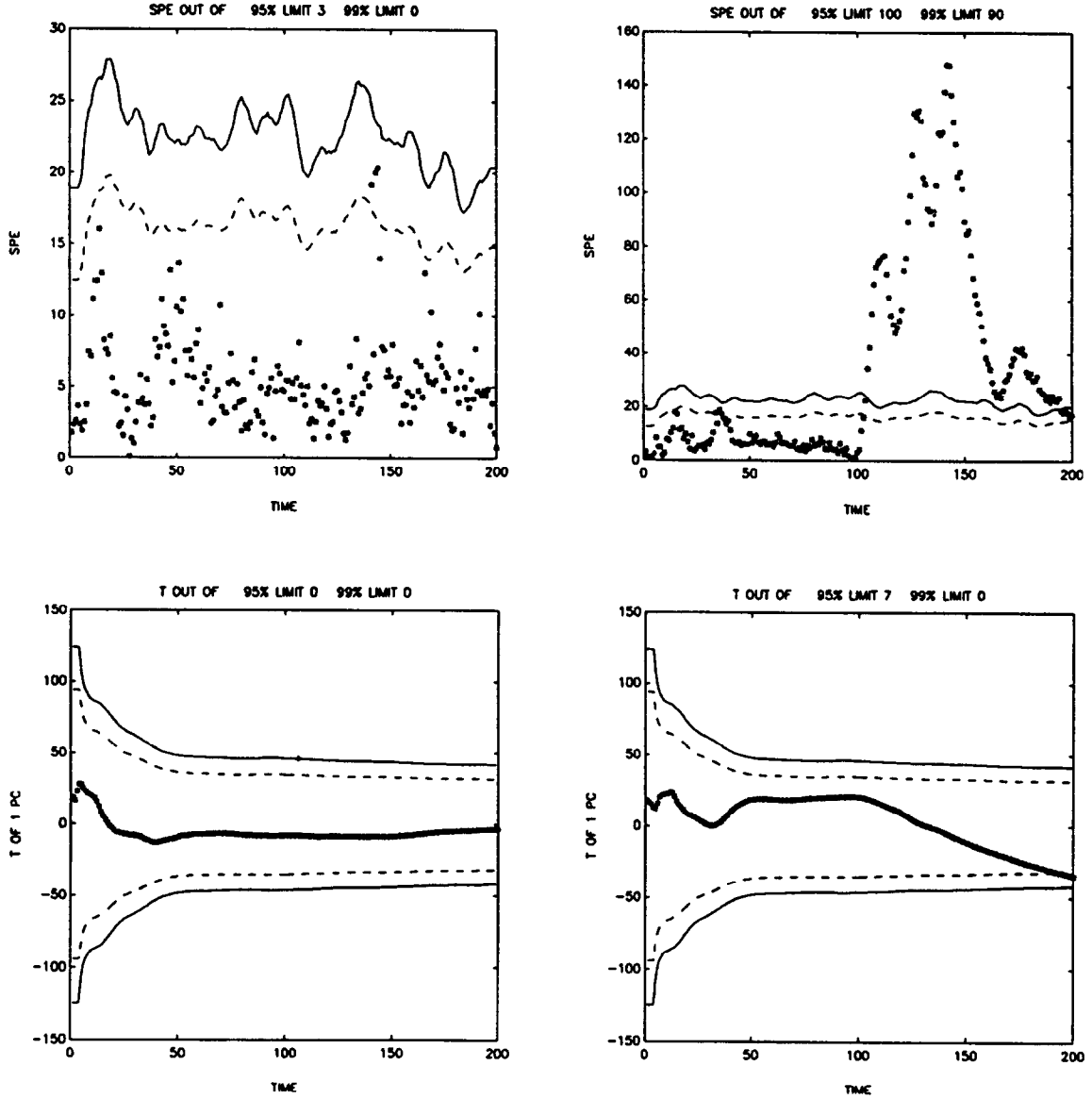

Fig. 3. Monitoring charts for the SPE and t_1-scores with their 95% and 99% control limits (dashed and solid lines) for the new 'good' batch (left hand side plots) and the for 'bad' batch with the problem half-way through its operation (right hand side plots). The abnormality in the 'bad' batch is clearly flagged in the SPE chart after time interval 105.

statistics can be used to monitor the overall performance of the t-scores and the predicted final qualities [10,25]. If an abnormal situation is detected by either of these charts, one can diagnose the fault by interrogating the underlying MPLS model to find which process variables are primarily responsible for the detected deviations. This diagnostic information can be found by checking the contribution of each process variable to the deviations observed in the t-scores and residuals [20,24,26,27].

