PCA with Efficient Statistical Testing Method for Process Monitoring

Fangping Mu and Venkat Venkatasubramanian*

Laboratory for Intelligent Process Systems, School of Chemical Engineering
Purdue University, West Lafayette, IN 47907, USA

Abstract

Principal component analysis (PCA) has been used successfully for fault detection and identification in processes with highly correlated variables. The fault detection decision used depends solely on the current sample though the results of previous samples are available and is based on a clear definition of normal operation region, which is difficult to define in reality. In the present work, a novel statistical testing algorithm is integrated with PCA for further improvement of fault detection and identification performance. We use the idea to decompose the scores space and residual space generated by PCA into several subsets so chosen that in each subset the detection problem can be solved with an efficient recursive change detection algorithm based on χ^2-generalized likelihood ratio (GLR) test. *Copyright Ó 2003 IFAC*

Keywords: Principal Component Analysis, Process Monitoring, Sequential Statistical Testing, Contribution Plots

1. INTRODUCTION

Today's chemical processes are becoming heavily instrumented to measure a large number of process variables and data are being recorded more frequently. These process measurements are highly correlated. Identifying and troubleshooting abnormal operating conditions are difficult task with these large amounts of data. The most commonly used technique is principal component analysis (PCA). Process monitoring using PCA is widely based on 'snap shot' Shewhart type control charts, such as T^2- and *SPE*-statistic control charts. The decision depends solely on the current sample though the results of previous sample are available. The implementation of this test is quite simple, but, as one might expect, one pays for this simplicity with rather severe limitations on performance. First, subtle failures are much difficult to detect with this simple scheme. Second, it is difficult to get a tradeoff between false alarm and quick fault detection.

Several extended methods have been proposed for fault detection and identification based on PCA algorithm. Ku et al. (1995) proposed dynamic PCA for process monitoring. Bakshi (1998) combined PCA and wavelet analysis (multiscale PCA) for fault detection. Kano et al. (2001) proposed moving PCA for process monitoring. Kano et al. (2002) described process monitoring based on dissimilarity of process data.

Several problems are not yet solved in these algorithms. First, a clear definition of normal operating condition is needed. This is not the case in reality. In general, there are large gray areas where incipient or small faults occur, while normal process can go to this region by chance. Second, multivariate CUSUM charts, which can detect small changes, are only available for scores. The monitoring of residuals is also very important. Third, though algorithms, which can detect small changes that affect the correlation structure such as Kano's dissimilarity based process monitoring scheme, have been proposed, such schemes cannot detect the variables that are responsible for the fault when the fault occurs.

In this paper, we describe an approach, which integrate PCA with efficient statistical testing algorithm which can solve the problems mentioned above. The outline of the paper is as follows. First, we give a brief introduction of PCA for process monitoring and fault detection. Then, we review

* Corresponding author. Tel: 1-765-494-0734.
E-mail: venkat@ecn.purdue.edu

several statistical testing algorithms. The integration of PCA with an efficient statistical testing algorithm is then presented. Case studies to demonstrate the proposed approach are provided. The paper is concluded with summary.

2. PCA FOR PROCESS MONIOTRING AND FAULT DETECTION

PCA technique is used to develop a model describing the expected variation under normal operating conditions (NOC). An appropriate reference *m*-dimensional data set *X* with *n* samples and *m* variables is chosen which defines the NOC for a particular process. After the data has been properly scaled, PCA can be used to divide the measurement space into two subspaces, one principal component subspace and one residual subspace as,

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

where *X* is the normal operating condition data and *E* represents residual error matrix.

The variance of each principal component is determined by the eigenvalues associated with the principal component. T^2 statistic is a Shewhart type chart defined based on principal component subspace as, $T_i^2 = t_i \Lambda^{-1} t_i^T = x_i P \Lambda^{-1} P^T x_i^T$. The matrix Λ is a diagonal matrix containing the eigenvalues associated with the *A* principal components retained in the model. Statistical confidence limits for T^2 can be calculated by means of the F-distribution as follows.

$$T_{A,n,a}^2 = \frac{A(n-1)}{n-A} F_a\,(A, n-A)$$

SPE-statistic is another Shewhart type chart defined on the residual subspace. A general assumption is that the variance is same in all directions, so *SPE*-statistic is defined as,

$$SPE_i = e_i I e_i^T = x_i (I - P_k P_k^T) x_i^T$$

The statistical confidence limits of *SPE*-statistic can be calculated from its approximate distribution. We can also approximate it as,

$$SPE_a = \left(\sum_{i=A+1}^{p} l_i \right) c_a^{\,2} (p - A)$$

When a vector of new data is available, T^2-and *SPE*-statistic can be calculated based on the model generated and are compared with the corresponding confidence limits. If either of the confidence limits is violated, a fault situation is detected.

Contribution plots (Nomikos, 1996) are a PCA approach to fault identification that takes into account the special correlation, thereby improving the univariate statistical techniques. PCA separates the observation space into two subspaces – the reduced space defined by the principal components of the model and the residual subspace. If T^2-statistic or *SPE*-statistic is out of limit, the contribution plots can be used to indicate the variables which are responsible the deviation.

3. STATISTICAL TESTING STRATEGIES

We assume that the measurement of the process follows an independent Gaussian multivariate distribution. For the measurement sequence, $\{X_i\}$, a vector of parameters θ, which is typically the process mean, describes the stochastic behavior of the process. Under the desirable conditions, this vector belongs to the set Θ_0. A control procedure is applied to this process for fault detection and monitoring. If a control procedure triggers a signal under desirable conditions, it is classified as false alarm. At some point in time, the parameters abruptly change to some value that belongs to a rejectable set, Θ_1. The control scheme is then supposed to detect this change as soon as possible.

The criteria of performance of a control scheme are usually related to the behavior of some characteristics of its distribution, most typically the average run length (ARL), which is the average number of observations required for the algorithm to signal that θ has changed. Ideally, the ARL should be large when the process is in control and small when the process is out of control.

An important tool used for fault detection is based on the logarithm of likelihood ratio. It is defined as $s(x) = \ln \frac{p_{q_1}(x)}{p_{q_0}(x)}$. Basseville and Nikiforov (1993) provided a good discussion on log-likelihood ratio strategy for fault detection. Under some general conditions, log-likelihood ratio schemes possess optimality properties in the sense that they provide the best sensitivity for a given rate of false alarms.

3.1 Page's CUSUM algorithm

To improve the sensitivity of the Shewhart charts, Page (1954) modified Wald's theory of sequential hypothesis testing to develop the CUSUM charts that have certain optimality. In this algorithm, post change parameter θ_1 is assumed known and the unknown change point is estimated by maximum likelihood in CUSUM scheme. The CUSUM criterion can be expressed recursively as

$$t = \inf\{n \geq 1 : g_n \geq h\}$$

$$g_n = (g_{n-1} + \log \frac{p_{q_1}(X_n)}{p_{q_0}(X_n)})^+, \; g_0 = 0$$

where a^+ =*a.I* ($a^3 0$), $p_q(.)$ is the distribution density function depending on parameter θ.

Moustakides (1986) has shown that Page's CUSUM scheme is optimal in the minmax sense: let *h* be so chosen that $E_0(N) = \gamma$ and let F_γ be the class of all monitoring schemes subject to the constraint $E_0(N) \geq \gamma$, where $E_0(N)$ is the expected ARL when the process is in control. Then the above CUSUM minimize the worst-case expected delay over all rules that belong to F_γ.

3.2 GLR algorithm

The parameter θ$_1$ after change is generally unknown. An obvious way to modify the CUSUM rule for the case with unknown post change parameter θ is to estimate it by maximum likelihood, leading to the Generalized Likelihood Ratio (GLR) rule.

$$N_G = \inf\{n : \max_{1\le k\le n} \sup_{q\in\Theta} \sum_{i=k}^{n} \log \frac{p_q(X_i)}{p_{q_0}(X_i)} \ge h\}$$

Siegmund and Venkatraman (1995) give asymptotic approximations to the ARL of the GLR under θ$_0$ and under θ≠θ$_0$, which shows that the GLR rule is asymptotically optimal in the minmax sense. For normal distribution with mean θ and variance 1, they have shown that $\log E_0(N) \sim h$ as $E_0(N) \to \infty$ for the GLR rule. This formula provides an estimation of *h* given $E_0(N)$.

Unlike CUSUM rule, the GLR rule doesn't have convenient recursive forms and the memory requirements and number of computations at time *n* grow to infinity with *n*.

3.3 c^2-GLR algorithm

If we know the post change parameter magnitude but not the direction, we can design an optimal recursive algorithm for the change detection. Here, the process is an independent Gaussian multivariate (*r*>1) sequence and its mean vector θ changes at an unknown time *n*.

$$L(X_t) = \begin{cases} N(q_0, \Sigma), & if \quad t < n \\ N(q_1, \Sigma), & if \quad t > n \end{cases}$$

We know the post change magnitude $(q_1 - q_0)^T \Sigma^{-1} (q_1 - q_0) = b^2$. It has been shown that χ^2-GLR can be calculated in recursive form, which greatly reduce the computational burden. The stopping time of GLR algorithm for this situation can be formulated in recursive form as (Nikiforov 2001),

$$\hat{N} = \inf\{n \ge 1 : S_n \ge h\}$$

$$S_n = -n_n \frac{b^2}{2} + b\left|c_n\right|, \; c_n^2 = V_n^T \Sigma^{-1} V_n$$

$$V_n = \mathbf{1}_{\{\hat{S}_{n-1} > 0\}} V_{n-1} + (X_n - q_0), \; n_n = 1_{\{\hat{S}_{n-1} > 0\}} n_{n-1} + 1$$

In this algorithm, the magnitude after change is assumed known, which is not true in practice. To deal with this problem, the GLR algorithm can be used. However, this algorithm is computationally expensive. Nikiforov (2001) proposed a suboptimal scheme to solve the computational burden problem. The idea is to decompose a given parameter space into several subsets so chosen that in each subset the detection problem can be solved with loss of a small part, ε, of optimality by a recursive change detection algorithm.

3.4 e-optimality algorithm

This algorithm is designed for detection of changes over a domain $\Theta_1 = \{q_1 : b_0^2 \le (q_1 - q_0)^T \Sigma^{-1} (q_1 - q_0) \le b_1^2\}$ using a collection of *L*-parallel recursive tests. Each subset is so chosen that the detection problem can be solved by a recursive χ^2-GLR algorithm.

The ε-optimality algorithm is summarized below.

1) Given the tuning parameters e, b_0, b_1, h, calculate the number of parallel tests *L*, which is the smallest integer $\ge \log \frac{b_1}{b_0} (\log \frac{1+\sqrt{e}}{1-\sqrt{e}})^{-1}$.
2) For *l* = 1,…, *L* compute $a_l = b_0 \frac{(1+\sqrt{e})^l}{(1-\sqrt{e})^{l-1}}$ and initialize the *L* parallel tests.
3) Take the next observations. For *l* = 1,.., *L*, compute $S_n(a_l)$.
4) Check if $\max\{S_n(a_1), \ldots, S_n(a_L)\} \ge h$ then declare alarm. Otherwise, go to step 3.

4. PCA WITH EFFICIENT STATISTICAL TESTING ALGORITHM FOR PROCESS MONIOTIRNG

In all the above algorithms, an inverse of covariance matrix Σ is needed for the fault detection procedure. However, when lots of process variables are measured and they are correlated, Σ can be singular or near singular. In such case, PCA can be used to divide the measurement space into two subspaces—a score subspace and a residual subspace.

Based on the PCA model, *T*^{2}-statistic is designed to detect abnormality in the scores subspace while *SPE*-statistic is for the residual subspace. In the conventional PCA procedure using *T*^{2} and *SPE* for fault detection, the overall type I error is controlled by the level of α. The type II error will be dependent on the post change parameter. Therefore, it is difficult for the procedure to detect small changes whose *T*^{2} and *SPE* statistics is inside the confidence limits. It is also difficult to get a good tradeoff between false alarm and quick detection based on this procedure. It has been shown that ε-optimality GLR algorithm can be used to detect small faults without increasing the false alarm rate. Here we proposed an algorithm to integrate PCA and ε-optimality GLR statistical testing algorithm for fault detection.

First, capability to detect changes of extremely high magnitude can frequently be improved by introducing an additional signal criterion, which calls for a signal at the moment *k* if testing statistic of a single observation x_k exceeds c, which is a predefined value. Here we choose 99.99% confidence limit for *T*^{2} and *SPE*-statistics as the c value for *T*^{2} and *SPE* statistics, respectively. We define the area between 68% and 99.99% confidence of *T*^{2}- and *SPE*-statistic as gray area in the scores and residuals subspace, respectively. Several parallel recursive tests based on ε-optimality algorithm can be designed for the gray area. The following is a summary of the proposed algorithm.

Offline stage

1) Collect normal operating condition (NOC) data X and build PCA model based on NOC data $X = \sum_{i=1}^{A} t_i p_i + E$, where A is the number of principal components used in the model.
2) Based on the PCA model, calculate the 68% and 99.99% confidence limits for T^2-statistic as T_{68} and $T_{99.99}$, and for SPE-statistic as SPE_{68} and $SPE_{99.99}$.
3) Given ε, calculate the number of parallel test for scores as $L_T = ceil(\log\sqrt{\frac{T_{99.99}}{T_{68}}}(\log\frac{1+\sqrt{e}}{1-\sqrt{e}})^{-1})$ and the number of parallel test for residuals as $L_{SPE} = ceil(\log\sqrt{\frac{SPE_{99.99}}{SPE_{68}}}(\log\frac{1+\sqrt{e}}{1-\sqrt{e}})^{-1})$, where *ceil(x)* rounds the elements of X to the nearest integers towards infinity.
4) Calculate the L optimal subdivisions for the test of scores and residuals. For l = 1,…, L, compute optimal subdivisions for T^2-statistic as $T_l = \sqrt{T_{68}}\frac{(1+\sqrt{e})^l}{(1-\sqrt{e})^{l-1}}$ and for SPE-statistic as $SPE_l = \sqrt{SPE_{68}}\frac{(1+\sqrt{e})^l}{(1-\sqrt{e})^{l-1}}$.
5) Given $E_0(N)$, which is the expected ARL when the process is in control, calculate the threshold for the parallel tests. For parallel tests of scores, $h_T = A\{\log(E_0(N))\}$. For parallel test of residuals, $h_{SPE} = \{\log(E_0(N))\}(\sum_{j=A+1}^{m} l_j)$, where l_j is the jth eigenvalue of covariance or correlation matrix of X.

Online stage

1) When new measurements x_i are available, calculate scores t_i and residuals e_i as $t_i = Px_i$, $e_i = x_i - t_i P$.
2) Calculate the T^2 and SPE-statistic for the new data based on scores and residuals as $T_i^2 = t_i^T \Lambda^{-1} t_i$ and $SPE_i = e_i^T e_i$. If $T_i^2 > T_{99.99}$ and/or $SPE_i > SPE_{99.99}$, an alarm is triggered.
3) Otherwise, calculate the testing statistic for each parallel test for scores and residuals.

   For each l = 1,…, L, compute $S_i(T_l)$ as

   $i := i+1,\ i_i = \mathbf{1}_{\{S_{i-1}(T_l)>0\}} i_{i-1} + 1$

   $V_{T^2,i} = 1_{\{S_{i-1}(T_l)>0\}} V_{T^2,i-1} + t_i, \quad c_i^2 = V_{T^2,i}^T \Lambda^{-1} V_{T^2,i}$

   $S_i(T_l) = -i_i \frac{T_l^2}{2} + T_l |c_i|$

   For each l = 1,…, L, compute $S_i(SPE_l)$ similarly as,

   $i := i+1,\ i_i = \mathbf{1}_{\{S_{i-1}(SPE_l)>0\}} i_{i-1} + 1$

   $V_{SPE,i} = 1_{\{S_{i-1}(SPE_l)>0\}} V_{SPE,i-1} + e_i, \quad c_i^2 = V_{SPE,i}^T V_{SPE,i}$

   $S_i(SPE_l) = -i_i \frac{SPE_l^2}{2} + SPE_l |c_i|$

   If $\max\{S_i(T_1),\ldots,S_i(T_L)\} \ge h_T$ and/or $\max\{S_i(SPE_1),\ldots,S_i(SPE_L)\} \ge h_{SPE}$, then an alarm is triggered.

