8th International IFAC Symposium on
Dynamics and Control of Process Systems
Preprints Vol.1, June 6-8, 2007, Cancún, Mexico

# TWO-DIMENSIONAL DYNAMIC PCA WITH AUTO-SELECTED SUPPORT REGION

**Yuan Yao, Ningyun Lu and Furong Gao***

*Dept. of Chemical Engineering*
*Hong Kong University of Science and Technology*
*Clear water bay, Kowloon, Hong Kong SAR, P. R. China*

Abstract: Support region selection is a key step of two-dimensional dynamic PCA modeling for batch process dynamics, as it can affect the accuracy of the model and the efficiency of monitoring and fault diagnosis. In this paper, an automatic method for support region selection is developed. This data-based method can be applied universally on different batch processes without any prior process knowledge. Simulation shows that developed method has good application potentials for both monitoring and fault diagnosis.


Keywords: Batch process, Modeling, Monitoring, PCA, Dynamic, 2-D, Support region.

## 1. INTRODUCTION

Nowadays, batch processes are widely applied in industrial manufacturing to manufacture high-value-added products. To ensure operation safety and product quality, the multivariate statistical monitoring methods, such as principal component analysis (PCA) and partial least squares (PLS), have been extended from continuous processes to batch processes for online monitoring (Nomikos and MacGregor, 1994; Nomikos and MacGregor, 1995; Wold, et al., 1996; Rännar, et al., 1998; Lu, et al., 2004). However, these methods all assume batch independency for the process. In fact, dynamics are inherent characteristics of batch processes. In some cases, such dynamics exist not only within a particular batch, but also from batch to batch. Causes of batch-wise dynamics, for example, are slow property changing of feed stocks, process characteristics drift, process controller designed in a way of run-to-run adjustment, effects of slow response variables, and so on. All these are common in batch processes. To take process dynamics into consideration, several modeling methods have been developed. Batch dynamic principal component analysis (BDPCA) (Chen and Liu, 2002) captures within-batch dynamic information, while the lifted state space model (Lee and Dorsey, 2004) and the method incorporating prior batches information into MPCA model building (Flores-Cerrillo and MacGregor, 2004) concern batch-wise dynamics.

To build both types of dynamic information into a single model simultaneously, a two-dimensional dynamic principal component analysis (2-D-DPCA) model (Lu, et al., 2005) has been developed by the authors. This model has a parsimonious 2-D structure which is easy to build and can provide efficient online faults detection. Small changes in correlation or process drifts can both be detected effectively. However, a remaining problem is on the determination of support region for such a 2-D-DPCA model. In our previous work, the support region (ROS) is assumed to be limited in the quarter plane and have a regular shape. This assumption may not be reasonable for certain batch processes. In this paper, the problem of ROS determination for 2-D-DPCA model is presented and resolved.

The article is organized as following. In section 2, the problem of ROS selection is illustrated. Then a data-driven method that can determine the ROS automatically is presented in section 3. In the section

* To whom correspondence should be addressed:
Tel: (852)-2358-7139, Fax: (852)-2358-0054, Email: kefgao@ust.hk

4, simulations are given to compare the monitoring and diagnosis results between 2-D-DPCA models with quarter plane ROS and with auto-selected ROS. Finally, a conclusion is given in section 5 to summarize the paper.