The control limits for the monitoring charts are derived from the statistical properties of the historical reference distribution of past normal batches [10]. Each of the fifty 'good' batches in the reference database was passed through the on-line monitoring algorithm given above, and their t-scores for each PLS component and SPE were collected at each time interval. This provided fifty observations at each time interval for the t-scores and SPE which were used to construct the control limits for future observations.

Two additional batches were simulated to provide examples for on-line monitoring. One was a 'good' batch, and the other a 'bad' one in which the level of organic impurities in the butadiene monomer feed to the reactor, increased by 50% halfway through the batch operation (at time interval 100). This latter fault is an incipient one, typical in industry, where the abnormal operation develops slowly. The final product from this 'bad' batch was slightly out of the acceptable quality region. Fig. 3 shows the on-line monitoring charts, with their 95% and 99% control limits, for the two new batches. The new 'good' batch shows no abnormality in any of these charts. The 'bad' batch with the problem half-way through its operation, is clearly flagged as abnormal in the SPE chart around time interval 105. After this time interval the observations from this batch move away from the reduced x-space. The MPLS model is not any longer valid, and one should treat the predicted t-scores with caution.

5. Confidence intervals for the final quality variables

Although the SPC charts based on MPLS for monitoring a batch process can be constructed similar to those based on MPCA, MPLS additionally provides predictions for the final product qualities. MPLS gives, at each time interval, predictions of the final quality variables ($\mathbf{Y}$) of the product. These predictions do not have anything to do with the actual values of the quality variables at the given time interval. They only refer to the values which the product quality variables will have upon completion of the batch. In this section we develop approximate confidence intervals for these predictions which can be easily calculated. We treat the problem in its general form and the prediction confidence intervals are applicable to any PLS study.

A major problem in the statistical analysis of PLS is the nonlinear extraction of the PLS components. PLS does not only look at the conditional distribution of the y-variables given the x-observations, but treats both x and y as random variables connected through the latent variables t. In the following we shall treat only the case with univariate $\mathbf{y}$, for the multivariate case ($\mathbf{Y}$) one has to treat each of the y-variables separately. The $\mathbf{X}$ and $\mathbf{Y}$ matrices are assumed to be mean centered. The statistical properties which we shall derive in this section, are based on the work of Searle [28] for regression based on generalized inverses.

The regression problem $\mathbf{y} = \mathbf{X}\beta$ can always get a solution in the following form:

$$\mathbf{b} = \mathbf{G}\mathbf{y} \tag{6}$$

where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}$.

PLS gives a right weak generalized inverse $\mathbf{G}$ of the PLS approximation of the $\mathbf{X}$ matrix $\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}'$ ($\hat{\mathbf{X}}\mathbf{G}\hat{\mathbf{X}} = \hat{\mathbf{X}}$, $\mathbf{G}\hat{\mathbf{X}}\mathbf{G} = \mathbf{G}$, $\hat{\mathbf{X}}\mathbf{G} = (\hat{\mathbf{X}}\mathbf{G})'$) which is given by:

$$\mathbf{G} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}' \tag{7}$$

Rao and Mitra [29], and Boullion and Odell [30] show that a right weak generalized inverse of $\hat{\mathbf{X}}$, which has the same rank as $\hat{\mathbf{X}}$, gives the least squares solution $\mathbf{b}$ for the problem $\mathbf{y} = \hat{\mathbf{X}}\beta$, in which:

$$|\hat{\mathbf{X}}\mathbf{b} - \mathbf{y}| \leq |\hat{\mathbf{X}}\mathbf{g} - \mathbf{y}|, \ \forall \mathbf{g} \tag{8}$$