If an alarm is triggered, variable contribution can be used to determine the process variable(s) that are responsible for the alarm. PCA divides the variable space into the score subspace and the residual subspace. Therefore, the variable contribution to T^2 should just use the information in the subspace captured by PCs. According to our knowledge, all of the definition of variable contribution to T^2 uses the information in the whole variable space. Here we provide a new definition of variable contribution to T^2 which using only the information in the subspace spanned by PCs. Given that $t = xP$, $\hat{x} = tP^T$ where $\hat{x}$ is the prediction based on PCA model,

$$T^2 = t^T \Lambda^{-1} t = \hat{x}^T P \Lambda^{-1} P^T \hat{x} = \left\|\Lambda^{-1/2} P^T \hat{x}\right\|^2 = \left\|\sum_{k=1}^{m} \Lambda^{-1/2} P_k \hat{x}_k\right\|^2,$$

so we can define the variable contribution to T^2 as

$$T_k^2 = \left\|\Lambda^{-1/2} P_k \hat{x}_k\right\| = \sum_{i=1}^{a} p_{k,i}^2 \hat{x}_k / l_i \cdot$$

If the alarm is triggered by T^2-statistic out of 99.99% confidence limits, the new definition of variable contribution to T^2 can be used to determine the variables that are most affected by the fault. If the alarm is triggered by one of the parallel tests in scores space. The following variable contribution definition to cumulative scores can be used.

$$V_{T^2,k}^2 = \sum_{j=1}^{a} p_{k,j}^2 (V_{T^2} P)_k^2 / l_i$$

If the alarm is triggered by SPE-statistic out of 99.99% confidence limits, variable contribution to SPE statistic can be used for fault identification. If the alarm is triggered by one of the parallel tests for residuals, we can define variable contributions based on the cumulative residuals as follows and use them for fault identification.

$$V_{SPE,k}^2 = (V_{SPE})_k^2$$

If $\max\{S_i(T_1),\ldots,S_i(T_L)\} \ge h_T$, and $V_{T^2,k}^2$ is large compare to others, then the kth variable is heavily affected by the fault. Similarly, if $\max\{S_i(SPE_1),\ldots,S_i(SPE_L)\} \ge h_{SPE}$ and $V_{SPE,k}^2$ is large than the others, then the kth variable is heavily affected by the fault.

When the proposed scheme detects a fault, it also provides a rough estimation of the fault magnitude based on the information which test is above the confidence limit.

Note that the proposed scheme is different from L-parallel tests of T^2- and SPE-statistic. In this scheme the multivariate nature of the process is considered during the design of the algorithm.

5. CASE STUDIES

5.1 AR process

In this section, we will demonstrate the use of the proposed algorithm for process monitoring of a simple multivariate process. The simple process is used to obtain statistically meaningful results. The data for this example are generated from a model suggested by Ku et al. (1995).

$$x(k)=\begin{bmatrix}0.118 & -0.191\\ 0.847 & 0.264\end{bmatrix}x(k-1)+\begin{bmatrix}1 & 2\\ 3 & -4\end{bmatrix}u(k-1),$$

$$y(k)=x(k)+v(k)$$

where u is the correlated input:

$$u(k)=\begin{bmatrix}0.811 & -0.226\\ 0.477 & 0.415\end{bmatrix}u(k-1)+$$

$$\begin{bmatrix}0.193 & 0.689\\ -0.320 & -0.749\end{bmatrix}w(k-1)$$

The input w is a random noise with zero mean and variance 1. The output y is equal to x plus the random noise v(k), with zero mean and variance 0.1. Both input u and output y are measured but v and w are not. Normal operating condition data consists of 200 measurements. 2 principal components are used to build the monitoring model. For the proposed algorithm, ε = 0.05, $E_0(N)$ = 10,000. 4 parallel tests are used for scores space and residuals space, respectively.

Case 1: This case is to monitor the normal process. 1000 normal operating condition data are simulated and used for monitoring based on the conventional PCA model and the proposed algorithm. T^2- and *SPE*-statistic for the conventional PCA model is shown in Figure 1. Though the process is normal, 36 samples are above the warning limit (95%) of T^2-statistic and 4 are above action limit (99%). For the SPE-statistic, 43 samples are above warning limit and 6 are above action limit. The proposed algorithm is used for the normal data. The results are shown in Figure 2. No alarms are generated for those samples.

Case 2: This case is to simulate the mean of w_1 shift from 0.0 to 0.5 introduced at sample 100. T^2- and *SPE*-statistic for conventional PCA model are shown in Figure 3. The conventional PCA cannot detect the fault effectively. The results of the 4 parallel tests for the scores and residuals subspace are shown in Figure 4. The fault is detected at sample 135 by tests in scores subspace and at sample 132 by tests in residuals subspace. This scheme can also provide an estimation of the magnitude of the fault.

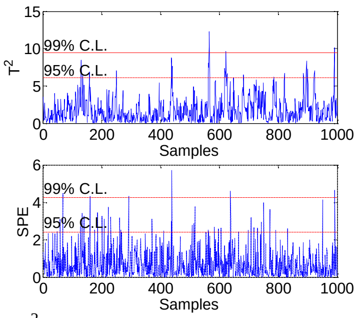

Figure 1. T^2 and SPE-statistic for conventional PCA.

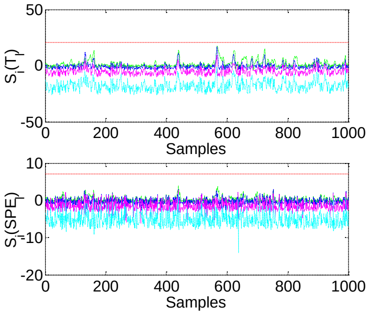

Figure 2. 4 parallel tests for scores and residuals subspace.

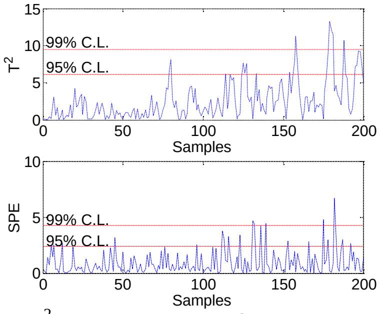

Figure 3. T^2 and SPE-statistic for conventional PCA.

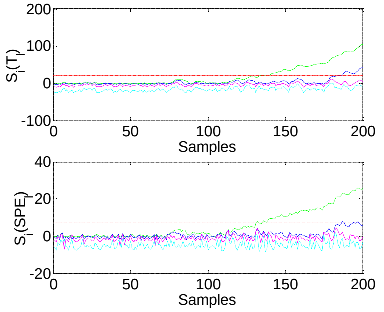

Figure 4. 4 parallel tests for scores and residuals subspace.

5.2 Tennessee Eastman Process

The Tennessee Eastman challenge problem is a simulation of a real chemical plant provided by the Eastman Company (Downs and Vogel, 1993). The process has five major units: the reactor, the product condenser, a vapor-liquid separator, a recycle compressor and a product stripper. The control system used for dynamic simulations is the decentralized PID control system designed by McAvoy and Ye (1994). A total of 16 variables, selected by Chen and McAvoy (1998) for monitoring purposes, are used for monitoring in this study. PCA model is built based on 48 hours of steady state simulation data. The sampling interval of the process variable is 3 min. 11 principal components are used to build the model. For the proposed algorithm, ε = 0.05, $E_0(N)$ = 10,000. 2 and 3 parallel tests are used for scores subspace and residuals subspace, respectively.

Case 1: This is the 3rd process disturbance designed in the original paper. It is to simulate a step change in the D feed temperature. The total simulation time is 48 hours and the disturbance is introduced into the system after 36 hours of steady state simulation.

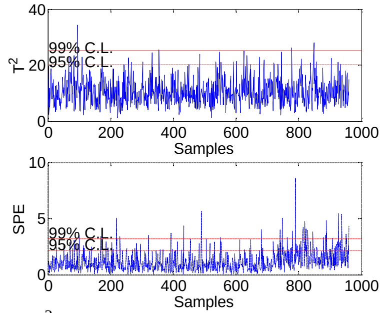

Figure 5. T^2 and SPE-statistic for conventional PCA.

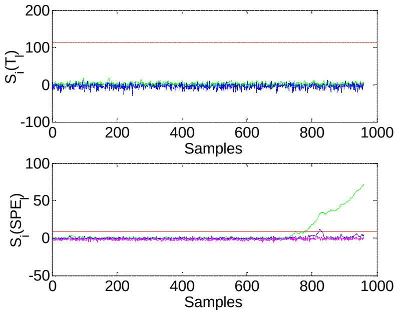

Figure 6. 2 parallel tests for scores subspace and 3 for residuals subspace.

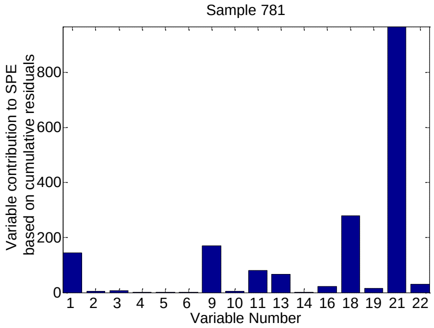

Figure 7. Variable contributions to the cumulative residuals.

T^2 and *SPE*-statistic for conventional PCA are shown in Figure 7. For the 36 hours of steady state simulation, 20 samples are above the warning limit (95%) of T^2-statistic and 2 are above action limit (99%). For the *SPE*-statistic, 54 samples are above warning limit and 10 are above action limit. Conventional PCA cannot detect the fault effectively. Results using the proposed algorithm are shown in Figure 8. There is no false alarm during the 36 hours of steady state simulation. The fault is identified at sample 781 by parallel tests of residuals subspace. Variable contribution to *SPE* based on cumulative residuals is shown in Figure 9. Based on this plots, we can find that variables 21 (Reactor cooling water outlet temperature), 18 (Stripper temperature) and 9 (Reactor temperature) contribute most to the out of control situation.

6. SUMMARY

An approach to integrate PCA with efficient statistical testing algorithm for process monitoring and fault detection has been presented. The fault detection decision depends not only on the current sample but the results of previous sample. A clear definition of normal operating condition is not needed. PCA can separate the observation space into a score subspace and a residual subspace. The two subspaces are divided into several subsets so chosen that in each subset the detection problem can be solved with an efficient recursive change detection algorithm based on χ^2-GLR test. Simulations show that the proposed algorithm can effectively suppress the false alarm and detect small changes in the process.

REFERENCE

Bakshi, B.R. (1998). Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE J.*, **44** 1596-1610

Basseville, M. and I. Nikiforov (1993). *Detection of Abrupt Changes*, Prentice-Hall, Englewood Cliffs, NJ

Chen, G. and T.J. McAvoy (1998). Predictive on-line monitoring of continuous processes. *J. of Process Control*, **8**, 409-420.

Downs, J.J. and E.F., Vogel (1993). A plant-wide Industrial Process Control Problem. *Comput. Chem. Engng*, **17**, 245-255

Kano, M., S. Hasebe, I. Hashimoto, and H. Ohno (2002). Statistical Process Monitoring Based on Dissimilarity of Process Data. *AIChE J.*, Vol.48, pp.1231-1240.

Kano. M., S. Hasebe, I. Hashimoto and H. Ohno (2001). A new multivariate statistical process monitoring method using principal component analysis. *Comput. Chem. Engng.*, **25** 1103-1113

Ku, W., R.H. Storer and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.*, **30**, 179-196

McAvoy, T.J. and N. Ye (1994). Base Control for the Tennessee Eastman Problem. *Comput. Chem. Engng*, **18** 383-413

Moustakides, G.V. (1986). Optimal stopping for detecting change in distribution. *Ann. Statist.*, **14**, 1379-1387

Nikiforov, I. (2001). A simple change detection scheme. *Signal Processing*, **81** 149-172

Nomikos, P. (1996). Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis. *ISA Trans.*, Vol.35, pp.259-266.

Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, **41**,100-114

Siegmund, D. and E.S. Venkatraman (1995). Using the generalized likelihood ratio statistics for sequential detection of a change-point. *Ann. Statist.*, **23** 255-271

COMPUTATION OF THE PERFORMANCE OF SHEWHART CONTROL CHARTS

Pieter Mulder, Julian Morris and Elaine B. Martin

Centre for Process Analytics and Control Technology,
School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK

Abstract: The performance of a control chart in statistical process control is often quantified in terms of the Average Run Length (ARL). The ARL enables a comparison to be undertaken between various monitoring strategies. These are often determined through Monte Carlo simulation studies. Monte Carlo simulations are time consuming and if too few runs are performed then the results will be inaccurate. An alternative approach is proposed based on analytical computation. The analytical results are compared with those of the Monte Carlo simulations for three case studies. *Copyright © 2003 IFAC*

Keywords: Statistical Process Control, Control charts, Serial correlation, Monte Carlo simulations

1. INTRODUCTION

In Statistical Process Control (SPC), a variety of control charts have been applied including Shewhart, CUSUM and EWMA (e.g. Montgomery, 1991; Wetherill and Brown, 1991). Each method has associated advantages and disadvantages that have been reported in the literature. Control chart performance is traditionally quantified in terms of the Average Run Length (ARL). Run length is defined as the number of observations from the start of the control chart to the first out-of-control signal.

Except for simple cases (e.g. Brook and Evans, 1972; Schmid, 1995), Monte Carlo simulations have been used to determine the ARL. This involves the realisation of a vector containing a random signal and then applying the control scheme and measuring the run length (time to the first alarm). This is repeated many times, each time a different random vector is generated, and finally the ARL is computed. The main issue with this method is the trade-off between computer time and accuracy of the results. A large number of realisations are necessary if the results are to be precise (Lowry *et al*, 1992; Wardell *et al*, 1994). Therefore it is believed that an analytical method for the computation of the ARL would be desirable. The density function of the run length of a control chart is first constructed based on the in-control probability of an observation. This approach is similar to that of Wetherill and Brown (1991). They assumed that the in-control probability was constant for every observation. In contrast, in this work, this constraint is relaxed. This allows the computation of the ARL for more complicated SPC monitoring strategies, and ultimately for correlated data.

The ARL is investigated in more detail for three case studies. The first case study looks at the ARL of a Shewhart control chart based on independent data and is derived for both the in-control and out-of-control situation. This example demonstrates the validity of the approach. The impact of serial correlation on the performance of control charts is well known (Alwan and Roberts, 1988; Montgomery and Mastralango, 1991). One solution is to estimate an ARMA model (Harris and Ross, 1991) for univariate systems, or a VAR model (Mulder *et al*, 2001) for multivariate systems, and to monitor the residuals, which are free of serial correlation. In the second case study, the focus is on the residuals of a first order AutoRegressive, AR(1), time series model as defined by Box *et al,* (1994).

For the third case study, the ARL of a correlated time series generated by an AR(1) model is computed. Schmid (1995) claimed that an explicit solution does not exist for the ARL of correlated data, and that only general statements about the ARL are possible. In this study it will be shown that although there is not an explicit solution for the ARL, there is a numerical approximation. The analytical results are compared with the results of Monte Carlo simulations for each of the three case studies.

2. CONTROL CHARTS

The run length of a control chart is defined as the number of observations until the first observation moves outside of the control limits. After this observation, the control chart is stopped and calculation of the run length is recommenced from the next in-control observation. In this section, the density function of the run length is constructed.