## 2. PROBLEM DESCRIPTION

2-D structure has already been widely applied in data filtering, image processing and some other areas. 2-D-DPCA model is the first 2-D method to model batch processes for process monitoring. It combines lagged regression model structure and PCA method to capture both 2-D dynamics and cross-correlation information among variables. To build such a 2-D model, a key step is to choose the support region which is also called the region of support (ROS) or the prediction region. The ROS should be a proper selection of a neighborhood that can provide a good prediction. In batch processes, all relationships are causal, so this neighborhood is in the past of the current sample. In our previous work (Lu, et al., 2005), the ROS is assumed to be within a quarter plane and restricted in a rectangular shape, which means that the current value of a variable $x_j(i,k)$ is to be predicted by a region including its past values in time direction $x_j(i,k-1)$ , $x_j(i,k-2)$ , …, $x_j(i,k-n)$ , its past values in batch direction $x_j(i-1,k)$ , $x_j(i-2,k)$ , …, $x_j(i-m,k)$ , and together with the values in the cross direction $x_j(i-1,k-1)$ , …, $x_j(i-m,k-n)$ , where $x_j(i,k)$ is process measurement of variable *j* at sampling interval *k* in batch run *i* (*i=1,..,I; j=1,...J; k=1,...,K*). The order pair (*m*,*n*) of the support region was selected by calculating some index, such as Akaike information criterion (AIC), minimum descriptive length (MDL) and minimum eigenvalue (MEV) (Aksasse and Radouane, 1999), based on the implied equivalence between 2-D-DPCA model and 2-D AR model.

However, the quarter plane assumption of the ROS may not be an optimal or proper choice. Although the quarter plane is widely used in image processing and other areas, it dose not mean it is proper in batch process modeling. The ROS may not locate in the quarter plane. Take injection molding process as an example. The mold temperature is a slow response process variable. The temperature values of the future time intervals in the past batches can give better prediction of the current sample than the variable values in the quarter plane. This suggests that the ROS of a batch process may be irregular and complicated. Since batch process dynamics may be caused by different reasons and each process could have some special characteristics of its own, it is hard to use a uniform ROS shape for all batch processes. If process cross/auto-correlation characteristics are known, the support region can be chosen based on such prior knowledge. However, this prior information is unlikely available for most batch processes. So, a complete data-driven method for ROS determination is needed to determine automatically a proper ROS including its shape and orders.

## 3. METHOD FOR SUPPORT REGION (ROS) AUTO-SELECTION

A proper ROS should include the necessary past data to provide good prediction of variables' current values. In a sense, ROS determination is a problem similar to the task of selecting proper independent variables to predict current sample in a prediction model development. All past samples are candidates of such independent variables. The task is to choose a reasonable sub-group of them in a tidy structure to provide good prediction.

For model regression, many commercial programs, which use stepwise regression, have been developed to determine the regression variables according to certain criterion, such as AIC. Several methods have also been developed for variable selection when multicollinearity is present (Gauchi and Chagnon, 2001; Lazraq, et al., 2003; Chong and Jun, 2005).

Enlightened by the variable selection problem, a data-driven method for ROS auto-selection is designed. A neighborhood of the current sample is chosen as the candidate region of ROS which can be called initial ROS. All samples in this region can be regarded as candidate independent variables for the prediction model of current sample which is regarded as the dependent variable in the model. The initial ROS choosing should be sufficiently large to contain the proper ROS as a subset. Since the initial ROS is larger than the proper one, the target ROS can be obtained by eliminating unnecessary independent variables. Therefore, backward elimination is proposed to be conducted iteratively. In each run, a regression model is built to relate the remained candidate independent variables to current sample's value as the regression function. An index, which is a model evaluation criterion, is calculated at each iteration run. One independent variable is eliminated from the candidate region based on the model coefficients at each iteration. In the end, the best choice of the ROS is determined based on the comparison of the index values calculated during the iteration. Details of the method are given in the next.

For the initial ROS selection, the candidate region should be sufficiently large, so that the entire final ROS should completely be contained as pointed out earlier. On the other hand, it should not be too big in order to control the computation burden in the later steps. An initial ROS is proposed to satisfy the above requirements. A common property of the lagged variables in the final proper ROS is that all of them are significantly correlated with current sample so that they can provide good prediction. The correlation information is provided by the auto- and cross-correlation functions. When there is a significant correlation, the value of function is near one, otherwise, it towards zero. What we can do is to calculate the values of auto- and cross-correlation functions, and find out the past region that contains

significant correlations to the current sample and take this region as the initial ROS.