PLS gives the above least squares solution which has unique minimum $|\hat{\mathbf{X}}\mathbf{b} - \mathbf{y}|$, but the $\mathbf{b}$ and $\mathbf{G}$ are not defined uniquely. The property of invariance of generalized inverses, guarantees that the predicted $\hat{\mathbf{y}}$ has a unique value ($\mathbf{X}\mathbf{b}$) no matter what right weak generalized inverse of $\hat{\mathbf{X}}$ one choose to use. Al-

though, PLS does not provide the minimum norm solution (min|**b**| which is unique), its solution is close to this, since the matrix $\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{P}'$ is generally close to being symmetric and thus **G** would have been also a left weak generalized inverse of $\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}\mathbf{G}\hat{\mathbf{X}} = \hat{\mathbf{X}}$, $\mathbf{G}\hat{\mathbf{X}}\mathbf{G} = \mathbf{G}$, $\mathbf{G}\hat{\mathbf{X}} = (\mathbf{G}\hat{\mathbf{X}})'$). PLS provides the Moore–Penrose generalized inverse of the original **X** for the full rank decomposition of **X**, and in the case where **X** has full column rank, PLS provides the ordinary least squares solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

To proceed with Searle's analysis [28], one needs a generalized inverse of $\hat{\mathbf{X}}'\hat{\mathbf{X}}$. PLS gives the following reflexive generalized inverse **Z** for $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{Z}\hat{\mathbf{X}}'\hat{\mathbf{X}} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$, $\mathbf{Z}\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{Z} = \mathbf{Z}$):

$$\mathbf{Z} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}(\mathbf{T}'\mathbf{T})^{-1}(\mathbf{W}'\mathbf{P})^{-1}\mathbf{W}' \tag{9}$$

Define the indempotent matrix $\mathbf{H} = \mathbf{G}\hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{P}'$ ($\mathbf{H}^2 = \mathbf{H}$), which has rank equal to the number of PLS components we have extracted (rank(**H**) = R). Under the assumption that the y-variable is distributed normally as $N(\mathbf{X}\beta, \sigma^2)$, and that **Z** is independent of the y-variable, we get the following statistical analysis:

$$\mathbf{b} = \mathbf{Z}\hat{\mathbf{X}}'\mathbf{y} + ((\mathbf{H} - \mathbf{I})\mathbf{g} \text{ for arbitrary } \mathbf{g}$$

$$(\text{from now on, assume } \mathbf{g} = 0) \tag{10}$$

$$E(\mathbf{b}) = \mathbf{H}\beta \quad \mathbf{b} \text{ is a biased estimator of } \beta \tag{11}$$

$$\text{var}(\mathbf{b}) = \mathbf{Z}\sigma^2 \tag{12}$$

The statistical test which we can derive from the above results, in the usual regression notation, is summarized in Table 2.

The statistic MSR/MSE is distributed as an F distribution with R and $I - R - 1$ degrees of freedom [28]. The problem is that this statistic does not check for significant regression ($\beta \neq 0$). It tests for the null hypothesis $(\mathbf{X}\beta) = 0$. The only conclusion we can derive, if this test is significant, is that the PLS model accounts for a significant portion of the variation in the y-variable. The β is not an estimable function, since **b** is not invariant to the generalized inverse of $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ that is used for **Z** (**b** has an infinite number of descriptions). The only estimable function is any quantity $\mathbf{c}'\beta$, in which $\mathbf{c}'\mathbf{H} = \mathbf{c}'$. This $\mathbf{c}'\beta$ has $\mathbf{c}'\mathbf{b}$ as its best linear unbiased estimator which is distributed normally as:

$$\mathbf{c}'\mathbf{b} \sim N(\mathbf{c}'\beta, \mathbf{c}'\mathbf{Z}\mathbf{c}'\sigma^2) \tag{13}$$

This shows that we are not able in general to test for the significance of each coefficient separately, but only certain linear combinations of them. In spite of this, PLS provides a way to derive confidence intervals for the predicted y-variable since a new set of observations $\mathbf{x}_{\text{new}}$ can be decomposed as $\hat{\mathbf{x}}_{\text{new}} = \hat{\mathbf{t}}\mathbf{P}'$, and $\hat{\mathbf{x}}_{\text{new}}\beta$ is an estimable function ($\hat{\mathbf{x}}_{\text{new}}\mathbf{H} = \hat{\mathbf{x}}_{\text{new}}$). The confidence intervals at significance level α for an individual y-response are given by:

$$\hat{y} \pm t_{I-R-1,\alpha/2}(\text{MSE})^{1/2}\left(1 + \hat{\mathbf{t}}(\mathbf{T}'\mathbf{T})^{-1}\hat{\mathbf{t}}'\right)^{1/2} \tag{14}$$

where **T** and MSE is the t-score matrix and mean squared error of the PLS analysis of the data upon the PLS model was built, and $t_{I-R-1,\alpha/2}$ is the critical value of the Studentized variable with I-R-1 degrees of freedom at significance level $\alpha/2$.