The probability that an observation, X_k, is in control at time point, *k*, is given by:

$$P(UCL > X_k > LCL) \qquad (1)$$

and the probability that at point, *k*, observation, X_k, is out-of-control is defined as:

$$\begin{aligned} &P(X_k > UCL \; or \; X_k < LCL) \\ &=1\text{-} P(UCL > X_k > LCL) \end{aligned} \qquad (2)$$

Also it is assumed that an observation is either in-control or out-of-control. In a control chart, an observation is only recorded if the previous point was in-control. That is an observation can only be deemed to be in-control at time point *k* if the observations at $1,2,\ldots,k-1$ were in control:

$$\begin{aligned} P_{IC,k} &= P(UCL > X_1 > LCL)\cdot P(UCL > X_2 > LCL)\cdot\ldots \\ &\cdot P(UCL > X_{k-1} > LCL)\cdot P(UCL > X_k > LCL) \end{aligned} \qquad (3)$$

Also an observation at time point *k* in a control chart is out-of-control if:

$$\begin{aligned} P_{OC,k} &= P(UCL > X_1 > LCL)\cdot P(UCL > X_2 > LCL)\cdot\ldots \\ &\cdot P(UCL > X_{k-1} > LCL)\cdot P(X_k > UCL \; or \; X_k < LCL) \end{aligned} \qquad (4)$$

Based on equations 3 and 4, the Average Run Length is the expectation of the out-of-control run length and is given by:

$$E(kP_{OC,k}) = \sum_{k=1}^{\infty} kP_{IC,k-1}P(X_k > UCL \; or \; X_k < LCL) \qquad (5)$$

Based on the following definition

$$P(UCL > X_k > LCL) = \beta_k \qquad (6)$$

Equation 2 is given as:

$$P(X_k > UCL \; or \; X_k < LCL) = 1-\beta_k \qquad (7)$$

and equation 5 is redefined as:

$$ARL = E(kP_{OC,k}) = \sum_{k=1}^{\infty} k(1-\beta_k)\prod_{j=1}^{k-1}\beta_j \qquad (8)$$

This is as described by Wetherill and Brown (1991), except that β in equation 8 can differ for each time point, *k*.

2.1 Case 1 - Independent Data

For independent data, the value of an observation is independent of its previous value, thus $\beta_k = \beta$ for $1,2,\ldots,k-1$, and equation 8 becomes:

$$ARL = \sum_{k=1}^{\infty} k(1-\beta)\beta^{k-1} = \frac{1}{1-\beta} \qquad (9)$$

This result agrees with that of Wetherill and Brown (1991).

2.2 Case 2 - Residuals from an AR(1) Model

An AR(1) time series model is given by:

$$\begin{aligned} y_t &= \xi_t + \eta_t \\ \xi_t &= \alpha\xi_{t-1} + e_t \end{aligned} \qquad (10)$$

where y_t is the observed data, ξ_t is the underlying correlated time series, with α as its autoregressive parameter, and e_t is a white noise vector with variance σ_e^2, which is assumed to have a Normal distribution, and η_t is the mean shift applied to the data vector y_t (Kaskavelis, 2000). The one-step ahead prediction errors of an AR(1) model are:

$$\begin{aligned} \hat{e}_t &= y_t - y_{t|t-1} \\ &= \alpha\xi_{t-1} + e_t + \eta_t - \alpha\xi_{t-1} - \alpha\eta_t \\ &= e_t + \eta_t - \alpha\eta_{t-1} \end{aligned} \qquad (11)$$

When a process is in-control, $\eta_t = 0$, for all *t*, then the probability that at time point *k*, the process is in-control is constant, $\beta_k = \beta$, $k \geq 1$. Therefore the in-control ARL is given by:

$$ARL = \sum_{k=1}^{\infty} k(1-\beta)\beta^{k-1} = \frac{1}{1-\beta} \tag{12}$$

From equation 12 and for the desired in-control ARL, the control limits for the Shewhart control charts can be derived. For the out-of-control case, it is assumed that a constant mean shift is applied to y_t, $\eta_t = \eta$ for all $k \geq 1$. Thus according to equation 11, at $k = 1$, the probability that, when a mean shift is applied to y_t, the process is in-control is β_1 and for $k \geq 2$, $\beta_k = \beta$. The mean shift in the residuals for $k = 1$ is different to that for $k \geq 2$. Apley and Shi (1999) termed this the fault signature of a step change in the residuals for univariate systems. This issue was not considered by Harris and Ross (1991) or Kaskavelis (2000). They assumed that the mean shift in the residuals is identical for all *k*. The probability that the control chart will give an out-of-control signal at some point in the future is:

$$\begin{aligned} P &= (1-\beta_1) + \sum_{k=2}^{\infty} \beta_1 (1-\beta)\beta^{k-2} \\ &= (1-\beta_1) + \beta_1(1-\beta)\sum_{k=0}^{\infty} \beta^k = 1 \end{aligned} \tag{13}$$

since $\sum_k^{\infty} \beta^k = (1-\beta)^{-1}$ for $0 \leq \beta < 1$. The out-of control ARL is given by:

$$\begin{aligned} ARL &= (1-\beta_1) + \sum_{k=2}^{\infty} k\beta_1(1-\beta)\beta^{k-2} \\ &= (1-\beta_1) + \frac{\beta_1(1-\beta)}{\beta}\left(\frac{1}{(1-\beta)^2} - 1\right) \end{aligned} \tag{14}$$

When $\beta_1 = \beta$, it can be seen that equation 14 is equivalent to equation 9 and 12.

2.3 *Case 3 - Serially Correlated Data*

In this section it will be shown how β_k can be computed for AR(1) processes via the probability distribution function. It is assumed that observations X_k are monitored using a Shewhart control chart with an Upper Control Limit (UCL) and a Lower Control Limit (LCL). The cumulative distribution function of observation X_k at time point *k* is defined as (Papoulis, 1991):

$$F_k(x) = \int_{-\infty}^{x} f_k(z)dz \tag{15}$$

where f_k is the probability density function of observation X_k at time point *k*. The time series structure is defined as in equation 10, where α has variance $1-\alpha^2$. A condition of the control chart is that an observation at k is only plotted if the observation at $k-1$ is in-control:

$$UCL > X_{k-1} > LCL \tag{16}$$

otherwise the control chart would have been terminated at $k-1$. The conditional distribution function of X at $k-1$ is given by:

$$f_{k-1,cond}(x) = \frac{f_{k-1}(x)}{F_{k-1,uncond}(UCL) - F_{k-1,uncond}(LCL)} \tag{17}$$

The probability distribution function of the term αX_{k-1} is defined as (Papoulis, 1991):

$$g_k(x) = \frac{1}{|\alpha|} \cdot f_{k-1,cond}(x) \tag{18}$$

The white noise term e_t is normally distributed with variance $1-\alpha^2$ so that the unconditional *X* has unit variance. The probability distribution function of e_t is denoted as $N_k(x)$. Since αX_{k-1} and e_t are independent, the probability distribution function of their sum, the probability distribution function of X_k, is given by the convolution product (Papoulis, 1991):

$$f_k(x) = \int_{-\infty}^{x} g_k(z) n_k(z-x) dz \tag{19}$$

The in-control probability is thus the probability that observation X_k lies between the control limits:

$$\beta_k = F_k(UCL) - F_k(LCL) = \int_{LCL}^{UCL} f_k(z)dz \tag{20}$$

At the start of a control chart, no other observations are known. Therefore X_1 can be regarded as the unconditional observation of *X*, and subsequently β_1 is computed from equation 15. For $k > 1$, β_k can be computed recursively by the procedure described above, equations 17 to 19.

3. RESULTS

In this section, the theoretical relationships derived in the previous section are compared with the results from Monte Carlo simulations. For all Monte Carlo simulations 10,000 realisations of the control charts were computed. Each realisation comprised 10,000 observations and the first observation outside the control limits was taken to define the run length for that realisation. Since the ARL is the mean value of the run lengths of the realised control charts, the standard error of the ARL is:

$$\sigma_{ARL} = \frac{ARL}{\sqrt{N}} \tag{21}$$

where N is the number of realisations. Error bars will thus indicate the standard error of the Monte Carlo simulations. For the three cases it is assumed that the metric, the data in case 1 and 3, and the residuals in case 2, are monitored using a Shewhart control chart. For each case it is assumed that the metric used in the control chart is normally distributed and that the desired in-control ARL is equal to 370. This corresponds to a Shewhart control chart with control limits at -3σ and $+3\sigma$.

3.1 Case 1 - Independent Data

The observations, X, are drawn from a population with a normal distribution that are offset by a mean shift, η. The mean of the distribution is equal to η and its variance is σ_X^2:

$$X \sim N\left(\eta, \sigma_X^2\right) \tag{22}$$

The probability that X lies between the control limits is given by:

$$\beta = F(UCL) - F(LCL) = \int_{LCL}^{UCL} f(\xi) d\xi \tag{23}$$

where f is the probability distribution function of X, the normal distribution. Subsequently the ARL can be computed from equation 15.

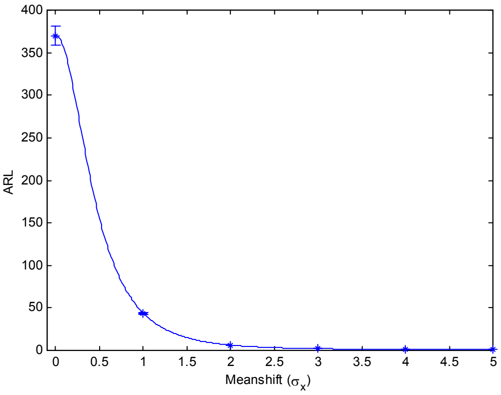

Fig. 1. Performance of Shewhart control chart for independent data.

The ARL as a function of the mean shift is shown in Fig. 1. The dots represent the results of the Monte Carlo simulation with error bars that indicate –3/+3 standard error of the mean and the solid line is the theoretical computation. It can be seen that the theoretical results correspond with those from the Monte Carlo simulations.

3.2 Case 2 - Residuals from an AR(1) Model.

The desired in-control ARL is 370. It is assumed that a step function of size η is superimposed on the time series y_t:

$$\begin{aligned} \eta_t &= 0 \quad t \le 0 \\ \eta_t &= \eta \quad t > 0 \end{aligned} \tag{24}$$

From equation 24, for $k = 1$, $\hat{e}_k \sim N\left(\eta, \sigma_e^2\right)$ and for $k \ge 1$, $\hat{e}_k \sim N\left(\eta(1-\alpha), \sigma_e^2\right)$. Together with Equation 15, this allows the computation of β for all k. The ARL is computed from equation 14.

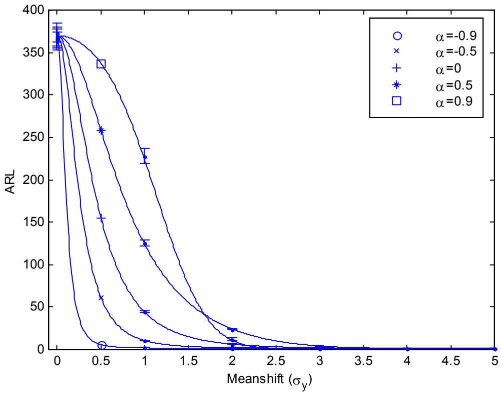

Fig. 2. Performance of Shewhart control chart for the residuals of an AR(1) model.

The ARL as a function of the mean shift for several values of alpha is shown in Fig 2, together with the results from the Monte Carlo simulations. The error bars indicate the results of the Monte Carlo simulations with –3/+3 standard error of the mean. The solid line indicates the theoretical calculation. Again it can be seen that the theoretical results correspond to those from the Monte Carlo simulations.

3.3 Case 3 - Serially Correlated Data.

The ARL as a function of the mean shift was determined in the previous two cases. In this case study, the influence of serial correlation on the in-control ARL was investigated. In contrast to cases 1 and 2, there is no direct analytical relationship for the ARL. Therefore a numerical approach was used. It is assumed that the probability distribution function of X for the first observation is that of the normal distribution with mean zero and unit variance. Based on equations 17 to 20, the probability distribution function of the second observation is computed. Then β_2 is computed from equation 15. These steps are repeated until β_k converges. The values of β_k

are then used in equation 8 to compute the ARL.

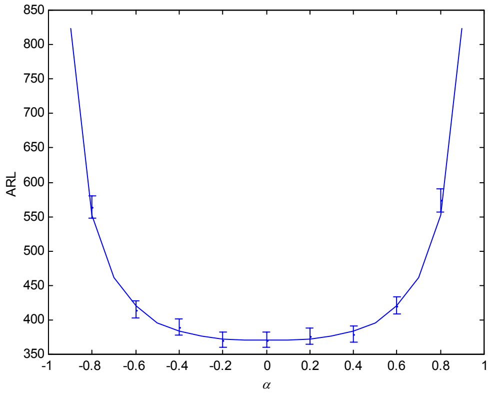

Fig. 3. The in-control ARL of Shewhart control charts for serial correlated data.

The ARL as a function of α is given in Fig. 3, together with the results from the Monte Carlo simulations. The dots represent the Monte Carlo simulations and the error bars indicate –3/+3 standard error of the mean and the solid line indicates the theoretical calculation. Again it can be seen that the theoretical results correspond to those from the Monte Carlo simulations.

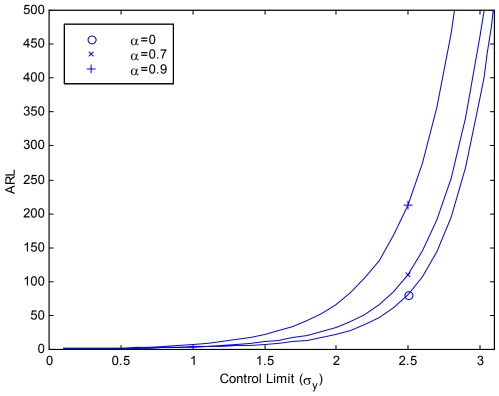

Fig. 4. The in-control ARL as function of the control limit for a selection of values for α.

As can be observed in Fig. 3, the in-control ARL only depends on the absolute value of α. It also can be seen that for increasing α, the in-control ARL increases. This means that on average it takes longer to detect a false alarm. Although this might appear advantageous, in practice the in-control ARL can be considered as a design parameter since it is implied by the choice of significance level δ. For independent and identical data, the in-control ARL is 1/δ, but for serially correlated data this relationship is not valid. Kaskavelis (2000) proposed an alternative philosophy which was to treat the in-control ARL as an explicit design parameter. The above method allows the rapid computation of the ARL for a range of values for the control limits. In Fig. 4, the ARL as a function of the control limit is shown for selected values of α. The control limit is given in terms of the standard deviation of y_t. From this figure the control limits for an AR(1) process can be determined.

Table 1 Control limits to guarantee an in-control ARL of 370 for AR(1) process

α	Control Limit	α	Control Limit
0	3.00	0.5	2.98
0.1	3.00	0.6	2.96
0.2	3.00	0.7	2.93
0.3	3.00	0.8	2.86
0.4	2.99	0.9	2.71

Since the calculations of the ARL closely match the results of the Monte Carlo simulation, Fig. 5, the impact of the mean shift on the ARL is investigated in greater detail without further comparisons being undertaken with simulations. In Fig. 5 the solid line refers to the theoretical calculations and the dots with arrow bars are the Monte Carlo simulations with the associated three standard errors of the mean

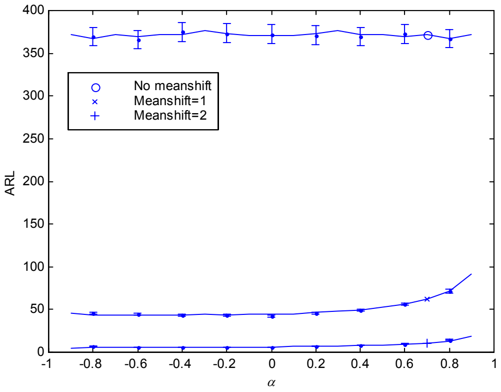

Fig. 5. The ARL as a function of alpha with adjusted control limits, with various mean shifts applied on the time series y_t.

The ARL of a Shewhart control chart with the control limits as given in Table 1 are shown in Fig. 5 where various mean shifts are applied to the time series data, y_t. In this situation, the control limit for negative α is equal to that of its positive counterpart. The calculations are in agreement with the Monte Carlo simulations. When no mean shift is applied, which corresponds to the in-control situation, the calculations do not give exactly 370, because of the rounding of the control limit to two decimals in Table 1. Compared with Fig 3, the in-control ARL does not deviate from 370.

In Fig. 6, the ARL is shown as a function of the mean shift and α. It can be observed that the ARL is only dependent on the mean shift and not on α, except for large positive values of α. Thus a Shewhart control chart with control limits adjusted to ensure the desired in-control ARL will exhibit the same sensitivity for equal sized mean shits regardless of the value of α. In practice the autoregressive parameter, α, is determined by either matching the autocorrelation function (Kaskavelis, 2000) or through estimation of the autoregressive parameter of an AR(1) process from the data.