In iterative backward eliminations, one independent variable corresponding to the smallest regression coefficient is eliminated from the candidate region in each run. The basis of doing so is that the coefficient of each independent variable denotes its contribution to dependent variable if all variables are normalized. So in each run, a variable contributing least to the prediction can be eliminated from the candidate region. This iteration is done until there are only one variable retained.

An index is used as a criterion to evaluate the regressions in each iteration. This index could be AIC, MDL, or any other indices used in variable selection. The AIC criterion is proposed by Akaike (1974) and can be calculated from the equation below,

$$\mathrm{AIC}(k) = \log(\hat{\sigma}_k^2) + 2\frac{k}{N}, \qquad (1)$$

where $\hat{\sigma}_k^2$ is the prediction error variance estimated from the regression model, $k$ is the number of independent variable retained in the candidate region for building the regression, and $N$ is the number of observations. Based on AIC criterion, the smaller AIC($k$) is, the better is the prediction. So the retained independent variables and $k$ corresponding to the smallest AIC($k$) is chosen to constitute the proper ROS and the number of lagged variables in the ROS. One important issue should be noted is that AIC often leads to over-fitting that results in a bigger ROS than the proper one. So if the results of equation (1) decay slowly after some iterations, the lagged variables retained in the candidate region corresponding to that run may be selected as the most proper ROS.

The detailed procedures of ROS auto-selection are listed below.

1. In the past half plane, select a candidate region (initial ROS) which is large enough and includes the target ROS to be determined. This initial selection can be based on prior process knowledge or auto- and cross-correlation functions between current sample and past measurements as introduced before. Fig. 1 shows the initial ROS for process variable $x_j$ which can be an irregular shape in the half plane including data points of $x_j(i,k-1)$, $x_j(i,k-2)$, …, $x_j(i,k-n_0(j))$, $x_j(i-1,k+r_1(j))$, $x_j(i-1,k+r_1(j)-1)$, …, $x_j(i-1,k)$, $x_j(i-1,k-1)$, …, $x_j(i-1,k-n_1(j))$, …, $x_j(i-m(j),k+r_{m(j)}(j))$, …, $x_j(i-m(j),k)$, $x_j(i-m(j),k-1)$, …, $x_j(i-m(j),k-n_{m(j)}(j))$, where $x_j(i,k)$ means the value of variable $j$ at the $k$th sampling time interval in batch $i$, $m(j)$ is the maximum number of lagged batches of variable $j$ included in the initial ROS, $n_v(j)$ is the maximum number of past sampling intervals of variable $j$ in the

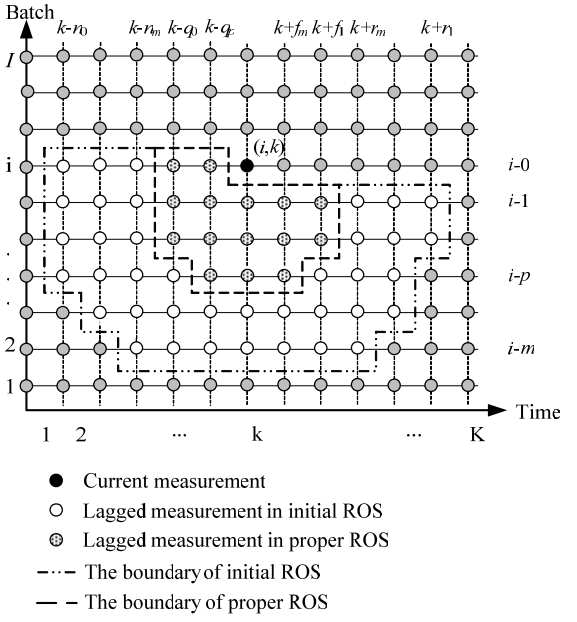


Fig. 1. Illustration of initial ROS and proper ROS

($i$-$v$)th batch included in the initial ROS, and $r_v(j)$ is the maximum number of future sampling intervals of variable $j$ in the ($i$-$v$)th batch included in the initial ROS.