The motivation behind the above analysis was to develop a simple expression for approximate confidence intervals for the PLS predictions. Eq. (14) is general for any PLS study. If one drops the (1) in the $(1 + \hat{\mathbf{t}}(\mathbf{T}'\mathbf{T})^{-1}\hat{\mathbf{t}}')$ term, one gets the equation for confidence intervals of the expected value of a y-response ($\hat{\mathbf{x}}_{\text{new}}\mathbf{b}$). Of course, the assumption that **Z** is not a function of the y-variable is incorrect. Phatak et al. [31] recognized this, and for the case of univariate **y** did a first order linear approximation of **b** around a set of observations ($\mathbf{X}_0$, $\mathbf{y}_0$) to get improved confidence intervals for $\hat{y}$. Although his approach is more accurate than the zero order approximation used here, it is computationally much more time consuming.

Fig. 4 shows the on-line predictions with their 95% and 99% confidence intervals, for three of the five quality variables for the two new batches. The predictions for the 'normal' batch, match well the actual final quality values of the product. For the 'bad' batch, the predictions capture its problem and stretch their values towards their actual final qualities. Quality variable two, which was poorly explained in the MPLS model, has the poorest predictions for the

Table 2
Statistical test

$\text{SSR} = \hat{\mathbf{y}}'\hat{\mathbf{y}}$	d.f. = R	$\text{MSR} = \text{SSR}/R$
$\text{SSE} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$	d.f. = $I - R - 1$	$\text{MSE} = \text{SSE}/(I - R - 1) = \hat{\sigma}^2$