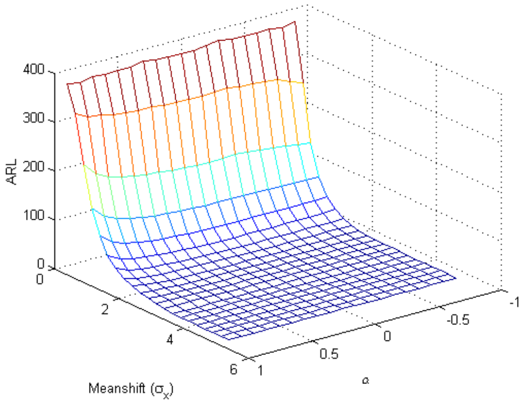

Fig. 6. The theoretically computed ARL as a function of alpha and the mean shift.

4. CONCLUSIONS

Within the paper, it is shown, based on the in-control probability at individual points in a control chart, how the density function of the run length of control charts can be determined. The density function can consequently be used to calculate the Average Run Length (ARL) of a control chart. The ARL is a widely used metric for comparing between monitoring strategies in SPC. The proposed approach is more generic than that described by Wetherill and Brown (1991).

The theoretical ARL for in-control data and out-of control data with step changes in the mean were calculated for three cases, independent data, the residuals of AR(1) models and serially correlated data. The theoretical results corresponded to the ARL obtained through Monte Carlo simulations. It is also shown that, in contrast to the claim of Schmid (1995), the ARL of serial correlated data in-control charts can be computed.

The in-control ARL's were computed as a function of the magnitude of the control limits. The control limits for Shewhart charts that realise an in-control ARL of 370 were determined for various values of autoregressive parameter for an AR(1) process. The impact of mean shifts on the performance of the ARL was subsequently investigated. It was found that the ARL depends only on the mean shift and not on α, except for large positive values of α.

The outcome of this work is that time consuming Monte Carlo simulations can now be replaced by the approach proposed for the assessment of the performance of control charts. This work can also be extended to more complicated SPC monitoring schemes, such as multivariate systems. However for multivariate problems, the problem is compounded by the fact that the parameter space may be large making the problem computationally intensive.

5. ACKNOWLEDMENTS

PM acknowledges the EU project PERFECT (Esprit Project 28870) for supporting his PhD studies.

6. REFERENCES

Alwan, L.C. and H.V. Roberts (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, **6**, pp. 87-95.

Apley, D.W. and Shi, J. (1999). The GLRT for statistical process control of autocorrelated processes. *IIE Transactions*, **31(12)**, pp. 1123-1134.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994). *Time series analysis: Forecasting and control*, Prentice Hall, Englewood Cliffs, NJ, USA.

Brook, D. and D.A. Evans (1972). An approach to the probability distribution of CUSUM run length. *Biometrika*, **59(3)**, pp. 539-549.

Harris, T. J., and Ross, W. H. (1991). Statistical process-control procedures for correlated observations. *Canadian Journal of Chemical Engineering*, **69(1)**, pp. 48-57.

Kaskavelis, E. (2000). Statistical monitoring and prediction of petrochemical processes. *PhD, University of Newcastle, Newcastle upon Tyne, UK.*

Lowry, C. A., W.H. Woodall, C.W. Champ and S.E. Rigdon (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, **34(1)**, pp. 46-53.

Montgomery, D.C. (1991). *Introduction to statistical quality control*, John Wiley & Sons, Singapore.

Montgomery, D.C. and C.M. Mastrangelo (1991). Some statistical process-control methods for autocorrelated data. *Journal of Quality Technology*, **23(3)**, pp. 179-193.

Mulder, P., E.B. Martin and A.J. Morris (2002). The monitoring of dynamic processes using multivariate time series. *Advances in Process Control 6*, York, UK, pp. 194-201.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes*, McGraw-Hill, New York.

Schmid, W. (1995). On the run-length of a shewhart chart for correlated data. *Statistical Papers*, **36(2)**, pp. 111-130.

Wardell, D. G., H. Moskowitz and R.D. Plante (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, **36(1)**, pp. 3-17.

Wetherill, G.B. and D.W. Brown (1991). *Statistical process control: Theory and practice,* Chapman and Hall, London, UK.

COMBINED MULTIVARIATE STATISTICAL PROCESS CONTROL

Manabu Kano [*] **Shouhei Tanaka** [*] **Shinji Hasebe** [*]
Iori Hashimoto [*] **Hiromu Ohno** [**]

[*] *Kyoto University, Kyoto 606-8501, Japan*
[**] *Kobe University, Kobe 657-0013, Japan*

Abstract: Multivariate statistical process control (MSPC) based on principal component analysis (PCA) has been widely used in chemical processes. Recently, the use of independent component analysis (ICA) was proposed to improve monitoring performance. In the present work, a new method, referred to as combined MSPC (CMSPC), is proposed by integrating PCA-based SPC and ICA-based SPC. CMSPC includes both MSPC methods as its special cases and thus provides a unified framework for MSPC. The effectiveness of CMSPC was demonstrated with its applications to a multivariable system and a CSTR process. *Copyright ©2003 IFAC*

Keywords: Statistical Process Control, Process Monitoring, Fault Detection, Independent Component Analysis, Principal Component Analysis

1. INTRODUCTION

The successful process operation often depends on the effectiveness of fault detection. On-line process monitoring plays an important role in detecting process upsets, equipment malfunctions, or other special events as early as possible. In chemical processes, statistical process control (SPC), which is a data-based approach for process monitoring, has been used widely and successfully. Well-known SPC techniques include Shewhart control charts, cumulative sum (CUSUM) control charts, and exponentially weighted moving average (EWMA) control charts. Such SPC charts are well established for monitoring univariate processes, but univariate SPC (USPC) does not function well for multivariable processes. In order to extract useful information from multivariate process data and utilize it for process monitoring, multivariate statistical process control (MSPC) based on principal component analysis (PCA) has been developed (Jackson and Mudholkar, 1979). In the last decade or so, many successful applications have been reported and various extensions of MSPC have been proposed (Kresta et al., 1991; Kano et al., 2002a).

PCA-based SPC (PCA-SPC) and its extensions have been widely accepted in process industries. However, their achievable performance is limited due to the assumption that monitored variables are normally distributed. Recently, to further improve the monitoring performance, a new MSPC method based on independent component analysis (ICA), referred to as ICA-SPC, was proposed by Kano et al. (2002b, 2003). They demonstrated the superiority of ICA-SPC over conventional methods.

ICA-SPC, however, does not always outperform PCA-SPC. ICA-SPC should be selected when process variables do not follow normal distribution. On the other hand, ICA-SPC likely will not improve the performance in comparison with PCA-SPC if process variables are normally distributed. In a practical case, where some variables follow normal distribution and others do not, which monitoring method should be selected? In the present work, to answer this question and propose a new framework for MSPC, combined MSPC (CMSPC) is developed by integrating PCA-SPC and ICA-SPC. The performance of CMSPC is evaluated with its applications to monitoring problems of a linear multivariable system and a CSTR process.

2. PCA-BASED MSPC

PCA, which is a tool for data compression and information extraction, finds linear combinations of variables that describe major trends in a data set. For monitoring a process by using PCA-SPC, control limits are set for two kinds of statistics, T^2 and Q, after a PCA model is developed. T^2 and Q are defined as

$$T^2 = \sum_{r=1}^{R} \frac{t_r^2}{\sigma_{t_r}^2} \tag{1}$$

$$Q = \sum_{p=1}^{P} (x_p - \hat{x}_p)^2 \tag{2}$$

where t_r is the r-th principal component score and $\sigma_{t_r}^2$ is its variance. x_p and $\hat{x}_p$ are a measurement of the p-th variable and its predicted (reconstructed) value, respectively. R and P denote the number of principal components retained in the PCA model and the number of process variables, respectively. The T^2 statistic is a measure of the variation within the PCA model, and the Q statistic is a measure of the amount of variation not captured by the PCA model.

3. ICA-BASED MSPC

ICA (Jutten and Herault, 1991) is a signal processing technique for transforming measured multivariate data into statistically independent components, which are expressed as linear combinations of measured variables. In this section, an ICA algorithm and ICA-SPC are briefly described.

3.1 *Problem Definition*

It is assumed that m measured variables $x_1, x_2, \ldots, x_m$ are given as linear combinations of $n\,(\leq m)$ unknown independent components $s_1, s_2, \ldots, s_n$. The independent components and the measured variables are mean-centered. The relationship between them is given by

$$\boldsymbol{x} = \boldsymbol{s}\boldsymbol{A} \tag{3}$$

$$\boldsymbol{x} = \left[\, x_1 \; x_2 \; \ldots \; x_m \,\right] \tag{4}$$

$$\boldsymbol{s} = \left[\, s_1 \; s_2 \; \ldots \; s_n \,\right] \tag{5}$$

where $\boldsymbol{A}$ is a full-rank matrix, called the mixing matrix. When k samples are available, the above relationship can be rewritten as $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$.

The basic problem of ICA is to estimate the original components $\boldsymbol{S}$ or to estimate the mixing matrix $\boldsymbol{A}$ from the measured data matrix $\boldsymbol{X}$ without any knowledge of $\boldsymbol{S}$ or $\boldsymbol{A}$. Therefore, the practical objective of ICA is to calculate a separating matrix $\boldsymbol{W}$ so that components of the reconstructed data matrix $\boldsymbol{Y}$, given as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}, \tag{6}$$

become as independent of each other as possible. The limitations of ICA are: 1) only non-Gaussian independent components can be estimated (just one of them can be Gaussian), and 2) neither signs, powers, nor orders of independent components can be estimated.

3.2 *Sphering with PCA*

Statistical independence is more restrictive than uncorrelation. Therefore, for performing ICA, measured variables $\{x_i\}$ are first transformed into uncorrelated variables $\{z_j\}$ with unit variance. This pretreatment can be accomplished by PCA and it is called sphering or prewhitening.

By defining the sphering matrix as $\boldsymbol{M}$, the relationship between $\boldsymbol{z}$ and $\boldsymbol{s}$ is given as

$$\boldsymbol{z} = \boldsymbol{x}\boldsymbol{M} = \boldsymbol{s}\boldsymbol{A}\boldsymbol{M} = \boldsymbol{s}\boldsymbol{B}^T \tag{7}$$

where $\boldsymbol{B}^T = \boldsymbol{A}\boldsymbol{M}$. Since s_i are mutually independent and z_j are mutually uncorrelated,

$$E\left[\boldsymbol{z}^T\boldsymbol{z}\right] = \boldsymbol{B}E\left[\boldsymbol{s}^T\boldsymbol{s}\right]\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{I} \tag{8}$$

is satisfied. Here $E[\cdot]$ denotes expectation. It is assumed here that the covariance matrix of s_i, $E\left[\boldsymbol{s}^T\boldsymbol{s}\right]$, is an identity matrix, because signs and powers of s_i remain arbitrary. Equation (8) means that $\boldsymbol{B}$ is an orthogonal matrix. Therefore, the problem of estimating a full-rank matrix $\boldsymbol{A}$ is reduced to the problem of estimating an orthogonal matrix $\boldsymbol{B}$ through the sphering.

3.3 *Fixed-Point Algorithm for ICA*

The fourth-order cumulant of zero-mean random variable y is defined as

$$\kappa_4(y) = E[y^4] - 3E[y^2]^2 \tag{9}$$

By minimizing or maximizing the fourth-order cumulant $\kappa_4(\boldsymbol{z}\boldsymbol{b})$ under the constraint of $\|\boldsymbol{b}\| = 1$, columns of the orthogonal matrix $\boldsymbol{B}$ are obtained as solutions for $\boldsymbol{b}$. Finding the local extrema of the fourth-order cumulant is equivalent to estimating the non-Gaussian independent components (Delfosse and Loubaton, 1995). In the present work, a fixed-point algorithm (Hyvarinen and Oja, 1997) is used to obtain $\boldsymbol{b}$ that minimizes or maximizes the fourth-order cumulant.

For estimating n independent components that are different from each other, the following orthogonal conditions are imposed.

$$\boldsymbol{b}_i^T \boldsymbol{b}_j = 0 \qquad (i \neq j) \tag{10}$$

Thus, the current solution $\boldsymbol{b}_i$ is projected on the space orthogonal to previously calculated $\boldsymbol{b}_j (j = 1, 2, \ldots, i-1)$. By defining

$$\boldsymbol{B} = \left[\, \boldsymbol{b}_1 \ \boldsymbol{b}_2 \ \ldots \ \boldsymbol{b}_n \,\right], \tag{11}$$

independent components $\boldsymbol{Y}$ can be obtained from

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{B} = \boldsymbol{X}\boldsymbol{M}\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}. \tag{12}$$

This means that the separating matrix $\boldsymbol{W}$ can be calculated from $\boldsymbol{W} = \boldsymbol{M}\boldsymbol{B}$.

The sphering matrix $\boldsymbol{M}$ uncorrelates $\boldsymbol{x}$ and scales it so that uncorrelated variables $\boldsymbol{z}$ have unit variances. Uncorrelated variables can be derived by using PCA. Therefore, the sphering matrix $\boldsymbol{M}$ can be decomposed into two parts: an uncorrelating matrix $\boldsymbol{P}$ and a scaling matrix $\boldsymbol{\Lambda}$. The uncorrelating matrix $\boldsymbol{P}$ is the same as the loading matrix of PCA. Therefore, Eq. (12) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}\boldsymbol{B} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{B}. \tag{13}$$

Both $\boldsymbol{P}$ and $\boldsymbol{B}$ are orthogonal matrices.

3.4 *Monitoring of Independent Components*

The procedure of ICA-SPC is the same as USPC. The only difference lies in the variables to be monitored. That is, independent components are monitored in ICA-SPC while correlated measured variables are monitored in USPC.

A separating matrix $\boldsymbol{W}$ in Eq. (12) and control limits must be determined in order to apply ICA-SPC to monitoring problems. For this purpose, the following procedure is adopted.

(1) Acquire time-series data when a process is operated under a normal condition. Normalize each column (variable) of the data matrix, i.e., adjust it to zero mean and unit variance, if necessary.

(2) Apply ICA to the normalized data, determine a separating matrix $\boldsymbol{W}$, and calculate independent components.

(3) Determine control limits of all independent components.

For on-line monitoring, a new sample of monitored variables is scaled with the means and the variances obtained at step 1. Then, it is transformed to independent components through the separating matrix $\boldsymbol{W}$. If one or more of the independent components are outside the corresponding control limits, the process is judged to be out of control.

4. COMBINED MSPC

ICA-SPC does not necessarily outperform PCA-SPC. ICA is based on the assumption that each measured variable is given as a linear combination of non-Gaussian variables that are independent of each other. Independent components, even if they can be calculated, are meaningless and ICA-SPC does not function well when this assumption is incorrect. In the present work, a new advanced MSPC method is proposed for further improving the monitoring performance by combining ICA-SPC and PCA-SPC. The proposed method is referred to as combined MSPC (CMSPC).

4.1 *CMSPC Algorithm*

The basic and important fact of PCA-SPC is that uncorrelated variables, i.e., principal components, are monitored. On the other hand, independent components are monitored in ICA-SPC. Since statistical independence is more restrictive than uncorrelation, ICA-SPC can outperform PCA-SPC. However, if process variables are normally distributed, the monitoring performance would not necessarily be improved by using ICA-SPC. Ideally, ICA-SPC should be used for monitoring non-Gaussian independent variables, and PCA-SPC is used for monitoring uncorrelated Gaussian variables. This conclusion motivates us to integrate PCA-SPC and ICA-SPC into a new MSPC method. For realizing this integration, non-Gaussian variables and Gaussian variables have to be distinguished.

In the present work, the fourth-order cumulant is used to evaluate the non-Gaussianity of components $\{y_l\}$ derived by using ICA. The fourth-order cumulant of any Gaussian random variable is zero. In addition, the absolute value of the fourth-order cumulant increases as the non-Gaussianity increases. Therefore, non-Gaussian independent variables can be selected from $\{y_l\}$ based on their fourth-order cumulants. When the fourth-order cumulant of an independent component is larger than the threshold determined in advance, it is judged to be non-Gaussian and independent.