2. Based on the chosen candidate region, augment data matrix as:

$$\mathbf{X} = \begin{bmatrix} X_{m+1,n+1} \\ \vdots \\ X_{m+1,K-r} \\ \vdots \\ X_{i,k} \\ \vdots \\ X_{I,K-r} \end{bmatrix}, \qquad (2)$$

where

$$X_{i,k} = [\mathbf{x}_1(i,k), \cdots, \mathbf{x}_j(i,k), \cdots, \mathbf{x}_J(i,k)]$$

$$\begin{aligned}\mathbf{x}_j(i,k) = [&x_j(i,k-1), \cdots, x_j(i,k-n_0(j)),\\ &x_j(i-1,k+r_1(j)), x_j(i-1,k+r_1(j)-1),\\ &\cdots, x_j(i-1,k), x_j(i-1,k-1),\\ &\cdots, x_j(i-1,k-n_1(j)),\\ &\cdots x_j(i-m(j),k+r_{m(j)}(j)),\\ &x_j(i-m(j),k+r_{m(j)}(j)-1),\\ &\cdots, x_j(i-m(j),k), x_j(i-m(j),k-1),\\ &\cdots, x_j(i-m(j),k-n_{m(j)}(j))]\end{aligned}$$

$$\begin{aligned}n = \max(&n_0(1), n_1(1), \cdots, n_m(1), \cdots, n_0(j), n_1(j),\\ &\cdots, n_m(j), \cdots, n_0(J), n_1(J), \cdots, n_m(J))\end{aligned}$$

$$\begin{aligned}r = \max(&r_1(1), r_2(1), \cdots, r_m(1), \cdots, r_0(j), r_1(j),\\ &\cdots, r_m(j), \cdots, r_0(J), r_1(J), \cdots, r_m(J))\end{aligned}$$

$$m = \max(m(1), \cdots m(j), \cdots m(J))$$

3. Normalize **X** and make each column has zero mean and unit variance.

4. Choose a variable $x_h$ where $h$=1,…,$J$ and $J$ is the number of process variables and augment data as:

$$\mathbf{Y_h} = \begin{bmatrix} x_h(m+1,n+1) \\ \vdots \\ x_h(m+1,K-r) \\ \vdots \\ x_h(i,k) \\ \vdots \\ x_h(I,K-r) \end{bmatrix}. \quad (3)$$

5. Normalize $\mathbf{Y_h}$ to zero mean and unit variance.

6. Take **X** as the matrix of independent variables, $\mathbf{Y_h}$ as the matrix of dependent variable, and build a regression model between **X** and $\mathbf{Y_h}$.

7. Calculate and store the value of evaluation index for this regression model. The index can be AIC (equation (1)) or other indices used in variable selection.

8. Eliminate a column in **X** corresponding to the smallest coefficient in the regression model.

9. Return to step 6 until all columns in **X** have been eliminated.

10. Compare all the values of evaluation index calculated and find the one indicating the most proper model. If AIC indices are used, to overcome the over-fitting problem, the one closer to the minimum value with fewer independent variables is selected.

11. Check **X** corresponding to the best model determined in step 10. The remaining columns of **X** give the most proper ROS of $x_h$. The selected proper ROS of $x_h$ is shown in Fig. 1. Similar to initial ROS, the proper ROS may have irregular shape in the past half plane, which includes $x_j(i,k-1)$, $x_j(i,k-2)$, …, $x_j(i,k-q_0(j))$, $x_j(i-1,k+f_1(j))$, $x_j(i-1,k+f_1(j)-1)$, …, $x_j(i-1,k)$, $x_j(i-1,k-1)$, $x_j(i-1,k-q_1(j))$, …, $x_j(i-p(j),k+f_{p(j)}(j))$, …, $x_j(i-p(j),k)$, $x_j(i-p(j),k-1)$, $x_j(i-p(j),k-q_{p(j)}(j))$, where $p(j)$ is the maximum number of lagged batches of variable $j$ included in the proper ROS, $q_v(j)$ is the maximum number of past sampling intervals of variable $j$ in the ($i$-$v$)th batch included in the proper ROS, and $f_v(j)$ is the maximum number of future sampling intervals of variable $j$ in the ($i$-$v$)th batch included in the proper ROS.