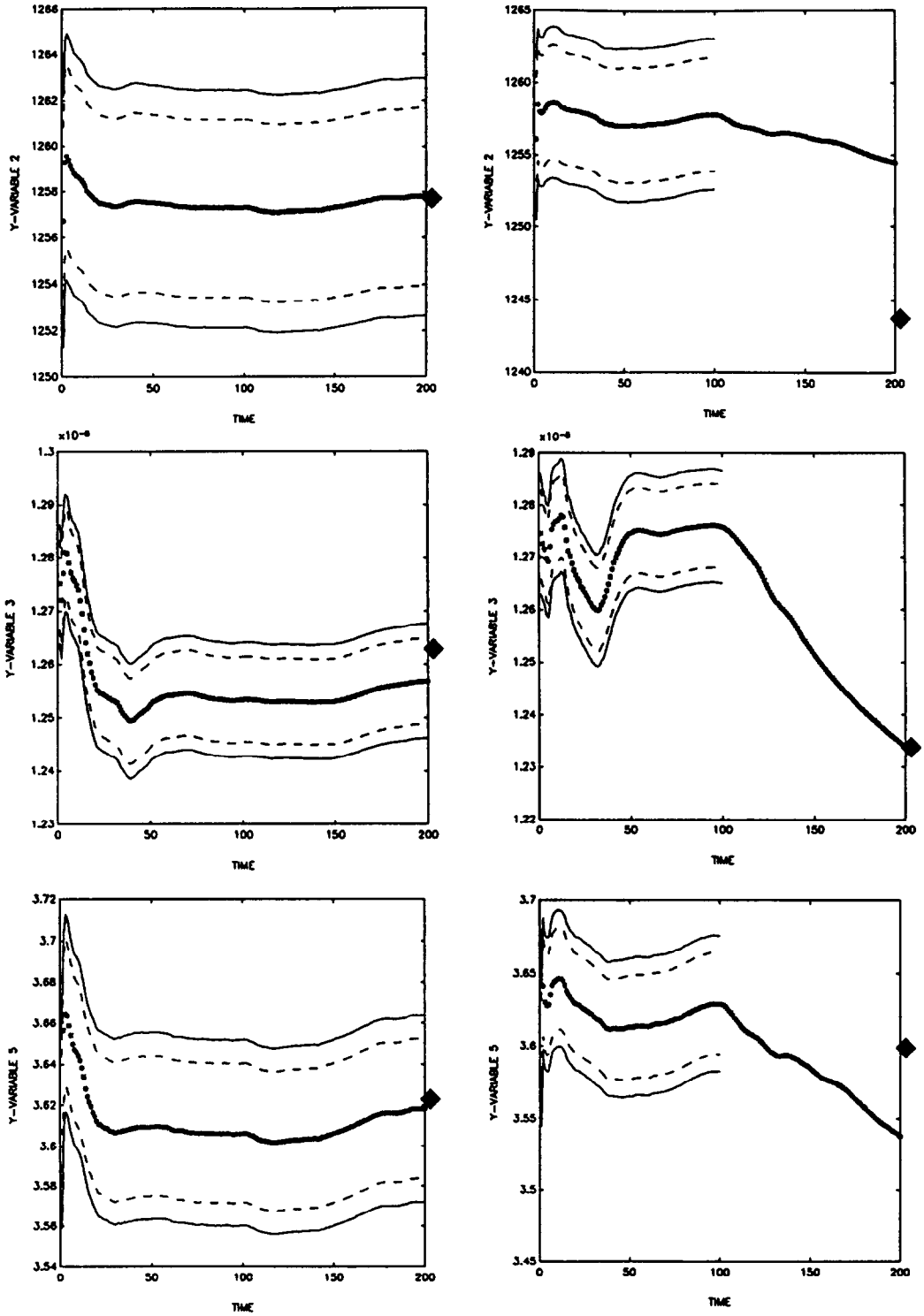

Fig. 4. On-line predictions, with their 95% and 99% confidence intervals (dashed and solid lines), for three of the five final product quality variables for the 'good' batch (left hand side plots) and for the 'bad' batch with the problem half-way through its operation (right hand side plots). The actual final product qualities are indicated by diamond marks. The PLS model for the 'bad' batch is not valid after time interval 105 (absence of confidence intervals) and its predictions are not generally trustworthy.

'bad' batch. On the other hand, quality variable three, which was very well explained in the MPLS model, has very good predictions and its final prediction is very close to the actual one for the 'bad' batch. Since the MPLS model is no longer valid for the 'bad' batch after time interval 105 (the SPE exceeds its control limits in Fig. 3), the predictions after this time interval are not trustworthy. The confidence limits given by Eq. (14) will no longer be valid, and they have not be plotted beyond this time interval in Fig. 4. Although the predictions may not be accurate, the directions that the quality variables will take can be trusted in general, and this can help considerably in diagnosing the source of the abnormality. This is another advantage of PLS with respect to other regression methods, it models both x- and y-spaces to give good predictions and also provides a measure through its residuals in the x-space of how well the PLS model can be trusted.

6. MPCA or MPLS

The question that arises in monitoring batch processes is whether to use MPCA or MPLS. MPCA uses only the information about the process operational behavior (x-data) and its model describes how the on-line process measurements deviate from their average trajectories when the process operates in an 'in-control' state. As a consequence, it will flag any abnormality in the process measurements even though it may be irrelevant to the quality of the product. As an example, a batch-run may have a slightly different agitator power profile because of a deterioration in its agitator mechanism. This event will cause an alarm in the MPCA monitoring. If the agitator power is not correlated with the final product qualities, the MPLS monitoring may not detect this deterioration in the agitator. Therefore, which approach one uses will depend upon whether or not one is primarily interested in events that will probably offset product quality or in any type of abnormal behavior. In general, it may be beneficial to try to detect all process deteriorations and correct them before they lead to permanent malfunctions.

A difficultly with using MPLS to analyze and monitor batch processes is having a sufficient number of quality variables which describe adequately the product quality. Batch quality variables typically are measurements of physical properties of the product, or variables which indicate if the product will have acceptable operation in the next stage, and sometimes variables from customer feedback. For an effective PLS, the **Y** matrix should have as columns, quality variables which are closely connected with the batch process, as to be well correlated with the process measurements. Also, these quality measurements should span a wide range of product properties because it is hard to believe that the whole batch operation can be reflected to a single quality measurement, or that one quality measurement can capture all the quality aspects of the final product. Another difficulty with the batch quality measurements is that they are usually susceptible to a significant amount of measurement error. In such cases, the uncertainty in the quality measurements can make the use of MPLS inappropriate.

7. Conclusions

Batch monitoring methods based on MPCA have been extended by using MPLS. This extension allows one not only to utilize the historical data on the measured process variable trajectories, but also on the final quality measurements at the end of each batch. In addition to monitoring the process variable space, MPLS gives on-line predictions for the final product qualities. Approximate confidence intervals have been developed for these predictions. The proposed monitoring schemes have shown, via a polymerization simulation, that they are able to detect clearly and quickly an abnormality.