Consider a data matrix $\boldsymbol{X} \in \Re^{k \times m}$, where k and m are the number of samples and that of variables, respectively. All variables are mean-centered. When r of m components $\{y_1, y_2, ..., y_r\}$ are judged to be non-Gaussian, these r independent components should be monitored independently. In other words, ICA-SPC should be applied to these r independent components. However, since the other $m-r$ components $\{y_{r+1}, y_{r+2}, ..., y_m\}$ are Gaussian, these $m-r$ variables should be monitored by using PCA-SPC. In practice, the ICA algorithm might not converge if more than one component is Gaussian. Therefore, a part of $\boldsymbol{X}$, which is explained by $\{y_{r+1}, y_{r+2}, ..., y_m\}$, needs to be derived without calculating these $m-r$ components.

From Eq. (13), the first r non-Gaussian independent components are given as

$$\boldsymbol{Y}_r = \left[\, \boldsymbol{y}_1 \ \boldsymbol{y}_2 \ \cdots \ \boldsymbol{y}_r \,\right] = \boldsymbol{X}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{B}_r. \qquad (14)$$

A part of $\boldsymbol{X}$, which is explained by the first r non-Gaussian independent components, can be reconstructed from $\boldsymbol{Y}_r$ or $\boldsymbol{X}$.

$$\begin{aligned} \boldsymbol{X}_r &= \boldsymbol{Y}_r \boldsymbol{B}_r^T \boldsymbol{\Lambda}^{-1} \boldsymbol{P}^T \\ &= \boldsymbol{X}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{B}_r\boldsymbol{B}_r^T \boldsymbol{\Lambda}^{-1} \boldsymbol{P}^T \end{aligned} \qquad (15)$$

As a result, $\boldsymbol{X}_{m-r}$, which cannot be explained by $\boldsymbol{Y}_r$, is calculated as follows:

$$\begin{aligned} \boldsymbol{X}_{m-r} &= \boldsymbol{X} - \boldsymbol{X}_r \\ &= \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{B}_r\boldsymbol{B}_r^T \boldsymbol{\Lambda}^{-1} \boldsymbol{P}^T). \end{aligned} \qquad (16)$$

Since $\boldsymbol{X}_{m-r}$ does not include significant non-Gaussian components, it can be monitored successfully by using PCA-SPC.

4.2 *CMSPC Procedure*

The procedure of CMSPC is summarized as follows:

(1) Acquire time-series data when a process is operated under a normal condition. Normalize each column (variable) of the data matrix, i.e., adjust it to zero mean and unit variance, if necessary.
(2) Apply ICA to the normalized data $\boldsymbol{X}$, and calculate independent components $\{y_l\}$.
(3) Calculate the fourth-order cumulant of independent components.
(4) Adopt independent components $\{y_1, y_2, \cdots, y_r\}$ with the fourth-order cumulant larger than the threshold (e.g. 0.1) as non-Gaussian independent components.
(5) The other components are regarded as Gaussian, and those variables are projected onto the original space through Eq. (16).
(6) Apply PCA to the reconstructed data $\boldsymbol{X}_{m-r}$, and calculate principal components $\{z'_1, z'_2, \cdots, z'_{m-r}\}$.
(7) Calculate T^2 and Q statistics.
(8) Determine control limits of independent components $\{y_1, y_2, \cdots, y_r\}$ and those of T^2 and Q.
(9) Monitor $\{y_1, y_2, \cdots, y_r\}$, T^2, and Q on-line.

CMSPC includes both PCA-SPC and ICA-SPC as its special cases. In fact, CMSPC is the same as PCA-SPC when no independent components are adopted in step (4), because PCA is applied to $\boldsymbol{X}_{m-r} = \boldsymbol{X}$ in such a case. On the other hand, CMSPC is the same as ICA-SPC when $r = rank(\boldsymbol{X})$. Therefore, CMSPC provides a unified framework for MSPC.

5. APPLICATION 1

In this section, USPC, PCA-SPC, ICA-SPC, and the proposed CMSPC are applied to fault detection problems of an eight-variable system:

$$\boldsymbol{x} = \boldsymbol{s}\boldsymbol{A} + \boldsymbol{v} \qquad (17)$$

$$\boldsymbol{A} = \begin{bmatrix} 0.95 & 0.82 & 0.94 & 0.14 \\ 0.23 & 0.45 & 0.92 & 0.20 \\ 0.61 & 0.62 & 0.41 & 0.20 \\ 0.49 & 0.79 & 0.89 & 0.60 \\ 0.89 & 0.92 & 0.06 & 0.27 \\ 0.76 & 0.74 & 0.35 & 0.20 \\ 0.46 & 0.18 & 0.81 & 0.02 \\ 0.02 & 0.41 & 0.01 & 0.75 \end{bmatrix}^T \qquad (18)$$

$$\boldsymbol{s} = \left[\, s_1 \ s_2 \ s_3 \ s_4 \,\right] \qquad (19)$$

where $\{s_i\}$ are uncorrelated random signals following uniform or normal distribution with unit variance ($\sigma_s = 1$). The output $\boldsymbol{x}$ is corrupted by measurement noise $\boldsymbol{v}$ following normal distribution ($\sigma_v = 0.1$). For evaluating the monitoring performance, mean shifts of $\{s_i\}$ or $\{x_j\}$ are investigated.

One data set, including 100,000 samples, obtained from a normal operating condition was used to build a PCA model, to determine a separating matrix, and also to determine control limits. To evaluate the monitoring performance, average run length (ARL) is used. ARL is the average number of points that must be plotted before a point indicates an out-of-control condition. To calculate ARL, 10,000 data sets were generated by changing seeds of the random signals $\boldsymbol{s}$ and $\boldsymbol{v}$ in each case shown in Table 1.

The control limit of each index or variable is determined so that the number of samples outside

Table 1. ARL Comparison.

Case 1				
	s_i :	uniform distribution		
	fault :	$\boldsymbol{s}_1$		
Shift	USPC	PCA-SPC	ICA-SPC	CMSPC
size	x_5	T_4^2	y_3	y_3
0	98.1	99.0	101	101
0.2	82.5	84.0	59.6	59.6
0.5	42.2	43.2	18.0	18.0
1.0	16.5	12.3	5.5	5.5
Case 2a				
	s_i :	normal distribution		
	fault :	$\boldsymbol{s}_1$		
Shift	USPC	PCA-SPC	ICA-SPC	CMSPC
size	x_5	T_4^2	y_3	T_4^2
0	96.0	101	97.3	101
0.2	91.9	96.0	91.7	96.0
1.0	33.5	36.6	37.6	36.6
2.0	8.9	8.1	10.8	8.1
Case 2b				
	s_i :	normal distribution		
	fault :	$\boldsymbol{s}_2$		
Shift	USPC	PCA-SPC	ICA-SPC	CMSPC
size	x_5	T_4^2	y_4	T_4^2
0	103	97.5	99.8	97.5
1.0	32.3	37.4	51.6	37.4
2.0	8.3	8.5	18.5	8.5
3.0	3.2	2.7	8.0	2.7
Case 3				
	s_1, s_2 :	uniform distribution		
	s_3, s_4 :	normal distribution		
	fault :	$\boldsymbol{x}_5$		
Shift	USPC	PCA-SPC	ICA-SPC	CMSPC
size	x_5	Q_4	y_8	Q_2
0	96.8	96.1	98.9	102
0.1	79.9	55.4	54.1	48.3
0.2	50.3	21.1	20.2	14.8
0.5	12.6	2.5	2.5	1.6

the control limit is 1% of the entire samples while the process is operated under a normal condition. The monitored indexes for PCA-SPC are T_4^2 and Q_4. The subscript 4 means that four principal components are retained in the PCA model. In ICA-SPC, however, each independent component is independently monitored. In CMSPC, the number of independent components and that of principal components retained in the PCA model depend on the cases. Four independent components and no principal components are retained in case 1, no independent components and four principal components in case 2, and two independent components and two principal components in case 3.

Fault detection results are summarized in Table 1. ARL decreases as the shift size increases, irrespective of the type of monitoring method. In case 1, the results have clearly shown the advantage of ICA-SPC and CMSPC over both USPC and PCA-SPC, and the ARL of ICA-SPC is the same as that of CMSPC because all original variables follow uniform distribution. On the other hand, in case 2, PCA-SPC and CMSPC are superior to ICA-SPC, and they achieve the same performance because all variables follow normal distribution. The difference between case 2a and 2b is the variable where the mean shift occurs. In case 2a, the monitoring performance of all four SPC methods is similar, and the advantage of using PCA-SPC over ICA-SPC is not clear. In case 2b, however, PCA-SPC outperforms ICA-SPC. Therefore, it is concluded that PCA-SPC functions better than or as well as ICA-SPC when all measured variables are Gaussian. Although PCA-SPC is better than ICA-SPC in case 2, these two methods do not outperform USPC. Even when measured variables are mutually correlated, USPC sometimes outperforms MSPC. In case 2b, USPC gives better performance than the others because a_{52}, which is the coefficient from s_2 to x_5 in $\boldsymbol{A}$, is larger than the others in the same row and thus the mean shift can be easily detected by monitoring x_5.

In case 3, two of four original variables follow uniform distribution and the other two variables follow normal distribution. In this case, CMSPC can detect the mean shift of x_5 earlier than the other methods. This result clearly shows the advantage of CMSPC over other SPC methods.

ICA-SPC functions well for generating and monitoring non-Gaussian independent variables, while PCA-SPC is suitable for monitoring Gaussian variables. Therefore, the answer to the question “Which MSPC method should be applied to our process?” depends on the process. However, the proposed CMSPC combines the advantages of both PCA-SPC and ICA-SPC, and thus it enables us to select the best solution automatically.

6. APPLICATION 2

In this section, four SPC methods are applied to monitoring problems of a CSTR process (Johannesmeyer and Seborg, 1999). The objective of this section is to show the usefulness of CMSPC with its application to a more realistic example.

The CSTR process used for dynamic simulations is shown in Fig. 1. The reactor is equipped with a cooling jacket. The process has two manipulated variables (valves) and five process measurements. A total of nine variables used for monitoring are listed in Table 2. Process data are generated from a normal operating condition and eight abnormal operating conditions listed in Table 3. All variables are measured every five seconds.

The control limit of each index or variable is determined in the same way as the previous section. Five principal components are retained in the PCA model for PCA-SPC. The number of non-Gaussian independent components is five,

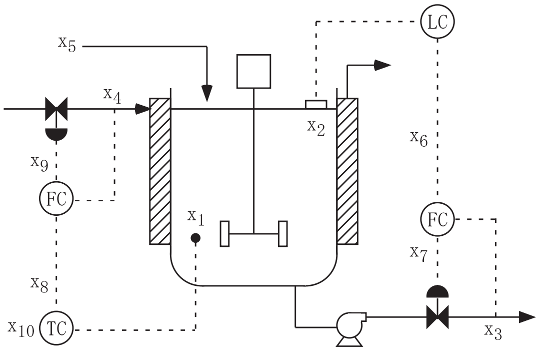

Fig. 1. CSTR with feedback control.

Table 2. Process variables.

x_1	reactor temperature
x_2	reactor level
x_3	reactor outlet flow rate
x_4	coolant flow rate
x_5	reactor feed flow rate
x_6	MV of level controller
x_7	MV of outlet flow controller
x_8	MV of temperature controller
x_9	MV of coolant flow controller

Table 3. Disturbances and faults.

Case	Operation Mode
N	normal operation
F1	dead coolant flow measurement
F2	bias in reactor temp. measurement
F3	coolant valve stiction
F4	feed flow rate - step
F5	feed concentration - ramp
F6	coolant feed temperature - ramp
F7	upstream pressure in coolant line - step
F8	downstream pressure in outlet line - step

Table 4. ARL Comparison (CSTR).

Case	USPC	PCA-SPC	ICA-SPC	CMSPC
N	94.7	115	96.6	95.4
F1	1.0	1.4	1.1	1.1
F2	7.9	8.7	8.8	8.8
F3	7.0	2.6	1.4	1.4
F4	49.5	23.1	1.0	1.0
F5	52.4	60.1	57.6	57.6
F6	60.0	61.8	55.8	55.8
F7	79.3	86.2	2.5	2.5
F8	61.7	6.6	1.0	1.0

and the other four components are monitored together by using T^2 in CMSPC. The results are summarized in Table 4. In this application, there is little or no difference of ARLs among four monitoring methods except F4, F7, and F8. In those three cases, ICA-SPC and CMSPC can detect the faults considerably earlier than USPC and PCA-SPC. In addition, ICA-SPC and CMSPC achieve the same performance in almost all cases because the faults tend to be detected at non-Gaussian independent components. The results clearly show the effectiveness of CMSPC as well as ICA-SPC.

7. CONCLUSIONS

A new advanced MSPC method, referred to as combined MSPC (CMSPC), was developed by integrating conventional PCA-SPC and recently proposed ICA-SPC. CMSPC includes both PCA-SPC and ICA-SPC as its special cases, and thus it provides a unified framework for MSPC. The application results show that CMSPC functions very well and it combines the advantages of both PCA-SPC and ICA-SPC.

REFERENCES

Delfosse, N. and P. Loubaton (1995). Adaptive Blind Separation of Independent Sources: A Deflation Approach. *Signal Processing*, **45**, 59–83.

Hyvarinen, A. and E. Oja (1997). A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, **9**, 1483–1492.

Jackson, J.E. and G.S. Mudholkar (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, **21**, 341–349.

Johannesmeyer, M. and D.E. Seborg (1999). Abnormal Situation Analysis Using Pattern Recognition Techniques and Historical Data. *AIChE Annual Meeting*, Dallas, TX, Oct.31-Nov.5.

Jutten, C. and J. Herault (1991). Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture. *Signal Processing*, **24**, 1–10.

Kano, M., K. Nagao, H. Ohno, S. Hasebe, I. Hashimoto, R. Strauss, and B.R. Bakshi (2002a). Comparison of Multivariate Statistical Process Monitoring Methods with Applications to the Eastman Challenge Problem. *Comput. & Chem. Engng*, **26**, 161–174.

Kano, M., S. Tanaka, H. Ohno, S. Hasebe, and I. Hashimoto (2002b). The Use of Independent Component Analysis for Multivariate Statistical Process Control. *Proc. of Int'l Symp. on Advanced Control of Industrial Processes (AdCONIP'02)*, 423–428, Kumamoto, Japan, June 10–11.

Kano, M., S. Tanaka, S. Hasebe, and I. Hashimoto, H. Ohno (2003). Monitoring of Independent Components for Fault Detection. *AIChE J.* (in press)

Kresta, J.V., J.F. MacGregor, and T.E. Marlin (1991). Multivariate Statistical Monitoring of Process Operating Performance. *Can. J. Chem. Eng.*, **69**, 35–47.

BATCH MONITORING THROUGH COMMON SUBSPACE MODELS

S. Lane, E.B. Martin and A.J. Morris

Centre for Process Analytics and Control Technology,
School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, England, UK

Abstract: Multi-way statistical projection techniques have typically been applied in the development of monitoring models for single recipe or single grade production. As defined, implementation of these techniques in multi-product applications necessitates the development of a large number of process models. This issue can be overcome through the use of common sub-space models constructed by pooling the individual variance-covariance matrices. A second issue with multi-way approaches is the difficulty of interpreting multi-way contribution plots. An alternative approach is the U^2 statistic. In this paper an extension is proposed, the V^2 statistic, based on the cumulative contribution of variables at each sample point. The methodologies are demonstrated on two industrial applications. *Copyright © 2002 IFAC*

Keywords:- Modelling; Statistical Analysis; Manufacturing Processes

1. INTRODUCTION

Over the last decade the emphasis in process manufacturing has changed. Quality and product consistency have become major consumer decision factors and are key elements in determining business success, growth and competitive advantage. Manufacturing products that meet their quality specifications first time result in higher productivity, reduced manufacturing costs through less re-work, give-away and waste. This all contributes to reducing the impact of the process on the environment by minimising raw materials and energy usage. The achievement of right first time production requires a reduction in process variability and thus the monitoring of process behaviour over time to ensure that the key process/product variables remain close to their desired (target) value is essential. This has led to a significant increase in the industrial application of statistical methods for interrogating the process to obtain an enhanced understanding of the process and the implementation of Statistical Process Control (SPC) for process monitoring and the early warning of the onset of changes in process behaviour.