12. Return to step 4. Choose another $x_h$, and find its support region.

13. When every variable's support region is determined, the combination of them is the complete ROS to be used in 2-D-DPCA model building.

After the complete proper ROS is determined, the data matrix can be augmented as a combination of current sample and samples in the support region. Then 2-D-DPCA is performed on this augmented matrix. The formulas of PCA modeling, data reconstruction, *SPE* and corresponding control limits calculation have been reported in our previous work (Lu, et al., 2005).

## 4. SIMULATION RESULTS

In this section, 2-D-DPCA monitoring and diagnosis efficiency with auto-selected support region (ROS) is compared with a previous quarter plane ROS for a batch process with 2-D dynamics. The process model is given as below.

$$\begin{aligned} x_1(i,k) &= 0.5 * x_1(i\text{-}1,k+1) + 0.8 * x_1(i,k\text{-}1) \\ &\quad - 0.3 * x_1(i\text{-}1,k) + w_1 \\ x_2(i,k) &= 0.44 * x_2(i\text{-}1,k+1) + 0.67 * x_2(i,k\text{-}1) \\ &\quad - 0.11 * x_2(i\text{-}1,k) + w_2 \\ x_3(i,k) &= 0.4 * x_3(i,k\text{-}1) + 0.25 * x_1(i,k) \\ &\quad + 0.35 * x_2(i,k) + w_3 \\ x_4(i,k) &= 0.8 * x_4(i,k\text{-}1) + 0.53 * x_1(i,k) \\ &\quad - 0.33 * x_2(i,k) + w_4 \end{aligned} \quad , \quad (4)$$

where $i$ is the batch index; $k$ is the time index; $x_1$ and $x_2$ are two independent signals with dynamics in time and batch directions; $x_3$ and $x_4$ are function of $x_1$, $x_2$ and their own values at one step before in the current batch; $w_j\,(j=1,2,3,4)$ are Gaussian noises with variance 0.01. 100 batches data are generated and there are 200 samples in each batch.

The ROS is determined with the proposed auto-selection method. The complete ROS includes $x_1(i,k-1)$, $x_1(i-1,k+1)$, $x_1(i-1,k)$, $x_2(i,k-1)$, $x_2(i-1,k+1)$, $x_2(i-1,k)$, $x_3(i,k-1)$ and $x_4(i,k-1)$, which agree with the process. The 2-D AIC method is also used to do the same job based on the quarter plane assumption of the ROS. The quarter plane selected ROS is made up of $x_1(i,k-1)$, $x_1(i-1,k-1)$, $x_1(i-1,k)$, $x_2(i,k-1)$, $x_2(i-1,k-1)$, $x_2(i-1,k)$, $x_3(i,k-1)$, $x_3(i-1,k-1)$, $x_3(i-1,k)$, $x_4(i,k-2)$, $x_4(i,k-1)$, $x_4(i-1,k-2)$, $x_4(i-1,k-1)$ and $x_4(i-1,k)$.

Two 2-D-DPCA models are built based on these two different choices of ROS. To determine the number of PCs retained in the model, cross-validation method (Wold, 1978) is used. 4 PCs are retained in the 2-D-DPCA model with auto-selected ROS while 2-D-DPCA model with quarter plane support assumption retains 5 PCs. Both models can explain more than 99 percent of variations and there is no significant correlation but only noise retained in residuals since retained PCs capture most dynamics.

With the 2-D-DPCA model of normal operation, the control limits of *SPE*-statistic can be calculated based on a weighted $\chi^2$ distribution assumption and be used to monitor future batches.