When additional information about the initial conditions and set-up of the batch process is available, one may use a multi-block method based on MPCA or MPLS [20,24,27,32] to incorporate this information into the monitoring scheme. Such prior information can be organized in a new matrix which may have variables like feed quality measurements, initial amounts of initiator or emulsifier, preprocessing conditions such as preheat duration, position of the batch in the cleaning cycle, operator on duty, etc.

The methodology presented here is generic in that it is easily applied to almost any batch or semi-batch process, and provides a base for continuous improve-

ment. The proposed monitoring charts are in accordance with the SPC requirements, in that they can be easily displayed and diagnose a fault. In addition, the data reduction and the light computational requirements of the proposed methods do not impose any problem in their implementation. The objective of the monitoring procedure is to detect faults, diagnose them, and eliminate their cause and thereby shrink the control limits and work towards a more consistent quality product.

References

[1] R. Iserman, Automatica, 20 (1984) 387.
[2] G.J. Birky and T.J. McAvoy, Comput. Chem. Eng., 14 (1990) 713.
[3] P. Nomikos and J.F. MacGregor, AIChE J., 40 (1994) 1361.
[4] G.J. Hahn and M.B. Cockrum, J. Appl. Stat., 14 (1987) 35.
[5] S.A. Vander Wiel, W.T. Tucker, F.W. Faltin and N. Doganaksoy, Technometrics, 34 (1992) 286.
[6] C.E. Marsh and T.W. Tucker, ISA Trans., 30 (1991) 39.
[7] J. Kresta, J.F. MacGregor and T.E. Marlin, Can. J. Chem. Eng., 69 (1991) 35.
[8] B. Skagerberg, J.F. MacGregor and C. Kiparissides, Chemom. Intell. Lab. Syst., 14 (1992) 341.
[9] B.M. Wise, N.L. Ricker, D.F. Veltkamp and B.R. Kowalski, Process Contr. Qual., 1 (1990) 41
[10] P. Nomikos and J.F. MacGregor, Technometrics, 37 (1995) 41.
[11] S. Wold, P. Geladi, K. Esbensen and J. Ohman, J. Chemom. 1 (1987) 41.
[12] P. Geladi and B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 1.
[13] P. Geladi, Chemom. Intell. Lab. Syst., 7 (1989) 11.
[14] A.K. Smilde and D.A. Doornbos, J. Chemom., 5 (1991) 345.
[15] A.K. Smilde, Chemom. Intell. Lab. Syst., 15 (1992) 143.
[16] E. Sanchez and B.R. Kowalski, J. Chemom., 4 (1990) 29.
[17] A. Hoskuldsson, J. Chemom., 2 (1988) 211.
[18] K.A. Kosanovich, M.J. Piovoso, K.S. Dahl, J.F. MacGregor and P. Nomikos, Am. Control Conf. '94, IFAC, June 29–July 1, Baltimore, Maryland, 1994.
[19] T.O. Broadhead, A.E. Hamielec and J.F. MacGregor, Makromol. Chem. Suppl., 10 (1985) 105.
[20] P. Nomikos, Multivariate Statistical Process Control of Batch Processes, Ph.D. Thesis, Department of Chemical Engineering, McMaster University, Canada, 1995.
[21] S. Wold, Technometrics, 20 (1978) 397.
[22] S. Wold, Chemom. Intell. Lab. Syst., 14 (1992) 71.
[23] P. Nelson, P.A. Taylor and J.F. MacGregor, Chemom. Intell. Lab. Syst., submitted.
[24] T. Kourti, P. Nomikos and J.F. MacGregor, J. Process Contr., submitted.
[25] N.D. Tracy, J.C. Young and R.L. Mason, J. Qual. Technol., 24 (1992) 88 .
[26] P. Miller, R.E. Swanson and C.E. Heckler, J. Qual. Technol., submitted.
[27] J.F. MacGregor, C. Jaeckle, C. Kiparissides and M. Koutoudi, AIChE J., 40 (1994) 826.
[28] S.R. Searle, Matrix Algebra Useful for Statistics, Wiley, New York, 1982.
[29] C.R. Rao and S.K. Mitra, Generalized Inverse of Matrices and its Applications, Wiley, New York, 1971.
[30] T.L. Boullion and P.L. Odell, Generalized Inverse Matrices, Wiley-Interscience, New York, 1971.
[31] A. Phatak, P.M. Reilly and A. Penlidis, Anal. Chim. Acta, 277 (1993) 495.
[32] L.E. Wangen and B.R. Kowalski, J. Chemom., 3 (1988) 3.