An area of rapidly growing interest for the monitoring of processes is that of Multivariate Process Performance Monitoring (MPPM). MPPM schemes have typically been based on the statistical projection techniques of Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) and their multi-way extensions for batch processes. Reported practical applications of MPPM have focused on the production of a single manufactured product i.e. one grade, one recipe, etc. with separate models being used to monitor different types of products (Kosanovich and Piovoso, 1995, Kourti *et al*, 1995, Rius *et al*, 1997, Martin *et al,* 1999). However, in recent years, process manufacturing has increasingly been driven by market forces and customer needs resulting in the necessity for flexible manufacturing to meet the requirements of changing markets and product diversification. Thus with many companies now producing a wide variety of products, there is a real need for process monitoring models which allow a range of products, grades or recipes to be monitored using a single process representation.

The elimination of between group variation is a prerequisite for statistical process monitoring, so that interest can focus on within process (product) variability. This normally requires constructing separate control charts for each type of product or grade to be monitored. In many process monitoring situations this may be impractical because of the large number of control charts required to monitor all the products being manufactured and the limited amount of data available from which to develop a process representation. An extension to multi-way PCA and multi-way PLS that allows the construction of a multiple group model is proposed based on combining the variance-covariance matrices of each of the individual groups. The loadings for the latent variables are then calculated from the pooled variance-covariance matrix of the individual groups. Previous work has been published on the multiple-group PCA algorithm, Lane *et al*, (2001) and thus the paper focuses on the multi-group PLS algorithm.

2. PROJECTION TO LATENT STRUCTURES

A brief overview of the PLS algorithm is presented. A more detail discussion of the methodology can be found in Garthwaite (1994). The objective of PLS is to determine a set of latent variable scores that "best" describe the variation in the process data set (**X**) data set that is most influential on the quality data set (**Y**) data set. Using these latent variables it is then possible to construct a set of latent variable scores for the process data i.e. **T** = **XW**, where **T** is the matrix of latent variable scores and **W** is the matrix containing the latent variable loadings. A number of different algorithms have been proposed to derive the loadings for the latent variables associated with PLS. One approach is based on the extensions to the NIPALS (Non-linear Iterative Partial Least Squares) method, which regresses the columns of **X** on **Y** directly. As a consequence, it is not feasible to combine a number of different data sets into a single model.

Lindgren *et a,l* (1993) presented a kernel algorithm for determining the latent variables that is based on the eigenvector decomposition of the variance-covariance matrix, $\mathbf{R} = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. By adapting the kernel algorithm, a multiple group model can be constructed by pooling the individual variance-covariance matrices ($\mathbf{R}_i$). In this way the formal statistical basis for the multiple group model, as given by Flury (1987), can be extended. The variance-covariance approach is based on the hypothesis that the first *a* eigenvectors of each of the individual variance-covariance matrices span the same common subspace. Although the model introduced by Flury (1987) related to common principal components, the hypothesis is also appropriate for PLS, since it is the variance-covariance matrices that are of interest. Krzanowski (1984) had previously shown that the common loadings for the latent variables could be extracted from a weighted sum of the individual variance-covariance matrices.

3. THE MULTIGROUP PLS ALGORITHM

The algorithm for constructing the multiple group model based on the kernel algorithm is as follows:

1. Calculate the kernel matrices for each individual group:

$$\mathbf{R}_i = \left(\frac{\mathbf{X}_{is}^T\mathbf{Y}_{is}}{n_i - 1}\right)\left(\frac{\mathbf{X}_{is}^T\mathbf{Y}_{is}}{n_i - 1}\right)^T \quad i = 1, \ldots., g \tag{1}$$

where $\mathbf{X}_{is}$ is the matrix of scaled process data for group *i*, $\mathbf{Y}_{is}$ is the matrix of scaled quality data for group *i*, n_i is the number of observations in group *i*, and *g* is the total number of groups.

2. Construct the pooled kernel matrix:

$$\mathbf{R}_p = \frac{\sum_{i=1}^{g}(n_i - 1)\mathbf{R}_i}{(N - g)} \qquad i = 1, \ldots., g \tag{2}$$

where $N = \sum_{i=1}^{g} n_i$

3. Calculate the loading vector ($\mathbf{w}_k$) for the process variables, where ($\mathbf{w}_k$) is the first eigenvector of the pooled kernel matrix ($\mathbf{R}_p$).

4. Once ($\boldsymbol{w}_k$) has been estimated, the latent variable scores for each group ($\boldsymbol{t}_{ik}$) can be calculated: -

$$\mathbf{t}_{ik} = \mathbf{X}_{is}\mathbf{w}_k \tag{3}$$

where $\mathbf{t}_{ik}$ is the matrix of principal component scores for group *i* and dimension *k*, $\mathbf{w}_k$ is the common latent variable loading for dimension *k* and $\mathbf{X}_{is}$ is the scaled data matrix for group *i*.

5. The loading vectors ($\mathbf{p}_i$) and ($\mathbf{q}_i$) are then calculated as:

$$\mathbf{p}_i = \frac{\mathbf{t}_{ik}^T\mathbf{X}_{is}}{\mathbf{t}_{ik}^T\mathbf{t}_{ik}} \qquad \mathbf{q}_i = \frac{\mathbf{t}_{ik}^T\mathbf{Y}_{is}}{\mathbf{t}_{ik}^T\mathbf{t}_{ik}} \tag{4}$$

where $\mathbf{Y}_{is}$ is the matrix of scaled quality data.

6. The process and quality data matrices are then deflated:

$$\mathbf{X}_{is\,new} = \mathbf{X}_{is} - \mathbf{t}_{ik}\mathbf{p}_i^T$$
$$\mathbf{Y}_{is\,new} = \mathbf{Y}_{is} - \mathbf{t}_{ik}\mathbf{q}_i^T \tag{5}$$

The next $(k + 1)^{\text{th}}$ latent variable is then calculated:

$$\mathbf{t}_{ik+1} = \mathbf{X}_{is\,new}\mathbf{w}_{k+1} \tag{6}$$

where $\mathbf{w}_{k+1}$ is the first eigenvector of the updated pooled kernel matrix:

$$\mathbf{R}_{p\,new} = \frac{\sum_{i=1}^{g}(n_i - 1)\mathbf{R}_{i\,new}}{N - g} \tag{7}$$

and

$$\mathbf{R}_i \ new = \left(\frac{\mathbf{X}^T_{is\,new}\mathbf{Y}_{is\,new}}{n_i - 1} \right)\left(\frac{\mathbf{X}^T_{is\,new}\mathbf{Y}_{is\,new}}{n_i - 1} \right)^T \qquad (8)$$

The iteration process continues with new values for $\mathbf{p}_i$ and $\mathbf{q}_i$ being calculated. Finally the data matrices $\mathbf{X}_{isnew}$ and $\mathbf{Y}_{isnew}$ are deflated. The iteration process steps (1 to 6) are repeated until the required numbers of latent variables have been extracted.

4. MPCA AND MPLS FOR MONITORING BATCH DATA

Batch processes differ from continuous processes in that each variable, *j*, is measured at *k* time intervals for a total of *I* batches. The data set is thus three-dimensional (*I* x *J* x *K*). As a consequence interest is in both the "between" and "within" batch variability. The application of MPCA or MPLS to the three-dimension data array associated with batch manufacturing is equivalent to performing standard PCA or PLS on a large two-dimensional data matrix formed by unfolding the original three-dimensional array. The unfoding approach adopted in this paper is that proposed by Kourti *et* al, (1995) and demonstrated in Fig. 1. This approach allows the variability between batches to be analysed by summarising the variability in the data with respect to both variables and their time variation. The data contained in the two-dimensional matrix is mean centred and scaled prior to applying either MPCA or MPLS. By subtracting the mean of each column from the two-dimensional data matrix the non-linearities are effectively removed from the data.

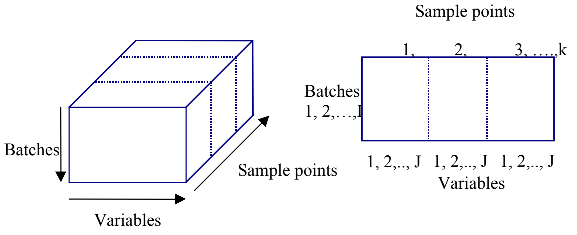

Fig. 1. Data unfolding

5. MULTIPLE GROUP MPCA AND MPLS

As described in Section 3, the pooled correlation (variance-covariance) approach is based on the existence of a common eigenvector subspace spanned by the first *a* eigenvectors of the individual correlation (variance-covariance) matrices. A formal statistical model was given by Flury (1987), who computed the common principal components using Maximum Likelihood Estimation (MLE). Krzanowski (1984) had previously demonstrated that the common principal components derived using the pooled correlation (variance-covariance) matrix were almost identical to those computed from MLE. In practice the pooled correlation (variance-covariance) approach proposed by Krzanowski (1984) is simpler to apply than the MLE approach, which requires the implementation of an iterative algorithm. The pooled correlation (variance-covariance) approach compares the subspaces defined by the eigenvectors associated with the largest eigenvalues. No conditions are placed on the MLE proposed by Flury (1987). This is a major consideration when determining the method to be used for calculating the latent variables for process monitoring. In process monitoring, it is convention to construct the process models using the eigenvectors corresponding to the largest eigenvalues. As a consequence determining the common latent variables from the pooled correlation (variance-covariance) matrix is more appropriate for industrial applications.

6. V^2 CONTRIBUTION PLOTS

The contribution plots introduced by Miller *et al*, (1998) are formulated from the weighted contribution of each variable to the principal component (latent variable) score at the sample points of interest. In the batch monitoring approach adopted in this paper there are a large number of variable contributions (variable x sample points) to analyse. In some situations this can make the contribution plots difficult to interpret. Furthermore the deviations usually impact on the manufacturing process over a number of sample points. As a consequence the development of a contribution plot that indicates the cumulative contribution of each variable to the principal component (latent variable) scores at each sampling point is desirable. The cumulative contribution of each variable is better related to the latent variable scores, whose deviation from the centre of the control region is usually caused by the cumulative affect of small deviations from the mean batch trajectory.

The V^2 statistic is an extension of the U^2 statistic of Runger (1996) and Runger and Alt (1996) and is proposed as a technique for examining the cumulative contribution of each variable individually or as a group of variables, at each sample point. The V^2 statistic is calculated as the difference between two T^2 statistics. The first includes the entire variable set and the second excludes the variable or groups of variables whose contribution is of interest. To examine the cumulative contribution to the batch scores, a V^2 statistic is calculated at each sample point this requiring the calculation of:

$$V_1^2 = T^2 - T_1^2 \qquad (9)$$

where T_1^2 excludes the variable or variables of interest at the first sample point. At the second sample point (V_2^2) is calculated from:

$$V_2^2 = T^2 - T_2^2 \qquad (10)$$

where T_2^2 excludes the variable(s) of interest at both the first and second sample points. The calculation is repeated at each sample point to obtain V_3^2 , V_4^2, etc. until the end of the batch run. Each individual V^2 statistic can then be plotted as a bar graph, which shows the cumulative contribution of each variable or group of variables at each sample point.

7. PROCESS PERFORMANCE MONITORING

7.1 Case Study 1

To demonstrate the application of multiple group multi-way PCA, three sets of data from a metal etcher process were considered (Wise *et al,* 1999). Data was supplied from an A1-stack etching process that was being performed using a Lam 9600 plasma-etching tool. The objective of the process is to etch the NiN/A1-0.5% Cu/TiN/oxide stack with an inductively coupled BCI_3/CI_2 plasma. The standard manufacturing process consists of a series of six steps. The first two are for the achievement of gas flow and stabilisation. Steps 3 and 4 are the brief plasma ignition step and the main etch of the A1 layer terminating at the A1 endpoint respectively. The next step acts as an over etching for the underlying TiN and oxide layers whilst the final step is associated with the venting of the chamber.

Etching of an individual wafer is analogous to a single batch in a chemical process. Changes in the process mean are a result of a residue building up on the inside of the chamber following the cleaning cycle, differences in the incoming materials resulting from changes in the upstream process and drift in the process monitoring sensors themselves. As a result of the changes in the process mean there are three distinct operating levels identified in the data set. When the data is combined into a single data set, the scores of principal component 1 and principal component 2 identify the discrete operating levels as seen in Fig. 2.

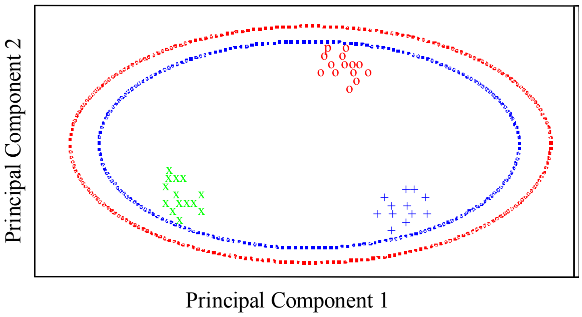

Fig. 2. Bivariate scores plot (Mixed covariance model)

In this case the major source of variation explained by the individual principal components is the variation of each variable from the overall mean of the data set and thus identifies the different operating conditions of each variable. This between group variation present in the data set causes the principal component scores to cluster according to which operating region, or grade of product, they represent. When such clustering occurs there are two issues that impact on process monitoring: (i) the control limits may be conservative and assignable cause process events may not be detected and (ii) an assignable cause reflected in the movement of a principal component score into another cluster when the operating conditions have not been changed may result in the real process event not being detected. As a consequence of the clustering observed and due to the changing mean levels, the process data was divided into three subsets one for each of the operating levels. The composition of each of these data sets is presented in Table 1.

Table 1. The Metal Etcher Data Sets

Operating level	Observations	Variables	Batches
1	90	17	17
2	90	17	16
3	90	17	18

A reference model for the multiple group application was then constructed using the three data sets. By analysing each of the data sets, it was inferred that the different operating levels share common characteristics that determine the process behaviour and as a consequence the use of the multiple group modelling approach was validated. Ten principal components were selected from cross-validation explaining 68%. A bivariate scores plot for principal components 1 and 2, Fig. 3, shows that the scores are independent and identically distributed. As a result it was inferred that the multiple group monitoring model provided a good representation of the overall etch process.

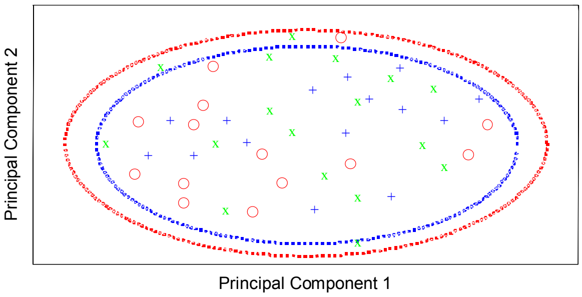

Fig. 3. Bivariate scores plot (Pooled covariance model)

To evaluate the detection and diagnostic capabilities of the multiple group model, a data set containing an increase in the TCP power was projected onto the reference model. This was done in a manner so as to simulate an on-line monitoring situation. Each observation that is projected onto the monitoring chart represents the status of an ‘on-line batch’ at successive sampling points during the etch process. The bivariate scores plot of principal components 1 and 2 (Fig. 4) detects the change in the operating conditions as a slow drift away from the centre of the control region.

At the beginning of the etch run, the principal component scores lie in the centre of the control region. After the first few sample points the scores gradually drift away from the centre towards the control limits. An out-of-statistical-control signal is flagged as the scores cross the action limits. In this particular example no remedial action was taken and the scores continue to drift away from the control region until the conclusion of the process run.