To check the models' ability of monitoring and diagnosis, a change in variable correlation structure is generated to simulate the fault. From batch 61, variable $x_2$ is formulated as:

$$x_2(i,k) = 0.1 * x_2(i-1,k+1) + 0.67 * x_2(i,k-1) - 0.05 * x_2(i-1,k) + w_2 \quad (5)$$

From Fig. 2, it can be found that this change in magnitude is quite insignificant although there are some differences between the parameters of the two models.

The monitoring results are shown in Fig. 3 and 4, respectively for auto-selected ROS and quarter plane ROS. Though the fault is insignificant, both 2-D-DPCA methods can detected the fault efficiently from batch 61 within which batch the fault begins. But the model with auto-selected ROS dose the detection job much better while the model with quarter plane ROS only give alarms at a few samples. Besides this, the difference in fault diagnosis results tells us another

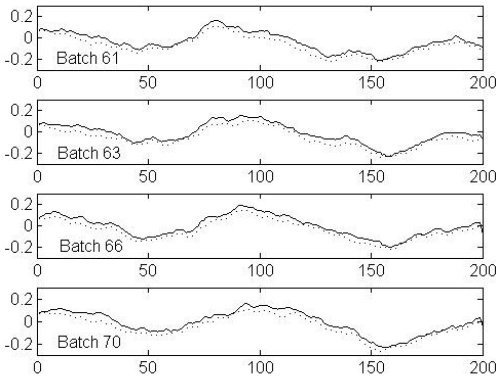


Fig. 2. Faulty trajectories of variable $x_2$ with changed correlation structure (Solid lines: normal trajectories; Dash lines: faulty trajectories)

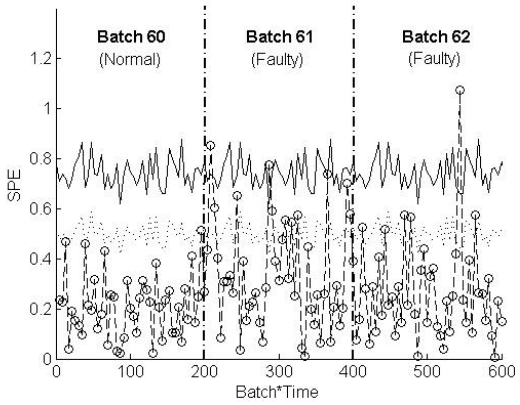


Fig. 3. Monitoring for a fault of correlation structure changing by 2-D-DPCA with auto-selected ROS (Solid line: 99% control limit; Dash line: 95% control limit)

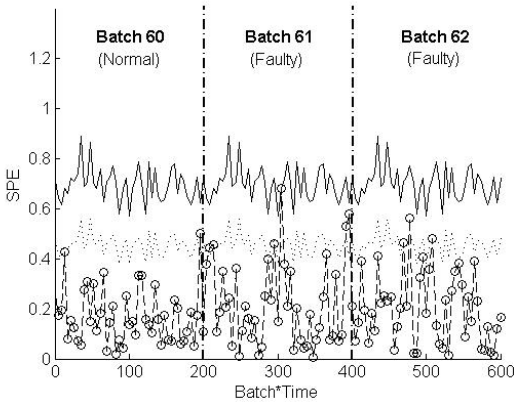


Fig. 4. Monitoring for a fault of correlation structure changing by 2-D-DPCA with quarter plane ROS (Solid line: 99% control limit; Dash line: 95% control limit)

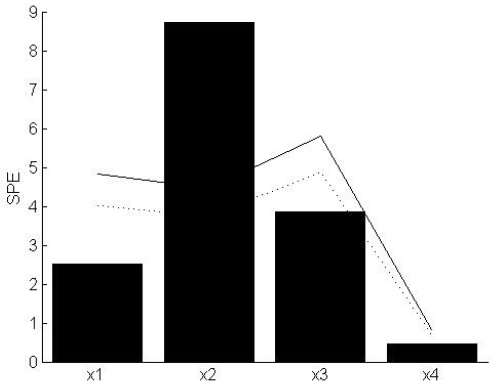


Fig. 5. *SPE* statistic contribution plot of Batch 61 based on 2-D-DPCA with auto-selected ROS (Solid line: 99% control limit; Dash line: 95% control limit)

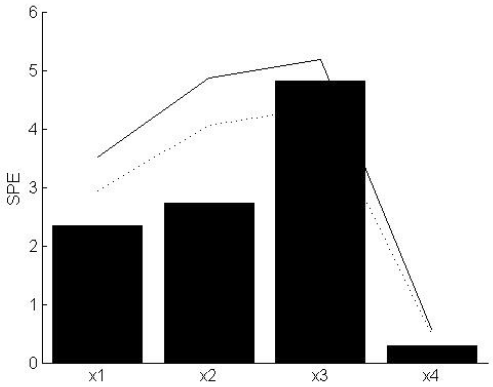


Fig. 6. *SPE* statistic contribution plot of Batch 61 based on 2-D-DPCA with quarter plane ROS (Solid line: 99% control limit; Dash line: 95% control limit)

important benefit of the 2-D-DPCA method with auto-selected ROS. The contribution plots for *SPE* of the batch 61 with control limits are drawn in Fig. 5 and 6, respectively for the method with auto-selected ROS and quarter plane ROS, based on the Westerhuis method (Westerhuis, et al., 2000). The 2-

D-DPCA model with auto-selected ROS shows clearly and correctly that the fault with variable $x_2$, while the one with quarter plane mistakenly takes $x_3$ as the faulty one.

In summary, incorrect structure of ROS can give distorted correlation information among process variables. This not only increases the chances of miss alarm, but also affects diagnosis results. 2-D-DPCA with auto-selected ROS solves such problems nicely, and this method is completely data-driven without the requirement of any prior process knowledge.

## 5. CONCLUSION

2-D-DPCA is a modeling method that can capture both time-wise and batch-wise dynamics of batch processes. Proper selection of support region (ROS) has shown to be important to monitoring efficiency and diagnosis accuracy. The idea of variable selection has been adopt in this work to solve the ROS selection problem nicely.

## REFERENCES


Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **19**, 716-723.

Aksasse B. and L. Radouane (1999). Two-dimensional autoregressive (2-D-AR) model order estimation. *IEEE Transactions on Signal Processing*, **47**, 2072-2077.

Chen J. and K.C. Liu (2002). On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chemical Engineering Science*, **57**, 63-75.

Chong I.G., and C.H. Jun (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, **78**, 103-112

Flores-Cerrillo J. and J.F. MacGregor (2004). Multivariate monitoring of batch processes using batch-to-batch information. *AIChE Journal*, **50**, 1219-1228.

Gauchi J. and P. Chagnon (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, **58**, 179-193.

Lazraq A., R. Cléroux, and J.P. Gauchi (2003) Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, **66**, 117-126.

Lee J.H. and A.W. Dorsey (2004). Monitoring of batch processes through state-space models. *AIChE Journal*, **50**, 1198-1210.

Lu N., F. Gao and F. Wang (2004). Sub PCA modeling and monitoring strategy for batch processes. *AIChE Jounal*, **50**, 255-259.

Lu N., Y. Yao and F. Gao (2005). Two-dimensional dynamic PCA for batch process monitoring. *AIChE Journal*, **51**, 3300-3304.

Nomikos P. and J.F. MacGregor (1994). Monitoring of batch processes using multiway principal component analysis. *AIChE Journal*, **40**, 1361-1375.

Nomikos P. and J.F. MacGregor (1995). Multiway partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, **30**, 97-108.

Rännar S., J.F. MacGregor and S. Wold (1998). Adaptive batch monitoring suing hierarchical PCA. *Chemometrics and Intelligent Laboratory systems*, **41**, 73-81.

Westerhuis J.A., S.P. Gurden, and A.K. Smilde (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, **51**, 95-114.

Wold S. (1978). Cross-validatory estimation of the number of components in Factor and principal components models. *Technometrics*, **20**, 397-405.

Wold S., N. Kettaneh and K. Tjessem (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, **10**, 463-482.