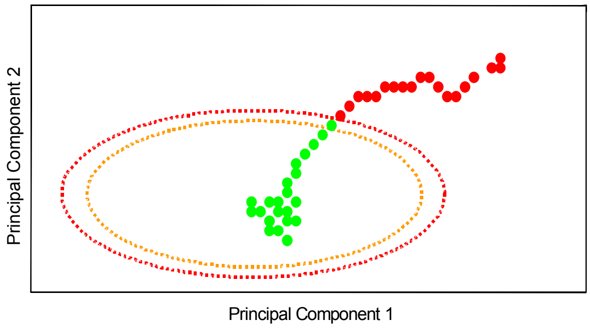

Fig. 4. Bivariate scores plot

The V^2 contribution plot, Fig. 5, identifies the variable indicative of the out-of-statistical-control signal. For clarity only the contribution from a single variable is plotted and it can be seen that the contribution from the variable follows a similar profile to the principal component scores shown in Fig. 4. These results demonstrate the power of the multi-group modelling approach and confirm the findings of Wise *et al*, (1999).

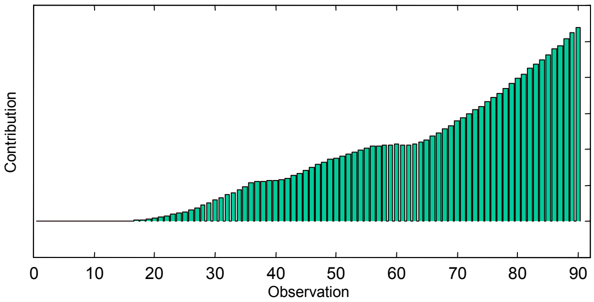

Fig. 5. V^2 Contribution Plot

7.2 Case Study 2

The industrial process used to demonstrate the on-line application of the multiple group multi-way PLS model is a polymer film manufacturing process. The manufacture of polymer film can be considered as a series of unit operations that are applied to convert polymer pellets to a rolled film product (e.g. Weighell *et al*, 2001). A number of different film types are manufactured using the same process equipment through changes in the operating conditions being made and the types of polymer pellets used. Following the production of each roll of film, a number of quality attributes are measured at the end of the roll. As a consequence each roll of film can be considered as a separate batch of product. In this example, 105 process variables and 3 quality variables are included in the model. These provide a description of the "well being" of the process and its manufacturing performance, although at present the process operators only monitor a few "key" process variables.

Separate performance monitoring charts were constructed for each unit within the manufacturing process. In this example the unit of interest is the sheet forming process. Two grades of film manufactured using two different production lines were used to construct the multiple group model. In this particular plant there are a number of different lines manufacturing polymer film and as a consequence there is both between line variation in the data as well as the between polymer grade variation. The composition of each data set is shown in Table 2. Again as in the previous Case study, initially all the data was combined into a single data matrix. For comparative purposes, standard single group MPLS was carried-out on the combined data matrix.

Table 2. The Films Production Line Data

Data Set	Line	Grade	Obs.	Proc Vars	Qual Vars	Btch's
1	1	1	100	19	3	23
2	1	2	100	19	3	23
3	2	1	100	19	3	19
4	2	2	100	19	3	19

Inspection of the bivariate scores plot of latent variables 1 and 2 showed, as expected, the scores to be clustered into four distinct regions (not shown). Combining the data into a single matrix and applying standard multi-way PLS does not result in a satisfactory model for on-line process monitoring (Lane *et al*, 2001). Inspection of the bivariate scores plot of latent variables 1 and 2 for the multiple group model showed the latent variable scores to be independent and normally distributed (not shown), implying that the multiple group model was appropriate for monitoring the polymer film manufacturing process.

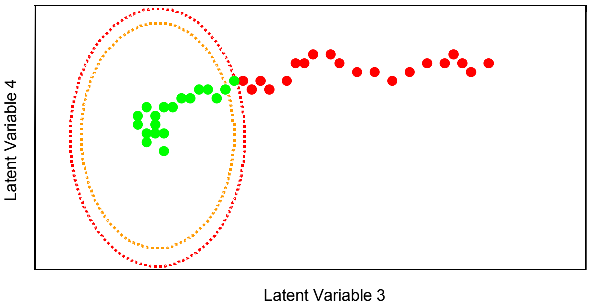

Fig. 6. Bivariate scores plot of latent variable 3 versus latent variable 4

A data set containing a process fault, a reduction in the pressure, was projected onto the reference model. Fig. 6, represents the on-line status of the batch at successive sample points during the manufacturing process. The bivariate scores plot of latent variables 3 and 4, Fig. 6, detects a process disturbance. As with the multiple group multi-way PCA Case Study, the scores are seen to drift away from the centre of the control region. Again in this particular application no remedial action could be taken when the abnormal situation was detected.

Once the out-of-statistical-control operation was signalled, a V^2 contribution plot was used to determine the variables indicative of the out-of-statistical-control signal. In this case a change in pressure impacted on two of the process variables. Fig. 7 shows the V^2 statistic for the joint contribution of these two variables. In this case the V^2 statistic is calculated by omitting both variables from the T_i^2 statistic (Equations 9 and 10) at each sample point.

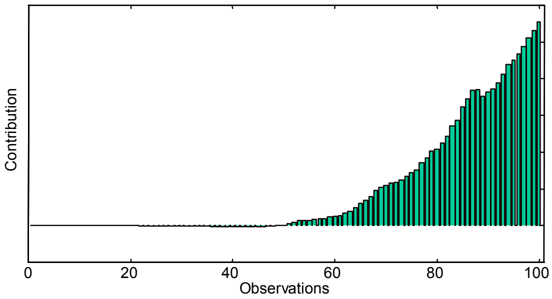

Fig. 7. V^2 contribution plot

8. CONCLUSIONS AND DISCUSSION

The applications presented in the paper are from manufacturing processes where the amount of data from each distinct set of operating conditions or product grade was limited. Prior analysis of the individual data sets suggested that the data available was not sufficient to give a true reflection of the underlying process variability. By applying the pooled correlation (variance-covariance) approach, all the products being manufactured could be monitored using a small number of monitoring charts. In this way the cost and time required to update the models can be significantly reduced allowing the faster on-line implementation of process performance monitoring schemes. The multiple group process monitoring techniques were also demonstrated to have good fault detection and diagnostic capabilities.

An alternative contribution plot for batch processes has also been proposed. The objective of the multi-group modelling methodologies proposed is to monitor the deviation of each variable from its mean trajectory. As a consequence, most out-of-statistical-control signals are the result of a variable, or group of variables deviating from their mean trajectories over a number of sample points. By examining the cumulative contribution of each variable, or group of variables, at each sample point a clear indication of their influence on the out-of-statistical-control signal can be obtained.

9. ACKNOWLEDGEMENTS

The authors acknowledge the support of the UK EPSRC Awards GR/M09971 and GR/R35100/01.

10. REFERENCES

Flury, B.N. (1987), Two generalisations of the common principal component model, *Biometrika*, **74**, pp. 59-69.

Garthwaite, P.H. (1994), An interpretation of partial least squares. *Journal of the American Statistical Association*, **89**, pp. 122-127.

Kosanovich, K.A., S. Dahl and M.J. Piovoso (1996). Improved process understanding using multi-way principal component analysis, *Ind. Eng. Chem. Res.*, **35**, pp. 138-146.

Kourti, T., P. Nomikos and J.F. MacGregor (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control,* **5**, pp. 277-284.

Krzanowski, W.J. (1984). Principal component analysis in the presence of group structure, *Applied. Statistics*, **33(2)**, pp. 164-168.

Lane, S., E.B. Martin, A.J. Morris and R.A.G. Kooijmans (2001). Performance monitoring of a multi-product semi-batch processes. *Journal of Process Control*, **11**, pp. 1-11

Lindgren, F., P. Geladi and S. Wold (1993). The kernel algorithm for PLS. *Journal of Chemometrics,* **7**, pp. 45-59.

Martin, E.B., A.J. Morris and C. Kiparrisides (1999). Manufacturing Performance Enhancement through Multivariate Statistical Process Control, *Annual Reviews in Control*, **23**, pp. 35 - 44.

Miller, P., R. E. Swanson and C.E. Heckler (1998). Contribution plots: A Missing Link in Multivariate Quality Control, *Applied Mathematics and Computer Science*, **8**, pp. 775-792.

Rius, A., M.P. Callao and F.X. Rius (1997). Multivariate statistical process control applied to sulfate determination by sequential injection analysis. *The Analyst,* **122**, 737-741.

Runger, G.C. (1996), Projections and the U^2 Multivariate Control Chart, *Journal of Quality Technology*, **28**, pp. 313-318.

Weighell, M., E.B. Martin and A.J. Morris, (2001), The statistical monitoring of a complex manufacturing process. *Journal of Applied Statistics,* **2**8(3&4), pp. 409-425.

Wise, B. M., N.B. Gallagher, S. Watts Butler, D.D. White jr. and G.G. Barna (1999). A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics*, **13**, pp. 379-396.

APPLICATION OF PLS-BASED REGRESSION FOR MONITORING BITUMEN RECOVERY IN A SEPARATION CELL

Harigopal Raghavan [*] **Sirish L. Shah** [*]
Ramesh Kadali [**] **Brian Doucette** [**]

[*] *Department of Chemical & Materials Engineering,*
University of Alberta, Edmonton, AB, Canada
[**] *Suncor Extraction, Fort McMurray, AB, Canada*

Abstract:
Partial Least Squares (PLS) is a technique used to perform regression between blocks of explanatory variables and dependent variables. PLS uses projections of original variables along directions which maximize the covariance between these blocks. It has been popular due to its data-reduction property and its ability to handle collinearity within the data blocks. In this paper some issues which arise in the the development of multivariate static models of industrial processes using PLS regression are studied. An industrial example of the application of PLS regression for the development of inferential sensors to predict the Bitumen Recovery in a separation cell is shown. Some of the challenges encountered in the development and online implementation of the inferential sensors and the proposed solutions are presented.

Keywords: Soft-sensor, Partial Least Squares regression, Online implementation, Bitumen Recovery, Monitoring

1. INTRODUCTION

In many chemical engineering applications, control variables may not be available as frequently as would be desired for satisfactory closed-loop control. For example, key product quality variables are available after several hours of lab analysis. Often, it is possible to estimate the quality variables using other process variables which are measured frequently. The relationship or the model that is used to predict quality variables using other process variables is often called a “soft-sensor”. The quality-variable estimator is called a soft-sensor since it is based on software calculations rather than a physical instrument. The soft-sensors developed in this way can be used for inferential control or process monitoring. Discussions on inferential control can be found in (Kresta *et al.*, 1994; Parrish and Brosilow, 1985; Amirthalingam *et al.*, 2000; Li *et al.*, 2002).

Multivariate statistical techniques such as Principal Components Analysis (PCA) and PLS have been applied for process monitoring, fault detection and static modelling in chemical processes (Kresta *et al.*, 1991; Qin and McAvoy, 1992; Qin, 1993; Nomikos and MacGregor, 1995; Ricker, 1988). In addition, extensions of these approaches for handling dynamic and auto-correlated data have been proposed (Ku *et al.*, 1995; Lakshminarayanan *et al.*, 1997). In particular, PLS regression is a popular technique used in the development of soft-sensors in the form of static models for multivariate processes. The main advantage of using PLS for process modelling comes from its ability to decompose the problem of obtaining model coefficients from multivariate data into a set of univariate regression problems. Univariate regression is performed on latent variables obtained by projecting the input and output data onto directions along which the covariance between these variables is maximized. The models obtained through this exercise can then be used for monitoring the current state of the process. The advantages in using static models for monitoring include the simplicity of the models and the ease of implementation and maintenance.

2. PLS REGRESSION

The commonly used procedure for PLS is as follows:

Consider the zero-mean, unit variance data matrices $\mathbf{X} \in \Re^{N \times m}$ and $\mathbf{Y} \in \Re^{N \times p}$ where N is the number of observations, m is the number of process variables and p the number of quality variables. A linear static model explaining $\mathbf{Y}$ based on $\mathbf{X}$ is given as follows:

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{1}$$

Using the well known Ordinary Least Squares regression (OLS) we obtain the solution:

$$\hat{\mathbf{C}}_{ols} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

However, because of the high degree of correlation among the variables within the predictor space the matrix $\mathbf{X}^T\mathbf{X}$ may be ill-conditioned. In addition we may be interested in obtaining the directions along which the common (second-moment) information between these blocks is concentrated. To satisfy these objectives, the following procedure is adopted in PLS regression. The matrix $\mathbf{X}$ is decomposed into a score matrix $\mathbf{T} \in \Re^{N \times a}$ and a loadings matrix $\mathbf{P} \in \Re^{m \times a}$, where a is the number of PLS components used. Hence the following decomposition is achieved:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \tag{3}$$

where $\mathbf{E}$ is a residual matrix. Similarly $\mathbf{Y}$ is decomposed as

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} \tag{4}$$

To obtain the loadings vectors the following algorithm is used:

(1) Initialize, $\mathbf{Y}_1 = \mathbf{Y}$ and $\mathbf{X}_1 = \mathbf{X}$ and $i = 1$.
(2) Perform SVD on $\mathbf{X}_i^T\mathbf{Y}_i$ and calculate $\mathbf{j}_i$, the left singular vector corresponding to the largest singular value ω_i and $\mathbf{q}_i$ the corresponding right singular vector. This SVD calculation corresponds to capturing the direction ($\mathbf{j}_i$,$\mathbf{q}_i$) which maximizes covariance between $\mathbf{X}_i$ and $\mathbf{Y}_i$.
(3) Let $\mathbf{t}_i$ and $\mathbf{u}_i$ be the corresponding scores. Perform a univariate regression between $\mathbf{t}_i$ and $\mathbf{u}_i$ to obtain $\mathbf{b}_i$.
(4) The loadings vector for $\mathbf{X}_i$ is given by

$$\mathbf{p}_i = \frac{\mathbf{X}_i^T\mathbf{t}_i}{\mathbf{t}_i^T\mathbf{t}_i} \tag{5}$$

(5) Deflate $\mathbf{Y}$ and $\mathbf{X}$ according to

$$\mathbf{Y}_{i+1} = \{\mathbf{Y}_i - \mathbf{b}_i\mathbf{t}_i\mathbf{q}_i^T\} \tag{6}$$
$$\mathbf{X}_{i+1} = \{\mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^T\} \tag{7}$$

(6) Set $i = i + 1$.
(7) Go to step 2.

After a stages the approximations are

$$\mathbf{X} \approx \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_a\mathbf{p}_a^T \tag{8}$$
$$\mathbf{Y} \approx \mathbf{u}_1\mathbf{q}_1^T + \mathbf{u}_2\mathbf{q}_2^T + \cdots + \mathbf{u}_a\mathbf{q}_a^T \tag{9}$$

Hence we get the PLS estimate of the model coefficients as:

$$\hat{\mathbf{C}}_{pls} = \mathbf{J}(\mathbf{P}^T\mathbf{J})^{-1}\mathbf{BQ}^T \tag{10}$$

where, the columns of $\mathbf{J}$ and $\mathbf{Q}$ contain the singular vectors of the SVD's carried out at each stage, the columns of $\mathbf{P}$ contain the loadings vectors of the $\mathbf{X}$ matrix and $\mathbf{B}$ is a diagonal matrix containing the latent variable regression coefficients from each stage.

3. PROCESS DESCRIPTION

An industrial example of the application of PLS regression is presented in this section. Soft-sensors were developed to predict the Bitumen Recovery in a separation cell. These soft-sensors have been implemented online at Suncor Energy's Extraction facility at Fort McMurray in Alberta, Canada. The separation cell is used in the extraction of bitumen from oil sands. Oil sands are deposits of bitumen, that must be treated to convert them into crude oil which can then be refined in conventional refineries. The main processes in converting the oil sands to crude oil are Mining, Extraction and Upgrading. In the mining stage, the oil sands are mined using trucks and shovels. This is followed by the extraction stage in which bitumen is separated from the sand using processes such as froth-flotation. The bitumen is then converted to crude oil in the upgrading stage.

The extraction operations can be briefly described as follows: The oil sand is first passed through a slurry preparation stage. The main operation in this stage is to form a slurry using hot water, oil sands and caustic. Heat is used to reduce the viscosity of the bitumen. Caustic helps in the attachment of bitumen to the air in the froth formation while releasing it from the sand particles. The bitumen then forms small globules that are important in the formation of froth. Agitation also aids in the breaking up the oil sand. The slurry passes through a series of vibrating screens that separate and reject any rocks or clumps of clay still present in the slurry. It is then pumped into separation cells.

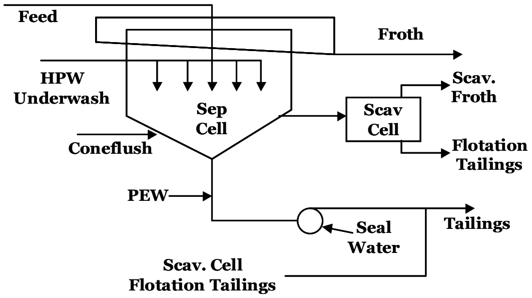

Fig. 1. Process Flow-sheet for Separation Cell

A schematic of a separation cell is shown in Fig. 1. The separation cell allows the slurry to settle out into its various layers, the most important layer being the froth layer which rises to the top. The tailings sand sinks to the bottom. The middle layer is called the middlings layer and consists of bitumen, clay and water. The middlings remain suspended between the sand and the bitumen froth until it is drawn off and sent through the secondary separation cell. The secondary separation vessel extracts the remaining bitumen from the middlings. The main objective in the operation of the separation cell is to maximize the amount of bitumen in the froth and minimize the amount of bitumen lost in the tailings and middlings streams. A measure of the efficiency of operation of the separation cell is given by the Bitumen Recovery which can be calculated from the predictions of quality variables using the following equation:

$$Rec = \frac{F_{fr}\rho_{fr}C_{fr}}{F_{fr}\rho_{fr}C_{fr} + F_t\rho_tC_t + F_{ft}\rho_{ft}C_{ft}} \quad (11)$$

where, Rec is the Bitumen Recovery in the cell, F_{fr}, F_t & F_{ft} refer to the Froth, Tailings and Flotation Tailings flows, ρ_{fr}, ρ_t & ρ_{ft} refer to the Froth, Tailings and Flotation Tailings densities and C_{fr}, C_t & C_{ft} refer to the concentrations of Bitumen in the Froth, Tailings and Flotation Tailings in wt% respectively. Hence the quality variables of interest are concentrations of Bitumen in the Froth, the Tailings and the Flotation Tailings. In our soft-sensor development, we used 25 process variables, measured every minute, to predict these 3 quality variables. Of the three product variables, one was available through lab analysis every 12 hours and the other two were available every 2 hours.

4. CHALLENGES IN SOFT-SENSOR DEVELOPMENT

While there have been other reported applications of PLS regression for developing soft sensors, we consider the current application to be especially challenging. Monitoring the extraction of bitumen from oil sands is a problem which poses some unique challenges. These include, in the words of a practicing engineer from this industry, "*changing process conditions, wide operating regions, bad data and lack of good software resources*". In addition we have encountered other challenging problems for which we have some suggested solutions. The challenges and the proposed solutions for bitumen recovery estimation are discussed below. Many of these solutions may also apply to other applications.

4.1 Sample consolidation

One of challenges encountered while developing these soft-sensors is due to the practice of physical consolidation of samples of the quality variables. It involves mixing a number of physical samples of the product collected at different time instants before performing lab analysis. For the process under consideration, consolidation is achieved using a flow totalizer and a triggering mechanism. When the cumulative flow in a line exceeds a set point it sets off a mechanism which leads to the collection of a sample in a container. The consolidation mechanism is illustrated in Fig. 2. This process continues for about 12 hours at the end of which, the container has a mixture of the samples collected over this period. This liquid is then stirred for homogeneity and the consolidated sample is used for analysis. In order to build realistic models using such samples, it is important that the modelling methodology including the data pre-treatment mimic the process as much as possible. Hence we resorted to time-averaging of the input data as dictated by the sample consolidation mechanism before the actual regression was performed.

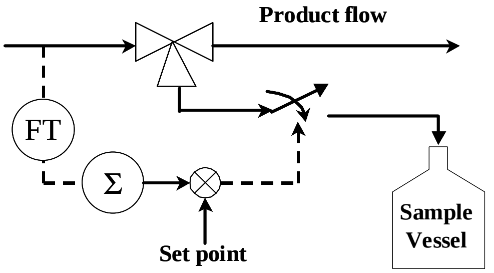

Fig. 2. Sample Consolidation mechanism

Let us assume that $k+1$ samples were collected at times $T_1, T_1+t_1, T_1+t_2, ..., T_2 = T_1+t_k$, where

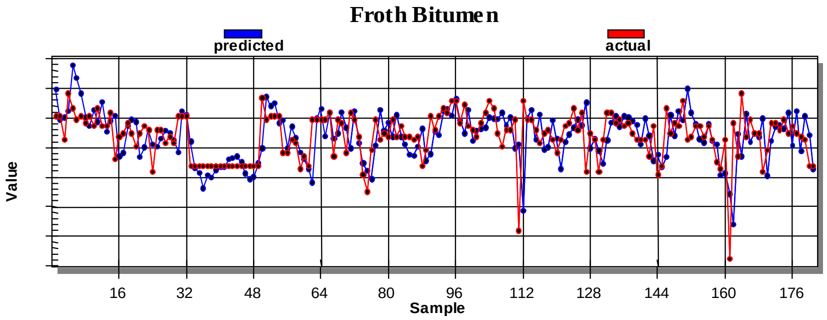

Fig 3. Predictions of Froth Bitumen using PLS Regression

T_1 and T_2 refer to times when the vessel was removed for analysis and $t_1, t_2, \ldots, t_k$, refer to the times when the trigger mechanism was engaged. Then, assuming that equal volumes of the product were sampled at the sample instant, the following equation holds approximately:

$$Y_{av} \approx \frac{1}{k} \sum_{t_i=t_1}^{t_k} Y(t_i)$$

Under the assumption that the process can be represented well using a linear static model of the form:

$$\begin{aligned} Y(t_i) = {} & a_1 u_1(t_i - t_{d1}) + a_2 u_2(t_i - t_{d2}) \\ & + \ldots + a_m u_m(t_i - t_{dm}) \end{aligned}$$

where, $a_1, \ldots, a_m$ are the static regression coefficients of the m input variables $u_1, \ldots, u_m$ and the d_i is the time delay between the i^{th} input and the output, we get the expression:

$$\begin{aligned} Y_{av} \approx \frac{1}{k} \{ & a_1 \sum_{t_i=t_1}^{t_k} u_1(t_i - t_{d1}) + \ldots \\ & + a_m \sum_{t_i=t_1}^{t_k} u_m(t_i - t_{d1}) \} \end{aligned}$$

Hence time-averaging can be used to mimic the sample consolidation mechanism.

4.2 Large sampling intervals and effect on data size

Another challenge is in the large sampling times for the quality variable. The sampling time for the froth bitumen is 12 hours. This means that even data collected over the course of a few months would yield very few values for the froth bitumen. For example we obtained only 60 samples over 30 days. In addition the ratio of sampling time of the process variables to that of the quality variable is 720. Developing multi-rate models with such large sampling ratios given that we have 25 inputs, is not practical. For static regression problems where we are interested in capturing spatial relationships between different variables rather than temporal relationships, we can use the data at the slow sample rates. This is the procedure adopted in the models developed in this exercise. As pointed before this reduces the number of samples available for modelling.

4.3 Using interpolated process data while identifying dynamic models

In problems where dynamic models are required, it has been pointed out in chemical engineering literature that one can use simple interpolation devices such as linear interpolation provided the measurements are not very noisy (Amirthalingam *et al.*, 2000). However, it is important to realize the potential dangers in using such interpolation devices. These interpolation devices introduce additional data where there is none. Hence the identification problem becomes one of identifying "correct" models from "wrong" data. The problem with ZOH interpolation is that, when the ZOH interpolation device is used, the output remains flat till the next sample arrives even though there might be changes in the inputs. The use of linear interpolation is generally accepted in the modelling phase even though it is a non-causal operation because it is carried out as an off-line exercise. However, the use of linear interpolation could lead to the identification of non-causal models for the particular input-output set considered. This is because the output starts to move in the direction of the next value even before the input starts moving. When using routine operating data for identification, there may be feedback induced (controller) correlations hidden in the data. In these correlations, the output is the cause and the manipulated input is the effect. Hence the coefficients being identified may be those of the controller rather than those related to the process. One may be further misled by the fact that the predictions of these models are quite good. Hence it is important to supplement and validate the results of "black-box" identification approaches using process knowledge of gain directions.

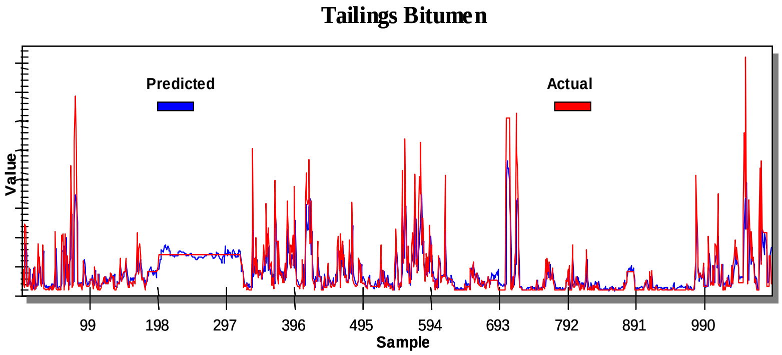

Fig 4. Predictions of Tailings Bitumen using PLS Regression

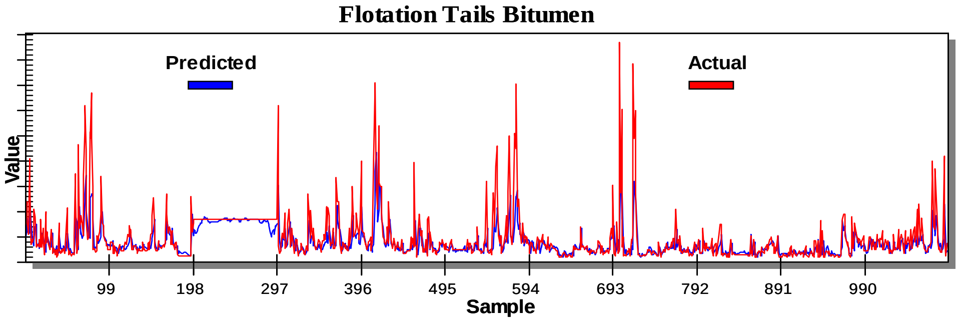

Fig 5. Predictions of Flotation Tails Bitumen using PLS Regression

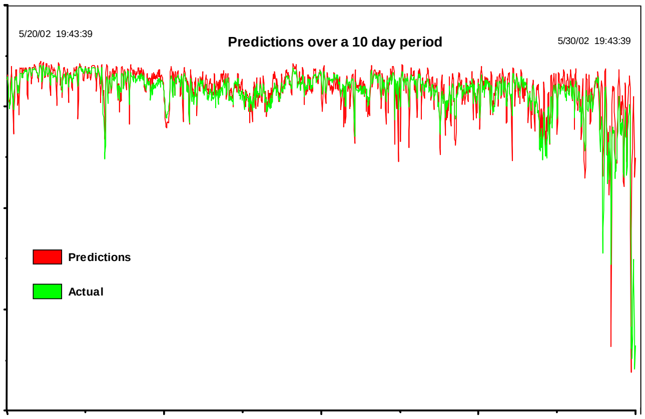

Fig 6. Online prediction of Bitumen Recovery

4.4 Estimating time delays in industrial processes

The problem of time-delay estimation was found to be particularly challenging. This is tackled in the identification literature using correlation analysis or by looking at cross-correlation between variables at different lags. However these are very difficult to apply in practice because of the non-stationarity of the signals, the multivariate nature of process data and correlations induced by operational and control strategies. In practice, using transport lags obtained from process knowledge or specific tests gives more reliable results. In this exercise we estimated the time delays using our knowledge of the physical locations of the sensors while making sure that material recycles were taken into account. We have assumed the transport delays to be constant. Hence, the variation in these transport lags due to varying throughput is of concern. It is not easy to fix this problem in the current framework and hence we have not attempted it here. However the problem of time delay estimation from routine operating data is an important problem which needs to be addressed by the chemical process community.

4.5 Nonlinear transformations

While developing models of systems using linear regression it is desirable to have normally distributed errors affecting the system and a linear relationship between the variables in the system. However, in practice these conditions may not hold. For example, the presence of a nonlinear relationship between the dependent and independent variables, or non-normality of the independent variables or the errors manifests itself as non-normality of the dependent variable. Hence it is important to check whether it might be inappropriate to identify a standard linear model using a given set of data. If non-linearity is suspected, we may need to use suitable transformations of the variables to coax the dependent variable to normality or to produce a linear relationship between X and Y. A dependent variable may not be normally distributed if its values are bounded, creating a skewed distribution. When it comes to inference of parameters from regression, it is important to ensure that the errors are normally distributed. A non-normal dependent variable does not necessarily mean a non-normal distribution of errors. However, the converse is often encountered. This argument is also supported by the common practice of drawing conclusions about the error distribution from the distribution of the residuals. When the dependent variable is found to be non-normal, one may consider using transformations to normalize the dependent variable. A few common transformations that can be used for dependent variables, include the logarithmic ($Z = log(Y)$), exponential ($Z = e^Y$), power ($Z = Y^p$) and logistic ($Z = \frac{log(Y)}{1-log(Y)}$) transformations.

For the Bitumen recovery separation cell, the distributions of two of the quality variables show significant deviation from normality. They are the Bitumen concentrations in the Tailings and Flotation tailings. These quality variables take non-negative values which are generally low, except during upsets, which are characterized by large spikes in these variables. Performing linear regression without transformation leads to poor prediction of these spikes. Due to the nature of the distribution a specific nonlinear transformation was applied on these dependent variables which led to a significant improvement in the quality of the predictions.

5. ONLINE RESULTS

The results of the predictions are shown in Fig. 3, 4 and 5, from which it is clear that there is great potential for the use PLS regression for predicting bitumen recovery.

The soft sensors developed using PLS regression have been implemented online in Suncor Extraction's Distributed Control System (DCS) and their Plant historian (Fig. 6) and the results are encouraging. These predictions are being used for monitoring the bitumen recovery in the separation cell. The plant personnel are happy to have a simple tool which gives them advance warning of a fall in the recovery and are satisfied with the performance of the soft-sensors.

6. CONCLUDING REMARKS

An industrial application of PLS regression techniques for developing soft-sensors for predicting infrequently measured quality variables in a Bitumen extraction process has been described. Some of the challenges in applying these techniques to industrial problems have been presented with some proposed solutions.

7. ACKNOWLEDGEMENTS

Financial support from NSERC, Matrikon Inc. and ASRA in the form of the Industrial Research Chair programme is gratefully acknowledged.

REFERENCES

Amirthalingam, R., S. W. Sung and J. H. Lee (2000). A two step procedure for data-based modeling for inferential predictive control system design. *AIChE Journal* **46**, 1974–1988.

Kresta, J. V., J. F. MacGregor and T. E. Marlin (1991). Multivariate statistical monitoring of processes. *Can. J. Chem. Eng.* **69**(1), 35–47.

Kresta, J. V., T. E. Marlin and J. F. MacGregor (1994). Development of inferential process models using PLS. *Computers Chem. Engng.* **18**, 597–611.

Ku, W., R.H. Storer and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **30**, 179–196.

Lakshminarayanan, S., S.L. Shah and K. Nandakumar (1997). Modelling and control of multivariable processes: The dynamic projection to latent structures approach. *AIChE Journal* **43**, 2307–2323.

Li, D., S.L. Shah and T. Chen (2002). Analysis of dual-rate inferential control systems. *Automatica* **38**(6), 1053–1059.

Nomikos, P. and J.F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics* **37**(1), 41–59.

Parrish, J.R. and C.B. Brosilow (1985). Inferential control algorithms. *Automatica* **21**(5), 527–538.

Qin, S. J. and T. McAvoy (1992). Nonlinear PLS modeling using neural networks. *Comput. Chem. Eng.* **16**(4), 379–391.

Qin, S.J. (1993). Partial least squares regression for recursive system identification. In: *Proceedings of the 32nd Conference on Decision and Control*.

Ricker, N. Lawrence (1988). The use of biased least-squares estimators for parameters in discrete-time pulse response models. *Ind. Eng. Chem. Res.* **27**, 343–350.